

# **REGRESIÓN APLICADA USANDO R**

**Edgar Acuña Fernandez**

**Departamento de Ciencias Matemáticas**

**Universidad de Puerto Rico**

**Recinto Universitario de Mayaguez**

**Enero 15, 2015**

**©2015, Derechos reservados por Edgar Acuña. Prohibida su reproducción sin  
permiso del autor**

## PREFACIO

La razón principal de escribir este libro es la carencia de un texto completo de regresión que cubra las diversas técnicas de regresión, especialmente aquellas que han tomado auge en la última década. Un par de buenos libros de regresión son el “Classical and Modern Regression with applications” de Myers y el “Applied linear Regression” de Weisberg, pero ambos cubren muy poco material u omiten temas importantes en regresión tales como: selección de variables, regresión logística, regresión robusta y la muy importante área de regresión no paramétrica. Existen por otro lado buenos textos cubriendo solamente Regresión Robusta como el “Robust Regression and Outlier Detection” de Rousseeuw y Leroy y otros que tratan exclusivamente Regresión no paramétrica como el “Applied Nonparametric Regression” de Haerdle. El objetivo de este texto es cubrir la parte más transcendental de los libros antes mencionados.

En el transcurso de los diez años que he venido desarrollando el texto he usado varios programas estadísticos tales como: MINITAB, SAS, MATLAB, S-PLUS y últimamente R. La meta final es desarrollar todo el texto usando el programa gratuito R, el cual está disponible en [www.r-project.org](http://www.r-project.org). Aún quedan en el texto algunas salidas de MINITAB. Las salidas de SAS, MATLAB y S-Plus han sido prácticamente eliminadas.

Aunque el texto es en regresión aplicada también se ha tratado de probar varias identidades y propiedades de estimadores que aparecen en regresión. Sin embargo no es nuestra intención llenar el texto con demostraciones teóricas. Dos buenos textos donde se ve el lado teórico de Regresión son “Linear Regression Analysis” de Seber y “Linear Statistical Inference and its Applications” de Rao.

El texto está organizado en 9 capítulos. El primer capítulo se enfoca en regresión lineal simple y el segundo en regresión lineal múltiple. En el tercer capítulo se discute los diversos métodos de diagnosticar si las suposiciones del modelo de regresión se están cumpliendo o no. En el capítulo 4 se estudian diferentes transformaciones que se pueden hacer de las variables predictoras y de la variable de respuesta con la finalidad de mejorar el modelo de regresión para que haga un mejor ajuste de los datos. En el capítulo 5 se discute modelos de regresión considerando la presencia de variables categóricas. Aquí se incluye el estudio de la regresión logística. El capítulo 6 está dedicada al importante problema de selección de variables en regresión y en el problema 7 se discute la forma de detectar y resolver el problema de multicolinealidad entre las variables predictoras. Los capítulos 8 y 9 están dedicados a regresión robusta y regresión no paramétrica respectivamente.

Los conjuntos de datos que aparecen en este texto pueden ser obtenidos en el siguiente sitio de la internet en [academic.uprm.edu/eacuna/class6205.htm](http://academic.uprm.edu/eacuna/class6205.htm).

Finalmente, deseo agradecer la ayuda de mi pasados asistentes de investigación por colaborar conmigo en la depuración de errores presentes en el texto, así como en la edición de algunos capítulos y en la preparación de las transparencias del texto.

Por favor para reportar cualquier sugerencia o error mandarme un e-mail a [edgar.acuna@upr.edu](mailto:edgar.acuna@upr.edu).

Mayagüez, Enero 15, 2015

# CONTENIDO

1	Regresión lineal simple.....	1
1.1	Introducción.....	1
1.1.1	Usos del Análisis de Regresión.....	5
1.2	El modelo de Regresión Lineal Simple.....	5
1.2.1	Estimación de la línea de regresión usando mínimos cuadrados.....	6
1.2.2	Interpretación de los coeficientes de regresión estimados.....	9
1.2.3	Propiedades de los estimadores minimos cuadraticos de regression.....	9
1.2.4	Distribución de los estimadores minimos cuadraticos.....	11
1.2.5	Propiedades de los residuales.....	11
1.2.6	Estimación de la varianza del error.....	12
1.2.7	Descomposición de la suma de cuadrados.....	13
1.2.8	El coeficiente de Determinación $R^2$ .....	16
1.3	Inferencia en Regresion Lineal Simple.....	16
1.3.1	Inferencia acerca de la pendiente y el intercepto usando la prueba t.....	17
1.3.2	El análisis de Varianza para regresión lineal simple.....	20
1.3.3	Intervalo de predicción e intervalo de confianza para el valor medio de la variable de respuesta.....	21
1.4	El coeficiente de Correlación.....	24
1.5	Análisis de Residuales.....	27
1.5.1	Cotejando Normalidad en los errores y detectando outliers.....	28
1.5.2	Cotejando que la varianza sea constante.....	30
1.5.3	Cotejando si los errores están correlacionados.....	32
2	Regresión Lineal Múltiple.....	41
2.1	Introducción.....	41
2.2	El Modelo de Regresión lineal múltiple.....	46
2.2.1	Estimación de B por minimos cuadrados.....	47
2.2.2	Propiedades del estimador $\hat{\beta}$ .....	48
2.2.3	Estimación de la varianza $\sigma^2$ .....	49
2.3	Inferencia en regresión lineal múltiple.....	51
2.3.1	Prueba de hipótesis acerca de un coeficiente de regresión individual.....	51
2.3.2	Prueba de Hipótesis de que todos los coeficientes de regresión sean ceros.....	52
2.3.3	Prueba de hipótesis para un subconjunto de coeficientes de regresión.....	54
2.3.4	Intervalo de Confianza y de Predicción en Regresión Lineal Múltiple.....	56
2.3.5	La prueba de Falta de Ajuste.....	57
3	Anomalías en regresión y medidas remediales.....	64
3.1	“Outliers”, puntos de leverage alto y valores influyentes.....	64
3.2	Residuales y detección de outliers”.....	67
3.2.1	Media y Varianza del vector de residuales.....	67
3.2.2	Residuales Estudentizados internamente.....	68
3.2.3	Residuales Estudentizados externamente.....	70
3.3	Diagnósticos para detectar “outliers” y puntos de leverage alto.....	75
3.4	Plot de Residuales para detectar el efecto de variables y casos influyentes.....	79

3.5	Plot de Residuales para detectar Normalidad.....	82
3.6	Detectando varianza no constante.....	85
3.7	Errores correlacionados en regresión.....	84
4	Transformaciones en Regresión.....	92
4.1	Transformaciones para linealizar modelos.....	92
4.2	Transformaciones para estabilizar la varianza .....	95
4.3	Transformaciones de las variables predictoras en regresión multiple .....	98
4.4	Transformaciones para mejorar la normalidad de la variable de respuesta.....	104
4.5	Mínimos cuadrados ponderados.....	108
4.6	Mínimos cuadrados generalizados .....	113
5	Regresión con variables cualitativas.....	117
5.1	Regresión con variables predictoras cualitativas.....	117
5.1.1	Regresión con una sola variable cualitativa.....	117
5.1.2	Comparando las líneas de regresión de mas de dos grupos.....	121
5.2	Regresión Logística.....	121
5.2.1	Estimación del modelo logístico.....	126
5.2.2	Medidas de confiabilidad del modelo.....	127
5.2.3	Medidas influenciales para regresión logística.....	128
5.2.4	Uso de regresión logística en clasificación.....	132
6	Selección de variables en Regresión.....	136
6.1	Métodos “Stepwise”.....	136
6.1.1	“Backward Elimination” (Eliminación hacia atrás).....	136
6.1.2	“Forward Selection” (Selección hacia adelante).....	137
6.1.3	“Stepwise Selección” (Selección Paso a Paso).....	138
6.2	Método de los mejores subconjuntos.....	142
6.3	Criterios para elegir el mejor modelo.....	142
6.3.1	El coeficiente de Determinación $R^2$ .....	142
6.3.2	El $R^2$ ajustado.....	143
6.3.3	La varianza estimada del error ( $s^2$ ).....	143
6.3.4	$C_p$ de Mallows.....	143
6.3.5	PRESS. Suma de cuadrados de Predicción.....	146
6.3.6	Validación Cruzada .....	148
6.3.7	AIC .....	149
6.3.8	BIC .....	154
6.3.9	Validación cruzada Generalizada.....	156
6.3.10	Otros Criterios.....	157
6.3.11	Recomendación para elegir el mejor modelo.....	157
6.4	Otros métodos de selección de variable.....	158
6.4.1	Métodos basados en remuestreo.....	158
6.4.2	Métodos basados en regresión penalizada.....	158
6.4.3	Métodos Bayesianos.....	159
6.4.4	Algoritmos Genéticos.....	159
7.	Multicolinealidad .....	161
7.1	Multicolinealidad.....	161
7.1.1	Efectos de Multicolinealidad.....	161
7.1.2	Diagnósticos de Multicolinealidad.....	163

7.1.3	Medidas remediales al problema de multicolinealidad.....	166
7.2	Regresión Ridge.....	166
7.2.1	Aplicación de Regresión Ridge a Selección de variables.....	174
7.3	Componentes principales para Regresión.....	176
8	Regresión Robusta.....	186
8.1	Introducción.....	186
8.2	Regresión L1.....	187
8.3	Regresión M.....	190
8.3.1	Cálculo de los estimadores M de regresión.....	198
8.4	Regresión GM o Regresión de Influencia acotada.....	201
8.5	Regresión de Medianas de Cuadrados Mínima.....	202
9	Regresión Noparamétrica.....	206
9.1	Introducción.....	206
9.2	Suavización bivariada o Suavizadores de diagramas de puntos .....	207
9.2.1	El regresograma.....	207
9.2.2	“Running Means” y “Running Lines”.....	208
9.2.3	Suavizador por los k vecinos más cercanos.....	210
9.2.4	Suavización por kernels.....	210
9.2.5	Regresión local ponderada, LOWESS.....	212
9.2.6	Regresión Polinomial .....	213
9.2.7	Regresión por Splines.....	215
9.2.8	Suavización por Splines.....	217
9.3	Suavización multidimensional.....	220
9.3.1	Modelos Aditivos generalizados, GAM.....	220
9.3.2	Regresión usando árboles de decisión (CART).....	222
	Apéndice A: Revisión de Matrices.....	231
	Apéndice B: Laboratorios en R para Regresión Aplicada.....	240
	Referencias.....	265

# CAPÍTULO 1

## REGRESIÓN LINEAL SIMPLE

### 1.1. Introducción

**Regresión** es un conjunto de técnicas que son usadas para establecer una relación entre una variable cuantitativa llamada *variable dependiente* y una o más variables independientes llamadas *variables predictoras*. Las variables independientes también deberían ser cuantitativas, sin embargo es permitido que algunas de ellas sean cualitativas. La ecuación que representa la relación es llamada el **modelo de regresión**. Si todas las variables independientes fueran cualitativas entonces el modelo de regresión se convierte en un modelo de **diseños experimentales**.

Ejemplos de modelos de regresión:

- La variable de respuesta puede ser la tasa de divorcio en tanto que una variable predictora puede ser el nivel de ingreso familiar.
- El precio de una casa puede ser la variable dependiente mientras que el área, el número de cuartos, el número de baños, y los años de antigüedad de la casa pueden ser usadas como variables predictoras.

Para estimar la ecuación del modelo se debe tener una muestra de entrenamiento. En el caso de una sola variable independiente, esta muestra consiste de  $n$  pares ordenados  $(x_i, y_i)$  para  $i=1, \dots, n$ . En el caso de varias variables independientes se deben tener  $n$  nuplas  $(\mathbf{x}_i, y_i)$ , para  $i=1, \dots, n$ , donde  $\mathbf{x}_i$  es el vector de mediciones de las variables predictoras para la  $i$ -ésima observación.

**Ejemplo 1.** En la siguiente tabla se muestra la tasa de mortalidad infantil (muertes de niños de 5 años o menos por cada 1,000 nacidos vivos) y el porcentaje de vacunación en veinte países del mundo. Los datos fueron tomados de un reporte de la UNICEF del año 1994.

	NACION	%INMUNIZACION	TASA_mor
1	"Bolivia"	77	118
2	"Brazil"	69	65
3	"Cambodia"	32	184
4	"Canada"	85	8
5	"China"	94	43
6	"Czech_Republic"	99	12
7	"Egypt"	89	55
8	"Ethiopia"	13	208
9	"Finland"	95	7
10	"France"	95	9
11	"Greece"	54	9
12	"India"	89	124
13	"Italy"	95	10
14	"Japan"	87	6
15	"Mexico"	91	33
16	"Poland"	98	16
17	"Russian_Federation"	73	32
18	"Senegal"	47	145
19	"Turkey"	76	87
20	"United_Kingdom"	90	9

El objetivo es hallar una ecuación que represente lo más preciso posible la relación entre la variable independiente: el porcentaje de inmunización, y la variable dependiente: la tasa de mortalidad. La figura 1.1 muestra el plot de los datos, obtenido usando el programa R, el comando usado aparece en el laboratorio 1( ver apéndice del texto). El plot sugiere que hay una aceptable relación lineal entre las variables. Además, la tasa de mortalidad tiende a bajar a medida que aumenta el porcentaje de inmunización.

### Relacion de la tasa de mortalidad con el porcentaje de inmunizacion

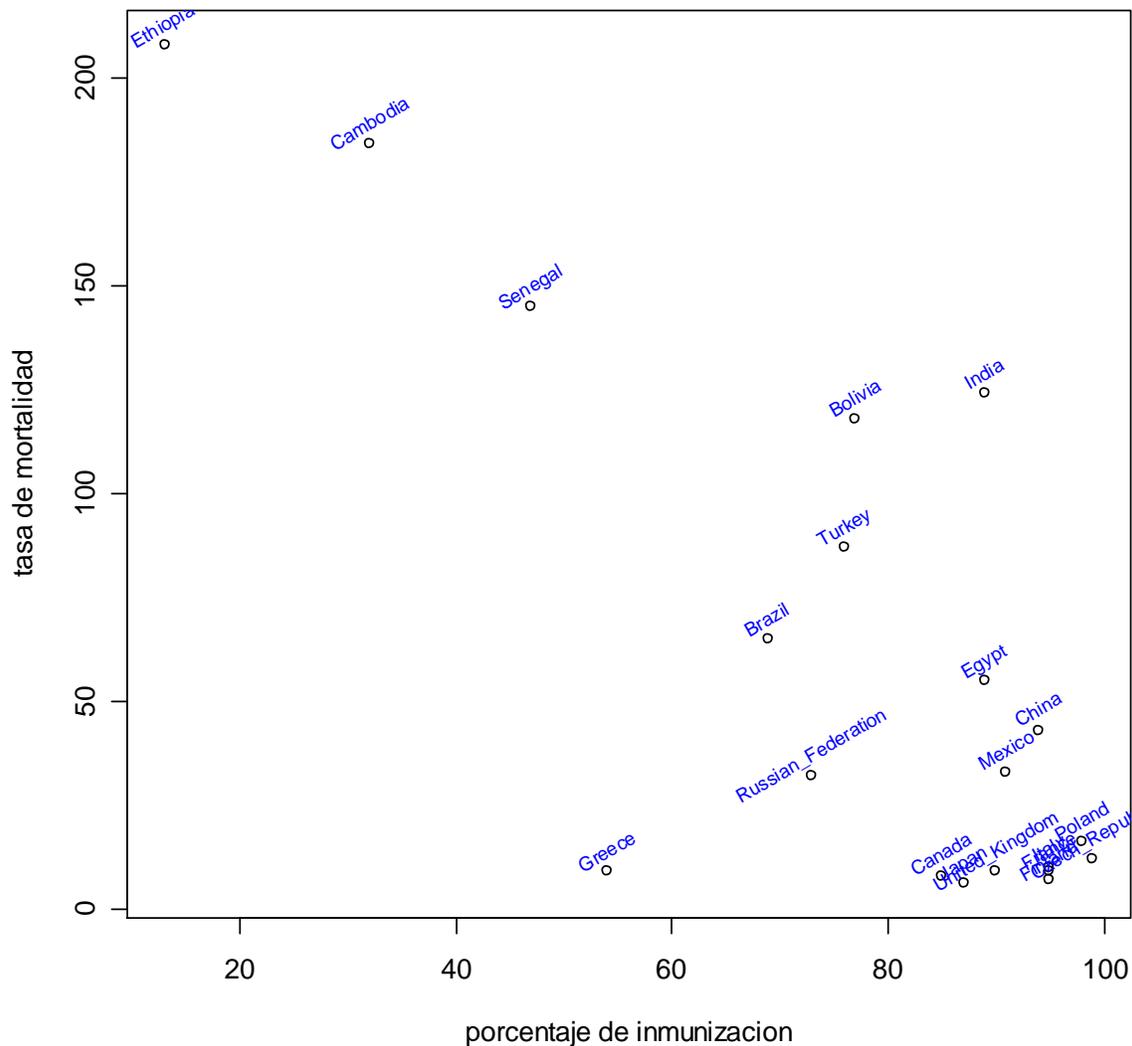


Figura 1.1 Plot que relaciona la tasa de mortalidad con el porcentaje de inmunización en cada país

De la figura 1.1 se puede ver que los países Ethiopia(8), Cambodia(3) y Senegal(18) parecen estar algo alejados de la mayoría de los datos. Igualmente, Greece(11) e India(12) aparecen algo fuera de la tendencia. No es muy obvio concluir que hay una relación lineal entre las variables. La figura 1.2 muestra la línea de regresión obtenida usando el programa R. Los comandos aparecen en el laboratorio 1 que aparece en el apéndice del texto.

La salida obtenida en R para la regresión lineal correspondiente es como sigue:

```
> l1<-lsfit(x,y)
> ls.print(l1)
Residual Standard Error=40.1393
R-Square=0.6258
F-statistic (df=1, 18)=30.1006
p-value=0
```

	Estimate	Std.Err	t-value	Pr(> t )
Intercept	224.3163	31.4403	7.1347	0
X	-2.1359	0.3893	-5.4864	0

De los resultados obtenidos, se tiene que la medida de confiabilidad del modelo, llamada **coeficiente de determinación (  $R^2$  )**, es sólo 62.58%, lo cual no es muy alto. Sin tomar en cuenta que esta medida se ve afectada por la presencia de los valores anormales, nos indica que la relación lineal entre las variables no es muy fuerte.

### Relacion de la tasa de mortalidad con el porcentaje de inmunización

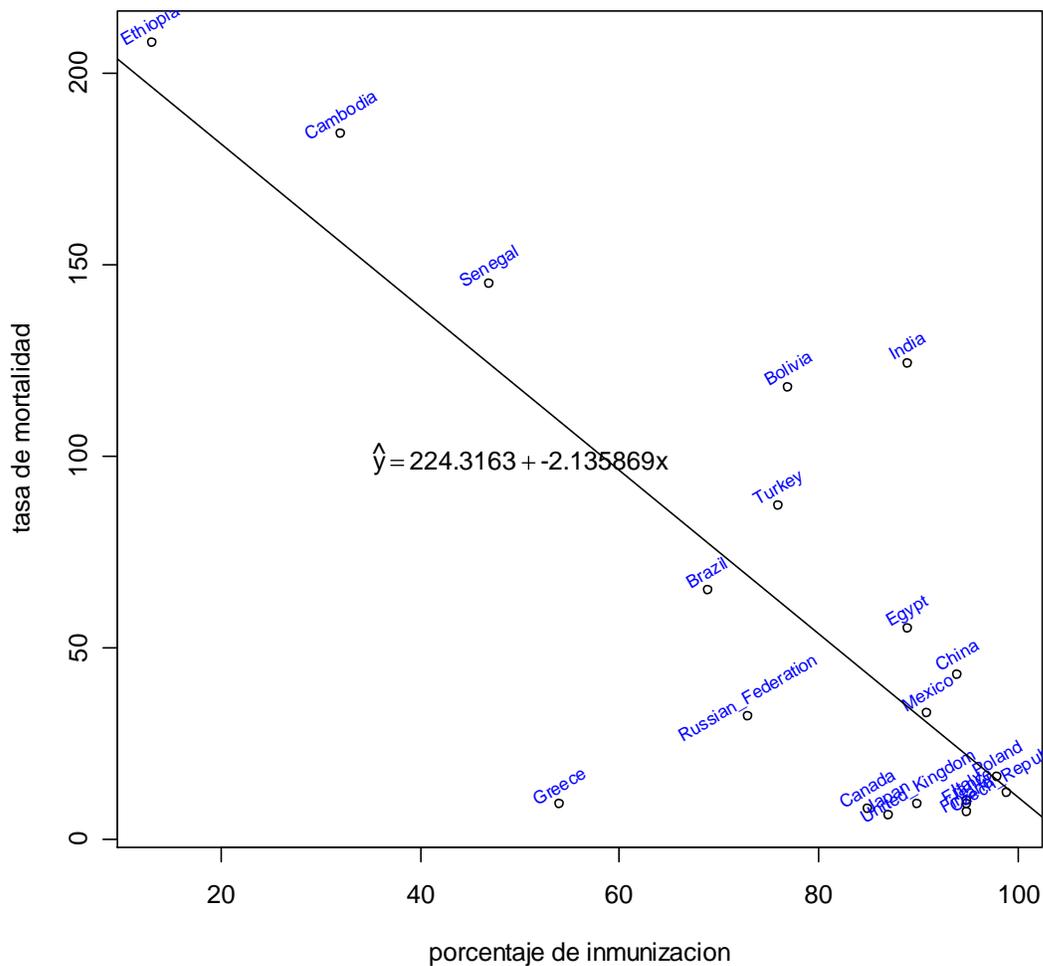


Figura 1.2 Línea de Regresión para los datos del ejemplo 1.

Si eliminamos las observaciones 11 y 12 la relación mejora notablemente, lo cual se puede ver en la siguiente salida de R

```
> l2<-lsfit(x1,y1)
> ls.print(l2)
Residual Standard Error=24.73
R-Square=0.8617
F-statistic (df=1, 16)=99.7027
p-value=0

      Estimate Std.Err t-value Pr(>|t|)
Intercept 251.4824 20.2188 12.4380    0
X         -2.4766  0.2480 -9.9851    0
```

Se observa que el  $R^2$  subió a un 86.2%, que es bastante aceptable. Asimismo en la figura 1.3 muestra la nueva línea de regresión que ajusta a mejor a los datos.

### Relacion entre tasa de mortalidad y porcentaje de inmunizacion

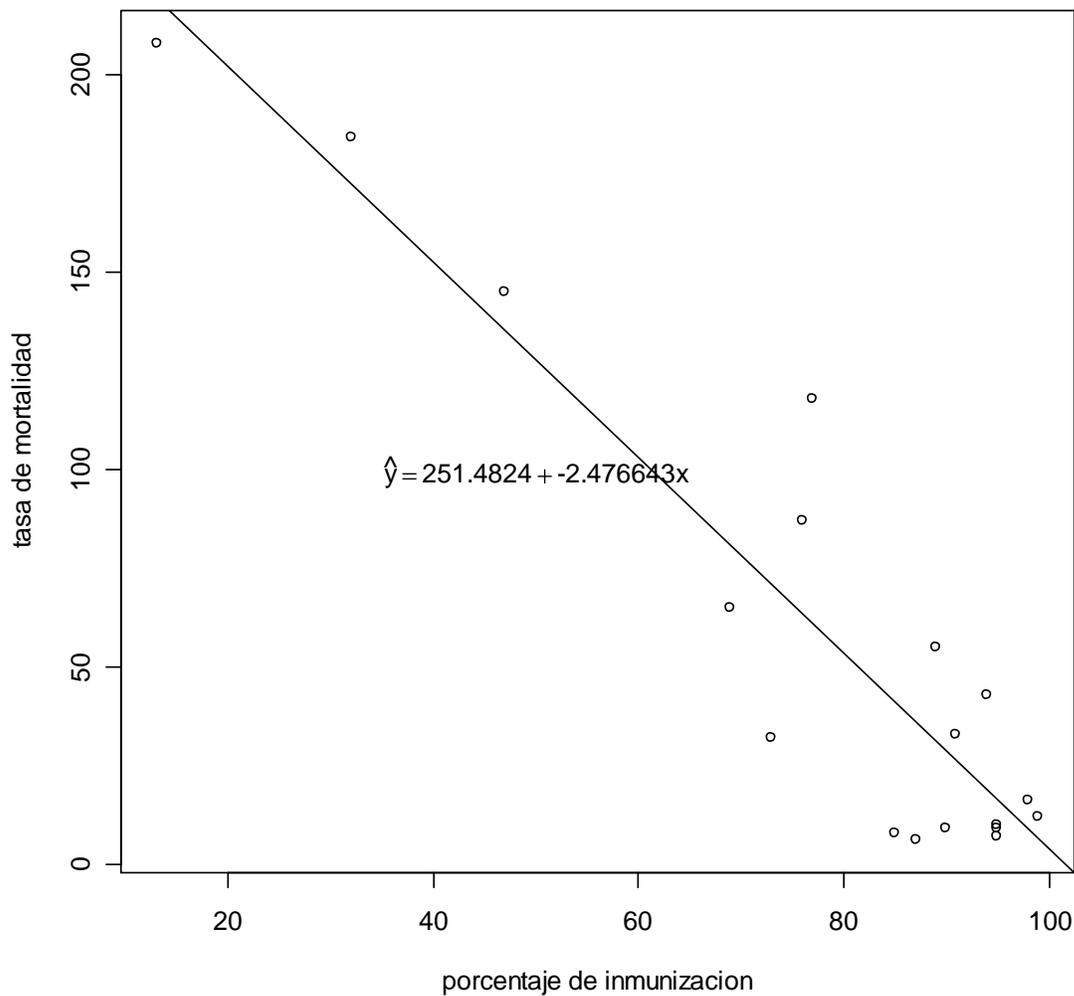


Figura 1.3 Línea de regresión después de eliminar las observaciones atípicas 11 y 12

El análisis de regresión es un proceso interactivo y el desarrollo de las computadoras en la última década ha facilitado e incentivado el uso de regresión en el análisis estadístico.

Regresión también es conocido como **Ajuste por cuadrados mínimos**, debido al método que se usa para estimar el modelo de regresión. Cuadrados Mínimos es acreditado a Karl Gauss y data desde los inicios de 1800. El nombre regresión fue introducido por Francis Galton a finales de 1800 cuando trató de relacionar las alturas de hijos y padres.

### 1.1.1 Usos del análisis de regresión:

Los siguientes son los principales usos de un modelo de regresión, aunque frecuentemente estos se dan al mismo tiempo en el análisis de un conjunto de datos:

- a) **Predicción:** El objetivo aquí es pronosticar valores de la variable de respuesta para valores futuros de la variables predictoras, es decir para valores más allá del rango de valores de las variables predictoras presentes en la muestra de entrenamiento. Tal vez ésta sea la razón principal para usar regresión en el análisis estadístico.
- b) **Descripción:** La idea es establecer una ecuación lineal o linealizable que describa la relación entre la variable dependiente y las variables predictoras.
- c) **Control:** Se busca controlar el comportamiento o variación de la variable de respuesta de acuerdo a los valores que asumen las variables predictoras. Por ejemplo, cuantas horas debería estudiar como mínimo un estudiante para sacar 90 puntos o más en un examen.
- d) **Selección de variables:** Inicialmente se pueden haber considerado muchas variables para explicar el comportamiento de la variable de respuesta a través de un modelo lineal, pero la presencia de muchas variables predictoras puede afectar el rendimiento del modelo además de que la computación del mismo se puede volver lenta. Por lo tanto, hay que usar técnicas para escoger solo las variables predictoras que sean más relevantes y aquellas que no sean redundantes en explicar la variación de la variable de respuesta.

## 1.2 El modelo de Regresión Lineal simple

En este caso se tiene una variable de respuesta o dependiente, denotada por  $Y$  y una sola variable predictora representada por  $X$ . El modelo de regresión lineal simple es de la forma

$$Y = \alpha + \beta X + \varepsilon \quad (1.1)$$

Aquí  $\alpha$  y  $\beta$  son el intercepto y la pendiente del modelo de regresión respectivamente y  $\varepsilon$  es un error aleatorio. El modelo es lineal porque la variable predictora no esta elevado a ninguna potencia o no es usada como argumento de otra funcion. Por otro lado, si se toma una muestra, que es representada por los  $n$  pares ordenados  $(X_i, Y_i)$  entonces el modelo se puede escribir como

$$Y_i = \alpha + \beta X_i + e_i \quad \text{para } i=1, \dots, n \quad (1.2)$$

Las constantes  $\alpha$  y  $\beta$  son los parámetros del modelo,  $e_i$  para  $i=1, 2, \dots, n$ , es una muestra aleatoria del error aleatorio  $\varepsilon$ , al igual que  $Y_i$  ( $i=1, 2, \dots, n$ ) es una muestra aletoria de la variable aletoria  $Y$ . Los parámetros  $\alpha$  y  $\beta$  son estimados en base a la muestra estimada y a la ecuación lineal

$$Y = \hat{\alpha} + \hat{\beta} X$$

es llamada la línea de regresión estimada.

En la figura 1.4 se muestra la línea de regresión estimada y los errores estimados para algunas de las observaciones de la muestra.

### Suposiciones del modelo:

- La variable predictora  $X$  es no aleatoria y se supone que ha sido medida con la mejor precisión posible. Sin embargo hay algunas situaciones donde también se supone que  $X$  es aleatoria.
- Los errores  $e_i$  son variables aleatorias con media 0 y varianza constante  $\sigma^2$ . Por ahora no se requerirá que los errores tengan una distribución normal.
- Los errores  $e_i$  y  $e_j$  ( $i \neq j = 1, \dots, n$ ) son independientes entre sí. Es decir,  $Cov(e_i, e_j) = 0$

Como en la ecuación del modelo solamente los  $e_i$ 's son aleatorios entonces las  $y_i$ 's deben tener también varianza constante  $\sigma^2$  y deben ser independientes por parejas.

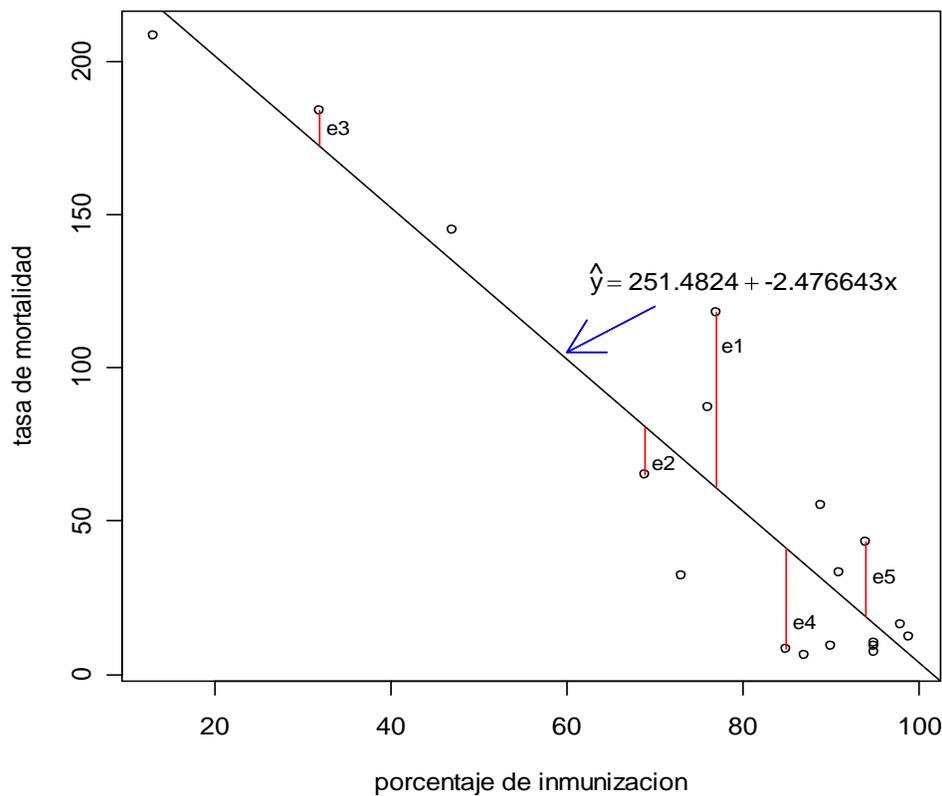


Figura 1.4 Errores con respecto a la línea de regresión para algunas de las observaciones del ejemplo 1

### 1.2.1 Estimación de la línea de regresión usando Mínimos Cuadrados

Si se toma el valor esperado de  $y_i$  para el valor  $x_i$  de  $x$  entonces de (1.2) se obtiene

$$E(y_i) = E(\alpha + \beta x_i + e_i) = \alpha + \beta x_i \quad (1.3)$$

O más específicamente que

$$E(y/x) = \alpha + \beta x \quad (1.4)$$

Es decir, la esperanza ( o media ) condicional de  $y$  dado que la variable predictora asume el valor de  $x$  es una ecuación lineal en  $x$ . La notación anterior es mas adecuada cuando se considera que  $x$  también es aleatoria. La ecuación (1.4) es llamada la línea de regresión poblacional.

Los parámetros  $\alpha$  y  $\beta$  deben ser estimados en base a la muestra tomada. El método usual para estimarlos es el de los cuadrados mínimos. La idea es minimizar la suma de los cuadrados de los errores  $e_i$ , con respecto a  $\alpha$  y  $\beta$ . Es decir,

$$Q(\alpha, \beta) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \quad (1.5)$$

Derivando parcialmente  $Q(\alpha, \beta)$  con respecto a  $\alpha$  y  $\beta$  e igualando a cero se obtienen las siguientes ecuaciones

$$\frac{\partial Q}{\partial \alpha} = -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \quad (1.6)$$

$$\frac{\partial Q}{\partial \beta} = -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) x_i = 0 \quad (1.7)$$

simplificando ambas ecuaciones se obtiene

$$n\alpha + \beta \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (1.8)$$

$$\alpha \sum_{i=1}^n x_i + \beta \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \quad (1.9)$$

este par de ecuaciones es conocido como las **ecuaciones normales del modelo**. Resolviendo este par de ecuaciones se obtiene que

$$\hat{\beta} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad (1.10)$$

lo cual es equivalente a  $\hat{\beta} = \frac{S_{xy}}{S_{xx}}$

donde:  $S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}$  es llamada la suma de productos corregida y

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \text{ es la llamada suma de cuadrados corregidos de X.}$$

De la primera ecuación normal es fácil ver que:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (1.11)$$

Por la forma de  $Q$ , es natural pensar que en el punto  $(\hat{\alpha}, \hat{\beta})$  hay un mínimo. Más formalmente, se podría aplicar el criterio de la segunda derivada para máximos y mínimos de la función bivariada. En este caso habría que cotejar que:

$$\frac{\partial^2 Q(\alpha, \beta)}{\partial^2 \alpha} > 0, \text{ y que } D = Q_{\alpha\alpha}(\alpha, \beta)Q_{\beta\beta}(\alpha, \beta) - (Q_{\alpha\beta}(\alpha, \beta))^2 > 0$$

$$\text{como } \frac{\partial^2 Q(\alpha, \beta)}{\partial^2 \alpha} = 2n > 0 \text{ y}$$

$$D = 4n \sum_{i=1}^n x_i^2 - 4 \left( \sum_{i=1}^n x_i \right)^2 = 4n \sum_{i=1}^n (x_i - \bar{x})^2 \geq 0, \text{ las condiciones requeridas se cumplen.}$$

Finalmente, la línea de regresión estimada, llamada también la línea ajustada, por cuadrados mínimos será:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x \quad (1.12)$$

Una vez que se ajusta la línea de regresión, el error aleatorio se vuelve un valor observado y es llamado **residual**, el cual es representado por  $r_i$  o por  $\hat{e}_i$ .

Sustituyendo el valor de  $\hat{\alpha}$  en la ecuación anterior se tiene:

$$\hat{y} = \bar{y} + \hat{\beta}(x_i - \bar{x}) \quad (1.13)$$

Esta ecuación puede ser considerada como la estimación de un modelo de regresión donde la variable predictora ha sido centrada.

También se puede usar el método de Máxima verosimilitud para estimar los coeficientes de la línea de regresión, pero se necesita considerar la suposición de que los errores aleatorios  $e_i$  se distribuyen normalmente con media cero y varianza  $\sigma^2$ . En este caso las estimaciones se obtienen maximizando la

función de verosimilitud  $L(\alpha, \beta, \sigma^2) = \prod_{i=1}^n f(y_i - \alpha - \beta x_i)$ , donde  $f$  representa la función de

densidad de una normal  $N(0, \sigma^2)$ . Sin embargo, es más frecuente trabajar con el logaritmo de la función de verosimilitud. Luego, se deriva  $\log L$  con respecto a  $\alpha$ ,  $\beta$ ,  $\sigma^2$  y se iguala a cero. Al resolver el sistema de ecuaciones se obtiene las estimaciones de los coeficientes  $\alpha$  y  $\beta$ , así como el estimador de la varianza  $\sigma^2$ .

Las propiedades de los estimadores mínimo cuadráticos de los coeficientes de regresión se discuten en las secciones 1.2.3 y 1.2.4.

### 1.2.2 Interpretación de los coeficientes de regresión estimados

La pendiente  $\hat{\beta}$  indica el cambio promedio en la variable de respuesta cuando la variable predictora aumenta en una unidad adicional. El intercepto  $\hat{\alpha}$  indica el valor promedio de la variable de respuesta cuando la variable predictora vale 0. Sin embargo, carece de interpretación práctica si es irrazonable pensar que el rango de valores de  $x$  incluye a cero. Cuando se tiene evidencia de que la variable de respuesta assume el valor 0 cuando la predictora es cero, entonces es más razonable ajustar una línea de regresión sin intercepto, véase ejercicio 4.

En el ejemplo 1, la ecuación de la línea de regresión estimada es

$$\text{Tasa\_mort} = 224.316 - 2.13587\% \text{inmuniz},$$

lo que significa que en promedio la tasa de mortalidad de niños menores de 5 años disminuirá en promedio en 2.13 cuando el % de inmunización aumenta en uno por ciento.

Por otro lado la tasa de mortalidad promedio de los países donde no hay inmunización será de 224.316. Aunque es difícil pensar que exista un país donde no se vacunen a los niños, ya que muchas veces la UNICEF dona las vacunas.

### 1.2.3 Propiedades de los estimadores mínimos cuadrados de regresión

a)  $\hat{\beta}$  es un estimador insegado de  $\beta$ . Es decir,  $E(\hat{\beta}) = \beta$

Recordar que:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad \text{Luego, como } x \text{ no es variable aleatoria y}$$

$E(y_i) = E(\alpha + \beta x_i + e_i) = \alpha + \beta x_i$ , por suposición b) del modelo, se obtiene que:

$$E(\hat{\beta}) = \frac{\sum_{i=1}^n (x_i - \bar{x})E(y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\alpha \sum_{i=1}^n (x_i - \bar{x}) + \beta \sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1.14)$$

como la suma de las desviaciones con respecto a la media es cero, se sigue que:

$$E(\hat{\beta}) = \beta \frac{\sum_{i=1}^n (x_i - \bar{x})x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta \frac{S_{xx}}{S_{xx}} = \beta$$

b)  $\hat{\alpha}$  es un estimador insegado de  $\alpha$ . Es decir,  $E(\hat{\alpha}) = \alpha$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (1.15)$$

Luego,

$$\begin{aligned} E(\hat{\alpha}) &= E(\bar{y}) - E(\hat{\beta})\bar{x} = E(\bar{y}) - \beta\bar{x} = \\ &= E\left(\frac{\sum_{i=1}^n y_i}{n}\right) - \beta\bar{x} = \frac{1}{n} \sum_{i=1}^n E(y_i) - \beta\bar{x} = \frac{1}{n} \sum_{i=1}^n (\alpha + \beta x_i) - \beta\bar{x} = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha \end{aligned} \quad (1.16)$$

c) La varianza de  $\hat{\beta}$  es  $\frac{\sigma^2}{Sxx}$  y la de  $\hat{\alpha}$  es  $\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right)$

Usando la propiedad que  $\text{Var}(cy) = c^2 \text{Var}(y)$  y el hecho de que la suposición de que  $\text{Cov}(e_i, e_j) = 0$  es equivalente a  $\text{Cov}(y_i, y_j) = 0$ , se tiene que  $\text{Var}\left(\sum_{i=1}^n c_i y_i\right) = \sum_{i=1}^n c_i^2 \text{Var}(y_i)$ . En consecuencia,

$$\text{Var}(\hat{\beta}) = \text{Var}\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \text{Var}(y_i)}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \sigma^2 \frac{Sxx}{(Sxx)^2} = \frac{\sigma^2}{Sxx} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1.17)$$

Por otro lado notar que  $\hat{\alpha}$  puede ser reescrita de la siguiente manera

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \sum_{i=1}^n \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{Sxx}\right] y_i. \quad (1.18)$$

Luego,

$$\text{Var}(\hat{\alpha}) = \sigma^2 \sum_{i=1}^n \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{Sxx}\right]^2 = \sigma^2 \sum_{i=1}^n \left[\frac{1}{n^2} - \frac{2\bar{x}(x_i - \bar{x})}{nSxx} + \frac{\bar{x}^2(x_i - \bar{x})^2}{(Sxx)^2}\right] \quad (1.19)$$

el segundo término de la suma se cancela y finalmente se obtiene que

$$\text{Var}(\hat{\alpha}) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2 Sxx}{(Sxx)^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{Sxx}\right] \quad (1.20)$$

Hay otra forma de calcular la varianza de  $\hat{\alpha}$ , usando el hecho que  $\text{Cov}(\bar{y}, \hat{\beta}) = 0$ , véase ejercicio 1. Las propiedades discutidas en esta sección serán usadas cuando se haga inferencia estadística para el modelo de regresión.

### 1.2.4 Distribución de los estimadores mínimos cuadráticos

Para efecto de hacer inferencia en regresión, se requiere asumir que los errores  $e_i$ , se distribuyen en forma normal e independientemente con media 0 y varianza constante  $\sigma^2$ . En consecuencia, también las  $y_i$ 's se distribuyen normalmente con media  $\alpha + \beta x_i$  y varianza  $\sigma^2$ .

En el cálculo de los valores esperados de  $\hat{\alpha}$  y  $\hat{\beta}$  se estableció que estos son una combinación lineal de las  $y_i$ 's. Esto es que  $\hat{\alpha} = \sum_{i=1}^n a_i y_i$  y  $\hat{\beta} = \sum_{i=1}^n b_i y_i$ . Por lo tanto, usando el hecho que una combinación lineal de variables aleatorias normales e independientes también se distribuye normalmente, y los resultados de la sección 1.2.3 se puede establecer que:

$$\begin{aligned} \text{i) } \hat{\beta} &\sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \\ \text{ii) } \hat{\alpha} &\sim N\left(\alpha, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right) \end{aligned} \quad (1.21)$$

### 1.2.5 Propiedades de los residuales

Los residuales  $r_i = y_i - \hat{y}_i$  son las desviaciones de los valores observados de la variable de respuesta con respecto a la línea de regresión estimada. Los residuales representan los errores aleatorios observados, y satisfacen las siguientes propiedades:

a) La suma de los residuales es 0. Es decir,  $\sum_{i=1}^n r_i = 0$

En efecto,  $\sum_{i=1}^n r_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i) = \sum_{i=1}^n y_i - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^n x_i = 0$ . La última igualdad se justifica por la primera ecuación normal.

b)  $\sum_{i=1}^n r_i x_i = 0$ . Similarmente, a la propiedad a) se tiene

$\sum_{i=1}^n r_i x_i = \sum_{i=1}^n (y_i - \hat{y}_i)x_i = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = \sum_{i=1}^n x_i y_i - \hat{\alpha}\sum_{i=1}^n x_i - \hat{\beta}\sum_{i=1}^n x_i^2 = 0$ . La última igualdad se justifica por la segunda ecuación normal.

c)  $\sum_{i=1}^n r_i \hat{y}_i = 0$ . Claramente,  $\sum_{i=1}^n r_i \hat{y}_i = \sum_{i=1}^n r_i (\hat{\alpha} + \hat{\beta}x_i) = \hat{\alpha}\sum_{i=1}^n r_i + \hat{\beta}\sum_{i=1}^n r_i x_i = 0$ . La última igualdad se justifica por a) y b).

### 1.2.6 Estimación de la varianza del error

La varianza del error, representada por  $\sigma^2$  es desconocida y debe ser estimada usando los residuales. Un estimador insesgado de  $\sigma^2$  es

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} \quad (1.22)$$

$s^2$  es llamado también **el cuadrado medio del error**. Existe una fórmula alterna para calcular  $s^2$ , pero esta será discutida más adelante cuando se haga el análisis de varianza para regresión simple.

#### Verificación de que $E(s^2) = \sigma^2$

En esta verificación consideraremos que  $(y_i - \hat{y}_i)$  es aleatoria,

Notar que

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)(y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{y}_i)y_i - \sum_{i=1}^n (y_i - \hat{y}_i)\hat{y}_i \quad (1.23)$$

Usando la propiedad c) de los residuales, la segunda de las sumas anteriores se cancela y usando las propiedades a) y b) se tiene que

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)y_i = \sum_{i=1}^n (y_i - \hat{y}_i)(\alpha + \beta x_i + e_i) = \sum_{i=1}^n (y_i - \hat{y}_i)e_i \quad (1.24)$$

Por otro lado,

$$(y_i - \hat{y}_i) = (\alpha + \beta x_i + e_i) - (\hat{\alpha} + \hat{\beta} x_i) = (\alpha - \hat{\alpha}) + (\beta - \hat{\beta})x_i + e_i \quad (1.25)$$

Asímismo,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = (\alpha + \beta\bar{x} + \bar{e}) - \hat{\beta}\bar{x} = \alpha + (\beta - \hat{\beta})\bar{x} + \bar{e} \quad (1.26)$$

Sustituyendo (1.26) en (1.25) se obtiene que

$$(y_i - \hat{y}_i) = (\beta - \hat{\beta})(x_i - \bar{x}) + e_i - \bar{e} \quad (1.27)$$

Reemplazando (1.27) en (1.24) se llega a

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)e_i = \sum_{i=1}^n [(\beta - \hat{\beta})(x_i - \bar{x})e_i + e_i^2 - e_i\bar{e}]$$

Tomado valores esperados en la última expresión y sustituyendo en la ecuación (1.23) se consigue

$$E\left[\sum_{i=1}^n (y_i - \hat{y}_i)^2\right] = \sum_{i=1}^n [(x_i - \bar{x})E((\beta - \hat{\beta})e_i) + E(e_i^2) - E(e_i\bar{e})] \quad (1.28)$$

Usando la suposiciones del modelo de regresión lineal es fácil ver que  $E(e_i^2) = \sigma^2$  y que

$$E(e_i\bar{e}) = E\left(e_i \frac{\sum_{j=1}^n e_j}{n}\right) = E\left(\frac{e_i^2}{n}\right) = \frac{\sigma^2}{n}. \quad (1.29)$$

Por otro lado, de la fórmula para  $\hat{\beta}$  se obtiene lo siguiente

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{j=1}^n (x_j - \bar{x})y_j}{S_{xx}} = \frac{\sum_{j=1}^n (x_j - \bar{x})(\alpha + \beta x_j + e_j)}{S_{xx}} = \beta + \frac{\sum_{j=1}^n (x_j - \bar{x})e_j}{S_{xx}}$$

Por lo tanto,

$$E[(\beta - \hat{\beta})e_i] = -\frac{E\sum_{j=1}^n (x_j - \bar{x})e_j e_i}{S_{xx}} = -\frac{(x_i - \bar{x})\sigma^2}{S_{xx}} \quad (1.30)$$

Finalmente, sustituyendo (1.29) y (1.30) en (1.28) se obtiene,

$$E\left[\sum_{i=1}^n (y_i - \hat{y}_i)^2\right] = -\frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{S_{xx}} + n\sigma^2 - \sigma^2 = (n-2)\sigma^2, \text{ con lo cual concluye la prueba.}$$

Una vez que se fija la línea de regresión y se estiman los errores por los residuales se tiene que un estimando de la varianza del error es

$$s^2 = \frac{\sum_{i=1}^n r_i^2}{n-2}$$

### 1.2.7 Descomposición de la suma de cuadrados total

Lo que se va hacer aquí es tratar de descomponer la variación total de Y en dos partes, una que se deba a la relación lineal de Y con X y otra a causas no controlables. Lo ideal es que gran parte de la variación de Y se explique por su relación lineal con X.

En la figura 1.5 se puede ver que la desviación de un valor observado  $y_i$  con respecto a la media  $\bar{y}$  se puede escribir como

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \quad (1.31)$$

Elevando al cuadrado en ambos lados de 1.31 y sumando sobre todas las observaciones se obtiene

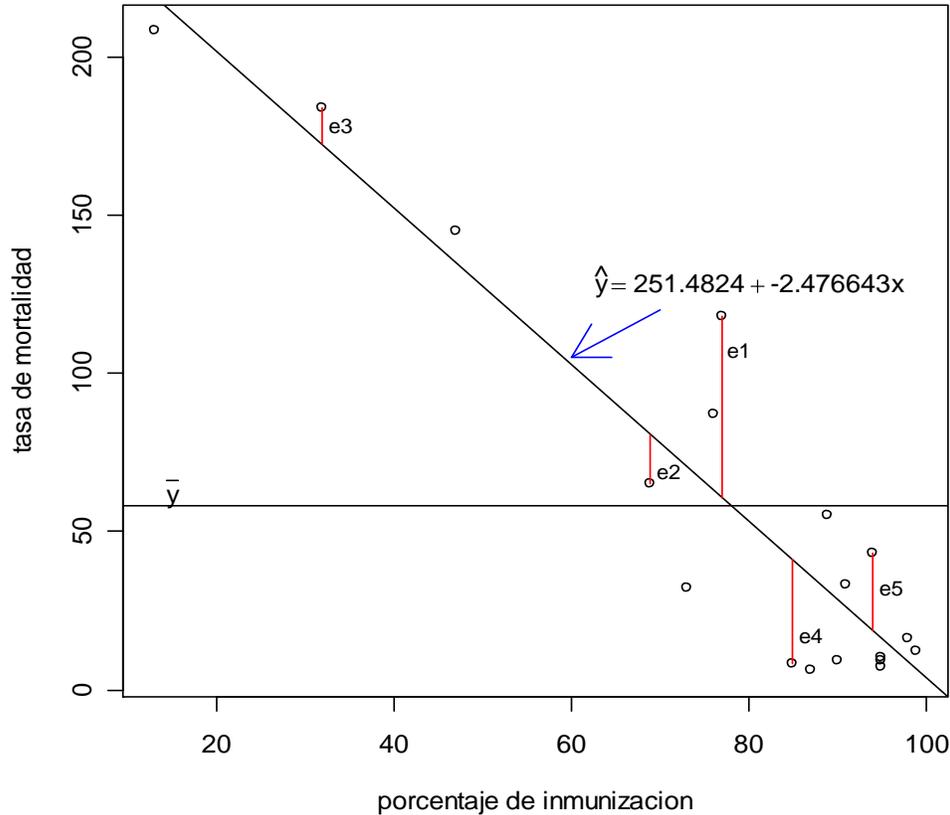


Figura 1.5 Diagrama para descomponer la desviación total en desviación debido a la regresión más desviación debido al error.

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \quad (1.32)$$

La suma de productos del lado derecho se puede escribir como,

$$\sum_{i=1}^n (y_i - \hat{y}_i)\hat{y}_i - \bar{y} \sum_{i=1}^n (y_i - \hat{y}_i) \quad (1.33)$$

la primera de las sumas es 0 por la propiedad c) de los residuales y la segunda es 0 por la propiedad a) de los residuales. En consecuencia (1.32) se reduce a

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (1.34)$$

donde,

$SST = \sum_{i=1}^n (y_i - \bar{y})^2$  es llamada la suma de cuadrados del total y representa la variación total de las

y's.

$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  es llamada la Suma de Cuadrados del Error o Residual y representa la variación de las y's que se debe a causas no controlables. Notar que el estimado de la varianza poblacional  $s^2$ , puede ser calculado por  $s^2 = \frac{SSE}{n-2}$

$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  es llamada la Suma de Cuadrados debido a la Regresión y representa la variación de la y's que es explicada por su relación lineal con X.

Las sumas de cuadrados definidas anteriormente son variables aleatorias, pues dependen de y, la cual es aleatoria. Así, si en SSR se sustituye  $\hat{y}_i$  por  $\hat{\alpha} + \hat{\beta}x_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$  se tiene que

$$SSR = \hat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.35)$$

Por otro lado, considerando las sumas de cuadrados como variables aleatorias y tomando valor esperado en cada lado de relación (1.35) se tiene

$$E(SSR) = S_{xx}E(\hat{\beta}^2) = S_{xx}[Var(\hat{\beta}) + (E(\hat{\beta}))^2] = S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta^2\right)$$

Luego,

$$E(SSR) = \sigma^2 + \beta^2 S_{xx} \quad (1.36)$$

Asimismo,  $E(SST) = E(SSR) + E(SSE)$ . Así que,

$$E(SST) = \sigma^2 + \beta^2 S_{xx} + (n-2)\sigma^2 = (n-1)\sigma^2 + \beta^2 S_{xx} \quad (1.37)$$

Las sumas de cuadrados juegan un papel muy importante cuando se hace inferencia en regresión, por eso es importante saber como es su distribución. Por teoría de modelos lineales, se puede establecer que las sumas de cuadrados son formas cuadráticas de las variables  $y_i$  y por lo tanto se distribuyen como una Ji-cuadrado. Más específicamente, se pueden establecer los siguientes resultados:

i).  $\frac{SST}{\sigma^2} \sim \chi^2_{(n-1)}$  (Ji-cuadrado no central con n-1 grados de libertad ). Los grados de libertad se pueden establecer de la fórmula de cálculo de  $SST$ , pues en ella se usan  $n$  datos, pero en ella aparece un valor estimado ( $\bar{y}$ ) por lo tanto se pierde un grado de libertad.

ii).  $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-2)}$  (Ji-cuadrado con n-2 grados de libertad ). Para calcular  $SSE$  se usan  $n$  datos pero hay presente un estimado  $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ , cuyo cálculo depende a su vez de dos estimaciones. Por lo tanto se pierden dos grados de libertad. También se puede escribir que  $\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{(n-2)}$

iii).  $\frac{SSR}{\sigma^2} \sim \chi^2_{(1)}$  (Ji-cuadrado no central con 1 grado de libertad y parámetro de no centralidad  $\frac{\beta^2 S_{xx}}{\sigma^2}$ ). De la ecuación (1.35) se puede notar que el cálculo de  $SSR$  envuelve el cuadrado de una variable distribuida normalmente. Por un resultado de Estadística Matemática se sabe que el cuadrado de una normal estándar es una Ji-cuadrado con un grado de libertad.

### 1.2.8 El Coeficiente de Determinación $R^2$

Es una medida de la bondad de ajuste del modelo. Está definido por

$$R^2 = \frac{SSR}{SST} * 100\%$$

También indica que porcentaje de la variación de la variable de respuesta es explicada por su relación lineal con la variable predictora. Un modelo de regresión con  $R^2$  mayor o igual a 75% se puede considerar bastante aceptable. Aunque se puede ser un poco flexible dependiendo del tipo de datos y de la cantidad de datos disponible.

En el ejemplo 1, sólo un 62.6% de la variación de la mortalidad infantil de niños menores de 5 años es explicada por su relación lineal con el porcentaje de inmunización, lo cual no es muy alto y hace poco confiable las predicciones.

Lamentablemente, el valor de  $R^2$  es afectado por la presencia de valores anormales. Así, un valor de  $R^2$  bien cercano al 100% no necesariamente garantiza una buena predicción del modelo. Pero si se puede decir que un modelo con  $R^2$  bajo es inadecuado para hacer predicciones.

### 1.3 Inferencia en Regresión Lineal Simple

En esta sección discutirá pruebas de hipótesis e intervalos de confianza acerca de los coeficientes de regresión del modelo de regresión poblacional. También se construirán intervalos de confianza de las predicciones y del valor medio de la variable de respuesta.

### 1.3.1 Inferencia acerca de la pendiente y el intercepto usando la prueba t.

Inferencia acerca de la pendiente de la línea de regresión se discutirá detalladamente, en lo que respecta al intercepto será tratado brevemente. Como se ha visto en la sección 1.2.8 si se asume que las  $y_i$ 's tienen una distribución normal para cada valor de la variable predictora  $x$  entonces el estimado

$\hat{\beta}$  de la pendiente de regresión se distribuye como una normal con media  $\beta$  y varianza  $\frac{\sigma^2}{S_{xx}}$ . Esto es

equivalente a decir, que el estadístico  $z = \frac{\hat{\beta} - \beta}{\frac{\sigma}{\sqrt{S_{xx}}}}$  se distribuye como una normal estándar,  $N(0,1)$ .

Desafortunadamente, este estadístico no se puede usar en la práctica, pues por lo general  $\sigma$  es desconocida. Por otro lado, también sabemos que el estadístico  $\chi^2_{(n-2)} = \frac{(n-2)s^2}{\sigma^2}$  se distribuye como una Ji-cuadrado con  $n-2$  grados de libertad. Por un resultado de Estadística Matemática y probando previamente que hay independencia entre  $\hat{\beta}$  y  $s^2$ , se tiene que

$$t = \frac{z}{\sqrt{\frac{\chi^2_{(n-2)}}{n-2}}} = \frac{\hat{\beta} - \beta}{\frac{s}{\sqrt{S_{xx}}}} \quad (1.38)$$

se distribuye como una  $t$  de Student con  $n-2$  grados de libertad. El estadístico  $t$  dado en 1.38 es usado para hacer prueba de hipótesis y calcular intervalos de confianza acerca de  $\beta$ .

Un intervalo de confianza del  $100(1-\alpha)\%$  para la pendiente poblacional  $\beta$  es de la forma

$$\left( \hat{\beta} - t_{(n-2, \alpha/2)} \frac{s}{\sqrt{S_{xx}}}, \hat{\beta} + t_{(n-2, \alpha/2)} \frac{s}{\sqrt{S_{xx}}} \right)$$

donde  $\alpha$ , que varía entre 0 y 1, es llamado el nivel de significación,  $t_{(n-2, \alpha/2)}$  es un valor  $t$  tal que el área debajo de la curva y a la derecha de dicho valor es igual a  $\alpha/2$ . La expresión  $\frac{s}{\sqrt{S_{xx}}}$  es llamada

el error estándar (propriadamente es un estimado del error estándar) de  $\hat{\beta}$ . Muy raros son los programas estadísticos que muestran, en sus salidas de análisis de regresión, intervalos de confianza para la pendiente, solamente dan el  $\hat{\beta}$  y su error estándar. Hay que calcular  $t_{(n-2, \alpha/2)}$  usando cálculos de percentiles (por computadora o en tablas) y luego se calcula la fórmula del intervalo.

**Ejemplo 2:** Para los datos del ejemplo 1. Calcular un intervalo de confianza del 95% para la pendiente poblacional.

**Solución.** Usando el laboratorio 2 en R que aparece en la página de internet del texto se obtienen los siguientes resultados

```
> summary(l2)
```

Call:

```
lm(formula = tasa.mort ~ porc.inmuniz, data = muertes)
```

Residuals:

```
   Min      1Q  Median      3Q      Max
-99.97934 -16.57854  0.06684  20.84946  89.77608
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  224.3163    31.4403   7.135 1.20e-06 ***
porc.inmuniz  -2.1359     0.3893  -5.486 3.28e-05 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.14 on 18 degrees of freedom

Multiple R-Squared: 0.6258, Adjusted R-squared: 0.605

F-statistic: 30.1 on 1 and 18 DF, p-value: 3.281e-05

Notar que  $\hat{\beta} = -2.1359$  y que su error estándar es 0.3893. Los grados de libertad del la t son  $20-2=18$  y el  $\alpha=0.05$ , luego hay que buscar el percentil  $t_{(0.025,18)}$ . Este percentil, o su simétrico correspondiente, puede ser obtenido usando el comando de R, `qt(.975,18)`, el cual da un valor de 2.1009. Usando nuevamente el laboratorio 2 resulta

```
> # Hallando el intervalo de confianza del 95% para la pendiente Beta
```

```
> bint<-c(beta-qt(.975,18)*eebeta,beta+qt(.975,18)*eebeta)
```

```
> bint
```

```
[1] -2.95290 -1.31890
```

Luego, el Intervalo de confianza del 95% para  $\beta$  será

$$(-2.95290, -1.31890)$$

Por lo tanto, hay un 95% de confianza de que la pendiente de regresión poblacional caiga entre  $-2.95$  y  $-1.32$ .

Haciendo una discusión análoga al caso de la pendiente se puede llegar a establecer que un intervalo de confianza del  $100(1-\alpha)\%$  para el intercepto  $\alpha$  de la línea de regresión poblacional es de la forma

$$\left( \bar{\alpha} - t_{(n-2,\alpha/2)} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}}, \bar{\alpha} + t_{(n-2,\alpha/2)} s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{Sxx}} \right)$$

Ahora se considerará prueba de hipótesis para los coeficientes de la línea de regresión. Desde el punto de vista clásico, la siguiente tabla muestra la manera de hacer pruebas de hipótesis para la pendiente  $\beta$ , asumiendo que su valor es  $\beta^*$  y con un nivel de significación  $\alpha$ .

Caso I	Caso II	Caso III
Ho: $\beta = \beta^*$	Ho: $\beta = \beta^*$	Ho: $\beta = \beta^*$
Ha: $\beta < \beta^*$	Ha: $\beta \neq \beta^*$	Ha: $\beta > \beta^*$

Prueba Estadística		
$t = \frac{\hat{\beta} - \beta^*}{\frac{s}{\sqrt{Sxx}}} \sim t_{(n-2)}$		
Regla de Decisión		
Rechazar Ho, si $t_{cal} < -t_{(\alpha, n-2)}$	Rechazar Ho, si $ t_{cal}  > t_{(\alpha/2, n-2)}$	Rechazar Ho, si $t_{cal} > t_{(\alpha, n-2)}$

Obviamente el caso más importante es el caso II cuando  $\beta^*=0$ . Porque de rechazarse la hipótesis nula sugeriría de que hay relación lineal entre las variables X y Y. En la manera clásica uno rechaza o acepta la hipótesis nula comparando el valor de la prueba estadística con un valor obtenido de la tabla de t para un nivel de significación  $\alpha$  dado, usualmente de 0.05 ó 0.01.

A inicios de los años 80's y con la ayuda de los programas de computadoras se comenzó a probar hipótesis usando la técnica del "P-value", que es el nivel de significación observado. Es decir, el valor de  $\alpha$  al cual se rechazaría la hipótesis nula si se usaría el resultado que da la prueba estadística. Un "P-value" cercano a cero, sugeriría rechazar la hipótesis nula. Sin embargo, existe un consenso en la mayoría de los autores a rechazar la hipótesis nula si el "P-value" es menor de 0.05.

**Ejemplo 3:** Para los datos del ejemplo 1, probar la hipótesis de que la pendiente poblacional es cero.

**Solución:** Usando los resultados del laboratorio 2 de R.

Coefficients:

```

Estimate Std. Error t value Pr(>|t|)
(Intercept) 224.3163 31.4403 7.135 1.20e-06 ***
porc.inmuniz -2.1359 0.3893 -5.486 3.28e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 40.14 on 18 degrees of freedom  
Multiple R-Squared: 0.6258, Adjusted R-squared: 0.605

Las hipótesis serán:

Ho:  $\beta=0$  ( Es decir, no hay relación lineal entre las variables)

Ha:  $\beta \neq 0$  ( Hay relación lineal entre las variables)

Se observa que el "P-value" correspondiente a porcentaje de inmunización es  $0.0000328 < 0.05$ . Por lo tanto, se concluye que hay relación lineal entre las variables, aunque no se puede decir aún que tan fuerte es esta relación.

Similarmente, se pueden hacer pruebas de hipótesis para el intercepto.

Las hipótesis serán:

Ho:  $\alpha=0$  ( La línea de regresión poblacional pasa por el origen)

Ha:  $\alpha \neq 0$  ( La línea de regresión poblacional no pasa por el origen)

Como el “P-value” es  $0.000012 < 0.05$  se concluye que hay suficiente evidencia de que la línea de regresión poblacional NO pasa por el origen.

### 1.3.2 El análisis de varianza para regresión lineal simple

El análisis de varianza para regresión consiste en descomponer la variación total de la variable de respuesta en varias partes llamadas fuentes de variación. Como se vió en la sección 1.2.7, para el caso de regresión lineal solo hay dos fuentes: Una variación debido a la Regresión y otra variación debido al error. Cada variación es cuantificada por una suma de cuadrados, las cuales como se mencionó anteriormente tienen una distribución Ji-cuadrado.

Al dividir las suma de cuadrados por sus grados de libertad se obtienen los **cuadrado medio**. Así se tienen tres cuadrados medios

$$\begin{aligned}\text{Cuadrado Medio de Regresión} &= \text{MSR} = \text{SSR}/1 \\ \text{Cuadrado Medio del Error} &= \text{MSE} = \text{SSE}/(n-2) \\ \text{Cuadrado Medio del Total} &= \text{MST} = \text{SST}/(n-1),\end{aligned}$$

Pero este último no es usado. Notar también que  $\text{MSE} = s^2$ .

Por otro lado, en la sección 1.2.6, se ha demostrado que  $E[\text{MSE}] = \sigma^2$  y en la ecuación 1.36 de la sección 1.2.7 se tiene que  $E[\text{MSR}] = \sigma^2 + \beta^2 S_{xx}$ . Si estuviésemos probando la hipótesis  $H_0: \beta = 0$  y ésta fuera cierta entonces  $E[\text{MSR}] = \sigma^2$ , y su distribución pasa a ser una Ji-Cuadrado (central) con 1 grado de libertad. Luego, tanto MSE como MSR estimarían a la varianza poblacional.

Por resultados de Estadística Matemática se puede mostrar que la división de dos Cuadrados medios independientes se distribuye como una F. Más precisamente,

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F_{(1, n-2)}$$

siempre que la hipótesis nula  $H_0: \beta = 0$  es cierta. Aquí el numerador tiene 1 grado de libertad y el denominador tiene n-2. La independencia descansa en el hecho que  $\text{COV}(\hat{Y}_i - \bar{Y}, \hat{Y}_i - Y_i) = 0$ , véase ejercicio 10.

Todos los cálculos se resumen en la siguiente tabla llamada **tabla de Análisis de Varianza**

Fuente de Variación	g.l.	Sumas de Cuadrados	Cuadrados Medios	F
Debido a la Regresión	1	SSR	$\text{MSR} = \text{SSR}/1$	$\frac{\text{MSR}}{\text{MSE}}$
Error	n-2	SSE	$\text{MSE} = \text{SSE}/(n-2)$	
Total	n-1	SST		

Desde el punto de vista clásico la hipótesis  $H_0: \beta = 0$  se rechazaría en favor de  $H_0: \beta \neq 0$  si el valor de la prueba de F es mayor que  $F_{(\alpha, 1, n-2)}$ . En la manera moderna de probar hipótesis se rechazaría la hipótesis nula si el “P-value” de la prueba de F es menor de 0.05.

Para los datos del ejemplo 1, la tabla de análisis de varianza obtenida al correr el programa del laboratorio 2 será como sigue:

```
> anova(12)
Analysis of Variance Table
```

```

Response: tasa.mort
          Df Sum Sq Mean Sq F value    Pr(>F)
porc.inmuniz  1  48497   48497  30.101 3.281e-05 ***
Residuals    18  29001    1611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Claramente se rechaza la hipótesis nula pues el p-value da 0.0000328. Notar que,  $t_{(n-2)}^2 = F_{(1,n-2)}$ .

### 1.3.3 Intervalo de confianza para el valor medio de la variable de respuesta e Intervalo de Predicción

Talvez el uso más frecuente que se le da a una línea de regresión es para hacer predicciones acerca de la variable de respuesta  $Y$  para un valor dado de  $x$ . Supongamos que queremos predecir el valor medio de las  $Y$  para un valor  $x_0$  de la variable predictora  $x$ . Es decir,  $E(Y/x = x_0) = \alpha + \beta x_0$ .

Es natural pensar que el estimado puntual sera  $\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$ . Sin embargo, es muy riesgoso predecir basado en un solo valor y es más conveniente usar un intervalo donde se espera que caiga el valor de  $Y$  con un cierto grado de confianza. Como  $\hat{\alpha}$  y  $\hat{\beta}$  se distribuyen normalmente, entonces  $\hat{y}_0$  también se distribuye normalmente con media  $\alpha + \beta x_0$  y varianza igual a:

$$Var(\hat{Y}_0) = Var(\hat{\alpha} + \hat{\beta}x_0) = Var(\hat{\alpha}) + x_0^2 Var(\hat{\beta}) + 2x_0 Cov(\hat{\alpha}, \hat{\beta})$$

Sustituyendo expresiones halladas en la sección 1.2.3 y el hecho que  $Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{x}\sigma^2}{Sxx}$  (ver ejercicio 5), se tiene:

$$Var(\hat{Y}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{Sxx} \right) + x_0^2 \frac{\sigma^2}{Sxx} + 2x_0 \left( \frac{-\bar{x}\sigma^2}{Sxx} \right)$$

de donde resulta

$$Var(\hat{Y}_0) = \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx} \right)$$

En consecuencia, estandarizando y sustituyendo la  $\sigma$  por  $s$  se tendrá que:

$$\frac{\hat{y}_0 - E(Y/x_0)}{s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}} \sim t_{(n-2)}$$

Usando el resultado previo se puede establecer que un intervalo de confianza del  $100(1-\alpha)\%$  para el valor medio de las  $y$ 's dado que  $x=x_0$  es de la forma

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{(\alpha/2, n-2)} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (1.39)$$

Pero, frecuentemente uno está interesado en estimar un valor individual de Y dado  $x=x_0$  y no un promedio de valores. Evidentemente, que hay un mayor riesgo de hacer de esto. La predicción del valor individual  $Y_0 = \alpha + \beta x_0 + e_0$ , está dada también por  $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}x_0$ . Trabajando con la diferencia  $Y_0 - \hat{Y}_0$ , se puede ver fácilmente que  $E(Y_0 - \hat{Y}_0) = 0$  y que

$$Var(Y_0 - \hat{Y}_0) = Var(Y_0) + Var(\hat{Y}_0) - 2Cov(Y_0, \hat{Y}_0)$$

Luego,

$$Var(Y_0 - \hat{Y}_0) = \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) - 2Cov(Y_0, \hat{Y}_0)$$

como  $Y_0$  y  $\hat{Y}_0$  son no correlacionados,  $Cov(Y_0, \hat{Y}_0) = 0$ . Entonces,

$$Var(Y_0 - \hat{Y}_0) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Haciendo cálculos similares a cuando se obtuvo el intervalo de confianza para el valor medio, se puede establecer que un intervalo de confianza de  $100(1-\alpha)\%$  (mas conocido como intervalo de predicción) para un valor individual de Y dado  $x=x_0$  es de la forma

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{(\alpha/2, n-2)} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (1.40)$$

Notar que este intervalo es más amplio que el intervalo de confianza, pues la varianza estimada incluye un termino adicional. Muchos programas estadísticos calculan unas curvas que se obtienen uniendo los limites superiores e inferiores de los intervalos de confianza (o de predicción) para varios valores de la variable predictora, y estas son llamadas **Bandas de confianza (o Bandas de predicción)**.

**Ejemplo 4:** a) Hallar un intervalo de confianza del 99% para la tasa de mortalidad promedio de niños menores de 5 años en los países cuyo porcentaje de inmunización es 80%. Hallar un intervalo de predicción del 95% para la tasas de mortalidad de niños menores de 5 años en los países cuyo porcentaje de inmunización sea del 80%.

**Solución:** Usando nuevamente los resultados producidos por el programa del laboratorio 2 se obtiene los siguientes resultados.

```
> predict(l2,porc.inmuniz,se.fit=T,interval=c("confidence"),level=.99)
$fit
      fit   lwr   upr
[1,] 53.44674 27.44776 79.44572

$se.fit
[1] 9.032315
```

```
$df
[1] 18

$residual.scale
[1] 40.13931

> predict(l2,porc.inmuniz,se.fit=T,interval=c("prediction"),level=.95)
$fit
      fit   lwr   upr
[1,] 53.44674 -32.9915 139.8850

$se.fit
[1] 9.032315

$df
[1] 18

$residual.scale
[1] 40.13931
```

**Interpretación:** Hay un 99% de confianza de que la tasa de mortalidad media de todos los países con porcentaje de inmunización del 80% caiga entre 27.45 y 79.45 y la tasa de mortalidad de un país, cuyo porcentaje de inmunización es 80% caerá entre  $-32.99$  y  $139.88$  con un 95% de confianza.

La figura 1.6 muestra las bandas de confianza y predicción del 95 por ciento para los datos del ejemplo 1.

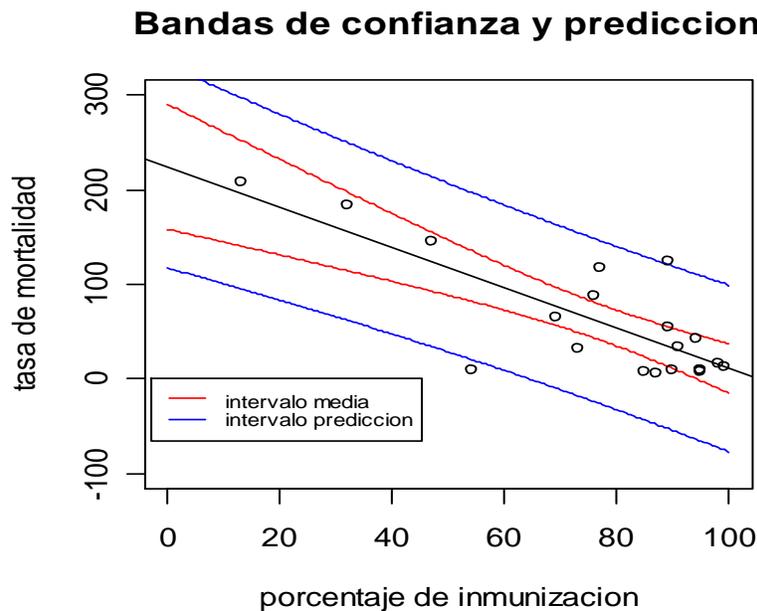


Figura 1.6 Bandas de confianza y predicción para los datos del ejemplo 1.

## 1.4 El Coeficiente de Correlación

Algunas veces se considera que tanto la variable de respuesta como la predictora son aleatorias. Por ejemplo, si se quiere relacionar horas de estudio ( $X$ ) y nota en un examen ( $Y$ ). La manera estándar sería establecer de antemano los posibles números de horas que se va a considerar y luego para cada una de las horas elegir al azar por lo menos un estudiante y registrarle su nota. Sin embargo, también se puede elegir al azar un estudiante y hacerle las dos preguntas: Cuántas horas estudió? y qué nota obtuvo en el examen?. En este caso  $(X, Y)$  se comporta como una variable aleatoria bivariada, que generalmente se distribuye como una normal bivariada. Una Normal Bivariada tiene cinco parámetros: las medias  $\mu_x$ ,  $\mu_y$ , las desviaciones estándares  $\sigma_x$  y  $\sigma_y$  y el coeficiente de correlación  $\rho$ .

El coeficiente de correlación, es un valor que mide el grado de asociación lineal entre las variables aleatorias  $X$  y  $Y$  y se define como

$$\rho = \frac{Cov(X, Y)}{\sigma_x \sigma_y} \quad (1.41)$$

Se puede mostrar que

- $-1 \leq \rho \leq 1$
- Si  $\rho^2=1$  entonces  $Y=A + BX$ , con probabilidad 1. Si  $Y=A + BX$ , donde  $A$  y  $B$  son constantes, entonces si  $A>0$ ,  $\rho=1$  y si  $A<0$ ,  $\rho=-1$ .
- Si la regresión de  $Y$  sobre  $X$  es lineal, esto es la media condicional de  $Y$  dado  $X$  es

$$E(Y/X) = \alpha + \beta x. \text{ Entonces, } \beta = \rho \frac{\sigma_y}{\sigma_x}, \text{ y } \alpha = \mu_y - \beta \mu_x. \text{ Notar que si la pendiente de la}$$

línea de regresión es cero entonces la correlación es 0, y que  $\beta$  y  $\rho$  varían en la misma dirección.

- Si  $(X, Y)$  se distribuye como una normal bivariada, entonces la varianza condicional de  $Y$  dado  $X$ , está dado por  $\sigma_{y/x}^2 = \sigma_y^2(1 - \rho^2)$ . Luego, si  $\rho = \pm 1$ , entonces  $\sigma_{y/x}^2 = 0$ , implicando que hay una perfecta relación lineal entre  $Y$  y  $X$ . Más específicamente, si  $\rho = 1$ , entonces  $X$  y  $Y$  crecen en la misma dirección y si  $\rho = -1$ ,  $Y$  decrece cuando  $X$  crece.

Todo lo anterior ocurre en la población, así que  $\rho$  es un parámetro que debe ser estimado. Suponiendo que se ha tomado una muestra de  $n$  pares  $(x_i, y_i)$ , entonces, el **coeficiente de correlación muestral** se calcula por

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (1.42)$$

Notar que  $r = \hat{\beta} \sqrt{\frac{S_{xx}}{S_{yy}}}$  y que  $r^2 = \frac{\hat{\beta}^2 S_{xx}}{S_{yy}} = \frac{SSR}{SST}$ . Es decir, que el cuadrado del coeficiente de

correlación es igual al coeficiente de determinación. Al igual que el parámetro poblacional, la correlación muestral varía entre  $-1$  y  $1$ . Por lo general, un  $r$  mayor de  $0.75$  en valor absoluto es considerado aceptable, aunque algunas veces debido a la naturaleza de los datos hay que exigir un valor más alto, digamos mayor de  $0.90$ .

En R, el comando `cor` permite calcular la correlación entre dos o más variables. Para el ejemplo 1, los resultados son:

```
> cor(muertes$tasa.mort,muertes$porc.inmuniz)
[1] -0.7910654
```

El valor de la correlación en valor absoluto es algo mayor de 0.75, lo que implicaría una aceptable relación lineal entre las variables, además cuando el porcentaje de inmunización aumenta la tasa de mortalidad disminuye.

**Advertencia:** *Correlación alta no implica necesariamente una relación causa efecto entre las variables. Usando la fórmula de correlación entre dos variables que en la vida real no tiene ninguna relación entre sí, (por ejemplo X: peso de los profesores y Y=salario del profesor) se puede obtener un r bastante alto cercano a 1 o -1 y eso no implica necesariamente que X explique el comportamiento de Y (podría darse el caso que mientras menos pesa un profesor gana menos).*

La figura 1.7 muestra varios diagramas de puntos y sus respectivas correlaciones. Los datos y comandos correspondientes aparecen el laboratorio 3 del texto.

### Ejemplos de correlaciones

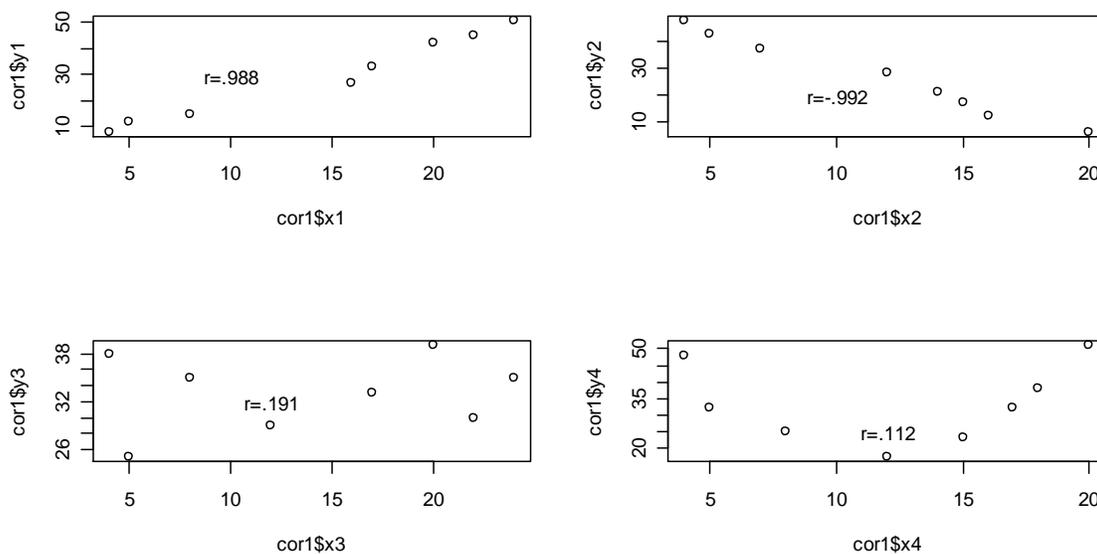


Figura 1.7 Cuatro distintos patrones de datos y sus correspondientes coeficientes de correlación

Notar que en los dos últimos plots la correlación es cercana a cero, pero en el primer caso no parece haber ningún tipo de relación entre las variables y en el otro no hay relación lineal pero si existe una relación cuadrática.

El valor de correlación es afectado por la presencia de valores anormales, en la figura 1.8 se puede ver el efecto de los valores anormales en el valor de la correlación para 4 diferentes relaciones. Los datos y comandos correspondientes aparecen el laboratorio 3 del texto.

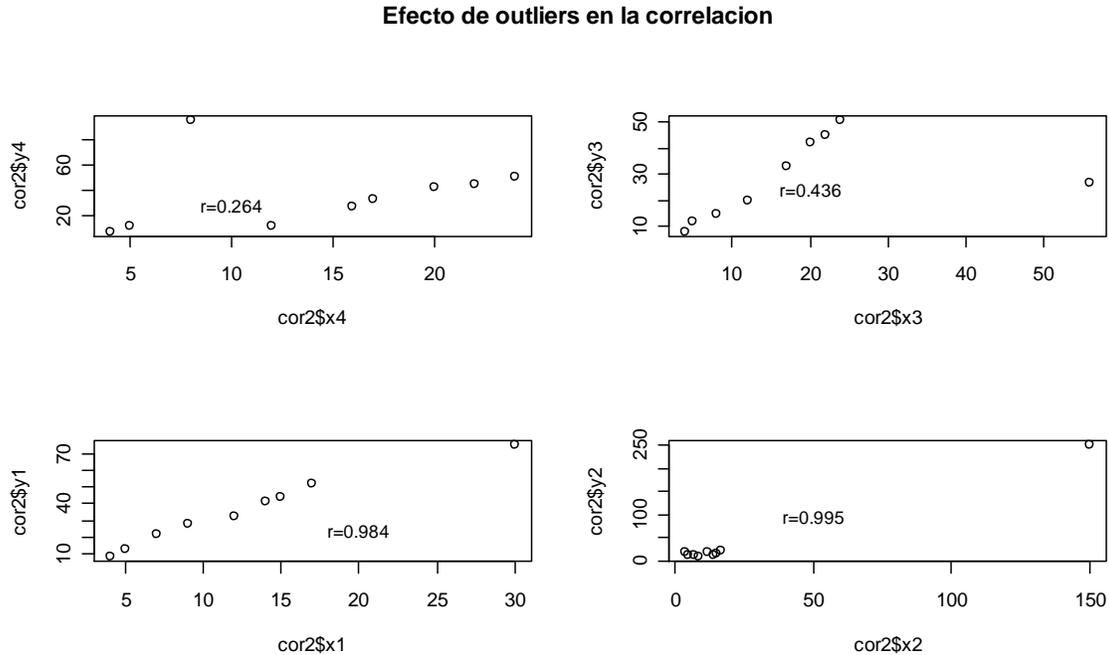


Figura 1.8 Cuatro distintos patrones de datos para mostrar el efecto de outliers en el coeficiente de correlación

**Interpretación de la figura 8:** En el primer caso existe un valor bastante anormal en la dirección vertical que hace que la correlación sea bastante baja a pesar de que los otros valores parecen estar bastante alineados.

En el segundo caso existe un valor bastante alejado horizontalmente de la mayor parte de los datos y que hace que la correlación sea relativamente baja a pesar de que los otros valores muestran una alta asociación lineal.

En el tercer caso hay una observación bastante alejado en ambas direcciones sin embargo no tiene ningún efecto en la correlación, cuyo valor de por sí es alto.

En el cuarto caso hay un valor bastante alejado en ambas direcciones y las restantes observaciones están poco asociadas, pero el valor anormal hace que el valor de la correlación sea bastante alto.

Debido a la relación entre la pendiente de la línea de regresión y el coeficiente de correlación, la prueba estadística para probar  $H_0: \rho=0$  (la correlación poblacional es cero) versus  $H_a: \rho \neq 0$  (hay correlación entre las poblaciones X e Y) es similar a la prueba de la pendiente de la línea de regresión: Es decir,

$$t = \frac{\hat{\beta}}{\frac{s}{\sqrt{S_{xx}}}} = \frac{r \sqrt{\frac{S_{yy}}{S_{xx}}}}{\sqrt{\frac{S_{yy}(1-r^2)}{n-2} \frac{1}{S_{xx}}}} = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)} \quad (1.43)$$

La prueba estadística para probar  $H_0: \rho = \rho_0$  (la correlación poblacional es de magnitud  $\rho_0$ ) versus  $H_a: \rho \neq \rho_0$  involucra el uso de una transformación del coeficiente de correlación muestral, llamada la **transformación z de Fisher**, ya que la distribución de  $r$  no es normal y tiende a ser asimétrica para valores grandes de  $\rho$ . La transformación está definida por

$$z = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right) = \tanh^{-1}(r) \quad (1.44)$$

la cual tiene una distribución aproximadamente normal con media

$$E(z) = \frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right) + \frac{\rho}{2(n-1)}$$

y varianza  $Var(z) = \frac{1}{n-3}$ . La aproximación es bastante buena si  $n > 50$ . En consecuencia, la prueba estadística será:

$$Z = \sqrt{n-3} \left[ \frac{1}{2} \log\left(\frac{1+r}{1-r}\right) - \frac{1}{2} \log\left(\frac{1+\rho_0}{1-\rho_0}\right) - \frac{\rho_0}{2(n-1)} \right]$$

## 1.5 Análisis de residuales

Los residuales,  $r_i = y_i - \hat{y}_i$  que son estimaciones de los errores aleatorios del modelo  $\hat{e}_i = y_i - \hat{y}_i = y_i - \alpha - \beta x_i$ , son importantes para establecer si las suposiciones del modelo se cumplen y para explorar el porqué de un mal ajuste del modelo. La manera más fácil de examinar los residuales es mediante plots los cuales permiten cotejar:

- Si la distribución de los errores es normal y sin “outliers”.
- Si la varianza de los errores es constante y si se requieren transformaciones de las variables.
- Si la relación entre las variables es efectivamente lineal o presenta algún tipo de curvatura
- Si hay dependencia de los errores, especialmente en el caso de que la variable predictora sea tiempo.

Existen varios tipos de residuales, por ahora solo introduciremos dos:

**i) Residual Estandarizado:** En este caso se divide el residual entre la desviación estándar del error. Es decir,

$$\text{Residual estandarizado} = \frac{y_i - \hat{y}_i}{s}$$

ii) **Residual Estudentizado:** En el residual estandarizado se está considerando de antemano que cada residual tiene la misma varianza, pero en realidad cada uno de ellos tiene su propia varianza como se muestra a continuación.

$$\text{Var}(y_i - \hat{y}_i) = \text{Var}(y_i) + \text{Var}(\hat{y}_i) - 2\text{Cov}(y_i, \hat{y}_i)$$

Usando resultados de la sección 1.3.3, lo anterior se puede escribir como

$$\text{Var}(y_i - \hat{y}_i) = \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) - 2\text{Cov}(y_i, \bar{y} + \hat{\beta}(x_i - \bar{x}))$$

calculando la covarianza, se obtiene

$$\text{Var}(y_i - \hat{y}_i) = \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) - 2\sigma^2 \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)$$

En consecuencia,

$$\text{Var}(y_i - \hat{y}_i) = \sigma^2 \left( 1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right)$$

Por lo tanto, usando el correspondiente estimado para  $\sigma^2$  se tiene que:

$$r_i^* = \frac{r_i}{s \sqrt{\left( 1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}}$$

el cual es llamado el *i*-ésimo **residual estudentizado**. En algunos programas estadísticos, los  $r_i^*$  son llamados residuales estandarizados. También son llamados residuales estudentizados internamente (ver más adelante, la sección 3.1).

### 1.5.1 Cotejando normalidad de los errores y detectando outliers

Normalidad de los errores es un requisito indispensable para que tengan validez las pruebas estadísticas de t y F que se usan en regresión. Existen varios métodos gráficos y pruebas estadísticas tanto paramétricas como no paramétricas para cotejar la normalidad de un conjunto de datos. La manera más fácil es usando gráficas tales como histogramas, “stem-and-leaf” o “Boxplots”.

Una gráfica más especializada es el plot de Normalidad. Aquí se plotea los residuales versus los valores que se esperarían si existiera normalidad, estos valores son llamados los scores normales. Dado el *i*-ésimo residual, su score normal se encontrará determinando primero a que percentil le corresponde en la distribución de los datos, se han propuesto varias maneras de hacer esto. Luego de determinar el percentil se halla el valor que le corresponde a dicha percentil en la distribución normal estándar. El *i*-ésimo score normal es aproximado en forma bastante precisa por

$$z_{(i)} = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$$

donde  $\Phi$  representa la función de distribución acumulada de una normal estándar y  $n$  es el número de observaciones en la muestra.

Habría normalidad si los puntos del plot se alinean cerca de una línea que pasa por el origen. Si se usan los residuales estudentizados la línea además de pasar por el origen debería tener pendiente cercana a 1.

**Ejemplo 5.** Cotejar si existe normalidad para los datos del ejemplo 1.

Considerando los residuales estudentizados y la funciones **hist** y **boxplot** de R se obtiene el histograma y “boxplot” correspondientes.

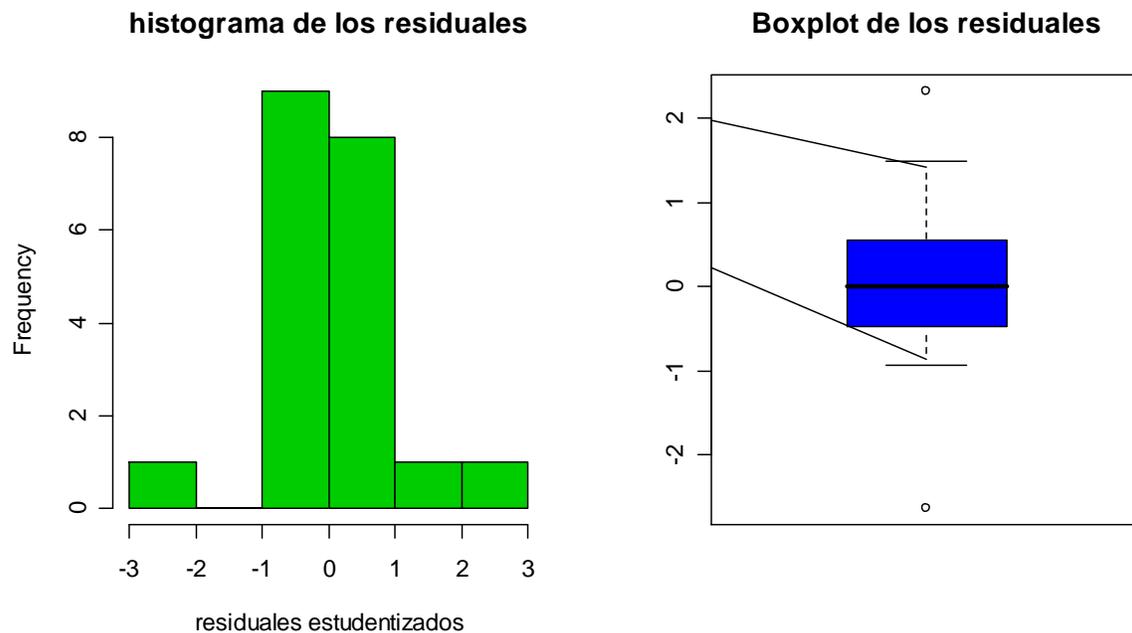


Figura 1.9 Histograma y boxplot de residuales de la regresión del ejemplo 1.

El histograma no parece ser de forma acampanada, es decir no hay normalidad, además parece haber un “outlier” inferior. El boxplot indica bastante simetría en el centro pero no así en los extremos de la distribución. Además se identifican dos outliers, uno superior y el otro inferior.

En R, el plot de normalidad se usa usando los comandos **qqnorm** y **qqline**. El plot de Normalidad correspondiente al ejemplo 1 se muestra en la figura 1.10.

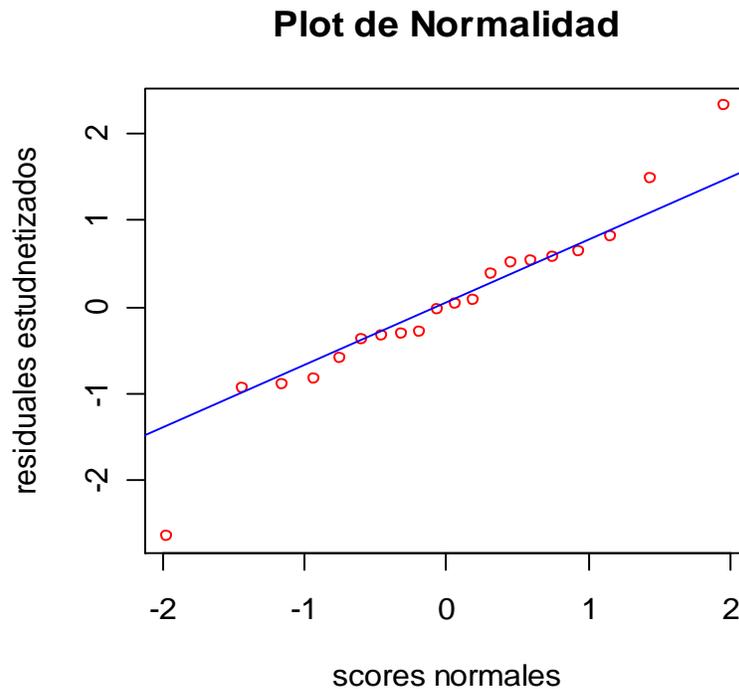


Figura 1.10 Plot de Normalidad para los residuales de la regresion del ejemplo 1.

En el plot de normalidad los puntos siguen una tendencia bastante lineal, especialmente en el centro. Pero lo que es más notorio es la presencia de un “outlier” inferior y dos probables “outliers” superiores.

Otra manera de detectar si hay “outliers” es cotejando si los residuales estudentizados son mayores que 2 en valor absoluto. En el capítulo dedicado a diagnósticos de regresión se hará una discusión más detallada de los criterios para detectar “outliers”.

### 1.5.2 Cotejando que la varianza sea constante

En este caso se plotea los residuales estandarizados versus los valores ajustados o versus la variable predictora  $X$ . No se plotea versus las  $y_i$  observadas porque los residuales y las  $y_i$  's se espera que estén correlacionados.

Si los puntos del plot caen en una franja horizontal alrededor de 0 entonces la varianza de los errores es constante. Si los puntos siguen algún patrón entonces se dice que la varianza de los errores no es constante.

**Ejemplo 6:** Hacer un plot de residuales para cotejar si hay varianza constate de los errores para los datos del ejemplo 1.

Los comandos en R para obtener estas graficas aparecen el laboratorio 4 del texto. Para los datos del ejemplo 1 se obtienen los plots que aparecen en las figuras 1.11 y 1. 12.

### Visualizando si la varianza es constante

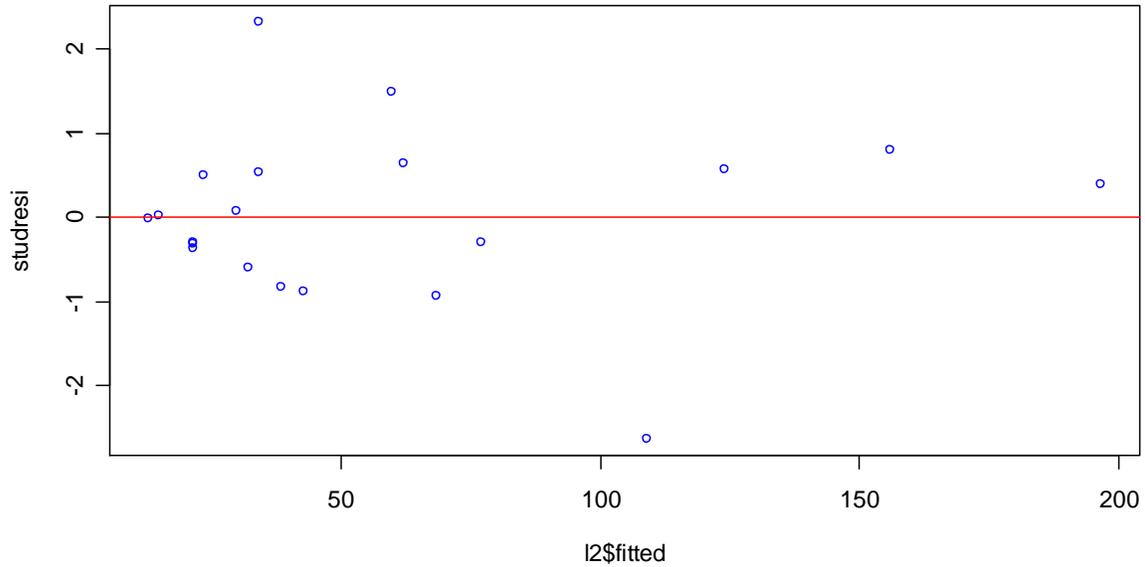


Figura 1.11 Plot de residuales para detectar si la varianza constante con respecto a los valores ajustados

En la Figura 1.11 se puede notar que los puntos se reparten equitativamente alrededor de la línea horizontal. Nuevamente lo que llama más la atención es la presencia del “outlier”. Por lo tanto, la varianza parece ser constante. Si plotamos los residuales versus la variable predictora en lugar de los valores ajustados, se obtiene la siguiente gráfica

### Visualizando la dependencia de la varianza y la predictora

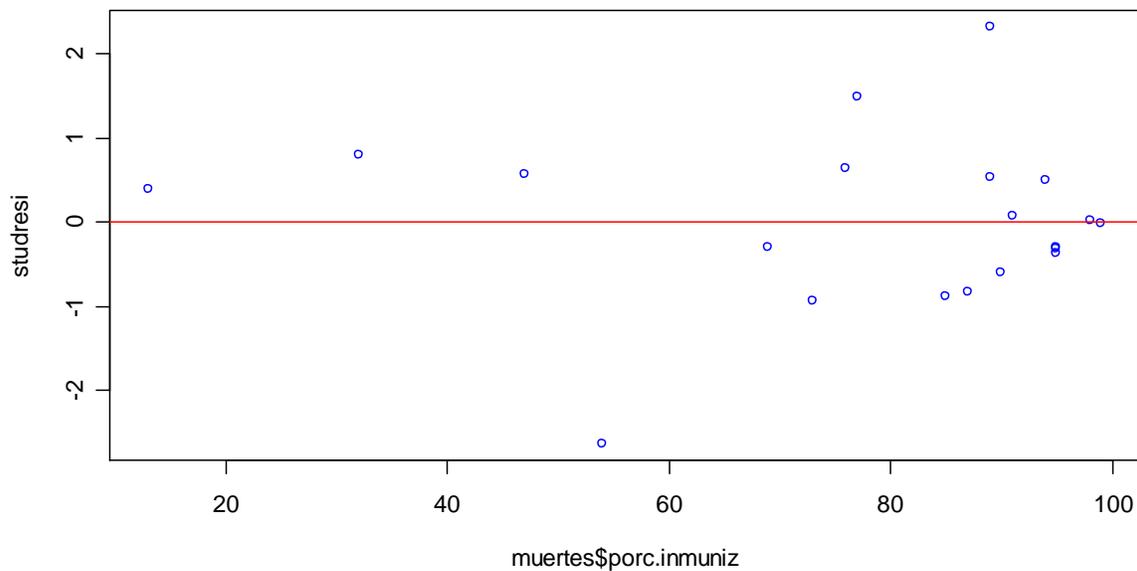


Figura 1.12 Plot de residuales para detectar si la varianza constante con respecto a la variable predictora

Al igual que en la figura anterior, se puede ver en la figura 1.12 que, excluyendo los dos “outliers”, los puntos parecen estar en una franja horizontal, por lo tanto se podría considerar que la varianza es constante con respecto a la predictora. Notar que también hay cuatro puntos alejados en la dirección horizontal. Estas observaciones también pueden tener influencia en los cálculos de la línea de regresión.

Si se observa algún patrón en el plot se puede hacer transformaciones en una o en ambas variables para estabilizar la varianza. Otra alternativa es usar *mínimos cuadrados ponderados*. Nuevamente esto será discutido más detalladamente en el capítulo 3 del texto cuando se haga análisis de residuales en regresión múltiple.

En R se puede hacer un plot simultáneo de los residuales. Usando el laboratorio 4 de R para el ejemplo 1 se obtiene la siguiente gráfica.

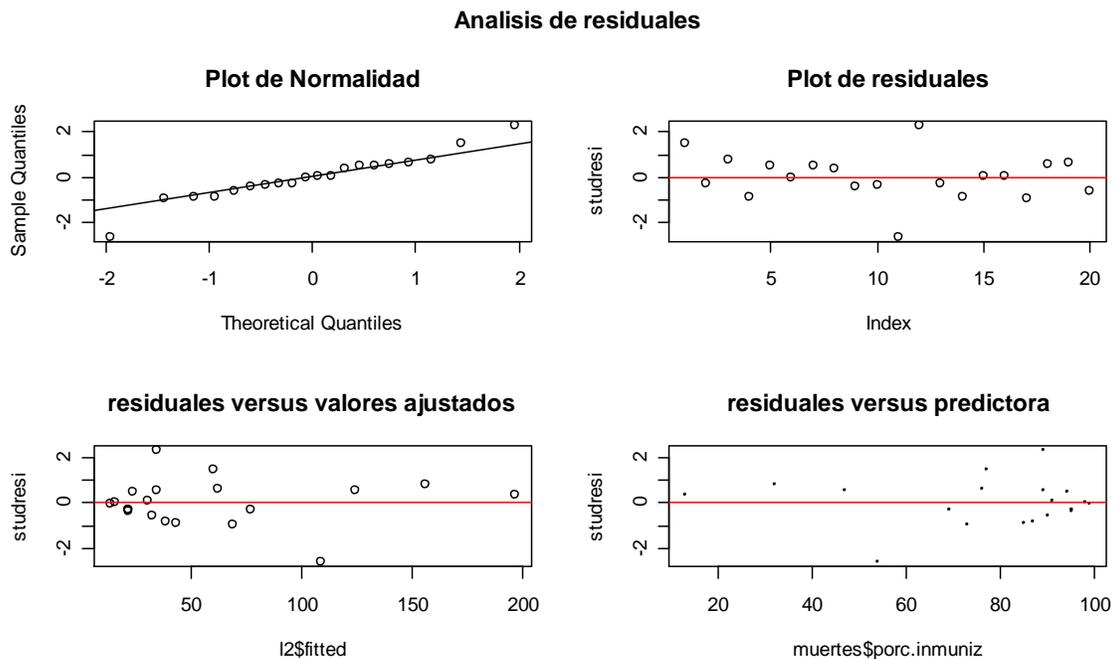


Figura 1.13 Plots para hacer análisis de residuales

### 1.5.3 Cotejando si los errores están correlacionados.

Cuando la variable predictora es tiempo, puede ocurrir que los errores estén correlacionados secuencialmente entre sí. Si en el plot de residuales versus valores ajustados se observa un patrón cíclico entonces hay correlación entre los errores.

Existe también la prueba de Durbin-Watson que mide el grado de correlación de un error con el que anterior y el posterior a él. El estadístico es

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

Notar que  $D$  es aproximadamente igual a  $2(1-r)$ , donde  $r$  representa la correlación lineal entre los errores  $e_i$ 's y  $e'_{i-1}$ 's. Usando ese hecho se puede mostrar que  $D$  varía entre 0 y 4. Si  $D$  es cerca de 0 los errores están correlacionados positivamente. Si  $D$  está cerca de 4 entonces la correlación es negativa. Además la distribución de  $D$  es simétrica con respecto a 2. Así que un valor de  $D$  cercano a 2 indica que no hay correlación de los errores. Más formalmente hay que comparar el valor de  $D$  con dos valores críticos  $D_L$  y  $D_U$  de una tabla.

Aplicando la función `dw` del laboratorio 4 de R a los datos del ejemplo 1 resulta

```
el estadístico Durbin Watson de la regresión lineal es= 2.678912
```

Como el valor está cerca de 2, se concluirá que no hay correlación entre los errores. También se puede ver en el plot de residuales, que no hay un patrón cíclico de los puntos.

## EJERCICIOS

1. Considerando un modelo de regresión lineal simple, calcular  $Cov(\bar{Y}, \hat{\beta})$
2. Probar que la línea de regresión estimada pasa por  $(\bar{X}, \bar{Y})$
3. En un modelo de regresión lineal simple calcular  $E[SST]=E[\sum_{i=1}^n (y_i - \bar{y})^2]$
4. **Regresión que pasa por el origen.** Algunas veces se conoce de antemano que la línea de regresión pasa por el origen. Es decir el modelo es de la forma  $y_i = \beta x_i + e_i$ .
  - a) Hallar el estimador por cuadrados mínimos de  $\beta$ . Cuál es su varianza?
  - b) Hallar el estimador de la varianza poblacional  $\sigma^2$
  - c) Establecer la fórmula para un intervalo de confianza del  $100(1-\alpha)\%$  de confianza para  $\beta$

5. Probar que  $Cov(\hat{\alpha}, \hat{\beta}) = \frac{-\bar{x}\sigma^2}{S_{xx}}$

6. En un estudio del desarrollo del conocimiento se registra la edad ( $X$ ) en meses a la que 21 niños dicen su primera palabra y el puntaje en la prueba de Gessell ( $Y$ ), un test de habilidad que toma posteriormente el niño (ver datos **Gessell** en la página de internet del curso). Los resultados son como siguen

Edad	Puntaje	Edad	Puntaje
15	95	9	96
26	71	10	83
10	83	11	84
9	91	11	102
15	102	10	100
20	87	12	105
18	93	42	57
10	100	17	121
8	104	11	86
20	94	10	100
7	113		

- a) Hallar la línea de regresión. e interpretar los coeficientes de la línea de regresión
  - b) Trazar la línea de regresión encima del diagrama de puntos.
  - c) Probar la hipótesis de que la pendiente es cero. Comentar su resultado
  - d) Interpretar el coeficiente de determinación  $R^2$
  - e) Hallar un intervalo de confianza del 99% para la pendiente de la línea de regresión poblacional
  - f) Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 95% para el valor individual y valor medio de la variable de respuesta e intprepar el resultado.
7. En un pueblo se eligen 15 personas al azar y se anota su salario mensual ( $X$ ), y la cantidad que ahorran mensualmente ( $Y$ ). Ambas cantidades están expresadas en dólares. (ver datos **salarios** en la página de internet del curso).

Salario	Ahorro
800	150
850	100

900	280
1200	400
1500	350
1700	500
1900	635
2000	600
2300	750
2500	680
2700	900
3000	800
3200	300
3500	1200
5000	1000

- Hallar la línea de regresión. e interpretar los coeficientes de la línea de regresión
- Trazar la línea de regresión encima del diagrama de puntos.
- Interpretar el coeficiente de determinación
- Probar la hipótesis de que la pendiente es cero. Comentar su resultado
- Hallar un intervalo de confianza del 95% para la pendiente de regresión poblacional.
- Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 90 para el valor individual y el valor medio de la variable de respuesta e intprepar el resultado.

8. Leer el conjunto de datos **brain** que aparece en la página de internet del texto y considerar las variables:

**MRI (X)**, conteo en pixels del 18 scans de resonancia magnetica del cerebro de una persona  
**Score\_IQ, (Y)** score en un test de inteligencia.

Mientras más alto sea el conteo de pixels mas grande es el cerebro de las personas.

- Hallar la línea de regresión ajustada. e interpretar los coeficientes de la línea de regresión
- Trazar la línea de regresión encima del diagrama de puntos.
- Probar la hipótesis de que la pendiente es cero (usando las pruebas t y F). Comentar su resultado
- Interpretar el Coeficiente de Determinación.
- Hallar un intervalo de confianza del 99% para la pendiente de la regresion poblacional e interpretar su resultado
- Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 90 por ciento para el valor individual y el valor medio de la variable de respuesta e intprepar el resultado.

9.

- Si  $Y=3.5-1.5X$  ,  $SST=219$  y  $SSE=59$ , hallar e interpretar el valor de la correlación entre X y Y
- Considerando los datos dados en a) y que la muestra de entrenamiento consiste de 36 datos, hallar el valor de la prueba estadística para probar que la pendiente de regresión es cero.

10. Considerando un modelo de regresión lineal simple, calcular

$$Cov(Y_i - \hat{Y}_i, \hat{Y}_i - \bar{Y})$$

11. Probar que el coeficiente de correlación muestral r cae entre  $-1$  y  $1$ .

12. Suponga que en el modelo de regresión lineal simple los valores  $x_i$  y  $y_i$  son reemplazados por  $ax_i+b$  y  $cy_i+d$  respectivamente donde a,b,c y d son constantes tales que  $a \neq 0$  y  $c \neq 0$ . Cuál es el efecto

de estas transformaciones en  $\hat{\alpha}$ ,  $\hat{\beta}$ ,  $\hat{\sigma}^2$ ,  $R^2$  y la prueba estadística para probar la hipótesis nula  $H_0: \beta=0$ ?

13. Considere el modelo de regresión lineal simple  $Y = \alpha + \beta X + \varepsilon$ , donde tanto X como Y y  $\varepsilon$  son variables aleatorias con varianzas  $\sigma_x^2$ ,  $\sigma_y^2$  y  $\sigma_\varepsilon^2$  respectivamente y  $\sigma_{xy}$  representa la covarianza entre X y Y. En la estimación mínimo cuadrática de  $\alpha$  y  $\beta$  se minimiza la suma de cuadrados de las **distancias verticales** de las observaciones a la línea ajustada. En **Regresión Ortogonal** la estimación de  $\alpha$  y  $\beta$  se hace considerando que la línea es ajustada de tal manera que se minimiza la **distancia** más corta de las observaciones a la línea ajustada. Hallar los estimadores de los coeficientes de la regresión ortogonal.

14. Los siguientes datos fueron recolectados por el físico James Forbes para estimar indirectamente la altura sobre el nivel del mar de acuerdo a las mediciones del punto de ebullición (boiling point) del agua. (ver datos **Forbes** en la página de internet del curso).

Columna 1: Numero de la observación

Columna 2: Boiling Point( °F)

Columna 3: Pressure (in. Hg)

Columna 3: Log(Pressure)

Columna 4: 100\*Log(Pressure)

1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3804	138.04
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

- Hacer un diagrama de puntos de Pressure versus Boiling point. Piensa Ud. que hay una tendencia lineal.
- Hacer un diagrama de puntos de 100\*log(Pressure) versus Boiling point. Piensa Ud. que se observa mejor la tendencia lineal que en a)
- Ajustar la línea de regresión de 100\*log(Pressure) versus Boiling point. Trazar la línea sobre el plot hallado en b). Comentar los coeficientes de regresión. Interpretar los “p-values” de la prueba t y el de la prueba F.
- Interpretar el Coeficiente de determinación  $R^2$
- Obtener un intervalo de confianza del 99% para  $\beta$ . Interpretar su resultado
- Obtener un intervalo de confianza del 99% para el valor predicho y un intervalo de confianza para el valor medio de 100\*log(Pressure) cuando el Boiling Point es de 195 °F.

15. Los siguientes datos fueron recolectado para tratar de pronosticar el nivel del agua del rio Snake en Wyoming. (ver datos **River** en la página de internet del curso).

Columna 1: Contenido de agua en la nieve caida hasta Abril 1, desde 1919 hast 1935

Columna 2: Produccion de agua (en pulgadas) en el rio Snake, entre los meses de abril y julio.

23.1 10.5  
 32.8 16.7  
 31.8 18.2  
 32.0 17.0  
 30.4 16.3  
 24.0 10.5  
 39.5 23.1  
 24.2 12.4  
 52.5 24.9  
 37.9 22.8  
 30.5 14.1  
 25.1 12.9  
 12.4 8.8  
 35.1 17.4  
 31.5 14.9  
 21.1 10.5  
 27.616.1

- Hacer un diagrama de puntos de la producción de agua versus cantidad de agua en la nieve. Piensa Ud. que hay una tendencia lineal.
- Ajustar la linea de regresión producción de agua versus cantidad de agua en la nieve. Trazar la linea sobre el plot hallado en a). Comentar los coeficientes de regresión. Interpretar los “p-values” de la prueba t y el de la prueba F.
- Interpretar el Coeficiente de determinación  $R^2$ .
- Hallar un intervalo de confianza del 95% para la pendiente. Interpretar su resultado.
- Obtener un intervalo de confianza del 95% para el valor predicho y un intervalo de confianza para el valor medio de la producción de agua cuando la cantida de agua en la nieve es de un 35%.

16. Los siguientes datos fueron registrados en el río Amazonas (Iquitos, Perú) para observar si la deforestación afecta los niveles del agua del río. (Ver datos **Amazonas** en la página de internet del curso).

columna 1: año

columna 2: nivel de agua máximo (en metros) en el rio Amazonas en Iquitos

columna 3: nivel de agua mínimo (en metros) en el rio Amazonas en Iquitos

1962 25.82 18.24  
 1963 25.35 16.50  
 1964 24.29 20.26  
 1965 24.05 20.97  
 1966 24.89 19.43  
 1967 25.35 19.31  
 1968 25.23 20.85  
 1969 25.06 19.54  
 1970 27.13 20.49  
 1971 27.36 21.91  
 1972 26.65 22.51

1973 27.13 18.81  
 1974 27.49 19.42  
 1975 27.08 19.10  
 1976 27.51 18.80  
 1977 27.54 18.80  
 1978 26.21 17.57

- Hacer un diagrama de puntos de Nivel máximo versus año, Nivel mínimo versus año y de Nivel Máximo versus Nivel mínimo. Piensa Ud. que hay una tendencia lineal?.
- Obtener la línea de regresión de Nivel máximo versus año, Nivel mínimo versus año y de Nivel Máximo versus Nivel mínimo. Piensa Ud. que hay una tendencia lineal. Interpretar los coeficientes y los “p-values” de las pruebas t y F.
- Interpretar el Coeficiente de determinación  $R^2$  para cada una de las 3 regresiones.
- Obtener un intervalo de confianza del 95% para el valor predicho y un intervalo de confianza para el valor medio del nivel máximo del agua para el año 1980.

17. Los siguientes datos se han recolectado para explicar el rendimiento en millas por gallon de varios modelos de carros.

Columna 1: Modelo de carro  
 Columna 2.VOL: Volumen de la cabina del carro.  
 Columna 3.HP: caballos de potencia del motor  
 Columna 4.MPG: millas promedio por galón  
 Columna 5 .SP: Velocidad máxima (mph)  
 Columna 6.WT: Peso de vehiculo (100 lb)

Modelo de carro	VOL	HP	MPG	SP	WT
GM/GeoMetroXF1	89	49	65.4	96	17.5
GM/GeoMetro	92	55	56.0	97	20.0
GM/GeoMetroLSI	92	55	55.9	97	20.0
SuzukiSwift	92	70	49.0	105	20.0
DaihatsuCharade	92	53	46.5	96	20.0
GM/GeoSprintTurbo	89	70	46.2	105	20.0
GM/GeoSprint	92	55	45.4	97	20.0
HondaCivicCRXHF	50	62	59.2	98	22.5
HondaCivicCRXHF	50	62	53.3	98	22.5
DaihatsuCharade	94	80	43.4	107	22.5
SubaruJusty	89	73	41.1	103	22.5
HondaCivicCRX	50	92	40.9	113	22.5
HondaCivic	99	92	40.9	113	22.5
SubaruJusty	89	73	40.4	103	22.5
SubaruJusty	89	66	39.6	100	22.5
SubaruJusty4wd	89	73	39.3	103	22.5
ToyotaTercel	91	78	38.9	106	22.5
HondaCivicCRX	50	92	38.8	113	22.5
ToyotaTercel	91	78	38.2	106	22.5
FordEscort	103	90	42.2	109	25.0
HondaCivic	99	92	40.9	110	25.0
PontiacLeMans	107	74	40.7	101	25.0
IsuzuStylus	101	95	40.0	111	25.0
DodgeColt	96	81	39.3	105	25.0
GM/GeoStorm	89	95	38.8	111	25.0
HondaCivicCRX	50	92	38.4	110	25.0

HondaCivicWagon	117	92	38.4	110	25.0
HondaCivic	99	92	38.4	110	25.0
Subaru Loyale	102	90	29.5	109	25.0
VolksJettaDiesel	104	52	46.9	90	27.5
Mazda323Protege	107	103	36.3	112	27.5
FordEscortWagon	114	84	36.1	103	27.5
FordEscort	101	84	36.1	103	27.5
GM/GeoPrism	97	102	35.4	111	27.5
ToyotaCorolla	113	102	35.3	111	27.5
EagleSummit	101	81	35.1	102	27.5
NissanCentraCoupe	98	90	35.1	106	27.5
NissanCentraWagon	88	90	35.0	106	27.5
ToyotaCelica	86	102	33.2	109	30.0
ToyotaCelica	86	102	32.9	109	30.0
ToyotaCorolla	92	130	32.3	120	30.0
ChevroletCorsica	113	95	32.2	106	30.0
ChevroletBeretta	106	95	32.2	106	30.0
ToyotaCorolla	92	102	32.2	109	30.0
PontiacSunbirdConv	88	95	32.2	106	30.0
DodgeShadow	102	93	31.5	105	30.0
DodgeDaytona	99	100	31.5	108	30.0
EagleSpirit	111	100	31.4	108	30.0
FordTempo	103	98	31.4	107	30.0
ToyotaCelica	86	130	31.2	120	30.0
ToyotaCamry	101	115	33.7	109	35.0
ToyotaCamry	101	115	32.6	109	35.0
ToyotaCamry	101	115	31.3	109	35.0
ToyotaCamryWagon	124	115	31.3	109	35.0
OldsCutlassSup	113	180	30.4	133	35.0
OldsCutlassSup	113	160	28.9	125	35.0
Saab9000	124	130	28.0	115	35.0
FordMustang	92	96	28.0	102	35.0
ToyotaCamry	101	115	28.0	109	35.0
ChryslerLebaronConv	94	100	28.0	104	35.0
DodgeDynasty	115	100	28.0	105	35.0
Volvo740	111	145	27.7	120	35.0
FordThunderbird	116	120	25.6	107	40.0
ChevroletCaprice	131	140	25.3	114	40.0
LincolnContinental	123	140	23.9	114	40.0
ChryslerNewYorker	121	150	23.6	117	40.0
BuickReatta	50	165	23.6	122	40.0
OldsTrof/Toronado	114	165	23.6	122	40.0
Oldsmobile98	127	165	23.6	122	40.0
PontiacBonneville	123	165	23.6	122	40.0
LexusLS400	112	245	23.5	148	40.0
Nissan300ZX	50	280	23.4	160	40.0
Volvo760Wagon	135	162	23.4	121	40.0
Audi200QuatroWag	132	162	23.1	121	40.0
BuickElectraWagon	160	140	22.9	110	45.0
CadillacBrougham	129	140	22.9	110	45.0
CadillacBrougham	129	175	19.5	121	45.0
Mercedes500SL	50	322	18.1	165	45.0
Mercedes560SEL	115	238	17.2	140	45.0
JaguarXJSCConvert	50	263	17.0	147	45.0
BMW750IL	119	295	16.7	157	45.0
Rolls-RoyceVarious	107	236	13.2	130	55.0

- a) Hacer un diagrama de puntos de MPG versus HP. Piensa Ud. que hay una tendencia lineal.
- b) Ajustar la línea de regresión de MPG versus HP. Trazar la línea sobre el plot hallado en a).  
Comentar los coeficientes de regresión. Interpretar los “p-values” de la prueba t y el de la prueba F.
- c) Interpretar el Coeficiente de determinación  $R^2$
- d) Obtener un intervalo de confianza del 99% para  $\beta$ . Interpretar su resultado
- e) Obtener un intervalo de confianza del 90% para el valor predicho y un intervalo de confianza para el valor medio de MPG cuando HP=100.
- f) Hacer un diagrama de puntos de MPG versus WT. Piensa Ud. que hay una tendencia lineal.
- g) Ajustar la línea de regresión de MPG versus WT. Trazar la línea sobre el plot hallado en a).  
Comentar los coeficientes de regresión. Interpretar los “p-values” de la prueba t y el de la prueba F.
- h) Interpretar el Coeficiente de determinación  $R^2$
- i) Obtener un intervalo de confianza del 99% para  $\beta$ . Interpretar su resultado
- j) Obtener un intervalo de confianza del 90% para el valor predicho y un intervalo de confianza para el valor medio de MPG cuando WT=35.

18.

- a) Si  $Y=3.5-1.5X$ ,  $SST=219$  y  $SSE=59$ , hallar e interpretar el valor de la correlación entre X y Y
- b) Considerando los datos dados en a) y que la muestra de entrenamiento consiste de 36 datos, hallar el valor de la prueba estadística para probar que la pendiente de regresión es cero.

19. Considere que  $(X,Y)$  tiene una distribución normal bivariada con parámetros  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x$ ,  $\sigma_y$  y el coeficiente de correlación  $\rho$ . Probar que

a) La media condicional de Y dado X es  $E(Y/X) = \alpha + \beta x$ . Donde  $\beta = \rho \frac{\sigma_y}{\sigma_x}$ , y  $\alpha = \mu_y - \beta \mu_x$ .

Notar que si la pendiente de la línea de regresión es cero entonces la correlación es 0, y que  $\beta$  y  $\rho$  varían en la misma dirección.

b) La varianza condicional de las Y dado X, está dado por  $\sigma_{y/x}^2 = \sigma_y^2(1 - \rho^2)$ . Luego, si  $\rho = \pm 1$ , entonces  $\sigma_{y/x}^2 = 0$ , implicando que hay una perfecta relación lineal entre Y y X. Más específicamente, si  $\rho = 1$ , entonces X y Y crecen en la misma dirección y si  $\rho = -1$ , Y decrece cuando X crece.

# CAPÍTULO 2

## REGRESIÓN LINEAL MULTIPLE

### 2.1 Introducción

Es evidente que lo más económico y rápido para modelar el comportamiento de una variable  $Y$  es usar una sola variable predictora y usar un modelo lineal. Pero, algunas veces es bastante obvio de que el comportamiento de  $Y$  es imposible que sea explicada en gran medida por solo una variable. Por ejemplo, es imposible tratar de explicar el rendimiento de un estudiante en un examen, teniendo en cuenta solamente el número de horas que se preparó para ella. Claramente, el promedio académico del estudiante, la carga académica que lleva, el año de estudios, son tres de las muchas otras variables que pueden explicar su rendimiento. Tratar de explicar el comportamiento de  $Y$  con más de una variable predictora usando una funcional lineal es el objetivo de regresión lineal múltiple.

Frecuentemente, uno no es muy familiar con las variables que están en juego y basa sus conclusiones solamente en cálculos obtenidos con los datos tomados. Es decir, si ocurre que el coeficiente de determinación  $R^2$  sale bajo (digamos menor de un 30%), considerando además que su valor no se ha visto afectado por datos anormales, entonces el modelo es pobre y para mejorarlo hay tres alternativas que frecuentemente se usan:

- a) Transformar la variable predictora, o la variable de respuesta  $Y$ , o ambas y usar luego un modelo lineal.
- b) Usar regresión polinómica con una variable predictora.
- c) Conseguir más variables predictoras y usar una regresión lineal múltiple.

En el primer caso, se puede perder el tiempo tratando de encontrar la transformación más adecuada y se podría caer en sobre-ajuste (“*overfitting*”), es decir, encontrar un modelo demasiado optimista, que satisface demasiado la tendencia de los datos tomados pero que es pobre para hacer predicciones debido a que tiene una varianza grande.

En el segundo caso el ajuste es más rápido, pero es bien fácil caer en “*overfitting*” y, además se pueden crear muchos problemas de cálculo ya que pueden surgir problemas de colinealidad, es decir relación lineal entre los términos del modelo polinómico.

El tercer caso es tal vez la alternativa más usada y conveniente. Tiene bastante analogía con el caso simple, pero requiere el uso de vectores y matrices.

En el siguiente ejemplo se mostrará el uso interactivo de las tres alternativas a través de seis modelos de regresión y servirá como un ejemplo de motivación para introducirnos en regresión lineal múltiple.

**Ejemplo 1:** Considerar el conjunto de datos **millaje** donde la variable de respuesta es  $Y$ =(MPG) millas promedio por galón de un auto, y las variables predictoras son;  $X_1$ =(VOL): Capacidad en volumen del carro,  $X_2$ =(HP): Potencia del Motor,  $X_3$ =(SP) :Velocidad Máxima y  $X_4$ =(WT): Peso del auto. Los datos fueron adaptados de la “Data and Story Library” ([lib.stat.cmu.edu/DASL/](http://lib.stat.cmu.edu/DASL/)) y están disponibles en [academic.uprm.edu/eacuna/millaje.txt](http://academic.uprm.edu/eacuna/millaje.txt).

Primero, se explorará las relaciones entre todas las parejas de variables, en particular la relación de  $Y$  con cada una de las variables predictoras. Esto se logra con una gráfica llamada **plot matricial**, la cual está disponible en la mayoría de programas estadísticos. La función **pairs** de **R**

produce el plot matricial para las variables del ejemplo 1 tal como se muestra en la siguiente figura:

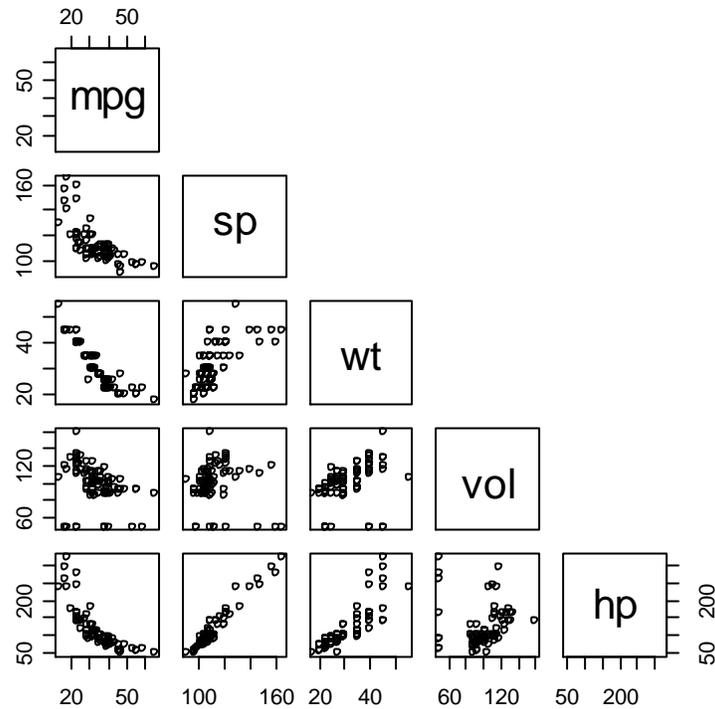


Figura 2.1. Plot matricial de las variables del conjunto de datos **millaje**.

Claramente se puede ver que la variable predictora (WT) es la que tiene mejor relación lineal con MPG y que VOL tiene una pobre relación lineal con MPG. En tanto que para HP y SP la relación lineal no es muy marcada.

Ahora, analicemos la relación entre HP y MPG. Un plot de estas variables se muestra en la figura 2.2. Si hacemos la regresión lineal entre las dos variables se obtiene

```
> l1<-lm(mpg~hp)
> l1

Call:
lm(formula = mpg ~ hp)

Coefficients:
(Intercept)          hp
    50.0661       -0.1390

> summary(l1)$r.squared
[1] 0.6239
```

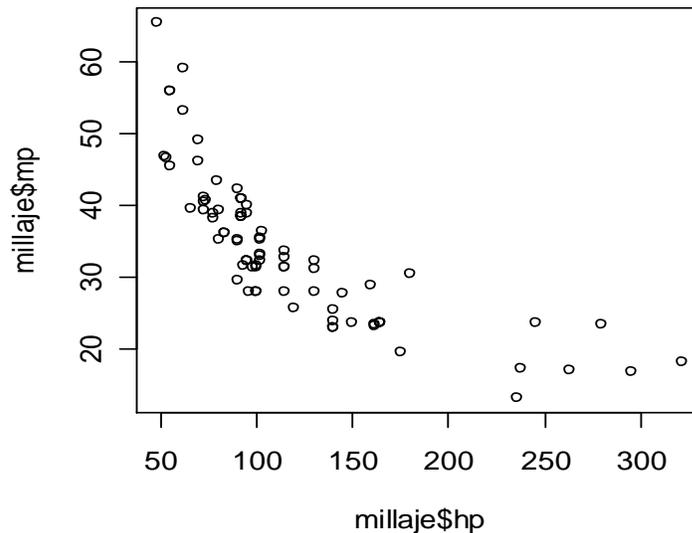


Figura 2.2. Plot de MPG versus HP.

El  $R^2 = 62.4\%$  no está bajo, pero hay que tratar de mejorarlo, usando las alternativas a y b. En la gráfica se observa una curvatura, así que se podría ajustar una regresión cuadrática. Los resultados usando  $hp$  y  $hp^2$  como variables predictoras son los siguientes:

```
> l2=lm ( mpg ~ hp + hp2)
> l2
```

Call:

```
lm(formula = mpg ~ hp + hp2)
```

Coefficients:

```
(Intercept)    hp    hp2
71.2313424 -0.4598708 0.0009707
```

```
> summary(l2)$r.squared
[1] 0.8067
```

El  $R^2$  resulta ser  $80.7\%$  lo que representa una gran mejora, pero hay un peligro de hacer predicciones porque al final la cuadrática tiende a subir, y se podría concluir que un auto con 400 HP podría tener un rendimiento de 42.59 millas por galón, similar al de un carro de 50 HP. Este es un ejemplo de un modelo sobreajustado (“overfitted”). Notar también el valor bien pequeño del coeficiente del término cuadrático, el cual podría causar problema en el cálculo de las predicciones.

Observando más detenidamente la gráfica de la figura 2.2 se puede ver que hay un comportamiento asintótico en la parte inferior, es decir, que después de cierto nivel de HP, el millaje tiende a estabilizarse. Esto sugiere que podríamos tratar un modelo hiperbólico de la

forma  $MPG = \alpha + \beta \frac{1}{HP}$  para ajustar los datos. Considerando la predictora  $hp1 = 1/hp$ , se obtienen los siguientes resultados en R.

```
> l3=lm(mpg~hp1)
> l3
```

Call:

```
lm(formula = mpg ~ hp1)
```

Coefficients:

```
(Intercept)    hp1
      9.73    2373.11
```

```
> summary(l3)$r.squared
[1] 0.8429
```

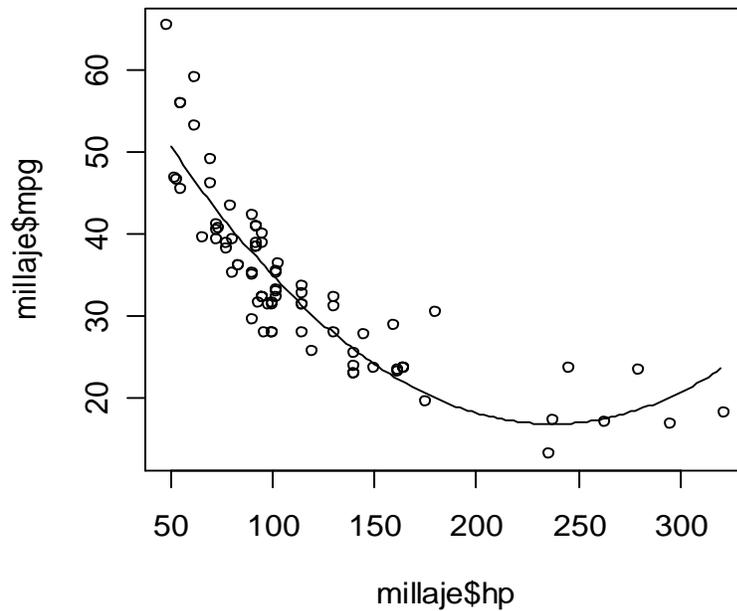


Figura 2.3. Regresión cuadrática de MPG versus HP

Notar que el  $R^2 = 84.3\%$  está bastante aceptable lo cual indica un buen ajuste del modelo. Así para un carro con 400 de HP su MPG será 15.66.

Alguién que no quiere perder el tiempo explorando relaciones polinómicas o haciendo transformación de variables, tratará de conseguir información acerca de otras variables

predictoras, con la esperanza de subir sustancialmente su  $R^2$  pero usando solamente modelos lineales.

Del plot matricial que aparece en la figura 2.1 no hay ninguna duda de que la variable a considerar conjuntamente con HP sería WT. Haciendo uso de R se obtiene

```
> l4<-lm(mpg~hp+wt)
> l4
```

Call:

```
lm(formula = mpg ~ hp + wt)
```

Coefficients:

```
(Intercept)    hp      wt
  66.85500   -0.02097  -0.99037
```

```
> summary(l4)$r.squared
[1] 0.8235
```

El cual sería el segundo mejor modelo usando el criterio de  $R^2$  ya que produce un valor de 82.4%. Si usamos el hecho de que la relación de Y con HP1 resulta ser bastante buena, podemos intentar ajustar un modelo lineal con HP1 y WT como las variables predictoras. Los resultados aparecen a continuación:

```
> l5<-lm(mpg~hp1+wt)
> l5
```

Call:

```
lm(formula = mpg ~ hp1 + wt)
```

Coefficients:

```
(Intercept)    hp1      wt
  36.5361   1387.1768  -0.5439
```

```
> summary(l5)$r.squared
[1] 0.8933
```

Este último sería el mejor modelo hasta ahora ya que su  $R^2=89.9$  es el mayor de todos. Así se puede seguir explorando más modelos, pero teniendo cuidado de no caer en “*overfitting*”.

Si ajustamos un modelo de regresión lineal múltiple con las 4 variables predictoras disponibles se obtiene

```
> l6<-lm(mpg~vol+hp+sp+wt)
> l6
```

Call:

```
lm(formula = mpg ~ vol + hp + sp + wt)
```

Coefficients:

```
(Intercept)    vol      hp      sp      wt
 192.43775   -0.01565   0.39221  -1.29482  -1.85980
```

```
> summary(l6)$r.squared
```



### 2.2.1 Estimación del vector de parámetros $\beta$ por Cuadrados Mínimos

Al igual que en regresión lineal simple hay que minimizar la suma de cuadrados de los errores. La suma de cuadrados de los errores puede ser expresada vectorialmente de la siguiente manera

$$Q(\beta) = \sum_{i=1}^n e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) \quad (2.3)$$

donde el símbolo ' indica transpuesta del vector o matriz (es decir, la matriz que se obtiene intercambiando las fila por columnas en la matriz original). Haciendo operaciones con los vectores y matrices se obtiene

$$Q(\beta) = \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta = \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \quad (2.4)$$

En la igualdad anterior se ha usado la propiedad  $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ . Derivando Q con respecto a  $\beta$  e igualando a cero se obtiene el sistema de ecuaciones normales;

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y} \quad (2.5)$$

de donde resolviendo para  $\beta$  se obtiene

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2.6)$$

aquí  $(\mathbf{X}'\mathbf{X})^{-1}$  representa la matriz inversa de  $(\mathbf{X}'\mathbf{X})$ . Notar que  $\mathbf{X}'\mathbf{X}$  es simétrica, pues su transpuesta da la misma matriz.

En la regresión lineal simple,  $p=1$  y el modelo puede ser escrito en forma matricial como

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \\ & \\ & \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

Manipulando las matrices se obtiene que

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdot & \cdot & \cdot & 1 \\ x_{11} & x_{21} & \cdot & \cdot & \cdot & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \\ & \\ & \\ 1 & x_{n1} \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \end{bmatrix}$$

Luego las ecuaciones normales se reducen a:

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \end{bmatrix}$$

Por comodidad podemos eliminar el segundo subíndice de las  $x$ 's ya que no afecta en nada. Como

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}$$

se concluye que

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{nS_{xx}} \begin{bmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ -\sum_{i=1}^n x_i \sum_{i=1}^n y_i + n \sum_{i=1}^n x_i y_i \end{bmatrix} \quad (2.7)$$

y haciendo manipuleo algebraico se llega a las formulas para los estimadores del intercepto y de la pendiente que se vieron en la sección 1.2 del capítulo 1.

## 2.2.2 Propiedades del estimador $\hat{\beta}$

En forma similar al caso simple, el estimador minimo cuadrático tiene las siguientes propiedades:

a)  $\hat{\beta}$  es insesgado, o sea  $E(\hat{\beta}) = \beta$ . En efecto,

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e})] = E[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}] \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{e}) \end{aligned}$$

como  $E(\mathbf{e})=0$ , se llega a  $E(\hat{\boldsymbol{\beta}})=\boldsymbol{\beta}$ .

$$b) \text{Var}(\hat{\boldsymbol{\beta}})=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Para probar esto debemos usar una propiedad de la matriz de varianza-covarianza de  $\mathbf{Az}$  donde  $\mathbf{A}$  es matriz y  $\mathbf{z}$  vector columna. La propiedad dice que  $\text{Var}(\mathbf{Az})=\mathbf{A}\text{Var}(\mathbf{z})\mathbf{A}'$ .

$$\text{Luego, } \text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{Var}(\mathbf{Y})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Usando los hechos que  $(\mathbf{X}')'=\mathbf{X}$  y que  $[(\mathbf{X}'\mathbf{X})^{-1}]'=(\mathbf{X}'\mathbf{X})^{-1}$ , por simetría de la matriz inversa de  $\mathbf{X}'\mathbf{X}$ , se obtiene que

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

c) Si no se asume normalidad el estimador mínimo-cuadrático  $\hat{\boldsymbol{\beta}}$  es el mejor estimador dentro de los estimadores lineales insesgados de  $\boldsymbol{\beta}$ , en el sentido que es el que tiene la varianza más pequeña. Este es llamado el **teorema de Gauss-Markov**.

d) Si se asume normalidad de los errores entonces  $\hat{\boldsymbol{\beta}}$  es el mejor estimador entre todos los estimadores insesgados de  $\boldsymbol{\beta}$

### 2.2.3 Estimación de la varianza $\sigma^2$

En un modelo de regresión lineal múltiple con  $p$  variables predictoras (con el intercepto habrían en total  $p+1$  parámetros a estimar), se tiene que un estimado de la varianza de los errores es

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} = \frac{\sum_{i=1}^n \hat{e}_i^2}{n-p-1} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p-1} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1} \quad (2.8)$$

El numerador de la expresión representa la suma de cuadrados de los residuales y puede ser escrito como:

$$SSE = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{Y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})'(\mathbf{Y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = \mathbf{Y}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

donde  $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  es conocida en regresión como la "*Hat Matrix*" (la matriz sombrero). Notar que  $\mathbf{H}'=\mathbf{H}$  y que  $\mathbf{H}^2=\mathbf{H}\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'=\mathbf{H}$ . En álgebra matricial se dice que  $\mathbf{H}$  es idempotente.  $\mathbf{H}$  tiene muy buenas propiedades una de ellas es que  $\text{Traza}(\mathbf{H})=\text{rango}(\mathbf{H})=p+1$ . Por otro lado,  $(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})=(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})=\mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{H}^2=\mathbf{I}-2\mathbf{H}+\mathbf{H}=\mathbf{I}-\mathbf{H}$ . Así que también  $\mathbf{I}-\mathbf{H}$  es también simétrica e idempotente.

En consecuencia, la varianza estimada de los errores puede ser escrita como:

$$\hat{\sigma}^2 = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}}{n - p - 1} = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}}{n - p - 1} \quad (2.9)$$

Más común es usar el símbolo  $s$  para la desviación estándar estimada de los errores y

$$s = \sqrt{\frac{SSE}{n - p - 1}} = \sqrt{MSE}$$

**Propiedad:** Sea  $\mathbf{Y}$  un vector aleatorio  $n$ -dimensional tal que  $E(\mathbf{Y}) = \boldsymbol{\mu}$  y  $\text{VAR}(\mathbf{Y}) = \mathbf{V}$  entonces

$$E(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = \text{Traza}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \quad (2.10)$$

Usando la propiedad anterior con  $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$  y  $\mathbf{V} = \sigma^2\mathbf{I}_n$  se puede mostrar que  $E[s^2] = \sigma^2$ .

**Ejemplo 2:** Hallar el modelo de regresión lineal múltiple para explicar el rendimiento en millaje promedio por galón (MPG) de los vehículos de acuerdo a cuatro variables predictoras: VOL, HP, SP y WT e interpretar los valores estimados.

Las variables predictoras que fueron usadas antes en el ejemplo 1 están definidas como sigue:

VOL: Capacidad de la cabina en pies cúbicos

HP: Potencia del motor

SP: Velocidad máxima (mph)

WT: Peso del vehículo (100 lb)

El modelo de regresión que ya fue obtenido en el ejemplo 1 es el siguiente:

$$\text{MPG} = 192 - 0.0156 \text{ VOL} + 0.392 \text{ HP} - 1.29 \text{ SP} - 1.86 \text{ WT}$$

### Interpretación de los coeficientes de regresión estimados:

$\hat{\beta}_1 = -0.0156$  significa que el millaje promedio por galón baja en promedio en 0.0156 cuando el volumen interior del carro aumenta en un pie cúbico, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_2 = 0.392$  significa que el millaje promedio por galón aumenta en promedio en 0.392 cuando la potencia del motor aumenta en 1 HP, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_3 = -1.29$  significa que el millaje promedio por galón baja en promedio en 1.29 cuando la velocidad máxima del carro aumenta en 1 milla por hora, asumiendo que las otras variables permanecen fijas.

$\hat{\beta}_4 = -1.86$  significa que el millaje promedio por galón baja en 1.86 cuando el peso del vehículo aumenta en 100 libras, asumiendo que las otras variables permanecen fijas.

En general, un coeficiente de regresión representa el cambio promedio en la variable de respuesta cuando la variable predictora correspondiente se incrementa en una unidad adicional y asumiendo que las otras variables predictoras permanecen fijas.

### 2.3. Inferencia en Regresión lineal múltiple

En esta sección se harán pruebas de hipótesis e intervalos de confianza acerca de los coeficientes del modelo de regresión poblacional. También se calcularán intervalos de confianza de las predicciones que se hacen con el modelo.

De ahora en adelante vamos a suponer que  $\mathbf{e} \sim \text{NI}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$  o equivalente que  $\mathbf{Y} \sim \text{NI}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ .

Al igual que en regresión lineal simple la variación total de Y se descompone en dos variaciones: una debido a la regresión y otra debido a causas no controlables. Es decir,

$$SST = SSR + SSE$$

Por teoría de modelos lineales se puede determinar que las sumas de cuadrados que aparecen en el análisis de regresión son formas cuadráticas de la variable de respuesta Y. Por lo tanto, éstas se distribuyen como una Ji-cuadrado. Más específicamente, se pueden establecer los siguientes resultados:

i).  $\frac{SST}{\sigma^2} \sim \chi^2_{(n-1)}$  Ji-cuadrado no central con n-1 grados de libertad. Los grados de libertad se

pueden establecer de la fórmula de cálculo de SST, pues en ella se usan n datos, pero en ella aparece un valor estimado ( $\bar{y}$ ) por lo tanto se pierde un grado de libertad.

ii).  $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-p-1)}$  Ji-cuadrado con n-p-1 grados de libertad. Para calcular SSE se usan n

datos pero hay presente un estimado  $\hat{y}_i$  cuyo cálculo depende a su vez de p+1 estimaciones. Por lo tanto se pierden p+1 grados de libertad. También se puede escribir que

$$\frac{(n-p-1)s^2}{\sigma^2} \sim \chi^2_{(n-p-1)}$$

iii).  $\frac{SSR}{\sigma^2} \sim \chi^2_{(p)}$  Ji-cuadrado no central con p grados de libertad

#### 2.3.1 Prueba de hipótesis acerca de un coeficiente de regresión individual

En este caso la hipótesis nula más importante es  $H_0: \beta_i = 0$  ( $i=1,2,\dots,p$ ), o sea la variable  $X_i$  no es importante en el modelo, versus la hipótesis alterna  $H_a: \beta_i \neq 0$ ; la variable  $X_i$  si merece ser considerada en el modelo. La prueba estadística es la prueba de t, definida por

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{s\sqrt{C_{ii}}}$$

donde el error estándar de  $\hat{\beta}_i$  se calcula por  $se(\hat{\beta}_i) = s\sqrt{C_{ii}}$ ,  $C_{ii}$  es el i-ésimo elemento de la diagonal de  $(\mathbf{X}'\mathbf{X})^{-1}$ . Esta t se distribuye como una t de Student con n-p-1 grados de libertad. R al igual que otros programas de computadoras, da el "P-value" de la prueba t. Para el ejemplo anterior se obtiene lo siguiente

```
> summary(l6)
```

Call:

```
lm(formula = mpg ~ vol + hp + sp + wt)
```

Residuals:

```
Min 1Q Median 3Q Max
-9.0108 -2.7731 0.2733 1.8362 11.9854
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775 23.53161 8.178 4.62e-12 ***
vol -0.01565 0.02283 -0.685 0.495
hp 0.39221 0.08141 4.818 7.13e-06 ***
sp -1.29482 0.24477 -5.290 1.11e-06 ***
wt -1.85980 0.21336 -8.717 4.22e-13 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Los "P-values" de la prueba de t sugieren que la variable **VOL** no contribuye al modelo, pues se acepta la hipótesis nula de que dicho coeficiente es cero. Las otras tres variables **WT**, **HP** y **SP** si parecen ser importantes en el modelo ya que los "P-values" de la prueba t correspondientes son menores que .05.

### 2.3.2 Prueba de Hipótesis de que todos los coeficientes de regresión son ceros.

En este caso la hipótesis nula es  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ , o sea que el modelo no sirve, versus la hipótesis alterna  $H_a$ : Al menos uno de los coeficientes es distinto de cero, o sea al menos una de las variables del modelo puede ser usada para explicar la variación de  $Y$ .

En la sección 2.2.3 se mencionó que  $E(s^2) = \sigma^2$ . La suma de cuadrados del error tiene  $n-p-1$  g.l. Nuevamente usando esperado de una formas cuadrática se puede mostrar que

$$E(SSR) = E[\mathbf{Y}'(\mathbf{H} - \mathbf{1}\mathbf{1}'/n)\mathbf{Y}] = p\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{H} - \mathbf{1}\mathbf{1}'/n)\mathbf{X}\boldsymbol{\beta} \quad (2.11)$$

Donde  $\mathbf{1}$  es un vector columna de  $n$  unos. Si la hipótesis nula se cumpliera entonces  $E(MSR) = \sigma^2$ . La suma de cuadrados de Regresión tiene  $p$  grados de libertad que es igual al número de variables predictoras en el modelo.

Se puede mostrar que si la hipótesis nula es cierta entonces :

$$F = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} = \frac{MSR}{MSE}$$

se distribuye como una  $F$  con  $p$  grados de libertad en el numerador y  $n-p-1$  g.l en el denominador.

La prueba de  $F$  se obtiene al hacer la tabla del análisis de varianza para la regresión múltiple, la cual se muestra a continuación:

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados Medios	F
Regresión	SSR	P	MSR=SSR/p	F=MSR/MSE
Error	SSE	n-p-1	MSE=SSE/n-p-1	
Total	SST	n-1		

Para el ejemplo 1, usando todas las variables predictoras, se tiene,

Residual standard error: 3.653 on 77 degrees of freedom

Multiple R-Squared: 0.8733, Adjusted R-squared: 0.8667

F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

Notar que la desviación estimada del error es  $s=3.653=\sqrt{MSE}=\sqrt{13.3}$ . El "P-value" de la Prueba de F es 0.0000, lo cual lleva a la conclusión de que al menos una de las variables predictoras presentes en el modelo es importante para predecir MPG.

El coeficiente de Determinación  $R^2$  tiene la misma interpretación que en regresión lineal simple y se calcula por  $R^2 = \frac{SSR}{SST}$ .

El  $R^2=87.3\%$ , lo que quiere decir que hay un ajuste bastante bueno asumiendo que no hay datos contaminados en el conjunto de datos. El 87.3% de la variación del millaje promedio por galón es explicada por su relación lineal con VOL, HP, SP y WT. El R-Sq(adj) llamado el  $R^2$  ajustado será definido más adelante en el capítulo de selección de variables.

La suma de cuadrados de regresión puede ser particionada en tantas partes como variables predictoras existen en el modelo. Esto es llamado un particionamiento secuencial de la suma de cuadrados de regresión y sirve para determinar la contribución de cada una de las variables predictoras al comportamiento de Y. Formalmente,

$$SSR(\beta_1, \beta_2, \dots, \beta_p / \beta_0) = SSR(\beta_1 / \beta_0) + SSR((\beta_2, \beta_1, \beta_0)) + \dots + SSR(\beta_p / \beta_{p-1}, \dots, \beta_1, \beta_0)$$

Aquí  $SSR(\beta_k / \beta_{k-1}, \dots, \beta_1, \beta_0)$  significa el incremento en la suma de cuadrados de regresión cuando la variable  $X_k$  es incluida en el modelo, el cual ya contiene las variables predictivas  $X_1, \dots, X_{k-1}$ .

La función **anova** de **R** produce estas sumas parciales. Para el ejemplo anterior se obtiene lo siguiente:

```
> 16
```

```
Call:
```

```
lm(formula = mpg ~ vol + hp + sp + wt)
```

```
Coefficients:
```

```
(Intercept)    vol      hp      sp      wt
 192.43775  -0.01565  0.39221 -1.29482 -1.85980
```

```
> anova(16)
```

```
Analysis of Variance Table
```

```

Response: mpg
      Df Sum Sq Mean Sq F value Pr(>F)
vol    1 1101.6 1101.6 82.563 8.172e-14 ***
hp     1 4731.1 4731.1 354.584 < 2.2e-16 ***
sp     1 233.6 233.6 17.509 7.515e-05 ***
wt     1 1013.8 1013.8 75.979 4.221e-13 ***
Residuals 77 1027.4 13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

La Suma de Cuadrados de Regresión es  $7080.1=1101.6+4731.1+233.6+1013.8$ . 233.6 significa que la suma de cuadrado de regresión aumenta en 233.6 cuando la variable SP es añadida al modelo, después que las variables VOL y HP ya están incluidas. El problema ahora es tratar de establecer pruebas para determinar si una variable predictora o un subconjunto de ellas efectivamente debe estar o no en el modelo.

Las sumas de cuadrados de regresión secuenciales varia si se cambia el orden de las anteriores predictoras al momento de ajustar el modelo. Así, si elegimos el orden WT, HP, SP y al final VOL se obtiene el siguiente resultado.

```

> l6<-lm(mpg~wt+hp+sp+vol)
> anova(l6)
Analysis of Variance Table

```

```

Response: mpg
      Df Sum Sq Mean Sq F value Pr(>F)
wt     1 6641.5 6641.5 497.7630 < 2.2e-16 ***
hp     1 35.4 35.4 2.6516 0.1075
sp     1 397.0 397.0 29.7522 5.739e-07 ***
vol    1 6.3 6.3 0.4698 0.4951
Residuals 77 1027.4 13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Notar que las variables HP y VOL no son significativas. Es claro que la variable VOL es la menos importante de las variables predictoras.

### 2.3.3 Prueba de hipótesis para un subconjunto de coeficientes de regresión

Algunas veces estamos interesados en probar si algunos coeficientes del modelo de regresión son iguales a 0 simultáneamente. Por ejemplo, si el modelo tiene  $p$  variables predictoras, quisieramos probar si los  $k$  primeros coeficientes son ceros, o sea  $H_0: \beta_1 = \dots = \beta_k = 0$ . De rechazarse la hipótesis nula implicaría que las  $k$  primeras variables predictoras pueden ser excluidas del modelo.

Al modelo en donde se consideran todas las  $p$  variables se le llama el *modelo completo* y al modelo que queda asumiendo que la hipótesis nula es cierta se le llama el *modelo reducido*.

Es decir, que el modelo reducido sería

$$Y = \beta_{k+1}X_{k+1} + \beta_{k+2}X_{k+2} + \dots + \beta_p X_p + e \quad (2.12)$$

Para probar si la hipótesis nula es cierta se usa una prueba de F que es llamada F-parcial. La prueba de F parcial se calcula por

$$F_p = \frac{\frac{SSR(C) - SSR(R)}{k}}{\frac{SSE(C)}{n - p - 1}} = \frac{SSR(C) - SSR(R)}{MSE(C)}$$

donde  $SSR(C)$  y  $MSE(C)$  representan la suma de cuadrados de regresión y el cuadrado medio del error del modelo completo respectivamente, y  $SSR(R)$  es la suma de cuadrados de regresión del modelo reducido.

$SSR(C) = SSR(\beta_1, \beta_2, \dots, \beta_p / \beta_0)$  y

$SSR(R) = SSR(\beta_{k+1}, \beta_{k+2}, \dots, \beta_p / \beta_0)$

$SSR(C) - SSR(R) = SSR(\beta_1, \beta_2, \dots, \beta_k / \beta_{k+1}, \beta_{k+2}, \dots, \beta_p)$

Esta última diferencia representa el incremento en la suma de cuadrados de regresión cuando  $X_1, \dots, X_k$  son añadidas al modelo en el cual ya están presentes  $X_{k+1}, \dots, X_p$  y la constante

Si  $F_p$  es mayor que  $F_{1-\alpha}$  usando  $k$  grados de libertad para el numerador y  $n-p-1$  para el denominador entonces se rechaza  $H_0$ , en caso contrario se acepta.

**R** no tiene una opción que haga directamente la prueba de  $F$  parcial. Hay que calcular los dos modelos de regresión y usar las sumas de cuadrados de regresión de ambos modelos para calcular la prueba de  $F$  parcial.

**Ejemplo 3:** En el ejemplo 1, probar que las variables VOL y HP no son importantes y pueden ser excluidas del modelo

Haciendo el análisis de regresión sin incluir VOL y HP se obtiene:

```
> l2
Call:
lm(formula = mpg ~ sp + wt, data = millaje)

Coefficients:
(Intercept)          sp           wt
   75.64938    -0.09816    -0.99738

> anova(l2)
Analysis of Variance Table

Response: mpg
      Df Sum Sq Mean Sq F value    Pr(>F)
sp     1  3842.6   3842.6   219.49 < 2.2e-16 ***
wt     1  2881.8   2881.8   164.61 < 2.2e-16 ***
Residuals 79 1383.0     17.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Luego  $SSR(R)=6724.4$  y  $SSR(C)-SSR(R)=7080.1-6724.4=355.7$ .

```
> #Suma de cuadrados de regresion del modelo completo
> a=sum(anova(11)$Sum[-(p+1)])
> #Suma de cuadrados de regresion del modelo reducido
> b=sum(anova(12)$Sum[-(k+1)])
> #Cuadrado Medio del error del modelo completo
> c=anova(11)$Mean[p+1]
> #Calculo del F parcial
> fp<-((a-b)/2)/c
> fp
[1] 13.32720
```

Luego la F-parcial será

$$F_p = \frac{355.63}{2} = \frac{177.81}{13.34} = 13.33$$

Usando un nivel de significación del 5%, Hay que comparar  $F_p$  con  $F(.95,2,77)$ .

```
> #Hallando el percentil de la F con alpha=.05
> qf(.95,k,n-p-1)
[1] 3.115366
```

Luego  $F_p > F(.95,2,77)=3.11$  por lo tanto se rechaza la prueba y se concluye que VOL y HP no pueden ser eliminadas simultáneamente, al menos una de ellas es importante.

### 2.3.4 Intervalos de Confianza y de Predicción en Regresión Lineal Múltiple.

Supongamos que se desea predecir el valor medio de la variable de respuesta  $Y$  para una combinación predeterminada de las variables predictoras  $X_1, \dots, X_p$ . Consideremos el vector  $\mathbf{x}'_o = (1, x_{1,0}, \dots, x_{p,0})$  donde  $x_{1,0}, \dots, x_{p,0}$  son los valores observados de  $X_1, \dots, X_p$  respectivamente.

El valor predicho para el valor medio de la variable de respuesta  $Y$  será  $\hat{y}_o = \mathbf{x}'_o \hat{\boldsymbol{\beta}}$ . De donde,

$Var(\hat{y}_o) = \mathbf{x}'_o \mathbf{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}_o = \sigma^2 \mathbf{x}'_o (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_o$ . En consecuencia asumiendo que los errores están normalmente distribuidos se tiene que un intervalo del  $100(1-\alpha)\%$  para el valor medio de  $Y$  dado que  $\mathbf{x}=\mathbf{x}_o$  es de la forma

$$\hat{y}_o \pm t_{(\alpha/2, n-p-1)} s \sqrt{\mathbf{x}'_o (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_o} \quad (2.13)$$

También usando la misma derivación que se hizo en el caso de regresión lineal simple se llega a establecer que un intervalo de confianza (más conocido como intervalo de predicción) del  $100(1-\alpha)\%$  para el valor individual de  $Y$  dado  $\mathbf{x}=\mathbf{x}_o$  es de la forma

$$\hat{y}_o \pm t_{(\alpha/2, n-p-1)} s \sqrt{1 + \mathbf{x}'_o (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_o} \quad (2.14)$$

**Ejemplo 4:** Usando el conjunto de datos **millaje**, hallar un intervalo de confianza del 95% para el millaje promedio por galón de todos los vehículos con capacidad interior de 90 pies cúbicos, un

HP de 50 una velocidad máxima de 1200 millas por galón y un peso de 20,000 libras. Hallar un intervalo de predicción para el millaje de un carro con las mismas características anteriores.

Usando R se obtiene

```
> # hallando el intervalo de confianza del 95% para el valor medio
> sp<-100
> wt<-20
> vol<-90
> hp<-50
> nuevo<-as.data.frame(cbind(sp,wt,vol,hp))
> nuevo
  sp wt vol hp
1 100 20 90 50
> predict.lm(l1,nuevo,se.fit=T,interval=c("confidence"),level=.95)
$fit
      fit   lwr   upr
[1,] 43.9624 42.41585 45.50894
```

```
$se.fit
[1] 0.7766682
```

```
$df
[1] 77
```

```
$residual.scale
[1] 3.652755
```

```
> #Hallando el ntervalo de prediccion del 99% para los mismos datos
> predict.lm(l1,nuevo,se.fit=T,interval=c("prediction"),level=.99)
$fit
      fit   lwr   upr
[1,] 43.9624 34.09908 53.82571
```

```
$se.fit
[1] 0.7766682
```

```
$df
[1] 77
```

```
$residual.scale
[1] 3.652755
```

Hay un 95% de confianza de que el millaje promedio de todos los carros con las características dadas caiga entre 42.41 y 45.50 millas por galón. Hay un 99% de confianza de que el rendimiento de millas por galón de un carro con las características mencionadas caiga entre 34.09 y 53.82

### 2.3.5 La prueba de Falta de Ajuste

Es una prueba que se usa para determinar si la forma del modelo que se está considerando es adecuada. Es decir, si la regresión debe o no incluir términos potencias o interacciones entre las

variables predictoras. En el caso de regresión simple la prueba requiere que haya por lo menos un valor de la variable predictora con varias observaciones de  $y$ . En regresión múltiple se debe suponer que hay  $m$  combinaciones distintas de las  $n$  observaciones de las  $p$  variables predictoras y que por cada una de esas combinaciones hay  $n_i$  ( $i=1, \dots, m$ ) observaciones de la variable de respuesta, es decir,  $\sum_{i=1}^m n_i = n$ .

La Suma de Cuadrados del Error se particiona de la siguiente manera

$$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_i - \hat{y}_i)^2 \quad (2.15)$$

donde  $\hat{y}_i$  es el valor predicho por el modelo de regresión para la  $i$ -ésima combinación de las variables predictoras, mientras que  $\bar{y}_i$  es el valor promedio de la variable predictora para la  $i$ -ésima combinación.

La primera suma de cuadrados del lado derecho es llamado la **Suma de Cuadrados del Error Puro (SSPE)** y tiene  $n-m$  grados de libertad. Si no hubiera varios valores de la variable de respuesta por cada combinación de las predictoras esta suma sería cero. Se puede demostrar que el valor esperado del cuadrado medio del error puro es igual a la varianza poblacional  $\sigma^2$ , sea o no sea el modelo de regresión adecuado.

La segunda suma de cuadrados que también puede ser escrita como  $\sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2$  es llamada la **Suma de Cuadrados de Falta de Ajuste (SSLOF)** y tiene  $m-p-1$  grados de libertad. Si el modelo especificado es correcto entonces el valor esperado del cuadrado medio de Falta de Ajuste es igual a  $\sigma^2$ . Si le faltan términos al modelo (por ejemplo: potencias, productos de variables, etc.) entonces el cuadrado medio de la falta de ajuste sobreestimaré a  $\sigma^2$ .

En resumen, la hipótesis nula será  $H_0$ : El modelo es adecuado (no hay falta de ajuste) versus  $H_a$ : el modelo no es adecuado y la prueba estadística es una prueba de  $F$  dada por

$$F = \frac{SSLOF / (m - p - 1)}{SSPE / (n - m)} = \frac{MSLOF}{MSPE}$$

que se distribuye como una  $F(m-p-1, n-m)$  si la hipótesis nula es cierta. La hipótesis nula es rechazada si el valor de la prueba estadística es mayor que  $F(1-\alpha, m-p-1, n-m)$ .

R no tiene una función para calcular directamente la prueba de bondad de ajuste. Hay que introducir una variable adicional que identifique los valores de  $y$  y correspondiente al mismo valor de la variable predictora.

**Ejemplo 5:** Usando el conjunto de datos **millaje**, hacer una prueba de Falta de Ajuste si se considera la variable de respuesta MPG y la variable predictora HP.

Usando el laboratorio 8 del apéndice del texto se obtiene los siguientes resultados

```
> millajelf=millaje[,c(1,5)]
> table(millajelf$hp)
```

49 52 53 55 62 66 70 73 74 78 80 81 84 90 92 93 95 96 98 100

```

1 1 1 3 2 1 2 3 1 2 1 2 2 4 7 1 5 1 1 4
102 103 115 120 130 140 145 150 160 162 165 175 180 236 238 245 263 280 295 322
5 1 5 1 3 4 1 1 1 2 4 1 1 1 1 1 1 1 1 1

```

```

># Hay m=40 valores distintos de la predictora
># anadiendo una columna var3 que identifica a que grupo pertenece cada
># observación
> millajelf[1:10,]
  mpg hp var3
1 65.4 49 1
2 56.0 55 4
3 55.9 55 4
4 49.0 70 7
5 46.5 53 3
6 46.2 70 7
7 45.4 55 4
8 59.2 62 5
9 53.3 62 5
10 43.4 80 11
.....
.....
> #haciendo la regresion lineal simple
> l1=lm(mpg~hp,data=millajelf)
> l1

```

```

Call:
lm(formula = mpg ~ hp, data = millajelf)

```

```

Coefficients:
(Intercept)      hp
 50.0661      -0.1390

```

```

> anova(l1)
Analysis of Variance Table

```

```

Response: mpg
      Df Sum Sq Mean Sq F value Pr(>F)
hp      1 5058.0 5058.0 132.69 < 2.2e-16 ***
Residuals 80 3049.4 38.1
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

> #Haciendo el analisis de varianza de claificacion simple de mpg
># entre los 40 grupos diferentes
> l2=lm(mpg~factor(var3),data=millajelf)
> anova(l2)
Analysis of Variance Table

```

```

Response: mpg
      Df Sum Sq Mean Sq F value Pr(>F)
factor(var3) 39 7794.4 199.9 26.809 < 2.2e-16 ***
Residuals 42 313.1 7.5

```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> # Haciendo el anova para comparar los dos modelos . Se extrae la suma de cuadrados del
># error Puro y la suma de cuadrados de falta de Ajuste.
```

```
>#
```

```
>anova(11,12)
```

```
Analysis of Variance Table
```

```
Model 1: mpg ~ hp
```

```
Model 2: mpg ~ factor(var3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	80	3049.44				
2	42	313.11	38	2736.33	9.6592	1.703e-11 ***

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La suma de cuadrados de Error Puro es 313.11 y la suma de cuadrados de Falta de ajuste es 2736.33. Si usamos el “P-value” de la prueba F de Falta de Ajuste se concluye que se rechaza la hipótesis nula, es decir, hay suficiente evidencia para concluir que faltan términos en el modelo.

MINITAB y otros programas estadísticos dan el “P-value” de esta prueba. MINITAB además da una prueba de Falta de Ajuste que no requiere que hayan varios valores de la variable de respuesta para cada combinación. La prueba de Falta de ajuste que da MINITAB es más informativa aún, dice que hay una posible curvatura en HP (ver el plot de la Figura 2.2), que hay outliers en la dirección de la variable predictora y finalmente da una prueba de ajuste global.

Consideremos ahora una prueba de ajuste usando todas las variable predictoras

```
> millajep=millaje[,2:5]
```

```
> dim(unique(millajep))
```

```
[1] 70 4
```

```
># Hay m=70 valores distintos de la predictora
```

```
># anadiendo una columna var4 que identifica a que grupo pertenece cada
```

```
># observación
```

```
> millajelf=cbind(millaje,var4)
```

```
> millajelf[1:10,]
```

	mpg	sp	wt	vol	hp	var4
1	65.4	96	17.5	89	49	1
2	56.0	97	20.0	92	55	2
3	55.9	97	20.0	92	55	2
4	49.0	105	20.0	92	70	3
5	46.5	96	20.0	92	53	4
6	46.2	105	20.0	89	70	5
7	45.4	97	20.0	92	55	2
8	59.2	98	22.5	50	62	6
9	53.3	98	22.5	50	62	6
10	43.4	107	22.5	94	80	7

```
># Haciendo ;la regresion lineal multiple
```

```
> l2=lm(mpg~sp+wt+vol+hp,data=millajelf)
```

```

> anova(12)
Analysis of Variance Table

Response: mpg
      Df Sum Sq Mean Sq F value   Pr(>F)
sp      1 3842.6  3842.6 287.9944 < 2.2e-16 ***
wt      1 2881.8  2881.8 215.9879 < 2.2e-16 ***
vol     1  46.0   46.0   3.4451  0.06727 .
hp      1 309.7   309.7  23.2093 7.131e-06 ***
Residuals 77 1027.4   13.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Haciendo el analisis de varianza de claificacion simple de mpg
># entre los 70 grupos diferentes

> l3=lm(mpg~factor(var4),data=millajelf)
> anova(13)
Analysis of Variance Table

Response: mpg
      Df Sum Sq Mean Sq F value   Pr(>F)
factor(var4) 69 7990.3  115.8  11.859 2.130e-05 ***
Residuals  12  117.2    9.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Haciendo el anova para comparar los dos modelos . Se extrae la suma de cuadrados del
># error Puro y la suma de cuadrados de falta de Ajuste.
>#
> anova(12,13)
Analysis of Variance Table

Model 1: mpg ~ sp + wt + vol + hp
Model 2: mpg ~ factor(var4)
  Res.Df  RSS Df Sum of Sq  F Pr(>F)
1     77 1027.38
2     12  117.18 65   910.20 1.434 0.2518

```

La clásica prueba de Falta de Ajuste acepta la hipótesis nula, es decir, que no hay suficiente evidencia para concluir que haya Falta de Ajuste. Sin embargo, la prueba de Falta de ajuste de MINITAB es mas infomativa y concluye que hay interacción entre las variables predictoras HP y SP, que hay que transformar WT y además hay outliers.

## EJERCICIOS

1. Propiedades de la matriz HAT  $H = X(X'X)^{-1}X'$

a) La traza de una matriz es igual a la suma de los elementos que están en su diagonal. Probar que  $\text{Traza}(H) = p'$  con  $p' = p + 1$ , donde  $p$  es el número de variables predictoras.

b) Probar que  $\frac{1}{n} \leq h_{ii} \leq 1$ , donde  $h_{ii}$  es el  $i$ -ésimo elemento de la diagonal de  $H$ .

c) Probar que  $H\mathbf{1}_n = \mathbf{1}_n$  aquí  $\mathbf{1}_n$  es un vector columna cuyo elementos son todos unos

2. Usar el conjunto de datos **Fuel** con variable de respuesta Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas. Los datos están disponible en la página de internet del texto

a) Hallar la variable que tiene correlación más alta con la variable de respuesta

b) Hacer un plot matricial para ver si no hay outliers y determinar si el coeficiente de correlación es confiable.

c) Hacer una regresión lineal de  $Y$  versus la variable determinada en los pasos a y b y tratar otros modelos: cuadrático, exponencial, logaritmico para mejorar el  $R^2$ , si es posible

d) Hallar un Intervalo de Confianza del 99% para el valor medio y el valor Predicho de  $Y$ , escogiendo un valor adecuado de la variables predictoras usando el modelo lineal. Trazar las bandas de confianza. Comentar sus resultados.

e) Hallar el modelo de regresión múltiple considerando todas las variables predictoras e interpretar los coeficientes de regresión.

f) Interpretar el coeficiente de Determinación  $R^2$ .

g) Probar que todos los coeficientes del modelo de regresión son ceros. Comentar el resultado.

h) Probar que cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.

3. Sea  $\hat{\beta} = (X'X)^{-1}X'Y$ . Probar que

$$(Y - X\hat{\beta})'(Y - X\hat{\beta}) = (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) + Y'(I - H)Y$$

y en consecuencia  $\hat{\beta}$  es el estimador mínimo cuadrático.

4. **Uso de la factorización QR en Regresión** Supongamos que tenemos una matriz ortogonal  $Q$  de orden  $n \times p'$  (es decir,  $Q'Q = I$ ) y una matriz triangular superior  $R$  tal que  $QR = X$

a) Probar que  $R'R = X'X$

b) Escribir el estimador minimocuadrático  $\hat{\beta}$  en términos de  $Q$ ,  $R$  y  $Y$ . Cual sería la ventaja de usar esta fórmula con respecto a la fórmula original.?

c) Expresar  $\hat{Y}$  y  $\hat{e}$  en términos de  $Y$  y  $Q$ .

5. Usar el conjunto de datos **Highway**, con variable de respuesta es TASA y todas las otras como variables predictoras para responder las siguientes preguntas. Los datos están en la página de internet del texto.

a) Hallar la variable que tiene correlación más alta con la variable de respuesta

- b) Hacer un plot matricial para ver si no hay outliers y determinar si el coeficiente de correlación es confiable
- c) Hacer una regresión lineal de Y versus la variable determinada en los pasos a y b y tratar otros modelos: cuadrático, exponencial, logaritmico para mejorar el  $R^2$ , si es posible
- d) Hallar el modelo de regresión múltiple considerando todas las variables predictoras e interpretar los coeficientes de regresión.
- e) Interpretar el coeficiente de Determinación  $R^2$ .
- f) Probar que todos los coeficientes del modelo de regresión son ceros. Comentar el resultado.
- g) Probar que cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.
- h) Hallar las dos variables que están menos correlacionadas con la variable de respuesta y probar la hipótesis de que ambas variables deben ser excluidas simultáneamente del modelo.
- i) Hallar un Intervalo de Confianza para el valor medio de Y y el valor Predicho del 99% para Y, escogiendo valores adecuados de las variables predictoras. Comentar sus resultados.
- j) Hacer un análisis de falta de ajuste usando como variable predictora, aquella obtenida en a).

6. Considerando un modelo de regresión lineal múltiple probar que  $E[s^2]=\sigma^2$  donde  $s^2$  es la varianza estimada del error definida en la ecuación 2.9.

7. Verificar la identidad de la ecuación 2.11

**8. Efecto de subajuste.** Supongamos que se ajusta el modelo  $Y=X\beta+e$  donde  $X$  es una matriz  $n \times r$  cuando en realidad el modelo verdadero incluye  $s$  adicionales variables predictoras contenidas en la matriz  $Z$ . Es decir, que el verdadero modelo es  $Y=X\beta+Z\gamma+e$ . Mostrar que en general el estimador mínimos cuadrático  $\hat{\beta}$  usando el modelo reducido es sesgado. Asimismo mostrar que el estimador de la varianza es sesgado. Bajo que condiciones ambos estimadores serían insesgado?

9. Supongamos que se ha obtenido la siguiente regresión usando una muestra de 75 observaciones

$$Y = -5.16 + .325X_1 + 5.55X_2 + .3X_3 + .01X_4 + 8.75X_5 - .97X_6$$

- a) Interpretar cualquiera de los coeficientes de las variables predictoras
- b) Hallar el valor de la prueba estadística de F si el coeficiente de determinación  $R^2=.95$
- c) Explicar detalladamente como se probaría la hipótesis  $H_0:\beta_1=\beta_3=\beta_4=\beta_6$

**10. Efecto de sobreajuste.** Supongamos que realmente el modelo  $Y=X\beta+e$  (1) ajusta a nuestro conjunto de datos. Reescribamos el modelo anterior por

$$Y=X_1\beta_1+X_2\beta_2+e, \text{ donde } X=(X_1 | X_2)$$

$X_1$  es de orden  $n \times k$  y  $X_2$  es de orden  $n \times (p-k)$ ,  $n$  es el número de observaciones y  $p$  es el número de parámetros del modelo, es decir el número de variables predictoras más el intercepto.

Consideremos que en lugar del modelo (1) se usa el siguiente modelo para ajustar los datos

$$Y=X_1\beta_1+e \quad (2)$$

- a) Hallar el esperado del estimador mínimo cuadrático de  $\beta_1$  usando el modelo (2), pero considerando que realmente (1) es el que se cumple.
- b) Hallar el esperado del estimador minimocuadrático de la varianza estimada usando el modelo (2).

# CAPÍTULO 3

## ANOMALIAS EN REGRESIÓN Y MEDIDAS REMEDIALES

En este capítulo se estudiarán diversos diagnósticos de regresión que nos permitan verificar si las suposiciones del modelo de regresión lineal se cumplen. Algunos de estos diagnósticos están basados en medidas que envuelven residuales y otros en plots de los residuales.

### 3.1 “Outliers”, puntos de leverage alto y valores influenciales

Una observación  $(y^*, x^*_1, \dots, x^*_p)$  es considerado un “outlier” si está bastante alejado de la mayoría de los datos sea en la dirección vertical o en la horizontal. Sin embargo, la mayoría de los textos llaman “outlier” a un valor alejado solamente en la dirección vertical y **punto de leverage alto** a una observación alejada en la dirección horizontal.

Una observación  $(y^*, x^*_1, \dots, x^*_p)$  es considerado un **valor inflencial** si su presencia afecta tremendamente el comportamiento del modelo. Por ejemplo, en el caso de regresión simple remover un valor inflencial podría cambiar dramáticamente el valor de la pendiente.

Consideremos el siguiente conjunto de datos, consistente de 8 observaciones

X	4	5	7	9	12	14	16	35
Y	6	7	12	15	18	21	28	65

La figura 3.1 muestra el plot de los datos. Notar que el punto O es un “outlier” y punto de leverage alto, pero a través de cálculos mostraremos que no es un valor inflencial.

Primero, calcularemos la ecuación de regresión con el dato “outlier”

```
> l1=lm(y~x)
> summary(l1)
```

```
Call:
lm(formula = y ~ x)
```

```
Residuals:
  Min   1Q Median   3Q   Max
-2.8825 -0.3140  0.4765  1.1130  1.4595
```

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.80152   1.03618  -2.704  0.0354 *
x           1.90600   0.06567  29.026 1.11e-07 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.727 on 6 degrees of freedom
Multiple R-Squared: 0.9929, Adjusted R-squared: 0.9918
```

F-statistic: 842.5 on 1 and 6 DF, p-value: 1.108e-07

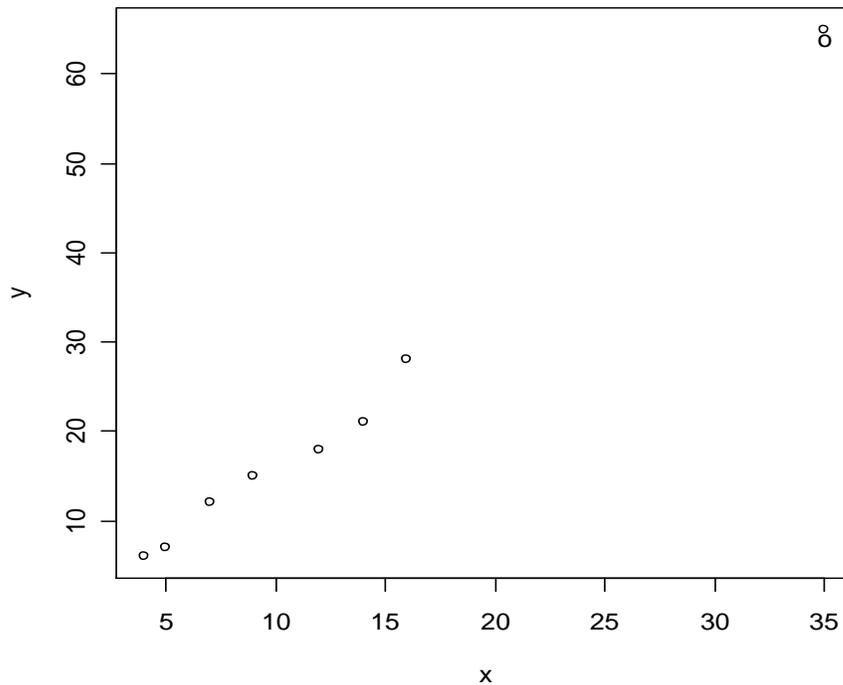


Figura 3.1. Ejemplo de una observación que es “outlier” y punto leverage alto pero que no es influyente.

Por otro lado la regresión sin el dato “outlier” es:

```
> x1=x[-8]
> y1=y[-8]
> l2=lm(y1~x1)
> summary(l2)
```

Call:

```
lm(formula = y1 ~ x1)
```

Residuals:

```
 1    2    3    4    5    6    7
0.1034 -0.5818  1.0477  0.6773 -1.3784 -1.7489  1.8807
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.8443    1.3465  -0.627  0.558
x1           1.6852    0.1286  13.101 4.62e-05 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.442 on 5 degrees of freedom

Multiple R-Squared: 0.9717, Adjusted R-squared: 0.966  
 F-statistic: 171.6 on 1 and 5 DF, p-value: 4.625e-05

Notar que la pendiente y el  $R^2$  han cambiado solo ligeramente. En consecuencia, la observación es un “outlier” y punto de leverage alto pero no es es influyente.

Supongamos ahora que al conjunto de datos anterior y al cual se le eliminó el “outlier”, se le agrega el dato (35,22) que es considerado un punto de leverage alto. El plot del conjunto de datos es mostrado en la figura 3.2, donde la observación 0 representa el dato con leverage alto. Mostraremos que esta observación si resulta ser influyente.

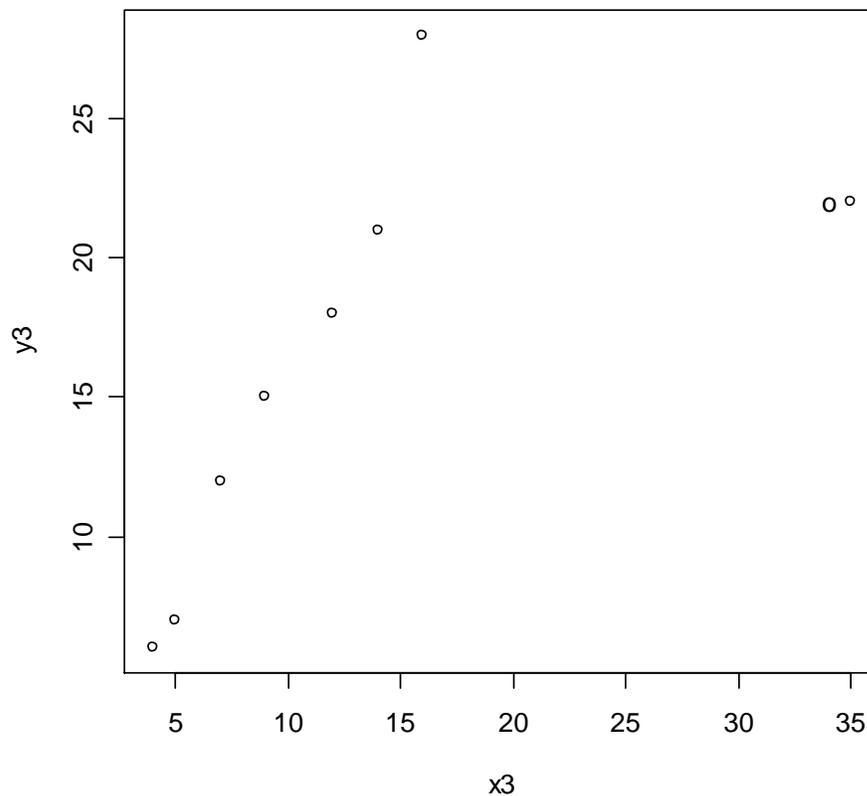


Figura 3.2. Ejemplo de una observación que es punto de leverage alto y que también es influyente.

La ecuación de regresión considerando el dato de leverage alto es

```
> l3=lm(y3~x3)
> summary(l3)
```

```
Call:
lm(formula = y3 ~ x3)
```

```
Residuals:
```

```

      Min      1Q  Median      3Q      Max
-5.7487 -5.1957 -0.1435  2.7556 10.1772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.4642      3.6183   2.616  0.0398 *
x3           0.5224      0.2293   2.278  0.0629 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.03 on 6 degrees of freedom
Multiple R-Squared:  0.4638,    Adjusted R-squared:  0.3745
F-statistic: 5.191 on 1 and 6 DF,  p-value: 0.06295

```

Se puede observar el gran efecto sobre el  $R^2$  que baja de 97.2% a 46.4% y un cambio drástico en la pendiente que cambia de 1.69 a 0.522..

En consecuencia un “outlier” vertical y/o punto de leverage alto puede ser influyente o no serlo. Por otro lado si una observación es influyente entonces es un “outlier” vertical o un punto de leverage alto.

### 3.2 Residuales y detección de “outliers”.

Consideremos el modelo de regresión lineal múltiple  $\mathbf{Y}=\mathbf{X}\mathbf{B}+\mathbf{e}$ , donde  $E(\mathbf{e})=\mathbf{0}$  y  $\text{Var}(\mathbf{e})=\sigma^2\mathbf{I}$ . Luego,  $\widehat{\mathbf{Y}}=\mathbf{X}\widehat{\boldsymbol{\beta}}$ , pero como  $\widehat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ , se tiene que  $\widehat{\mathbf{Y}}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=\mathbf{H}\mathbf{Y}$ , ésta es la razón por la que a  $\mathbf{H}$  se le llama la matriz HAT (sombrero), ya que actúa como una transformación de  $\mathbf{Y}$  a  $\widehat{\mathbf{Y}}$ . En particular,  $\widehat{y}_i=\sum_{j=1}^n h_{ij}y_j$ , donde  $h_{ij}$  es el elemento de la matriz  $\mathbf{H}$  que está en la  $i$ -ésima fila y  $j$ -ésima columna.

Luego, el vector de residuales  $\widehat{\mathbf{e}}=\mathbf{Y}-\widehat{\mathbf{Y}}=\mathbf{Y}-\mathbf{H}\mathbf{Y}=(\mathbf{I}-\mathbf{H})\mathbf{Y}$ . En particular,

$$\widehat{e}_i=y_i-\sum_{j=1}^n h_{ij}y_j.$$

#### 3.2.1 Media y Varianza del vector de residuales

Notar que

$$E(\widehat{\mathbf{e}})=(\mathbf{I}-\mathbf{H})E(\mathbf{Y})=(\mathbf{I}-\mathbf{H})\mathbf{X}\mathbf{B}=\mathbf{X}\mathbf{B}-\mathbf{H}\mathbf{X}\mathbf{B}=\mathbf{X}\mathbf{B}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{B}=\mathbf{X}\mathbf{B}-\mathbf{X}\mathbf{B}=\mathbf{0}$$

Por otro lado,

$$\text{Var}(\widehat{\mathbf{e}})=\text{Var}[(\mathbf{I}-\mathbf{H})\mathbf{Y}]=(\mathbf{I}-\mathbf{H})\text{Var}(\mathbf{Y})(\mathbf{I}-\mathbf{H})'=\sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})'=\sigma^2(\mathbf{I}-\mathbf{H})^2=\sigma^2(\mathbf{I}-\mathbf{H})$$

Aquí se ha usado el hecho que  $\mathbf{I}-\mathbf{H}$  es simétrica e idempotente, como se vió en la sección 2.2.3.

En particular,  $\text{Var}(\widehat{e}_i)=\sigma^2(1-h_{ii})$ . Esta varianza es estimada por  $s^2(1-h_{ii})$ .

Asimismo,  $Cov(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2$

Notar que :

- a) Tanto los errores  $e_i$  como los residuales tienen media 0.
- b) La varianza de los errores es constante, pero la de los residuales no lo es.
- c) Los errores no están correlacionados, pero los residuales sí lo están.

### 3.2.2 Residuales Estudentizados internamente

Para reducir el efecto de las varianzas de los residuales es más conveniente trabajar con versiones estandarizadas de ellos. Así, el **residual estudentizado internamente** se define por

$$r_i^* = \frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}} \quad (3.1)$$

La covarianza de los residuales estudentizados es igual a

$$Cov(r_i^*, r_j^*) = Cov\left(\frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}}, \frac{\hat{e}_j}{\sigma\sqrt{1-h_{jj}}}\right) = \frac{Cov(\hat{e}_i, \hat{e}_j)}{\sigma^2\sqrt{(1-h_{ii})(1-h_{jj})}} = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}$$

En algunos programas estadísticos como MINITAB y el toolbox estadístico de MATLAB los  $r_i^*$  son llamados **residuales estandarizados**.

### 3.2.3 Residuales estudentizados externamente

Supongamos que la  $i$ -ésima observación es eliminada del conjunto de datos y que se ajusta el modelo lineal con las  $n-1$  observaciones que quedan. Sean  $\hat{\boldsymbol{\beta}}_{(i)}$  y  $s_{(i)}^2$  las estimaciones de los parámetros del modelo y de la varianza de los errores respectivamente. Usando la siguiente identidad debido a Gauss

$$(\mathbf{X}'_{(i)} \mathbf{X}_{(i)})^{-1} = (\mathbf{X}' \mathbf{X})^{-1} + \frac{(\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_i' (\mathbf{X}' \mathbf{X})^{-1}}{1-h_{ii}} \quad (3.2)$$

donde  $\mathbf{X}_{(i)}$  representa a la matriz  $\mathbf{X}$  sin su  $i$ -ésima fila  $\mathbf{x}_i'$ , se puede establecer las siguientes relaciones entre  $\hat{\boldsymbol{\beta}}$  y  $\hat{\boldsymbol{\beta}}_{(i)}$  y entre  $s^2$  y  $s_{(i)}^2$

$$i) \hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i \hat{e}_i}{1-h_{ii}}$$

$$ii) s_{(i)}^2 = \frac{n-p-1}{n-p-2} s^2 - \frac{\hat{e}_i^2}{(n-p-2)(1-h_{ii})}$$

La identidad de Gauss es un caso particular de la **Identidad de Sherman-Morrison-Woodbury** (1950)

$$(\mathbf{A} \pm \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} \mp \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 \pm \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}} \quad (3.3)$$

donde  $\mathbf{A}$  es una matriz cuadrada no singular  $n \times n$ , y  $\mathbf{u}$  y  $\mathbf{v}$  son dos vectores de dimensión  $n$ .

En efecto, puesto que  $\mathbf{X}'_{(i)}\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}'_i\mathbf{x}_i$ , donde  $\mathbf{x}'_i$  es la  $i$ -ésima fila de  $\mathbf{X}$ , se puede tomar  $\mathbf{A} = \mathbf{X}'\mathbf{X}$  y  $\mathbf{u} = \mathbf{v} = \mathbf{x}_i$ , y se obtiene (3.2).

Si  $\tilde{y}_i$  representa el valor estimado de la variable de respuesta para la  $i$ -ésima observación entonces  $\tilde{y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{(i)}$ . Como la  $i$ -ésima observación no fue usada en la estimación del modelo entonces  $y_i$  y  $\tilde{y}_i$  son independientes. Luego la varianza del residual  $y_i - \tilde{y}_i$  está dada por

$$\text{Var}(y_i - \tilde{y}_i) = \text{Var}(y_i) + \text{Var}(\tilde{y}_i) = \sigma^2 + \sigma^2 \mathbf{x}'_i (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} \mathbf{x}_i \quad (3.4)$$

Estimando  $\sigma^2$  por  $s_{(i)}^2$  y considerando que si  $y_i$  no es un outlier entonces  $E(y_i - \tilde{y}_i) = 0$  se obtiene

$$t_i = \frac{y_i - \tilde{y}_i}{s_{(i)} \sqrt{1 + \mathbf{x}'_i (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} \mathbf{x}_i}} \quad (3.5)$$

$t_i$  es llamado un **residual estudentizado externamente** y tiene  $n-p-2$  grados de libertad.

**Propiedad:** Relación entre el residual usual y el residual usando un modelo eliminando la  $i$ -ésima observación

$$y_i - \tilde{y}_i = \frac{\hat{\epsilon}_i}{1 - h_{ii}} \quad (3.6)$$

**Prueba:** Sustituyendo  $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} \mathbf{X}'_{(i)}\mathbf{y}_{(i)}$  en

$$y_i - \tilde{y}_i = y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}_{(i)} \quad (3.7)$$

y usando luego la identidad de Gauss (3.2) se obtiene

$$\begin{aligned} y_i - \tilde{y}_i &= y_i - \mathbf{x}'_i \left[ (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}} \right] \mathbf{X}'_{(i)} \mathbf{y}_{(i)} \\ &= y_i - \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_{(i)} \mathbf{y}_{(i)} - \frac{h_{ii} \mathbf{x}'_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_{(i)} \mathbf{y}_{(i)}}{1 - h_{ii}} \end{aligned}$$

$$\begin{aligned}
&= \frac{(1-h_{ii})y_i - (1-h_{ii})\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)} - h_{ii}\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}}{1-h_{ii}} \\
&= \frac{(1-h_{ii})y_i - \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}}{1-h_{ii}}
\end{aligned}$$

Si se usa luego el hecho que  $\mathbf{X}'_{(i)}\mathbf{y}_{(i)} + \mathbf{x}'_i y_i = \mathbf{X}'\mathbf{y}$ , la anterior relación es equivalente a .

$$y_i - \tilde{y}_i = \frac{(1-h_{ii})y_i - \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{x}'_i y_i)}{1-h_{ii}}$$

como  $\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \hat{y}_i$  se obtiene

$$\begin{aligned}
y_i - \tilde{y}_i &= \frac{(1-h_{ii})y_i - \hat{y}_i + h_{ii}y_i}{1-h_{ii}} \\
&= \frac{y_i - \hat{y}_i}{1-h_{ii}} = \frac{\hat{e}_i}{1-h_{ii}}
\end{aligned}$$

Lo cual concluye la prueba.

Asímismo, se puede establecer la siguiente relación entre los distintos tipos de residuales

$$t_i = \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}} = r_i^* \left( \frac{n-p-2}{n-p-1-r_i^{*2}} \right)^{1/2} \quad (3.8)$$

### 3.3 Diagnósticos para detectar “outliers” y puntos de leverage alto

Ahora consideraremos diagnósticos basados en medidas y que servirán para detectar si una observación es un ”outlier” o un punto de leverage alto. Los diagnósticos más básicos son:

Si  $|h_{ii}| > 2p/n$  (algunos usan  $3p/n$ . Aquí  $p$  es el número de parámetros) entonces la  $i$ -ésima observación es considerado un “punto leverage” y pudiera ser influyente

Si  $|t_i| > 2$  ( o si  $|r_i| > 2$ ) entonces la  $i$ -ésima observación es considerada un “outlier” y también puede ser influyente.

A continuación definiremos otros diagnósticos más sofisticados:

**i) La Distancia Cook** (Cook, 1977): Mide el cambio que ocurriría en el vector  $\hat{\beta}$  de coeficientes estimados de regresión (y por lo tanto en el valor ajustado de la variable de respuesta) si la  $i$ -ésima observación fuera omitida. Se calcula por

$$CD_i^2 = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)})}{ps^2} = \frac{(\hat{Y} - \hat{Y}_{(i)})' (\hat{Y} - \hat{Y}_{(i)})}{ps^2} = r_i^{*2} \frac{h_{ii}}{p(1-h_{ii})} \quad (3.9)$$

La primera igualdad resulta del hecho que

$$(\hat{\beta} - \hat{\beta}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)}) = [\mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)})]' [\mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)})] = (\hat{Y} - \hat{Y}_{(i)})' (\hat{Y} - \hat{Y}_{(i)})$$

La segunda identidad resulta de la relación entre  $\hat{\beta}$  y  $\hat{\beta}_{(i)}$  mencionada en la sección 3.2.3, lo cual implica que

$$CD_i^2 = \frac{e_i^2}{(1-h_{ii})^2} \frac{x_i' (\mathbf{X}' \mathbf{X})^{-1} x_i}{ps^2} = \frac{h_{ii}}{p(1-h_{ii})} \frac{e_i^2}{(1-h_{ii})s^2} = r_i^{*2} \frac{h_{ii}}{p(1-h_{ii})}$$

Notar que si el residual estandarizado es muy grande y si el valor leverage es grande, es decir si la observación está bien alejado en la dirección vertical y horizontal entonces su distancia Cook es bien grande. En general un  $CD_i^2 > 1$  indica que la  $i$ -ésima observación es potencialmente influyente. Una observación con  $CD_i^2 < 0.1$  no merece ninguna discusión y si su  $CD_i^2 < 0.5$  merece un poco de atención. Más específicamente una observación con  $CD_i^2 > F(0.50, p, n-p)$  es considerado como un valor influyente, la razón es que  $\hat{\beta}$  cae en un elipsoide de confianza centrado en  $\hat{\beta}$  de radio  $F(\alpha, p, n-p)$ . Aquí  $p$  es el número de coeficientes en el modelo. Sin embargo si todos los  $CD_i^2$  son menores que 1 es mejor plotear los valores  $CD_i^2$  para detectar si hay observaciones con valores grandes comparados con los demás.

**ii) DFFITS** (Belsley, Kuh, y Welsch, 1980). Es similar a la Distancia Cook, excepto por un factor de escala y el remplazo de la varianza estimada  $s^2$  por  $s_{(i)}^2$ , la varianza estimada del error excluyendo la  $i$ -ésima observación en los cálculos. Más precisamente,

$$DFFITS_i^2 = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})' (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{s_{(i)}^2} = t_i^2 \frac{h_{ii}}{(1-h_{ii})} \quad (3.10)$$

En forma similar a lo realizado en la segunda identidad de la Distancia Cook, se obtiene que

$$DFFITS_i^2 = \frac{h_{ii} e_i^2}{(1-h_{ii})^2 s_{(i)}^2} = t_i^2 \frac{h_{ii}}{(1-h_{ii})}$$

Un  $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$  indica un posible valor influyente. Notar que

$$CD_i^2 = \frac{r_i^2}{pt_i^2} DFFITS_i^2 \quad (3.11)$$

**iii) DFBETAS** (Belsley, Kuh, y Welsch, 1980). Mide la influencia de la  $i$ -ésima observación en cada uno de los coeficientes de regresión. Se calcula por

$$(DFBETAS)_{ji} = \frac{\beta_j - \beta_{j,(i)}}{s_{(i)} \sqrt{c_{jj}}} \quad (3.12)$$

( $i=1,..n, j=0,..p$ ), donde  $c_{jj}$  es el  $j$ -ésimo elemento de la diagonal de  $(\mathbf{X}'\mathbf{X})^{-1}$ .

Un  $|DFBETAS|_{ji} > \frac{2}{\sqrt{n}}$  indica un posible valor influyente.

**iv) COVRATIO** (Belsley, Kuh, y Welsch, 1980)

Mide el efecto en la variabilidad de los coeficientes de regresión al remover la  $i$ -ésima observación. Se define por

$$COVRATIO_i = \frac{\det[s_{(i)}^2 (X'_{(i)} X_{(i)})^{-1}]}{\det[s^2 (X' X)^{-1}]} \quad (3.13)$$

para  $i=1,..n$ . Donde  $\det[A]$  significa el determinante de la matriz  $A$ . Usando propiedades de determinantes, se puede obtener la siguiente equivalente fórmula

$$(COVRATIO)_i = \left( \frac{s_{(i)}^2}{s^2} \right)^p \frac{1}{(1-h_{ii})} \quad (3.14)$$

Si  $(COVRATIO)_i > 1+3p/n$  o si  $(COVRATIO)_i < 1-3p/n$  entonces la  $i$ -ésima observación tiene un valor influyente grande.

**Ejemplo 1:** Aplicar los diagnósticos de regresión al conjunto de datos **millaje**.

La siguiente es una lista completa de los diagnósticos para todas las observaciones obtenida usando SAS..

Obs	Dep Var	Predict Value	Std Err Predict	Std Err Residual	Student Residual	Cook's D
1	65.4000	53.4146	1.267	11.9854	3.426	3.499   *****   0.335
2	56.0000	49.7766	1.029	6.2234	3.505	1.776   ***   0.054
3	55.9000	49.7766	1.029	6.1234	3.505	1.747   ***   0.053
4	49.0000	45.3013	0.742	3.6987	3.577	1.034   **   0.009
5	46.5000	50.2870	1.092	-3.7870	3.486	-1.086   **   0.023

6	46.2000	45.3482	0.725	0.8518	3.580	0.238			0.000
<b>7</b>	<b>45.4000</b>	<b>49.7766</b>	<b>1.029</b>	<b>-4.3766</b>	<b>3.505</b>	<b>-1.249</b>		**	<b>0.027</b>
<b>8</b>	<b>59.2000</b>	<b>47.2349</b>	<b>1.213</b>	<b>11.9651</b>	<b>3.445</b>	<b>3.473</b>		*****	<b>0.299</b>
<b>9</b>	<b>53.3000</b>	<b>47.2349</b>	<b>1.213</b>	<b>6.0651</b>	<b>3.445</b>	<b>1.760</b>		***	<b>0.077</b>
10	43.4000	41.9529	0.639	1.4471	3.596	0.402			0.001
11	41.1000	44.4650	0.646	-3.3650	3.595	-0.936		*	0.006
<b>12</b>	<b>40.9000</b>	<b>39.5790</b>	<b>1.263</b>	<b>1.3210</b>	<b>3.428</b>	<b>0.385</b>			<b>0.004</b>
13	40.9000	38.8124	0.976	2.0876	3.520	0.593		*	0.005
14	40.4000	44.4650	0.646	-4.0650	3.595	-1.131		**	0.008
15	39.6000	45.6039	0.735	-6.0039	3.578	-1.678		***	0.024
16	39.3000	44.4650	0.646	-5.1650	3.595	-1.437		**	0.013
17	38.9000	42.5103	0.605	-3.6103	3.602	-1.002		**	0.006
<b>18</b>	<b>38.8000</b>	<b>39.5790</b>	<b>1.263</b>	<b>-0.7790</b>	<b>3.428</b>	<b>-0.227</b>			<b>0.001</b>
19	38.2000	42.5103	0.605	-4.3103	3.602	-1.197		**	0.008
20	42.2000	38.4951	0.631	3.7049	3.598	1.030		**	0.007
21	40.9000	38.0473	0.651	2.8527	3.594	0.794		*	0.004
22	40.7000	42.5157	0.729	-1.8157	3.579	-0.507		*	0.002
23	40.0000	37.8978	0.682	2.1022	3.589	0.586		*	0.002
24	39.3000	40.2540	0.504	-0.9540	3.618	-0.264			0.000
25	38.8000	38.0856	0.647	0.7144	3.595	0.199			0.000
<b>26</b>	<b>38.4000</b>	<b>38.8139</b>	<b>1.152</b>	<b>-0.4139</b>	<b>3.466</b>	<b>-0.119</b>			<b>0.000</b>
27	38.4000	37.7657	0.849	0.6343	3.553	0.179			0.000
28	38.4000	38.0473	0.651	0.3527	3.594	0.098			0.000
<b>29</b>	<b>29.5000</b>	<b>38.5108</b>	<b>0.623</b>	<b>-9.0108</b>	<b>3.599</b>	<b>-2.504</b>		*****	<b>0.038</b>
<b>30</b>	<b>46.9000</b>	<b>43.5274</b>	<b>1.173</b>	<b>3.3726</b>	<b>3.459</b>	<b>0.975</b>		*	<b>0.022</b>
31	36.3000	34.9973	0.659	1.3027	3.593	0.363			0.001
32	36.1000	39.0892	0.661	-2.9892	3.592	-0.832		*	0.005
33	36.1000	39.2925	0.549	-3.1925	3.611	-0.884		*	0.004
34	35.4000	36.0564	0.512	-0.6564	3.617	-0.181			0.000
35	35.3000	35.8061	0.649	-0.5061	3.595	-0.141			0.000
36	35.1000	39.4107	0.564	-4.3107	3.609	-1.194		**	0.007
37	35.1000	37.8083	0.448	-2.7083	3.625	-0.747		*	0.002
38	35.0000	37.9647	0.497	-2.9647	3.619	-0.819		*	0.003
39	33.2000	34.1686	0.598	-0.9686	3.603	-0.269			0.000
40	32.9000	34.1686	0.598	-1.2686	3.603	-0.352			0.001
41	32.3000	30.8137	0.828	1.4863	3.558	0.418			0.002
42	32.2000	34.8852	0.512	-2.6852	3.617	-0.742		*	0.002
43	32.2000	34.9947	0.465	-2.7947	3.623	-0.771		*	0.002
44	32.2000	34.0747	0.524	-1.8747	3.615	-0.519		*	0.001
45	32.2000	35.2763	0.576	-3.0763	3.607	-0.853		*	0.004
46	31.5000	35.5677	0.478	-4.0677	3.621	-1.123		**	0.004
47	31.5000	34.4756	0.454	-2.9756	3.624	-0.821		*	0.002
48	31.4000	34.2879	0.491	-2.8879	3.620	-0.798		*	0.002
49	31.4000	34.9234	0.444	-3.5234	3.626	-0.972		*	0.003
50	31.2000	30.9076	0.857	0.2924	3.551	0.082			0.000
51	33.7000	29.7337	0.610	3.9663	3.601	1.101		**	0.007
52	32.6000	29.7337	0.610	2.8663	3.601	0.796		*	0.004
53	31.3000	29.7337	0.610	1.5663	3.601	0.435			0.001
54	31.3000	29.3738	0.633	1.9262	3.598	0.535		*	0.002
<b>55</b>	<b>30.4000</b>	<b>23.9641</b>	<b>1.094</b>	<b>6.4359</b>	<b>3.485</b>	<b>1.847</b>		***	<b>0.067</b>
56	28.9000	26.4784	0.715	2.4216	3.582	0.676		*	0.004

57	28.0000	27.4881	0.670	0.5119	3.591	0.143				0.000
58	28.0000	31.4862	0.929	-3.4862	3.533	-0.987		*		0.013
59	28.0000	29.7337	0.610	-1.7337	3.601	-0.481				0.001
60	28.0000	30.4341	0.853	-2.4341	3.552	-0.685		*		0.005
61	28.0000	28.8107	0.737	-0.8107	3.578	-0.227				0.000
62	27.7000	27.1006	0.599	0.5994	3.603	0.166				0.000
63	25.6000	24.7507	0.902	0.8493	3.540	0.240				0.001
64	25.3000	23.2965	0.787	2.0035	3.567	0.562		*		0.003
65	23.9000	23.4217	0.741	0.4783	3.577	0.134				0.000
66	23.6000	23.4906	0.679	0.1094	3.589	0.030				0.000
<b>67</b>	<b>23.6000</b>	<b>24.0105</b>	<b>1.520</b>	<b>-0.4105</b>	<b>3.321</b>	<b>-0.124</b>				<b>0.001</b>
<b>68</b>	<b>23.6000</b>	<b>23.0093</b>	<b>0.628</b>	<b>0.5907</b>	<b>3.598</b>	<b>0.164</b>				<b>0.000</b>
69	23.6000	22.8059	0.726	0.7941	3.580	0.222				0.000
70	23.6000	22.8684	0.684	0.7316	3.588	0.204				0.000
71	23.5000	20.7522	1.295	2.7478	3.415	0.805		*		0.019
<b>72</b>	<b>23.4000</b>	<b>19.9118</b>	<b>1.692</b>	<b>3.4882</b>	<b>3.237</b>	<b>1.077</b>		<b>**</b>		<b>0.063</b>
73	23.4000	22.7989	0.824	0.6011	3.559	0.169				0.000
74	23.1000	22.8458	0.781	0.2542	3.568	0.071				0.000
<b>75</b>	<b>22.9000</b>	<b>18.7231</b>	<b>1.281</b>	<b>4.1769</b>	<b>3.421</b>	<b>1.221</b>		<b>**</b>		<b>0.042</b>
76	22.9000	19.2081	1.113	3.6919	3.479	1.061		<b>**</b>		0.023
77	19.5000	18.6925	0.888	0.8075	3.543	0.228				0.001
<b>78</b>	<b>18.1000</b>	<b>20.6117</b>	<b>2.033</b>	<b>-2.5117</b>	<b>3.035</b>	<b>-0.828</b>		*		<b>0.061</b>
79	17.2000	19.0194	1.082	-1.8194	3.489	-0.521		*		0.005
<b>80</b>	<b>17.0000</b>	<b>20.7779</b>	<b>1.593</b>	<b>-3.7779</b>	<b>3.287</b>	<b>-1.149</b>		<b>**</b>		<b>0.062</b>
<b>81</b>	<b>16.7000</b>	<b>19.3010</b>	<b>1.871</b>	<b>-2.6010</b>	<b>3.137</b>	<b>-0.829</b>		*		<b>0.049</b>
<b>82</b>	<b>13.2000</b>	<b>12.7102</b>	<b>1.636</b>	<b>0.4898</b>	<b>3.266</b>	<b>0.150</b>				<b>0.001</b>

	Hat Diag		Cov	INTERCEP		VOL	HP	SP	WT
Obs	Rstudent	H	Ratio	Dffits	Dfbetas	Dfbetas	Dfbetas	Dfbetas	Dfbetas
<b>1</b>	<b>3.7900</b>	<b>0.1204</b>	<b>0.5107</b>	<b>1.4021</b>	<b>1.1286</b>	<b>0.3421</b>	<b>1.1002</b>	<b>-1.0801</b>	<b>-1.2007</b>
<b>2</b>	<b>1.8014</b>	<b>0.0794</b>	<b>0.9408</b>	<b>0.5289</b>	<b>0.4117</b>	<b>0.1168</b>	<b>0.3936</b>	<b>-0.3937</b>	<b>-0.4227</b>
<b>3</b>	<b>1.7712</b>	<b>0.0794</b>	<b>0.9472</b>	<b>0.5200</b>	<b>0.4048</b>	<b>0.1148</b>	<b>0.3871</b>	<b>-0.3871</b>	<b>-0.4156</b>
4	1.0346	0.0413	1.0383	0.2146	0.0142	0.0628	0.0176	-0.0026	-0.0793
<b>5</b>	<b>-1.0877</b>	<b>0.0894</b>	<b>1.0852</b>	<b>-0.3408</b>	<b>-0.2762</b>	<b>-0.0706</b>	<b>-0.2633</b>	<b>0.2655</b>	<b>0.2751</b>
6	0.2365	0.0394	1.1072	0.0479	0.0028	0.0098	0.0030	0.0001	-0.0162
<b>7</b>	<b>-1.2534</b>	<b>0.0794</b>	<b>1.0468</b>	<b>-0.3680</b>	<b>-0.2865</b>	<b>-0.0813</b>	<b>-0.2739</b>	<b>0.2739</b>	<b>0.2941</b>
<b>8</b>	<b>3.7569</b>	<b>0.1103</b>	<b>0.5120</b>	<b>1.3229</b>	<b>0.6042</b>	<b>-0.9409</b>	<b>0.4143</b>	<b>-0.5252</b>	<b>-0.2150</b>
<b>9</b>	<b>1.7852</b>	<b>0.1103</b>	<b>0.9771</b>	<b>0.6286</b>	<b>0.2871</b>	<b>-0.4471</b>	<b>0.1969</b>	<b>-0.2496</b>	<b>-0.1021</b>
10	0.4002	0.0306	1.0897	0.0711	-0.0108	0.0168	-0.0098	0.0144	-0.0103
11	-0.9352	0.0312	1.0407	-0.1679	-0.0799	-0.0135	-0.0724	0.0708	0.0925
<b>12</b>	<b>0.3833</b>	<b>0.1195</b>	<b>1.2008</b>	<b>0.1412</b>	<b>-0.0790</b>	<b>-0.0946</b>	<b>-0.0883</b>	<b>0.0881</b>	<b>0.0772</b>
13	0.5906	0.0714	1.1236	0.1637	-0.0988	0.0457	-0.0888	0.1039	0.0358
14	-1.1327	0.0312	1.0135	-0.2034	-0.0968	-0.0163	-0.0877	0.0858	0.1121
15	-1.6984	0.0405	0.9235	-0.3488	-0.2276	-0.0140	-0.2046	0.2117	0.2174
16	-1.4468	0.0312	0.9620	-0.2598	-0.1236	-0.0208	-0.1121	0.1096	0.1432
17	-1.0022	0.0274	1.0279	-0.1683	0.0055	-0.0228	0.0073	-0.0152	0.0341
<b>18</b>	<b>-0.2259</b>	<b>0.1195</b>	<b>1.2084</b>	<b>-0.0832</b>	<b>0.0465</b>	<b>0.0558</b>	<b>0.0520</b>	<b>-0.0519</b>	<b>-0.0455</b>

19	-1.1999	0.0274	0.9993	-0.2014	0.0066	-0.0272	0.0087	-0.0182	0.0409
20	1.0302	0.0299	1.0267	0.1808	-0.0646	0.0702	-0.0573	0.0697	0.0059
21	0.7918	0.0318	1.0582	0.1434	-0.0703	0.0342	-0.0659	0.0752	0.0262
22	-0.5048	0.0399	1.0934	-0.1029	-0.0643	-0.0462	-0.0627	0.0631	0.0693
23	0.5833	0.0349	1.0817	0.1108	-0.0518	0.0359	-0.0464	0.0551	0.0135
24	-0.2621	0.0191	1.0834	-0.0365	-0.0010	-0.0045	0.0005	-0.0006	0.0053
25	0.1975	0.0314	1.0993	0.0356	-0.0191	-0.0029	-0.0192	0.0209	0.0110
<b>26</b>	<b>-0.1186</b>	<b>0.0995</b>	<b>1.1844</b>	<b>-0.0394</b>	<b>0.0149</b>	<b>0.0330</b>	<b>0.0191</b>	<b>-0.0174</b>	<b>-0.0204</b>
27	0.1774	0.0540	1.1261	0.0424	-0.0138	0.0283	-0.0102	0.0140	-0.0028
28	0.0975	0.0318	1.1019	0.0177	-0.0087	0.0042	-0.0081	0.0093	0.0032
<b>29</b>	<b>-2.5951</b>	<b>0.0291</b>	<b>0.7192</b>	<b>-0.4492</b>	<b>0.1645</b>	<b>-0.1604</b>	<b>0.1481</b>	<b>-0.1778</b>	<b>-0.0220</b>
<b>30</b>	<b>0.9746</b>	<b>0.1032</b>	<b>1.1187</b>	<b>0.3306</b>	<b>0.2568</b>	<b>0.0056</b>	<b>0.2237</b>	<b>-0.2568</b>	<b>-0.1632</b>
31	0.3605	0.0325	1.0941	0.0661	-0.0372	0.0246	-0.0336	0.0380	0.0167
32	-0.8304	0.0327	1.0550	-0.1528	-0.0745	-0.0869	-0.0745	0.0754	0.0828
33	-0.8828	0.0226	1.0379	-0.1343	-0.0712	-0.0194	-0.0619	0.0689	0.0567
34	-0.1803	0.0197	1.0866	-0.0255	0.0115	-0.0017	0.0114	-0.0124	-0.0065
35	-0.1399	0.0316	1.1009	-0.0253	0.0075	-0.0156	0.0056	-0.0075	0.0010
36	-1.1978	0.0238	0.9960	-0.1871	-0.0980	-0.0197	-0.0829	0.0950	0.0717
37	-0.7449	0.0151	1.0452	-0.0921	-0.0066	-0.0029	-0.0007	0.0036	0.0052
38	-0.8175	0.0185	1.0411	-0.1122	-0.0018	0.0484	0.0112	-0.0038	-0.0162
39	-0.2672	0.0268	1.0918	-0.0443	0.0172	0.0274	0.0221	-0.0188	-0.0260
40	-0.3500	0.0268	1.0881	-0.0581	0.0225	0.0359	0.0289	-0.0246	-0.0341
41	0.4155	0.0514	1.1127	0.0967	-0.0773	-0.0125	-0.0742	0.0797	0.0576
42	-0.7403	0.0197	1.0505	-0.1048	-0.0011	-0.0458	0.0004	0.0027	0.0028
43	-0.7693	0.0162	1.0438	-0.0987	0.0024	-0.0136	0.0083	-0.0023	-0.0116
44	-0.5161	0.0206	1.0711	-0.0749	0.0310	0.0333	0.0380	-0.0332	-0.0418
45	-0.8513	0.0249	1.0441	-0.1361	0.0128	0.0818	0.0318	-0.0171	-0.0542
46	-1.1252	0.0171	1.0000	-0.1485	-0.0196	0.0088	-0.0054	0.0187	-0.0107
47	-0.8192	0.0155	1.0377	-0.1027	0.0263	0.0168	0.0341	-0.0280	-0.0375
48	-0.7959	0.0180	1.0430	-0.1079	0.0192	-0.0439	0.0192	-0.0182	-0.0108
49	-0.9715	0.0148	1.0187	-0.1191	0.0061	-0.0042	0.0138	-0.0069	-0.0182
50	0.0818	0.0550	1.1292	0.0197	-0.0156	-0.0056	-0.0154	0.0162	0.0127
51	1.1029	0.0279	1.0144	0.1869	-0.0247	-0.0723	-0.0442	0.0224	0.0905
52	0.7940	0.0279	1.0538	0.1345	-0.0177	-0.0521	-0.0318	0.0162	0.0652
53	0.4326	0.0279	1.0848	0.0733	-0.0097	-0.0284	-0.0173	0.0088	0.0355
54	0.5329	0.0300	1.0802	0.0937	-0.0038	0.0428	-0.0033	-0.0008	0.0107
<b>55</b>	<b>1.8767</b>	<b>0.0897</b>	<b>0.9350</b>	<b>0.5891</b>	<b>-0.3824</b>	<b>0.2079</b>	<b>-0.3056</b>	<b>0.3776</b>	<b>0.1730</b>
56	0.6736	0.0383	1.0775	0.1345	-0.0672	0.0567	-0.0507	0.0648	0.0279
57	0.1416	0.0337	1.1033	0.0264	-0.0114	0.0139	-0.0097	0.0103	0.0079
58	-0.9867	0.0647	1.0710	-0.2595	-0.0242	0.1519	0.0149	0.0245	-0.0963
59	-0.4790	0.0279	1.0818	-0.0812	0.0107	0.0314	0.0192	-0.0097	-0.0393
60	-0.6829	0.0545	1.0951	-0.1640	0.0147	0.0954	0.0382	-0.0143	-0.0853
61	-0.2252	0.0407	1.1091	-0.0464	0.0170	0.0034	0.0207	-0.0156	-0.0283
62	0.1653	0.0269	1.0950	0.0275	-0.0157	0.0071	-0.0136	0.0151	0.0114
63	0.2385	0.0610	1.1327	0.0608	-0.0055	-0.0121	-0.0108	0.0030	0.0284
64	0.5592	0.0464	1.0967	0.1233	-0.0241	0.0444	-0.0217	0.0164	0.0376
65	0.1329	0.0411	1.1120	0.0275	-0.0064	0.0038	-0.0067	0.0049	0.0118
66	0.0303	0.0345	1.1056	0.0057	-0.0009	0.0011	-0.0008	0.0006	0.0018
<b>67</b>	<b>-0.1228</b>	<b>0.1732</b>	<b>1.2899</b>	<b>-0.0562</b>	<b>0.0111</b>	<b>0.0513</b>	<b>0.0170</b>	<b>-0.0125</b>	<b>-0.0305</b>
68	0.1631	0.0296	1.0982	0.0285	-0.0067	0.0034	-0.0054	0.0054	0.0092
69	0.2205	0.0395	1.1079	0.0447	-0.0072	0.0229	-0.0031	0.0046	0.0046

70	0.2026	0.0351	1.1033	0.0386	-0.0071	0.0158	-0.0040	0.0050	0.0065
71	0.8027	0.1258	1.1707	0.3045	-0.0236	0.1381	0.0289	0.0196	-0.0810
<b>72</b>	<b>1.0786</b>	<b>0.2145</b>	<b>1.2597</b>	<b>0.5637</b>	<b>-0.1260</b>	<b>-0.2019</b>	<b>-0.0820</b>	<b>0.1434</b>	<b>0.0472</b>
73	0.1678	0.0509	1.1227	0.0389	-0.0044	0.0252	-0.0005	0.0020	0.0006
74	0.0708	0.0457	1.1184	0.0155	-0.0020	0.0092	-0.0005	0.0011	0.0008
<b>75</b>	<b>1.2250</b>	<b>0.1230</b>	<b>1.1039</b>	<b>0.4587</b>	<b>-0.0118</b>	<b>0.2285</b>	<b>-0.0020</b>	<b>-0.0218</b>	<b>0.0669</b>
76	1.0621	0.0928	1.0932	0.3398	-0.0327	-0.0212	-0.0519	0.0138	0.1490
77	0.2265	0.0591	1.1307	0.0567	0.0005	0.0121	0.0020	-0.0042	0.0104
<b>78</b>	<b>-0.8259</b>	<b>0.3097</b>	<b>1.4789</b>	<b>-0.5532</b>	<b>-0.2271</b>	<b>0.1223</b>	<b>-0.2680</b>	<b>0.2168</b>	<b>0.2449</b>
79	-0.5190	0.0877	1.1497	-0.1610	-0.0557	-0.0540	-0.0760	0.0613	0.0679
<b>80</b>	<b>-1.1518</b>	<b>0.1903</b>	<b>1.2091</b>	<b>-0.5584</b>	<b>-0.1786</b>	<b>0.3170</b>	<b>-0.1807</b>	<b>0.1674</b>	<b>0.0829</b>
<b>81</b>	<b>-0.8274</b>	<b>0.2624</b>	<b>1.3839</b>	<b>-0.4935</b>	<b>-0.2092</b>	<b>-0.2396</b>	<b>-0.2846</b>	<b>0.2187</b>	<b>0.3211</b>
<b>82</b>	<b>0.1490</b>	<b>0.2007</b>	<b>1.3336</b>	<b>0.0747</b>	<b>0.0359</b>	<b>-0.0182</b>	<b>0.0340</b>	<b>-0.0386</b>	<b>-0.0086</b>

*Nota: Las observaciones en negritas pueden ser influenciales según al menos uno de los diagnósticos que se describen a continuación:*

Una lista de las observaciones con al menos uno de los diagnósticos sobrepasando los valores críticos es obtenida usando R (ver el laboratorio 10 del texto).

Potentially influential observations of  
lm(formula = mpg ~ ., data = millaje) :

	dfb.1_	dfb.sp	dfb.wt	dfb.vol	dfb.hp	dffit	cov.r	cook.d	hat
1	1.13_*	-1.08_*	-1.20_*	0.34_*	1.10_*	1.40_*	0.51_*	0.34	0.12
2	0.41_*	-0.39_*	-0.42_*	0.12	0.39_*	0.53	0.94	0.05	0.08
3	0.40_*	-0.39_*	-0.42_*	0.11	0.39_*	0.52	0.95	0.05	0.08
5	-0.28_*	0.27_*	0.28_*	-0.07	-0.26_*	-0.34	1.09	0.02	0.09
7	-0.29_*	0.27_*	0.29_*	-0.08	-0.27_*	-0.37	1.05	0.03	0.08
8	0.60_*	-0.53_*	-0.21	-0.94_*	0.41_*	1.32_*	0.51_*	0.30	0.11
9	0.29_*	-0.25_*	-0.10	-0.45_*	0.20	0.63	0.98	0.08	0.11
12	-0.08	0.09	0.08	-0.09	-0.09	0.14	1.20_*	0.00	0.12
15	-0.23_*	0.21	0.22	-0.01	-0.20	-0.35	0.92	0.02	0.04
18	0.05	-0.05	-0.05	0.06	0.05	-0.08	1.21_*	0.00	0.12
26	0.01	-0.02	-0.02	0.03	0.02	-0.04	1.18_*	0.00	0.10
29	0.16	-0.18	-0.02	-0.16	0.15	-0.45	0.72_*	0.04	0.03
30	0.26_*	-0.26_*	-0.16	0.01	0.22_*	0.33	1.12	0.02	0.10
55	-0.38_*	0.38_*	0.17	0.21	-0.31_*	0.59	0.94	0.07	0.09
67	0.01	-0.01	-0.03	0.05	0.02	-0.06	1.29_*	0.00	0.17
72	-0.13	0.14	0.05	-0.20	-0.08	0.56	1.26_*	0.06	0.21_*
75	-0.01	-0.02	0.07	0.23_*	0.00	0.46	1.10	0.04	0.12
78	-0.23_*	0.22	0.24_*	0.12	-0.27_*	-0.55	1.48_*	0.06	0.31_*
80	-0.18	0.17	0.08	0.32_*	-0.18	-0.56	1.21_*	0.06	0.19_*
81	-0.21	0.22	0.32_*	-0.24_*	-0.28_*	-0.49	1.38_*	0.05	0.26_*
82	0.04	-0.04	-0.01	-0.02	0.03	0.07	1.33_*	0.00	0.20_*

De acuerdo a los residuales estudentizados internamente o externamente, las observaciones 1, 8 y 29 son “outliers”.

De acuerdo a los valores leverages  $h_{ii}$ , las observaciones 72, 78, 80, 81 y 82 son puntos leverages, pues tiene  $h_{ii} > 3p/n = 0.1829$

De acuerdo a la Distancia Cook no hay ninguna observación que tenga gran influencia pues todos los  $CD_i^2$  son menores que 1, más aún son menores que  $F(0.50,5,77)=0.878$ . Sin embargo, las observaciones 1 y 8 tienen un  $CD_i^2$  mucho mayor que las otras y deberían ser consideradas cuidadosamente.

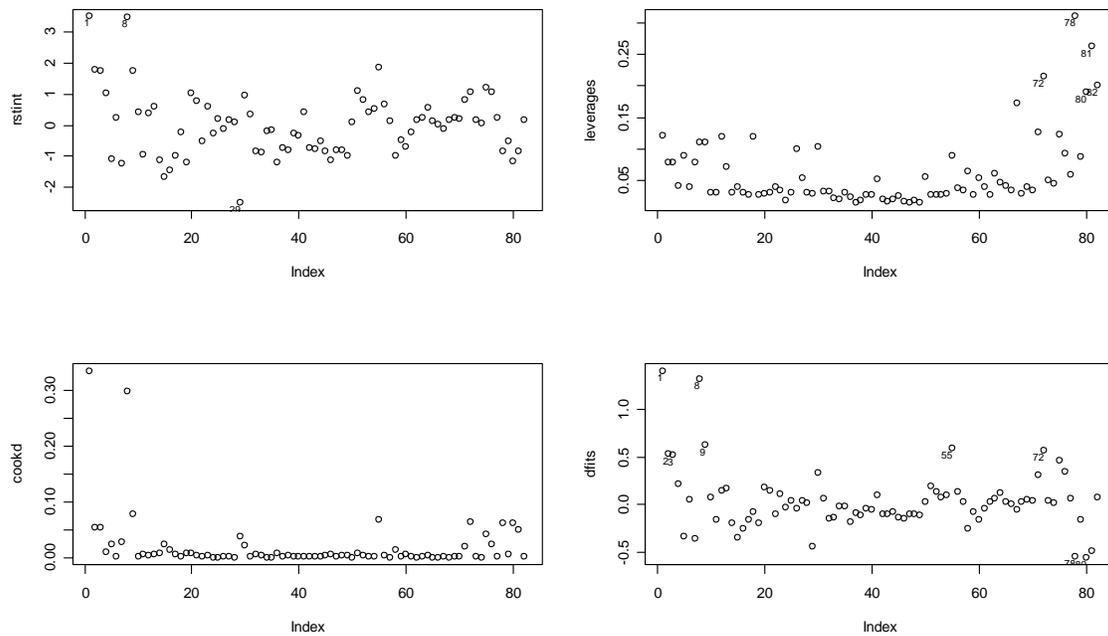


Figura 3.3. Plot de varios diagnosticos de regresión para cada observación del ejemplo 1.

De acuerdo al DFFITS serían influyentes las observaciones 1, 2, 3, 8, 9, 55, 72, 78 y 80, pues su

$$|DFFITS_i| > 2 \sqrt{\frac{5}{82}} = .49386$$

De acuerdo al COVRATIO serían influyentes las observaciones 1, 8, 12, 18, 26, 29, 72, 78, 80, 81 y 82, pues su  $COVRATIO > 1.1829$  ó  $< 0.8171$ . Los  $COVRATIO$  de las observaciones 12, 18, 26 y 29 están bastante cerca de los puntos de corte.

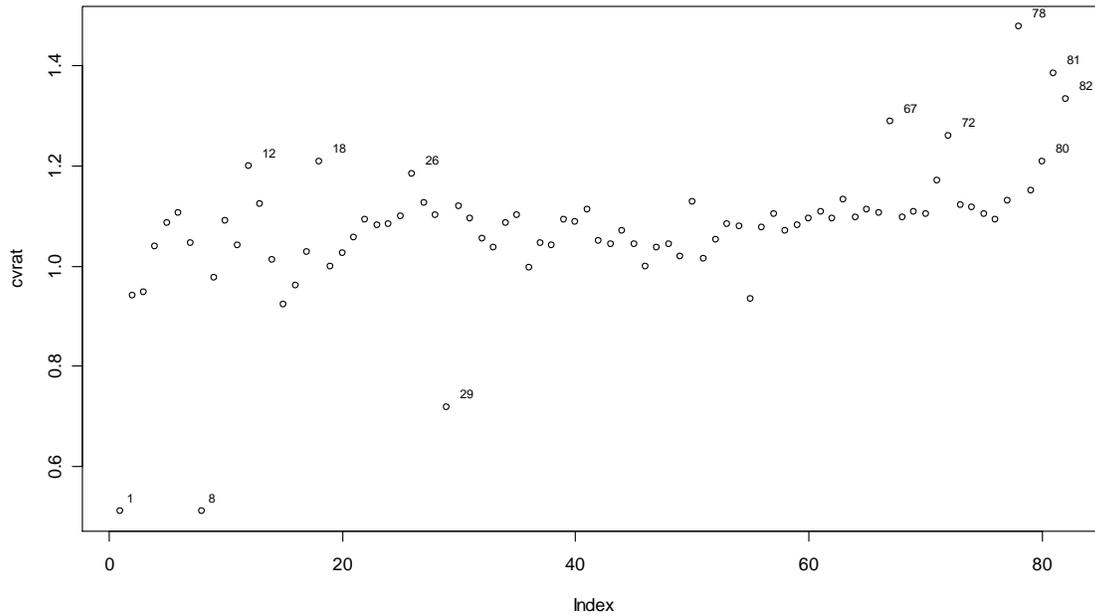


Figura 3.4. Plot de COVRATIOs para cada observación del ejemplo 1.

De acuerdo a los DFBETAS una observación es influyente si su valor absoluto es mayor  $2/\sqrt{82} = 0.22086$ . Las observaciones 1, 8, 9, 80 y 81 parecen afectar el comportamiento del coeficiente  $\beta_1$ , el valor DFBETAS de la observación 75 está muy cerca del punto de corte y no ha sido considerado. Las observaciones 1, 2, 3, 5, 7, 8, 30, 55, 78 y 81 afectan el comportamiento de  $\beta_2$ , en tanto que 1, 2, 3, 5, 7, 8, 9, 30, y 55 parecen tener influencia en  $\beta_3$  y las observaciones 1, 2, 3, 5, 7, 78 y 81 afectan el comportamiento de  $\beta_4$ .

Plot de los DFBETAS

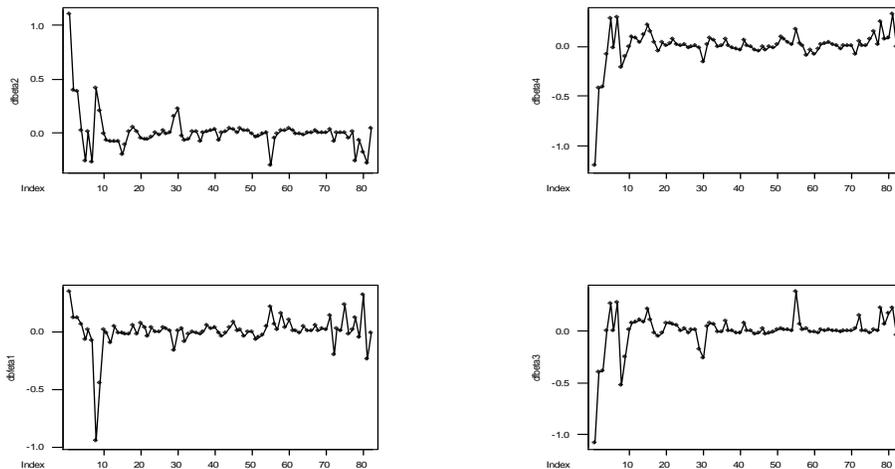


Figura 3.5. Plot de los DFBETAS para el conjunto datos millaje.

En conclusión, las observaciones 1, 2, 3, 5, 7, 8, 9, 30, 55, 72, 78., 80, 81 y 82 parecen ser las más influyentes.

### 3.4 Plot de Residuales para detectar el efecto de variables y casos influyentes

Existen ciertos plots de residuales que se usan para estudiar el efecto de añadir una nueva variable predictora en un modelo. Estos plots también permiten detectar la presencia de casos influyentes. Supongamos que queremos ver la importancia de la variable predictora  $x_j$ . Consideremos el modelo

$$Y = X_j B_j + \beta_j x_j + e$$

Donde  $X_j$  es la matriz  $X$  sin incluir la columna  $j$ . Se puede mostrar que

$$\hat{\beta}_j = \frac{\mathbf{x}_j' (\mathbf{I} - \mathbf{H}_{-j}) \mathbf{Y}}{\mathbf{x}_j' (\mathbf{I} - \mathbf{H}_{-j}) \mathbf{x}_j} \quad (3.15)$$

Definamos los siguientes residuales

- a)  $\hat{e}_{Y/X_{-j}} = (\mathbf{I} - \mathbf{H}_{-j}) \mathbf{Y}$
- b)  $\hat{e}_{Y/x, X_{-j}} = (\mathbf{I} - \mathbf{H}) \mathbf{Y}$
- c)  $\hat{e}_{x_j/X_{-j}} = (\mathbf{I} - \mathbf{H}_{-j}) \mathbf{X}_j$

en el caso a) se han considerado en el modelo todas las predictoras excepto  $x_j$ , en el caso b) están consideradas todas las variables predictoras y en el caso c) son los residuales de la regresión de  $x_j$  versus las otras variables no consideradas en el modelo.

Hay cuatro tipos de plots de residuales que permiten ver el impacto de cada variable predictora  $x_j$  en el modelo. Estos son:

- a) Plot de Residuales versus las variables predictoras
- b) Plot de regresión parcial (o plot de variable añadida)
- c) Plot de residuales parciales
- d) Plot de residuales parciales aumentados.

#### a) Plot de residuales versus las variables predictoras.

Aquí se plotea

$$\hat{e}_{Y/x, X_{-j}} \text{ versus } x_j$$

Si el modelo es adecuado los puntos se deberían alinear a lo largo de una franja horizontal. Si se observa algún patrón no lineal entonces la variable predictora debería ser transformada.

Para el ejemplo de Millaje estos son los plots de residuales que resultan.

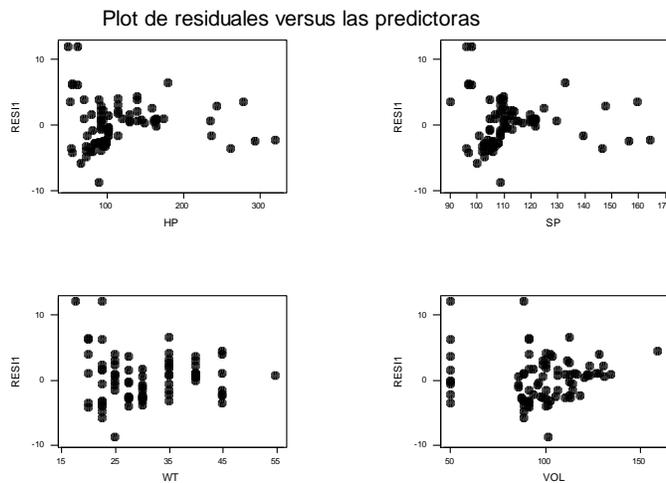


Figura 3.6. Plot de residuales versus las predictoras para el conjunto de datos millaje

Lo que más se destaca en estas gráficas es la presencia de varios valores “outliers” y puntos de leverage alto. También se observa que no todos los puntos se alinean uniformemente alrededor del eje horizontal 0, pero es difícil detectar la tendencia para usar una transformación no lineal.

### b) Plots de regresión parciales ( o plot de variable añadida)

Aquí se plotea los residuales  $\hat{e}_{Y/X_{-j}}$  versus  $\hat{e}_{x_j/X_{-j}}$

En el plot de regresión parcial se plotea los residuales de la regresión de  $y$  considerando todas las variables predictoras excepto  $x_j$  versus los residuales de la regresión de  $x_j$  contra todas las variables predictoras distintas a ella.

Si la variable  $x_j$  entra al modelo en forma lineal entonces su plot de regresión parcial debería mostrar una tendencia lineal que pasa por el origen. Si se observa una tendencia no lineal habría que considerar una transformación de  $x_j$ . También se puede localizar a los puntos que afecta el cálculo del coeficiente de regresión correspondiente.

Consideremos por ejemplo el plot de regresión parcial para la variable HP del conjunto de datos Millaje, donde se asume que el modelo contiene ya las otras 3 variables predictoras.

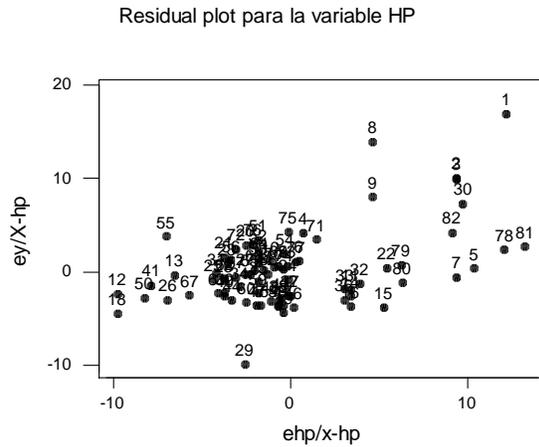


Figura 3.7. Plot de regresión parcial considerando la variable predictora HP

Se observan muchos valores influyentes y la tendencia lineal es bien pobre. Si usamos  $1/HP$  en lugar de HP la cosa no mejora mucho.

En realidad el efecto de este plot se observa mejor si consideramos primero la regresión de MPG con la variable VOL, que es la que tiene menos correlación y si consideramos luego añadir WT. El plot de regresión parcial que se obtiene es

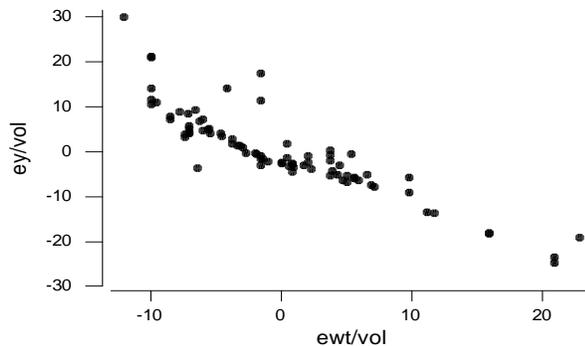


Figura 3.8. Plot de regresión parcial considerando la variable WT asumiendo que el modelo solo contiene a VOL.

Se puede observar que hay bastante linealidad en el plot, y que la línea estimada pasaría por el origen. Luego, se debería usar una regresión lineal múltiple con dos variables predictoras.

Consideremos la misma situación anterior pero en lugar de WT ahora queremos incluir HP. El plot de la Figura 3.7 ya no se ve tan lineal sino parece como una rama de una hipérbola equilátera.

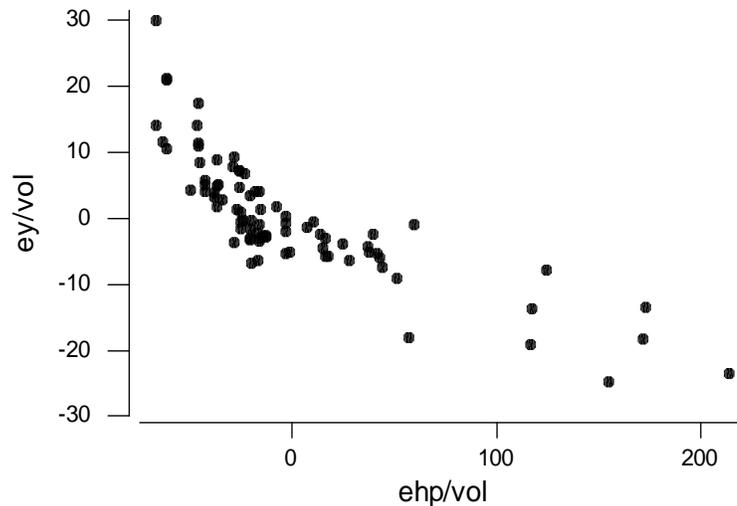


Figura 3.9. Plot de regresión parcial considerando la variable HP asumiendo que el modelo solo contiene a VOL.

Así que sería mejor usar  $1/HP$  en lugar de HP en el modelo

### c) Plot de residuales parciales o de residuales más componente

Aquí se plotea

$$\hat{e}_{Y/x, X_{-j}} + x_j \beta_j \text{ versus } x_j$$

Es más efectivo para detectar no linealidad que el plot de regresión parcial. No es muy adecuado para detectar casos influyentes.

### d) Plot de residuales parciales aumentados

Aquí se plotea

$$\hat{e}_{y/X_{-j}, x_j^2} + x_j \beta_j + x_j^2 \beta_{jj} \text{ versus } x_j$$

Este plot fue propuesto por Mallows (1986) y es el más adecuado para cotejar si la variable  $x_j$  debe entrar en forma cuadrática al modelo.

## 3.5 Plot de residuales para detectar Normalidad

La suposición de la normalidad de los errores es bien importante para el proceso de hacer inferencia en regresión lineal múltiple. Al igual que en regresión lineal simple esto puede ser cotejado haciendo un plot de normalidad para los errores estudentizados internamente.

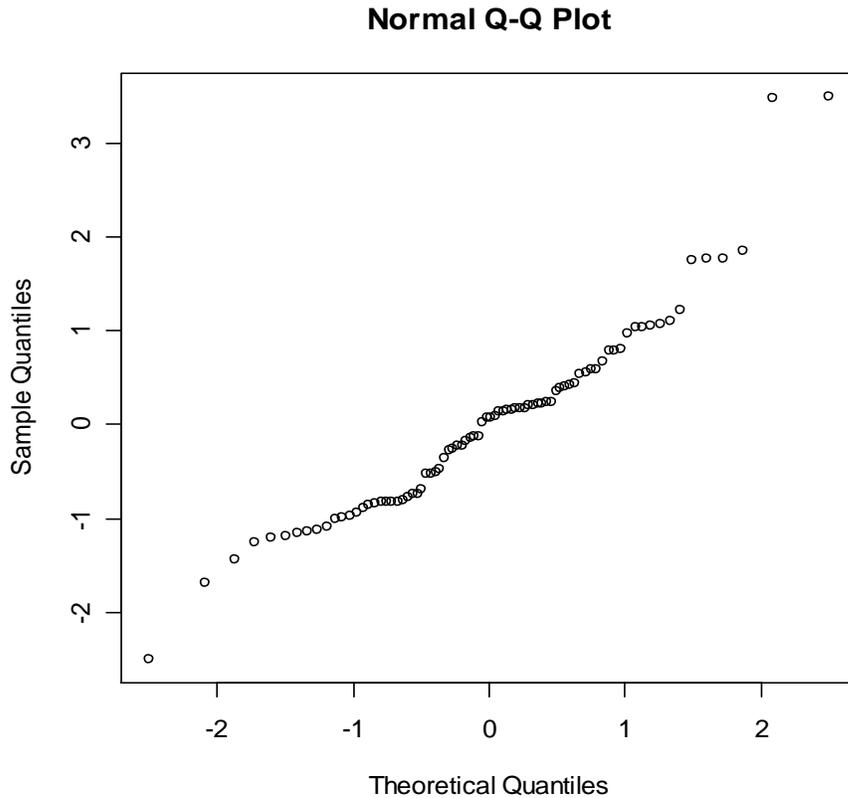


Figura 3.10. Plot de normalidad para los residuales del conjunto de datos millaje.

El plot de Normalidad consiste en un plot de los **scores normales** (estadísticos de orden normales) versus los residuales estandarizados ordenados. Los scores normales representan los valores esperados de observaciones ordenadas que provienen de una distribución normal estándar. El  $i$ -ésimo score normal es aproximado en forma bastante precisa por

$$z_{(i)} = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right)$$

donde  $\Phi$  representa la función de distribución acumulada de una normal estándar y  $n$  ( $n > 5$ ) es el número de observaciones en la muestra.

Para que haya normalidad los puntos deben estar alineados alrededor de una recta que pasa por el origen. En la figura 3.8 se muestra el plot de normalidad de los residuales correspondientes a la regresión lineal múltiple del conjunto de datos **millaje**, se observa que los puntos están bastante alineados pero se observan varios “outliers” en ambos extremos de la distribución.

Si la tendencia de los puntos es curvada entonces la distribución es asimétrica. El tipo de asimetría es determinada por el lado donde está la parte curvada. Un plot de normalidad que produce una curva en forma de S indica que la distribución tiene una cola pesada o liviana dependiendo de la forma de la S. Si la S es alargada entonces la cola es liviana.

También se podría aplicar una prueba no paramétrica como la de Kolmogorov-Smirnov o Shapiro-Wilks para detectar normalidad.

```
> l1=lm(mpg~.,data=millaje)
> resi=rstandard(l1)
> ks.test(resi,"pnorm")
```

#### One-sample Kolmogorov-Smirnov test

```
data: resi
D = 0.0881, p-value = 0.519
alternative hypothesis: two-sided
```

```
> boxplot(resi)
> shapiro.test(resi)
```

#### Shapiro-Wilk normality test

```
data: resi
W = 0.945, p-value = 0.001542
```

El “p-value” de la prueba de Kolmogorov-Smirnov es mayor que 0.05 por lo tanto se acepta la hipótesis de que hay normalidad de los residuales. Sin embargo, la prueba de Shapiro-Wilks indica que no hay normalidad puesto que el “p-value” de la prueba es pequeño.

### 3.6 Detectando varianza no constante

La suposición de que en el modelo de regresión lineal múltiple, los errores tienen varianza constante es importante para que los estimadores mínimos cuadrados sean óptimos. Por lo general varianza no constante viene acompañado del hecho que no hay normalidad.

Para detectar si la varianza es constante o no se hace un plot de residuales estudentizados versus los valores ajustados  $\hat{y}_i$ 's. Si los puntos aparecen alineados arbitrariamente alrededor de una franja horizontal centrada en la línea horizontal en cero entonces hay indicación de varianza constante (homocedasticidad). Si los puntos forman algún tipo de patrón como el que se muestra en la figura 3.9 entonces indica una violación de la suposición de homocedasticidad. Aquí la varianza varía en forma proporcional a la media de la variable de respuesta Y. Este plot es típico cuando los errores siguen una distribución Poisson o log-normal.

Algunas veces la varianza puede variar de acuerdo a los variables de una variable predictora. Para detectar esta situación hay que hacer un plot de residuales versus cada variable predictora.

Si hay indicación de que la varianza poblacional  $\sigma^2$  no es constante entonces hay dos remedios posibles:

- i) Usar mínimos cuadrados ponderados donde los pesos que se usan son hallados en base a los datos tomados.
- ii) Transformar la variable de respuesta Y usando transformación que estabiliza la varianza

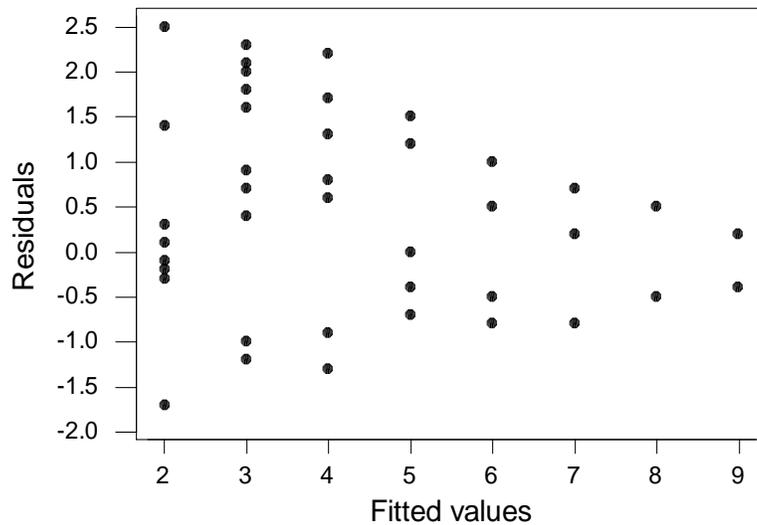


Figura 3.11. Este plot muestra que la varianza de los errores no es constante y que varia en forma proporcional a la media de la variable de respuesta

Las medidas remediales para varianza no constante serán discutidas en el capítulo 4 del texto.

### 3.7 Errores correlacionados en Regresión

Una de las suposiciones que se hace en regresión lineal es que  $Cov(e_i, e_j) = E(e_i e_j) = 0$  para  $i \neq j$ . Es decir que los errores no se correlacionan entre sí. Hay un caso en regresión cuando la variable predictora es tiempo, donde puede haber una dependencia del comportamiento con respecto al tiempo. Por ejemplo, las ventas de una compañía de ropas pueden seguir un patrón que depende de la época del año y pudiera ocurrir entonces que  $E(e_i, e_{i+k}) \neq 0$  para un cierto  $k$  en este caso se dice que los errores tienen una correlación serial y están autocorrelacionados. Si se gráfica los residuales versus la variable predictora tiempo y se observa mucho cambio de signo entonces la autocorrelación es negativa si el cambio de signo no es muy frecuente entonces la autocorrelación es positiva.

**Ejemplo 2.** Consideremos las siguientes series de tiempo

year	y1	y2	y3
1	10	15	5
2	20	10	10
3	35	18	15
4	50	12	18
5	12	24	20
6	24	15	32

7	40	32	37
8	50	18	39
9	14	40	42
10	25	20	35
11	40	55	30
12	60	25	27

cuyas graficas aparecen en la siguiente figura

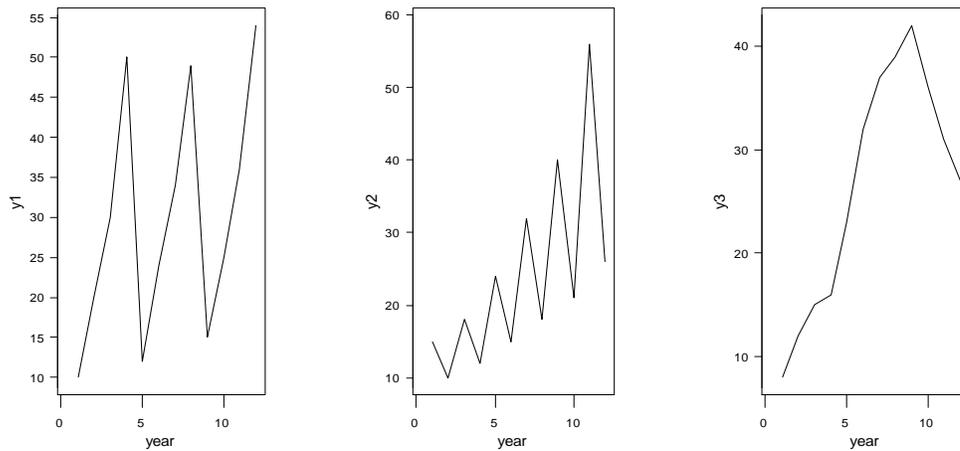


Figura 3.12. Gráfica de las 3 series de tiempo del ejemplo

En los dos primeros plots la autocorrelación es negativa y en la última es positiva  
Los plots de residuales correspondientes se muestran en la figura 3.11

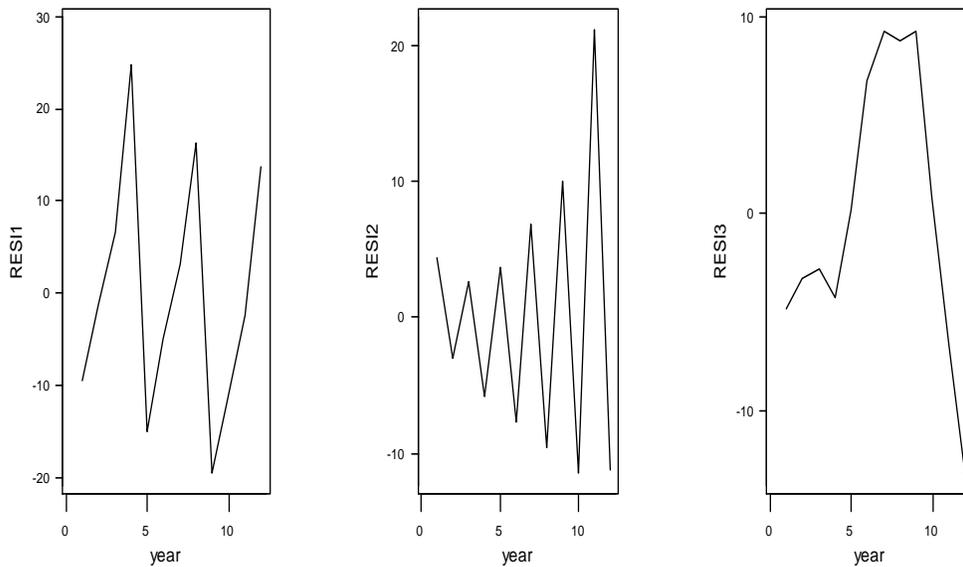


Figura 3.13. Plot de los residuales de las tres series de tiempo del ejemplo 2

Cuando los residuales cambian frecuentemente de signo hay autocorrelación negativa y si hay un conglomerado de residuales de un mismo signo antes de cambiar a otro entonces la autocorrelación es positiva. Lo anterior se puede observar más claramente si se plotea los residuales en el tiempo  $t$  versus los residuales en el tiempo  $t-1$ . La siguiente figura muestra estos plots para los datos correspondiente a las dos últimas graficas de la figura anterior.

El Modelo autorregresivo de primer orden para los errores se define

$$e_t = \rho e_{t-1} + u_t$$

donde se supone que las  $u_t$  son variables aleatorias distribuidas normalmente con media 0 y varianza constante.

La prueba de Durbin-Watson se usa para detectar si hay correlación positiva. Es decir para probar  $H_0: \rho=0$  vs  $H_a: \rho>0$ .

La prueba está dada por

$$D = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

Existen tablas de la prueba, que dan valores límites DL y DU para que se rechaze la hipótesis nula. La tabla considera número de casos:(n), número de predictoras y tres niveles de significación. Las decisiones se toman así: Se rechaza Ho si  $D < DL$ , se acepta Ho si  $D > DU$  y la prueba no lleva a ninguna conclusión si  $DL < D < DU$ .

La prueba de Durbin-Watson no detecta autocorrelaciones de segundo orden o mayor. R no tiene una función que calcule este estadístico, pero hemos construido una función **dw** para calcularla. Sin embargo, el estadístico es dado por muchos programas entre ellos MINITAB.

Para los datos cuyas gráficas de residuales aparecen en la figura 3.11 se obtienen los siguientes resultados para el estadístico de Durbin-Watson.

```
> l1=lm(y1~year, data=corrdata)
> dw(l1$residuals)
[1] 1.835639
> l2=lm(y2~year, data=corrdata)
> dw(l2$residuals)
[1] 3.537003
> l3=lm(y3~year, data=corrdata)
> dw(l3$residuals)
[1] 0.4571476
>
```

Buscando en la tabla de Durbin-Watson con  $n=12$  (aprox con  $n=15$ ),  $k=1$  y  $\alpha=0.5$  resulta ser que  $D_L=1.08$  y  $D_U=1.36$ , por lo tanto, no se rechaza la hipótesis nula en los dos primeros casos ya que  $DW=1.835$  y  $DW=3.53$  respectivamente son mayores que 1.36, y se concluye que no hay autocorrelación de primer orden entre los errores. En el último caso si se rechaza la hipótesis nula puesto que  $DW=.457 < 1.08$  y se concluye que hay autocorrelación positiva de primer orden.

Si se desea probar una hipótesis de dos lados  $H_0: \rho=0$ , versus  $H_a: \rho \neq 0$  entonces se rechaza  $H_0$ : si  $D < D_L$  ó  $4-D < D_L$ , al nivel de significación de  $2\alpha$ . Si  $D > D_U$  y  $4-D > D_U$  entonces no se rechaza. Para cualquier otro valor de D la prueba no llega a ninguna conclusión.

Una regla práctica es que cuando el estadístico de Durbin-Watson sale cerca de 2 entonces es probable que no hay autocorrelación.

Si hubiera autocorrelación positiva de primer orden entonces una forma de resolver el problema sería considerar el modelo

$$y_t = \beta_0 + \beta_1 t + \beta_2 y_{t-1} + e_t$$

donde  $y_{t-1}$  son los valores de la variable de respuesta en el tiempo anterior. Estos modelos son llamados modelos de series de tiempo y son discutidos en textos especializados.



## Ejercicios

1. Probar las siguientes identidades

$$i) \hat{\beta}_{(i)} = \hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i}{1-h_{ii}}$$

$$ii) s_{(i)}^2 = \frac{n-p-1}{n-p-2}s^2 - \frac{\hat{e}_i^2}{(n-p-2)(1-h_{ii})}$$

2. Probar las siguientes relaciones

$$i) t_i = \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

$$ii) t_i = r_i^* \left( \frac{n-p-2}{n-p-1-r_i^{*2}} \right)^{1/2}$$

3. Probar la siguiente formula equivalente para calcular COVRATIOS

$$(COVRATIO)_i = \left( \frac{s_{(i)}^2}{s^2} \right)^p \frac{1}{(1-h_{ii})}$$

4. Usar el conjunto de datos **Fuel** con variable de respuesta: Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto.

- Cotejar las suposiciones del modelo de regresión múltiple mediante un plot de residuales.
- Determinar outliers y puntos con leverage alto usando los diagnósticos de regresión.
- Usar plot de residuales para evaluar el efecto de añadir la segunda variable predictora con la correlación más alta con Fuel al modelo que ya tiene incluido la variable más altamente correlacionada con Fuel.

5. Usar el conjunto de datos **Headcirc** con variable de respuesta: headcirc (circunferencia de la cabeza del bebe) para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto

- Cotejar las suposiciones del modelo de regresión múltiple mediante un plot de residuales.
- Determinar outliers y puntos con leverage alto usando los diagnósticos de regresión.
- Usar plot de residuales para evaluar el efecto de añadir la segunda variable predictora con la correlación más alta con Headcirc al modelo que ya tiene incluido la variable más altamente correlacionada con Headcir.

6. Usar el conjunto de datos **Grasa** con variable de respuesta: grasa (porcentaje de grasa en el cuerpo) para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto.

- Cotejar las suposiciones del modelo de regresión múltiple mediante un plot de residuales.
- Determinar outliers y puntos con leverage alto usando los diagnósticos de regresión.
- Usar plot de residuales para evaluar el efecto de añadir al modelo con la variable predictor más altamente correlacionada con grasa, la segunda variable más altamente correlacionada con grasa.

7. Supongamos que ajustamos un modelo de regresión múltiple con intercepto y se define la

distancia (cuadrada) Mahalanobis de  $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})$  a  $\bar{\mathbf{x}} = \sum_{i=1}^n \frac{\mathbf{x}_i}{n}$  por

$D_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{C}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ , donde  $\mathbf{C}^{-1}$  es la inversa de la matriz de covarianza de las  $x$ 's.

Establecer una relación entre los valores leverages  $h_{ii}$  y  $D_i^2$ .

# CAPÍTULO 4

## TRANSFORMACIONES EN REGRESIÓN

### 4.1 Transformaciones para linealizar modelos

Consideremos por ahora solo modelos con una variable predictora. La idea es tratar de aumentar la medida de ajuste  $R^2$  del modelo, sin incluir variables predictoras adicionales. Lo primero que hay que hacer es un plot para observar el tipo de tendencia, pueden resultar plots como los que aparecen en las figuras 4.1 y 4.2.

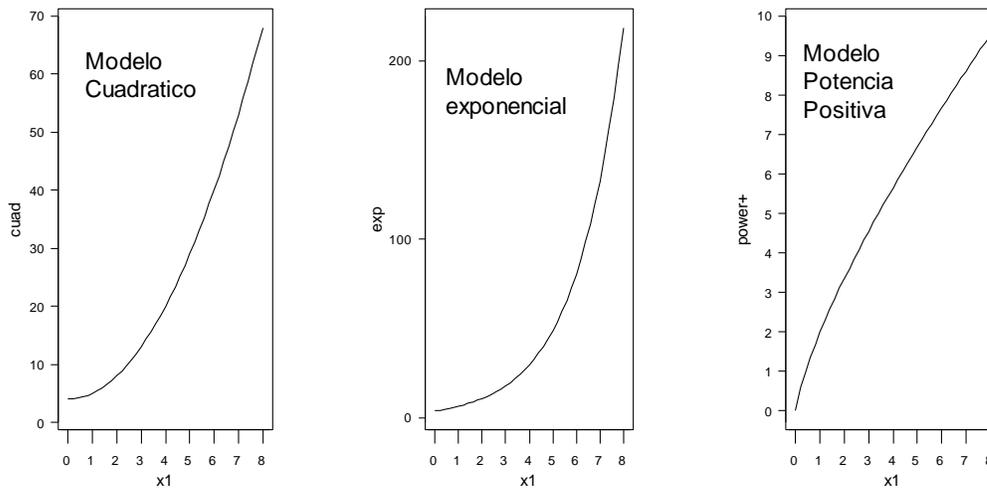


Figura 4.1. Gráficas de tres modelos no lineales.

En la primera gráfica de la figura 4.1 se ha ajustado un modelo cuadrático, que es de la forma general  $y=a+bx+cx^2$  y es el caso más sencillo de regresión polinómica. Esto puede ser modelado como una regresión múltiple con dos variables predictoras.

La segunda gráfica corresponde a un modelo exponencial de la forma  $y=\alpha e^{\beta x}$  con  $\alpha$  y  $\beta$  positivos. Este modelo es muy adecuado para modelar crecimientos poblacionales.

La tercera gráfica corresponde a un modelo potencial o doblemente logarítmico de la forma  $y=\alpha x^{\beta}$ , con  $\beta$  positivo.

La primera gráfica de la figura 4.2 corresponde a un modelo hiperbólico o inverso de la forma  $y=\alpha+\beta/x$ , con  $x > 0$ .

La segunda gráfica corresponde a un modelo logarítmico de la forma  $y=\alpha+\beta \log(x)$  con  $x > 0$ .

La tercera gráfica corresponde a un modelo potencia pero con  $\beta > 0$ .

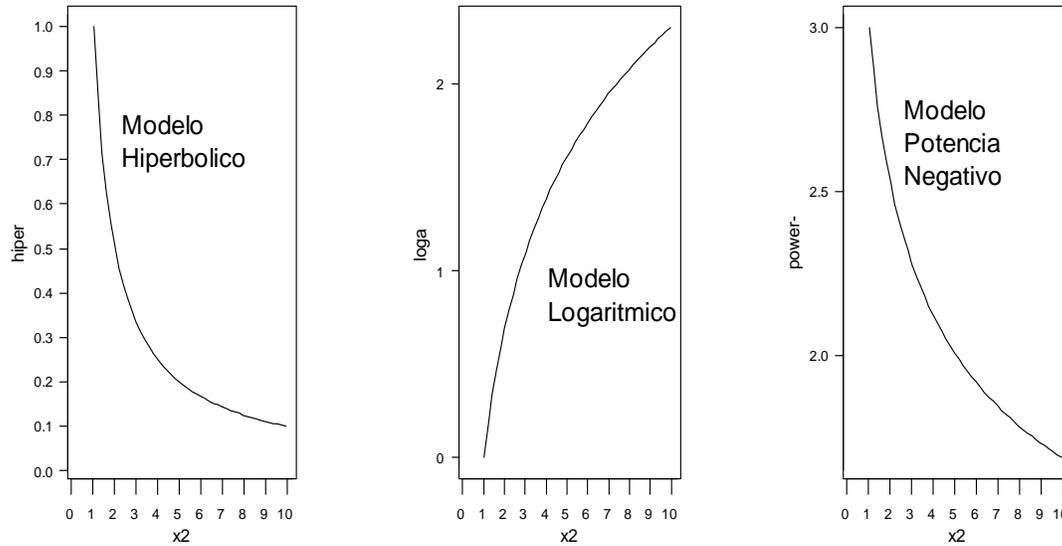


Figura 4.2. Mas gráficas de modelos no lineales

La siguiente tabla muestra las transformaciones de la variable predictora y/o respuesta que se requieren para linealizar varios modelos.

Nombre del modelo	Ecuación del Modelo	Transformación	Modelo Linealizado
Exponencial	$Y = \alpha e^{\beta X}$	$Z = \text{Log } Y$ $X = X$	$Z = \text{Log } \alpha + \beta X$
Logaritmico	$Y = \alpha + \beta \text{Log } X$	$Y = Y$ $W = \text{Log } X$	$Y = \alpha + \beta W$
Doblemente Logarítmico o Potencia	$Y = \alpha X^{\beta}$	$Z = \text{Log } Y$ $W = \text{Log } X$	$Z = \text{Log } \alpha + \beta W$
Hiperbólico	$Y = \alpha + \beta/X$	$Y = Y$ $W = 1/X$	$Y = \alpha + \beta W$
Doblemente Inverso	$Y = 1/(\alpha + \beta X)$	$Z = 1/Y$ $X = X$	$Z = \alpha + \beta X$

El primer y tercer modelo son válidos bajo la suposición de que los errores son multiplicativos y habría que cotejar haciendo análisis de residuales. Si el logaritmo de los errores tiene una media de cero y varianza constante entonces se cumplirían los supuestos. Si los errores no son multiplicativos entonces deberían aplicarse técnicas de regresión no lineal las cuales no son consideradas en este texto.

**Ejemplo 1.** Los siguientes datos representan como ha cambiado la población en Puerto Rico desde 1930.

```

year  poblacion
1930  1552000
1940  1877800

```

1950	2218000
1960	2359800
1970	2716300
1980	3196520
1990	3527796

Se desea establecer un modelo para predecir la población de Puerto Rico en el año 2000.

**Solución:** Observando el diagrama de puntos de población versus años que aparece en la figura de abajo

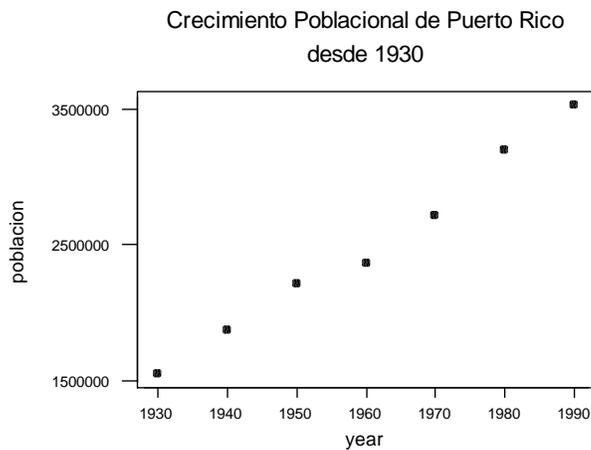


Figura 4.3 Crecimiento poblacional de Puerto Rico desde 1930

El plot sugiere que podemos ajustar los datos al modelo exponencial

$$\text{Poblac} = \alpha e^{\beta \text{year}}$$

Y el modelo linealizado da como ecuación

$$\text{Ln}(\text{Poblac}) = -11.4 + 0.0133 \text{ year}$$

con un  $R^2$  del 98.9%, mejorando el  $R^2$  del modelo lineal que era de 98.7%. Para predecir la población para el año 2000 se obtiene que

$$\text{Ln}(\text{Poblac}) = -11.4 + 0.0133(2000) = -11.4 + 26.6 = 15.2$$

luego  $\text{Poblac} = e^{15.2} = 3,992,787$  será la población de PR estimada para el año 2000.

## 4.2 Transformaciones para estabilizar la varianza.

Algunas veces el comportamiento de la varianza varía según la variable de respuesta. Una de las medidas remediales para hacer constante la varianza es transformar la variable de respuesta. La siguiente tabla muestra las transformaciones de la variable de respuesta que hay que hacer para hacer que la varianza sea constante

Situación	Transformación
$\text{Var}(e_i) \propto E(y_i)$	$\sqrt{y}$
Igual que el caso anterior	$\sqrt{y} + \sqrt{y+1}$
$\text{Var}(e_i) \propto (E(y_i))^2$	$\text{Log}(Y)$
Igual que el caso anterior	$\text{Log}(y+1)$
$\text{Var}(e_i) \propto (E(y_i))^4$	$1/y$
Igual que el caso anterior	$1/(y+1)$
$\text{Var}(e_i) \propto E(y_i)[1-E(y_i)]$	$\text{Sin}^{-1}(\sqrt{y})$

Las transformaciones se justifican de la siguiente manera:

Expandiendo en series de Taylor una función  $h(Y)$  alrededor de  $\mu=E(Y)$  se obtiene

$$h(Y) \approx h(\mu) + h'(\mu)(Y - \mu) + h''(\mu)(Y - \mu)^2 / 2 \quad (4.1)$$

Tomando varianza a ambos lados y considerando solamente la aproximación lineal se obtiene

$$\text{Var}(h(Y)) \approx [h'(E(y))]^2 \text{Var}(Y) \quad (4.2)$$

Por ejemplo, si  $\text{Var}(Y) \propto [E(y)]^2$  se tendrá que  $[h'(E(Y))]^2 \approx \text{constante}/[E(y)]^2$ . Luego,  $h'(\mu) \approx 1/\mu$ , de donde por integración resulta  $h(\mu) \approx \log(\mu)$ .

Haciendo un plot de residuales versus los valores ajustados de  $Y$  se puede estimar la transformación más adecuada. Aunque es mejor agrupar la variable  $Y$ , y calcular medias y desviaciones estándar para cada uno de los grupos y luego estimar la mejor línea que pasa por los puntos  $(\log \bar{Y}_g, \log S_g^2)$ .

**Ejemplo 2.** Aplicar una transformación para estabilizar la varianza en el modelo de regresión para el conjunto de datos **millaje**

**Solución.** Si observamos el plot de residuales versus valores ajustados por el modelo de regresión, el cual aparece en la figura 4.4 podemos ver que la varianza está cambiando de alguna manera con los valores  $\hat{y}$ . Se ha explorado varias transformaciones del tipo potencia para la variable de respuesta y la que ha dado mejores resultados es la transformación  $h(y)=y^{-1/2}$  que es

aquella correspondiente a la situación cuando la varianza de los errores es proporcional al cubo de la media de la variable de respuesta.

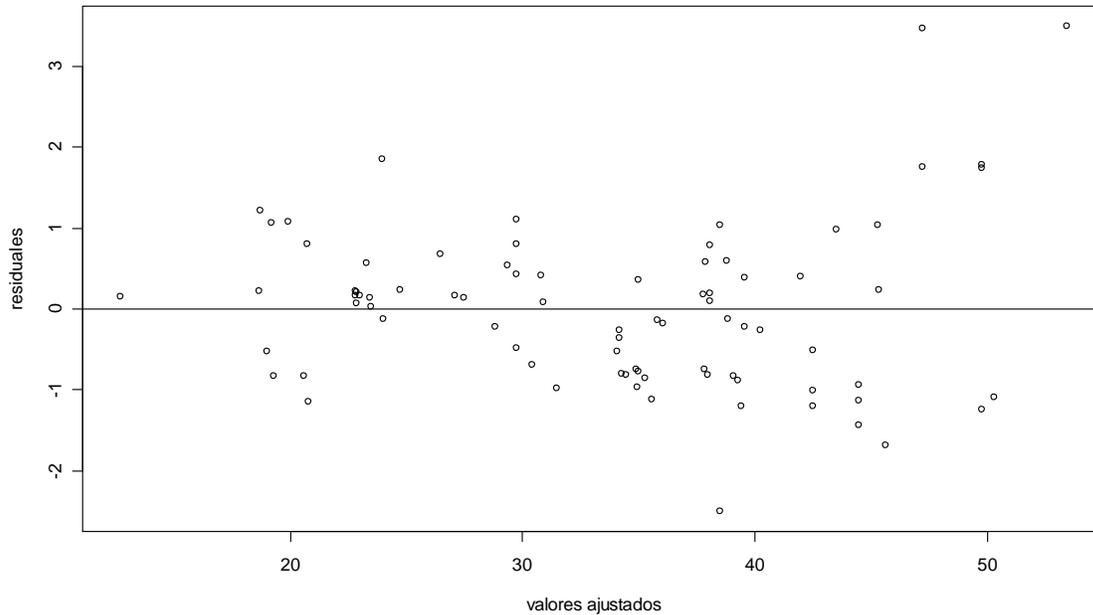


Figura 4.4. Plot de residuales estandarizados versus valores ajustados para el conjunto de datos **millaje**

```
> # El lsfitt indica que la varianza es proporcional a la media al cuadrado
> # una transformacion logaritmica en la variable de respuesta es recomendada
> mpglog<-log(millaje$mpg)
> millaje1<-cbind(millaje,mpglog)

> l2<-lm(mpglog~sp+wt+vol+hp,data=millaje1)
> summary(l2)
```

Call:

```
lm(formula = mpglog ~ sp + wt + vol + hp, data = millaje1)
```

Residuals:

```
   Min      1Q  Median      3Q      Max
-0.273816 -0.058032 -0.008837  0.038624  0.253079
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.7725247  0.5647743  10.221 5.49e-16 ***
sp          -0.0130542  0.0058747  -2.222  0.0292 *
wt          -0.0370088  0.0051209  -7.227 3.08e-10 ***
vol         -0.0003088  0.0005478  -0.564  0.5746
hp           0.0029479  0.0019540   1.509  0.1355
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08767 on 77 degrees of freedom  
 Multiple R-Squared: 0.9211, Adjusted R-squared: 0.917  
 F-statistic: 224.8 on 4 and 77 DF, p-value: < 2.2e-16

```
# Considerando que la varianza es proporcional a la media al cubo
# una transformacion h(y)=y^-0.5 es realizada
mpg05<-millaje$mpg^-0.5
millaje2<-cbind(millaje,mpg05)
l3<-lm(mpg05~sp+wt+vol+hp,data=millaje2)
summary(l3)
```

> summary(l3)

Call:

lm(formula = mpg05 ~ sp + wt + vol + hp, data = millaje2)

Residuals:

Min	1Q	Median	3Q	Max
-0.019083	-0.003005	0.001039	0.003944	0.024431

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9.141e-02	5.014e-02	1.823	0.0722 .
sp	-7.386e-05	5.215e-04	-0.142	0.8878
wt	2.398e-03	4.546e-04	5.275	1.18e-06 ***
vol	1.751e-05	4.863e-05	0.360	0.7198
hp	1.621e-04	1.735e-04	0.935	0.3529

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007783 on 77 degrees of freedom  
 Multiple R-Squared: 0.9266, Adjusted R-squared: 0.9228  
 F-statistic: 243.1 on 4 and 77 DF, p-value: < 2.2e-16

El plot de residuales versus valores ajustados es como en la Figura 4.5. Notar que no se observa ningún patron de los puntos y hay dos “outliers” bien distinguibles.

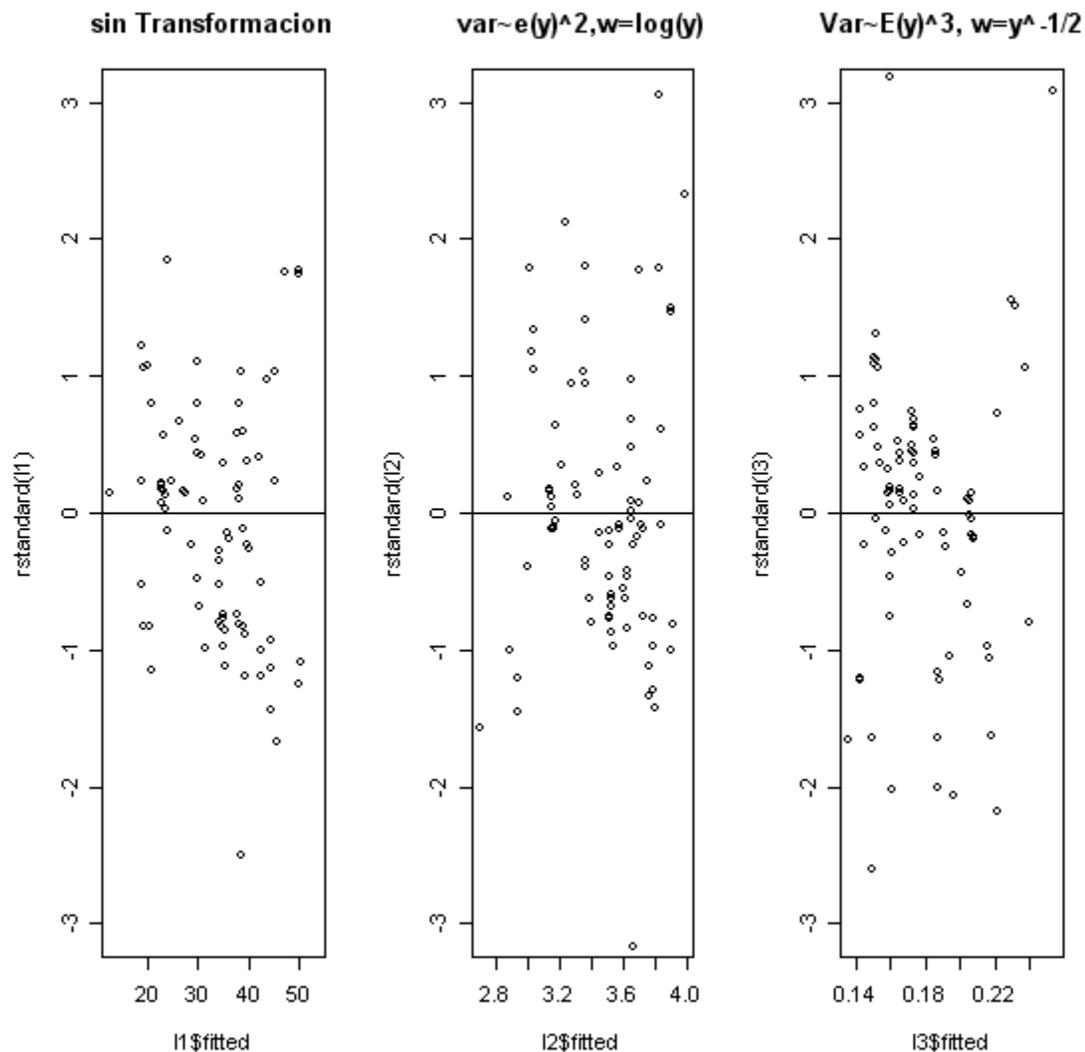


Figura 4.5. Plot de residuales versus valores ajustados después de la transformación.

### 4.3 Transformaciones de las variables predictoras en regresión múltiple

Supongamos que se tiene una variable de respuesta  $Y$  y varias variables predictoras, las cuales asumen valores positivos. Se desea hacer transformaciones en las variables predictoras para mejorar la medida de ajuste del modelo. Lo primero que uno intenta es hacer un plot matricial y de allí extraer las relaciones de  $y$  con cada una de las variables predictoras. Pero estas transformaciones se pueden ver afectadas por la colinealidad (dependencia lineal) existente entre las variables predictoras. Este mismo problema afecta al plot de regresión parcial o de variables añadidas.

En 1962, Box y Tidwell, propusieron un método para transformar las variables predictoras pero solo usando potencia de ellas. Mas específicamente, ellos consideraron el modelo

$$y = \beta_0 + \beta_1 w_1 + \dots + \beta_k w_k + e \quad (4.3)$$

donde  $w_j = x_j^{\alpha_j}$  si  $\alpha_j \neq 0$  y  $w_j = \ln(x_j)$  si  $\alpha_j = 0$ . El método está basado en el desarrollo en series de Taylor del modelo anterior con respecto a  $\alpha = (\alpha_1, \dots, \alpha_k)$  y alrededor del punto  $\alpha_0 = (\alpha_{1,0}, \dots, \alpha_{k,0}) = (1, \dots, 1)$ . Haciendo las derivaciones respectivas, el modelo (4.1) se reduce a:

$$y \cong \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + (\alpha_1 - 1)\beta_1 x_1 \ln x_1 + (\alpha_2 - 1)\beta_2 x_2 \ln x_2 + \dots + (\alpha_k - 1)\beta_k x_k \ln x_k$$

el cual es equivalente a

$$y \cong \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \gamma_1 z_1 + \gamma_2 z_2 + \dots + \gamma_k z_k \quad (4.4)$$

donde  $\gamma_j = (\alpha_j - 1)\beta_j$  y  $z_j = x_j \ln x_j$  para  $j=1,2,\dots,k$ .

El procedimiento para la estimación de los  $\alpha_j$  se puede resumir como sigue:

- Hacer la regresión lineal múltiple considerando las variables predictoras originales  $x_j$  y denotar los estimados de los coeficientes por  $b_j$ .
- Hacer una regresión lineal múltiple de  $y$  versus las variables predictoras originales mas las variables  $z_j = x_j \ln(x_j)$  y denotar los estimados de los coeficientes de  $z_j$  por  $\hat{\gamma}_j$ .
- Estimar  $\alpha_j$  por  $\hat{\alpha}_j = \frac{\hat{\gamma}_j}{b_j} + 1$

El procedimiento se puede repetir varias veces usando en cada etapa las nuevas variables transformadas y la siguiente relación de recurrencia

$$\hat{\alpha}_j^{(m+1)} = \left( \frac{\hat{\gamma}_j^{(m)}}{b_j^{(m)}} + 1 \right) \hat{\alpha}_j^{(m)} \quad (4.5)$$

Terminando el proceso cuando  $|\alpha_j^{(m+1)} - \alpha_j^{(m)}| < TOL$ , donde TOL es una cantidad de tolerancia muy cercana a cero.

Sin embargo, muy a menudo un solo paso es suficiente.

**Ejemplo 3.** Aplicar la técnica sugerida por Box and Tidwell al conjunto de datos **millaje**.

**Solución.** Usando R se obtiene

```
> l1<-lm(mpg~.,data=millaje)
> betas<-l1$coeff
> betas
(Intercept)    sp      wt      vol      hp
192.43775332 -1.29481848 -1.85980373 -0.01564501 0.39221231
> summary(l1)
```

Call:

```
lm(formula = mpg ~ ., data = millaje)
```

Residuals:

```
  Min    1Q  Median    3Q   Max
-9.0108 -2.7731  0.2733  1.8362 11.9854
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
sp          -1.29482    0.24477  -5.290 1.11e-06 ***
wt          -1.85980    0.21336  -8.717 4.22e-13 ***
vol         -0.01565    0.02283  -0.685  0.495
hp           0.39221    0.08141   4.818 7.13e-06 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.653 on 77 degrees of freedom

Multiple R-Squared: 0.8733, Adjusted R-squared: 0.8667

F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

Notar que la predictora VOL no es significativa.

La regresión con las variables originales resulta ser

$$\text{MPG} = 192.4 - 0.0156 \text{ VOL} + 0.392 \text{ HP} - 1.294 \text{ SP} - 1.859 \text{ WT}$$

Ahora creamos cuatro variables predictoras  $z_1=x_1\ln x_1$ ,  $z_2=x_2\ln x_2$ ,  $z_3=x_3\ln x_3$  y  $z_4=x_4\ln x_4$ . Haciendo la regresión múltiple con las 8 variables predictoras se obtiene

```
> l2<-lm(mpg~.,data=millaje1)
> betas2<-l2$coeff
> betas2
(Intercept)    sp      wt      vol      hp      z1
1048.2022263 -38.8522423 -17.9023484 -1.0023285  5.4675149  6.3624693
      z2      z3      z4
  3.3262799  0.1803016 -0.8006012
> summary(l2)
```

Call:

```
lm(formula = mpg ~ ., data = millaje1)
```

Residuals:

```
  Min    1Q  Median    3Q   Max
-8.0797 -1.4479 -0.1852  1.4320 10.1958
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 1048.2022  268.3693   3.906 0.000208 ***
sp          -38.8522  11.8106  -3.290 0.001546 **
wt          -17.9023   4.3238  -4.140 9.2e-05 ***
vol         -1.0023   0.5916  -1.694 0.094470 .
```

```

hp      5.4675  1.8491  2.957 0.004185 **
z1      6.3625  1.9713  3.228 0.001871 **
z2      3.3263  0.8739  3.806 0.000291 ***
z3      0.1803  0.1086  1.660 0.101185
z4     -0.8006  0.2744 -2.917 0.004690 **

```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.247 on 73 degrees of freedom  
Multiple R-Squared: 0.905, Adjusted R-squared: 0.8946  
F-statistic: 86.97 on 8 and 73 DF, p-value: < 2.2e-16

La ecuación de regresión estimada resulta ser

$$\text{MPG} = 1048.2 - 38.852 \text{ SP} - 17.902 \text{ WT} - 1.002 \text{ VOL} + 5.467 \text{ HP} + 6.362 x_1 \ln x_1 + 3.326 x_2 \ln x_2 + 0.180 x_3 \ln x_3 - 0.800 x_4 \ln x_4$$

Notar que tanto VOL como la variable  $z_3$ , relacionada a ella, son no significativas.

Aplicando el paso c) del algoritmo se tendría que

```

> gammas<-betas2[c(6:9)]
> #Hallando los alfas
> alfas<-(gammas/betas1)+1
> alfas
      z1      z2      z3      z4
-3.9137925 -0.7885113 -10.5245410 -1.0412443

```

Haciendo la regresión con las nuevas variables  $\text{vol}^{-10.52}$ ,  $\text{hp}^{-1.04}$ ,  $\text{sp}^{-3.91}$  y  $\text{wt}^{-0.79}$  se obtiene

```

> sp1<-millaje1$sp^alfas[1]
> wt1<-millaje1$wt^alfas[2]
> vol1<-millaje1$vol^alfas[3]
> hp1<-millaje1$hp^alfas[4]
> #regresion con todas las variables transformadas
> l3<-lm(millaje1$mpg~sp1+wt1+vol1+hp1)
> summary(l3)

```

Call:

```
lm(formula = millaje1$mpg ~ sp1 + wt1 + vol1 + hp1)
```

Residuals:

```

      Min       1Q   Median       3Q      Max
-8.34348 -1.62938 -0.07744  1.35872 10.15980

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.298e+00  4.420e+00  -0.520 0.604656
sp1         -1.465e+08  4.698e+08  -0.312 0.755972
wt1          3.329e+02  9.382e+01   3.548 0.000665 ***
vol1         1.843e+18  8.827e+17   2.088 0.040082 *

```

```
hp1      1.668e+03 8.078e+02 2.065 0.042325 *
```

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.095 on 77 degrees of freedom
Multiple R-Squared: 0.909, Adjusted R-squared: 0.9043
F-statistic: 192.4 on 4 and 77 DF, p-value: < 2.2e-16
```

Hay problemas con la variable transformada de VOL, su coeficiente estimado es enormemente grande, porque todas sus entradas se hacen demasiado pequeñas.

Repitiendo el proceso, eliminado VOL antes de aplicar el método de Box and Tidwell se obtiene que

```
> millaje2<-cbind(millaje2,z11,z21,z31)
> l21<-lm(mpg~.,data=millaje2)
> betas22<-l21$coeff
> gammas1<-betas22[c(5:7)]
> #Hallando los alfas1
> alfas1<-(gammas1/betas12)+1
> alfas1
      z11      z21      z31
-4.3033518 -0.9219605 -1.0966294
```

Luego,  $\alpha_1=-1.09$ ,  $\alpha_2=-4.30$  y  $\alpha_3=-0.92$

```
> #Creando las nuevas variables
> sp11<-millaje2$sp^alfas1[1]
> wt11<-millaje2$wt^alfas1[2]
> hp11<-millaje2$hp^alfas1[3]
> l5<-lm(millaje2$mpg~sp11+wt11+hp11)
> summary(l5)
```

Call:

```
lm(formula = millaje2$mpg ~ sp11 + wt11 + hp11)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-8.60068 -1.61086  0.08952  1.18229 12.43902
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.084e+00 3.681e+00  0.566 0.57286
sp11      -1.507e+09 2.800e+09 -0.538 0.59186
wt11       4.503e+02 1.345e+02  3.348 0.00125 **
hp11       2.146e+03 1.052e+03  2.040 0.04475 *
```

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.167 on 78 degrees of freedom
Multiple R-Squared: 0.9035, Adjusted R-squared: 0.8998
```

F-statistic: 243.4 on 3 and 78 DF, p-value: < 2.2e-16

Luego la regresión estimada es

MPG = 2.084 + 2146 hp11 + 450.3 wt11 -1.507e+09 sp11

Observese que la predictora SP11 no es significativa y podríamos sacarla del modelo. El cual se reduciría ahora a

```
> #Haciendo la regresion con solo las dos variables significativas
> l6<-lm(millaje2$mpg~wt11+hp11)
> summary(l6)
```

Call:

lm(formula = millaje2\$mpg ~ wt11 + hp11)

Residuals:

Min	1Q	Median	3Q	Max
-8.67047	-1.66461	0.04419	1.21415	12.53739

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.3744	1.8522	0.202	0.84
wt11	511.1816	72.4007	7.060	5.73e-10 ***
hp11	1600.0983	280.2234	5.710	1.90e-07 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.153 on 79 degrees of freedom

Multiple R-Squared: 0.9031, Adjusted R-squared: 0.9007

F-statistic: 368.3 on 2 and 79 DF, p-value: < 2.2e-16

La ecuación de regression estimada es:

MPG = 0.374 + 1600.0 hp1 + 511.1 wt1

Donde  $hp1=1/hp^{1.09}$  y  $wt1=1/wt^{0.92}$ . En la sección 2.3.5 habíamos llegado a establecer que el mejor modelo era de MPG versus  $hpo=1/hp$  y  $wto=1/wt$ . Los resultados eran como sigue:

```
> reg1=lm(mpg~hp0+wt0)
> summary(reg1)
```

Call:

lm(formula = mpg ~ hp0 + wt0)

Residuals:

Min	1Q	Median	3Q	Max
-8.70343	-1.70579	-0.01131	1.20593	12.60609

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.217	1.573	0.773	0.442

```
hp0    1131.489  200.586  5.641 2.54e-07 ***
wt0    610.370   87.913  6.943 9.61e-10 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.165 on 79 degrees of freedom
```

```
Multiple R-Squared: 0.9024,    Adjusted R-squared: 0.8999
```

```
F-statistic: 365.3 on 2 and 79 DF,  p-value: < 2.2e-16
```

Notar pues que la transformación de Box y Tidwell parece ser bastante eficiente.

#### 4.4. Transformaciones para mejorar la normalidad de la variable de respuesta

En 1964, Box y Cox introdujeron una transformación de la variable de respuesta con el objetivo de satisfacer la suposición de normalidad del modelo de regresión. La transformación es de la forma  $y^\lambda$  (transformación potencia), donde  $\lambda$  es estimada con los datos tomados. Más

específicamente, la transformación está definida por  $w = \frac{y^\lambda - 1}{\lambda}$  si  $\lambda \neq 0$  y  $w = \ln(y)$  si  $\lambda = 0$ . Notar

que  $\lim_{\lambda \rightarrow 0} \frac{y^\lambda - 1}{\lambda} = \ln y$ . En la figura 4.6 se muestra la gráfica de la transformación Box-Cox para

cinco valores distintos de lambda.

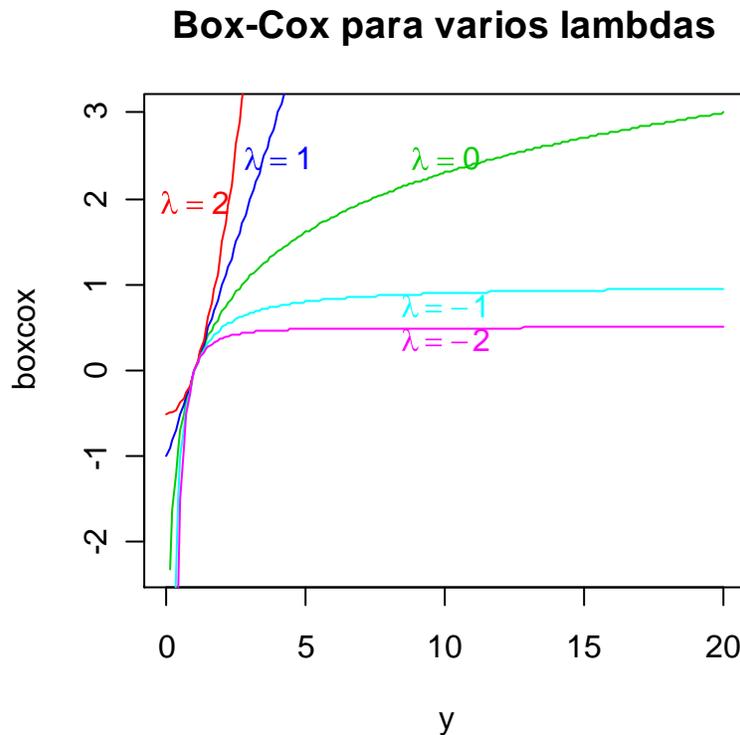


Figure 4.6 Transformación Box-Cox para varios valores de lambda

El parámetro  $\lambda$  se estima conjuntamente con los coeficientes del modelo de regresión lineal múltiple usando el método de Máxima verosimilitud

$$w = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + e \quad (4.6)$$

Notar que  $Var(w) \approx \mu_y^{2(\lambda-1)} \sigma_y^2$ . Como se quiere  $Var(w_1) = \dots = Var(w_n) = C$ , se tiene que

$\prod_{i=1}^n \mu_{y_i}^{2(\lambda-1)} \sigma_{y_i}^2 = C^n$ . La transformación estandarizada de los  $w$ 's se define por

$$z_i = \frac{w_i}{\tilde{y}^{\lambda-1}} \quad (4.7)$$

donde  $\tilde{y} = (\prod_{i=1}^n y_i)^{1/n}$ , es la media geométrica de las  $y$ 's. Luego, el modelo (4.6) se convierte en

$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ . El método asume que para algún  $\lambda$  las  $z_i$ 's son normales e independientes con varianza común  $\sigma^2$ .

Escribiendo la función de verosimilitud, correspondiente al modelo transformado, en términos de las  $z_i$ 's se tiene que

$$L(\boldsymbol{\beta}, \lambda) = \frac{e^{-\frac{1}{2\sigma^2} \mathbf{e}'\mathbf{e}}}{(2\pi\sigma^2)^{n/2}} = \frac{e^{-\frac{1}{2\sigma^2} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})}}{(2\pi\sigma^2)^{n/2}}$$

Luego se puede establecer que el máximo del logaritmo de la función de verosimilitud está dado por:

$$LnL(\hat{\boldsymbol{\beta}}, \lambda) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) \quad (4.8)$$

donde  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{z}$ , y  $\hat{\sigma}^2 = SSE/n = (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})/n$ . Luego,

$$LnL(\hat{\boldsymbol{\beta}}, \lambda) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\hat{\sigma}^2) - \frac{n}{2} \equiv -\frac{n}{2} \ln(\hat{\sigma}^2) \quad (4.9)$$

Claramente (4.9) depende de  $\lambda$  puesto que  $\hat{\sigma}^2$  depende de  $\mathbf{z}$  y ésta a su vez de  $\lambda$ .

El procedimiento para estimar el parámetro  $\lambda$  es el siguiente:

- 1) Seleccionar una conjunto de valores de  $\lambda$  entre  $-2$  y  $2$ , usualmente entre 10 y 20 valores
- 2) Para cada valor de  $\lambda$ , ajustar el modelo

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- 3) Plotear  $\max[Ln L(\boldsymbol{\beta}, \lambda)]$  versus  $\lambda$ .
- 4) Escoger como parámetro  $\lambda$  aquel valor que da el mayor valor para  $\max[Ln L(\boldsymbol{\beta}, \lambda)]$ .

Varios programas estadísticos, entre ellos S-Plus y R, tienen funciones que permiten estimar el parámetro  $\lambda$  de la transformación Box-Cox. Además del plot del paso 3 producen un intervalo de confianza para  $\lambda$ .

**Ejemplo 4.** Aplicar la transformación de Box y Cox al conjunto de datos **millaje**

**Solución:** Haremos uso de R, cuya librería MASS incluye la función **boxcox**. Los resultados usando las variables originales son como sigue:

```
> reg1<-lm(MPG~VOL+HP+SP+WT,data=MILLAJE)
> summary(reg1)
```

```
Call: lm(formula = MPG ~ VOL + HP + SP + WT, data = MILLAJE)
```

```
Residuals:
```

```
  Min   1Q Median   3Q   Max
-9.011 -2.773 0.2733 1.836 11.99
```

```
Coefficients:
```

```
      Value Std. Error t value Pr(>|t|)
(Intercept) 192.4378  23.5316   8.1778  0.0000
      VOL   -0.0156   0.0228  -0.6854  0.4951
      HP    0.3922   0.0814   4.8176  0.0000
      SP   -1.2948   0.2448  -5.2899  0.0000
      WT   -1.8598   0.2134  -8.7166  0.0000
```

```
Residual standard error: 3.653 on 77 degrees of freedom
```

```
Multiple R-Squared: 0.8733
```

```
F-statistic: 132.7 on 4 and 77 degrees of freedom, the p-value is 0
```

```
Correlation of Coefficients:
```

```
  (Intercept)  VOL   HP   SP
VOL  0.1049
HP  0.9814   0.2324
SP -0.9961  -0.1501 -0.9837
WT -0.8658  -0.4260 -0.9228  0.8555
```

```
Aplicando la función boxcox
```

```
> boxcox(reg1,lambda=seq(-.6,.6,length=20),plotit=T)
```

Se obtiene el plot de la siguiente figura, donde el parámetro  $\lambda$  puede ser estimado por  $-0.22$ , con un intervalo de confianza de  $(-0.54,0.10)$

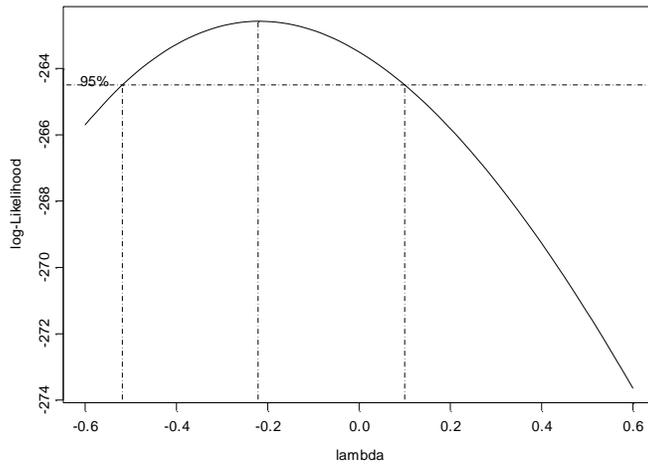


Figura 4.7. Plot de log-likelihood para varios valores de  $\lambda$

Ahora veremos el efecto de la transformación

```
> millaje1=millaje
> millaje1$mpg<-((millaje$mpg)^-0.22-1)/-0.22
> reg2<-lm(mpg~vol+hp+sp+wt,data=millaje1)
> summary(reg2)
```

Call: lm(formula = mpg ~ vol + hp + sp + wt, data = millaje1)

Residuals:

Min	1Q	Median	3Q	Max
-0.128	-0.0237	-0.004189	0.01595	0.1096

Coefficients:

	Value	Std. Error	t value	Pr(> t )
(Intercept)	3.2290	0.2600	12.4214	0.0000
vol	-0.0001	0.0003	-0.4866	0.6279
hp	0.0004	0.0009	0.4573	0.6488
sp	-0.0033	0.0027	-1.2145	0.2283
wt	-0.0152	0.0024	-6.4641	0.0000

Residual standard error: 0.04035 on 77 degrees of freedom

Multiple R-Squared: 0.9252

F-statistic: 238.2 on 4 and 77 degrees of freedom, the p-value is 0

Correlation of Coefficients:

(Intercept)	vol	hp	sp	
vol	0.1049			
hp	0.9814	0.2324		
sp	-0.9961	-0.1501	-0.9837	
wt	-0.8658	-0.4260	-0.9228	0.8555

Los plots para cotejar normalidad de los residuales se muestra en la figura 4.8

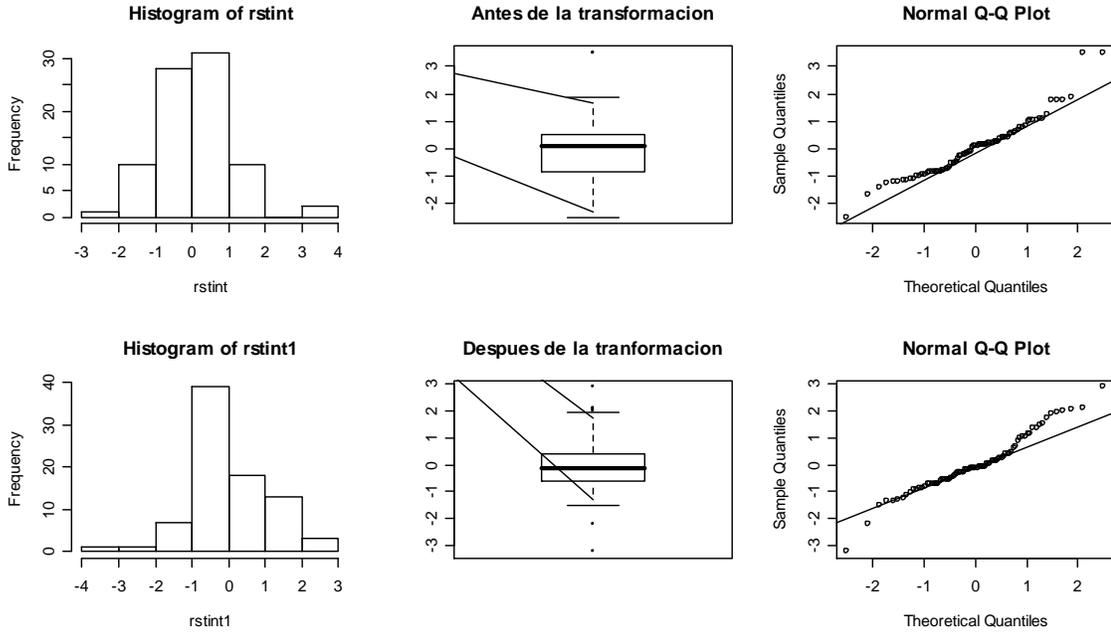


Figura 4.8. Plots para ver el efecto de la transformación Box-Cox en la distribución de los residuales de la regresión para el conjunto de datos millaje.

Notar que los puntos están mejor alineados que en plot con las variables originales (ver figura 4.8) especialmente en la parte central. Se observan claramente dos “outliers” inferiores y uno superior. Notar que el  $R^2$  ha subido de 87.33% a 92.52%, mejorando el efecto de transformar las variables predictoras que se llevó a cabo en el ejemplo 3.

**4.5 Mínimos cuadrados ponderados.**

Otra manera de tratar de remediar la falta de homogeneidad de varianza de los errores es usar mínimos cuadrados ponderados, suponiendo que los errores son todavía no correlacionados. En

este caso se minimiza  $\sum_{i=1}^n w_i (y_i - \hat{y}_i)^2$ , donde  $w_i$  representa el peso asignado a la  $i$ -ésima

observación. Por ejemplo, si en el plot de residuales versus la variable predictor se observa que la dispersión aumenta cuando  $x$  aumenta sería conveniente usar  $w_i = \frac{1}{\sigma_i^2}$ . Aquí,  $\sigma_i^2$  son las

varianzas poblacionales de la  $Y$  para cada observación  $x_i$  en caso de regresión lineal simple o para cada combinación de las variables predictoras en el caso de regresión lineal múltiple. Obviamente estas varianzas no son conocidas y deben ser estimadas por sus varianzas muestrales  $s_i^2$ . Si hay solamente una observación y para el valor  $x_i$  entonces se consideran valores de  $y$  y correspondientes a valores cercanos a  $x_i$ . En otras palabras la variable  $x$  es considerada agrupada.

Esta no es la única manera de escoger los pesos, en regresión robusta que será tratada en el capítulo 8, se hacen distintos cálculos de los pesos con la idea de dar a las observaciones

anómalas un menor peso. El cálculo de los pesos está basado mayormente en los diagnósticos de reesión.

Consideremos el modelo de regresión lineal múltiple

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (4.10)$$

con  $\text{Var}(\mathbf{e}) = \mathbf{V}\sigma^2$ , donde  $\mathbf{V}$  es una matriz diagonal. Es decir,

$$\mathbf{V} = \begin{bmatrix} k_1^2 & 0 & \cdot & \cdot & 0 \\ 0 & k_2^2 & \cdot & \cdot & 0 \\ 0 & 0 & k_3^2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & k_n^2 \end{bmatrix}$$

Sea  $\mathbf{W} = (\mathbf{V}^{1/2})^{-1}$ , claramente,  $\mathbf{W}'\mathbf{W} = \mathbf{V}^{-1}$ . Multiplicando ambos lados del modelo lineal (4.10) por  $\mathbf{W}$  se obtiene

$$\mathbf{W}\mathbf{y} = \mathbf{W}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{e} \quad (4.11)$$

Sea  $\mathbf{y}^* = \mathbf{W}\mathbf{y}$ ,  $\mathbf{e}^* = \mathbf{W}\mathbf{e}$  y  $\mathbf{X}^* = \mathbf{W}\mathbf{X}$ , entonces el modelo (4.11) se convierte en el modelo

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^* \quad (4.12)$$

Notar que  $\text{Var}(\mathbf{e}^*) = \text{Var}(\mathbf{W}\mathbf{e}) = \mathbf{W}\text{Var}(\mathbf{e})\mathbf{W}' = \mathbf{W}\mathbf{V}\mathbf{W}'\sigma^2 = \mathbf{I}\sigma^2$ , así que la varianza de los errores es constante. Luego el estimador mínimo cuadrático de  $\boldsymbol{\beta}$  será

$$\boldsymbol{\beta}^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^* = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}.$$

Se puede ver fácilmente que  $\mathbf{E}(\boldsymbol{\beta}^*) = \boldsymbol{\beta}$  y que

$$\text{Var}(\boldsymbol{\beta}^*) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\text{Var}(\mathbf{Y})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$$

**Ejemplo 5:** Consideremos las variables MPG y WT del conjunto de datos **millaje** y que además la primera y última observacion han sido eliminadas. El plot de residuales versus la variable predictora WTO (WT excluyendo la primera y última observación) aparece en la figura 4.9.

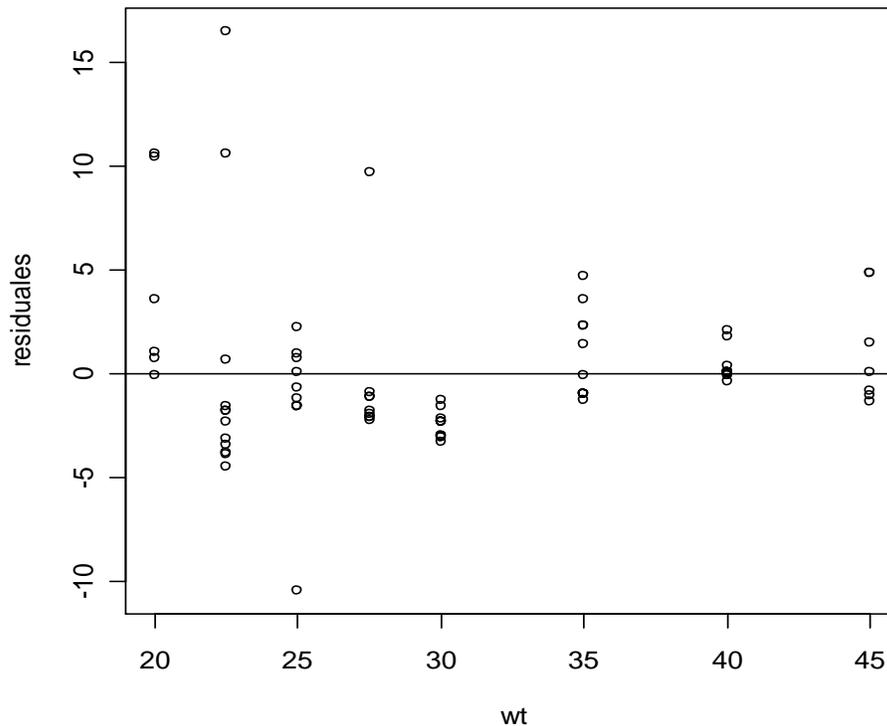


Figura 4.9. Plot de residuales versus la variable WTO donde se observa que la varianza no es homogénea

Aunque es difícil verlo en forma definitiva la variabilidad de los residuales está disminuyendo cuando la variable predictora aumenta. Algo más formal sería calcular la varianza de las  $y$ 's por cada valor de  $X$ . esto produce la siguiente tabla de valores.

$X_i$	$n_i$	$s_i^2$
20.0	6	23.8987
22.5	12	42.7533
25.0	10	12.0049
27.5	9	14.5586
30.0	12	0.3961
35.0	12	4.4627
40.0	12	0.5973
45.0	7	7.2948

Haciendo un plot de  $x_i$  versus  $s_i^2$  parece haber una buena relación cuadrática entre ambas variables. En la figura 4.9 se observa el plot de puntos y la regresión cuadrática. La ecuación del modelo resulta ser

$$s_i^2 = 148.482 - 7.81488 X_i + 0.103609 X_i^{**2}$$

El  $R^2=64.7$ .

Para determinar los pesos hay dos alternativas:

**Primera alternativa:** (Myers pag.281 Weisberg pag 85). Usar  $w_i = \frac{1}{s_i^2}$ , los  $s_i^2$  están dados en

la tabla anterior. Para usar esta alternativa debería haber un número razonable de observaciones y's para cada  $X_i$ .

**Segunda Alternativa:** (Draper y Smith, pag 226). Usar la ecuación de la regresión cuadrática para estimar las varianzas muestrales  $s_i^2$  para cada  $x_i$ . Luego, escogemos los pesos como el

recíproco de la varianzas muestrales estimadas. Es decir,  $w_i = \frac{1}{\hat{s}_i^2}$ , donde  $\hat{s}_i^2$  es el valor correspondiente a un  $X_i$  en el modelo cuadrático.

Los resultados que se obtienen en R son los siguientes:

#### a) Análisis sin usar regresión ponderada.

```
> millaje1<-millaje[-c(1,82),c(1,3)]
> l1=lm(mpg~wt,data=millaje1)
> summary(l1)
```

Call:

```
lm(formula = mpg ~ wt, data = millaje1)
```

Residuals:

```
   Min     1Q  Median     3Q    Max
-10.484 -1.992 -1.017  0.720 16.474
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.40171   1.78634   37.73 <2e-16 ***
wt          -1.09671   0.05635  -19.46 <2e-16 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.825 on 78 degrees of freedom

Multiple R-Squared: 0.8293, Adjusted R-squared: 0.8271

F-statistic: 378.8 on 1 and 78 DF, p-value: < 2.2e-16

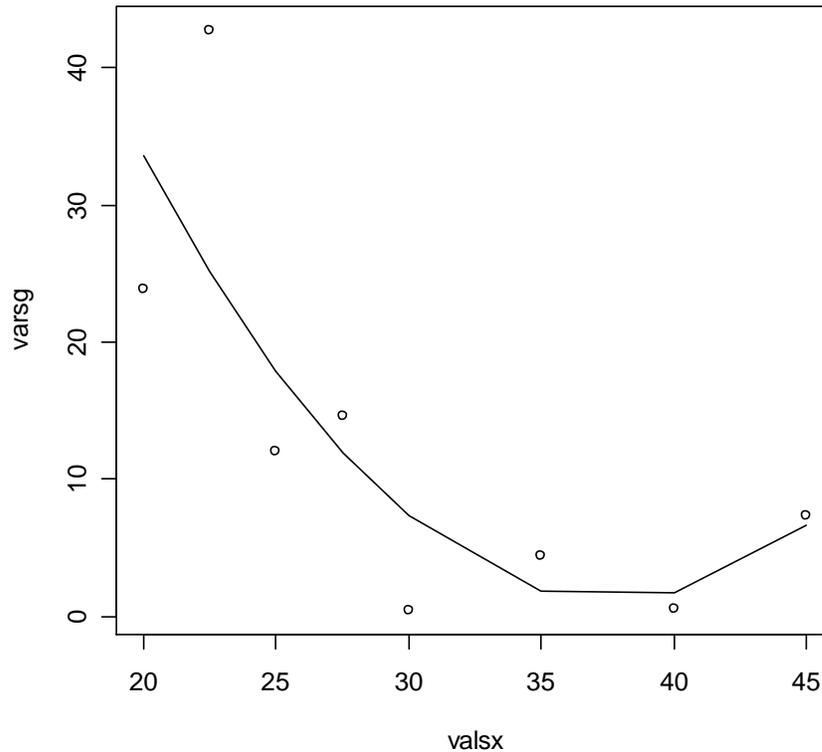


Figura 4.10. Ajuste cuadrático de la varianza versus la variable predictora

### B) Análisis de regresión ponderada con la alternativa a).

```
> summary(lw1)
```

Call:

```
lm(formula = mpg ~ ., data = millaje1, weights = pesos)
```

Residuals:

```
   Min    1Q  Median    3Q   Max
-2.1007 -0.1768  0.1953  1.0526  3.2224
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.51692   1.06539   54.92 <2e-16 ***
wt          -0.86954   0.03107  -27.99 <2e-16 ***
---
```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.206 on 78 degrees of freedom

Multiple R-Squared: 0.9094, Adjusted R-squared: 0.9083

F-statistic: 783.4 on 1 and 78 DF, p-value: < 2.2e-16

### C) Análisis de Regresión ponderada usando la alternativa b)

```
> lw2<-lm(mpg~.,data=millaje1,weights=pesos1)
> summary(lw2)
```

Call:

```
lm(formula = mpg ~ ., data = millaje1, weights = pesos1)
```

Residuals:

```
   Min     1Q   Median     3Q    Max
-2.3321 -0.7047 -0.3122  0.2965  3.4499
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.94848   1.72500   37.65 <2e-16 ***
wt          -1.02365   0.04738  -21.61 <2e-16 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.082 on 78 degrees of freedom

Multiple R-Squared: 0.8568, Adjusted R-squared: 0.855

F-statistic: 466.8 on 1 and 78 DF, p-value: < 2.2e-16

Observese que cuando se hace la regresión ponderada con la alternativa a) se obtiene una mejora del 7.0% en el  $R^2$  mientras que con la alternativa b) solo se mejora un 2.8% .

## 4.6 Mínimos Cuadrados generalizados

Consideremos ahora la situación más general de que los errores no tiene varianza constante y además que son correlacionados. Sea el modelo de regresión lineal múltiple

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Supongamos ahora que  $\text{Var}(\mathbf{e}) = \mathbf{V}\sigma^2$ , donde  $\mathbf{V}$  es una matriz simétrica y definida positiva. Un caso particular de  $\mathbf{V}$  es cuando los errores tienen distinta varianza y no están correlacionados.

Siempre es posible encontrar una matriz nonsingular y simétrica  $\mathbf{T}$  tal que  $\mathbf{T}\mathbf{T} = \mathbf{T}^2 = \mathbf{V}$ . Mutiplicando ambos lados del modelo anterior por  $\mathbf{T}^{-1}$  se obtiene

$$\mathbf{T}^{-1}\mathbf{y} = \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{T}^{-1}\mathbf{e}$$

Sea  $\mathbf{e}^* = \mathbf{T}^{-1}\mathbf{e}$ , notando que  $\text{Var}(\mathbf{e}^*) = \text{Var}(\mathbf{T}^{-1}\mathbf{e}) = \mathbf{T}^{-1}\text{Var}(\mathbf{e})\mathbf{T}^{-1} = \mathbf{I}\sigma^2$  entonces el estimador mínimo cuadrático de  $\boldsymbol{\beta}$  se obtiene minimizando

$$\mathbf{e}^{*\prime}\mathbf{e}^* = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

---

lo cual produce  $\beta^* = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{Y}$ . Se puede ver fácilmente que  $\mathbf{E}(\beta^*) = \beta$  y que  $\mathbf{Var}(\beta^*) = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{Var}(\mathbf{Y}) \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \sigma^2$ .

## Ejercicios

1. Usar el conjunto de datos **Highway**, con variable de respuesta es RATE y todas las otras como variables predictoras para responder las siguientes preguntas.
  - a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el  $R^2$  de la regresión usando como predictora la que tiene mas alta correlacion
  - b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
  - c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
  - d) Aplicar una transformación tipo potencia para estabilizar la varianza
  - e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.
  
2. Usar el conjunto de datos **Fuel** con variable de respuesta es Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas.
  - a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el  $R^2$  de la regresión usando como predictora la que tiene mas alta correlacion
  - b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
  - c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
  - d) Aplicar una transformación tipo potencia para estabilizara la varianza
  - e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.
  
3. Usar el conjunto de datos **Headcirc** con variable de respuesta es headcirc (circunferencia de la cabeza del bebe) para responder a las siguientes preguntas.
  - a) Hacer 4 transformaciones que linealizan un modelo para ver si se puede incrementar el  $R^2$  de la regresión usando como predictora la que tiene mas alta correlacion
  - b) Aplicar la transformación de Box y Tidwell. Interpretar sus resultados
  - c) Aplicar la transformación Box-Cox a su modelo. Interpretar sus resultados.
  - d) Aplicar una transformación tipo potencia para estabilizar la varianza.
  - e) Hacer una regresión por mínimos cuadrados ponderados usando una variable predictora adecuada.
  
4. Verificar la relación 4.7
  
5. Prueba para detectar varianza no constante (Cook y Weisberg, 1983). Consiste de los siguientes pasos:
  - a) Calcular la regression de Y versus todas las variables predictoras y guardar los residuales  $\hat{e}_i$ .
  - b) Calcular los residuales cuadrados escalados  $u_i$  definidos por  $u_i = \frac{\hat{e}_i}{\tilde{\sigma}^2}$ , donde  $\tilde{\sigma}^2 = \frac{\sum_{i=1}^n \hat{e}_i^2}{n}$  es el estimado máximo verosímil de  $\sigma^2$ .
  - c) Calcular la regresión de  $u_i$  versus las  $\mathbf{z}_i$  incluyendo el intercepto. Los  $\mathbf{z}_i$  son las variables de las que se sospecha que depende la varianza  $\sigma^2$ . Así,  $\mathbf{z}_i = \hat{y}_i$  indica que la varianza varia con la variable de respuesta,  $\mathbf{z}_i = x_i$  indica que la varianza varía con la predictora  $x_i$ . También  $\mathbf{z}_i$  pueden contener todas las variables predictoras o un subconjunto de ellas. Guardar la Suma de cuadrado de la regresión (SSR).

d) Calcular la prueba  $S=SSR/2$ . Si  $S$  es grande entonces hay indicación de varianza no constante. Mas formalmente  $S$  se distribuye asintóticamente como una Ji cuadrado con  $q$  grados de libertad, donde  $q$  es el número de componentes de  $z_i$ , bajo la hipótesis nula de varianza constante. Aplicar la prueba definida por los pasos a –d a los datos de los ejercicios 1 y 3.

6) Deducir que transformación de la variable de respuesta hace que la varianza  $\sigma^2$  sea constante cuando ella es proporcional a  $[E(y)]^4$

## CAPÍTULO 5

### REGRESIÓN CON VARIABLES CUALITATIVAS

#### 5.1 Regresión con variables predictoras cualitativas.

Frecuentemente se considera que entre las variables predictoras, que explican el comportamiento de la variable de respuesta, hay algunas que son cualitativas o categóricas. Por ejemplo, si en una empresa se trata de explicar el salario de un empleado hay muchas variables predictoras a considerar algunas de ellas cuantitativas y otras cualitativas. Entre las variables cuantitativas estarán años de experiencia en la empresa, años de educación, edad, etc. y entre las variables cualitativas estarán el sexo del empleado, estado civil, jerarquía del empleado, etc.

Cuando una variable cualitativa asume solamente dos valores es llamada variable indicadora, variable binaria o variable “dummy”. Estas variables son codificadas numéricamente con 0’s y 1’s.

Algunas veces la variable cualitativa puede asumir más de dos valores. Por ejemplo, la variable Opinión: “A favor”, “Indeciso”, “En contra”. Se podría codificar los valores como 0, 1 y 2 pero esto estaría implicando una suposición de ordenamiento y además implicaría que el efecto de cambiar de “A favor” a “Indeciso” es lo mismo que cambiar de “Indeciso” a “En contra” (o sea se está suponiendo igual espaciamiento). Ambas suposiciones no son justificables. Una mejor alternativa es definir dos variables indicadoras

$A_1=1$  “A favor”, 0 en otro caso

$A_2= 1$  “En contra”, 0 en otro caso

La combinación  $A_1=1$  y  $A_2=0$  representaría que la variable Opinión asume el valor “A favor”, la combinación  $A_1=0$  y  $A_2=1$  representaría que la variable Opinión asume el valor “En contra”. Usar una tercera variable es redundante, puesto que los indecisos pueden ser representados por  $A_1=A_2=0$ . Estas variables cualitativas, donde el orden es irrelevante, son llamadas más propiamente variables nominales. Si la variable nominal asume  $k$  valores distintos habría que usar  $k+1$  variables indicadoras para representar todos sus valores. Es decir, el número de variables predictoras en el modelo se incrementaría en  $k$ . Si se tuviera un gran número de variables predictoras nominales el modelo de regresión se volvería bastante complejo ya que tendría un gran número de parámetros que estimar.

Las variables cualitativas donde el orden si interesa se le conoce como variables ordinales y en ese caso es más frecuente codificar la variable como una secuencia ordenada de números enteros.

En un problema de regresión debe haber por lo menos una variable predictora cuantitativa. Si todas las variables predictoras fueran cualitativas entonces el problema se convierte en uno de **diseños experimentales**.

#### 5.1.1 Regresión con una sola variable predictora cualitativa

Consideremos un modelo de regresión con una sola variable cualitativa  $A$  y una variable cuantitativa  $X$ . Es decir,

$$Y=\beta_0+\beta_1X+\beta_2A + \varepsilon \quad (5.1)$$

Notar que si  $A=0$  se obtiene el modelo lineal simple

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (5.2)$$

Y que si  $A=1$  se obtiene el modelo

$$Y = (\beta_0 + \beta_2) + \beta_1 X + \varepsilon \quad (5.3)$$

Notar que las líneas estimadas de los modelos (5.2) y (5.3) serán paralelas (igual pendiente). El valor estimado de  $\beta_2$  representa el cambio promedio en la variable de respuesta al cambiar el valor de la variable “dummy”.

**Ejemplo 1.** En el conjunto de datos **bajopeso**, disponible en la página de internet del texto, se trata de relacionar el peso de lo recién nacidos con los pesos de sus madres y la condición de fumar de las mismas. El conjunto de datos contiene 189 observaciones y será tratado en forma más completa más adelante.

**Solución.** Considerando que la variable fumar asume el valor 0 si la persona no fuma y 1 si la persona fuma se obtiene los siguientes resultados

```
> l2<-lm(pbebe~pmama+fuma)
> summary(l2)

Call:
lm(formula = pbebe ~ pmama + fuma)

Residuals:
    Min       1Q   Median       3Q      Max
-2030.90  -445.69   29.16   521.76  1967.76

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2501.125    230.836  10.835  <2e-16 ***
pmama         4.237      1.690   2.507  0.0130 *
fuma        -272.081    105.591  -2.577  0.0107 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 707.8 on 186 degrees of freedom
Multiple R-Squared:  0.06777,    Adjusted R-squared:  0.05775
F-statistic: 6.761 on 2 and 186 DF,  p-value: 0.001464
```

Notar que el  $R^2$  es bajísimo. El coeficiente de regresión estimado de fumar es  $-272$  y significa que si la mamá fuma en promedio el peso del bebé disminuirá en 272 gramos.

Podemos hacer la regresión por grupos. Es decir, una regresión para las madres que no fuman y otras para las que si fuman. Se obtienen los siguientes resultados.

Para madres no fumadoras:

```
> l3<-lm(pbebe0~pmama0)
> summary(l3)
```

Call:

```
lm(formula = pbebe0 ~ pmama0)
```

Residuals:

Min	1Q	Median	3Q	Max
-2029.87	-550.86	28.23	551.78	1976.84

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2350.578	326.583	7.197	7.21e-11 ***
pmama0	5.387	2.439	2.209	0.0292 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 740.2 on 113 degrees of freedom

Multiple R-Squared: 0.04139, Adjusted R-squared: 0.03291

F-statistic: 4.88 on 1 and 113 DF, p-value: 0.02919

Para madres fumadoras:

```
> summary(l4)
```

Call:

```
lm(formula = pbebe1 ~ pmama1)
```

Residuals:

Min	1Q	Median	3Q	Max
-2038.8	-414.4	35.0	473.6	1490.2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2391.962	301.229	7.941	1.98e-11 ***
pmama1	2.965	2.274	1.304	0.196

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 656.5 on 72 degrees of freedom

Multiple R-Squared: 0.02307, Adjusted R-squared: 0.0095

F-statistic: 1.7 on 1 and 72 DF, p-value: 0.1964

Notar que ambos casos los  $R^2$  son más bajos que el  $R^2$  anterior. Los plots son mostrados en la figura 5.1

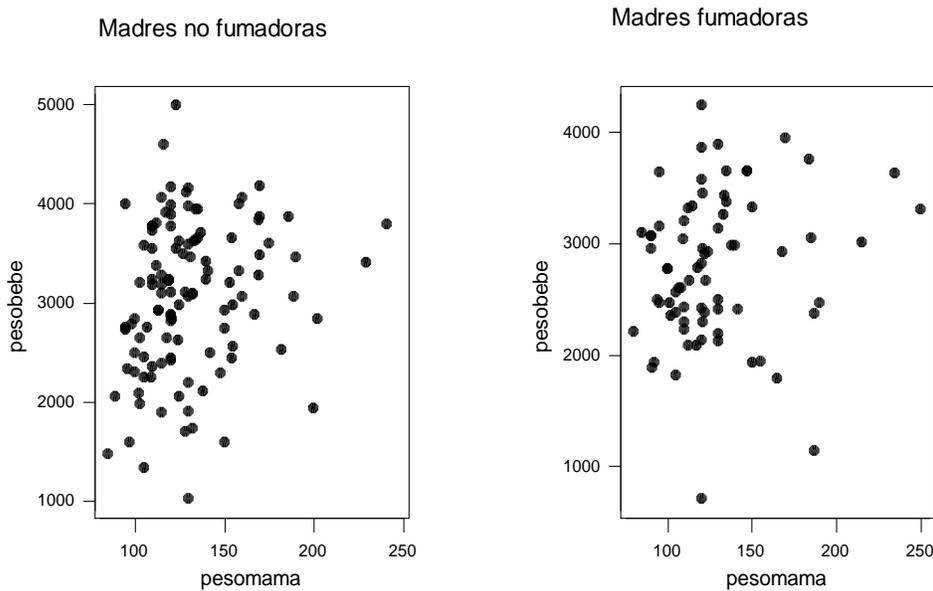


Figura 5.1. Plots de la relación pesobebe versus pesomama según la condición de fumar de la madre

En ambos plots se puede ver que no parece haber relación entre el peso del bebe y peso de la madre aunque esto es más evidente para las madres fumadoras. Los “outliers” parecen afectar más la regresión del peso bebe versus peso mama entre las madres no fumadoras.

Si se desea comparar las pendientes de las línea de regresión de los dos grupos se puede usar una prueba de t similar a la prueba de comparación de dos medias y asumiendo que hay homogeneidad de varianza. También se puede usar una prueba de F parcial o de t para probar la hipótesis  $H_0: \beta_3=0$  en el siguiente modelo

$$Y = \beta_0 + \beta_1 A + \beta_2 X + \beta_3 AX + e$$

Cuando la hipótesis nula no es rechazada se concluye que la pendiente de regresión de ambos grupos son iguales y el uso del modelo 5.1 sería adecuado.

Si no hubiera igualdad de varianza de los dos grupos, habría que usar una prueba de t aproximada similar al problema de Behrens-Fisher. La prueba está definida por

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\frac{s_1^2}{Sxx_1} + \frac{s_2^2}{Sxx_2}}}$$

Donde,  $\hat{\beta}_1$  and  $\hat{\beta}_2$  son los pendientes estimadas de cada línea de regresión y  $s_1^2$  y  $s_2^2$  son las estimaciones de la varianza del error en cada modelo.

Los grados de libertad de la prueba t se aproximan por

$$gl = \frac{(c_1 + c_2)^2}{\frac{c_1^2}{m-1} + \frac{c_2^2}{n-1}}$$

$$\text{con } c_1 = \frac{s_1^2}{m} \text{ y } c_2 = \frac{s_2^2}{n}.$$

Donde m y n son los grados de libertad de la suma de cuadrados del error en cada modelo.

### 5.1.2 Comparando las líneas de regresión de más de dos grupos.

Supongamos que se tiene una variable predictora continua X para explicar el comportamiento de Y en tres grupos. Luego hay tres modelos de regresión que se pueden comparar. Estos son:

$$\text{i) } Y = \beta_{01} + \beta_{11}X + \varepsilon$$

$$\text{ii) } Y = \beta_{02} + \beta_{12}X + \varepsilon$$

$$\text{iii) } Y = \beta_{03} + \beta_{13}X + \varepsilon$$

Para relacionar las líneas de regresión hay que introducir 3 variables “dummy” para identificar los grupos  $G_1$ ,  $G_2$ , y  $G_3$  y 3 variables adicionales  $Z_1 = G_1X$ ,  $Z_2 = G_2X$ , y  $Z_3 = G_3X$ . Otra alternativa sería usar solo dos variables “dummy”. Hay 4 posibilidades que podrían ocurrir:

a) Que las líneas se intersecten en un punto cualquiera, ya que tendrían diferente intercepto y pendiente. En este caso se ajusta el modelo  $Y = \beta_{01}G_1 + \beta_{11}Z_1 + \beta_{02}G_2 + \beta_{12}Z_2 + \beta_{03}G_3 + \beta_{13}Z_3 + \varepsilon$ .

Usando dos variables “dummy” este modelo sería  $Y = \beta_0 + \beta_1X + \beta_{01}G_1 + \beta_{02}G_2 + \beta_{11}Z_1 + \beta_{12}Z_2 + \varepsilon$ .

b) Que las líneas sean paralelas (homogeneidad de pendientes). En este caso se ajusta el modelo  $Y = \beta_{01}G_1 + \beta_{02}G_2 + \beta_{03}G_3 + \beta X + \varepsilon$

c) Que las líneas tengan el mismo intercepto con el eje Y pero distintas pendientes (homogeneidad de interceptos). En este caso se ajusta el modelo  $Y = \beta_0 + \beta_{11}Z_1 + \beta_{12}Z_2 + \beta_{13}Z_3 + \varepsilon$

d) Que las tres líneas coincidan. En este caso se ajusta el modelo  $Y = \alpha + \beta X + \varepsilon$

Para probar la hipótesis  $H_0$ : el modelo satisface b) o c) o d) versus

$H_a$ : el modelo satisface a)

Se usa una prueba de F parcial dada por

$$F_m = [(SSE_m - SSE_a) / (gl_m - gl_a)] / [SSE_a / gl_a]$$

Donde m, representa los modelos b, c, o d, y  $gl_m$  y  $gl_a$  representan los grados de libertad del error del modelo m y del modelo a, respectivamente. La F parcial se distribuye como una f con  $(gl_m - gl_a, gl_a)$  grados de libertad.

## 5.2 Regresión Logística

Consideraremos ahora que la variable de respuesta, Y, es una del tipo binario y que se tiene p variables predictoras x's, las cuales son consideradas aleatorias. Es decir, que el conjunto de

datos consiste de una muestra de tamaño  $n=n_1+n_2$ , donde  $n_1$  observaciones son de una clase  $C_1$  y  $n_2$  son de una clase  $C_2$ . Así, para cualquier observación  $\mathbf{x}_j$  la variable de respuesta  $Y$  es igual a 1 si  $\mathbf{x}_j$  es de la clase  $C_1$ , que contiene las observaciones donde el evento que estamos interesados se cumple. Mientras que  $Y$  es igual a 0 si  $\mathbf{x}_j$  pertenece a la clase  $C_2$ .

**Ejemplo 2:** El conjunto de datos **bajopeso** contiene los pesos de 189 bebés recién nacidos. Para determinar si el niño es de bajo peso (menos de 2500 gramos) o no lo es, o sea **bajopeso**=1 si peso bebe<2500 y **bajopeso**=0 en otro caso, se han medido las siguientes variables predictoras

**Edad:** edad de la madre

**Pesomama:** peso de la madre en su último período muestral

**Raza:** raza de la madre: 1=blanca, 2=negra, 3=otro

**Fuma:** 0 si la madre no fuma, 1 si lo hace.

**Prematur:** número de partos prematuros de la madre

**Hiperten:** 0 si la madre no sufre de hipertensión, 1 si sufre

**Uterirrit:** 0 si no tiene útero irritado, 1 si lo tiene.

**Chequeos:** número de visitas al médico en los tres primeros meses del embarazo.

Suponiendo que **pesomama** es la variable predictora más importante, podríamos explorar su relación con **bajopeso**. Haciendo un plot de **bajopeso** versus **pesomama** y ajustando una línea de regresión lineal simple se obtiene la siguiente figura

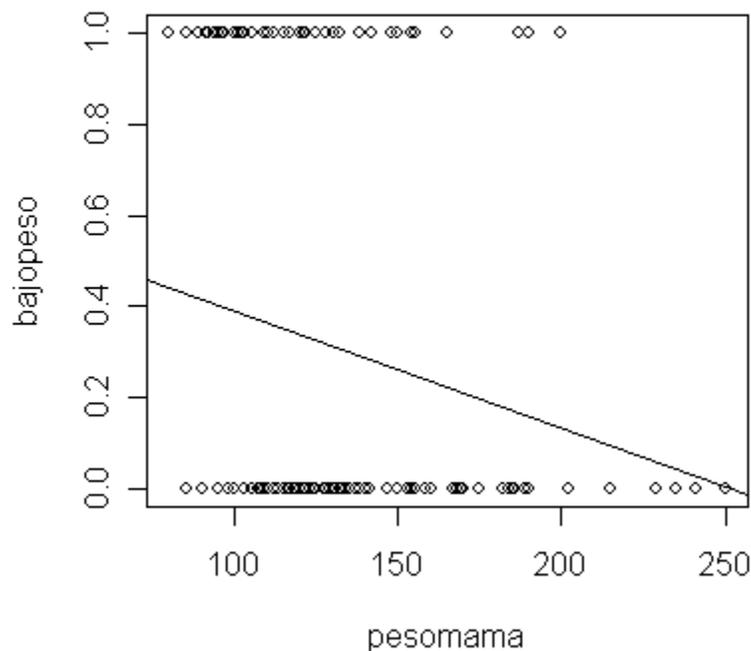


Figura 5.2. Regresión lineal estimada de los datos del ejemplo 2.

Como se puede ver es imposible que la línea de regresión represente la tendencia de los puntos. Además, la línea de regresión puede predecir valores de **bajopeso** que no son necesariamente 0 y

1, lo cual es totalmente ilógico. Asimismo, la suposición de varianza constante para la variable de respuesta no se cumple, como lo muestra el plot de residuales de la siguiente figura.

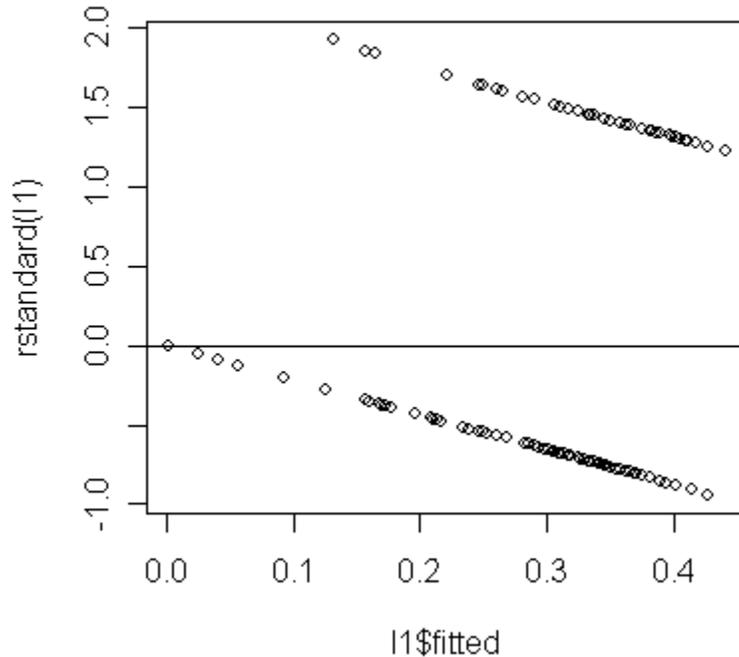


Figura 5.3. Plot de los residuales correspondiente a la regresión estimada del ejemplo 2.

Se podría usar mínimos cuadrados reponderados para remediar esta situación pero aún así conseguir predicciones 0 y 1 usando el modelo lineal sería imposible. En lugar del modelo de regresión lineal es más conveniente modelar la probabilidad de que la variable de respuesta asuma los valores 0 y 1 basado en las mediciones de las variables predictoras.

Notar que una curva en forma de S ajustaría bien los datos. Por otro lado, existe un modelo bien conocido en crecimiento poblacional cuya curva tiene esta forma y este modelo es llamado el modelo logístico y el cual se muestra en la figura 5.4. Propiamente, se ha graficado

$f(x) = \frac{1}{1+e^{-x}}$ ,  $-10 < x < 10$ . Esta curva también puede ser considerada como la gráfica de la

distribución acumulada correspondiente a la densidad logística  $p(x) = \frac{e^{-x}}{(1+e^{-x})^2}$

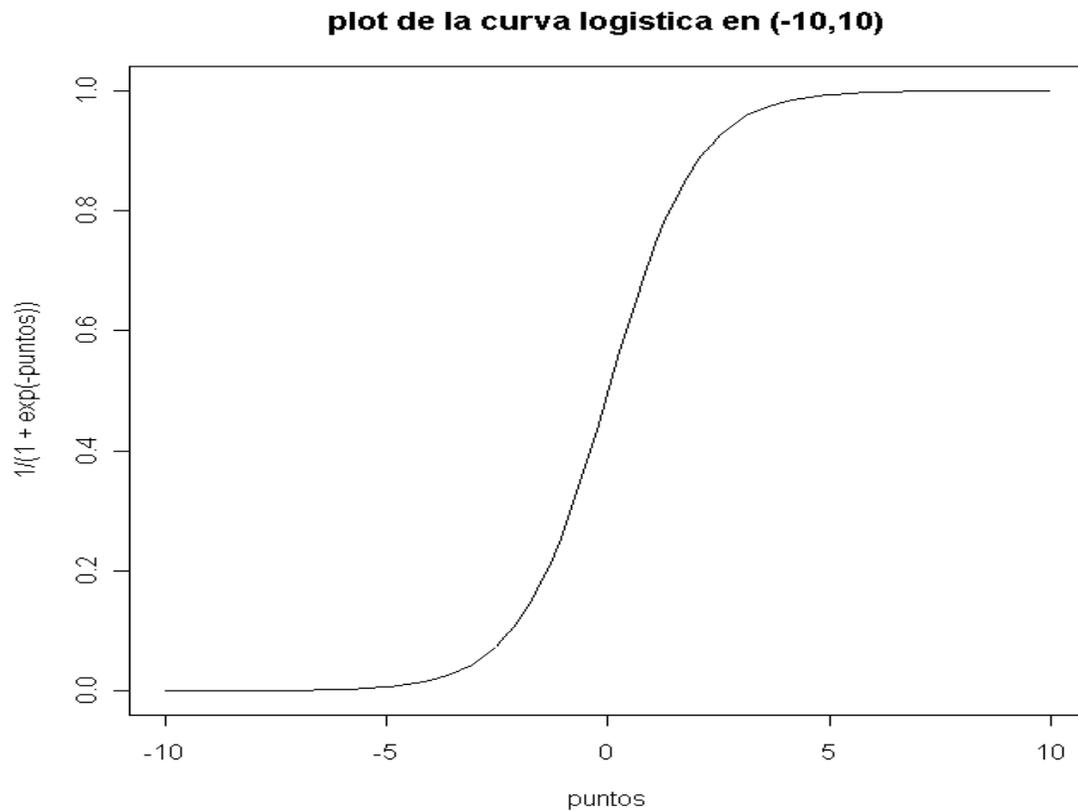


Figura 5.4. Plot de la función de distribución logistica acumulada en el intervalo (-10,10).

Algunas veces la variable de respuesta viene dada en forma agrupada en  $g$  grupos, con  $n_i$  observaciones en el  $i$ -ésimo grupo, como en el siguiente ejemplo.

**Ejemplo 3.** Se trata de relacionar la variable  $Y$ : el estudiante aprueba o no aprueba un examen con la variable  $X$ : horas de estudio para el examen. Para ello se recolecta información de los resultados de 430 estudiantes que tomaron el examen. La siguiente tabla resume dichos resultados.

X:Horas de estudio para el examen	Número de estudiantes	Estudiantes que aprueban el examen	Proporcion que aprueban
0	20	1	.05
1	50	5	.10
2	80	25	.3125
3	120	40	.3333
4	90	45	.5000
5	40	30	.7500
6	20	17	.8500
7	10	9	.9000

En total 172 son de la clase 1: aprobar el examen y 258 de la clase 0: fracasar el examen. Los estudiantes se agruparon en 8 grupos de acuerdo a sus horas de estudio.

Sea  $f(\mathbf{x}/C_i)$  ( $i=1,2$ ) la función de densidad del vector aleatorio  $p$ -dimensional  $\mathbf{x}$  en la clase  $C_i$ , en el modelo logístico se asume que

$$\log\left(\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}\right) = \alpha + \boldsymbol{\beta}' \mathbf{x} \quad (5.4)$$

Aquí  $\boldsymbol{\beta}$  es un vector de  $p$  parámetros y  $\alpha$  representa el intercepto.

Notar que si las variables  $\mathbf{x}$  en cada clase se distribuyen normalmente con igual matriz de covarianza  $\Sigma$  entonces se satisface la suposición (5.4) ya que

$$\log\left(\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}\right) = (\mathbf{u}_1 - \mathbf{u}_2)' \Sigma^{-1} (\mathbf{x} - \mathbf{1}/2(\mathbf{u}_1 + \mathbf{u}_2)) \quad (5.5)$$

En este caso  $\alpha = -(\mathbf{u}_1 - \mathbf{u}_2)' \Sigma^{-1} (\mathbf{u}_1 + \mathbf{u}_2) / 2$  y  $\boldsymbol{\beta} = (\mathbf{u}_1 - \mathbf{u}_2)' \Sigma^{-1}$ .

La suposición (5.4) se cumple también para otros tipos de distribuciones distintas de la normal multivariada tales como distribuciones de Bernoulli, y mezclas de éstas.

Por otro lado, sea  $p = P(Y=1/\mathbf{x})$  la probabilidad a posteriori de que  $Y$  sea igual a 1 para un valor observado de  $\mathbf{x}$ , entonces haciendo uso de probabilidad condicional se tiene que:

$$\frac{p}{1-p} = \frac{P\{Y=1\}f(\mathbf{x}/y=1)}{P\{Y=0\}f(\mathbf{x}/y=0)} = \frac{\pi_1 f(\mathbf{x}/C_1)}{\pi_2 f(\mathbf{x}/C_2)} \quad (5.6)$$

donde  $\pi_i$  representa la *probabilidad a priori* de que  $\mathbf{x}$  pertenezca a la clase  $C_i$ . La expresión

$\frac{p}{1-p}$  es llamado la razón de apuestas (*odds ratio*). Tomando logaritmos en ambos lados de

(5.6) se obtiene

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) + \log\left(\frac{f(\mathbf{x}/C_1)}{f(\mathbf{x}/C_2)}\right)$$

Usando la suposición (5.4), la ecuación anterior puede ser escrita como

$$\log\left(\frac{p}{1-p}\right) = \alpha + \boldsymbol{\beta}' \mathbf{x} \quad (5.7)$$

y  $\log\left(\frac{p}{1-p}\right)$  es llamado la transformación *logit*.

Despejando  $p$  de la expresión anterior se obtiene

$$p = \frac{\exp(\alpha + \boldsymbol{\beta}' \mathbf{x})}{1 + \exp(\alpha + \boldsymbol{\beta}' \mathbf{x})} \quad (5.8)$$

La ecuación (5.8) representa el modelo de la regresión logística, que fue introducida en 1944 por J. Berkson.

Un coeficiente  $b_k$  en el modelo de regresión logística estimado representa el cambio promedio de la logit función cuando la variable  $X_k$  cambia en una unidad adicional asumiendo que las otras variables permanecen constantes.

También se puede considerar que  $\exp(b_k)$  es una razón de cambio de la razón de apuestas cuando  $X_k$  varía en una unidad adicional. Si  $X_k$  es binaria entonces  $\exp(b_k)$  es el cambio en la razón de apuestas cuando ella asume el valor 1.

Cuando el modelo tiene una sola variable predictora, que además es binaria entonces existe una relación entre la regresión logística y el análisis de una tabla de contingencia 2 X 2.

### 5.2.1 Estimación del modelo logístico.

El método más usado para estimar  $\alpha$  y  $\beta$  es el método de máxima verosimilitud. Dada una observación  $\mathbf{x}$ , las probabilidades de que ésta pertenezca a las clases  $C_1$  y  $C_2$  son :

$$P(C_1 / \mathbf{x}) = \frac{\exp(\alpha + \beta' \mathbf{x})}{1 + \exp(\alpha + \beta' \mathbf{x})} \quad (5.9)$$

$$P(C_2 / \mathbf{x}) = 1 - P(C_1 / \mathbf{x}) = \frac{1}{1 + \exp(\alpha + \beta' \mathbf{x})} \quad (5.10)$$

respectivamente.

Considerando una muestra de tamaño  $n=n_1+n_2$  y un parámetro binomial  $p$  igual a  $\exp(\alpha + \beta' \mathbf{x}) / (1 + \exp(\alpha + \beta' \mathbf{x}))$  la función de verosimilitud es de la forma

$$L(\alpha, \beta) = \prod_{i=1}^{n_1} \frac{\exp(\alpha + \mathbf{x}_i' \beta)}{1 + \exp(\alpha + \mathbf{x}_i' \beta)} \cdot \prod_{j=n_1+1}^n \frac{1}{1 + \exp(\alpha + \mathbf{x}_j' \beta)} \quad (5.11)$$

asumiendo que las primeras  $n_1$  observaciones son de la clase  $C_1$  y las restantes son de la clase  $C_2$ . Los estimados  $\tilde{\alpha}$  y  $\tilde{\beta}$  son aquellos que maximizan la función anterior y son encontrados aplicando métodos iterativos tales como Newton-Raphson (SAS) o mínimos cuadrados ponderados iterativos (MINITAB, R/S-Plus).

La solución de la ecuación de verosimilitud puede no ser única si existe una marcada separación entre las dos clases.

Otra forma de hacer la estimación es como sigue: Los parámetros  $\alpha$  y  $\beta$  pueden ser estimados haciendo la regresión lineal múltiple de  $\logit(\hat{p})$  versus  $x_1, x_2, \dots, x_p$ . Usando los resultados de la sección 4.2 para aproximación de la varianza de una transformación se tiene que

$$Var\left[\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right)\right] \cong \left[\frac{1}{p(1-p)}\right]^2 \frac{p(1-p)}{n_1} = \frac{1}{n_1 p(1-p)} \quad (5.12)$$

Como  $p=p(\mathbf{x})$  se llega a un problema donde la varianza no es constante y se puede usar mínimos cuadrados ponderados con pesos  $w_i(\mathbf{x})=n_i \hat{p}(\mathbf{x})(1-\hat{p}(\mathbf{x}))$  para estimar los parámetros  $\alpha$  y  $\beta$  del modelo logístico.

La regresión logística es un caso particular de los modelos lineales generalizados (GLM) propuesto por Nelder y Wedderburn (1972). Los modelos lineales generalizados extienden los modelos lineales en dos sentidos: Primero, con la especificación de una **función link** que relaciona el esperado de la variable de respuesta con las predictoras lineales y segundo con la especificación de una función de distribución de los errores que es distinta de la Gaussiana. La forma de un modelo lineal generalizado es

$$\eta[E(y)] = \alpha + \beta \mathbf{x}$$

donde  $\eta(\cdot)$  es la función link. En un modelo de regresión clásico,  $\eta(t)=t$  y la distribución de los errores es Gaussiana o Normal. En la regresión logística  $\eta(t)=\log(t/(1-t))$  y la distribución de los errores es binomial. El modelo de regresión viene dado por  $E(y) = \eta^{-1}(\alpha + \beta \mathbf{x})$

En **MINITAB** el menú de **Regresión** contiene tres tipos de regresión logística: regresión logística binaria (aplicada a dos clases), regresión logística ordinal (si hay mas de dos clases) y regresión logística nominal (si hay mas de dos clases no ordenadas). Para ajustar un modelo logístico en **SAS** se usa el procedimiento **LOGISTIC**, mientras que en **R** se usa el procedimiento **glm** (modelos lineales generalizados) con la opción `family=binomial`. Aquí **family** representa el tipo de distribución de los errores.

### 5.2.2 Medidas de Confiabilidad del Modelo

Las siguientes son unas medidas que cuantifican el nivel de ajuste del modelo logístico al conjunto de datos:

**a) La Devianza Residual:** Es similar a la suma de cuadrados del error de la regresión lineal y se define como el negativo de dos veces la función de verosimilitud maximizada. Para los casos cuando la variable de respuesta  $Y$  no está agrupada se tiene que:

$$D = -2 \left\{ \sum_{i: y_i=1}^n \log(\hat{p}_i) + \sum_{i: y_i=0}^n \log(1 - \hat{p}_i) \right\}$$

Donde  $\hat{p}_i$  es el valor estimado de la ecuación (5.8).  $D$  es equivalente a la prueba de razón de verosimilitud para probar la validez del modelo logístico. El estadístico  $D$  se distribuye como una Ji-Cuadrado con  $n-p-1$  grados de libertad, donde  $p$  es el número de variables predictoras. Si  $D$  es mayor que una Ji-Cuadrado con  $n-p-1$  grados de libertad para un nivel de significación dado, entonces el modelo logístico no es confiable.

**b) El Pseudo- $R^2$ .** Se han propuesto versiones similares al  $R^2$  de la regresión lineal. Aquí definimos el propuesto por McFadden

$$Pseudo - R^2 = 1 - \frac{Devianza Residual}{Devianza Nula}$$

donde la Devianza Nula es la Devianza considerando solamente el intercepto y que se distribuye como una Ji-Cuadrado con  $n-1$  grados de libertad. Para hallar la Devianza Nula se hace una

regresión logística considerando que hay una sola variable predictora cuyos valores son todos unos. Un Pseudo- $R^2$  mayor de .3 es considerado como aceptable.

c) **El Criterio de Información de Akaike (AIC):** Se define por

$$AIC = D + 2(p+1)$$

Donde p es el número de variables predictoras. Un modelo es mejor que otro si su AIC es más pequeño.

d) **La Prueba de Bondad de Ajuste de Hosmer-Lemeshov.** En esta prueba los valores ajustados son agrupados en g grupos. La prueba es una del tipo  $\chi^2$  y se define por

$$C = \sum_{i=1}^g \frac{(O_i - n'_i \bar{p}_i)^2}{n'_i \bar{p}_i (1 - \bar{p}_i)}$$

donde g es el número de grupos de los valores ajustados ( g varia entre 6 y 10),  $n'_i$  es el número de observaciones en el i-ésimo grupo.  $O_i$  es la suma de las y's en el i-ésimo grupo y  $\bar{p}_i$  es el promedio de las proporciones estimadas  $\hat{p}_i$  del evento que está siendo considerado en el i-ésimo grupo.

Si C es mayor que  $\chi^2_{\alpha}$  con g-2 grados de libertad entonces se concluye que el modelo logístico no es adecuado.

### 5.2.3 Estadísticas Influenciales para regresión logística

Existen varios tipos de residuales que permiten cotejar si una observación es influyente o no para la regression logística.

a) **Residuales de Pearson:** Están definidos por

$$r_i = \frac{y_i - m_i \hat{p}_i}{\sqrt{m_i \hat{p}_i (1 - \hat{p}_i)}}$$

donde, si los valores de la variable de respuesta están agrupadas,  $y_i$  representa el número de veces que  $y=1$  entre las  $m_i$  repeticiones de  $X_i$ . Si los datos no están agrupados  $m_i=1$  para todo i.

El residual de Pearson es similar al residual estudentizado usado en regresión lineal. Así un residual de Pearson en valor absoluto mayor que 2 indica un dato atípico.

b) **Residuales de Devianza:** Están dados por

$$D_i = -\sqrt{2 |\log(1 - \hat{p}_i)|} \text{ si } y_i=0 \text{ y por } D_i = \sqrt{2 |\log(\hat{p}_i)|} \text{ si } y_i=1.$$

Si el residual de devianza es mayor que 2 en valor absoluto entonces la observación correspondiente es atípico. Estos son los residuales dados por R.

**Ejemplo 4.** Aplicar modelos de regresión logística a los datos del ejemplo 2 y basados en la medidas de bondad de ajuste seleccionar el mejor modelos entre ellos.

```
> # Haciendo la regresion logistica simple con la predictora pesomama
> logis1<-glm(bajopeso~pesomama,data=pesobebe,family=binomial)
> summary(logis1)
```

Call:

```
glm(formula = bajopeso ~ pesomama, family = binomial, data = pesobebe)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0951	-0.9022	-0.8018	1.3609	1.9821

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.99831	0.78529	1.271	0.2036
pesomama	-0.01406	0.00617	-2.279	0.0227 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom  
 Residual deviance: 228.69 on 187 degrees of freedom  
 AIC: 232.69

Number of Fisher Scoring iterations: 4

```
> #Haciendo la regresion logistica multiple usando todas las variables predictoras
> logis2<-glm(bajopeso~.,data=pesobebe,family=binomial)
> summary(logis2)
```

Call:

```
glm(formula = bajopeso ~ ., family = binomial, data = pesobebe)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8832	-0.8178	-0.5574	1.0288	2.1451

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.078975	1.276254	-0.062	0.95066
edad	-0.035845	0.036472	-0.983	0.32569
pesomama	-0.012387	0.006614	-1.873	0.06111 .
raza	0.453424	0.215294	2.106	0.03520 *
fuma	0.937275	0.398458	2.352	0.01866 *
prematuros	0.542087	0.346168	1.566	0.11736
hipertensio	1.830720	0.694135	2.637	0.00835 **
uteroirrit	0.721965	0.463174	1.559	0.11906
chequeos	0.063461	0.169765	0.374	0.70854

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom  
Residual deviance: 204.19 on 180 degrees of freedom  
AIC: 222.19

Number of Fisher Scoring iterations: 4

```
> #Haciendo otra vez la regresion logistica incluyendo solo las variables mas significativas
> logis3<-glm(bajopeso~pesomama+raza+fuma+hipertensio,data=pesobebe,family=binomial)
> summary(logis3)
```

Call:

```
glm(formula = bajopeso ~ pesomama + raza + fuma + hipertensio,
     family = binomial, data = pesobebe)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7988	-0.8865	-0.5847	1.0997	2.2503

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-0.357536	1.010584	-0.354	0.72350
pesomama	-0.015354	0.006523	-2.354	0.01858 *
raza	0.489555	0.207324	2.361	0.01821 *
fuma	1.080020	0.383735	2.814	0.00489 **
hipertensio	1.744272	0.687563	2.537	0.01118 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom  
Residual deviance: 211.04 on 184 degrees of freedom  
AIC: 221.04.

Number of Fisher Scoring iterations: 4

Observando el valor de la devianza residual y del AIC el tercer modelo seria el mejor modelo. Notar que la Devianza Residual=211.04 el cual habría que compararlo con una Ji-Cuadrado con 184 grados de libertad para un nivel de significación dado. Usando un nivel de significación del 5%, la Ji-Cuadrado da 216.64 que es mayor que la Devianza Residual. En consecuencia, los datos parecen ajustarse a un modelo logístico. El pseudo-R<sup>2</sup> da .1007.

```
pihat=logis3$fit
pihatc=cut(pihat,br=c(0,quantile(pihat,p=seq(.1,.9,.1)),1))
table(pihatc)
pihatc=cut(pihat,br=c(0,quantile(pihat,p=seq(.1,.9,.1)),1),labels=F)
table(pihatc)
E=matrix(0,nrow=10,ncol=2)
O=matrix(0,nrow=10,ncol=2)
```

```

for(j in 1:10){
  E[j,2]=sum(pihat[pihatc==j])
  E[j,1]=sum((1-pihat)[pihatc==j])
  O[j,2]=sum(pesobebe$bajopeso[pihatc==j])
  O[j,1]=sum((1-pesobebe$bajopeso)[pihatc==j]) }
>sum((O-E)^2/E)
[1] 5.744907
> 1-pchisq(sum((O-E)^2/E),8)
[1] 0.6757812

```

Viendo el “p-value” de la prueba de  $\chi^2$ , se concluye que hay suficiente evidencia estadística para aceptar que el tercer modelo satisface el modelo logístico.

De acuerdo a los residuales de Pearson las siguientes observaciones pueden ser influenciales

```

> y=pesobebe$bajopeso
> pihat=logis3$fit
> rp=(y-pihat)/sqrt(pihat*(1-pihat))
> rp[abs(rp)>2]
      13      132      147      152      155      170      183
-2.010543 2.539516 3.402709 2.048327 2.700368 2.178067 2.345670

```

De acuerdo a los residuales de devianza las siguientes observaciones pueden ser influenciales

```

> r1=sqrt(2*abs(log(pihat[y==1])))
> r2=-sqrt(2*abs(log(1-pihat[y==0])))
> rd=c(r2,r1)
> rd[abs(rd)>2]
      132      147      155
2.004045 2.250326 2.056837

```

**Ejemplo 4.** Aplicar regresión logística a los datos de la tabla del ejemplo 3.

```

> horas
[1] 0 1 2 3 4 5 6 7
> # número de estudiantes que aprueban el examen por hora de estudio
> est.aprob=c( 1, 5, 25, 40, 45, 30, 17, 9)
> # número de estudiantes que fracasan el examen por hora de estudio
> est.frac=c(19,45,55,80,45,10,3,1)
> logis4=glm(cbind(est.aprob,est.frac)~horas,family=binomial)
> summary(logis4)

```

Call:

```
glm(formula = cbind(est.aprob, est.frac) ~ horas, family = binomial)
```

Deviance Residuals:

```

      Min       1Q   Median       3Q      Max
-0.8166 -0.6576 -0.1846  0.4510  1.7415

```

Coefficients:

```

Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.59099 0.29779 -8.701 < 2e-16 ***
horas 0.68489 0.08474 8.082 6.35e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 93.0406 on 7 degrees of freedom  
Residual deviance: 5.4156 on 6 degrees of freedom  
AIC: 38.111

Number of Fisher Scoring iterations: 4

Comparando la Devianza residual que es 5.41 con el valor de la Ji-Cuadrado al 5% de significación y 6 grados de libertad que resulta 12.59, se concluye de que hay evidencia de que los datos ajustan bien a un modelo logístico. El pseudo- $R^2$  resulta ser bastante bueno .9417.

#### 5.2.4 Uso de la regresión logística en Clasificación:

Para efectos de clasificación la manera más fácil de discriminar es considerar que si  $p > 0.5$  entonces la observación pertenece a la clase que uno está interesado. Pero algunas veces esto puede resultar injusto sobre todo si se conoce si una de las clases es menos frecuente que la otra.

Métodos alternos son:

- Plotear el porcentaje de observaciones que están en la clase de interés y que han sido correctamente clasificadas (**Sensitividad**) versus distintos niveles de probabilidad y el porcentaje de observaciones de la otra clase que han sido correctamente clasificadas (**especificidad**) versus los mismos niveles de probabilidad anteriormente usados, en la misma gráfica. La probabilidad que se usará para clasificar las observaciones se obtiene intersectando las dos curvas.
- Usar la curva ROC (receiver operating characteristic curva). En este caso se grafica la sensibilidad versus (1-especificidad)100%, y se coge como el p ideal aquel que está más cerca a la esquina superior izquierda, o sea al punto (0,100).

Ahora aplicaremos la regresión logística como un clasificador a los datos del ejemplo anterior. En lo que sigue vamos a considerar los resultados del segundo modelo obtenido en el ejemplo 3. Prediciendo las clases con el segundo modelo usando el método más simple es decir comparando el valor ajustado por la regresión logística con  $p=0.5$  y asignando la observación a la clase 1 se obtienen 52 de las 189 observaciones mal clasificadas lo cual representa una tasa de mala clasificación del 27.51%

Haciendo la clasificación con el método más complicado calculando la sensibilidad y especificidad se obtiene la siguiente tabla

Sensitividad	Especificidad	P	(1-especificidad)%
100.00	15.38	0.10	84.62
88.14	36.15	0.20	63.85
79.66	53.85	0.25	46.15

67.80	67.69	0.30	32.31
55.93	77.69	0.35	22.31
47.46	83.08	0.40	16.92
35.59	89.23	0.50	10.77
18.64	96.92	0.60	3.08
13.56	98.46	0.70	1.54
1.69	99.23	0.80	0.77
0.00	100.00	0.90	0.00

Notar que para  $p=0.30$  la curva está más cerca a la esquina superior izquierda. La tasa de mala clasificación óptima es  $= 0.3227513$

Las gráficas de los dos métodos aparecen en las figuras 5.5 y 5.6 respectivamente y en ambos casos el  $p$ -óptimo a usarse es  $p=0.3$

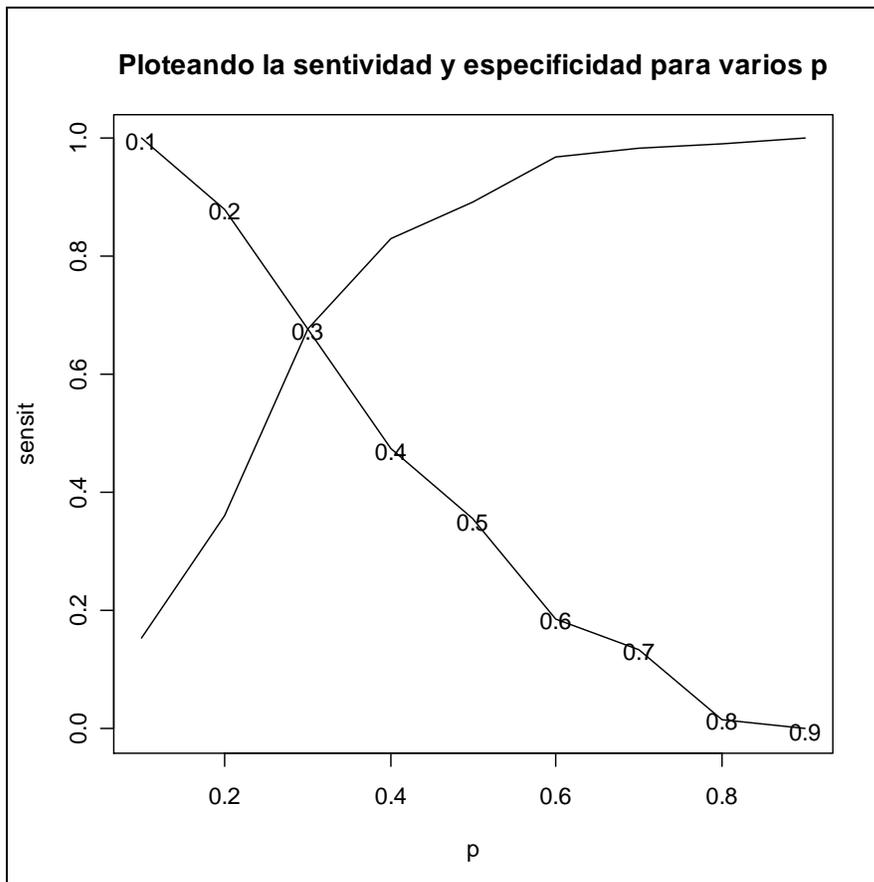


Figura 5.5. Determinación del  $p$  óptimo usando el método a

La regresión logística se puede extender al caso donde hay más de dos clases y recibe el nombre de regresión logística politémica. Este tipo de regresión es estudiada más detalladamente en un curso de clasificación. También existe una relación entre regresión logística y redes neuronales. El comando **multinom** de la librería nnet de R lleva a cabo regresión logística politémica.

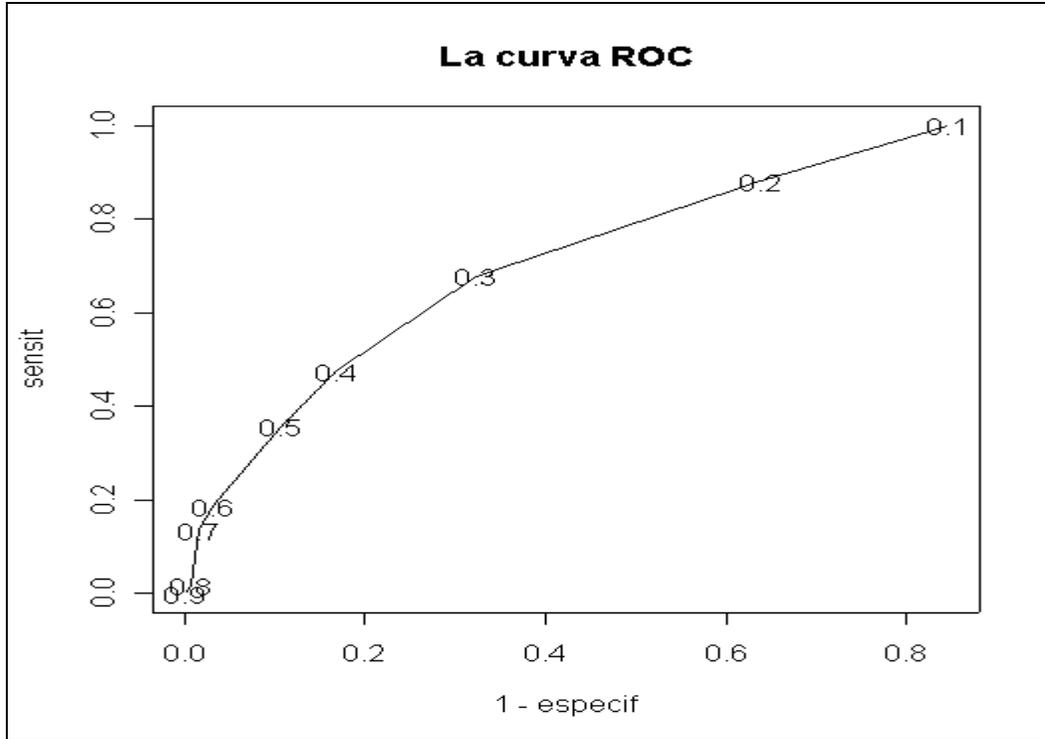


Figura 5.6. Determinación del p óptimo usando la curva ROC

## Ejercicios

**1. Comparando líneas de regresión.** Considerar el conjunto de datos bajopeso disponible en la página de internet del texto y tomar al peso del bebé como  $Y$  y a peso de la mamá como  $X$ . Comparar las pendientes y los interceptos de la línea de regresión en los tres grupos de raza de la madre.

**2.** Considerar el conjunto de datos **heartc** disponible en [math.uprm.edu/~edgar/datosclass.html](http://math.uprm.edu/~edgar/datosclass.html) en el cual se toman 13 mediciones a 297 pacientes para clasificarlos en propensos o no propensos a sufrir ataque cardíaco. Las clases están en la última columna y están codificadas como 1 y 2.

- Usar los criterios de la Devianza y del AIC para determinar un modelo de logístico óptimo
- Determinar la bondad de ajuste del modelo
- Identificar posibles valores influyentes
- Determinar la tasa de mala clasificación según las distintas maneras consideradas en el texto.

**3.** Considerar el conjunto de datos **breastw** disponible en [math.uprm.edu/~edgar/datosclass.html](http://math.uprm.edu/~edgar/datosclass.html) en el cual se toman 9 mediciones a 699 mujeres para clasificarlas en propensas o no propensas a tener cáncer al seno. Las clases están en la última columna y están codificadas como 1 y 2.

- Usar los criterios de la Devianza y del AIC para determinar un modelo de logístico óptimo
- Determinar la bondad de ajuste del modelo
- Identificar posibles valores influyentes
- Determinar la tasa de mala clasificación según las distintas maneras consideradas en el texto.

**4. Regresión logística con datos agrupados.** En una Universidad se registra el número de estudiantes que pasaron con A de acuerdo a las veces que habían tomado de antemano un curso de estadística. Los resultados se muestran en la siguiente tabla

Veces que había tomado el curso anteriormente	Número de estudiantes	Estudiantes pasando con A
0	300	30
1	150	25
2	80	20
3	35	5
4	20	3
5	5	1

- Construir una regresión logística para predecir la probabilidad de que un estudiante obtenga A en la clase de acuerdo a las veces que la ha tomado antes.
- Probar si la variable predictora: número de veces que el estudiante ha tomado antes el curso es significativa o no?
- Determinar la bondad de ajuste del modelo.

## CAPÍTULO 6

### SELECCIÓN DE VARIABLES EN REGRESIÓN

Selección de variables o también llamado selección de un subconjunto de predictoras es un procedimiento estadístico que es importante por diversas razones, entre estas están:

- a) No todas las variables predictoras tienen igual importancia, por lo tanto es más eficiente trabajar con un modelo donde las variables importantes estén presentes y las que tienen poca importancia no aparezcan.
- b) Algunas variables pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras, luego se hace necesario eliminarlas ya que son redundantes.
- c) Computacionalmente es más fácil trabajar con un conjunto de variables predictoras pequeño.
- d) Es más económico recolectar información para un modelo con pocas variables.
- e) Si se reduce el número de variables entonces el modelo se hace más **parsimonioso**. Se dice que un modelo es **parsimonioso** si consigue ajustar bien los datos pero usando la menor cantidad de variables predictoras posibles. Es más conveniente porque sus predicciones son más confiables y además es más robusto que el modelo original.

Desde que empezó a trabajarse en esta área en los años 60 y gracias al desarrollo de las computadoras se han introducido muchos métodos de selección de variables. Aquí describiremos sólo algunos de ellos.

#### 6.1 Metodos “Stepwise”

La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo (o excluyendo) una sola variable predictora en cada paso de acuerdo a ciertos criterios. El proceso secuencial termina cuando una regla de parada se satisface.

Hay tres algoritmos más comúnmente usados, los cuales serán descritos a continuación.

##### 6.1.1 “Backward Elimination” (Eliminación hacia atrás).

En este caso se comienza con el modelo completo y en cada paso se va eliminando una variable. Si resultara que todas las variables predictoras son importantes, es decir tienen “p-value” pequeños para la prueba t, entonces no se hace nada y el mejor modelo es el que tiene todas las variables predictoras disponibles. En caso contrario, en cada paso la variable que se elimina del modelo es aquella que satisface cualquiera de estos requisitos equivalentes entre sí:

- a) Aquella variable que tiene el estadístico de t, en valor absoluto, más pequeño entre las variables incluidas aún en el modelo. Es decir, aquella variable con el F parcial más pequeño. El F parcial está definido por:

$$F_p = [SSR_k - SSR_{k-1}] / MSE_k$$

donde  $SSR_k$  es la suma de cuadrados debido a la regresión con k variables y  $SSR_{k-1}$  es la misma suma con k-1 variables.  $MSE_k = SSE_k / (n - k - 1)$  es el cuadrado medio del error del modelo que incluye k variables. Hay que calcular el  $F_p$  para cada una de las variables presentes aún en

el modelo y se elimina del modelo aquella variable que da el  $F_p$  mas pequeño. Se puede mostrar que  $t^2 = F_p$ . En realidad todo el proceso se entiende mucho mejor con la  $t$  que con la  $F$ .

- b) Aquella variable que produce la menor disminución en el  $R^2$  al ser eliminada del modelo. Es decir, aquella variable que produce el mas pequeño incremento en la suma de cuadrados del error.
- c) Aquella variable que tiene la correlación parcial (en valor absoluto) más pequeña con la variable de respuesta, tomando en cuenta las variables que quedarían en el modelo. La correlación parcial de  $Y$  con la variable  $X_i$  se define como la correlación entre los residuales de la regresión de  $Y$  con todas las variables predictoras, excepto  $X_i$  y los residuales de la regresión de  $X_i$  con todas las otras restantes variables predictoras.

El método “Backward” padece del efecto de anidamiento ya que toda variable que es eliminada del modelo ya no vuelve a entrar a él.

El proceso termina cuando se cumple una de las siguientes condiciones:

- a) Se llega a un modelo con un número prefijado  $p^*$  de variables predictoras.
- b) El valor de la prueba de  $F$  parcial para todas las variables incluidas en el modelo son mayores que un número prefijado  $F$ -out (por lo general este valor es 4). O en forma equivalente, se para cuando el valor absoluto del estadístico de  $t$  para cada variable es mayor que la raíz cuadrada de  $F$ -out (por lo general,  $|t| > 2$ ). Algunas veces se prefija de antemano un nivel de significación dado  $\alpha^*$  (digamos del 10%) para la prueba de  $t$  o de  $F$  parcial en cada paso y en este caso se termina el proceso cuando todos los  $p$ -values son menores que  $\alpha^*$ .

### 6.1.2 “Forward Selection” (Selección hacia adelante).

Aquí se empieza con la regresión lineal simple que considera como variable predictora a aquella que esta más altamente correlacionada (sin tomar en cuenta el signo) con la variable de respuesta. Si esta primera variable no es significativa entonces se considera el modelo  $\hat{Y} = \bar{Y}$  y se para el proceso, de lo contrario se sigue y en el siguiente paso se añade al modelo la variable que reúne cualquiera de estos requisitos equivalentes entre sí:

- a) Aquella variable que tiene el estadístico de  $t$ , en valor absolute, más grande entre las variables no incluidas aún en el modelo. Es decir, la variable con el  $F$ -parcial más grande.
- b) Aquella variable que produce el mayor incremento en el  $R^2$  al ser añadida al modelo. Es decir, aquella variable que produce la mayor reducción en la suma de cuadrados del error.
- c) Aquella variable que tiene la correlación parcial más alta (en valor absoluto) con la variable de respuesta, tomando en cuenta las variables ya incluidas en el modelo.

Aquí también está presente el efecto de anidamiento ya que toda variable que es añadida al modelo ya no puede salir del mismo.

El proceso termina cuando se cumple una de las siguientes condiciones:

- a) Se llega a un modelo con un número prefijado  $p^*$  de variables predictoras.
- b) El valor de la prueba de  $F$  parcial para cada una de las variables no incluidas aun en el modelo es menor que un número prefijado  $F$ -in (por lo general este valor es 4). O en forma equivalente se para cuando el valor absoluto del estadístico de  $t$  es menor que la raíz cuadrada de  $F$ -in (por lo general,  $|t| < 2$ ). Algunas veces se prefija de antemano un nivel de significación

dado  $\alpha^*$  (digamos del 15%) para la prueba de t o de F parcial en cada paso y en este caso se termina el proceso cuando todos los p-values de la prueba t de las variables no incluidas aún son mayores que  $\alpha^*$ .

### 6.1.3 “Stepwise Selección” (Selección Paso a Paso)

Fue introducido por Efroymson (1960) para subsanar el problema de anidamiento de los dos métodos anteriores. Se puede considerar como una modificación del método “Forward”. Es decir, se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable en forma similar al método forward, pero se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada. Aquí se usan F-out y F-in con  $F\text{-in} \leq F\text{-out}$ .

El proceso termina cuando ninguna de las variables, que no han entrado aún, tienen importancia suficiente como para entrar al modelo.

Prácticamente todos los programas estadísticos ejecutan los procedimientos “stepwise”. En R se puede usar la función `regsubsets` de la librería **leaps**. En S-Plus existe la función **stepwise** que tiene la opción **method**, la cual permite elegir entre el método backward, forward y stepwise (Efroymson). En MINITAB se sigue la secuencia **STAT** ▶ **Regression** ▶ **Stepwise**.

**Ejemplo 1:** Aplicar los métodos “stepwise” al conjunto de datos **grasa**. La variable de respuesta grasa: porcentaje de grasa en el cuerpo. En 252 sujetos se midieron las siguientes variables predictoras:

- edad ( en años)
- peso ( en libras)
- altura (en pulgadas)
- cuello (circunferencia en cms)
- pecho (circunferencia en cms)
- abdomen (circunferencia en cms)
- cadera (circunferencia en cms)
- muslo (circunferencia en cms)
- rodilla (circunferencia en cms)
- tobillo (circunferencia en cms)
- biceps (circunferencia en cms)
- antebrazo (circunferencia en cms)
- muneca (circunferencia en cms)

para predecir su porcentaje de grasa en el cuerpo

A continuación se muestra la salida en el R para el método de selección hacia adelante. Hemos creado dos funciones **selforw** y **backelim** que ejecutan el método de selección hacia adelante y el método de eliminación hacia atrás, respectivamente. Las funciones están disponibles en la página de internet del texto. Para el primer método hemos usado un nivel de significación del 15 por ciento y para el segundo un nivel de significación del 10 por ciento.

```
> grasa=read.table("http://math.uprm.edu/~edgar/grasa.txt",header=T)
> selforw(grasa[,2:14],grasa[,1],.15)
Loading required package: leaps
```

## Selección Forward

p=numero de coeficientes en el modelo, p=1 es por el intercepto

nvar=p-1=numero de variables predictoras

add.var=la variable que ha sido añadida al modelo actual

pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

p	nvar	add.var	pvmax	s	r2	r2adj	Cp
2	1	abdomen	0.0000	4.877	0.662	0.660	72.869
3	2	peso	0.0000	4.456	0.719	0.717	20.691
4	3	muneca	0.0047	4.393	0.728	0.724	14.210
5	4	antebrazo	0.0098	4.343	0.735	0.731	9.314
6	5	cuello	0.1000	4.328	0.738	0.733	8.559
7	6	edad	0.1098	4.314	0.741	0.734	7.973
8	7	muslo	0.1098	4.291	0.744	0.737	6.338

La variable mas importante, seleccionada en el primer paso es abdomen, la cual da un valor del coeficiente de determinación del 66.2% al hacer la regresión simple con la variable de respuesta grasa. Después son seleccionadas en estricto orden, peso, muneca, antebrazo, cuello, edad y muslo. Las restantes variables no son escogidas porque sus “p-values” correspondientes a la prueba de F-parcial son mayores del 15% que se había elegido de antemano. El coeficiente de determinación,  $R^2$ , del modelo de regresión múltiple incluyendo las siete variables predictoras es del 74.4%. Notar también que la desviación estándar estimada del error va disminuyendo a medida que se añaden mas variables en el modelo. Las otras estadísticas  $R^2$  ajustado y el Cp de Mallows serán explicadas mas adelante.

La siguiente tabla lista los valores de t y los p-values correspondientes para cada una de las variables no incluidas en el modelo, cuando se hace la regresión considerando las variables ya incluidas y cada una de las que falta incluir.

	t-value	P-value
altura	-0.48	0.629
pecho	0.20	0.840
cadera	-1.41	0.159
rodilla	-0.01	0.991
tobillo	0.83	0.406
biceps	1.18	0.240

Notar que todas las variables no incluidas aún en el modelo tienen “P-value” grande, mayor de 0.15, lo cual indica que ellas son no significativas. En consecuencia, el proceso termina. El valor de F crítico es  $F(1,243,.15)=2.085$  y el de t crítico correspondiente es 1.44 y se puede ver que todos los t-values en valor absoluto son menores que 1.44.

El método de eliminación hacia atrás da los siguientes resultados:

```
> backelim(grasa[,2:14],grasa[,1],.10)
```

Eliminacion hacia atras

p=numero de coeficientes en el modelo

nvar=p-1=numero de variables predictoras

rem.var=la variable a ser removida, el modelo actual no incluye

esta variable

pvmin=pvalue de la F parcial correspondiente a la variable menos importante en cada paso

p	nvar	rem.var	pvmin	s	r2	r2adj	Cp
14	14	13	rodilla	0.9497	4.296	0.749	0.736 12.004
13	13	12	pecho	0.8045	4.288	0.749	0.737 10.065
12	12	11	altura	0.4928	4.283	0.748	0.738 8.533
11	11	10	tobillo	0.3957	4.281	0.748	0.738 7.250
10	10	9	biceps	0.2888	4.282	0.747	0.738 6.367
9	9	8	cadera	0.1594	4.291	0.744	0.737 6.338

La primera de las trece predictoras que es removida del modelo es rodilla, porque es la que da el “p-value” mas grande para la prueba de F-parcial. Luego se eliminan las variables: pecho, altura, tobillo, bíceps y cadera. Las restantes siete variables se quedan en el modelo, porque sus “p-values” deben ser menores que 0.10. Mas especificamente el valor crítico de F corresponde a una  $F(1, n-k-1, \alpha) = F(1, 244, .10) = 2.72$ , aqui  $k=7$  número de variables presentes en el modelo en el paso 7, y el correspondiente valor critico de t es 1.65. Notar que las variables que se quedan en el modelo son las mismas que son elegidas con el método de selección hacia adelante, pero que el  $R^2$  ha bajado muy poco. Observar también que la desviación estándar estimada del error inicialmente disminuye cuando se eliminan variables de modelo, pero al final comienza a crecer.

Si se escoge un nivel de significación del 5% entonces el proceso termina en 10 pasos y solo quedan cuatro variables en el modelo: peso, abdomen, antebrazo y muñeca.

Las salidas en S-Plus para el método de eliminación hacia atrás, método de selección hacia delante, y método “stepwise” respectivamente son como sigue:

```
> grasa.y<-grasa[,1]
> grasa.x<-grasa[,2:14]

> breg<-stepwise(grasa.x,grasa.y,method="back")
> breg
$rss:
[1] 4411.522 4412.655 4421.330 4434.613 4455.324 4491.849 4553.520 4619.874
4658.236 4786.054 4943.245 5947.463 17578.990

$size:
[1] 12 11 10 9 8 7 6 5 4 3 2 1 0

$which:
edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
12(- 9) T T T T T T T T F T T T T
11(- 5) T T T T F T T T F T T T T
10(- 3) T T F T F T T T F T T T T
9(-10) T T F T F T T T F F T T T
8(-11) T T F T F T T T F F F T T
7(- 7) T T F T F T F T F F F T T
6(- 4) T T F F F T F T F F F T T
5(- 8) T T F F F T F F F F F T T
4(- 1) F T F F F T F F F F F T T
3(-12) F T F F F T F F F F F F T
```

```

2(-13) F T F F F T F F F F F F F
1(- 2) F F F F F T F F F F F F F
0(- 6) F F F F F F F F F F F F F

```

\$f.stat:

```

[1] 0.003988103 0.061378633 0.471848985 0.723998939 1.130246734 1.992092066
3.350023992 3.570128594 4.042716405
[10] 6.777492645 8.145201631 50.584224182 488.928083209

```

\$method:

```
[1] "backward"
```

```
> freg<-stepwise(grasa.x,grasa.y,method="forw")
```

```
> freg
```

\$rss:

```

[1] 5947.463 4943.245 4786.054 4658.236 4607.169 4559.235 4491.849 4455.324 4434.613
4421.330 4412.655 4411.522 4411.448

```

\$size:

```
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13
```

\$which:

```

      edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
1(+ 6)  F  F  F  F  F  T  F  F  F  F  F  F  F
2(+ 2)  F  T  F  F  F  T  F  F  F  F  F  F  F
3(+13)  F  T  F  F  F  T  F  F  F  F  F  F  T
4(+12)  F  T  F  F  F  T  F  F  F  F  F  T  T
5(+ 4)  F  T  F  T  F  T  F  F  F  F  F  T  T
6(+ 1)  T  T  F  T  F  T  F  F  F  F  F  T  T
7(+ 8)  T  T  F  T  F  T  F  T  F  F  F  T  T
8(+ 7)  T  T  F  T  F  T  T  T  F  F  F  T  T
9(+11)  T  T  F  T  F  T  T  T  F  F  T  T  T
10(+10) T  T  F  T  F  T  T  T  F  T  T  T  T
11(+ 3)  T  T  T  T  F  T  T  T  F  T  T  T  T
12(+ 5)  T  T  T  T  T  T  T  T  F  T  T  T  T
13(+ 9)  T  T  T  T  T  T  T  T  T  T  T  T  T

```

\$f.stat:

```

[1] 488.928083209 50.584224182 8.145201631 6.777492645 2.726691325 2.575843689
3.660481076 1.992092066 1.130246734
[10] 0.723998939 0.471848985 0.061378633 0.003988103

```

\$method:

```
[1] "forward"
```

```
> stepreg<-stepwise(grasa.x,grasa.y,method="efroymsn")
```

```
> stepreg
```

\$rss:

```
[1] 5947.463 4943.245 4786.054 4658.236 4607.169 4559.235 4491.849
```

\$size:

```
[1] 1 2 3 4 5 6 7
```

```
$which:
```

```
edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps antebrazo muneca
1(+ 6) F F F F F T F F F F F F F
2(+ 2) F T F F F T F F F F F F F
3(+13) F T F F F T F F F F F F T
4(+12) F T F F F T F F F F F T T
5(+ 4) F T F T F T F F F F F T T
6(+ 1) T T F T F T F F F F F T T
7(+ 8) T T F T F T F T F F F T T
```

```
$f.stat:
```

```
[1] 488.928083 50.584224 8.145202 6.777493 2.726691 2.575844 3.660481
```

```
$method:
```

```
[1] "efroymsn"
```

Notar que S-Plus da toda la secuencia de como todas las variables son removidas en el método “backward” y de como son añadidas en el método “forward” pero no selecciona las mejores variables. Sin embargo, en el método “stepwise” (Efroymsn) si se reportan las mejores variables.

En MINITAB, aplicaremos el método “stepwise”, con un nivel de significación del 15% para remover una variable y del 15% para que entre una variable. Este último porcentaje puede ser menor que el nivel de significación para remover variables. Los resultados son como siguen:

### Stepwise Regression: grasa versus edad, peso, ...

```
Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15
```

```
Response is grasa on 13 predictors, with N = 252
```

Step	1	2	3	4	5	6	7
Constant	-39.28	-45.95	-27.93	-34.85	-30.65	-25.89	-33.26
abdomen	0.631	0.990	0.975	0.996	1.008	0.945	0.918
T-Value	22.11	17.45	17.37	17.76	17.89	13.82	13.21
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
peso		-0.148	-0.114	-0.136	-0.123	-0.094	-0.119
T-Value		-7.11	-4.84	-5.48	-4.75	-2.98	-3.51
P-Value		0.000	0.000	0.000	0.000	0.003	0.001
muneca			-1.24	-1.51	-1.25	-1.59	-1.53
T-Value			-2.85	-3.40	-2.66	-3.09	-3.00
P-Value			0.005	0.001	0.008	0.002	0.003
antebraz				0.47	0.53	0.57	0.55
T-Value				2.60	2.86	3.08	2.99
P-Value				0.010	0.005	0.002	0.003
cuello					-0.37	-0.40	-0.40

T-Value					-1.65	-1.81	-1.83
P-Value					0.100	0.072	0.068
edad						0.046	0.068
T-Value						1.60	2.21
P-Value						0.110	0.028
muslo							0.22
T-Value							1.91
P-Value							0.057
S	4.88	4.46	4.39	4.34	4.33	4.31	4.29
R-Sq	66.17	71.88	72.77	73.50	73.79	74.06	74.45
R-Sq(adj)	66.03	71.65	72.44	73.07	73.26	73.43	73.71
C-p	72.9	20.7	14.2	9.3	8.6	8.0	6.3

Notar que el método “stepwise” produjo exactamente los mismos resultados que la selección hacia adelante. Usar un nivel de significación menor del 15% para que una variable entre al modelo trae como consecuencia elegir un modelo final mas pequeño.

## 6.2 Método de los mejores subconjuntos

Para problemas con un número pequeño de variables predictoras  $k$  (con  $k$  menor que 8), se podrían calcular uno o dos criterios de selección para las  $2^k$  regresiones posibles, luego se escogerían unos cuantos de estos modelos para un análisis más detallado y decidir sobre el mejor modelo. Lamentablemente hoy en día existen modelos con un gran número de variables predictoras, fácilmente se pueden encontrar problemas con más de 200 variables predictoras y ajustar  $2^{200}$  modelos sería un trabajo computacional bien pesado. Basándose en el algoritmo “Branch and Bound” (Ramificación y acotamiento) Hocking and Leslie (1967) propusieron un método para acelerar la búsqueda de los mejores subconjuntos de acuerdo a cierto criterio. Más tarde en 1974, Furnival and Wilson, propusieron un algoritmo llamado “Leaps and Bound” (Brincando y acotando) que permite elegir los mejores subconjuntos más eficientemente y este es el algoritmo adoptado por la mayoría de los programas estadísticos de computadoras.

## 6.3 Criterios para elegir el mejor modelo:

### 6.3.1 El coeficiente de Determinación $R^2$

La manera más básica de determinar el mejor modelo es eligiendo aquél que da un  $R^2$  bastante alto con el menor número de variables predictoras posibles. Aparte del efecto de datos anormales que pueden afectar este criterio, hay otro problema pues un modelo con pocas variables siempre tendrá un  $R^2$  menor o igual que un modelo que incluye un mayor número de variables, en consecuencia este criterio tendería a sugerirnos un modelo que contiene una buena cantidad de variables. Como una regla práctica se debería elegir un modelo con  $k$  variables si al incluir una variable adicional el  $R^2$  no se incrementa sustancialmente, algo como un 5%, en términos relativos.

### 6.3.2 El $R^2$ ajustado

Para subsanar la tendencia del  $R^2$  de elegir como mejor modelo aquel que tiene un gran número de variables predictoras, se ha definido un  $R^2$  ajustado de la siguiente manera:

$$R_{ajus}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{n-1}{n-p}(1-R^2) \quad (1)$$

donde  $p$  es el número de parámetros en el modelo. El  $R^2$  ajustado podría disminuir al incluirse una variable adicional en el modelo. Nuevamente, el modelo que se busca es aquel que tiene un  $R^2$ -ajustado alto con pocas variables.

### 6.3.3 La varianza estimada del error ( $s^2$ ).

El mejor modelo será aquel que tenga la varianza estimada (o desviación estándar) del error más pequeña.

### 6.3.4 El $C_p$ de Mallows.

La idea de este criterio, introducido por Mallows en 1973, es que el mejor modelo es aquel que no tiene ni mucha falta de ajuste (“underfitting”) ni mucho sobreajuste (“overfitting”) al ajustar los datos. Cuando hay falta de ajuste el estimado del valor predicho de la variable de respuesta tiene mucho sesgo y poca varianza, mientras que cuando hay “overfitting” la varianza del estimado del valor predicho es bastante alta, pero el sesgo es bajo. El cuadrado medio del error para un valor predicho sumando sobre todas las observaciones y considerando que hay  $p$  predictoras en el modelo, está dado por

$$\sum_{i=1}^n \frac{MSE(\hat{y}(x_i))}{\sigma^2} = \sum_{i=1}^n \frac{E[\hat{y}(x_i) - y(x_i)]^2}{\sigma^2} = \sum_{i=1}^n \frac{Var(\hat{y}(x_i)) + Sesgo^2(\hat{y}(x_i))}{\sigma^2} \quad (2)$$

Puede ser demostrado que

$$\sum_{i=1}^n \frac{Var(\hat{y}(x_i))}{\sigma^2} = p \quad (3)$$

y que

$$\sum_{i=1}^n \frac{Sesgo^2(\hat{y}(x_i))}{\sigma^2} = (n-p) \left( \frac{E(s_p^2) - \sigma^2}{\sigma^2} \right) \quad (4)$$

El criterio de Mallows trata de encontrar un modelo donde tanto el sesgo como la varianza sean moderados. El estimado del lado derecho de la ecuación (2) es llamado el estadístico de Mallows y está dado por

$$C_p = p + (n-p) \frac{s_p^2}{s^2} - (n-p) = \frac{SSE_p}{s^2} - (n-2p) \quad (5)$$

donde  $SSE_p$  es la suma de cuadrados del error del modelo que contiene  $p$  parámetros, incluyendo el intercepto, y  $s^2$  es la varianza estimada con el modelo completo. Si un modelo con  $p$  parámetros es adecuado entonces  $E(SSE_p) = (n-p)\sigma^2$ . Luego,  $E[SSE_p/s^2]$  es aproximadamente  $(n-p)\sigma^2/\sigma^2 = (n-p)$ . En consecuencia si el modelo fuera adecuado  $E(C_p) = p$ . Para decidir

acerca del valor de  $p$  se acostumbra a plotear  $C_p$  versus  $p$ . Los valores  $p$  más adecuados serán aquellos cercanos a la intersección de la gráfica con la línea  $C_p=p$

La library **leaps** de R selecciona los mejores subconjuntos usando los criterios de  $R^2$ ,  $R^2$  ajustado y el  $C_p$  de Mallows. Aquí solo mostramos los resultados para el criterio  $C_p$ .

**Ejemplo 2.** Elegir los mejores subconjuntos de variables predictoras para el conjunto de datos **grasa** usando los criterios anteriores.

```
># El numero maximo de variables a entrar sera igual al numero de
> # predictoras del conjunto original
> maxvar<-dim(grasa)[2]
> #Mejor modelo usando Cp de mallows
> bcp<-leaps(grasa.x,grasa.y,method="Cp",nbest=1,names=nombres)
> bcp
$which
  edad peso altura cuello pecho abdomen cadera muslo rodilla tobillo biceps
1 FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
2 FALSE TRUE  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
3 FALSE TRUE  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
4 FALSE TRUE  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
5 FALSE TRUE  FALSE TRUE  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
6  TRUE TRUE  FALSE FALSE FALSE  TRUE FALSE TRUE  FALSE FALSE FALSE
7  TRUE TRUE  FALSE TRUE  FALSE  TRUE FALSE TRUE  FALSE FALSE FALSE
8  TRUE TRUE  FALSE TRUE  FALSE  TRUE TRUE  TRUE  FALSE FALSE FALSE
9  TRUE TRUE  FALSE TRUE  FALSE  TRUE TRUE  TRUE  FALSE FALSE TRUE
10 TRUE TRUE  FALSE TRUE  FALSE  TRUE TRUE  TRUE  FALSE TRUE TRUE
11 TRUE TRUE  TRUE  TRUE  FALSE  TRUE TRUE  TRUE  FALSE TRUE TRUE
12 TRUE TRUE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  FALSE TRUE TRUE
13 TRUE TRUE  TRUE  TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE TRUE

  antebrazo muneca
1  FALSE FALSE
2  FALSE FALSE
3  FALSE TRUE
4  TRUE TRUE
5  TRUE TRUE
6  TRUE TRUE
7  TRUE TRUE
8  TRUE TRUE
9  TRUE TRUE
10 TRUE TRUE
11 TRUE TRUE
12 TRUE TRUE
13 TRUE TRUE

$label
[1] "(Intercept)" "edad"    "peso"    "altura"  "cuello"
[6] "pecho"        "abdomen" "cadera"  "muslo"   "rodilla"
[11] "tobillo"     "biceps"   "antebrazo" "muneca"

$size
[1] 2 3 4 5 6 7 8 9 10 11 12 13 14
```

```
$Cp
```

```
[1] 72.868837 20.690746 14.210205 9.314331 8.559272 7.664855 6.337654
```

```
[8] 6.367146 7.249744 8.533156 10.065111 12.003988 14.000000
```

```
> p<-2:maxvar
```

```
> plot(p,bcp$Cp,type="l")
```

```
> title("Grafica del Cp de Mallows segun el tamaño del modelo")
```

```
> lines(2:maxvar,2:maxvar)
```

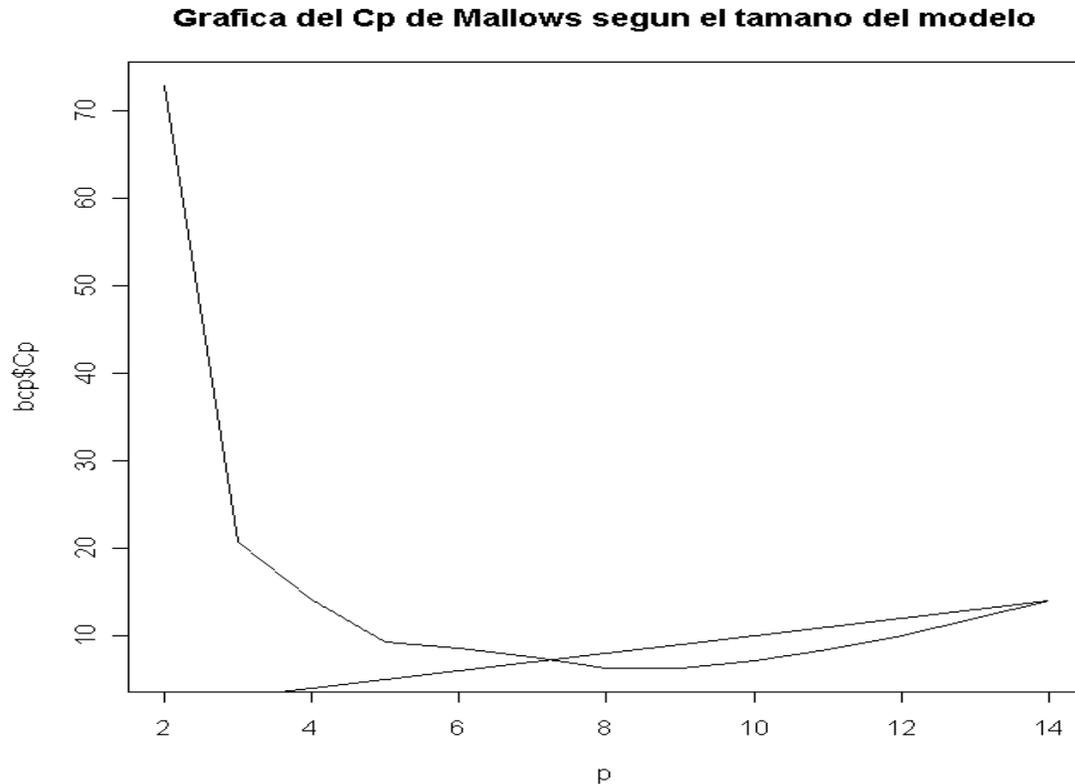


Figura 6.1 Plot del Cp de Mallows para determinar el número óptimo de variables predictoras.

Notar que la curva y la línea se intersectan alrededor de  $p=7$ .

MINITAB permite seleccionar los mejores subconjuntos basados en los criterios mencionados anteriormente. Se debe usar la secuencia **STAT** ▶ **Regression** ▶ **Best Subsets**.

### Best Subsets Regression: grasa versus edad, peso, ...

Response is grasa

```

a
  r t n
a c a c o o b t m
l u p d a m d b i e u
e p t e e o d u i i c b n

```

Vars	R-Sq	R-Sq(adj)	C-p	S	d e u l c m e s l l e r e a s r l h e r l l l p a c d o a o o n a o a o s z a																	
1	66.2	66.0	72.9	4.8775																		X
1	49.4	49.2	232.2	5.9668																		X
2	71.9	71.7	20.7	4.4556						X												X
2	70.2	70.0	36.6	4.5866																		X
3	72.8	72.4	14.2	4.3930						X												X
3	72.4	72.0	18.0	4.4251						X	X											X
4	73.5	73.1	9.3	4.3427						X												X X
4	73.3	72.8	11.4	4.3609						X												X X
5	73.8	73.3	8.6	4.3276						X	X											X X
5	73.7	73.2	9.2	4.3336						X	X											X X
6	74.1	73.5	7.7	4.3111						X	X											X X
6	74.1	73.4	8.0	4.3138						X	X	X										X X
7	74.4	73.7	6.3	4.2906						X	X	X										X X
7	74.3	73.6	7.4	4.2998						X	X	X										X X X
8	74.7	73.8	6.4	4.2819						X	X	X										X X
8	74.6	73.8	7.0	4.2872						X	X	X										X X X
9	74.8	73.8	7.2	4.2808						X	X	X										X X X
9	74.7	73.8	7.7	4.2851						X	X	X	X									X X
10	74.8	73.8	8.5	4.2832						X	X	X	X									X X X X
10	74.8	73.8	8.7	4.2850						X	X	X	X									X X X
11	74.9	73.7	10.1	4.2879						X	X	X	X									X X X X
11	74.8	73.7	10.5	4.2920						X	X	X	X	X								X X X X
12	74.9	73.6	12.0	4.2963						X	X	X	X	X	X							X X X X
12	74.9	73.6	12.1	4.2968						X	X	X	X	X	X	X						X X X X
13	74.9	73.5	14.0	4.3053						X	X	X	X	X	X	X	X					X X X X

Lo que se muestra aquí son los dos mejores subconjuntos de variables para cada número de variables predictoras, excepto cuando se tiene el modelo que incluye todas las variables. De acuerdo al  $R^2$  y  $R^2$  ajustado el mejor modelo sería aquel que incluye solo dos variables predictoras: peso y abdomen. De acuerdo al  $C_p$  de Mallows se escogería el modelo que incluye 6 variables predictoras: Edad, peso, abdomen, muslo, antebrazo y muñeca. El  $C_p$  es de 7.7.

De acuerdo a la varianza estimada del error se escogería el modelo que incluye 4 variables predictoras: peso, abdomen, antebrazo y muñeca.

**6.3.5 PRESS ( Suma de cuadrados de Predicción)**

El criterio suma de cuadrados de Predicción [PRESS], introducido por Allen en 1974, es una combinación de todas las regresiones posibles, análisis de residuales y “leave-one-out” (ver más adelante validación cruzada).

Supongamos que hay  $p$  parámetros en el modelo y que tenemos  $n$  observaciones disponibles para estimar los parámetros. En cada paso se deja de lado la  $i$ -ésima observacion del conjunto de datos y se calculan todas las regresiones posibles (más eficientemente se podrían calcular solamente los mejores subconjuntos de regresión que resultan de aplicar algún criterio, tal como el  $C_p$  de Mallows). Luego se calcula la predicción  $\hat{y}_{(i)}$  para la observación que no fue incluida y se calcula el residual correpondiente  $e_{(i)} = y_i - \hat{y}_{(i)}$ , el cual es llamado el residual PRESS. Ya se vió en la sección 3.1.4 que la relación entre el residual PRESS y el residual usual  $\hat{e}_i$  es

$$e_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \tag{6}$$

donde los  $h_{ii}$  representan los elementos de la diagonal de la matriz  $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ .

Si la diferencia entre el residual PRESS y el residual usual de una observación es bien grande entonces se considera que dicha observación es influyente.

La medida PRESS para el modelo de regresión que contiene  $p$  parámetros se define por:

$$PRESS = \sum_{i=1}^n e_{(i)}^2 \quad (7)$$

Otra forma, equivalente de cálculo sería

$$PRESS = \sum_{i=1}^n \left( \frac{\hat{e}_i}{1 - h_{ii}} \right)^2 \quad (8)$$

Según el criterio PRESS el mejor modelo será aquel que tenga el valor de PRESS más bajo. Algunas veces se acostumbra usar un PRESS promedio, que es simplemente el PRESS dividido por el número de observaciones del conjunto de datos. No existe una función en R que calcule el PRESS pero ésta puede ser programada fácilmente. MINITAB calcula el PRESS cuando se hace regresión y cuando se hace el “stepwise”.

**Ejemplo 3.** Calcular el PRESS para los mejores modelos según los criterios discutidos anteriormente del conjunto de datos grasa

```
> grasa=read.table("http://math.uprm.edu/~edgar/grasa.txt",header=T)
> PRESS=function(x)
{#x es un objeto que sale de aplicar lm
  sum(resid(x)^2/(1 - lm.influence(x)$hat)^2)
}
> lm1=lm(grasa~peso+abdomen, data=grasa)
> PRESS(lm1)
[1] 5109.1
> lm2=lm(grasa~peso+abdomen+antebrazo+muneca, data=grasa)
> PRESS(lm2)
[1] 4908.053
> lm3=lm(grasa~peso+abdomen+antebrazo+muneca+edad+muslo, data=grasa)
> PRESS(lm3)
[1] 4877.671
> lm4=lm(grasa~peso+abdomen+antebrazo+muneca+edad+muslo+cuello, data=grasa)
[1] 4840.64
```

Si buscamos un modelo parsimonioso, sería mejor elegir aquel que incluye 4 variables ya que si bien tiene un PRESS mayor que el de 6 variables, la diferencia no es mucha. Esta selección coincide con la que da el método de eliminación hacia atrás al nivel de significación del 5%.

### 6.3.6 Validación Cruzada (CV)

Fue introducido por Stone en 1974. La idea aquí es estimar el error de predicción dividiendo al azar el conjunto de datos en varias partes. En cada paso una de las partes se convierte en una muestra de prueba que sirve para validar el modelo y las restantes partes constituyen lo que es llamado una muestra de entrenamiento que sirve para construir el modelo. Por lo general se usan 10 partes y eso es llamado una “10 fold cross-validation”, ó  $n$  partes y en ese caso es llamado el método “leave-one-out”(dejar uno afuera). Este último se relaciona bastante con el PRESS. El cálculo del error por validación cruzada usando  $K$  partes estará dado por:

$$CV = \frac{\sum_{i=1}^K \sum_{j=1}^{N_i} (y_j - \hat{y}_j^{(-i)})^2}{n}$$

donde  $\hat{y}_j^{(-i)}$  representa el valor predicho para la  $j$ -ésima observación de la parte  $N_i$  usando una línea de regresión que ha sido estimada sin haber usado las observaciones de dicha parte. La idea es escoger el mejor modelo como aquel que tiene el más error de validación cruzada promedio más pequeño. En el caso de “:leave-one-out” el error de predicción promedio es  $PRESS/n$ .

El cálculo de validación cruzada para regresión no está disponible en ninguno de los programas estadísticos usados en este texto. Se debe tener que escribir un programa para obtenerlo, o usar solamente el método “leave-one-out”. Nosotros hemos escrito la función **CV10reg** que estima el error promedio de predicción usando validación cruzada 10.

#### Ejemplo 4. Aplicar la función CV10reg al conjunto grasa.

```
> #Leyendo el conjunto de datos pero excluyendo los nombres de las columnas
> grasa<-read.table(file="c:/grasa.txt",header=F,skip=1)
> dim(grasa)
[1] 252 14
> #Estimando el error promedio de prediccion con todas las predictoras
> CV10reg(grasa,10)
Los estimados del error promedio de prediccion en cada repeticion son
[1] 20.51001 19.91260 20.20943 19.94021 20.48995 20.54480 20.29945 20.40238
[9] 19.98185 20.60561
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 20.28963
> # Estimando el error promedio de prediccion usando: peso, abdomen
> CV10reg(grasa[,c(1,3,7)],10)
Los estimados del error promedio de prediccion en cada repeticion son
[1] 20.21294 20.15848 20.43202 20.43359 20.31295 20.27680 20.15327 20.26616
[9] 20.30443 20.24162
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 20.27923
> # Estimando el error promedio de prediccion usando: peso, abdomen, antebrazo y muneca.
> CV10reg(grasa[,c(1,3,7,13,14)],10)
Los estimados del error promedio de prediccion en cada repeticion son
[1] 19.26377 19.48586 19.63418 19.66251 19.63294 19.35995 19.40967 19.58761
[9] 20.51080 19.54337
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 19.60907
> # Estimando el error promedio de prediccion usando: edad, peso, abdomen, muslo, antebrazo y
muneca.
>CV10reg(grasa[,c(1:3,7,9,13,14)],10)
Los estimados del error promedio de prediccion en cada repeticion son
[1] 19.61473 19.29554 19.23239 19.69176 19.52733 19.36840 19.66784 19.21728
[9] 19.24110 19.15240
El estimado del error promedio de prediccion por VC con el numero de repeticiones dado es
[1] 19.40088
```

>

De los resultados previos el criterio CV nos sugiere que el mejor modelo es aquel que incluye las siguientes seis variables; edad, peso, abdomen, muslo, antebrazo y muñeca.

### 6.3.7 AIC

El criterio de información de Akaike (Akaike, 1973), tiene su origen en conceptos de teoría de información y está basado en la minimización de la distancia Kullback-Leibler entre la distribución de la variable de respuesta  $Y$  bajo el modelo reducido y bajo el modelo completo. Se define como,

$$\text{AIC} = -2 \cdot \text{máximo de la log likelihood} + 2p \quad (9)$$

Donde  $p$  es el número de parámetros dle modelo. En particular para el caso de regresión, asumiendo que la varianza de las  $y$ 's es estimada por  $\text{SSE}/n$ , la fórmula anterior se reduce a:

$$\text{AIC} = n \log[\text{SSE}_p/n] + 2p \quad (10)$$

Existen otras variantes a la formula (10). Un buen modelo es aquel con bajo AIC.

MINITAB no da el AIC, pero si aparece en SAS y S-Plus (aunque la versión que calculan es  $\text{AIC} = [\text{SSE}_p/s^2] + 2p$ . Tanto en R como en S-Plus están disponibles la funciones `step` y `stepAIC` (de la librería MASS) que calcula el mejor modelo por el método "stepwise" basado en el criterio AIC.

**Ejemplo 5.** Seleccionar el mejor modelo de regresión para el conjunto grasa usando el criterio AIC usando los métodos "forward" y "backward".

```
># Metodo "backward"
># Primero hay que hallar la regresion con todas las variables predictoras
> #Hallando el mejor subconjunto usando stepwise y el criterio AIC
> l1<-lm(grasa~.,data=grasa)
> step(l1,scope=~.,direction="backward")
Start: AIC= 749.36
grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +
      muslo + rodilla + tobillo + biceps + antebrazo + muneca
```

	Df	Sum of Sq	RSS	AIC
- rodilla	1	0.1	4411.5	747.4
- pecho	1	1.1	4412.5	747.4
- altura	1	9.7	4421.2	747.9
- tobillo	1	11.4	4422.9	748.0
- biceps	1	20.9	4432.3	748.5
<none>			4411.4	749.4
- cadera	1	37.5	4448.9	749.5
- muslo	1	49.6	4461.0	750.2
- peso	1	50.6	4462.1	750.2
- edad	1	68.3	4479.7	751.2
- cuello	1	76.0	4487.4	751.7
- antebrazo	1	95.5	4507.0	752.8

- muñeca 1 170.1 4581.6 756.9  
 - abdomen 1 2261.0 6672.4 851.6

Step: AIC= 747.36

grasa ~ edad + peso + altura + cuello + pecho + abdomen + cadera +  
 muslo + tobillo + biceps + antebrazo + muñeca

	Df	Sum of Sq	RSS	AIC
- pecho	1	1.1	4412.7	745.4
- altura	1	9.7	4421.2	745.9
- tobillo	1	12.1	4423.6	746.1
- biceps	1	20.8	4432.3	746.5
<none>			4411.5	747.4
- cadera	1	37.4	4448.9	747.5
- peso	1	53.1	4464.6	748.4
- muslo	1	54.9	4466.4	748.5
- edad	1	74.1	4485.6	749.6
- cuello	1	78.4	4490.0	749.8
- antebrazo	1	96.8	4508.3	750.8
- muñeca	1	170.5	4582.1	754.9
- abdomen	1	2269.9	6681.4	850.0

Step: AIC= 745.43

grasa ~ edad + peso + altura + cuello + abdomen + cadera + muslo +  
 tobillo + biceps + antebrazo + muñeca

	Df	Sum of Sq	RSS	AIC
- altura	1	8.7	4421.3	743.9
- tobillo	1	12.4	4425.1	744.1
- biceps	1	20.1	4432.8	744.6
<none>			4412.7	745.4
- cadera	1	36.3	4449.0	745.5
- muslo	1	60.1	4472.7	746.8
- peso	1	70.8	4483.5	747.4
- edad	1	73.8	4486.5	747.6
- cuello	1	79.5	4492.1	747.9
- antebrazo	1	95.6	4508.3	748.8
- muñeca	1	170.0	4582.6	753.0
- abdomen	1	2879.4	7292.1	870.0

Step: AIC= 743.92

grasa ~ edad + peso + cuello + abdomen + cadera + muslo + tobillo +  
 biceps + antebrazo + muñeca

	Df	Sum of Sq	RSS	AIC
- tobillo	1	13.3	4434.6	742.7
- biceps	1	22.4	4443.7	743.2
- cadera	1	30.4	4451.8	743.6
<none>			4421.3	743.9
- muslo	1	68.8	4490.1	745.8
- cuello	1	77.1	4498.4	746.3

```
- edad 1 81.3 4502.6 746.5
- antebrazo 1 98.1 4519.4 747.5
- peso 1 119.6 4540.9 748.6
- muñeca 1 181.3 4602.6 752.0
- abdomen 1 3178.5 7599.9 878.4
```

Step: AIC= 742.68

grasa ~ edad + peso + cuello + abdomen + cadera + muslo + biceps +  
antebrazo + muñeca

	Df	Sum of Sq	RSS	AIC
- biceps	1	20.7	4455.3	741.9
- cadera	1	31.7	4466.4	742.5
<none>		4434.6	742.7	
- muslo	1	72.3	4506.9	744.8
- edad	1	77.6	4512.2	745.1
- cuello	1	87.3	4521.9	745.6
- antebrazo	1	97.4	4532.0	746.2
- peso	1	107.2	4541.8	746.7
- muñeca	1	168.0	4602.6	750.0
- abdomen	1	3182.0	7616.7	877.0

Step: AIC= 741.85

grasa ~ edad + peso + cuello + abdomen + cadera + muslo + antebrazo +  
muñeca

	Df	Sum of Sq	RSS	AIC
<none>		4455.3	741.9	
- cadera	1	36.5	4491.8	741.9
- cuello	1	79.1	4534.4	744.3
- edad	1	83.8	4539.1	744.5
- peso	1	93.0	4548.3	745.1
- muslo	1	100.7	4556.0	745.5
- antebrazo	1	140.5	4595.8	747.7
- muñeca	1	166.8	4622.2	749.1
- abdomen	1	3163.0	7618.3	875.0

Call:

lm(formula = grasa ~ edad + peso + cuello + abdomen + cadera + muslo + antebrazo +  
muñeca, data = grasa)

Coefficients:

(Intercept)	edad	peso	cuello	abdomen	cadera	muslo
-22.65637	0.06578	-0.08985	-0.46656	0.94482	-0.19543	0.30239
	antebrazo	muñeca				
	0.51572	-1.53665				

Las variables eliminadas son: rodilla, pecho, altura, tobillo, y bíceps en ese orden. El mejor modelo según el método “backward” y usando el criterio AIC es el considera las siguientes 8 variables predictoras: edad, peso, cuello, abdomen , cadera, muslo, antebrazo, y muñeca.

```
>#Metodo "Forward"
> #Hallando primero la regresion con la variable predictor a mas correlacionada V7
> l2=lm(grasa~abdomen,data=grasa)
>
step(l2,scope=~.+edad+peso+altura+cuello+pecho+cadera+muslo+rodilla+tobillo+biceps+antebr
azo+muneca,direction="forward")
Start: AIC= 800.65
grasa ~ abdomen
```

	Df	Sum of Sq	RSS	AIC
+ peso	1	1004.2	4943.2	756.0
+ muneca	1	709.2	5238.3	770.6
+ cuello	1	614.5	5332.9	775.2
+ cadera	1	548.2	5399.2	778.3
+ altura	1	458.8	5488.7	782.4
+ rodilla	1	318.7	5628.8	788.8
+ tobillo	1	233.3	5714.1	792.6
+ edad	1	200.9	5746.5	794.0
+ pecho	1	195.5	5752.0	794.2
+ muslo	1	174.6	5772.9	795.1
+ biceps	1	135.3	5812.2	796.8
+ antebrazo	1	54.3	5893.2	800.3
<none>			5947.5	800.6

```
Step: AIC= 756.04
grasa ~ abdomen + peso
```

	Df	Sum of Sq	RSS	AIC
+ muneca	1	157.2	4786.1	749.9
+ cuello	1	86.9	4856.3	753.6
+ muslo	1	81.4	4861.9	753.9
+ antebrazo	1	66.9	4876.4	754.6
+ biceps	1	63.8	4879.4	754.8
+ altura	1	40.3	4903.0	756.0
<none>			4943.2	756.0
+ rodilla	1	9.7	4933.5	757.5
+ edad	1	1.9	4941.3	757.9
+ tobillo	1	1.5	4941.7	758.0
+ pecho	1	0.01017	4943.2	758.0
+ cadera	1	0.00529	4943.2	758.0

```
Step: AIC= 749.9
grasa ~ abdomen + peso + muneca
```

	Df	Sum of Sq	RSS	AIC
+ antebrazo	1	127.8	4658.2	745.1
+ biceps	1	88.7	4697.3	747.2
+ muslo	1	40.5	4745.6	749.8
<none>			4786.1	749.9
+ cuello	1	25.2	4760.9	750.6
+ altura	1	23.4	4762.6	750.7

+ edad	1	21.2	4764.9	750.8
+ rodilla	1	20.5	4765.5	750.8
+ tobillo	1	15.0	4771.1	751.1
+ cadera	1	9.2	4776.8	751.4
+ pecho	1	1.3	4784.8	751.8

Step: AIC= 745.07

grasa ~ abdomen + peso + muñeca + antebrazo

	Df	Sum of Sq	RSS	AIC
+ cuello	1	51.1	4607.2	744.3
+ edad	1	38.4	4619.9	745.0
<none>		4658.2	745.1	
+ biceps	1	33.9	4624.4	745.2
+ muslo	1	27.2	4631.0	745.6
+ rodilla	1	19.8	4638.4	746.0
+ tobillo	1	18.2	4640.1	746.1
+ altura	1	18.0	4640.2	746.1
+ cadera	1	3.5	4654.7	746.9
+ pecho	1	0.5	4657.7	747.0

Step: AIC= 744.3

grasa ~ abdomen + peso + muñeca + antebrazo + cuello

	Df	Sum of Sq	RSS	AIC
+ edad	1	47.9	4559.2	743.7
+ biceps	1	45.9	4561.2	743.8
<none>		4607.2	744.3	
+ muslo	1	25.1	4582.1	744.9
+ altura	1	18.9	4588.3	745.3
+ cadera	1	11.0	4596.2	745.7
+ tobillo	1	10.7	4596.5	745.7
+ rodilla	1	10.4	4596.8	745.7
+ pecho	1	0.009572	4607.2	746.3

Step: AIC= 743.66

grasa ~ abdomen + peso + muñeca + antebrazo + cuello + edad

	Df	Sum of Sq	RSS	AIC
+ muslo	1	67.4	4491.8	741.9
+ biceps	1	48.1	4511.1	743.0
<none>		4559.2	743.7	
+ altura	1	19.0	4540.3	744.6
+ tobillo	1	14.8	4544.5	744.8
+ rodilla	1	6.6	4552.7	745.3
+ cadera	1	3.2	4556.0	745.5
+ pecho	1	0.8	4558.4	745.6

Step: AIC= 741.91

grasa ~ abdomen + peso + muñeca + antebrazo + cuello + edad +  
muslo

	Df	Sum of Sq	RSS	AIC
+ cadera	1	36.5	4455.3	741.9
<none>			4491.8	741.9
+ biceps	1	25.5	4466.4	742.5
+ tobillo	1	12.8	4479.1	743.2
+ altura	1	4.3	4487.5	743.7
+ pecho	1	0.8	4491.1	743.9
+ rodilla	1	0.002584	4491.8	743.9

Step: AIC= 741.85

grasa ~ abdomen + peso + muñeca + antebrazo + cuello + edad +  
muslo + cadera

	Df	Sum of Sq	RSS	AIC
<none>			4455.3	741.9
+ biceps	1	20.7	4434.6	742.7
+ altura	1	11.7	4443.6	743.2
+ tobillo	1	11.6	4443.7	743.2
+ rodilla	1	3.651e-02	4455.3	743.8
+ pecho	1	9.904e-05	4455.3	743.9

Call:

lm(formula = grasa ~ abdomen + peso + muñeca + antebrazo + cuello + edad + muslo +  
cadera, data = grasa)

Coefficients:

(Intercept)	abdomen	peso	muñeca	antebrazo	cuello	edad
-22.65637	0.94482	-0.08985	-1.53665	0.51572	-0.46656	0.06578
	muslo	cadera				
0.30239	-0.19543					

Las variables que son seleccionadas en cada paso son: abdomen, peso, muñeca, antebrazo, cuello, edad, muslo y cadera en ese orden. El mejor modelo obtenido con el método “forward” es el mismo que se obtuvo con el método “backward”.

### 6.3.8 BIC

Este criterio fue introducido por Schwarz en 1978 y está basado en argumentos bayesianos. Se define por

$$\text{BIC} = n \log[\text{SSE}_p/n] + p \log(n)$$

Notar que el BIC se diferencia del AIC solo en el último término. Los criterios AIC y  $C_p$  de Mallows tienden a dar modelos óptimos más grandes que el criterio BIC. R y S-Plus dan una versión modificada del BIC. MINITAB no da el BIC, pero sí aparece en SAS (donde es llamado SBC).

**Ejemplo 6:** La siguiente salida en R muestra los valores de AIC y BIC, según las fórmulas dadas y siguiendo un método “forward”.

```
> forwabic(grasa[,2:14],grasa[,1],.15)
Selección Forward
```

p=numero de coeficientes en el modelo, p=1 es por el intercepto

nvar=p-1=numero de variables predictoras

add.var=la variable que ha sido añadida al modelo actual

pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

```
  p nvar add.var pvmax  aic  bic
2 2  1 abdomen 0.0000 800.645 807.704
3 3  2  peso  0.0000 756.040 766.628
4 4  3 muneca 0.0047 749.896 764.014
5 5  4 antebrazo 0.0098 745.075 762.722
6 6  5 cuello  0.1000 744.297 765.473
7 7  6 edad   0.1098 743.661 768.367
8 8  7 muslo  0.1098 741.909 770.144
>
```

El modelo elegido con el criterio AIC sería el mismo que fue elegido con el clásico método “forward”. El modelo elegido con el criterio BIC sería aquel que consider a las predictoras: abdomen, peso, muñeca y antebrazo.

**Ejemplo 7.** A continuación se muestran los 15 mejores modelos para el conjunto de datos **grasa**, ordenados de acuerdo al Cp de Mallows, mostrando además los valores del AIC, BIC y SBC. Los resultados fueron obtenidos usando SAS version 8.

Number in Model	C(p)	R-Square	AIC	BIC	SBC	Variables in Model
<b>**7</b>	<b>6.3377</b>	<b>0.7445</b>	<b>741.9088</b>	<b>744.5436</b>	<b>770.14425</b>	<b>edad peso cuello abdomen muslo antebraz muneca</b>
8	6.3671	0.7466	741.8514	744.7178	773.61622	edad peso cuello abdomen cadera muslo antebraz muneca
8	6.9626	0.7459	742.4748	745.2951	774.23968	edad peso cuello abdomen muslo biceps antebraz muneca
9	7.2497	0.7477	742.6772	745.7373	777.97144	edad peso cuello abdomen cadera muslo biceps antebraz muneca
7	7.3761	0.7434	742.9864	745.5508	771.22181	edad peso cuello abdomen biceps antebraz muneca
8	7.6488	0.7452	743.1915	745.9589	774.95637	edad peso cuello abdomen muslo tobillo antebraz muneca
<b>*6</b>	<b>7.6649</b>	<b>0.7410</b>	<b>743.3451</b>	<b>745.7048</b>	<b>768.05114</b>	<b>edad peso abdomen muslo antebraz muneca</b>
9	7.7333	0.7472	743.1859	746.2041	778.48021	edad peso altura cuello abdomen cadera muslo antebraz muneca
9	7.7402	0.7472	743.1933	746.2108	778.48755	edad peso cuello abdomen cadera muslo tobillo antebraz muneca
6	7.9732	0.7406	743.6612	746.0031	768.36724	edad peso cuello abdomen antebraz muneca
6	8.0815	0.7405	743.7721	746.1077	768.47815	peso cuello abdomen biceps

8	8.1043	0.7447	743.6660	746.3984	775.43091	antebraz muneca edad peso altura cuello abdomen muslo antebraz muneca
9	8.1748	0.7468	743.6497	746.6295	778.94397	edad peso cuello abdomen muslo tobillo biceps antebraz muneca
8	8.2969	0.7445	743.8664	746.5841	775.63130	edad peso cuello pecho abdomen muslo antebraz muneca
8	8.3375	0.7445	743.9087	746.6232	775.67353	edad peso cuello abdomen muslo rodilla antebraz muneca

(\*) sería el mejor modelo con el  $C_p$  de Mallows, porque aún cuando no es el modelo con el menor  $C_p$ , es un error bastante común elegir como el mejor modelo aquel con el menor  $C_p$ , su valor 7.66 está cerca a  $6+1=7$  (número de parámetros del modelo).

(\*\*) sería el mejor modelo de acuerdo al AIC, porque aunque es el modelo con el segundo AIC más pequeño, se está eligiendo solo 7 variables.

### 6.3.9 Validación Cruzada Generalizada (CGV)

Este criterio fue introducido en 1979, por Golub, Heath and Waha. El cálculo de validación cruzada “leave-one out” es computacionalmente pesado y el GCV es una aproximación al “leave-one-out”, que puede ser calculado más rápidamente.

Se define por

$$GCV = \frac{nSSE_p}{[n - tr(H_p)]^2} = \frac{nSSE_p}{(n - p)^2}$$

donde  $H_p$  es la matriz HAT para el modelo que incluye a  $p$  variables. Luego,  $tr(H_p)=p$ . El modelo óptimo será aquel que incluye las  $p$  variables predictoras que hacen que GCV sea mínimo. El cálculo del GCV puede ser fácilmente programable en R o S-Plus. SAS da esta medida pero en su procedimiento para Regresión Noparamétrica.

Los resultados de GCV para el ejemplo grasa son como siguen:

```
> forwabic(grasa[,2:14],grasa[,1],.15)
Loading required package: leaps
Seleccin Forward
```

$p$ =numero de coeficientes en el modelo,  $p=1$  es por el intercepto

$nvar=p-1$ =numero de variables predictoras

$add.var$ =la variable que ha sido anadida al modelo actual

$pvmax=p$ -value de F-parcial correspondiente a la variable mas importante en cada paso

```
  p nvar add.var pvmax   aic   bic   gcv
2 2  1 abdomen 0.0000 800.645 807.704 23.9802
```

```

3 3 2  peso  0.0000 756.040 766.628 20.0916
4 4 3  muñeca 0.0047 749.896 764.014 19.6099
5 5 4  antebrazo 0.0098 745.075 762.722 19.2410
6 6 5  cuello  0.1000 744.297 765.473 19.1851
7 7 6  edad   0.1098 743.661 768.367 19.1408
8 8 7  muslo  0.1098 741.909 770.144 19.0128
>

```

### 6.3.10 Otros Criterios

Recientemente se han introducido muchos otros criterios para selección de variables en regresión entre los más conocidos están:

**MDL:** Longitud de descripción Mínima (Rissanen, 1978).

**RIC:** Criterio de Inflación del Riesgo (Foster y George, 1994)

**CIC:** Criterio de Inflación del Covarianza (Tibshirani and Knight, 1999)

Para más detalles acerca estos métodos veáse el texto “Subset selection in regression” por Alan Miller (2002).

### 6.3.11 Recomendaciones para elegir el mejor modelo

En cualquier problema las variables predictoras pueden ser clasificadas en 3 grupos:

- Las que son importantes.
- Las que uno no está seguro de su importancia. Es decir, podrían ser redundantes.
- Las que no son relevantes para explicar el comportamiento de la variable de respuesta.

Lo que se recomienda es eliminar las variables tipo c) eligiendo un buen subconjunto de variables predictoras usando para ello los criterios  $C_p$ , AIC o BIC y luego aplicar “stepwise” para descartar las variables tipo b) y quedarnos con las variables tipo a) que son las que nos interesan. Aplicando esta metodología a nuestro conjunto de datos grasa, vamos a considerar que las relevantes variables que se eligieron con el  $C_p$  son: edad, peso, abdomen, muslo, antebrazo y muñeca. Luego, aplicaremos el clásico “stepwise” con estas variables predictoras solamente. Usando MINITAB, obtenemos

#### Stepwise Regression: grasa versus edad, peso, ...

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is grasa on 6 predictors, with N = 252

Step	1	2	3	4
Constant	-39.28	-45.95	-27.93	-34.85
abdomen	0.631	0.990	0.975	0.996
T-Value	22.11	17.45	17.37	17.76
P-Value	0.000	0.000	0.000	0.000
peso		-0.148	-0.114	-0.136

T-Value	-7.11	-4.84	-5.48	
P-Value	0.000	0.000	0.000	
muneca		-1.24	-1.51	
T-Value		-2.85	-3.40	
P-Value		0.005	0.001	
antebrazo			0.47	
T-Value			2.60	
P-Value			0.010	
S	4.88	4.46	4.39	4.34
R-Sq	66.17	71.88	72.77	73.50
R-Sq(adj)	66.03	71.65	72.44	73.07
Mallows Cp	72.0	20.0	13.5	8.6

Solo quedan: abdomen, peso, muñeca y antebrazo como las variables predictoras más importantes. Así que las variables edad y muslo parecen ser redundantes, esto se explica porque estas variables están correlacionadas con otras predictoras. Este problema sera discutido en el próximo capítulo.

#### 6.4 Otros métodos de Selección de variables

Existen muchos otros métodos de selección de variables en regresión, solo mencionaremos cuatro de ellos.

##### 6.4.1. Método basados en remuestreo.

**Bootstrapping** (Efron, 1983)

**El pequeño Boostrapping** (Breiman, 1992)

##### 6.4.2. Métodos basados en Regresión Penalizada

La idea de estos métodos es estimar la importancia de cada uno de las variables predictoras y luego considerar solamente en el modelo aquellas que no tienen una relevancia despreciable. Se relaciona con la metodología de regression ridge que se verá en el próximo capítulo. Dos de los métodos más usados son:

**La Garrote** (Breiman, 1995)

**El Lasso** (Tibshirani, 1996)

##### 6.4.3 Métodos Bayesianos

Es considerado en gran detalle por Mitchel y Beauchamp (JASA, 1988). Supongamos que ya se tiene un conjunto de buenos modelos. La idea se basa en asignar probabilidades a priori a los coeficientes de cada uno de estos modelos que incluyen solo un subconjunto de predictoras e igualmente se asignan probabilidades a priori a cada uno de los modelos. Finalmente, se elige como mejor modelo aquel que tiene la probabilidad posterior más alta con respecto a la variable de respuesta.

**6.4.4. Algoritmo Genéticos:** En este caso el problema de selección de variables es considerado como un problema de optimización con respecto al número de variables predictoras que deben incluirse en el modelo. Luego, el problema de optimización es resuelto usando algoritmos Genéticos.

## Ejercicios

1. Supongamos que se desea omitir la variable predictora  $X_j$  de un modelo de regresión con  $p$  parámetros (incluyendo el intercepto) y  $n$  observaciones. Si  $F_j$  es el estadístico para probar la hipótesis  $H: \beta_j=0$  demostrar que:

$$C_{p-1} = \frac{F_j SSE_p}{s^2(n-p)} + C_p - 2$$

2. Hacer selección de variables predictoras usando el conjunto de datos

**Berkeley:** La variable de respuesta es SOMA y las predictoras son WT2, HT2, WT9, HT9, LG9, ST9. Disponible en la página de internet del texto.

- Los metodos “stepwise”. Explicar los pasos de los procesos y justificar la terminación del mismo.
- Los mejores usando los mejores subconjuntos con por lo menos 6 criterios. Explicar los resultados
- Comparar los resultados obtenidos en a y b. Dar su seleccion final.

3. Hacer selección de variables predictoras usando el conjunto de datos

**Highway:** La variable de respuesta es TASA y todas las otras son predictoras

- Los metodos “stepwise”. Explicar los pasos de los procesos y justificar la terminación del mismo.
  - Los mejores usando los mejores subconjuntos con por lo menos 6 criterios. Explicar los resultados
  - Comparar los resultados obtenidos en a) y b). Dar su selección final.
4. Verificar las ecuaciones (3) y (4) de las sección 6.3.4 del texto.
5. Investigar la relación entre los criterios AIC, BIC,  $C_p$  de Mallows y  $R^2$ .

# CAPÍTULO 7

## MULTICOLINEALIDAD

### 7.1 Multicolinealidad.

Dos predictoras  $X_1$  y  $X_2$  son exactamente colineales si existe una relación lineal tal que  $C_1X_1+C_2X_2=C_0$  para algunas constantes  $C_1$ ,  $C_2$  y  $C_0$ . Si la ecuación se cumple aproximadamente para los datos observados entonces se dice que hay colinealidad aproximada. Una medida comúnmente usada para detectar colinealidad es el coeficiente de determinación. Se dice que  $X_1$  y  $X_2$  son colineales si  $R_{12}^2$  es bastante cercano a 1 ( ó 100%). Sin embargo, cuando existen “outliers” esta medida no es completamente adecuada.

La definición se extiende al caso cuando hay más de dos variables predictoras. Un conjunto de predictoras  $X_1, X_2, \dots, X_p$  son colineales si para constantes  $c_0, c_1, \dots, c_p$ , la ecuación

$$c_1X_1+c_2X_2+\dots+c_pX_p=c_0 \quad (1)$$

se cumple aproximadamente. De la ecuación anterior se desprende que cuando hay multicolinealidad una de las predictoras puede ser determinada de las otras. Es decir,

$$X_k = (c_0 - \sum_{j \neq k} c_j X_j) / c_k \quad (2)$$

si el coeficiente de determinación  $R_k^2$  de la regresión de  $X_k$  con las otras variables predictoras es cercano a 1 se puede concluir tentativamente que hay multicolinealidad.

#### 7.1.1 Efectos de multicolinealidad

Si consideramos el modelo de regresión lineal múltiple

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_pX_p + e$$

entonces se puede mostrar que la varianza del j-ésimo coeficiente de regresión estimado es

$$\text{var}(\hat{\beta}_j) = \sigma^2 \left( \frac{1}{1-R_j^2} \right) \left( \frac{1}{S_{X_jX_j}} \right) \quad (3)$$

donde  $R_j^2$  es el coeficiente de Determinación de la regresión lineal de  $X_j$  contra todas las demas

predictoras. La cantidad  $\frac{1}{1-R_j^2}$  es llamado el j-ésimo **Factor de inflación de la varianza**, or

VIF<sub>j</sub> (Marquardt, 1970). Si  $R_j^2$  es cercano a 1 entonces la varianza de  $\hat{\beta}_j$  aumentará grandemente. El VIF representa el incremento en la varianza del coeficiente de una variable predictor debido a la presencia de multicolinealidad. Una variable predictor con un VIF mayor de 10 (esto es equivalente a un  $R^2=.90$ ), puede causar multicolinealidad. La mayoría de los programas estadísticos da los valores VIF. Otra forma practica de encontrar los VIF's es

hallando la matriz  $C^{-1}$ , que es la inversa de la matriz de correlaciones  $C$ . y luego los VIF's serán elementos que están en la diagonal de  $C^{-1}$ .

**Ejemplo 1.** Usar R para calcular los VIF para los datos del conjunto **millaje**.

```
> # Hallando la matriz de correlaciones de las variables predictoras
> mcor<-cor(millaje[,2:5])
> mcor
           sp          wt          vol          hp
sp  1.00000000  0.6785339 -0.04306242  0.96654517
wt  0.67853388  1.0000000  0.38495423  0.83222021
vol -0.04306242  0.3849542  1.00000000  0.07647905
hp  0.96654517  0.8322202  0.07647905  1.00000000
>

># Hallando la inversa de la matriz de correlaciones
> invcorr=solve(mcor)
> invcorr
           sp          wt          vol          hp
sp  71.675777  30.997855 -1.584250 -94.95376
wt  30.997855  18.318201 -2.272839 -45.03178
vol -1.584250 -2.272839  1.554038  3.30390
hp -94.953756 -45.031780  3.303900 130.00077
>
```

Los elementos de la diagonal de la inversa de la matriz de correlaciones son los VIF's.

```
> #Hallando los VIFs
> vif<-diag(solve(mcor))
> cat("los VIF's son:\n")
los VIF's son:
> vif
           sp          wt          vol          hp
71.675777  18.318201  1.554038 130.000772
```

Las variables HP y SP tiene un VIF bastante alto.

Eliminando la variable HP y recalculando los VIF's se obtiene

```
> #recalculando los VIF's sin la variable hp
> diag(solve(cor(millaje[,2:4])))
           sp          wt          vol
2.320683  2.719362  1.470071
```

Con lo cual parece resolverse el problema de multicolinealidad.

Para ver el efecto de la multicolinealidad, compararemos la regresión considerando todas las predictoras con la regresión que excluye la predictora hp.

```
> l1=lm(mpg~.,data=millaje)
> summary(l1)
```

Call:

```
lm(formula = mpg ~ ., data = millaje)
```

Residuals:

```

Min   1Q   Median   3Q   Max
-9.0108 -2.7731 0.2733 1.8362 11.9854

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775  23.53161  8.178 4.62e-12 ***
sp          -1.29482   0.24477 -5.290 1.11e-06 ***
wt          -1.85980   0.21336 -8.717 4.22e-13 ***
vol         -0.01565   0.02283 -0.685  0.495
hp           0.39221   0.08141  4.818 7.13e-06 ***
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.653 on 77 degrees of freedom  
Multiple R-Squared: 0.8733, Adjusted R-squared: 0.8667  
F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

```

> l2=lm(mpg~sp+wt+vol,data=millaje)
> summary(l2)

```

Call:

```
lm(formula = mpg ~ sp + wt + vol, data = millaje)
```

Residuals:

```

Min   1Q   Median   3Q   Max
-10.0003 -2.7013 -0.5674  1.2842 16.7766

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) 81.18296   5.12341 15.845 < 2e-16 ***
sp          -0.13484   0.04992 -2.701 0.00847 **
wt          -0.91127   0.09318 -9.780 3.35e-15 ***
vol         -0.04121   0.02516 -1.638 0.10554
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.14 on 78 degrees of freedom  
Multiple R-Squared: 0.8351, Adjusted R-squared: 0.8287  
F-statistic: 131.7 on 3 and 78 DF, p-value: < 2.2e-16

Notar el gran cambio en los coeficientes de regresión y en la desviación estándar de los mismos. La multicolinealidad también afecta al valor predicho y a la varianza de las predicciones.

### 7.1.2 Diagnósticos de Multicolinealidad.

De acuerdo a Besley, et al. (1991) se pueden seguir los siguientes pasos para detectar multicolinealidad

1) Cotejar si hay coeficientes de regresión con valores bien grandes o de signo opuesto a lo que se esperaba que ocurriera.

- 2) Cotejar si las variables predictoras que se esperaban sean importantes tienen valores de  $t$  pequeños para las hipótesis de sus coeficientes.
- 3) Cotejar si la eliminación de una fila o columna de la matriz  $X$  produce grandes cambios en el modelo ajustado.
- 4) Cotejar las correlaciones entre todas las parejas de variables predictoras para detectar las que son bastante altas.
- 5) Examinar el VIF. Si el VIF es grande, mayor que 10, entonces puede haber multicolinealidad.
- 6) Sea  $U$  una matriz tal que  $Z=XU$  y que  $Z'Z=U'X'XU=D$  donde  $D$  es una matriz diagonal con elementos positivos  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ . Los  $\lambda$ 's son llamados los eigenvalues (valores propios) de  $X'X$  y las columnas de  $U$  son los eigenvectors (vectores propios) de  $X'X$ . Se puede mostrar que  $U$  es ortogonal, es decir  $U'U=UU'=I$ . Las columnas de  $Z=XU$  son llamados **componentes principales** (ver sección 7.3). Algunos autores prefieren usar los “eigenvalues” de la matriz  $X'X$ , que se obtiene cuando cada columna de  $X$  es centrada solamente (matriz de covarianza), o cuando cada columna de  $X$  tiene norma 1, o sea se ha dividido cada columna de la matriz por la suma de sus cuadrados (matriz de correlación). Es decir,  $X^*X^*$ .

La matriz  $X^*$  es obtenida restando a cada columna de  $X$  la media correspondiente y dividiendo luego entre la raíz de la suma de cuadrados corregida por la media de la misma columna. La matriz de correlación es de la forma

$$\begin{bmatrix} 1 & r_{12} & \cdot & \cdot & r_{1p} \\ r_{21} & 1 & \cdot & \cdot & r_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{p1} & r_{p2} & \cdot & \cdot & 1 \end{bmatrix}$$

donde  $r_{ij}$  representa la correlación entre las variables  $X_i$  y  $X_j$ .

La presencia de “eigenvalues” cerca de 0 es también una indicación de multicolinealidad.

- 7) Usar el número condición de la matriz correlación  $X^*X^*$ . El número condición de la matriz  $X^*$  está definido por

$$K = (\text{mayor "eigenvalue"} / \text{menor "eigenvalue"})^{1/2} \quad (4)$$

Los “eigenvalues” que se usan son de la matriz de correlación. Notar que  $K \geq 1$ . Weisberg sugiere que un  $K > 30$  indica presencia de multicolinealidad. Sin embargo, basados en varios experimentos efectuados, pensamos que  $K > 10$  es suficiente para indicar multicolinealidad.

A continuación encontraremos el número condición correspondiente a la matriz de variables predictoras del conjunto de datos millaje/

```
> # Hallando los valores y vectores propios
> ev<-eigen(cor(millaje[,2:5]))
```

```
> ev
$values
[1] 2.689421048 1.100439312 0.205497158 0.004642483
```

```
$vectors
      [,1] [,2] [,3] [,4]
[1,] -0.5616362 0.2921753 0.5229999 0.57067460
[2,] -0.5526927 -0.2365918 -0.7530524 0.26733338
[3,] -0.1387771 -0.9134537 0.3820731 -0.01906127
[4,] -0.5998637 0.1557567 0.1157728 -0.77620876
```

```
> evals<-ev$values
> #Hallando el numero condicion
> cond<-sqrt(evals[1]/ev$values[4])
> cond
[1] 24.06879
```

Como el número condición es cercano a 30 se concluiría que hay algo de multicolinealidad presente.

8) Usar los *índices de condición*, definidos por  $\left(\frac{\lambda_{\max}}{\lambda_j}\right)^{1/2}$ . Un índice condición grande indica

presencia de multicolinealidad.

9) Descomponer la varianza de cada coeficiente en proporciones debido a cada una de los valores propios. Una variable con una proporción grande (mayor que .5) correspondiente a un índice condición grande implica que dicha variable está produciendo la multicolinealidad.

R no da directamente estos dos últimos diagnósticos, pero SAS si los muestra.

**Ejemplo 2.** Usar R para hallar los diagnósticos de multicolinealidad para el conjunto de datos **millaje**.

```
> indice_cond=sqrt(ev$values[1]/ev$values)
> indice_cond
[1] 1.000000 1.563314 3.617650 24.068790
> ev$values
[1] 2.689421048 1.100439312 0.205497158 0.004642483
> phi=ev$vectors^2/(ev$values)^2
> > colnames(phi)=colnames(millaje)[-1]
> phi
      sp      wt      vol      hp
[1,] 4.361065e-02 1.180238e-02 0.03781687 4.502560e-02
[2,] 2.522524e-01 4.622397e-02 0.46829360 5.901660e-02
[3,] 4.560623e-01 1.975884e+01 3.45685577 8.603831e-03
[4,] 1.669570e+04 1.125623e+03 621.88892248 2.795481e+04
> b=apply(phi,1,sum)
> prop.var=phi/b
> prop.var
```

```

      sp      wt      vol      hp
[1,] 0.31543522 0.08536643 0.27352886 0.3256694942
[2,] 0.30546924 0.05597568 0.56708794 0.0714671344
[3,] 0.01925909 0.83439773 0.14597985 0.0003633319
[4,] 0.35983645 0.02426016 0.01340335 0.6025000395
> apply(phi/b,1,sum)
[1] 1 1 1 1

```

Notar que la variable HP es la que tiene la proporción mas alta correspondiente al indice condicion mayor y por lo tanto da problemas de multicolinealidad.

### 7.1.3 Medidas remediales al problema de multicolinealidad

Básicamente hay tres propuestas:

- Regresión Ridge (Hoerl and Kennard, 1970)
- Componentes principales (Hotelling, 1965)
- Mínimos Cuadrados Parciales (H. Wold, 1975)

Sin embargo el problema de multicolinealidad también está relacionado con el proceso de selección de variables y esto puede ser considerado como una cuarta manera de resolver el problema de multicolinealidad. En las próximas dos secciones se discutirá regresión ridge y componentes principales. Mínimos cuadrados parciales, que es una técnica muy usada cuando hay pocas observaciones y muchas variables no será considerado en este texto.

## 7.2 Regresión Ridge

Consideremos la suma de las varianzas de los coeficientes de regression estimados por minimos cuadrados,  $\hat{\beta}$ , dada por  $E(\hat{\beta} - \beta)'(\hat{\beta} - \beta)$ . Hoerl y Kennard (1970) mostraron que

$$E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = E[e'X^*(X^*X^*)^{-2}X^*e] = \sigma^2 \text{Traza}(X^*X^*)^{-1} = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} \quad (5)$$

Donde  $\lambda_j$ ,  $j=1, \dots, p$ , son los valores propios de la matriz de correlación de las variables predictoras. Notar que si un valor propio (eigenvalue) es cercano a cero la suma de las varianzas se hace muy grande. Por otro lado, de la ecuación (5) se puede establecer que

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} \quad (6)$$

De donde,

$$E(\hat{\beta}'\hat{\beta}) \geq \beta'\beta + \sigma^2 \frac{1}{\lambda_p} \quad (7)$$

Es decir, que aún cuando  $\hat{\beta}$  es insesgado, se tiene que el cuadrado de su norma,

$\|\hat{\beta}\|^2 = \hat{\beta}'\hat{\beta} = \sum_{i=1}^p \hat{\beta}_i^2$  es un estimador sesgado.

La idea en regresión Ridge es encontrar un estimador  $\tilde{\beta}$  que aunque sea sesgado tenga menor longitud que  $\hat{\beta}$ . Es decir,  $\tilde{\beta}'\tilde{\beta} < \hat{\beta}'\hat{\beta}$ , lo cual significa que es un estimador más encogido hacia el origen que el estimador mínimo cuadrático.

Hoerl y Kennard, en 1970, propusieron el siguiente estimador:

$$\tilde{\beta} = (X'X + kI)^{-1} X'Y \quad (8)$$

donde el parámetro de encogimiento  $k$  (por lo general,  $0 < k < 1$ ) debe ser estimado de los datos tomados. Si  $k=0$  se obtiene el estimador mínimo cuadrático y a medida que  $k$  aumenta el estimador se aleja del estimador mínimo cuadrático y se hace más sesgado.

Se puede mostrar que el estimador ridge se obtiene al resolver

$$\begin{aligned} & \text{Min}_{\mathbf{B}} (\mathbf{y} - \mathbf{XB})'(\mathbf{y} - \mathbf{XB}) \\ & \text{Sujeto a que } \|\mathbf{B}\|^2 < k^2 \end{aligned} \quad (9)$$

Cuando la restricción (9) se sustituye por  $|\mathbf{B}| < k$  se obtiene el estimador *Lasso* (Tibshirani, 1996).

La fórmula anterior se puede usar con las variables predictoras y/o de respuesta en su forma original o en su forma estandarizadas. Sin embargo, desde el punto de vista computacional se recomienda trabajar con variables estandarizadas. Existen diversos tipos de estandarizaciones, el que más se usa es centrar los datos (restando por la media) y luego rescalarlos (dividiendo entre la desviación estándar o la raíz cuadrada de las sumas de cuadrados). También existen variantes de la formulación original de la regresión ridge, uno de ellos es considerar el parámetro  $k$  como un vector o una matriz, a estos métodos se le llama regresión ridge generalizada.

Hay varias propuestas acerca de la elección de  $k$ , pero lo que más se recomienda consiste en hacer un plot de los coeficientes del modelo para varios valores de  $k$  (generalmente, entre 0 y 1) este plot es llamado la **Traza Ridge**. Para elegir  $k$  hay que considerar los siguientes aspectos

- Que los valores de los coeficientes de regresión se estabilicen.
- Que los coeficientes de regresión que tenían un valor demasiado grande comiencen a tener valores razonables.
- Que los coeficientes de regresión que inicialmente tenían el signo equivocado cambien de signo.

En R la librería MASS tiene la función **lm.ridge** que ejecuta la regresión ridge aunque con otro tipo de estandarización. Nosotros hemos construido una función **acunaridge** que da los resultados iguales a los producidos por la mayoría de los otros programas estadísticos. MINITAB no hace regresión Ridge, pero se puede preparar un macro donde se calcule  $\tilde{\beta}$  según la fórmula dada para varios valores de  $k$ . y luego se plotea la traza Ridge. SAS si produce los coeficientes de la regresión ridge.

Para los datos de **millaje** se obtienen los siguientes resultados.

```
> rr1=acunaridge(mpg~.,data=millaje,lambda=seq(0,0.1,.01))
> rr1
```

\$coef

	0.00	0.01	0.02	0.03	0.04	0.05
Intercept	192.437753	108.396120	92.873600	86.491901	83.081559	80.992916
sp	-1.294818	-0.421319	-0.261868	-0.197614	-0.164265	-0.144632
wt	-1.859804	-1.127997	-0.975158	-0.901146	-0.853794	-0.819022
vol	-0.015645	-0.036861	-0.042665	-0.046188	-0.048825	-0.050967
hp	0.392212	0.096143	0.040194	0.016449	0.003285	-0.005096
	0.06	0.07	0.08	0.09	0.10	
Intercept	79.599544	78.612442	77.880438	77.317184	76.870043	
sp	-0.132191	-0.123935	-0.118291	-0.114362	-0.111604	
wt	-0.791393	-0.768332	-0.748448	-0.730909	-0.715181	
vol	-0.052777	-0.054338	-0.055703	-0.056905	-0.057970	
hp	-0.010906	-0.015171	-0.018435	-0.021011	-0.023095	

\$lambda

[1] 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10

\$kHKB

[1] 0.0006234958

```
> #Haciendo la Traza Ridge
> matridge<-cbind(rr1$lambda,t(rr1$coef[-1,]))
> win.graph()
> plot(matridge[,1],matridge[,2],ylim=c(2,0.5),type="l",xlab="lamb",ylab="coefs",col=2)
> title("La traza Ridge para Millaje")
> lines(matridge[,1],matridge[,3],col=3)
> lines(matridge[,1],matridge[,4],col=4)
> lines(matridge[,1],matridge[,5],col=6)
```

### La traza Ridge para Millaje

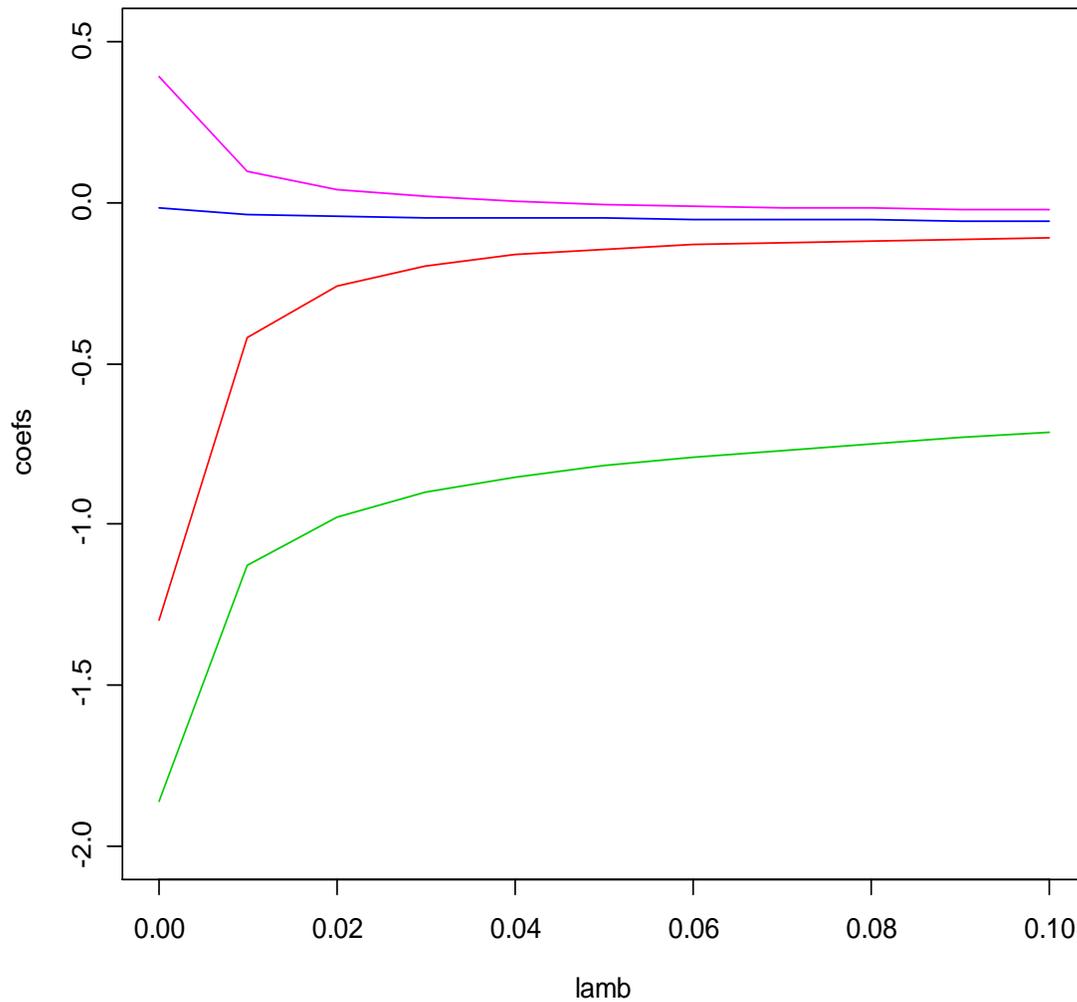


Figura 7.1. La Traza Ridge para los coeficientes de regresión del conjunto de datos millaje

En este plot se nota que el  $k$ -óptimo parece ser menor que 0.05. Detallando más el plot hemos llegado a establecer un  $k$ -óptimo de alrededor de 0.015. Un segundo método de elegir el parámetro  $k$  propuesto por Hoerl, Kennard y Baldwin (1975) está basado en procedimientos bayesianos. En este caso, el  $k$  óptimo es un estimado de la razón entre la varianza poblacional  $\sigma^2$  y la varianza del estimador ridge. Más específicamente

$$k_{opt} = \frac{ps^2}{\sum_{i=1}^p b_i^{*2}(0)} \quad (9)$$

Donde  $p$  es el número de variables predictoras,  $s^2$  es la estimación de la varianza de los errores del modelo de mínimos cuadrados trabajando con las variables originales y sin usar ningún tipo de estandarización. Finalmente,  $b_i^{*2}(0)$ , es el cuadrado del  $i$ -ésimo coeficiente de la regresión por mínimos cuadrados pero donde se ha usado la siguiente estandarización:

- a) La variable de respuesta Y ha sido centrada, es decir se le ha restado la media a cada dato.  
 b) Las variables predictoras han sido centradas y escaladas. Es decir a cada dato se le ha restado la media de la variable y dividido entre la raíz cuadrada de la suma de cuadrados con respecto a la media de cada variable. Es decir, para cada variable predictora x se ha aplicado la siguiente transformación.

$$z = \frac{x - \bar{x}}{\sqrt{Sxx}} \quad (10)$$

En el ejemplo anterior,  $s^2=13.34$ ,  $p=4$  y  $\sum_{i=0}^p b_i^{*2}(0) = 85598.8$  dando  $k_{opt}=0.000623$ .

**Ejemplo 3.** El conjunto de datos **Pollution** es uno de los más usados en los trabajos de regression ridge. Aplicar el método anterior para elegir el k-óptimo.

Fuente: McDonald, G.C. and Schwing, R.C. (1973) 'Instabilities of regression estimates relating air pollution to mortality', *Technometrics*, vol.15, 463-482.

Variabes en orden:

PREC Average annual precipitation in inches  
 JANT Average January temperature in degrees F  
 JULT Same for July  
 OVR65 % of 1960 SMSA population aged 65 or older  
 POPN Average household size  
 EDUC Median school years completed by those over 22  
 HOUS % of housing units which are sound & with all facilities  
 DENS Population per sq. mile in urbanized areas, 1960  
 NONW % non-white population in urbanized areas, 1960  
 WWDRK % employed in white collar occupations  
 POOR % of families with income < \$3000  
 HC Relative hydrocarbon pollution potential  
 NOX Same for nitric oxides  
 SOD Same for sulphur dioxide  
 HUMID Annual average % relative humidity at 1pm  
 MORT Total age-adjusted mortality rate per 100,000

En algunos textos se ha elegido el k-óptimo como 0.2 , simplemente observando a la traza ridge que es mostrado en la siguiente figura

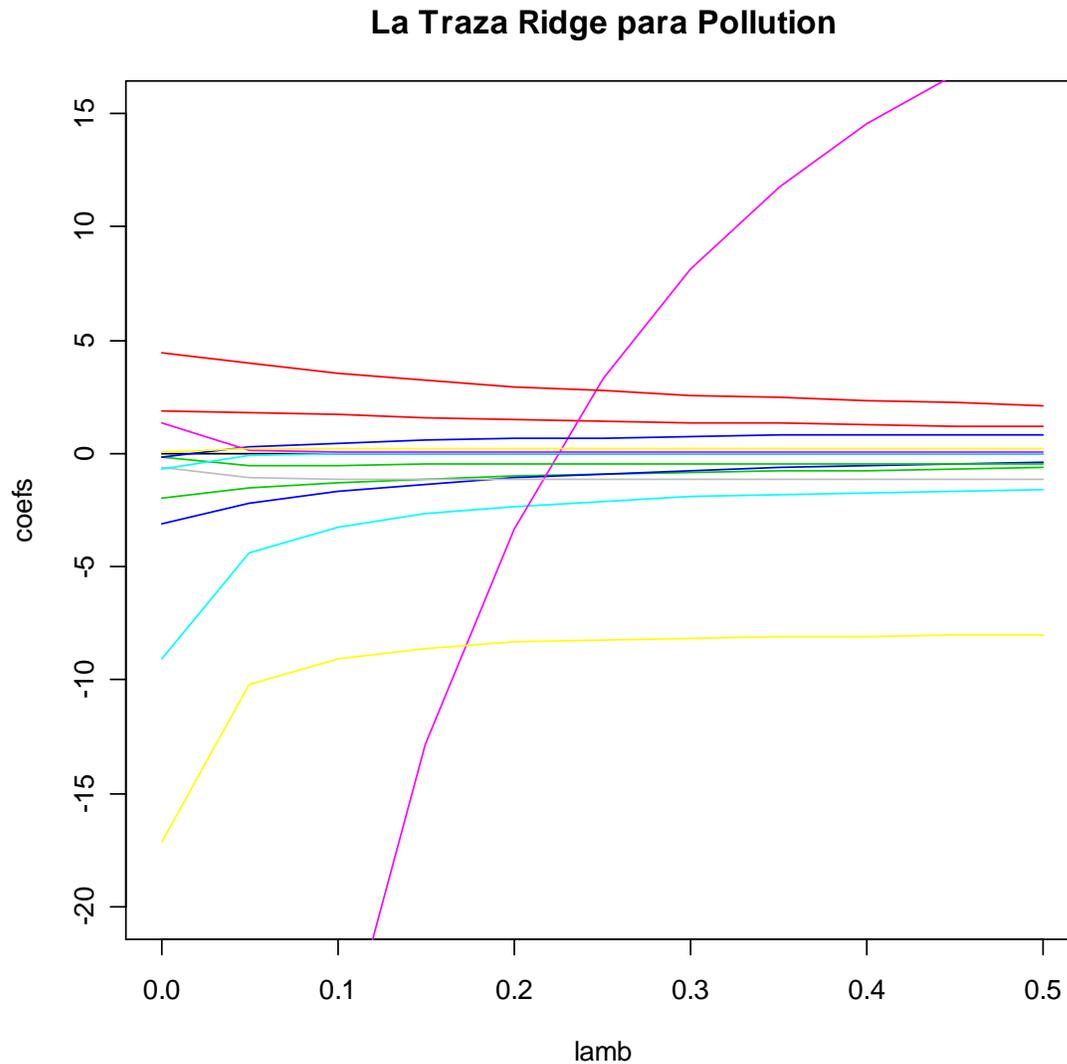


Figura 7.2. La Traza Ridge para los coeficientes de regresión del conjunto de datos Pollution.

Usando R se obtienen los siguientes resultados

A) La regresión mínimo cuadrática con las variables originales

```
> pollution=read.table("c://pollution.txt",header=T,skip=22)
> summary(lm(MORT~.,data=pollution))
```

Call:

```
lm(formula = MORT ~ ., data = pollution)
```

Residuals:

```
   Min     1Q  Median     3Q    Max
-68.0663 -18.0171  0.9117 19.2237 86.9612
```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.764e+03 4.373e+02 4.034 0.000215 ***
PREC        1.905e+00 9.237e-01 2.063 0.045071 *
JANT       -1.938e+00 1.108e+00 -1.748 0.087413 .
JULT       -3.100e+00 1.902e+00 -1.630 0.110159
OVR65      -9.065e+00 8.486e+00 -1.068 0.291230
POPN      -1.068e+02 6.978e+01 -1.531 0.132952
EDUC      -1.716e+01 1.186e+01 -1.447 0.155085
HOUS      -6.511e-01 1.768e+00 -0.368 0.714393
DENS       3.600e-03 4.027e-03 0.894 0.376147
NONW       4.460e+00 1.327e+00 3.360 0.001618 **
WWDRK     -1.871e-01 1.662e+00 -0.113 0.910883
POOR      -1.676e-01 3.227e+00 -0.052 0.958807
HC        -6.721e-01 4.910e-01 -1.369 0.177985
NOX       1.340e+00 1.006e+00 1.333 0.189506
SO.       8.625e-02 1.475e-01 0.585 0.561745
HUMID     1.068e-01 1.169e+00 0.091 0.927644

```

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.93 on 44 degrees of freedom  
Multiple R-squared: 0.7649, Adjusted R-squared: 0.6847  
F-statistic: 9.542 on 15 and 44 DF, p-value: 2.193e-09

```

> mcor=cor(pollution[,1:15])
> vif<-diag(solve(mcor))
> cat("los VIF's son:\n")
los VIF's son:
> vif
  PREC   JANT   JULT   OVR65   POPN   EDUC   HOUS
4.113888 6.143551 3.967774 7.470045 4.307618 4.860538 3.994781
  DENS   NONW   WWDRK   POOR   HC   NOX   SO.
1.658281 6.779599 2.841582 8.717068 98.639935 104.982405 4.228929
  HUMID
1.907092
>

```

De aquí se obtiene  $s^2=34.93^2=1220$  y notar que solo las predictoras HC y NOX parecen dar problemas de multicolinealidad.

```

> #recalculando los VIF's sin las variable HC(12) y NOX(13)
> diag(solve(cor(pollution[,c(1:11,14:15)]))))
  PREC   JANT   JULT   OVR65   POPN   EDUC   HOUS   DENS
3.188402 4.550720 3.448561 6.526173 4.216491 4.334704 3.888674 1.656427
  NONW   WWDRK   POOR   SO.   HUMID
5.665828 2.529818 8.342648 1.713090 1.550716

```

Parece haberse resuelto el problema de multicolinealidad

```
> #Hallando el numero condicion
```

```
> ev2<-eigen(cor(pollution[,1:15]))
> evals2<-ev2$values
> evals2
[1] 4.528391602 2.754841543 2.054464043 1.348389581 1.223219959 0.960443977
0.612741552 0.472011722 0.370853024 0.216394684 0.166350401 0.127005110
[13] 0.113986775 0.046039741 0.004866287
> cond2<-sqrt(evals2[1]/evals2[15])
> cond2
[1] 30.50515
```

El valor del número condición indica cierto grado de multicolinealidad.

B) La siguiente es la salida con la Y centrada y las predictoras centradas y escaladas

```
> summary(lm(MORT~.-1,data=as.data.frame(polluesca)))
```

Call:

```
lm(formula = MORT ~ . - 1, data = as.data.frame(polluesca))
```

Residuals:

```
Min      1Q  Median      3Q      Max
-68.0663 -18.0171  0.9117  19.2237  86.9612
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
PREC  146.134    70.053  2.086 0.04267 *
JANT  -151.345    85.607 -1.768 0.08386 .
JULT  -113.434    68.798 -1.649 0.10615
OVR65 -101.981    94.398 -1.080 0.28575
POPN  -110.980    71.683 -1.548 0.12858
EDUC  -111.399    76.145 -1.463 0.15042
HOUS  -25.714    69.031 -0.372 0.71127
DENS   40.214    44.476  0.904 0.37072
NONW  305.592    89.929  3.398 0.00143 **
WWDRK  -6.628    58.221 -0.114 0.90987
POOR   -5.357   101.973 -0.053 0.95834
HC   -474.866   343.026 -1.384 0.17308
NOX   476.931   353.882  1.348 0.18450
SO.    41.997    71.026  0.591 0.55728
HUMID   4.405    47.696  0.092 0.92682
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.54 on 45 degrees of freedom

Multiple R-squared: 0.7649, Adjusted R-squared: 0.6865

F-statistic: 9.759 on 15 and 45 DF, p-value: 1.204e-09

El vector de coeficientes  $b_i^*$  (0) será

```
> a$coeff
```

```

PREC   JANT   JULT   OVR65   POPN   EDUC
146.133717 -151.345128 -113.434215 -101.980618 -110.980457 -111.399273
HOUS   DENS   NONW   WWDRK   POOR   HC
-25.713683 40.214442 305.592481 -6.628012 -5.356901 -474.866318
NOX    SO.    HUMID
476.930673 41.996977 4.405377

```

```

> sum(a$coeff*a$coeff)
[1] 642736
>

```

Obteniendose,  $\sum_{i=0}^p b_i^{*2}(0) = 642736$ .

Finalmente, el k-óptimo será  $k=15*1220/642736=0.028472$ . Bastante más pequeño de lo que dice la literatura.

Existen otros métodos para elegir la traza Ridge, que están relacionados con los criterios usados para seleccionar los mejores subconjuntos tales como el  $C_p$  de Mallows, Validación Cruzada, PRESS, Validación cruzada generalizada, BIC, AIC, etc.

### 7.2.1 Aplicación de Regresión Ridge a Selección de variables

Según Hoerl y Kennard la regresión ridge puede usarse para seleccionar variables de la siguiente manera:

- Eliminar las variables cuyos coeficientes sean estables pero de poco valor. Si se trabaja con variables previamente estandarizadas, se pueden comparar directamente los coeficientes.
- Eliminar las variables con coeficientes inestables que tienden a cero.
- Eliminar las variables con coeficientes inestables.

**Ejemplo 4.** Usar el conjunto de datos **Pollution** los datos y aplicar regresión ridge para seleccionar variables predictoras.

**Solución:**

```

> # Hallando la regresion ridge
> rr2<-acunaridge(MORT~.,data=pollution,lambda=seq(0,.5,.05))

> rr2
$coef
      0.00   0.05   0.10   0.15   0.20
Intercept 1763.981364 1422.594815 1280.234719 1195.201816 1139.487870
V1      1.905419  1.804597  1.688618  1.593370  1.512140
V2     -1.937620 -1.517852 -1.298273 -1.142933 -1.024087
V3     -3.100426 -2.170975 -1.675098 -1.343193 -1.101898
V4     -9.065397 -4.364218 -3.266429 -2.685603 -2.331372
V5    -106.825702 -50.653231 -27.077625 -12.803395 -3.345964
V6    -17.157175 -10.198638 -9.057573 -8.569353 -8.331938
V7     -0.651117 -1.092326 -1.144082 -1.161332 -1.164510
V8      0.003600  0.004581  0.004992  0.005187  0.005272

```

```

V9      4.459600  3.951995  3.513477  3.192037  2.944157
V10     -0.187055 -0.554213 -0.508021 -0.478066 -0.462956
V11     -0.167643  0.256645  0.435579  0.554168  0.637346
V12     -0.672145 -0.075996 -0.047136 -0.036609 -0.031148
V13      1.340097  0.134323  0.084192  0.066940  0.057695
V14      0.086252  0.231817  0.239971  0.240491  0.238161
V15      0.106804  0.214495  0.315819  0.365177  0.386587
      0.25  0.30  0.35  0.40  0.45
Intercept 1100.736981 1072.600049 1051.488503 1035.230407 1022.440730
V1       1.441505  1.379284  1.323922  1.274251  1.229365
V2      -0.929050 -0.850769 -0.784879 -0.728488 -0.679579
V3      -0.916990 -0.770002 -0.649965 -0.549901 -0.465123
V4      -2.096823 -1.932272 -1.811562 -1.719735 -1.647702
V5       3.283430  8.117017 11.744709 14.527586 16.698532
V6      -8.206389 -8.133610 -8.085801 -8.049106 -8.016414
V7      -1.160604 -1.152825 -1.142859 -1.131657 -1.119786
V8       0.005296  0.005282  0.005245  0.005193  0.005131
V9       2.745250  2.580801  2.441687  2.321862  2.217139
V10     -0.457394 -0.457622 -0.461238 -0.466733 -0.473156
V11      0.698272  0.744311  0.779873  0.807764  0.829857
V12     -0.027810 -0.025558 -0.023932 -0.022699 -0.021726
V13      0.051561  0.046982  0.043318  0.040255  0.037620
V14      0.234481  0.230125  0.225445  0.220636  0.215816
V15      0.392254  0.388724  0.379733  0.367518  0.353463
      0.50
Intercept 1012.198883
V1       1.188549
V2      -0.636694
V3      -0.392352
V4      -1.589678
V5      18.414109
V6      -7.984059
V7      -1.107594
V8       0.005062
V9       2.124517
V10     -0.479911
V11      0.847457
V12     -0.020936
V13      0.035307
V14      0.211053
V15      0.338451

```

\$lambda

[1] 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50

\$kHKB

[1] 0.02847204

Las variables DENS, WWDRK , SO., y HUMID tienen coeficientes estables pero pequeños. Las variables HC y NOX tienen coeficientes inestables y que tienden a 0 y las variables JULT,

OVR65, POPN y POOR tienen coeficientes inestables. Así que solo deberían quedar en el modelo las variables PREC, JANT, EDUC, HOUS, NONW.

Usando los métodos stepwise se eligen: PREC, JANT, JULT, EDUC, NONW y SO. y usando los mejores subconjuntos se eligen: PREC, JANT, EDUC, NONW, y SO., que es el que tiene el BIC mas pequeño.

### 7.3. Componentes principales para Regresión

El objetivo del análisis por componentes principales ((Hotelling, 1933) es hacer una reducción de la información disponible. Es decir, la información contenida en  $p$  variables predictoras  $\mathbf{X}=(X_1, \dots, X_p)$  puede ser reducida a  $\mathbf{Z}=(Z_1, \dots, Z_p)$ , con  $p' < p$  y donde las nuevas variables  $Z_i$ 's llamadas las **componentes principales** no están correlacionadas. Es decir, los componentes principales de un vector aleatorio  $\mathbf{X}$  son los elementos de una transformación lineal ortogonal de  $\mathbf{X}$

Geoméricamente, hablando la aplicación de componentes principales equivale a hacer una rotación de los ejes coordenados.

Consideremos el modelo de regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + e$$

Estandarizemos todas las variables predictoras  $X_j$  usando  $X_j^* = \frac{x_j - \bar{x}_j}{\sqrt{SXX_j}}$ . Sea  $X^*$  la matriz

obtenida usando las  $X_j^*$  como columnas. Luego,  $X^{*'}X^*$  viene a ser la matriz de correlación de las variables predictoras  $X_j$

Para determinar los componentes principales hay que hallar una matriz ortogonal  $V$  tal que  $Z=X^*V$  y para la cual  $Z'Z=(X^*V)'(X^*V)=V'X^{*'}X^*V=\text{diag}(\lambda_1, \dots, \lambda_p)$ , y  $VV'=V''V=I$ , donde los  $\lambda_j$  son los valores propios de la matriz de correlación  $X^{*'}X^*$ . Luego, la  $j$ -ésima componente principal  $Z_j$  tiene desviación estándar igual a  $\sqrt{\lambda_j}$  y puede ser escrita como

$$Z_j = v_{j1}X_1^* + v_{j2}X_2^* + \dots + v_{jp}X_p^* \quad (12)$$

donde  $v_{j1}, v_{j2}, \dots, v_{jp}$  son los elementos de la  $j$ -ésima fila de  $V$ . La matriz  $V$  es llamada la matriz de cargas ("loadings"), y contiene los coeficientes de las variables en cada componente principal. Los valores calculados de las componentes principales  $Z_j$  son llamados los valores rotados o simplemente "scores".

El número máximo de componentes principales que se puede construir es igual a  $p$  el número de variables predictoras. Sin embargo usando solo algunas de ellas se consiguen buenos resultados. Decidir acerca del número de componentes principales que se deben usar es un gran problema.

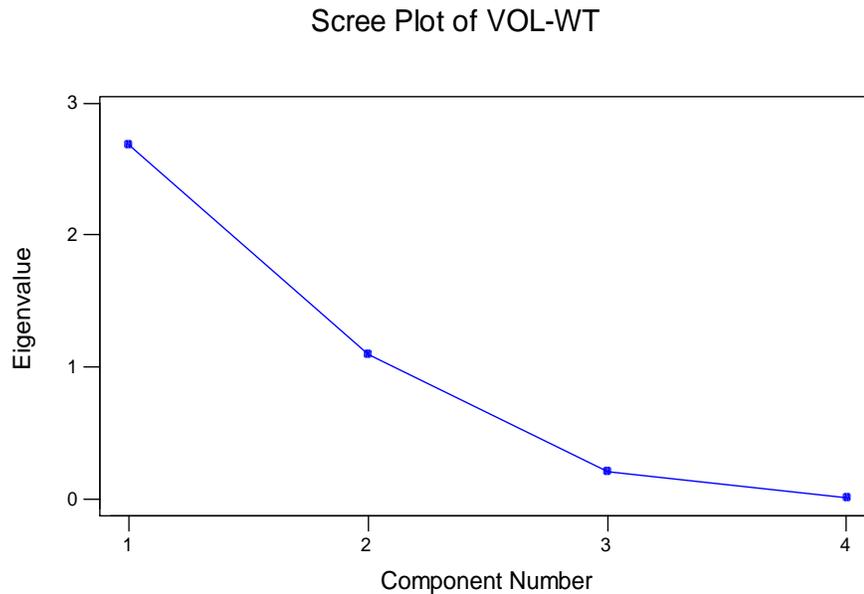


Figura 7.3. Scree Plot para elegir el número de componentes principales en el conjunto de datos millaje

#### Elección del número de componentes principales:

Por lo general se usan las siguientes dos alternativas:

- Elegir el número de componentes hasta donde se ha acumulado por lo menos 75% de la proporción de los valores propios.
- Elegir hasta la componente cuyo valor propio sea mayor que 1. Para esto se puede ayudar del “Scree Plot”. (Ver figura)

Las funciones **prcomp** y **princomp** de la librería **mva** de R calculan los componentes principales.

#### Ejemplo 5. Hallar los componentes principales del conjunto de datos de **millaje**

Para el conjunto millaje se obtiene los siguientes resultados

```
> pc<-prcomp(millaje[,2:5],scale=T,retx=T)
```

```
> pc
```

Standard deviations:

```
[1] 1.63994544 1.04901826 0.45331794 0.06813577
```

Rotation:

	PC1	PC2	PC3	PC4
sp	0.5616362	-0.2921753	0.5229999	0.57067460
wt	0.5526927	0.2365918	-0.7530524	0.26733338
vol	0.1387771	0.9134537	0.3820731	-0.01906127

```
hp 0.5998637 -0.1557567 0.1157728 -0.77620876
> summary(pc)
Importance of components:
      PC1  PC2  PC3  PC4
Standard deviation  1.640 1.049 0.4533 0.06814
Proportion of Variance 0.672 0.275 0.0514 0.00116
Cumulative Proportion 0.672 0.947 0.9988 1.00000
```

De acuerdo al screeplot de la figura habría que usar solamente dos componentes principales. Componentes principales se puede aplicar en regresión aunque con reservas ya que ellos no han usado para nada la variable de respuesta en sus cálculos. Hay dos opciones

- La opción clásica: Hacer la regresión solamente con las componentes principales que no han sido eliminadas, o
- La opción moderna: Hacer la regresión con el número de componentes principales determinadas por un proceso de selección de variables para elegir el mejor modelo.

Haciendo la regresión de  $Y^*$  versus todas las componentes principales se obtiene el modelo estimado

$$\hat{Y}^* = \hat{\alpha}_1 Z_1 + \dots + \hat{\alpha}_p Z_p \quad (13)$$

En forma matricial, el modelo  $Y=XB+e$  se transforma en el modelo  $Y=ZA+e'$  donde  $Z=XV$  y  $A=V'B$ . Se puede mostrar que la varianza del vector de coeficientes  $A$  está dado por  $\Lambda^{-1}\sigma^2$ , donde  $\Lambda$  es una matriz diagonal cuyos elementos son los valores propios. Es decir, la varianza estimada del  $j$ -ésimo coeficiente de regresión con los componentes principales viene dada por  $\text{Var}(\alpha_j)=s^2/\lambda_j$ .

**Ejemplo 6.** Aplicar regresión usando componentes principales a los datos de millaje

#### A) Regresión con las variables originales:

```
> summary(lm(mpg~.,data=millaje))
```

Call:

```
lm(formula = mpg ~ ., data = millaje)
```

Residuals:

```
  Min   1Q Median   3Q   Max
-9.0108 -2.7731 0.2733 1.8362 11.9854
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775  23.53161  8.178 4.62e-12 ***
sp          -1.29482   0.24477 -5.290 1.11e-06 ***
wt          -1.85980   0.21336 -8.717 4.22e-13 ***
vol         -0.01565   0.02283 -0.685  0.495
hp           0.39221   0.08141  4.818 7.13e-06 ***
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.653 on 77 degrees of freedom  
 Multiple R-squared: 0.8733, Adjusted R-squared: 0.8667  
 F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

### B) Regresión con las variables originales estandarizadas:

```
> mpgesca=scale(millaje,center=T,scale=T)
> summary(lm(mpg~.-1,data=as.data.frame(mpgesca)))

Call:
lm(formula = mpg ~ . - 1, data = as.data.frame(mpgesca))

Residuals:
    Min       1Q   Median       3Q      Max
-0.90066 -0.27718  0.02732  0.18354  1.19799

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
sp    -1.81681      0.34124   -5.324 9.49e-07 ***
wt    -1.51345      0.17251   -8.773 2.98e-13 ***
vol   -0.03466      0.05025   -0.690  0.492
hp     2.22834      0.45957    4.849 6.21e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3628 on 78 degrees of freedom
Multiple R-squared: 0.8733, Adjusted R-squared: 0.8668
F-statistic: 134.4 on 4 and 78 DF, p-value: < 2.2e-16
```

**Importante:** Cuando se estandarizan las variables predictoras y la variable de respuesta. Los coeficientes estandarizados son iguales a  $\frac{s_j}{s} \hat{\beta}$ , donde  $s$  es la desviación estándar de las  $y$ 's, y  $s_j$  es la desviación estándar de la  $j$ -ésima variable. Los valores de la prueba  $t$  y  $F$  no cambian y la varianza estimada  $s^2$  cambia a  $s^2/S_{yy}$ . En este ejemplo,  $s^2=13.3/8107.5=0.00164046$

### c) Regresión con todas las componentes principales sin estandarizar la respuesta.

```
> mpgpc=cbind(millaje[,1],pc$x)
> mpgpc=as.data.frame(mpgpc)
> summary(lm(V1~.,data=mpgpc))

Call:
lm(formula = V1 ~ ., data = mpgpc)

Residuals:
    Min       1Q   Median       3Q      Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 33.7817 0.4034 83.747 < 2e-16 ***
PC1 -5.2521 0.2475 -21.222 < 2e-16 ***
PC2 2.0608 0.3869 5.327 9.62e-07 ***
PC3 4.3445 0.8953 4.853 6.23e-06 ***
PC4 31.7186 5.9567 5.325 9.69e-07 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.653 on 77 degrees of freedom  
 Multiple R-squared: 0.8733, Adjusted R-squared: 0.8667  
 F-statistic: 132.7 on 4 and 77 DF, p-value: < 2.2e-16

```
> vif=diag(solve(cor(mpgpc[,2:5])))
> vif
PC1 PC2 PC3 PC4
  1  1  1  1
>
```

Observese que los VIF son iguales a 1. Lo que indica que ya no hay multicolinealidad. El  $R^2$  de este modelo y del modelo con las variables originales también es el mismo, ya que el coeficiente de determinación es invariante a transformaciones lineales.

Si sustituimos los componentes principales PC1-PC4 por sus respectivas ecuaciones como en la ecuación (12) se obtendrá la ecuación original. Así por ejemplo, el coeficiente  $-1.29$  de la variable SP resulta de la siguiente ecuación

$$[5.252*(-0.562)+2.06*(0.292)-4.344*(-0.53)-31.719*(0.571)]/14.038$$

aquí 14.038 es la desviación estándar de la variable SP.

#### d) Regresión con solo las dos primeras componentes principales, sin estandarizar la respuesta.

```
> summary(lm(V1~PC1+PC2,data=mpgpc))
```

Call:

```
lm(formula = V1 ~ PC1 + PC2, data = mpgpc)
```

Residuals:

```
Min 1Q Median 3Q Max
-8.290 -3.250 -0.540 2.318 18.740
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.7817 0.5153 65.56 < 2e-16 ***
PC1 -5.2521 0.3161 -16.61 < 2e-16 ***
PC2 2.0608 0.4942 4.17 7.75e-05 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.666 on 79 degrees of freedom  
 Multiple R-squared: 0.7879, Adjusted R-squared: 0.7825  
 F-statistic: 146.7 on 2 and 79 DF, p-value: < 2.2e-16

Aquí si no se obtiene la ecuación original al substituir los componentes principales. Por ejemplo el coeficiente de SP será

$$[5.252*(-0.562)+2.06*(0.292)]/14.038=-0.167$$

Similarmente para VOL, HP y WT se obtienen:

$$[5.252*(-0.139)+2.06*(-.913)]/22.166=-0.117$$

$$[5.252*(-0.600)+2.06*(0.156)]/56.841=-0.049$$

$$[5.252*(-0.553)+2.06*(-.237)]/8.1414=-0.416$$

Y el intercepto dará

$$33.8+5.25*((0.562*112.41)/14.038+(0.139*98.805)/22.166+(0.6*117.13)/56.841+(0.553*30.915)/8.1414)+2.06*((-0.292*112.41)/14.038+(0.913*98.805)/22.166+(-0.156*117.13)/56.841+(0.237*30.915)/8.1414)=82.953$$

Aquí los coeficientes de regresión son estimaciones sesgadas de los coeficientes de regresión poblacional.

Los coeficientes de la regresión con los componentes principales no son fáciles de interpretar porque en este caso las predictoras son combinaciones lineales de las predictoras originales. Lo que hay que hacer es tratar de expresar la regresión en términos de las variables originales para luego hacer la interpretación.

Aplicando el procedimiento de selección hacia adelante (“Forward”) se obtiene los siguientes resultados

```
> selforw(mpgpc[,2:5],mpgpc[,1],.15)
Loading required package: leaps
Selección Forward
```

p=numero de coeficientes en el modelo, p=1 es por el intercepto  
 nvar=p-1=numero de variables predictoras  
 add.var=la variable que ha sido anadida al modelo actual  
 pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

```
  p nvar add.var pvmax  s  r2 r2adj  Cp
2 2  1  PC1 0e+00 5.122 0.741 0.738 79.273
3 3  2  PC2 1e-04 4.666 0.788 0.782 52.902
4 4  3  PC4 1e-04 4.147 0.835 0.828 26.547
5 5  4  PC3 1e-04 3.653 0.873 0.867  5.000
>
```

El mejor modelo según el método “forward” es aquel que incluye todas las componentes principales. Esto no debe ser tomado como regla general como se verá en el siguiente ejemplo. Notar sin embargo que el método clásico recomendada la regresión con solamente dos componentes principales.

**Ejemplo 7.** Aplicar regresión usando componentes principales a los datos de pollution.

```
> pc<-prcomp(pollution[,1:15],scale=T,retx=T)
```

> pc

Standard deviations:

[1] 2.12800179 1.65977153 1.43334017 1.16120178 1.10599275 0.98002244  
 [7] 0.78277810 0.68703109 0.60897703 0.46518242 0.40786076 0.35637776  
 [13] 0.33761928 0.21456873 0.06975877

Rotation:

	PC1	PC2	PC3	PC4	PC5	PC6
V1	0.34547927	0.10264363	-0.02681399	0.33283614	-0.12232216	0.18274943
V2	0.06525347	-0.48216048	0.10601047	0.32881026	0.08515790	0.07812527
V3	0.34448633	-0.19541410	0.07810160	0.02480357	-0.39821611	-0.11519796
V4	-0.16298426	0.36487164	-0.15617683	0.52026608	-0.03511173	-0.13982856
V5	0.29727366	0.06598555	-0.03197863	-0.55971905	0.23078383	0.02608699
V6	-0.28650500	-0.17285588	0.42939325	-0.11014376	-0.13509909	0.03026891
V7	-0.36076129	-0.05042984	0.05492805	-0.14835149	-0.19310286	0.16212540
V8	-0.07150671	0.02139797	-0.44022545	0.05612602	-0.41769236	0.38843232
V9	0.30201160	-0.36887764	-0.06118170	-0.10245806	0.01779670	0.27224559
V10	-0.19645501	-0.24022842	0.33332901	0.04563508	-0.38386153	0.19059877
V11	0.35785982	-0.26832700	-0.02372060	0.27904384	0.13303123	-0.07130608
V12	-0.28277142	-0.36714007	-0.23942847	0.05276707	0.21682314	-0.22451250
V13	-0.26496868	-0.36339126	-0.31021837	0.04325345	0.20245716	-0.18868002
V14	-0.06750489	-0.09517245	-0.51943188	-0.17778397	-0.10168888	0.30101327
V15	-0.11330703	0.08216240	0.19118053	0.18035659	0.52057896	0.67584786
	PC7	PC8	PC9	PC10	PC11	PC12
V1	-0.012230021	0.48626851	-0.51151942	-0.04319658	0.11633732	-0.17601926
V2	-0.361931107	0.06880509	0.23356631	-0.38821397	-0.06182071	-0.24101083
V3	-0.091622038	0.12678624	0.28233138	0.39809622	0.53750325	-0.09978717
V4	0.209995070	0.12454673	-0.12808606	-0.06730038	-0.04328775	-0.05467298
V5	-0.035724146	0.01805908	-0.28659385	-0.36302122	0.15102108	-0.37866246
V6	0.151159442	0.07790867	-0.19969300	0.41776749	-0.35046171	-0.54147683
V7	-0.495214019	0.48593563	-0.01830232	-0.13397384	-0.14977007	0.24312314
V8	-0.313594638	-0.52825093	-0.20656213	0.06686120	-0.04282252	-0.21038293
V9	0.158490150	0.04544805	-0.31499478	0.24095995	-0.24032037	0.52782208
V10	0.469359106	-0.17048776	-0.11833883	-0.48350232	0.26923221	0.12389839
V11	0.095532450	-0.16036667	0.10332675	-0.07357918	-0.42073670	-0.11595861
V12	-0.016119010	-0.01245514	-0.24925081	0.07037528	0.26548752	0.00712984
V13	0.069695675	0.05738475	-0.21192889	0.09671983	0.20129628	-0.08334923
V14	0.434739103	0.37759030	0.42183348	-0.05049111	-0.11492698	-0.20487054
V15	0.000956947	-0.05265114	0.11737073	0.20389626	0.30082115	-0.04085240
	PC13	PC14	PC15			
V1	0.30468163	2.695860e-01	0.010001781			
V2	0.20427995	-4.311370e-01	0.006663022			
V3	-0.31350645	-6.884818e-02	0.005120874			
V4	-0.47526602	-4.595407e-01	0.044591308			
V5	-0.35047575	-1.844467e-01	0.021337034			
V6	-0.01563165	-1.022079e-01	0.048169108			
V7	-0.39963813	1.932746e-01	-0.021202610			
V8	-0.03883112	-9.330356e-05	0.001478613			
V9	-0.08284538	-3.951190e-01	0.015575980			
V10	-0.09162586	1.127369e-01	-0.027259184			
V11	-0.45365318	5.013989e-01	0.011177279			

```

V12 -0.02784298 1.198508e-01 0.688877704
V13 -0.04530960 6.280479e-02 -0.712151167
V14 0.07094674 2.958961e-03 0.108311401
V15 -0.16750266 7.181622e-02 -0.007742795
>
> summary(pc)
Importance of components:
      PC1  PC2  PC3  PC4  PC5  PC6  PC7  PC8  PC9  PC10  PC11  PC12
PC13  PC14  PC15
Standard deviation      2.128 1.660 1.433 1.1612 1.1060 0.980 0.7828 0.6870 0.6090 0.4652
0.4079 0.35638 0.3376 0.21457 0.06976
Proportion of Variance 0.302 0.184 0.137 0.0899 0.0815 0.064 0.0408 0.0315 0.0247 0.0144
0.0111 0.00847 0.0076 0.00307 0.00032
Cumulative Proportion 0.302 0.486 0.623 0.7124 0.7940 0.858 0.8988 0.9303 0.9550 0.9695
0.9805 0.98901 0.9966 0.99968 1.00000
> pollupc=cbind(pollution[,16],pc$x)
> pollupc=as.data.frame(pollupc)
> summary(lm(V1~.,data=pollupc))

```

Call:

```
lm(formula = V1 ~ ., data = pollupc)
```

Residuals:

```

  Min      1Q  Median      3Q      Max
-68.0663 -18.0171  0.9117 19.2237 86.9612

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) 940.3584    4.5093 208.540 < 2e-16 ***
PC1          15.5877    2.1369   7.295 4.24e-09 ***
PC2         -3.2915    2.7397  -1.201 0.23603
PC3        -19.8282    3.1725  -6.250 1.45e-07 ***
PC4         -2.7007    3.9160  -0.690 0.49404
PC5         -0.7183    4.1115  -0.175 0.86212
PC6         21.0846    4.6400   4.544 4.28e-05 ***
PC7         17.3389    5.8092   2.985 0.00462 **
PC8          7.0209    6.6188   1.061 0.29459
PC9        -18.6637    7.4671  -2.499 0.01624 *
PC10         13.3772    9.7753   1.368 0.17811
PC11        -14.6239   11.1492  -1.312 0.19644
PC12         29.2219   12.7598   2.290 0.02686 *
PC13         15.4357   13.4687   1.146 0.25797
PC14          4.6223   21.1928   0.218 0.82835
PC15        -87.1199   65.1861  -1.336 0.18826
---

```

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 34.93 on 44 degrees of freedom
Multiple R-squared: 0.7649, Adjusted R-squared: 0.6847
F-statistic: 9.542 on 15 and 44 DF, p-value: 2.193e-09

```

```
> summary(lm(V1~PC1+PC2+PC3+PC4+PC5,data=pollupc))
```

Call:

```
lm(formula = V1 ~ PC1 + PC2 + PC3 + PC4 + PC5, data = pollupc)
```

Residuals:

```
   Min      1Q  Median      3Q      Max
-115.008 -27.079  -5.946  36.286 111.809
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 940.3584    5.9149 158.980 < 2e-16 ***
PC1          15.5877    2.8030   5.561 8.53e-07 ***
PC2          -3.2915    3.5938  -0.916  0.364
PC3         -19.8282    4.1615  -4.765 1.47e-05 ***
PC4          -2.7007    5.1368  -0.526  0.601
PC5          -0.7183    5.3932  -0.133  0.895
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.82 on 54 degrees of freedom

Multiple R-squared: 0.5035, Adjusted R-squared: 0.4575

F-statistic: 10.95 on 5 and 54 DF, p-value: 2.621e-07

```
> selforw(pollupc[,2:16],pollupc[,1],.15)
```

Loading required package: leaps

Seleccion Forward

p=numero de coeficientes en el modelo, p=1 es por el intercepto

nvar=p-1=numero de variables predictoras

add.var=la variable que ha sido anadida al modelo actual

pvmax=p-value de F-parcial correspondiente a la variable mas importante en cada paso

```
  p nvar add.var pvmax   s  r2 r2adj  Cp
2 2  1  PC1 0.0000 53.076 0.284 0.272 77.927
3 3  2  PC3 0.0000 45.060 0.493 0.475 40.864
4 4  3  PC6 0.0002 40.210 0.603 0.582 22.216
5 5  4  PC7 0.0083 38.061 0.651 0.626 15.307
6 6  5  PC9 0.0204 36.528 0.684 0.655 11.060
7 7  6  PC12 0.0271 35.196 0.712 0.680  7.815
>
```

Notar que la mejor regression se consigue con seis componentes principales: PC1,PC3, PC6,PC7, PC9 y PC 12, y no por las primeras seis componentes principales.

## Ejercicios

- Probar que el estimador Ridge  $\hat{\beta}_R$  es un estimador sesgado del vector de coeficientes de regresión  $\beta$ .
  - Descomponer el error cuadrático medio total  $\sum_{j=1} E(\hat{\beta}_{jR} - \beta_j)^2$  en la suma de la varianza total de los estimados ridge más su sesgo al cuadrado. Encontrar explícitamente ambos términos.
- Hallar la varianza del estimador de regresión Ridge.
- Probar que el estimador ridge es la solución del problema de optimización

$$\text{Min } (y - XB)'(y - XB)$$

$$\text{Sujeto a que: } B'B \leq k^2$$

- Aplicar 5 diagnósticos de multicolinealidad al conjunto de datos BERKELEY disponible en la página de internet del curso, y aplicar regresión ridge para resolver el problema creado. Evaluar el efecto
- Considerando el conjunto de datos CRIMEN, disponible en la página de internet del curso
  - Hallar los componentes principales para su conjunto de predictoras
  - Hallar la regresión usando los más importantes componentes principales
  - Hallar la regresión usando "forward". Comparar sus resultados con el resultado anterior.

# CAPÍTULO 8

## REGRESIÓN ROBUSTA

### 8.1 Introducción

El método de mínimos cuadrados de ajustar una tendencia lineal es una de las herramientas más usadas desde que fue introducida en los inicios de los 1800. Sin embargo es un hecho reconocido por muchos que la presencia de datos “outliers” (verticales u horizontales) tiene un gran influencia en el ajuste por mínimos cuadrados. Un “outlier” mientras más exagerado sea hará que el ajuste lineal tienda a pasar cerca de él y el análisis de los residuales no sería muy confiable ya que estos darían la impresión de que nada malo estuviera pasando. Cuando hay presente datos influyentes las alternativas a mínimos cuadrados son las siguientes:

0. Seguir usando mínimos cuadrados como si nada hubiese pasado.
1. Hacer un análisis exploratorio de datos antes de ajustar el modelo. Esto incluye el uso de diagnósticos, transformaciones, gráficas dinámicas, etc.
2. Usar modelos lineales generalizados.
3. Usar métodos de regresión robusta, las cuales son modificaciones de los mínimos cuadrados y tiene como objetivo ajustar un modelo que resista la influencia de los “outliers”.

Los orígenes de la regresión robusta se remontan a 1973, cuando se publicó en el *Annals of Statistics* el artículo de Peter Huber. Existen una variedad de métodos robustos de regresión que son agrupados en tres:

- a) M-Regresión (M es por Máxima verosimilitud),
- b) R-Regresión (R es por Rangos) y
- c) L-Regresión (L es por combinación lineal de estadísticos de orden).

Existen también varias modificaciones de éstos. Los que han alcanzado más popularidad han sido los estimadores M de Regresión Robusta. En este texto solamente se discutirán los estimadores M y sus variantes.

Actualizando los comentarios hechos por Hogg en 1979. con la gran cantidad de métodos robustos existentes uno podría pensar que el uso de regresión robusta debería ser bastante amplio, pero este no ha sido el caso. Los principales programas estadísticos, excepto por S-Plus prácticamente ignoran regresión robusta. Nada más en la regresión M existen tantas maneras de elegir la función peso, tantas maneras de elegir las constantes de afinamiento y tantas posibilidades de elegir la constante de escala (variabilidad) que se le hace bien difícil al simple usuario elegir la estimación más adecuada. Para complicar más la cosa el cálculo de los estimadores se hace iterativamente y la elección del punto inicial y del algoritmo iterativo es crucial para alcanzar rápidamente convergencia. Finalmente, se ha avanzado muy poco en la parte inferencial

Después de los textos de Hampel, et al. (1986) y de Rousseeuw y Leroy (1987) no ha aparecido ningún otro texto importante en regresión robusta. Muchos de los investigadores pioneros de la regresión robusta se han movido a otras áreas y desde los mediados de los 90's la investigación en esta área ha disminuido. Con el desarrollo de las computadoras la investigación se ha enfocado más ahora en regresión no paramétrica, que se verá en el próximo capítulo.

## 8.2 Regresión $L_1$

Consideremos el modelo de regresión lineal múltiple  $y = \mathbf{X}\boldsymbol{\beta} + e$ . En mínimos cuadrados para estimar el vector de parámetros  $\boldsymbol{\beta}$  se debe minimizar  $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2$  donde  $\mathbf{x}'_i$  es una fila de la matriz  $\mathbf{X}$ .

En este caso si el residual es pequeño se le da poco peso y si es grande se le da mayor peso siguiendo una ecuación cuadrática. En 1887, Edgeworth propuso reemplazar los errores cuadráticos por errores en valor absoluto. Más específicamente, la ecuación de regresión se

obtendría minimizando  $\sum_{i=1}^n |e_i|$ . Esta regresión es llamada **regresión  $L_1$  o regresión de suma**

**absoluta mínima**. Sin embargo es usada muy restringidamente por las siguientes razones.

- El vector de coeficientes estimados no es único.
- La regresión  $L_1$  resiste la presencia de valores anormales en la dirección vertical. Pero es poco efectivo para valores anormales en la dirección  $x$ .
- La eficiencia del estimador disminuye a medida que aumenta el número de casos.
- Para obtener las estimaciones del coeficiente de regresión hay que resolver un problema de programación lineal, el cual es muy lento computacionalmente.

S-Plus tiene una función **lfit** que calcula la regresión  $L_1$ .

**Ejemplo 1.** El siguiente conjunto de datos llamado **Kootenay**, es clásico en los libros de regresión robusta. Aquí se mide el caudal del río Kootenay entre 1971 y 1973 en dos lugares uno llamado: Libby y otro que viene más adelante llamado Newgate. La idea es predecir el caudal en Newgate una vez que se observa lo que pasa en Libby. Ajustaremos una línea por mínimos cuadrados y una línea de regresión  $L_1$  usando S-Plus.

	Libby	Newgate
1	27.1	19.7
2	20.9	18.0
3	33.4	26.1
4	77.6	15.7
5	37.0	26.1
6	21.6	19.9
7	17.6	15.7
8	35.1	27.6
9	32.6	24.9
10	26.0	23.4
11	27.6	23.1
12	38.7	31.1
13	27.8	23.8

En la siguiente gráfica se muestra el diagrama de puntos de los datos

Notar el punto en la parte derecha inferior que está bien alejado de la mayor parte de los datos.

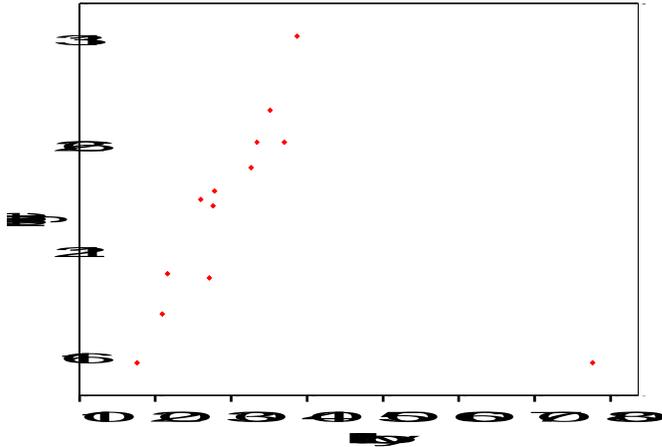


Figura 8.1. Scatterplot del conjunto de datos Kootenay

Haciendo la regresión por mínimos cuadrados (**lsfit**) y la regresión  $L_1$  (**l1fit**) se obtiene

S-PLUS : Copyright (c) 1988, 2001 Insightful Corp.

S : Copyright Lucent Technologies, Inc.

Professional Edition Version 6.0.3 Release 2 for Microsoft Windows : 2001

Working data will be in C:\Program Files\Insightful\splus6\users\Administrator

> Kootenay

Libby Newgate

1 27.1 19.7

2 20.9 18.0

3 33.4 26.1

4 77.6 15.7

5 37.0 26.1

6 21.6 19.9

7 17.6 15.7

8 35.1 27.6

9 32.6 24.9

10 26.0 23.4

11 27.6 23.1

12 38.7 31.1

13 27.8 23.8

> koo.ls<-lsfit(Kootenay\$Libby,Kootenay\$Newgate)

> koo.ls

\$coef:

Intercept	X
23.16443	-0.01427324

\$residuals:

[1] -3.0776245 -4.8661186 3.4122969 -6.3568256 3.4636806

[6] -2.9561273 -7.2132203 4.9365615 2.2008784 0.6066749

[11] 0.3295121 8.4879451 1.0323668

\$intercept:

[1] T

\$qr:

\$qr\$qt:

```
[1] -81.8460140 -0.7407183  3.9430373 -8.8018133
[5]  3.7520540 -1.6309619 -5.6187582  5.3528507
[9]  2.7854780  1.6356141  1.2507326  8.6618674
[13]  1.9401224
```

\$qr\$qr:

	Intercept	X
[1,]	-3.6055513	-117.31909150
[2,]	0.2773501	51.89557562
[3,]	0.2773501	-0.03935572
[4,]	0.2773501	-0.89106609
[5,]	0.2773501	-0.10872579
[6,]	0.2773501	0.18802397
[7,]	0.2773501	0.26510184
[8,]	0.2773501	-0.07211381
[9,]	0.2773501	-0.02394015
[10,]	0.2773501	0.10323833
[11,]	0.2773501	0.07240718
[12,]	0.2773501	-0.14148389
[13,]	0.2773501	0.06855329

\$qr\$ql:

```
[1] 1.277350 1.201513
```

\$qr\$rank:

```
[1] 2
```

\$qr\$pivot:

```
[1] 1 2
```

\$qr\$tol:

```
[1] 1e-007
```

```
> koo.l1<-l1fit(Kootenay$Libby,Kootenay$Newgate)
```

```
> koo.l1
```

\$coefficients:

Intercept	X
14.29149	0.319149

\$residuals:

```
[1] -3.2404251 -2.9617021  1.1489357 -23.3574467
[5]  0.0000000 -1.2851073 -4.2085114  2.1063848
[9]  0.2042551  0.8106378  0.0000000  4.4574466
[13]  0.6361693
```

```
># Los siguientes comandos sirven para mostrar el diagama de puntos y las lineas de
```

```
># regresion por minimos cuadrados y L1 simultaneamente
```

```
> plot(Kootenay$Libby,Kootenay$Newgate)
> abline(koo.ls)
> abline(koo.l1)
```

De la salida se obtiene que la regresión por mínimos cuadrados es

Newgate= 23.16443 -0.01427324 Libby

En tanto que la regresión L1 será

Newgate= 14.29149 +0.319149 Libby

A continuación se muestra el diagrama de puntos y las dos líneas de regresión.

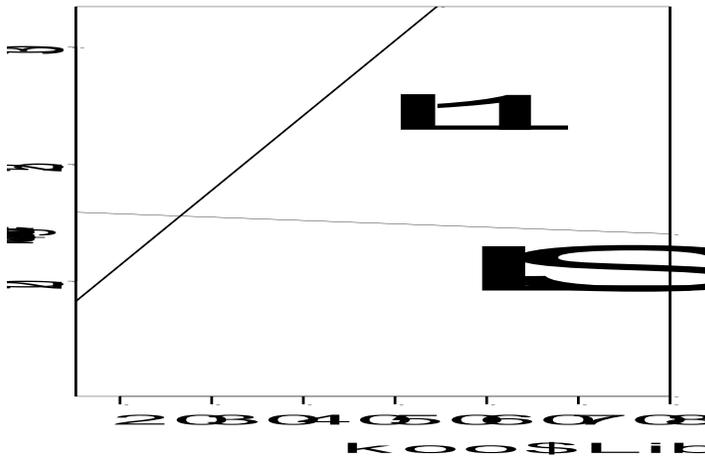


Figura 8.2. Plot que muestra la regresión mínimo cuadrática y la regresión L1.

En la gráfica se puede ver el gran efecto del punto leverage sobre la línea de mínimos cuadrados.

### 8.3. Regresión M

En 1973, Huber propuso un nuevo método de regresión que era como una combinación de la regresión L1 y la regresión por mínimos cuadrados. La idea se basaba en que para residuales pequeños se le da un peso cuadrático y para residuales grandes se le da un peso lineal. Más específicamente, la propuesta de Huber era minimizar con respecto a B

$$\sum_{i=1}^n \rho(e_i) = \sum_{i=1}^n \rho(y_i - \mathbf{x}'_i \boldsymbol{\beta}) \quad (1)$$

donde la función peso  $\rho$  se define por

$$\rho(t) = \begin{cases} \frac{t^2}{2} & |t| \leq c \\ c|t| - \frac{c^2}{2} & |t| > c \end{cases} \quad (2)$$

$c$  es una constante de afinamiento que depende del nivel de eficiencia. Para una eficiencia del 95% se escoge  $c=1.345$ . Como se verá más adelante, la función peso de Huber  $\rho$  es la más usada entre las muchas funciones pesos existentes. Por otro lado, debido a que la varianza  $\sigma^2$  de la variable de respuesta afecta la estimación de  $\beta$ , la función a minimizar es la que sigue

$$\sum_{i=1}^n \rho\left(\frac{y_i - \mathbf{x}'_i \beta}{\sigma}\right) \quad (3)$$

$\rho$  es una función simétrica con mínimo en 0 y  $\sigma$  es desconocido. Derivando con respecto a  $\beta$  e igualando a cero se obtiene el sistema de  $p$  ecuaciones no lineales.

$$\sum_{i=1}^n \psi\left(\frac{y_i - \mathbf{x}'_i \beta}{\sigma}\right) \mathbf{x}'_i = \mathbf{0} \quad (4)$$

donde  $\psi = \rho'$ . En el caso de mínimos cuadrados  $\rho(t) = t^2/2$  y  $\psi(t) = t$ .

Se puede mostrar que también se llega a la ecuación (4) a través del método de máxima verosimilitud. En particular la propuesta de Huber es obtenida si se asume que los errores siguen una distribución normal contaminada con una doble exponencial, es decir, una mezcla de una normal y una doble exponencial.

El parámetro de escala (o de variabilidad)  $\sigma$  puede ser estimado de antemano y sustituido en (4). La MAD, *desviación absoluta con respecto a la mediana*, de los residuales es la medida que más se usa para estimar  $\sigma$ . Se define por

$$\text{MAD} = 1.4825 * \text{mediana}|e_i - \text{mediana}(e_i)| \quad (5)$$

Otra alternativa más complicada es estimar  $\sigma$  simultáneamente con  $\beta$ .

Existen varias propuestas para la función peso  $\psi$ , entre las principales están

**Andrews**  $\psi(u) = c * \sin(u/c)$  si  $|u| \leq \pi * c$  y 0 en otro caso. El valor default de  $c=1.339$ .

**Bisquare (o Biweight)**  $\psi(u) = u * (1 - (u/c)^2)^2$  if  $|u| \leq c$  y 0 en otro caso. El valor default de  $c=4.685$ .

**Cauchy**  $\psi(u) = u / (1 + (u/c)^2)$  con  $c=2.385$  como valor default.

**Fair**  $\psi(u) = u / (1 + |u|/c)$  y  $c=1.4$  es el valor default.

**Hampel** Tiene 3 constantes de afinamiento,  $a$ ,  $b$  y  $c$ . Se define por  $\psi(u) = u$  si  $|u| \leq a$ ;  $a * \text{sign}(u)$  si  $a < |u| \leq b$ ;  $a * (c - |u|) / (c - b) * \text{sign}(u)$  si  $b < |u| \leq c$ ; y 0 en otro caso. Los valores defaults para  $a$ ,  $b$  y  $c$  son 2, 4 y 8 respectivamente.

**Huber**  $\psi(u) = u$  si  $|u| < c$  y  $c \cdot \text{sign}(u)$  en otro caso. El valor default de  $c=1.345$ .

**Logistic**  $\psi(u) = c \cdot \tanh(u/c)$  con valor default  $c=1.205$ .

**Talworth**  $\psi(u) = u$  si  $|u| \leq c$  y  $0$  en otro caso. El default es  $c=2.795$ .

**Welsch**  $\psi(u) = u \cdot \exp(-(u/c)^2)$ . El valor default es  $c = 2.985$ .

La función **rreg** de S-Plus y la función **rlm** del package **MASS** de **R** permiten calcular los estimadores  $m$  de regresión

**Ejemplo 2.** Calcular la regresión de Huber para el conjunto de datos **kootenay**.

Usando S-Plus se obtiene los siguientes resultados

```
> koo.rob<-rreg(koo$Libby,koo$Newgate,method=wt.huber)
> koo.rob
$coefficients:
(Intercept)      x
 23.19752 -0.01872314

$residuals:
 [1] -2.990126 -4.806209  3.527830 -6.044607  3.595234 -2.893103 -7.167995  5.059660
 [2] 2.312852  0.689279
 [11] 0.419236  8.627063  1.122981

$fitted.values:
 [1] 22.69013 22.80621 22.57217 21.74461 22.50477 22.79310 22.86800 22.54034 22.58715
 [2] 22.71072 22.68076
 [12] 22.47294 22.67702

$w:
 [1] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 0.9799193 1.0000000
 [2] 1.0000000 1.0000000
 [11] 1.0000000 0.8149970 1.0000000

$int:
 [1] T

$convc:
 [1] 0.029915705 0.005059456 0.001065885

$status:
 [1] "converged"

> plot(koo$Libby,koo$Newgate)
> abline(koo.ls)
> abline(koo.rob)
> plot(koo$Libby,koo$Newgate)
> abline(koo.ls)
```

```
> abline(koo.rob)
```

la ecuación de regresión Huber será

$$\text{Newgate} = 23.19752 - 0.01872314 \text{Libby}$$

Ploteando la regresión minimocuadrática y la Huber se obtiene la siguiente gráfica

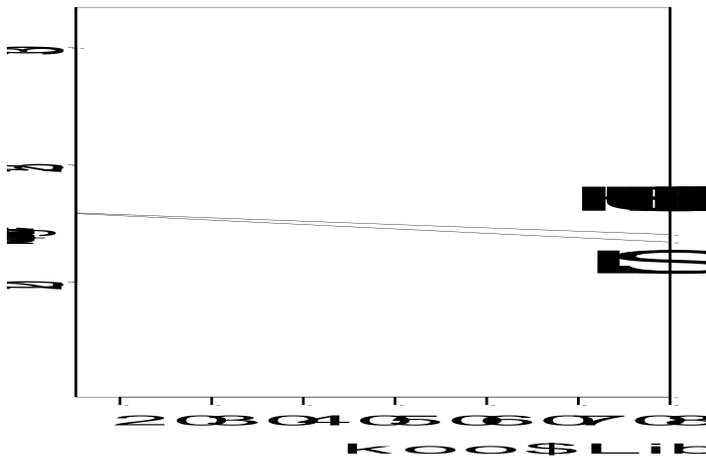


Figura 8.3. Plot que muestra la regresión la regresión mínimo cuadrática y la regresión Huber.

Notar que la regresión Huber también es afectada por el valor influyente.

En R todo lo anterior se ejecuta de la siguiente manera

```
R : Copyright 2001, The R Development Core Team
Version 1.3.1 (2001-08-31)
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type `license()' or `licence()' for distribution details.
```

```
> # asumiendo que los datos estan en el archivo koo.txt en el directorio de R
> uno<-read.table("koo.txt",header=T)
> uno
  Libby Newgate
1  27.1   19.7
2  20.9   18.0
3  33.4   26.1
4  77.6   15.7
5  37.0   26.1
6  21.6   19.9
7  17.6   15.7
8  35.1   27.6
9  32.6   24.9
10 26.0   23.4
```

```
11 27.6 23.1
12 38.7 31.1 lsfit(uno$Libby,uno$Newgate)
$coefficients
  Intercept      X
23.16442942 -0.01427324

$residuals
[1] -3.0776245 -4.8661186 3.4122969 -6.3568256 3.4636806 -2.9561273
[7] -7.2132203 4.9365615 2.2008784 0.6066749 0.3295121 8.4879451
[13] 1.0323668

$sintercept
[1] TRUE

$sqr
$sqr$sqr
[1] -81.8460140 -0.7407183 3.9430373 -8.8018133 3.7520540 -1.6309619
[7] -5.6187582 5.3528507 2.7854780 1.6356141 1.2507326 8.6618674
[13] 1.9401224

$sqr$sqr
  Intercept      X
[1,] -3.6055513 -117.31909150
[2,] 0.2773501 51.89557562
[3,] 0.2773501 -0.03935572
[4,] 0.2773501 -0.89106609
[5,] 0.2773501 -0.10872579
[6,] 0.2773501 0.18802397
[7,] 0.2773501 0.26510184
[8,] 0.2773501 -0.07211381
[9,] 0.2773501 -0.02394015
[10,] 0.2773501 0.10323833
[11,] 0.2773501 0.07240718
[12,] 0.2773501 -0.14148389
[13,] 0.2773501 0.06855329

$sqr$sqraux
[1] 1.277350 1.201513

$sqr$rank
[1] 2

$sqr$pivot
[1] 1 2

$sqr$tol
[1] 1e-07
> rlm(Newgate~.,data=uno,psi=psi.huber)
Call:
rlm.formula(formula = Newgate ~ ., data = uno)
Converged in 5 iterations
```

Coefficients:

```
(Intercept)  Libby
23.19413503 -0.01863951
```

Degrees of freedom: 13 total; 11 residual

Scale estimate: 5.23

```
> r1<-rlm(Newgate~.,data=uno)
> summary(r1)
```

Call: rlm.formula(formula = Newgate ~ ., data = uno)

Residuals:

```
  Min   1Q  Median   3Q   Max
-7.1661 -2.9890  0.6905  3.5284  8.6272
```

Coefficients:

```
      Value Std. Error t value
(Intercept) 23.1941  3.8521   6.0212
Libby      -0.0186  0.1083  -0.1722
```

Residual standard error: 5.231 on 11 degrees of freedom

Correlation of Coefficients:

```
(Intercept)
Libby -0.9145
> attributes(r1)
$names
[1] "coefficients" "residuals"  "effects"    "rank"
[5] "fitted.values" "assign"     "qr"         "df.residual"
[9] "w"           "s"          "psi"        "k2"
[13] "conv"        "converged" "x"          "call"
[17] "terms"       "xlevels"    "model"
```

\$class

```
[1] "rlm" "lm"
```

> r1\$residuals

```
  1    2    3    4    5    6    7
-2.9890044 -4.8045694  3.5284245 -6.0477094  3.5955267 -2.8915217 -7.1660797
  8    9   10   11   12   13
 5.0601116  2.3135129  0.6904921  0.4203153  8.6272139  1.1240432
```

**Ejemplo 3.** El siguiente conjunto de datos llamado **Stackloss**, es otro también bien usado en textos de regresión robusta, consiste de 21 observaciones de 4 variables; Stackloss(y), rate( flujo del aire) temp(temperatura del agua) y Acid ( concentración de acido ). A continuación se muestran los datos, la regresión robusta usando la función bisquare y el plot de residuales de la regresión mínimo cuadrática y de la regresión bisquare.

```
stackloss
```

```
stackloss rate Temp Acid
```

```
1  42 80 27 89
2  37 80 27 88
3  37 75 25 90
4  28 62 24 87
5  18 62 22 87
6  18 62 23 87
7  19 62 24 93
8  20 62 24 93
9  15 58 23 87
10 14 58 18 80
11 14 58 18 89
12 13 58 17 88
13 11 58 18 82
14 12 58 19 93
15  8 50 18 89
16  7 50 18 86
17  8 50 19 72
18  8 50 19 79
19  9 50 20 80
20 15 56 20 82
21 15 70 20 91
```

```
>stackloss<-as.matrix(stackloss)
```

```
>stackloss1<-stackloss[,2:4]
```

```
>stackloss2<-stackloss[,1]
```

```
> robtukey<-rreg(stackloss1, stackloss2, method=wt.bisquare)
```

```
> robtukey
```

```
$coefficients:
```

```
(Intercept)  rate  Temp  Acid
-42.28459 0.9273997 0.6510081 -0.1123065
```

```
$residuals:
```

```
[1] 2.5106733 -2.6016332 3.5619944 6.9322787 -1.7657050 -2.4167132 -1.3938823 -
0.3938823
[9] -1.7071145 -0.2382194 0.7725391 0.3112407 -3.0136064 -1.4292430 2.1917365
0.8548170
[17] -0.3684821 0.4176634 0.8789617 1.5391767 -10.4336602
```

```
$fitted.values:
```

```
[1] 39.489327 39.601633 33.438006 21.067721 19.765705 20.416713 20.393882 20.393882
16.707114 14.238219
[11] 13.227461 12.688759 14.013606 13.429243 5.808264 6.145183 8.368482 7.582337
8.121038 13.460823
[21] 25.433660
```

```
$w:
```

```
[1] 0.892416394 0.885432076 0.789930362 0.336110641 0.946231809 0.900393417
0.966212323 0.997263673
[9] 0.949531055 0.999032961 0.989517299 0.998266923 0.847699690 0.964641536
0.917808001 0.987286682
[17] 0.997611460 0.996972889 0.986607555 0.958974835 0.002377249
```

\$int:  
[1] T

\$conv:  
[1] 0.1623974286 0.0569754561 0.0822263512 0.0455994131 0.0419303121 0.0411232225  
0.0213469351 0.0036786480  
[9] 0.0008243514

\$status:  
[1] "converged"

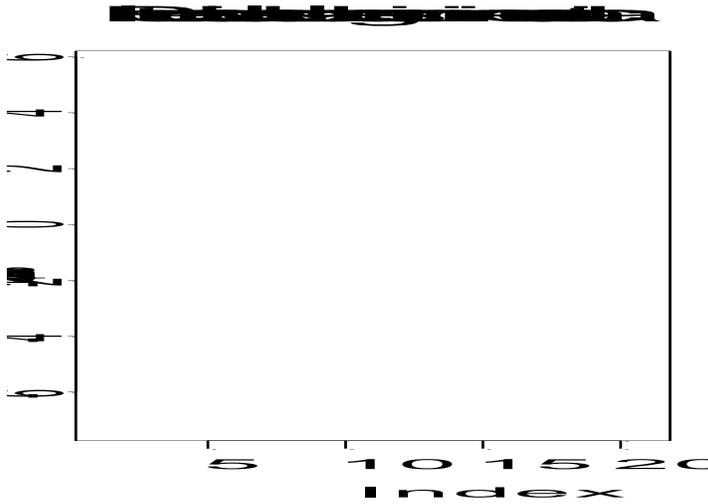


Figura 8.4. Plot de residuales de la regresión mínimo cuadrática.

Se notan como 4 outliers

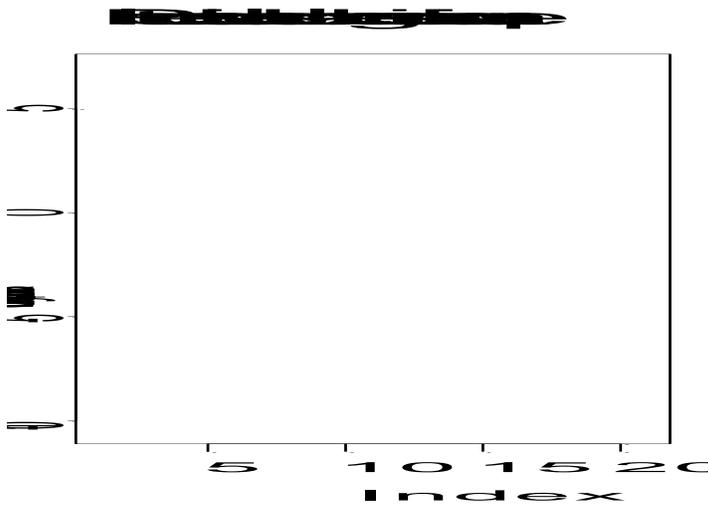


Figura 8.5. Plot de residuales de la regresión robusta bisquare.

Los 4 “outliers” se destacan un poco más en el plot de residuales correspondiente a la regresión bisquare, especialmente el que aparece en la parte inferior.

### 8.3.1 Cálculo de los estimadores M de regresión.

Los estimadores de regresión M se obtienen resolviendo el sistema de ecuaciones no lineales (4). La solución es obtenida mediante un proceso iterativo. Existen 3 propuestas para hacer esto:

a) *El método de Newton Raphson:* Cuya fórmula de iteración es:

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \sigma \left[ \sum_i \mathbf{x}_i' \psi' \left( \frac{y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}^{(m)}}{\tilde{\sigma}} \right) \mathbf{x}_i \right]^{-1} \mathbf{X}' \psi \left( \frac{\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{(m)}}{\tilde{\sigma}} \right) \quad (6)$$

b) *El método H de Huber:* Teniendo en cuenta que el estimador mínimo cuadrático  $\hat{\boldsymbol{\beta}}$  puede ser escrito como  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{e}$  entonces se puede deducir el siguiente proceso iterativo

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \hat{\mathbf{e}}^{(m)} \quad (7)$$

Huber (1977) sugirió el siguiente proceso iterativo

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \tilde{\sigma} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \psi \left( \frac{\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}^{(m)}}{\tilde{\sigma}} \right) \quad (8)$$

La idea aquí es reemplazar en cada iteración el residual  $\hat{\mathbf{e}} = \mathbf{y}_i - \mathbf{x}_i \hat{\boldsymbol{\beta}}$  por el residual modificado  $\psi(r_i / \hat{\sigma}) \hat{\sigma}$ .

c) *El método de Mínimos Cuadrados reponderados (IRLS)*

Beaton y Tukey(1974) se basaron en los mínimos cuadrados para proponer la siguiente fórmula iterativa

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \hat{\boldsymbol{\beta}}^{(m)} + \tilde{\sigma} (\mathbf{X}' \mathbf{W}^{(m)} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{(m)} (\mathbf{y} - \mathbf{X}' \boldsymbol{\beta}^{(m)}) \quad (9)$$

donde  $\mathbf{W}^{(m)}$  es una matriz diagonal cuyos elementos son:

$$w_i^{(m)} = \frac{\psi((y_i - x_i \beta^{(m)}) / \tilde{\sigma})}{(y_i - x_i \beta^{(m)}) / \tilde{\sigma}} \quad (10)$$

Este último procedimiento es el que implementa S-Plus y también puede programarse usando el procedimiento IML de SAS.

El valor inicial del proceso iterativo depende de la forma de la función  $\psi$ , si ésta es monótona se podría comenzar con los coeficientes mínimos cuadráticos, pero si no lo es se recomienda

comenzar con los coeficientes de la regresión  $L_1$  o de otra regresión robusta. Cuando se usa como valores iniciales los coeficientes de otra regresión robusta  $M$  entonces al estimador final se le llama estimador  $MM$ , nombre introducido por Yohai y Zammar (1990), propiamente el estimador inicial de escala también es estimado robustamente y en ese caso el estimador inicial  $M$  es llamado un estimador robusto  $S$ .

A continuación se muestran algunos resultados en R de estimadores robustos para varias funciones pesos y parámetros iniciales

```
> data(stackloss)
> r1<-rlm(stack.loss~.,data=stackloss)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss)
Converged in 9 iterations
```

```
Coefficients:
(Intercept) Air.Flow Water.Temp Acid.Conc.
-41.0265311  0.8293739  0.9261082 -0.1278492
```

```
Degrees of freedom: 21 total; 17 residual
Scale estimate: 2.44
> r1<-rlm(stack.loss~.,data=stackloss,psi=psi.huber)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.huber)
Converged in 9 iterations
```

```
Coefficients:
(Intercept) Air.Flow Water.Temp Acid.Conc.
-41.0265311  0.8293739  0.9261082 -0.1278492
```

```
Degrees of freedom: 21 total; 17 residual
Scale estimate: 2.44
> r1<-rlm(stack.loss~.,data=stackloss,psi=psi.hampel)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.hampel)
Converged in 4 iterations
```

```
Coefficients:
(Intercept) Air.Flow Water.Temp Acid.Conc.
-40.4747324  0.7410830  1.2250793 -0.1455251
```

```
Degrees of freedom: 21 total; 17 residual
Scale estimate: 3.09
> library(lqs)
> r1<-rlm(stack.loss~.,data=stackloss,method="MM",psi=psi.huber)
> r1
Call:
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.huber,
  method = "MM")
```

Converged in 11 iterations

Coefficients:

(Intercept) Air.Flow Water.Temp Acid.Conc.  
-41.5230431 0.9388404 0.5794524 -0.1129150

Degrees of freedom: 21 total; 17 residual

Scale estimate: 1.91

```
> r1<-rlm(stack.loss~.,data=stackloss,method="MM")
```

```
> r1
```

Call:

```
rlm.formula(formula = stack.loss ~ ., data = stackloss, method = "MM")
```

Converged in 12 iterations

Coefficients:

(Intercept) Air.Flow Water.Temp Acid.Conc.  
-41.7072712 0.9372710 0.5940633 -0.1129477

Degrees of freedom: 21 total; 17 residual

Scale estimate: 1.98

```
> r1<-rlm(stack.loss~.,data=stackloss,method="MM",psi=psi.bisquare)
```

```
> r1
```

Call:

```
rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.bisquare,  
method = "MM")
```

Converged in 11 iterations

Coefficients:

(Intercept) Air.Flow Water.Temp Acid.Conc.  
-41.5230726 0.9388402 0.5794546 -0.1129150

Degrees of freedom: 21 total; 17 residual

Scale estimate: 1.91

```
> summary(r1)
```

```
Call: rlm.formula(formula = stack.loss ~ ., data = stackloss, psi = psi.bisquare,  
method = "MM")
```

Residuals:

Min	1Q	Median	3Q	Max
-10.50957	-1.43820	-0.09084	1.02517	7.23167

Coefficients:

	Value	Std. Error	t value
(Intercept)	-41.5231	9.3070	-4.4615
Air.Flow	0.9388	0.1055	8.8983
Water.Temp	0.5795	0.2879	2.0125
Acid.Conc.	-0.1129	0.1223	-0.9234

Residual standard error: 1.912 on 17 degrees of freedom

Correlation of Coefficients:

```

      (Intercept) Air.Flow Water.Temp
Air.Flow  0.1793
Water.Temp -0.1489  -0.7356
Acid.Conc. -0.9016  -0.3389  0.0002
> attributes(r1)
$names
[1] "coefficients" "residuals"  "effects"    "rank"
[5] "fitted.values" "assign"      "qr"        "df.residual"
[9] "w"            "s"          "psi"       "k2"
[13] "conv"        "converged"  "x"         "call"
[17] "terms"       "xlevels"    "model"

$class
[1] "rlm" "lm"

```

### Algoritmo para calcular la regresión robusta usando IRLS:

- Calcular la MAD.
- Hacer la regresión por mínimos cuadrados y guardar los coeficientes como  $\hat{\beta}^{(0)}$  y los residuales como  $\hat{\mathbf{e}}_o$ .
- Estandarizar los residuales como  $\mathbf{e}_o^* = \frac{\hat{\mathbf{e}}_o}{MAD}$
- Calcular los pesos  $\mathbf{w}_o = \psi(\mathbf{e}_o^*) / \mathbf{e}_o^*$ .
- Calcular la regresión por mínimos cuadrados ponderados usando la matriz diagonal de pesos  $\mathbf{W}_o$  cuyos elementos en la diagonal son  $w_o$ . Guardar los coeficientes de regresión como  $\hat{\beta}^{(1)}$  y los residuales como  $\hat{\mathbf{e}}^{(1)}$ . Aquí.

$$\hat{\beta}^{(1)} = (\mathbf{X}' \mathbf{W}_o \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}_o \mathbf{y}$$

Sustituir el índice 0 por 1 y repetir los pasos c, d y e. Continuar el proceso hasta que la diferencia entre  $\hat{\beta}^{(n+1)}$  y  $\hat{\beta}^{(n)}$  sea despreciable.

De los tres algoritmos, el que converge más rápidamente es el de Newton-Raphson, pero tiene el problema que hay que calcular la derivada de  $\psi$ . El método de Huber es el que converge más lentamente pero tiene la ventaja de que  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  se calcula una sola vez.

### 8.4 Regresión GM o Regresión de Influencia acotada.

La *función influencia* introducida por Hampel en 1974, mide el efecto de una observación en el cálculo de un estimador, asumiendo que los datos siguen una distribución conocida  $F$ . En el caso de regresión la influencia (IF) puede ser escrita como el producto de la influencia en la dirección vertical (IR) por la diferencia en la dirección horizontal (IP). Por ejemplo, para mínimos cuadrados se obtiene que la influencia del estimador  $\hat{\beta}$  en el punto  $(\mathbf{y}, \mathbf{x})$ , asumiendo que la distribución  $F$  es la Normal está dada por

$$\text{IF}(y, \mathbf{x}, F, \hat{\boldsymbol{\beta}}) = (y - \mathbf{x}\hat{\boldsymbol{\beta}})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}' \quad (11)$$

Además  $\text{IR} = (y - \mathbf{x}\hat{\boldsymbol{\beta}})$  y  $\text{IP} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'$ . Tanto IP como IR no son acotados por lo tanto el estimador mínimo cuadrático no es robusto.

En el caso de los estimadores M se puede mostrar que IR es acotado pero no IP. El problema con la regresión M es que protege contra observaciones que son “outliers” verticales pero no contra aquellas observaciones que tienen un leverage alto. Para subsanar esta deficiencia se han propuesto modificaciones a los estimadores M, los cuales son llamados *estimadores M generalizados (GM)*. Básicamente hay 3 propuestas.

**Estimador de Mallows (1975):** Se obtiene modificando la ecuación (4) de la siguiente manera

$$\sum_{i=1}^n w(\mathbf{x}_i) \psi(r_i / \tilde{\sigma}) \mathbf{x}_i = \mathbf{0} \quad (12)$$

donde  $w$  representa una función peso

**Estimador de Schweppe (1975):** En este caso la ecuación (4) se modifica a:

$$\sum_{i=1}^n w(\mathbf{x}_i) \psi(r_i / (w(\mathbf{x}_i) \tilde{\sigma})) \mathbf{x}_i = \mathbf{0} \quad (13)$$

donde  $w(\mathbf{x}_i) = \sqrt{1 - h_{ii}}$ . Para calcular este estimador se puede usar mínimos cuadrados ponderados con  $w_i = \min(1, 2/|t_i|)$ .

**Estimador de Welsh (1982):** Usa la misma ecuación anterior pero con  $w(\mathbf{x}_i) = (1 - h_{ii}) / h_{ii}^{1/2}$ . Para calcular este estimador se puede usar mínimos cuadrados ponderados con  $w_i = \min(1, 2/|\text{DFFIT}_i|)$ , con  $c = 2(p/n)^{1/2}$ .

Se ha establecido que cuando el número de variables predictoras se incrementa, la resistencia de los estimadores GM a los puntos leverages se va deteriorando.

### 8.5 Regresión de Mediana de Cuadrados Mínima (LMS)

Fue introducida por Rousseeuw en 1984. Es bien conocido que la mediana, es una medida más resistente que la media cuando hay presente “outliers”. En efecto su “Breakdown Point” es 50%, mientras que el de la media es  $(1/n) * 100\%$ . Los mínimos cuadrados están relacionados al uso de la media (Si se minimiza suma de cuadrados con respecto a un punto se puede mostrar que ese punto es la media). La propuesta de Rousseeuw, consiste en minimizar la mediana de los cuadrados de los residuales. Es decir

$$\text{Min Mediana}(y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \quad (14)$$

Resulta ser que la regresión LMS resiste bastante bien los “outliers” verticales y horizontales. Esta regresión ha sido implementada en S-Plus, usando la función **lmsreg**.

Pero la desventaja de la regresión LMS es que no es eficiente aparte de ser computacionalmente pesada. Para resolver este problema, Rousseeuw (1984) introdujo la regresión por Sumas de Cuadrados podada mínima (LTS), en este caso se cuadrar los residuales, luego se los ordena y se

suma solamente los que no son muy grandes, se minimiza y el resultado dará la regresión LTS. Esta regresión ha sido implementada en S-Plus, usando la función **ltsreg**

**Ejemplo:** Calcular las regresiones LMS y LTS para los datos de Stackloss

```
> reglms<-lmsreg(stackloss1, stackloss2)
> reglms
$coefficients:
Intercept rate Temp      Acid
  -39.25 0.75  0.5 -2.586466e-017

$scale:
  Y
1.207615

$residuals:
[1] 7.75 2.75 7.50 8.75 -0.25 -0.75 -0.25 0.75 -0.75 0.75 0.75 0.25 -2.25 -1.75 0.75 -0.25
0.25
[18] 0.25 0.75 2.25 -8.25

$intercept:
[1] T

$method:
[1] "Least Median of Squares with 3143 samples of size 4 , 143 were singular."
```

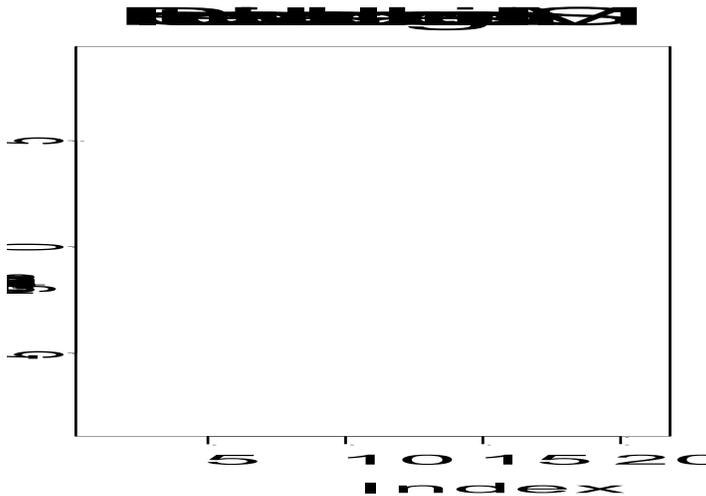


Figura 8.6. Plot de residuales de la regresión LMS.

Notar que los 4 “outliers” se destacan bastante.

```
reglts<-ltsreg(stackloss1, stackloss2)
> reglts
```

Method:  
Least Trimmed Squares Robust Regression.

Call:  
ltsreg.default(stackloss1, stackloss2)

Coefficients:  
Intercept rate Temp Acid  
-39.8935 0.9317 0.5205 -0.1145

Scale estimate of residuals: 2.041

Total number of observations: 21

Number of observations that determine the LTS estimate: 18

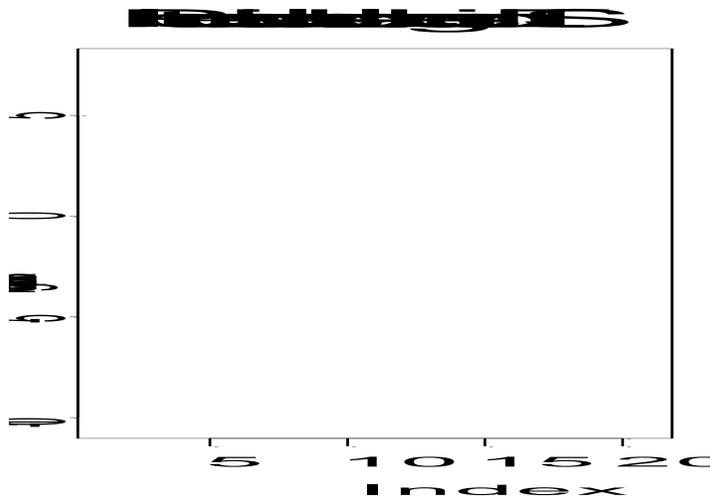


Figura 8.7. Plot de residuales de la regresión LTS.

Los “outliers” no se destacan tanto como en el plot anterior.

## EJERCICIOS

1. Usar el conjunto de datos Gessell, disponible en la página de internet del curso y realizar los siguientes análisis.
  - a) Calcular las regresiones L1, M con pesos Hampel o Bisquare, LMS y LTS. Hallar las sumas de cuadrados de residuales y compararlas con la suma de cuadrados residuales del ajuste mínimo cuadrático. Plotear todas las líneas halladas junto con la de mínimos cuadrados.
  - b) Plotear sus residuales y compararlos con los de la regresión mínimo cuadrática..
2. Usar el conjunto de datos Highway, disponible en la página de internet del curso y realizar los siguientes análisis.
  - a) Elegir una variable predictora adecuada (debería tener “outliers”) y calcular las regresiones L1, M con pesos Hampel o Bisquare, LMS y LTS. Hallar las sumas de cuadrados de residuales y compararlas con la suma de cuadrados residuales del ajuste mínimo cuadrático. Plotear todas las líneas halladas junto con la de mínimos cuadrados.
  - b) Hallar las regresiones L1, M de Hampel y la regresión LMS usando todas las variables predictoras. Plotear sus residuales y compararlos con los de la regresión mínimo cuadrática..

# CAPÍTULO 9

## REGRESIÓN NOPARAMÉTRICA

### 9.1 Introducción

La suavización de un conjunto de  $n$  datos  $\{X_i, Y_i\}$  para  $i=1,2,\dots,n$ . consiste en aproximar la función  $g$  en la siguiente relación de regresión.

$$Y_i = g(X_i) + \varepsilon_i \quad (9.1)$$

Donde  $g$  es la curva de respuesta media, llamada también “signal” y el error aleatorio  $\varepsilon$  es llamado “noise”. En regresión lineal  $g(X_i) = \alpha + \beta X_i$  y para efectos de hacer inferencia se asume que  $\varepsilon$  se distribuye normalmente. Más precisamente,  $g(x) = E(Y/X)$  es una media condicional, es decir el promedio de todas las  $y$ 's para un valor dado de  $X$ , donde  $(X, Y)$  no necesariamente aparece en la muestra.

En regresión noparamétrica, la forma de la función  $g$  y la distribución de los errores es determinada usando los datos que se han tomado.

En general hay dos maneras de atacar el problema:

- Ajustar los datos localmente (o sea haciendo uso de vecindades o “WINDOWS”) a través de modelos bien sencillos. Por ejemplo, aproximar la media condicional por un promedio de los valores observados de  $y$  que están en la vecindad del valor de  $x$ . La suavización por kernel es un ejemplo de este caso
- Ajustar un modelo que incluye una parte paramétrica (tal como un modelo polinomial) y otra parte noparamétrica sujeta a una penalidad por complejidad para prevenir el “overfitting”. Cuando ocurre “overfitting” es porque se ha tratado de ajustar el modelo más al “noise” que al “signal”. La suavización por splines es un ejemplo de este caso.

Ambos métodos requieren la selección de un parámetro de suavización

- Cuando se usa modelos locales se requiere estimar el ancho de banda (“bandwidth”) o ancho de ventana.
- Cuando se usa estimación penalizada hay que tratar de estimar la penalidad por complejidad, tratando de balancear la bondad de ajuste del modelo y la complejidad del mismo (la complejidad está relacionada al número de parámetros que hay que estimar en el modelo).

Primero consideraremos el caso de regresión noparamétrica cuando hay una sola variable predictora y una sola variable de respuesta y luego el caso de regresión noparamétrica multidimensional donde hay varias variables predictoras y una sola de respuesta. También existe el caso donde hay varias variables de respuesta y varias predictoras, siendo los más conocidos regresión por “projection pursuit” y MARS.

## 9.2 Suavización bivariada o Suavizadores de diagramas de dispersión (Scatterplot Smoothers)

Entre los métodos más usados están:

- i) El Regresograma (Tukey, 1961),
- ii) “Running means”(Promedios móviles),
- iii) “Running line”,
- iv) Suavización usando los k vecinos más cercanos, K-nn smoothing .
- v) Suavización por kernels, Nadaraya-Watson (1964)
- vi) Regresión local ponderada, LOESS (Cleveland, 1979)
- vii) Regresión polinomial,
- viii) Suavización por splines, (Wabba, 1975)
- ix) Regresión por splines, (Stone and Koo, 1985)

**Ejemplo 1.** A continuación se describirán cada uno de estos métodos usando el conjunto de datos **air**. Este es un conjunto de datos de 111 observaciones y 4 variables tomados de un estudio del medio ambiente en donde se midió 4 variables (columnas): ozone, solar radiation, temperature, y wind speed por 111 días seguidos.

Ozone: surface concentration of ozone in New York, in parts per million.

Radiation: solar radiation

Temperature: observed temperature, in degrees Fahrenheit.

Wind: wind speed, in miles per hour.

Fuente: John M. Chambers and Trevor J. Hastie, (eds.) Statistical Models in S, Wadsworth and Brooks, Pacific Grove, CA 1992, pg. 348.

**9.2.1 El regresograma.** Aquí se divide el intervalo de los valores de la variable predictora en varios subintervalos (usualmente 5). La amplitud de los subintervalos se elige de tal manera que haya aproximadamente igual número de datos en cada uno de ellos. Luego se promedia los valores de la variable de respuesta en cada subintervalo. Esto determina varios segmentos de línea que al unirlos forma el regresograma. Lo malo de este estimador es que no es suave porque hay saltos en cada punto de corte. A continuación se muestra una función **regresorg** en R que calcula el regresograma.

```
regresorg<-function (x,y,k)
{# *****
# Funcion que calcula el regresograma
# Input : El vector x de valores de la variable predictora
#         El vector y de valores de la variable de respuesta
#         k, el número de subintervalos a usar
#Output: Los valores de las medias de y en cada subintervalo y el plot
#         del regresograma superpuesto sobre el diagram de puntos
# Edgar Acuna, Mayo 2003
# *****
n<-length(x)
x<-sort(x)
y<-y[order(x)]
```

```

xpoints<-x[1]
nint<-floor(n/k)
ymeans<-rep(0,k)
for(j in 1:k)
{
ind<-((j-1)*nint+1):(j*nint)
if(j<k)
xpoints<-c(xpoints,x[j*nint])
if(j==k)
{ind<-((j-1)*nint+1):n
xpoints<-c(xpoints,x[n])
}
ymeans[j]<-mean(y[ind])
}
xpoints<-c(xpoints,xpoints[2:k])
xpoints<-sort(xpoints)
ymeans1<-rep(ymeans,each=2)
plot(x,y)
lines(xpoints,ymeans1)
cat("\las medias de y en cada subintervalo son:\n")
ymeans
}

```

En la siguiente figura se muestra el regresograma para los datos del ejemplo1 usando ozone como respuesta y radiation como predictora

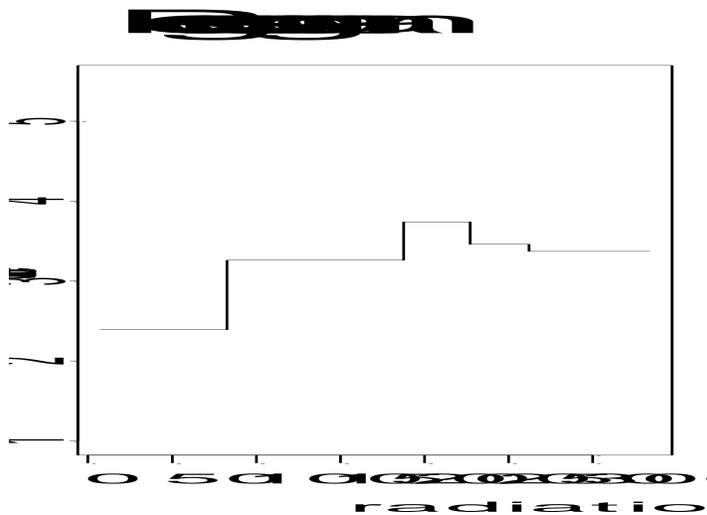


Figura 9.1. Regresograma para el conjunto de datos air.

**9.2.2 “Running Means”, “running Medians” y “Running Lines”** Aquí para cada valor  $x_i$  se define una vecindad simétrica  $N(x_i)$  que contenga a dicho punto. La simetría es el sentido que tiene igual número de puntos  $k$  tanto a la derecha como izquierda del punto dado, en los extremos esto no se puede lograr, pero se trata de estar lo mas cerca posible. El conjunto de índices de la vecindad simétrica para la observación  $x_i$ , varía entre  $\max(i-k, 1)$  hasta  $\min(i+k, n)$ . Luego se calcula el suavizador por “running means” en el punto  $x_i$  de la siguiente manera:

$$s(x_i) = \text{promedio de las } y\text{'s en } N(x_i)$$

El suavizador por “running medians” en el punto  $x_i$  está definido de la siguiente manera:

$$s(x_i) = \text{median de las } y\text{'s en } N(x_i)$$

En tanto que el suavizador por “running lines” se calcula por

$s(x_i)$  = valor estimado de la regresión mínimo cuadrática para  $x = x_i$  que se obtiene usando los puntos  $(x_i, y_i)$  con  $x_i$  que cae en  $N(x_i)$ .

MINITAB calcula la suavización por “running medians” para datos de series de tiempos igualmente espaciadas usando la secuencia **STAT** ▶ **EDA** ▶ **RSMOOTH**.

A continuación se muestra la función **runmeans** escrita en R que calcula la suavización por “running means”, usando vecindades con  $k$  observaciones a cada lado

```
runmeans<-function (x,y,k)
{
#*****
#Funcion que calcula el suavizador por running means
#Inputs: la variable de respuesta y, la variable predictora x y
#      k, el numero de vecinos a cada lado de una observacion
#Output: El plot del suavizador superpuesto en el diagrama de dispersion
#Edgar Acuna, mayo 2003
#*****
n<-length(x)
rm<-rep(0,n)
for(i in 1:n)
{ind1<-max(i-k,1)
ind2<-min(i+k,n)
tempo<-y[ind1:ind2]
rm[i]<-mean(tempo)
}
plot(x,y)

lines(sort(x),rm,type="l")
title("Running means")
}
```

La siguiente grafica muestra el suavizador running means para ozone versus radiation en el conjunto **air**, usando vecindades con  $k=3$  observaciones a cada lado del centro.

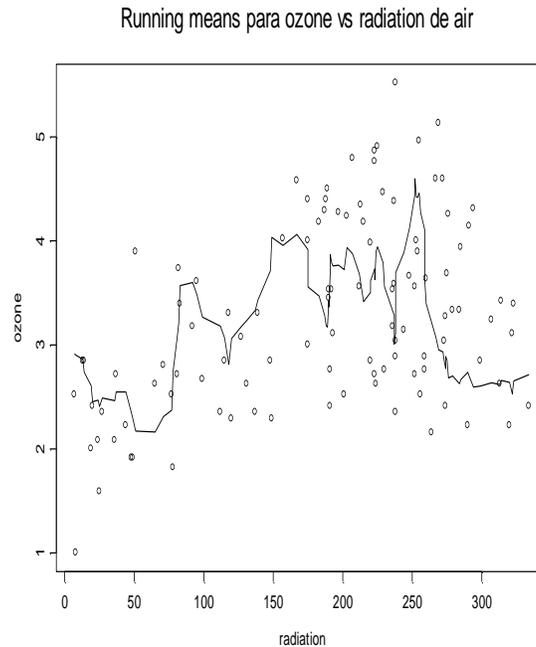


Figura 9.2. Running means para el conjunto de datos air.

**9.2.3 Suavizador por los  $k$  vecinos más cercanos.** Aquí para cada valor de  $x_i$  se define una vecindad  $N_k(x_i)$  que contiene los  $k$  valores de  $x$  que están más cercanos a  $x_i$ . La cercanía se determina usando una función distancia (por ejemplo la euclídeana). El valor de  $k$  generalmente es impar. Luego el suavizador se calcula por

$$s(x_i) = \text{promedio de las } y\text{'s en } N(x_i)$$

La mayoría de los programados estadísticos no tienen incluido este tipo de suavización en su menú. Pero pueden ser programados sin mucha dificultad.

### 9.2.4 Suavización por kernels.

Considerando que tanto  $x$  como  $y$  son aleatorias se puede escribir  $g(x) = E(y/x) = \int yf(y/x)dy$  donde  $f(y/x)$  representa la función de densidad condicional de  $y$  dado  $x$ . Usando la definición de densidad condicional lo anterior se puede re-escribir como

$$g(x) = \frac{\int yf(x, y)dy}{f(x)} \quad (9.2)$$

En la suavización por kernel la función de densidad de  $x$  y la función de densidad conjunta de  $(x,y)$  son estimadas usando los datos  $(x_i,y_i)$  de la muestra. Más específicamente,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \quad (9.3)$$

y

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) K\left(\frac{y-y_i}{h}\right) \quad (9.4)$$

Aquí  $K(t)$  es llamado el kernel y es una función no negativa, simétrica con respecto a 0 y con valor máximo en dicho punto. Además,  $\int_{-\infty}^{\infty} K(t)dt = 1$ . El kernel actúa como una función de peso, que otorga peso grande a los puntos cercanos al punto donde se va a suavizar y bajo peso a los puntos que están alejados del mismo. Hay bastantes funciones que se pueden considerar como Kernel, pero el más usado es el kernel Gaussiano. definido por

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

El parámetro  $h$  es llamado ancho de banda “bandwidth” y es estimado usando los datos. Hay bastantes propuestas para estimar  $h$ . Sustituyendo (3) y (4) en (2) se obtiene la estimación por el método de kernel para  $g$

$$\hat{g}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \quad (9.5)$$

Este estimador fue introducido independientemente en 1964 por Nadaraya y Watson.

La función ***ksmooth*** de la librería ***modreg*** de R halla la suavización basada en kernels. En SAS también se puede obtener este suavizador usando el modulo SAS/INSIGHT. Para estimar el ancho de banda se usa GCV (validación cruzada generalizada). A continuación se muestran los comandos y la gráfica que da S-Plus.

```
> plot(radiation, ozone)
> lines(ksmooth(radiation, ozone, kernel="normal", bandwidth=5))
```

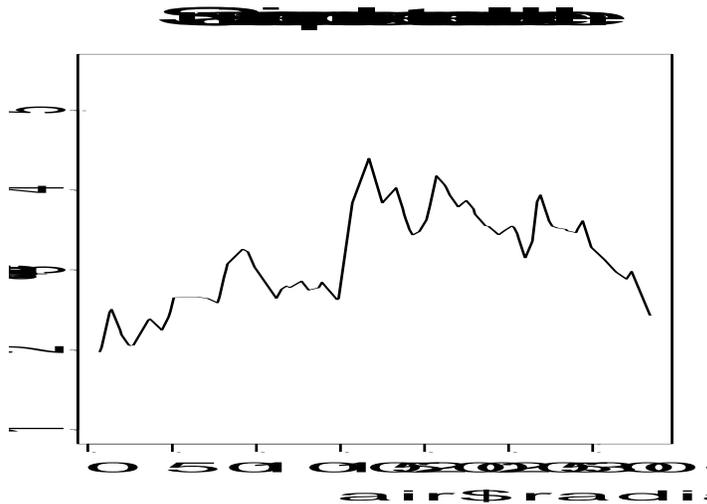


Figura 9.3. Suavización usando el método de kernels para el conjunto de datos air.

### 9.2.5 Regresión local ponderada, LOWESS

En este método, si  $x_0$  es un punto donde se desea hallar la suavización, entonces primero se halla una vecindad usando los  $k$  vecinos más cercanos y luego se halla una regresión ponderada en dicha vecindad el valor ajustado de  $y$  en  $x_0$  será el valor del suavizador. Más detalladamente el método trabaja así:

- i) Se identifican los  $k$  vecinos más cercanos de  $x_0$  y se denota la vecindad por  $N(x_0)$
- ii) Se calcula la distancia a  $x_0$  del punto más alejado que está dentro de la vecindad  $N(x_0)$  y se lo representa por  $\Delta(x_0)$ .
- iii) Para cada punto  $x_i$  en la vecindad  $N(x_0)$  se calcula los pesos  $w_i$  usando la función peso tricúbica definida por:

$$W(t, x_0) = \left[1 - \left(\frac{|t - x_0|}{\Delta(x_0)}\right)^3\right]^3 \text{ siempre que } |t - x_0| < \Delta(x_0)$$

- iv) Se define el suavizador  $s$  en  $x_0$  por:

$s(x_0)$ =valor ajustado en  $x_0$  de la regresión ponderada de  $y$  versus  $x$  en la vecindad  $N(x_0)$ , usando los pesos definidos en iii).

Cleveland, también propuso que se podría usar las funciones pesos de la regresión robusta para protegerse de la presencia de outliers.

Este suavizador es calculado por la función *lowess* (local weighted scatterplot smoother) de S-plus. SAS lo calcula en SAS/INSIGHT y en versión hay el procedimiento LOESS que también lo hace. LOESS es la generalización de LOWESS y permite usar más de una variable predictora. MINITAB también calcula la suavización LOWESS en **GRAPH ▶ PLOT ▶ Display ▶ lowess** A continuación se muestra LOWESS en MINITAB y S-PLUS para el ejemplo 1.

S-PLUS:

```
loess(ozone~radiation)
```

```
Call:
```

```
loess(formula = ozone ~ radiation)
```

```
Number of Observations:    111
Equivalent Number of Parameters: 4.5
Residual Standard Error:    0.7447
Multiple R-squared:         0.33
Residuals:
  min 1st Q median 3rd Q max
-1.382 -0.5808 -0.05205 0.5523 1.851
```

```
> plot(radiation,ozone)
> lines(lowess(radiation,ozone))
>
```

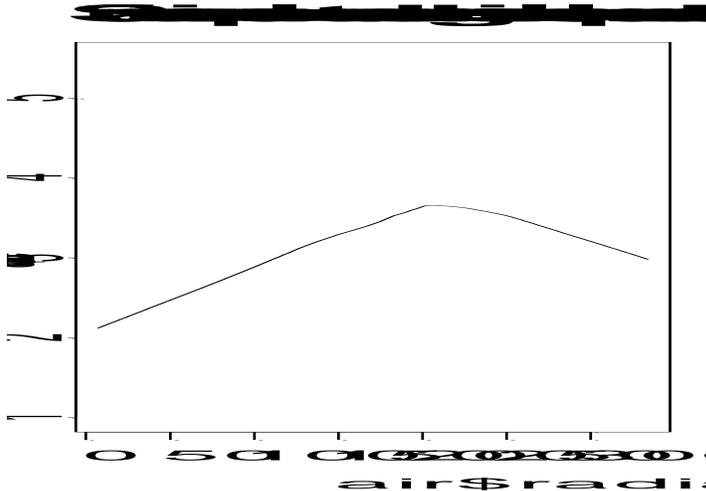


Figura 9.4 . Suavización del conjunto de datos air usando el método de regresión ponderada local

### 9.2.6 Regresión Polinomial.

Aquí se ajustan los datos  $(x_i, y_i)$  para  $i=1, \dots, n$ , a un polinomio de la forma

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k$$

$n$  debe ser mayor que  $k+1$  de lo contrario se tendría un “overfitting” total como lo muestra la siguiente figura

Regresion Polinomial de grado 4

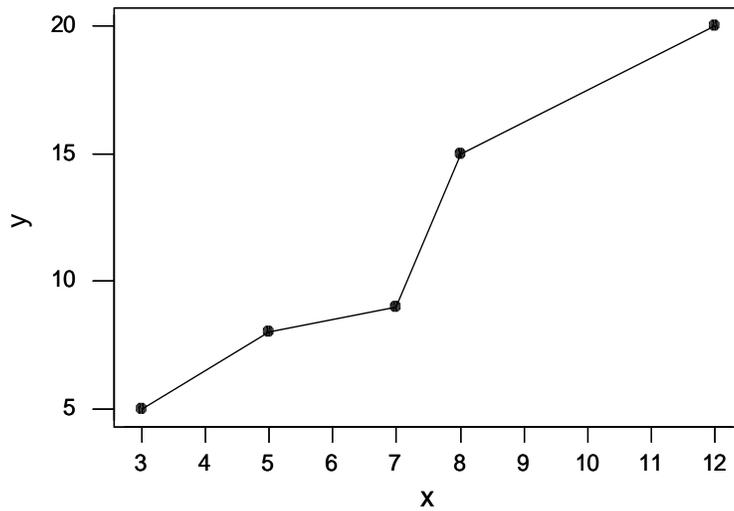


Figura 9.5 Ajuste de una regresión polinomial de grado 4 a un conjunto de 5 observaciones.

Para obtener una suavización polinomial de grado 3 en S-Plus se ejecuta los siguientes comandos

```
plot(radiation, ozone)
lines(airsort[,1], fitted(lm(airsort[,2]~poly(airsort[,1],3))))
```

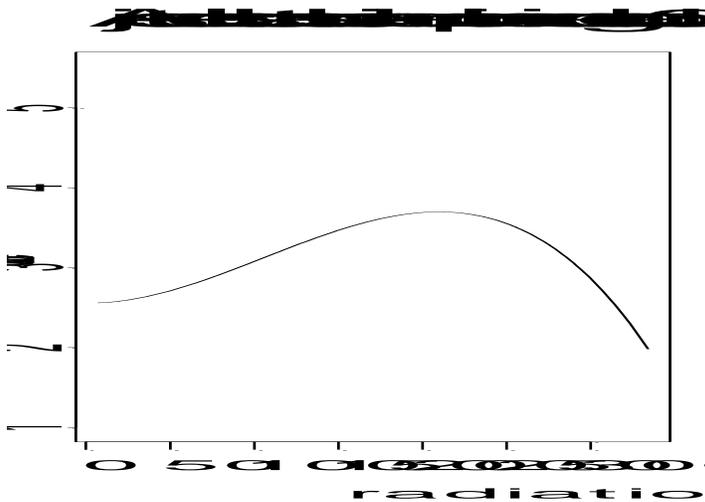


Figura 9.6. Ajuste de los datos air usando una regresión polinomial de grado 3.

### 9.2.7 Regresión por Splines

Un spline (Schoenberg, 1964) de orden  $p$  con  $k$  nudos,  $t_1, \dots, t_k$  en el intervalo  $[a, b]$  es una función que se obtiene dividiendo primero el intervalo  $[a, b]$  en los subintervalos  $[x_0, x_1], \dots, [x_k, x_{k+1}]$ , con  $x_0 = a$  y  $x_{k+1} = b$  y usando luego un polinomio de grado menor o igual que  $p$  en cada uno de los subintervalos, además estos pedazos polinomiales deben unirse suavemente en cada uno de los nudos. Más formalmente, el spline  $s(x)$  está definido por

$$s(x) = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{j=1}^K \beta_{p+j} (x - t_j)_+^p + \varepsilon \quad (9.6)$$

donde  $\beta_0, \beta_1, \dots, \beta_{p+1}, \dots, \beta_K$  son constantes a determinar, y

$$(t - x)_+^p = \begin{cases} (t - x)^p & t > x \\ 0 & t \leq x \end{cases}$$

es llamada la función potencia truncada de orden  $p$

En particular el spline cúbico está dado por

$$s(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^K \beta_{2+j} (x - t_j)_+^3 + \varepsilon \quad (9.7)$$

Un spline es llamado un natural de orden  $2m$  con nudos en  $x_1, \dots, x_k$  si además de lo arriba mencionado el suavizador es un polinomio de grado  $m$  fuera del intervalo  $[x_1, x_k]$

Las funciones  $1, x, x^2, \dots, x^p, (x - t_1)_+^p, \dots, (x - t_k)_+^p$  forman una base de funciones del spline. Lamentablemente esta base tiende a crear problemas de multicolinealidad, por la que se recomienda explorar otras bases. Una alternativa son los B-splines cuya base de funciones son calculadas recursivamente (ver Boor, 1978).

Las funciones  $ns$  y  $bs$  de S-Plus calculan las bases de funciones par el spline natural y el B-spline respectivamente. Los nudos son tomados como los cuantiles. Es decir, si hay un solo nudo éste sería la mediana. Si hay dos nudos entonces estos son los percentiles del 33% y 66%. Si hay 3 nudos estos son los cuartiles y así sucesivamente. El grado de la parte polinomial es tomado por "default" como igual a 3. Si se usa natural spline la relación entre nudos y grados de libertad está dada por nudos =  $gl - 1$  - intercepto, donde intercepto es igual a 1 si se considera el intercepto y a cero si no se considera el intercepto. Para los B-splines la relación es nudos =  $gl - 1$  - grado, donde grado = 3 por "default". En las siguientes gráficas se muestran los 6 natural splines y los 6 B-Splines para ajustar ozone versus radiation

```
>air2<-sort.col(air1,c(1,2),2)
>nair<-ns(air2$radiation,df=6,intercept=T)
>plot(air2$radiation,nair[,6],type="l")
> lines(air2$radiation,nair[,5])
> lines(air2$radiation,nair[,4])
> lines(air2$radiation,nair[,3])
> lines(air2$radiation,nair[,2])
> lines(air2$radiation,nair[,1])

>bair<-bs(air2$radiation,df=6,intercept=T)
```

```

>plot(air2$radiation,bair[,6],type="l")
> lines(air2$radiation,bair[,5])
> lines(air2$radiation,bair[,4])
> lines(air2$radiation,nair[,3])
> lines(air2$radiation,bair[,2])
>lines(air2$radiation,bair[,1])

```

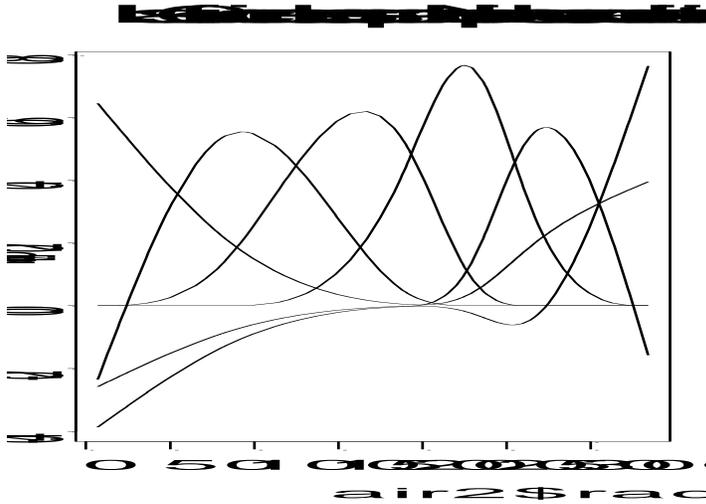


Figura 9.7. Las 6 funciones base par natural splines usando radiación

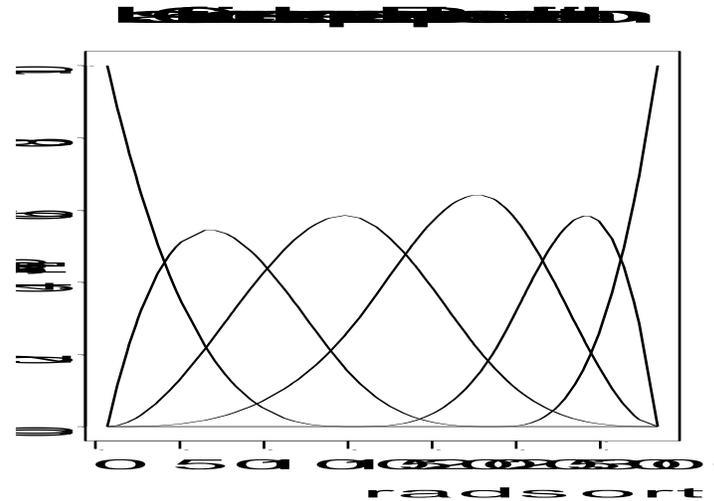


Figura 9.8. Las 6 funciones bases para B splines usando radiación

La regresión por spline usando  $k$  nudos  $t_j$  se define por:

$$y = \beta_0 + \beta_1 x + \dots + \beta_p x^p + \sum_{j=1}^K \beta_{p+j} (x - t_j)_+^p + \varepsilon \quad (9.8)$$

El modelo puede ser linealizado mediante transformaciones y hay que estimar  $p+K+1$  parámetros. El problema es determinar el número de nodos  $K$ . La idea básica es añadir el máximo número de nudos posibles y luego ir eliminando uno por uno tratando de maximizar la bondad de predicción del modelo y minimizando su complejidad.

Una vez determinada la base de los splines se puede hacer la regresión usando las funciones `lsfit` o `lm` de S-Plus.

A continuación se muestran las regresiones usando spline natural y B-Spline.

```
>plot(radiation,ozone)
> lines(airsort[,1],fitted(lm(airsort[,2]~ns(airsort[,1],df=5))))
> lines(airsort[,1],fitted(lm(airsort[,2]~bs(airsort[,1],df=6))))
```

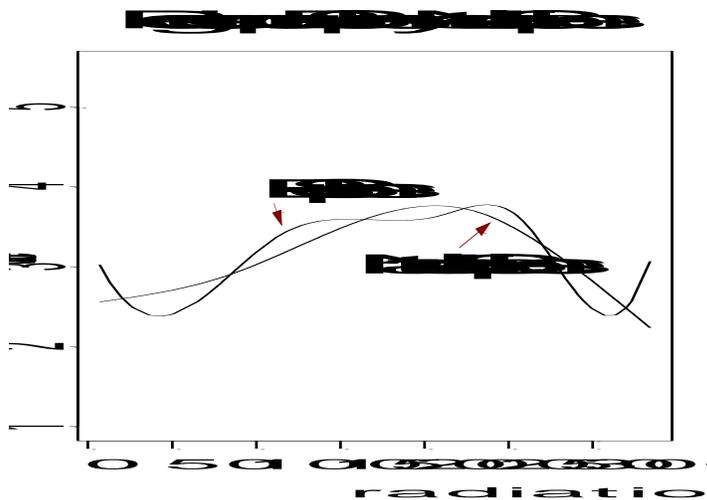


Figura 9.9. Regresión usando B-Splines y Natural splines de ozono versus radiación.

### 9.2.8 Suavización por Splines

El suavizador por splines se obtiene minimizando

$$\sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int [f''(t)]^2 dt \quad (9.9)$$

El primer término es una medida de la bondad de ajuste del modelo y el segundo término es una medida del grado de suavidad. El parámetro de suavidad  $\lambda$  es positivo y gobierna el intercambio entre la suavidad y la bondad de ajuste del suavizador. Cuando  $\lambda = \infty$  se obtiene una aproximación polinomial y cuando  $\lambda = 0$  se obtiene una regresión por spline.

Considerando que  $X_i^t = \{1, X_i, \dots, X_i^p, (X_i - t_1)_+^p, \dots, (X_i - t_k)_+^p\}$

$$\mathbf{X} = \begin{pmatrix} X_1' \\ \cdot \\ X_n' \end{pmatrix} \quad \text{y} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \cdot \\ \beta_{k+p} \end{pmatrix}$$

Entonces la ecuación anterior se puede escribir como

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}' \boldsymbol{\Omega} \boldsymbol{\beta} \quad (9.10)$$

donde  $\boldsymbol{\Omega}$  es una matriz tal que  $\{\Omega\}_{jk} = \int X_j''(t) X_k''(t) dt$

Reinsch (1967) mostró que existe un único mínimo de (9), y que éste es un spline cúbico natural con knots en los únicos valores de  $x_i$ .

Minimizando la expresión (10) con respecto a  $\boldsymbol{\beta}$  se obtiene que

$$\hat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Omega})^{-1} \mathbf{X}'\mathbf{y}$$

que es un resultado bien similar a Regresión Ridge.

Recordando que  $f = \mathbf{X}\boldsymbol{\beta}$  se tendría que

$$\hat{f} = \mathbf{X}'(\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Omega})^{-1} \mathbf{X}'\mathbf{y}$$

Aquí la matriz  $\mathbf{H}(\lambda) = \mathbf{X}'(\mathbf{X}'\mathbf{X} + \lambda\boldsymbol{\Omega})^{-1} \mathbf{X}'$  es llamada la matriz "HAT".

Los grados de libertad de la suavización es igual a la traza de  $\mathbf{H}(\lambda)$ . Esto es bastante similar al número de variables predictoras en un modelo de regresión.

## Elección del parámetro $\lambda$

### a) Usando validación cruzada

Sea  $s(x; \hat{\boldsymbol{\beta}}(\lambda))$  el spline ajustado con parámetro de suavización  $\lambda$

Sea  $s_{-i}(x; \hat{\boldsymbol{\beta}}(\lambda))$  el spline ajustado con parámetro de suavización  $\lambda$  pero sin usar la observación  $(x_i, y_i)$  entonces se define la función de validación cruzada como

$$CV(\lambda) = \sum_{i=1}^n \{s_i - s_{-i}(x_i, \hat{\boldsymbol{\beta}}(\lambda))\}^2$$

el valor  $\lambda$  que minimiza  $CV(\lambda)$  es el valor que se escoge como parámetro de suavización.

El problema con CV es que es computacionalmente caro calcularlo. Una mejor alternativa es usar GCV

### b) Usando validación cruzada generalizada (GCV)

El GCV en realidad no es una generalización del CV sino por el contrario una aproximación. Se define por

$$GCV(\lambda) = \frac{\sum_{i=1}^n \{y_i - s(x_i, \hat{\beta}(\lambda))\}^2}{[1 - \text{tr}(H(\lambda)/n)]^2} \quad (9.11)$$

el valor  $\lambda$  que minimiza  $GCV(\lambda)$  es el valor que se escoge como parámetro de suavización. Este es el procedimiento que usan SAS y S-Plus para estimar el parámetro de suavización.

A continuación se muestran los resultados en S-plus

```
>smooth.spline(air$radiation,air$ozone)
```

Call:

```
smooth.spline(x = air$radiation, y = air$ozone)
```

Smoothing Parameter (Spar): 0.01844406

Equivalent Degrees of Freedom (Df): 4.065246

Penalized Criterion: 48.79781

GCV: 0.5747841

```
> plot(radiation,ozone)
```

```
> lines(smooth.spline(radiation,ozone))
```

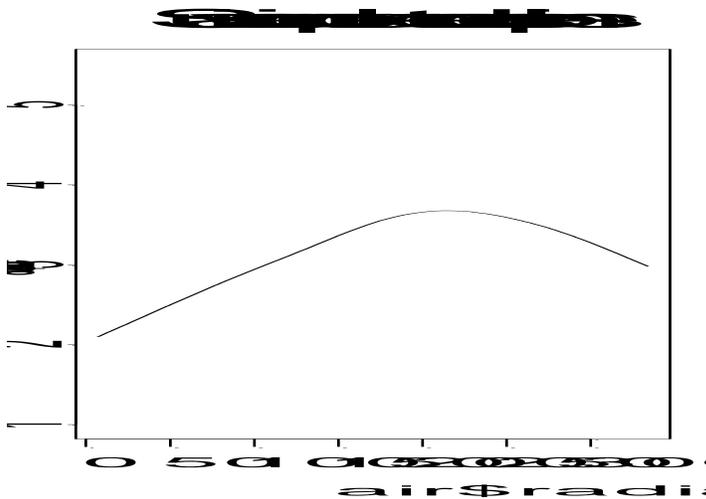


Figura 9.10. Suavización usando splines de radiación versus oone.

El program estadístico **SAS** también hace las suavizaciones usando LOESS, kernel, y splines.

### 9.3 Suavización multidimensional

Existen muchos métodos para hacer suavizaciones cuando existen varias variables predictoras. Entre los métodos más populares están.

- i) Modelos Aditivos generalizados, GAM (Hastie y Tibshirani, 1985)
- ii) Regresión por Projection Pursuit, PPR (Friedman, Stuetzel, 1981)
- iii) Regresión por arboles, CART (Breiman, Friedman, Olsen y Stone, 1984)
- iv) Regresión multivariada adaptativa usando Splines, MARS (Friedman, 1991)
- v) Esperados Condicionales Alternantes, ACE (Breiman y Friedman, 1985)
- vi) Neural Networks (Barron)
- vii) Wavelets smoothing (Donoho y Johnstone, 1995)

A continuación explicaremos brevemente algunos de ellos.

#### 9.3.1 Modelos Aditivos generalizados (GAM)

Un modelo aditivo generalizado es de la forma

$$y = f_1(x_1) + f_2(x_2) + \dots + f_p(x_p) + e$$

aquí las  $f_j$  son estimadas usando cualquiera de los suavizadores bivariados.

El modelo es ajustado usando el algoritmo “local scoring”, el cual iterativamente ajusta modelos aditivos ponderados usando “backfitting”. El algoritmo “backfitting” es un método de Gauss-Seidel para ajustar modelos aditivos usando residuales parciales de suavización iterativamente

#### Algoritmo “Backfitting”

1. En el paso inicial se define las funciones  $f_j^{(0)} \equiv 1$
2. En la  $i$ -ésima iteration, se estima  $f_j^{(i+1)}$  por

$$f_j^{(i+1)} = s\left(y - \sum_{k \neq j} f_k^i(x_k)\right) \text{ para } j=1, \dots, p$$

3. Cotejar si  $|f_j^{(i+1)} - f_j^i| < \delta$  para todo  $j=1, \dots, p$ , donde  $\delta$  es una constante de tolerancia. Si no se cumple la condición volver al paso 2. En caso contrario parar y usar  $f_j^{(i)}$  como  $f_j$  en el modelo aditivo.

Si bien terminos de suavización tales como lowess, bs, ns, kernel o k-nn pueden ir mezclados en una fórmula, es más conveniente usar el mismo suavizador para ahorrar memoria del computador.

S-Plus y R tiene la función **gam** para estimar un modelo aditivo generalizado. Aquí se aplica a los datos del ejemplo 1, usando ozone como variable de respuesta, y radiation y temperature como variables predictoras con regresión splines. La función gam de S-Plus es más general y permite usar otros suavizadores.

```
> gam1 <- gam(ozone ~ s(radiation) + s(temperature), data = air)
```

```

> gam1
Call:
gam(formula = ozone ~ lo(radiation) + bs(temperature),
    data = air)

Degrees of Freedom: 111 total; 103.433 Residual
Residual Deviance: 29.23443
> attributes(gam1)
$names:
 [1] "coefficients"  "residuals"
 [3] "fitted.values" "R"
 [5] "rank"          "smooth"
 [7] "nl.df"         "df.residual"
 [9] "var"           "assign"
[11] "terms"         "call"
[13] "formula"       "family"
[15] "nl.chisq"      "y"
[17] "weights"       "iter"
[19] "additive.predictors" "deviance"
[21] "null.deviance"  "contrasts"

$class:
 [1] "gam" "glm" "lm"

> gam1$fitted.values
 1   2   3   4   5   6   7   8   9  10  11  12
2.65671 2.54582 2.773776 2.597299 2.611294 2.393684 2.045584 2.808312 2.645998 2.736994 2.345038
2.512691
 13  14  15  16  17  18  19  20  21  22  23  24
2.596751 2.458358 2.596977 2.10186 2.094763 2.824071 2.065725 2.284439 1.997066 3.595331
3.432517 3.137973
 25  26  27  28  29  30  31  32  33  34  35  36
3.311191 4.253792 3.917691 3.382748 3.177898 2.868866 2.052431 2.610131 2.868028 3.814197
3.947865 3.606233
 37  38  39  40  41  42  43  44  45  46  47  48
3.586038 3.599523 4.138319 4.472126 4.461953 4.094639 2.979593 3.4159 2.86658 3.578421 3.632034
3.767283
 49  50  51  52  53  54  55  56  57  58  59  60
3.988882 3.933699 2.283529 3.914839 3.939381 3.141197 3.487908 4.17284 3.972782 3.76244 3.590106
3.062994
 61  62  63  64  65  66  67  68  69  70  71  72
2.869332 3.128421 4.267184 4.362023 4.321021 3.93223 3.634388 3.25018 2.70473 2.795085 2.785039
3.365681
 73  74  75  76  77  78  79  80  81  82  83  84
3.248156 3.261354 2.265019 3.409377 3.607601 4.00857 4.772059 4.647696 4.774663 4.528619
4.213954 4.44216
 85  86  87  88  89  90  91  92  93  94  95  96
4.436223 4.46699 3.613728 3.347621 3.512145 3.347788 3.141048 2.992318 3.606233 3.188219
3.287716 2.173835
 97  98  99  100  101  102  103  104  105  106  107  108
2.469837 3.363355 2.76363 2.486444 2.797151 3.630543 2.746346 2.140027 3.267639 2.174337
2.003213 2.770275
 109  110  111

```

```

3.035249 2.847853 2.78659
>
gam1$coefficients
(Intercept) lo(radiation) bs(temperature)1 bs(temperature)2 bs(temperature)3
  2.777342    2.448357   -1.033315    1.147127    1.8088

```

También se puede usar la función `predict.gam` para predecir nuevos valores de la variable de respuesta.

A continuación se presenta un programa para hacer un plot de la superficie estimada por el modelo aditivo generalizado.

```

gtemp<-seq(min(air$temperature),max(air$temperature),length=50)
gradiation<-seq(min(air$radiation),max(air$radiation),length=50)
grid1<-list(radiation=gradiation,temperature=gtemp)
grid1<-expand.grid(grid1)
estimado1<-predict.gam(gam1,grid1)
matest1<-matrix(estimado1,50,50)
persp(gtemp,gradiation,matest1, theta=45, phi=30, xlab="temperature", ylab="radiation",
zlab="ozone")

```

### 9.3.2 Regresión usando árboles de decisión (CART)

En este método, introducido formalmente por Breiman y otros (1984), la superficie de regresión es estimada usando el siguiente modelo aditivo

$$s(\mathbf{x}) = \sum_{i=1}^n c_i I_{N_i}(\mathbf{x})$$

las  $c_i$  son constantes y  $I_{N_i}(\mathbf{x})=1$  si  $\mathbf{x} \in N_i$  y es igual 0 en otro caso. Los  $N_i$  son hiperrectángulos disjuntos con lados paralelos a los ejes coordenados. Los hiperrectángulos son construidos usando por particionamiento recursive del espacio muestral. Estos hiperrectángulos pueden ser representados como una estructura de árbol.

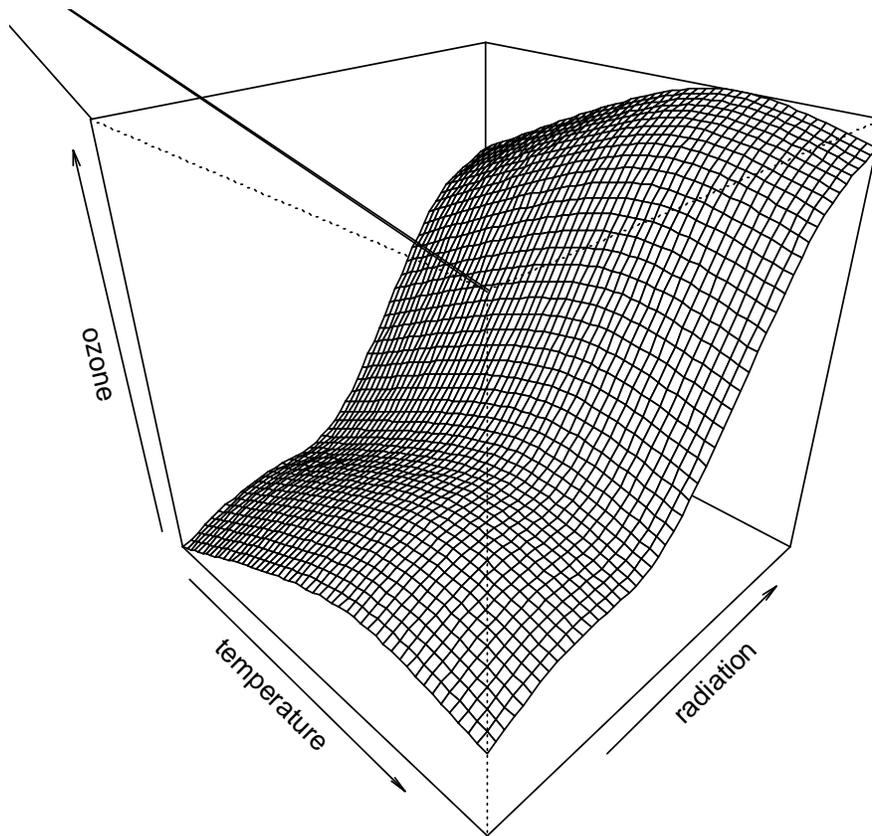


Figura 9.11. Superficie estimados para los datos *air* usando un modelo aditivo generalizado.

**Ejemplo 2:** Obtener la regresión usando arboles de decision para el conjunto de datos *air*

```
> arbol<-tree(ozone~radiation+temperature,data=air)
> arbol
node), split, n, deviance, yval
* denotes terminal node
```

```
1) root 111 87.2100 3.248
2) temperature<82.5 77 35.3700 2.828
4) temperature<77.5 50 11.0600 2.572
8) radiation<85 16 3.8930 2.238
16) temperature<61.5 5 1.3990 1.805 *
17) temperature>61.5 11 1.1310 2.435
34) radiation<25.5 5 0.4135 2.540 *
```

```

35) radiation>25.5 6 0.6169 2.347 *
9) radiation>85 34 4.5340 2.730
18) temperature<72.5 19 2.6920 2.807
36) radiation<208 6 0.8644 3.009 *
37) radiation>208 13 1.4700 2.714
74) radiation<279 6 0.1971 2.583 *
75) radiation>279 7 1.0840 2.825 *
19) temperature>72.5 15 1.5840 2.631
38) temperature<75.5 7 0.4731 2.430 *
39) temperature>75.5 8 0.5792 2.808 *
5) temperature>77.5 27 15.0100 3.301
10) radiation<82 5 2.4310 2.624 *
11) radiation>82 22 9.7670 3.454
22) temperature<81.5 17 7.7900 3.603
44) radiation<221.5 7 0.9025 3.378 *
45) radiation>221.5 10 6.2840 3.761
90) radiation<241 5 4.0340 4.134 *
91) radiation>241 5 0.8587 3.388 *
23) temperature>81.5 5 0.3218 2.949 *
3) temperature>82.5 34 7.4700 4.199
6) temperature<87.5 17 4.0370 3.929
12) radiation<203.5 6 0.7499 3.723 *
13) radiation>203.5 11 2.8940 4.041
26) radiation<272 6 1.5190 4.285 *
27) radiation>272 5 0.5891 3.748 *
7) temperature>87.5 17 0.9438 4.470
14) radiation<205 7 0.1251 4.365 *
15) radiation>205 10 0.6890 4.543 *
> summary(arbol)

```

Regression tree:

```
tree(formula = ozone ~ radiation + temperature, data = air)
```

Number of terminal nodes: 18

Residual mean deviance: 0.1919 = 17.85 / 93

Distribution of residuals:

```

  Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
-0.959 -0.2394 -0.01998 -7.001e-017  0.2474  1.384

```

>

El siguiente es un ejemplo de árbol obtenido con la función plot.tree de S-Plus

```
>plot.tree(arbol, type="u")
```

```
> text(arbol)
```

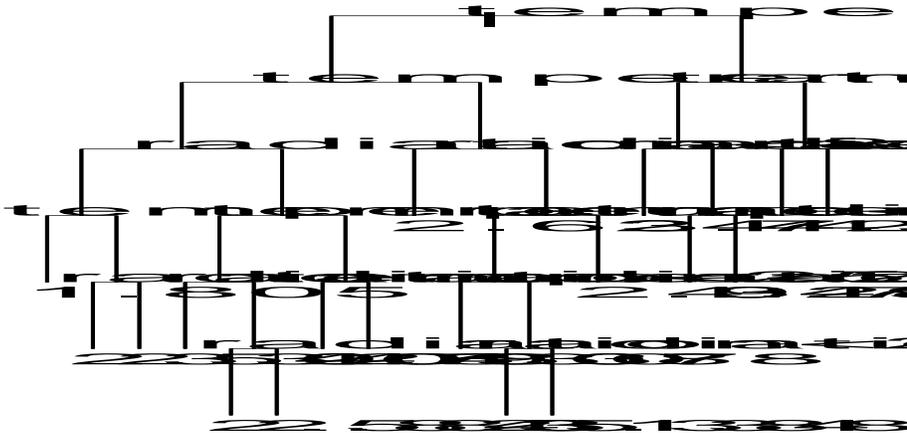


Figura 9.12. Representación como estructura de árbol de la regresión para los datos de air.

Se puede recortar el árbol en forma similar a hacer selección de variables usando la función `prune.tree`.

```
mejorarbol<-prune.tree(arbol,best=5)
```

```
> mejorarbol
```

```
node), split, n, deviance, yval
```

```
* denotes terminal node
```

```
1) root 111 87.210 3.248
 2) temperature<82.5 77 35.370 2.828
   4) temperature<77.5 50 11.060 2.572
    8) radiation<85 16 3.893 2.238 *
    9) radiation>85 34 4.534 2.730 *
   5) temperature>77.5 27 15.010 3.301
    10) radiation<82 5 2.431 2.624 *
    11) radiation>82 22 9.767 3.454 *
  3) temperature>82.5 34 7.470 4.199 *
```

```
> plot.tree(mejorarbol, type="u")
> text(mejorarbol)
```



Figura 9.13. Árbol podado para los datos de air.

Para obtener una superficie de la suavización por árboles se ejecuta los siguientes comandos

```
> gtemp<-seq(min(air$temperature),max(air$temperature),length=50)
> gradiation<-seq(min(air$radiation),max(air$radiation),length=50)
> grid<-cbind(gtemp,gradiation)
> grid1<-list(radiation=gradiation,temperature=gtemp)
> grid1<-expand.grid(grid1)
> estimado<-predict.tree(arbol,grid1)
> matest<-matrix(estimado,50,50)
> persp(gradiation,gtemp,matest, theta=30, phi=45, xlab="radiation", ylab="temperature",
zlab="ozone")
```

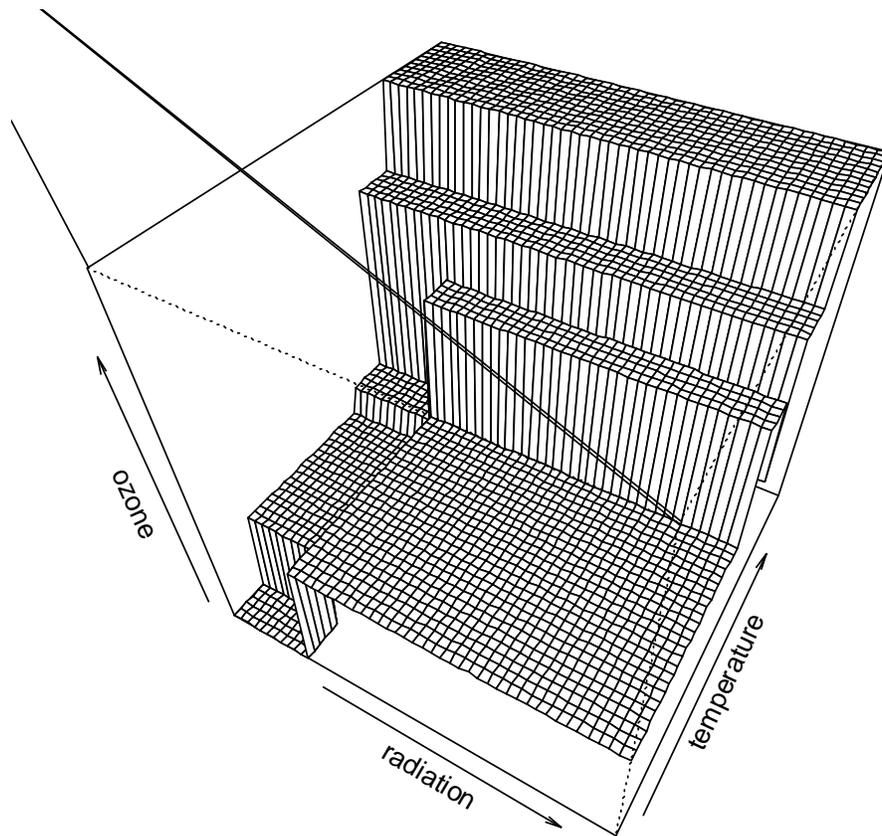


Figura 9.14. Superficie de respuesta de la regresión usando árboles para los datos *air*.

Para hacer predicciones para nuevos datos se usa la función *predict.tree*.

**Ejemplo 3:** Obtener la regresión usando arboles de decision para el conjunto de datos *headcirc*.

Usando R se obtiene:

```
> library(tree)
> arbol<-tree(headcirc~birthwt+gestage,data=headcirc)
> arbol
node), split, n, deviance, yval
* denotes terminal node
```

- 1) root 100 634.800 26.45
- 2) birthwt < 840 21 25.810 23.10
- 4) gestage < 27.5 15 9.333 22.67 \*

```

5) gestage > 27.5 6 6.833 24.17 *
3) birthwt > 840 79 309.800 27.34
6) gestage < 30.5 54 97.500 26.50
12) birthwt < 1180 30 27.200 25.60
24) gestage < 26.5 5 2.800 24.20 *
25) gestage > 26.5 25 12.640 25.88 *
13) birthwt > 1180 24 15.630 27.63
26) gestage < 29.5 17 7.529 27.29 *
27) gestage > 29.5 7 1.714 28.43 *
7) gestage > 30.5 25 91.360 29.16
14) birthwt < 1430 15 63.330 28.67 *
15) birthwt > 1430 10 18.900 29.90 *
> summary(arbol)

```

Regression tree:

```
tree(formula = headcirc ~ birthwt + gestage, data = headcirc)
```

Number of terminal nodes: 8

Residual mean deviance: 1.338 = 123.1 / 92

Distribution of residuals:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.667e+00	-6.667e-01	-3.333e-02	2.842e-16	5.714e-01	6.333e+00

```

> win.graph()
> plot.tree(arbol, type="u")
> text(arbol)

```

Obteniendo un mejor árbol de tamaño 5.

```

> mejorarbol<-prune.tree(arbol,best=5)
> mejorarbol
node), split, n, deviance, yval
* denotes terminal node

```

```

1) root 100 634.80 26.45
2) birthwt < 840 21 25.81 23.10 *
3) birthwt > 840 79 309.80 27.34
6) gestage < 30.5 54 97.50 26.50
12) birthwt < 1180 30 27.20 25.60
24) gestage < 26.5 5 2.80 24.20 *
25) gestage > 26.5 25 12.64 25.88 *
13) birthwt > 1180 24 15.63 27.63 *
7) gestage > 30.5 25 91.36 29.16 *
> win.graph()
> plot.tree(mejorarbol, type="u")
> text(mejorarbol)

```

```

> ggest<-seq(min(headcirc$gestage),max(headcirc$gestage),length=50)
> gbw<-seq(min(headcirc$birthwt),max(headcirc$birthwt),length=50)
> grid<-cbind(ggest,gbw)
> grid1<-list(gestage=ggest,birthwt=gbw)
> grid1<-expand.grid(grid1)
> estimado<-predict.tree(arbol,grid1)

```

```
> #grid2<-as.data.frame(grid1)
> matest<-matrix(estimado,50,50)
> persp(ggest, gbw, matest, theta=30, phi=45, xlab="gestage", ylab="birthwt",
zlab="headcir",col="lightgreen")
```

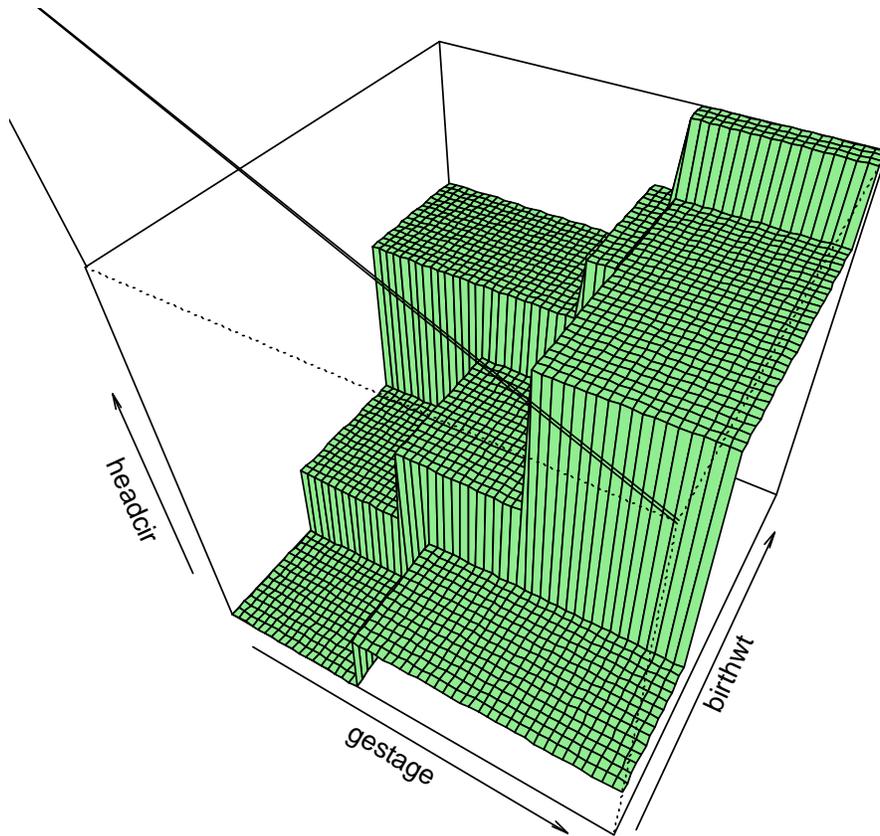


Figura 9.15. Superficie de la regresión usando árboles para los datos headcir.

## EJERCICIOS

1. Considerar el conjunto Berkeley, disponible en la página de internet del curso. Elegir una variable predictora y hallar las suavizaciones por kernel, lowess y splines. Calcular en cada caso la suma de cuadrados de los residuales y plotear las curvas suavizadas
2. Hacer un programa de preferencia en R que haga la suavización por “running lines” ( $k=5$ ) y aplicarlo al conjunto de datos Highway.
3. Hacer un programa de preferencia en R que haga la suavización por los  $k$  vecinos más cercanos ( $k=3$ ) y aplicarlo al conjunto de datos Fuel.

## Apéndice A. Repaso de Matrices

**1.-Definición:** Una matriz es un arreglo rectangular de números reales dispuestos en filas y columnas. Una matriz con  $m$  filas y  $n$  columnas se dice que es de orden  $m \times n$  de la matriz. Cuando  $m=n$  la matriz es llamada matriz cuadrada. Una matriz usualmente es denotada por una letra mayúscula y un elemento cualquiera de ella es llamado una entrada de la matriz. Así  $A=(a_{ij})$  representa a la matriz  $A$  y  $a_{ij}$  es la entrada en la fila  $i$  columna  $j$ .

**Ejemplo 1:** La matriz

$$A = \begin{bmatrix} 1 & 4 & 9 \\ 0.5 & 5 & 7 \\ 2 & 6 & 11 \end{bmatrix}$$

La entrada  $a_{21}=0.5$  y la entrada  $a_{32}=6$ .

Cuando la matriz tiene una sola columna es llamado un **vector columna** y si la matriz tiene una sola fila es llamado un **vector fila**. Así una matriz de orden  $m \times n$  se puede descomponer en  $m$  vectores filas o  $n$  vectores columnas. El número de elementos del vector es llamada la **dimensión** del vector.

**Ejemplo 2.** Escribir un vector fila y un vector columna de la matriz  $A$

**Solución:**

$$\mathbf{a}=[2 \ 6 \ 11] \quad \text{y} \quad \mathbf{b}=\begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

Los elementos de un vector solo tienen un subíndice el cual indica su posición.

### 2.- Operaciones con matrices:

Se pueden sumar y restar matrices siempre que éstas sean del mismo orden. Para obtener la matriz suma (resta) simplemente se suman y restan las entradas correspondientes.

O sea,  $A+B=(a_{ij}+b_{ij})$ .

**Multiplicación de dos vectores.** El producto (interno o escalar) de dos vectores de igual dimensión se obtiene sumando los productos de sus correspondientes elementos. Más específicamente, si  $\mathbf{a}=(a_1, \dots, a_n)$  y  $\mathbf{b}=(b_1, \dots, b_n)$  entonces  $\mathbf{ab}=a_1b_1+\dots+a_nb_n$

*Notar que la multiplicación de dos vectores produce un número y no un vector.*

**Ejemplo 3.** Hallar el producto de los vectores  $\mathbf{a}$  y  $\mathbf{b}$  del ejemplo 2.

**Solución:**  $ab=(2)(4)+(6)(5)+(11)(6)=8+30+66=104$ .

**Multiplicación de matrices.** Para que dos matrices se puedan multiplicar el número de columnas de la primera debe coincidir con el número de filas de la segunda. Así una matriz de orden  $5 \times 3$  se puede multiplicar con una matriz de orden  $3 \times 4$ , pero no con una matriz  $4 \times 4$ .

Sea  $A$  de orden  $m \times n$  y  $B$  de orden  $n \times q$  entonces el producto  $AB=C$  en donde  $C$  es una matriz  $m \times q$  cuya entrada en la posición  $(i,j)$  se obtiene multiplicando la  $i$ -ésima fila de  $A$  con la  $j$ -ésima columna de  $B$ .

Se debe notar que  $AB \neq BA$

**Ejemplo 4.** Calcular  $AB$  si  $A$  es la matriz del ejemplo 1 y  $B = \begin{bmatrix} 1 & 0 \\ 2 & 3 \\ 5 & 4 \end{bmatrix}$

**Solución:**

$$C=AB = \begin{bmatrix} 54 & 48 \\ 45.5 & 43 \\ 69 & 62 \end{bmatrix}$$

Por ejemplo la entrada  $c_{21}=45.5$  se obtuvo multiplicado el vector  $\mathbf{a}_2=[0.5 \ 5 \ 7]$  con

$$\mathbf{b}_1 = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}.$$

**Nota:** La división por matrices no está definida

**Transpuesta de una matriz:** La transpuesta de una matriz se obtiene intercambiando sus filas por sus columnas. La transpuesta de la matriz  $A$  se representa por  $A'$ .

**Ejemplo 5.** Hallar la transpuesta de la matriz  $B$

**Solución**

$$B' = \begin{bmatrix} 1 & 2 & 5 \\ 0 & 3 & 4 \end{bmatrix}$$

La transpuesta de un vector columna da un vector fila. Para ser compatible con el producto de matrices, el producto de dos vectores (columnas)  $\mathbf{a}$  y  $\mathbf{b}$  de igual dimensión es representado más adecuadamente por  $\mathbf{a}'\mathbf{b}$ .

**Propiedades.**

- i)  $(A+B)'=A'+B'$   
 ii)  $(AB)'=B'A'$

**Ejemplo 6.** Verificar la propiedad (ii) usando R y las matrices

$$A = \begin{bmatrix} 3 & 8 \\ 4 & 9 \\ 5 & 12 \end{bmatrix} \quad \text{y} \quad C = \begin{bmatrix} 5 & 1 \\ 3 & 7 \end{bmatrix}$$

**Solución:**

```
> A<-c(3,4,5,8,9,12)
> A<-matrix(A,3,2)
> A
  [,1] [,2]
[1,]  3  8
[2,]  4  9
[3,]  5 12
> # Hallando la transpuesta de A
> B<-t(A)
> B
  [,1] [,2] [,3]
[1,]  3  4  5
[2,]  8  9 12
> C<-c(5,3,1,7)
> C<-matrix(C,2,2)
> C
  [,1] [,2]
[1,]  5  1
[2,]  3  7
> #Multiplicando A por C
> A%%C
  [,1] [,2]
[1,] 39 59
[2,] 47 67
[3,] 61 89
> # Calculando transpuesta de A por C
> t(A%%C)
  [,1] [,2] [,3]
[1,] 39 47 61
[2,] 59 67 89
> #Multiplicando C' por A'
> t(C)%*%t(A)
  [,1] [,2] [,3]
[1,] 39 47 61
```

[2,] 59 67 89

**3. Norma (euclidea) de un Vector.** Dado el vector  $n$  –dimensional  $\mathbf{a}=(a_1,\dots,a_n)$  entonces su norma euclidea se define por

$$\|\mathbf{a}\|=\sqrt{a_1^2 + \dots + a_n^2} = \sqrt{\mathbf{a}'\mathbf{a}}$$

En general si  $p$  es un número real mayor o igual que uno se define la norma  $p$  del vector  $\mathbf{a}$  por

$$\|\mathbf{a}\|_p=(|a_1|^p+|a_2|^p+\dots+|a_n|^p)^{1/p}$$

Si  $p=1$  se obtiene la norma Manhattan y si  $p=\infty$  se obtiene la norma Chebyshev.

**4. Normas de matrices:** Dada una matriz  $A$  de orden  $m \times n$ , y una norma vectorial  $\|\cdot\|$  se define la norma  $p$  de una matriz por

$$\|A\|_p=\max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

En particular, para  $p=1$  se obtiene  $\|A\|_1=\max(\sum_{i=1}^m |a_{ij}|)$ , la mayor de la suma de las

columnas, para  $p=\infty$  se obtiene  $\|A\|_\infty=\max(\sum_{j=1}^n |a_{ij}|)$ , la mayor de la suma de las filas.

Para  $p=2$ , se obtiene,

$$\|A\|_2=\max_{\mathbf{x} \neq 0} \sqrt{\frac{\mathbf{x}'A'A\mathbf{x}}{\mathbf{x}'\mathbf{x}}} = (\text{mayor eigenvalue de } A'A)^{1/2}$$

También es bastante usada la norma de Frobenius.

**5. Matriz Identidad.** Es una matriz cuadrada cuyos elementos de su diagonal son todos unos y los que no están en la diagonal son todos ceros. La matriz identidad de orden  $n$  se denota por  $I_n$ . Por ejemplo,

$$I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

**Propiedad:** Si  $A$  es de orden  $m \times n$  entonces  $A I_n = A$  y  $I_m A = A$ .

**6. Inversa de una matriz.** La inversa de una matriz cuadrada  $A$  se representa por  $A^{-1}$  y es tal que  $AA^{-1}=A^{-1}A=I$ . Existen varios métodos de calcular inversas de matrices. El comando **solve** de R calcula la inversa de una matriz.

**Ejemplo 7.** Calcular usando R la inversa de la matriz  $A$  del ejemplo 1 y verificarla.

```
> A
  x1 x2 x3
[1,] 1.0 4 9
[2,] 0.5 5 7
[3,] 2.0 6 11
> inva<- solve(A)
> inva
  [,1] [,2] [,3]
x1 -0.81250 -0.6250 1.06250
x2 -0.53125 0.4375 0.15625
x3 0.43750 -0.1250 -0.18750
> inva%%A
  x1      x2      x3
x1 1.000000e+00 2.220446e-15 4.218847e-15
x2 -2.775558e-16 1.000000e+00 -1.471046e-15
x3 1.665335e-16 6.106227e-16 1.000000e+00
> A%%inva
  [,1] [,2] [,3]
[1,] 1.000000e+00 3.053113e-16 8.049117e-16
[2,] -2.775558e-16 1.000000e+00 8.326673e-17
[3,] -5.551115e-17 4.718448e-16 1.000000e+00
>
```

**Propiedad:** Si  $A$  y  $B$  son dos matrices cuadradas entonces

$$(AB)^{-1}=B^{-1}A^{-1}$$

**7. Traza de una matriz.** Si  $A$  es una matriz cuadrada entonces su traza es la suma de los elementos que están en su diagonal.

**Ejemplo 8.** La traza de la matriz  $A$  del ejemplo 1 es  $1+5+11=17$ .

**Propiedades.**

- (i)  $\text{tr}(A+B)=\text{tr}(A)+\text{tr}(B)$
- (ii)  $\text{tr}(AB)=\text{tr}(BA)$  siempre que  $AB$  y  $BA$  puedan efectuarse.

**8. Rango de una matriz.** Indica el número de columnas ( o filas) independientes que tiene una matriz. Algunas veces ocurre que una columna (o fila ) de una matriz es una combinación lineal de las otras columnas o fila. Si el rango de la matriz es igual al número de columnas entonces se dice que la matriz es de rango completo.

**Propiedad** Una matriz cuadrada de rango completo tiene inversa.

**9. Matriz Simétrica.** Una matriz cuadrada A es simétrica si es igual a su transpuesta. Es decir, al intercambiar filas por columnas se obtiene la misma matriz.

Por ejemplo, la matriz  $M = \begin{bmatrix} 1 & 7 & 5 \\ 7 & 4 & 9 \\ 5 & 9 & 2 \end{bmatrix}$  es simétrica. La matriz  $P = M'M$  también es simétrica

**Propiedad.** Si una matriz A es simétrica entonces también lo es su inversa  $A^{-1}$ . O sea, si A es simétrica  $(A^{-1})' = A^{-1}$ .

### 10. Determinante de una matriz cuadrada.

El determinante de una matriz cuadrada A consiste de la suma de ciertos productos de los elementos de A, cada uno de los productos es multiplicado por +1 o -1 de acuerdo a ciertas reglas. El determinante de la matriz A se representa por  $|A|$  o  $\det(A)$ .

**Ejemplo 9.** El determinante de la matriz A de orden 2x2 est'a dado por

$$\det(A) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Si una matriz tiene determinante cero se dice que es singular.

#### Propiedades

i) Si una matriz es singular entonces no tiene inversa

ii)  $|AB| = |A||B|$

iii)  $|A^{-1}| = 1/|A|$

**11. Matriz Idempotente.** Una matriz A es idempotente si es simétrica y si  $A^2 = A$

**12. Matriz Triangular.** Si los elementos debajo de la diagonal de la matriz son todos ceros entonces se dice que es del tipo triangular superior y si los elementos por encima de la diagonal son todos ceros entonces es llamada matriz triangular inferior. Sistemas de ecuaciones lineales asociados con matrices triangulares son fáciles de resolver.

El determinante de una matriz triangular es el producto de los elementos que están en su diagonal

**13. Matriz Ortogonal.** Una matriz  $A$  cuyos vectores columnas son de norma uno y ortogonales ( es decir su producto interno da 0) es llamada una matriz ortogonal. Si una matriz cuadrada  $A$  es ortogonal entonces  $A'A=A'A=I$ , o equivalentemente  $A^{-1}=A'$ .

**Propiedad.** El determinante de una matriz ortogonal es  $+1$  o  $-1$ .

**14. Forma Cuadrática.** Dado un vector columna  $\mathbf{z}$  de dimensión  $n$  y una matriz cuadrada  $A$  de dimensión  $n \times n$ . Entonces,

$$\mathbf{z}' \mathbf{A} \mathbf{z} = \sum_{i=1}^n a_{ii} z_i^2 + 2 \sum_{i < j} a_{ij} z_i z_j$$

es llamada una forma cuadrática en  $\mathbf{z}$  con matriz  $A$ . Notar que la forma cuadrática es un escalar. Se dice que la matriz  $A$  es definida positiva si  $\mathbf{z}' \mathbf{A} \mathbf{z} > 0$  para todo  $\mathbf{z} \neq \mathbf{0}$  y es semi definida positiva si  $\mathbf{z}' \mathbf{A} \mathbf{z} \geq 0$  para todo  $\mathbf{z}$ , pero  $\mathbf{z}' \mathbf{A} \mathbf{z} = 0$  para algún  $\mathbf{z} \neq \mathbf{0}$ .

**15. Valores propios y Vectores propios.** Sea  $A$  una matriz de orden  $n \times n$ . Los valores propios  $\lambda_1, \lambda_2, \dots, \lambda_n$  de la matriz  $A$  se obtiene resolviendo la ecuación

$$|A - \lambda I| = 0$$

Asociado con el  $i$ -ésimo valor propio  $\lambda_i$ , hay un vector  $\mathbf{v}_i$  que resulta de resolver

$$(A - \lambda_i I) \mathbf{v}_i = \mathbf{0}$$

**Propiedades:**

- Si la matriz  $A$  es simétrica entonces todos sus valores propios serán reales.
- $\text{Traza}(A) = \sum_{i=1}^n \lambda_i$
- Si  $V$  es una matriz cuyas columnas son los vectores propios de  $A$  entonces se cumple que  $V' A V = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ .

**16. La descomposición de una matriz en valores singulares (SVD).** Sea  $A$  una matriz real de dimension  $m$  por  $n$ . Existen matrices ortogonales  $U$  de orden  $m$  por  $m$  y  $V$  de orden  $n$  por  $n$  tales que:

$$V' A U = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_p) \text{ con } p = \min(m, n) \text{ y donde } \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0 \text{ son llamados los valores singulares de } A.$$

**Propiedad:** Los valores propios de la matriz  $A'A$  son los cuadrados de los valores singulares de  $A$ .

## APENDICE B

### Laboratorios en R para Regresion Apicada

```
#####
#Laboratorio 1: Regresion lineal simple
# Creado por Edgar Acuna
# Enero del 2010
#####
#Leyendo el conjunto de datos mortalidad de la internet
muertes=read.table("http://academic.uprm.edu/eacuna/mortalidad.txt",header=T)
#Mostrando los datos
muertes
#Haciendo un plot de tasa de mortalidad versus porcentaje de inmunizacion
x=muertes$porc.inmuniz
y=muertes$tasa.mort
win.graph()
plot(x,y,xlab="porcentaje de inmunizacion", ylab="tasa de mortalidad")
title("Relacion de la tasa de mortalidad con el porcentaje de inmunizacion")
pais=muertes$nacion
text(x,y,labels=as.character(pais),cex=.65,col="blue",pos=3, srt=30)
#Haciendo el ajuste por minimos cuadrados
l1<-lsfit(x,y)
#Mostrando los resultados del ajuste minimo cuadratico
l1
#Imprimiendo un resultado mas corto del ajuste minimocuadratico
ls.print(l1)
#Trazando la linea de regresión sobre el plot de puntos
abline(l1)
alfa=l1$coeff[1]
beta=l1$coeff[2]
text(50,100,bquote(hat(y)==.(alfa)+.(beta)*x))
#Extrayendo las observaciones anormales 11 y 12 y creando un
#nuevo conjunto muertes1
muertes1<-muertes[-c(11,12),]
#Haciendo el ajuste por minimos cuadrados excluyendo las
#observaciones anormales y ploteando la linea de regresion para el nuevo
#conjunto de datos
x1=muertes1$porc.inmuniz
y1=muertes1$tasa.mort
l2<-lsfit(x1,y1)
win.graph()
plot(x1,y1,xlab="porcentaje de inmunizacion", ylab="tasa de mortalidad")
abline(l2)
alfa1=l2$coeff[1]
beta1=l2$coeff[2]
text(50,100,bquote(hat(y)==.(alfa1)+.(beta1)*x))
ls.print(l2)
#####
```

```

# *****
# Laboratorio 2: Inferencia en regresion lineal simple
# Edgar Acuna
# Febrero 2010
# *****
# Leyendo el conjunto de datos mortalidad
muertes<-read.table("http://math.uprm.edu/~edgar/mortalidad.txt",header=T)
# Calculo de la linea de regresion usando el comando lsfit
l1<-lsfit(muertes$porc.inmuniz,muertes$tasa.mort)
# Calculo de la linea de regresion usando el comando lm
l2<-lm(tasa.mort~porc.inmuniz,data=muertes)
l2
# resumiendo los resultados de la linea de regresion
summary(l2)
# Imprimiendo los coeficientes estimados
summary(l2)$coef
beta=summary(l2)$coef[2,1]
eebeta=summary(l2)$coef[2,2]
# Hallando el intervalo de confianza del 95% para la pendiente Beta
bint<-c(beta-qt(.975,18)*eebeta,beta+qt(.975,18)*eebeta)
bint
# analisis de varianza para regresion
anova(l2)
#Hallando el intervalo de confianza para la media y
# el intervalo de prediccion de Y para varios valores de X.
porc.inmuniz<-c(79,80,89,107)
porc.inmuniz<-as.data.frame(porc.inmuniz)
# Intervalo de confianza para la media de Y
predict(l2,porc.inmuniz,se.fit=T,interval=c("confidence"),level=.99)
# Intervalo de prediccion para Y
predict(l2,porc.inmuniz,se.fit=T,interval=c("prediction"),level=.95)
# Haciendo bandas de confianza y bandas de prediccion
porc.inmuniz=seq(0,100,.5)
porc.inmuniz=as.data.frame(porc.inmuniz)
limic=predict(l2,porc.inmuniz,interval=c("confidence"),level=.95)
limip=predict(l2,porc.inmuniz,interval=c("prediction"),level=.95)
limites=cbind(porc.inmuniz,limic[,2:3],limip[,2:3])
plot(limites[,1],limites[,2],xlab="porcentaje de inmunizacion",ylab="tasa de
mortalidad",ylim=c(-100,300),type="l",col=2)
title("Bandas de confianza y prediccion")
points(limites[,1],limites[,3],type="l",col=2)
points(limites[,1],limites[,4],type="l",col=4)
points(limites[,1],limites[,5],type="l",col=4)
points(muertes$porc.inmuniz,muertes$tasa.mort)
bandas <- expression("intervalo media", "intervalo prediccion")
legend(-3, 1, bandas, lty=1, col=c(2,4), cex=.7)
abline(l2)
# *****

```

```

#*****
# Laboratorio 3: Analisis de Correlacion
# Febrero 2011,
# Edgar Acuna
#*****
#Leyendo el archivo de datos corrs directamente desde la internet
cor1=read.table(file="http://math.uprm.edu/~edgar/corrs.dat",header=T)
col=cor(cor1$x1,cor1$y1)
cat("\n","el coeficiente de correlacion entre X1 y Y1 es=",col,"\n")
co2=cor(cor1$x2,cor1$y2)
cat("\n","el coeficiente de correlacion entre X2 y Y2 es=",co2,"\n")
co3=cor(cor1$x3,cor1$y3)
cat("\n","el coeficiente de correlacion entre X3 y Y3 es=",co3,"\n")
co4=cor(cor1$x4,cor1$y4)
cat("\n","el coeficiente de correlacion entre X4 y Y4 es=",co4,"\n")
win.graph()
par(mfrow=c(2,2),oma=c(1,1,1,1))
plot(cor1$x1,cor1$y1)
text(10,30,"r=.988")
plot(cor1$x2,cor1$y2)
text(10,20,"r=-.992")
plot(cor1$x3,cor1$y3)
text(12,32,"r=.191")
plot(cor1$x4,cor1$y4)
text(12,25,"r=.112")
title("Ejemplos de correlaciones",outer=TRUE)
#Leyendo el archvo de datos corROUT
cor2<-read.table(file="http://math.uprm.edu/~edgar/corROUT.dat",header=T)
attributes(cor2)
c1<-cor(cor2$x1,cor2$y1)
cat("\n ","el coeficiente de correlacion entre X1 y Y1 es=",c1,"\n")
c2<-cor(cor2$x2,cor2$y2)
cat("\n ","el coeficiente de correlacion entre X2 y Y2 es=",c2,"\n")
c3<-cor(cor2$x3,cor2$y3)
cat("\n ","el coeficiente de correlacion entre X3 y Y3 es=",c3,"\n")
c4<-cor(cor2$x4,cor2$y4)
cat("\n ","el coeficiente de correlacion entre X4 y Y4 es=",c4,"\n")
win.graph()
par(mfrow=c(2,2),oma=c(1,1,1,1))
plot(cor2$x4,cor2$y4)
text(10,30,"r=0.264",cex=.9)
plot(cor2$x3,cor2$y3)
text(20,25,"r=0.436",cex=.9)
plot(cor2$x1,cor2$y1)
text(20,25,"r=0.984",cex=.9)
plot(cor2$x2,cor2$y2)
text(50,100,"r=0.995",cex=.9)
title("Efecto de outliers en la correlacion",outer=TRUE)
#*****

```

```

#*****
#Laboratorio 4: Analisis de residuales en regresion lineal simple
#Marzo 2008
#Edgar Acuna
# *****
# llamando al conjunto de datos mortalidad infantil
muertes<-read.table("http://math.uprm.edu/~edgar/mortalidad.txt",header=T)
# Ajustando el modelo de regresion
l2<-lm(tasa.mort~porc.inmuniz,data=muertes)
# Listado de las componentes del objeto l2
attributes(l2)
# imprimiendo los residuales usuales
l2$res
# imprimiendo un resumen de la regresion
l3<-summary(l2)
# imprimiendo un listado de las componentes del objeto l3
attributes(l3)
# Extrayendo la desviacion estandar estimada
l3$sigma
# Calculando los residuales estandarizados
l2$res/l3$sigma
# Calculando los residuales estudentizados (internamente)
studresi<-rstandard(l2)
# Cotejando la normalidad de los residuales estudentizados
win.graph()
par(mfrow=c(1,3),oma=c(1,1,1,1))
hist(studresi)
boxplot(studresi,main="boxplot de residuales")
qqnorm(studresi)
qqline(studresi)
title("Cotejando normalidad de residuales",outer=TRUE)
win.graph()
par(mfrow=c(2,2),oma=c(1,1,1,1))
qqnorm(studresi, main="Plot de Normalidad")
qqline(studresi)
plot(studresi, main="Plot de residuales")
abline(h=0,col=2)
plot(l2$fitted,studresi)
title("residuales versus valores ajustados",cex=.5)
abline(h=0,col=2)
plot(muertes$porc.inmuniz,studresi, main="residuales versus
predictora",cex=.5)
abline(h=0,col=2)
title("Analisis de residuales",outer=TRUE)
# Construyendo la funcion para calcular el estadistico Durbin-Watson
dw<-function(e)
{
# eliminando el primer dato del vector original
e1<-e[-1]
# eliminando el ultimo dato del vector original
e2<-e[-length(e)]
# creando las diferencias
diff<-e1-e2
dw<-sum(diff^2)/sum(e^2)
dw
}
#Leyendo el archivo de datos de mortalidad directamente de la internet

```

```

muertes<-
read.table(file="http://math.uprm.edu/~edgar/mortalidad.txt",header=T)
attributes(muertes)
#Ajustando el modelo lineal
l1<-lm(tasa.mort~porc.inmuniz,data=muertes)
#Calculando el estadistico de Durbin Watson
dw1<-dw(l1$res)
cat("\n ", "el estadistico Durbin Watson de la regresion lineal es=", dw1, "\n")
#*****

#*****
# Laboratorio 5. Introduccion a Regresion lineal multiple
# Edgar Acuna
# Marzo 2012
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
#millaje<-read.table(file="c:/millaje.txt",header=T)
attributes(millaje)
#Haciendo el plot matricial
win.graph()
pairs(millaje)
#haciendo el plot de mpg versus hp
win.graph()
plot(millaje$hp,millaje$mp)
#Extrayendo las variables del conjunto millaje
mpg<-millaje$mpg
hp<-millaje$hp
wt<-millaje$wt
sp<-millaje$sp
vol<-millaje$vol
#Ajustando el modelo lineal mpg vs hp
l1<-lm(mpg~hp)
summary(l1)
#Ajustando el modelo cuadratico de mpg versus hp
hp2<-hp^2
l2<-lm(mpg~hp+hp2)
summary(l2)
#Ajustando el modelo hiperbolico de mpg versus hp
hp1<-1/hp
l3<-lm(mpg~hp1)
summary(l3)
#Ajustando el modelo lineal de mpg versus las predictoras hp y wt
l4<-lm(mpg~hp+wt)
summary(l4)
#Ajustando el modelo lineal de mpg versus las predictoras hp1 y wt
l5<-lm(mpg~hp1+wt)
summary(l5)
#Ajustando el modelo lineal de mpg versus todas las predictoras
l6<-lm(mpg~vol+hp+sp+wt)
summary(l6)
#*****

```

```

#*****
# Laboratorio 6. Calculo matricial en Regresion lineal multiple
# Edgar Acuna,
# Marzo 2012
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
#millaje<-read.table(file="c:/millaje.txt",header=T)
mdatos<-as.matrix(millaje)
dim(mdatos)
# Construyendo la matriz X, cuya primera columna es de unos
mx<-millaje[,c(2:5)]
coll<-rep(1,82)
mx<-cbind(coll,mx)
mx<-as.matrix(mx)
#calculo del número de observaciones y variables predictoras
n<-dim(mx)[1]
p<-dim(mx)[2]-1
# Calculando la transpuesta de mx
t(mx)
# Calculando la transpuesta de mx y mutiplicandola por mx
t(mx)%*%mx
#Calculo de la matriz sombrero
h<-mx%*%solve(t(mx)%*%mx)%*%t(mx)
#Calculo de los valores ajustados yhat=hat*y
yhat<-h%*%mdatos[,1]
#Calculo de los residuales
resid<-mdatos[,1]-yhat
#Calculo de la suma de cuadrados de los errores
sse<-sum(resid^2)
#calculo del cuadrado medio del error=mse
df<-n-p-1
mse<-sse/df
#Calculo de la desviacion estandar estimada
s<-mse^.5
#*****

#*****
# Laboratorio 7:Inferencia en Regresion lineal multiple
# Edgar Acuna,
# Marzo 2012
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
#millaje<-read.table(file="c:/millaje.txt",header=T)
l1<-lm(mpg~.,data=millaje)
#Hallando los p-values para las pruebas de hipotesis
summary(l1)
anova(l1)
#Modelo reducido para probar B(vol)=B(hP)=0
l2<-lm(mpg~sp+wt, data=millaje)
anova.lm(l2)
p=dim(millaje)[2]-1
n=dim(millaje)[1]
k=2

```

```

#Suma de cuadrados de regresion del modelo completo
a=sum(anova(l1)$Sum[-(p+1)])
#Suma de cuadrados de regresion del modelo reducid
b=sum(anova(l2)$Sum[-(k+1)])
#Cuadrado Medio del error del modelo completo
c=anova(l1)$Mean[p+1]
#Calculo del F parcial
fp<-((a-b)/2)/c
fp
#Hallando el percentil de la F con alpha=.05
qf(.95,k,n-p-1)
#hallando el intervalo de confianza del 95% para el valor medio
sp<-100
wt<-20
vol<-90
hp<-50
nuevo<-as.data.frame(cbind(sp,wt,vol,hp))
nuevo
predict.lm(l1,nuevo,se.fit=T,interval=c("confidence"),level=.95)
#Hallando el ntervalo de prediccion del 99% para los mismos datos
predict.lm(l1,nuevo,se.fit=T,interval=c("prediction"),level=.99)
#*****

#*****
#Laboratorio 8. Falta de ajuste en Regresion
#Edgar Acuna,
#Marzo 2012
#*****
#Leyendo el conjunto de datos
millaje=read.table("http://math.uprm.edu/~edgar/millaje.txt",header=T)
millajelf=millaje[,c(1,5)]
table(millajelf$hp)
# Hay m=40 valores distintos de la predictora
# anadiendo una columna var3 que identifica a que grupo pertenece cada
# observación
var3=factor(millajelf$hp,labels=1:40)
millajelf=cbind(millajelf,var3)
millajelf[1:10,]
#haciendo la regresion lineal simple
l1=lm(mpg~hp,data=millajelf)
l1
anova(l1)
#Haciendo el analisis de varianza de claificacion simple de mpg
# entre los 40 grupos diferentes
l2=lm(mpg~factor(var3),data=millajelf)
anova(l2)
#Haciendo el anova para comparar los dos modelos. Se extrae la suma de
cuadrados del
# error Puro y la suma de cuadrados de falta de Ajuste.
anova(l1,l2)
#Consideremos ahora todas las variable predictoras
millajep=millaje[,2:5]
#Encontrando el numero de combinaciones distintas
dim(unique(millajep))

```

```

# Hay m=70 valores distintos de la predictora
# anadiendo una columna var4 que identifica a que grupo pertenece cada
# observación
#
millajelf=gruposlf(millaje[,2:5])
# Haciendo la regresion lineal multiple
l3=lm(mpg~sp+wt+vol+hp,data=millaje)
anova(l3)
#Haciendo el analisis de varianza de clasificacion simple de mpg
# entre los 70 grupos diferentes
mpg=millaje[,1]
millajelf=cbind(mpg,millajelf)
millajelf[1:3,]
l4=lm(mpg~factor(varg),data=as.data.frame(millajelf))
anova(l4)
#Haciendo el anova para comparar los dos modelos . Se extrae la suma de
cuadrados del
# error Puro y la suma de cuadrados de falta de Ajuste.
#anova(l2,l4)
#*****
#*****
gruposlf=function (data)
{# Esta funcion encuentra el numero de grupos distintos entre las filas
# de la matriz X y le anadae una columna varg que identifica al grupo
#
data=as.matrix(data)
p=dim(data) [2]
datau=unique(data)
rowsu=dim(datau) [1]
cat("\nnumber of distinct instances",rowsu)
cat("\n")
rowsrep=1:rowsu
n=dim(data) [1]
varg=rep(0,n)
data1=cbind(data,varg)
for(j in 1:rowsu)
{tempo=row.matches(datau[j,],data)
data1[tempo,p+1]=j
}
return(data1)
}
#*****
row.matches=function(y, X) {
#*****
#This functions finds out the rows in matrix X
#that are equal to the vector y
#y: a vector
#X: a matrix
#*****
  i <- seq(nrow(X))
  j <- 0
  while(length(i) && (j <- j + 1) <= ncol(X))
    i <- i[X[i, j] == y[j]]
  i
}
#*****

```

```

*****
# Laboratorio 9. Diagnosticos de casos influnciales en Regresion
#lineal multiple.
#Nota: Ejecutar primero la funcion acuinflu que aparece al final
#Edgar Acuna, Marzo 2008
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
#millaje<-read.table(file="c:/millaje.txt",header=T)
# Hallando el modelo de regresion ajustado
l1<-lm(mpg~.,data=millaje)
#Hallando los residuales estudentizados internamente
rstint<-rstandard(l1)
rstint
#Hallando los residuales estudentizados externamente
rstext<-rstudent(l1)
rstext
#Determinado las observaciones que son outliers
which(abs(as.vector(rstint))>2)
#Ploteando los residuales y mostrando los outliers segun rsint
out1=rstint[abs(rstint)>2]
win.graph()
par(mfrow=c(2,2))
plot(rstint)
text(names(out1),out1,names(out1),adj=c(1,1))
#mostrando los outliers segun rstext
which(abs(as.vector(rstext))>2)
# Construyendo la matriz X, cuya primera columna es de unos
mx<-millaje[,c(2:5)]
coll<-rep(1,82)
mx<-cbind(coll,mx)
mx<-as.matrix(mx)
p<-dim(mx)[2]
ndatos<-dim(mx)[1]
# Hallando los valores leverages(diagonal de la matriz Hat) de cada
observacion
leverages<-hat(mx)
#Determinando las observaciones que tienen un leverage alto
indlev=which(abs(leverages)>3*p/ndatos)
print(indlev)
#Ploteando los valores leverages
lev1=leverages[indlev]
#win.graph()
plot(leverages)
text(indlev,lev1,indlev,adj=c(1,1))
#calculando la distancia Cook de cada observacion
cookd<-cooks.distance(l1)
#win.graph()
plot(cookd)
#Determinando las observaciones influnciales segun distancia Cook
which(cookd>qf(0.50,p,ndatos-p))
#calculando los dffits
dfits<-dffits(l1)
#Determinando las observaciones influnciales segun dffits

```

```

which(abs(as.vector(dfits))>2*(p/ndatos)^.5)
dif1=dfits[abs(dfits)>2*(p/ndatos)^.5]
#win.graph()
plot(dfits)
text(names(dif1),dif1,names(dif1),adj=c(1,1))
#Calculado los dbfetas
dfb<-dfbetas(l1)
#Determinando las observaciones influenciadas para cada uno de los
coeficientes
for( i in 1:p)
{tempo<-which(abs(as.vector(dfb[,i]))>2*(ndatos)^(-.5))
cat("valores que influncian el coeficiente",i,"\n")
print(tempo)
}
#Calculando los covratios de cada observacion
cvrat<-covratio(l1)
#Determinando las observaciones influenciadas segun los covratios
indcvr=which(abs(1-as.vector(cvrat))>3*p/ndatos)
print(indcvr)
cvr1=cvrat[indcvr]
win.graph()
plot(cvrat)
text(indcvr,cvr1,indcvr,adj=c(-1,-1),cex=.7)
#Determinando todas las observaciones influenciadas con por lo menos uno
#de los criterios
#Nota: la funcion acuinflu es una version corregida de la funcion
#influence.measures disponible en R.
tempo<-acuinflu(l1)
summary(tempo)
#*****

acuinflu<-function (lm.obj)
{#Esta funcion modifica la fucion influence.measures disponible en R.
# Edgar Acuna, Marzo 2003
  is.influential <- function(infmat, n) {
    k <- ncol(infmat) - 4
    if (n <= k)
      stop("Too few cases, n < k")
    absmat <- abs(infmat)
    result <- cbind(absmat[, 1:k] > 2/sqrt(n), absmat[, k + 1] >
      3 * sqrt(k/n), abs(1 - infmat[, k + 2]) > (3 *
      k)/n, infmat[, k + 3]>qf(0.5, k, n - k),
      infmat[, k + 4] > (3 * k)/n)
    dimnames(result) <- dimnames(infmat)
    result
  }
  infl <- lm.influence(lm.obj)
  p <- lm.obj$rank
  e <- weighted.residuals(lm.obj)
  s <- sqrt(sum(e^2, na.rm = TRUE)/df.residual(lm.obj))
  xxi <- chol2inv(lm.obj$qr$qr, lm.obj$qr$rank)
  si <- infl$sigma
  h <- infl$shat
  dfbetas <- infl$coefficients/outer(infl$sigma, sqrt(diag(xxi)))
  vn <- variable.names(lm.obj)
  vn[vn == "(Intercept)"] <- "1_"
  colnames(dfbetas) <- paste("dfb", abbreviate(vn), sep = ".")
}

```

```

dffbts <- e * sqrt(h)/(si * (1 - h))
cov.ratio <- (si/s)^(2 * p)/(1 - h)
cooks.d <- ((e/(s * (1 - h)))^2 * h)/p
dn <- dimnames(lm.obj$qr$qr)
infmat <- cbind(dfbetas, dffit = dffbts, cov.r = cov.ratio,
  cook.d = cooks.d, hat = h)
is.inf <- is.influential(infmat, sum(h > 0))
ans <- list(infmat = infmat, is.inf = is.inf, call = lm.obj$call)
class(ans) <- "infl"
ans
}
#*****

#*****
# Laboratorio 10. Plots de residuales
# Edgar Acuna,
# Marzo 2012
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
#millaje<-read.table(file="c:/millaje.txt",header=T)
millaje
#Hallando la matriz de correlaciones
cor(millaje)
# Construyendo la matriz X, cuya primera columna es de unos
mx<-millaje[,c(2:5)]
coll<-rep(1,82)
mx<-cbind(coll,mx)
mx<-as.matrix(mx)
#Construyendo la matriz sombrero H
H<-mx%*%solve(t(mx)%*%mx)%*%t(mx)
#Hallando los residuales
n<-dim(mx)[1]
In=diag(0,n,n)
ehat<-(In-H)%*%millaje[,1]
#hallando los plots de residuales
win.graph()
par(mfrow=c(2,2))
plot(mx[,2],ehat,xlab="sp")
plot(mx[,3],ehat,xlab="wt")
plot(mx[,4],ehat,xlab="vol")
plot(mx[,5],ehat,xlab="hp")
#Hallando los residuales sin incluir la primera predictora(sp)
mx1<-mx[,-2]
H1<-mx1%*%solve(t(mx1)%*%mx1)%*%t(mx1)
ehat1<-(In-H1)%*%millaje[,1]
#Hallando los residuales de la primera predictora vs las otras
H1<-mx1%*%solve(t(mx1)%*%mx1)%*%t(mx1)
ehat1r<-(In-H1)%*%mx[,2]
#Hallando los residuales de incluir solo la variable volumen
mvol<-mx[,c(1,4)]
hvol<- mvol%*%solve(t(mvol)%*%mvol)%*%t(mvol)
ehvol<-(In-hvol)%*%millaje[,1]
#Hallando los residuales de peso (wt) versus volumen

```

```

ehwt.vol<- (In-hvol)%*%mx[,3]
#Haciendo el plot de regresion parcial de wt versus vol
win.graph()
plot(ehwt.vol,ehvol)
#Hallando los residuales de peso (hp) versus volumen
ehp.vol<- (In-hvol)%*%mx[,5]
#Haciendo el plot de regresion parcial de wt versus vol
win.graph()
plot(ehp.vol,ehvol)
# Cotejando la suposicion de normalidad
l1<-lm(mpg~wt,data=millaje)
summary(l1)
rstint<-rstandard(l1)
win.graph()
  par(mfrow=c(1,3))
  hist(rstint)
  boxplot(rstint)
  qqnorm(rstint)
  qqline(rstint)
#aplicando pruebas noparametricas
shapiro.test(rstint)
ks.test(rstint,"pnorm",0,1)
#Cotejando si la varianza es constante
win.graph()
plot(l1$fitted,rstint)
# Extra: Graficando los puntos y la curva ajustada encima
millaje$z<-1/millaje$wt
l2<-lm(mpg~z,data=millaje)
summary(l2)
win.graph()
#plot(millaje$z,millaje$mpg)
puntosx<-seq(15,60,length=100)
puntosy<-l2$coef[1]+l2$coef[2]/puntosx
plot(millaje$wt,millaje$mpg)
lines(puntosx,puntosy)
#*****

#*****
# Laboratorio 11.  Transformacion para estabilizar la varianza
# Edgar Acuna,
# Abril 2012
#
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
#millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
millaje<-read.table(file="c:/millaje.txt",header=T)
l1<-lm(mpg~.,data=millaje)
# Agrupando la variable de respuesta mpg en intervalos
nclases<-nclass.scott(millaje$mpg)
clases<-cut(millaje$mpg,nclases)
dmilla<-cbind(millaje$mpg,clases)
#Hallando la potencia de la transformacion
medias<-rep(0,nclases)
vars<-rep(0,nclases)

```

```

for(j in 1:nclases)
{medias[j]<-mean(dmillla[dmilla[,2]==j,])
vars[j]<-var(as.vector(dmillla[dmilla[,2]==j,]))
}
lmeans<-log(medias)
lvars<-log(vars)
lsfit(lmeans,lvars)
# El lsfit indica que la varianza es proporcional a la media al cuadrado
# una transformacion logaritmica en la variable de respuesta es recomendada
mpglog<-log(millaje$mpg)
millaje1<-cbind(millaje,mpglog)
l2<-lm(mpglog~sp+wt+vol+hp,data=millaje1)
summary(l1)
summary(l2)
# Considerando que la varianza es proporcional a la media al cubo
# una transformacion h(y)=y^-0.5 es realizada
mpg05<-millaje$mpg^-0.5
millaje2<-cbind(millaje,mpg05)
l3<-lm(mpg05~sp+wt+vol+hp,data=millaje2)
summary(l3)
win.graph()
par(mfrow=c(1,3))
plot(l1$fitted,rstandard(l1),main="sin Transformacion",ylim=c(-3,3))
abline(h=0)
plot(l2$fitted,rstandard(l2),main="var~e(y)^2,w=log(y)",ylim=c(-3,3))
abline(h=0)
plot(l3$fitted,rstandard(l3),main="Var~E(y)^3, w=y^-1/2",ylim=c(-3,3))
abline(h=0)
#*****

#*****
# Laboratorio 12. Transformacion de Box y Tidwell en un modelo de lineal
# multiple.
# Edgar Acuna,
# Marzo 2012
#
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
#millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
millaje<-read.table(file="c:/millaje.txt",header=T)
# Hallando el modelo de regresion ajustado
l1<-lm(mpg~.,data=millaje)
summary(l1)
betas<-l1$coeff
#Excluyendo el intercepto del vector de betas
betas1<-betas[-1]
#hallando las nuevas variables zetas
z1<-millaje$sp*log(millaje$sp)
z2<-millaje$wt*log(millaje$wt)
z3<-millaje$vol*log(millaje$vol)
z4<-millaje$hp*log(millaje$hp)
millaje1<-cbind(millaje,z1,z2,z3,z4)
l2<-lm(mpg~.,data=millaje1)
betas2<-l2$coeff
gammas<-betas2[c(6:9)]
#Hallando los alfas

```

```

alfas<- (gammas/betas1)+1
alfas
#Creando las nuevas variables
sp1<-millaje1$sp^alfas[1]
wt1<-millaje1$wt^alfas[2]
vol1<-millaje1$vol^alfas[3]
hp1<-millaje1$hp^alfas[4]
#regresion con todas las variables transformadas
l3<-lm(millaje1$mpg~sp1+wt1+vol1+hp1)
summary(l3)
#
#Haciendo nuevamente la transformacion desde el inicio pero sin usar vol
millaje2<-millaje[, -4]
l11<-lm(mpg~., data=millaje2)
summary(l11)
betas11<-l11$coeff
#Excluyendo el intercepto del vector de betas1
betas12<-betas11[-1]
#hallando las nuevas variables zetas
z11<-millaje2$sp*log(millaje2$sp)
z21<-millaje2$wt*log(millaje2$wt)
z31<-millaje2$hp*log(millaje2$hp)
millaje2<-cbind(millaje2, z11, z21, z31)
l21<-lm(mpg~., data=millaje2)
betas22<-l21$coeff
gammas1<-betas22[c(5:7)]
#Hallando los alfas1
alfas1<- (gammas1/betas12)+1
alfas1
#Creando las nuevas variables
sp11<-millaje2$sp^alfas1[1]
wt11<-millaje2$wt^alfas1[2]
hp11<-millaje2$hp^alfas1[3]
l5<-lm(millaje2$mpg~sp11+wt11+hp11)
summary(l5)
#Haciendo la regresion con solo las dos variables significativas
l6<-lm(millaje2$mpg~wt11+hp11)
summary(l6)
#*****

#*****
# Laboratorio 13. Transformacion de Box y Cox para remediar la falta de
# normalidad.
# Edgar Acuna,
# Marzo 2012
#
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
#millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt", header=T)
millaje<-read.table(file="c:/millaje.txt", header=T)
# Hallando el modelo de regresion ajustado
l1<-lm(mpg~., data=millaje)
summary(l1)
#Cargando la libreria MASS que contiene la funcion boxcox
library(MASS)
#Haciendo la transformacion

```

```

bc<-boxcox(l1,lambda=seq(-.6,.6,length=20),plotit=T)
#Transformando la respuesta
millaje1<-millaje
millaje1$mpg<-((millaje$mpg)^-0.22-1)/-0.22
l2<-lm(mpg~.,data=millaje1)
summary(l2)
# Viendo el efecto de la transformacion Box-Cox
rstint<-rstandard(l1)
rstint1<-rstandard(l2)
win.graph()
par(mfrow=c(2,3))
hist(rstint)
boxplot(rstint)
title("Antes de la transformacion")
qqnorm(rstint)
qqline(rstint)
hist(rstint1)
boxplot(rstint1)
title("Despues de la tranformacion")
qqnorm(rstint1)
qqline(rstint1)
#*****

#*****
# Laboratorio 14. Minimos cuadrados ponderados
# Edgar Acuna,
# Marzo 2012
#
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
#millaje<-read.table(file="c:/millaje.txt",header=T)
#
#Extrayendo la primera y ultima observacion y consideramdo solo
#las columnas de variable de respuesta(1) y de wt(3).
#
millaje1<-millaje[-c(1,82),c(1,3)]
win.graph()
plot(millaje1[,2],millaje1[,1])
#Agrupando los valores de x
tabla<-table(millaje1[,2])
valsx<-dimnames(tabla)
valsx<-as.double(unlist(valsx))
nvalsx<-length(valsx)
#Calculando la varianza de cada grupo
varsg<-rep(0,nvalsx)
for(i in 1:nvalsx)
{grupox<-millaje1[millaje1[,2]==valsx[i],]
varsg[i]<-var(grupox[,1])
}
varsg
#Asignando los pesos a cada observacion (metodo I)
win.graph()
plot(varsg, valsx)
ndatos<-dim(millaje1)[1]
pesos<-rep(0,ndatos)

```

```

for(i in 1:ndatos)
{for(j in 1:nvalsx)
{if(millaje1[i,2]==valsx[j])
{pesos[i]<-1/varsg[j]}
}
}
pesos
#resultados de regresion sin ponderar
lw1<-lm(mpg~.,data=millaje1,weights=pesos)
summary(lw1)
#Asignando los pesos a cada observacion (metodo II)
valsx2<-valsx^2
s2<-lm(varsg~valsx+valsx2)
s2
valsx<-millaje1[,2]
valsx2<-millaje1[,2]^2
nuevo<-data.frame(valsx, valsx2)
pesos1<-predict.lm(s2,nuevo)
pesos1<-1/pesos1
pesos1
#resultados de regresion ponderada
lw2<-lm(mpg~.,data=millaje1,weights=pesos1)
summary(lw2)
#*****
#Aplicando regresion ponderada con la variable sp
#*****
nclases<-nclass.scott(millaje[,2])
clases<-cut(millaje[,2],nclases)
millaje2<-cbind(millaje[,1],clases)
#Calculando la varianza de cada grupo
varsg2<-rep(0,nclases)
for(i in 1:nclases)
{grupox<-millaje2[millaje2[,2]==i,]
varsg2[i]<-var(grupox[,1])
}
varsg2
#Asignando los pesos a cada observacion
win.graph()
plot(varsg2,1:nclases)
nd<-dim(millaje)[1]
pesos2<-rep(0,nd)
for(i in 1:nd)
{for(j in 1:nclases)
{if(millaje2[i,2]==j)
{pesos2[i]<-1/varsg2[j]}
}
}
pesos2
#Resultados usando regresion ponderada
lw3<-lm(mpg~sp,data=millaje[,c(1,2)],weights=pesos2)
summary(lw3)
#Resultados sin usar regresion ponderada
lw4<-lm(mpg~sp,data=millaje)
summary(lw4)
#*****

```

```

#*****
# Laboratorio 15. Regresion con variables predictoras cualitativas
# Edgar Acuna,
# Marzo 2013
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
pesobebe<-
read.table(file="http://academic.uprm.edu/eacuna/bajopeso.txt",header=T)
#pesobebe<-read.table(file="c:/pesobebe.txt",header=T)
pbebe<-pesobebe[,10]
pmama<-pesobebe[,3]
fuma<-pesobebe[,5]
#regresion considerando que la relacion entre pesomama y pesobebe no
# es afectada por la variable fumar
l1<-lm(pbebe~pmama)
summary(l1)
#regresion considerando que fumar puede afectar la relacion pesomama-pesobebe
l2<-lm(pbebe~pmama+fuma)
summary(l2)
#haciendo una regresion separada para cada grupo
pbebe0<-pbebe[fuma==0]
pmama0<-pmama[fuma==0]
win.graph()
plot(pmama0,pbebe0)
l3<-lm(pbebe0~pmama0)
summary(l3)
pbebe1<-pbebe[fuma==1]
pmama1<-pmama[fuma==1]
win.graph()
plot(pmama1,pbebe1)
l4<-lm(pbebe1~pmama1)
summary(l4)
#Probando la hipotesis de pendientes iguales
g10<-summary(l3)$df[2]
g11<-summary(l4)$df[2]
var0<-summary(l3)$sigma^2
var1<-summary(l4)$sigma^2
#Probando igualdad de varianzas
razon=var0/var1
if(razon<qf(.95,g10,g11))
{cat("se acepta igualdad")}
#Calculo de la varianza ponderada
varp<-(g10*var0+g11*var1)/(g10+g11)
varp
***Calculo de las sumas de cuadrados sxx para cada grupo
ssx0<-sum(pmama0^2)-length(pmama0)*mean(pmama0)^2
ssx1<-sum(pmama1^2)-length(pmama1)*mean(pmama1)^2
#*Calculo de la prueba t
tcal<-(l3$coeff[2]-l4$coeff[2])/sqrt(varp*(1/ssx0+1/ssx1))
tcal
#Decision
if(abs(tcal)>qt(0.925,g10+g11))
{cat("Se rechaza hipotesis de igualdad\n")}
cat("Se acepta que las pendientes son iguales\n")
l5=lm(pbebe~pmama+fuma+fuma*pmama,data=pesobebe)
summary(l5)
#*****

```

```

#*****
# Laboratorio 16: Regresion logistica
# Edgar Acuna,
# Marzo 2013
#
#*****
#Leyendo el archivo de datos de millaje directamente de la internet
pesobebe<-
read.table(file="http://academic.uprm.edu/eacuna/bajopeso.txt",header=T)
#pesobebe<-read.table(file="c:/bajopeso.txt",header=T)
# la primera columna de pesobebe contiene la variable bajopeso que se obtiene
#codificando la ultima variable peso. Si peso<2500 entonces bajopeso es 0 de
lo
#contario vale 1
#Haciendo la regresion de bajopeso versus pesomama
l1=lm(bajopeso~pesomama,data=pesobebe)
#Ploteando bajopeso versus pesomama y la linea de regresion
win.graph()
plot(pesobebe$pesomama,pesobebe$bajopeso)
title("bajopeso versus pesomama")
abline(l1)
#Ploteando los residuales de la regresion
win.graph()
par(mfrow=c(1,2))
plot(l1$fitted,rstandard(l1))
abline(h=0)
plot(pesobebe$pesomama,rstandard(l1))
abline(h=0)
#Ploteando la curva logistica F(x)=1/(1+exp(-x)) en el rango (-10,10)
puntos<-seq(-10,10,length=100)
win.graph()
par(mfrow=c(1,1))
plot(puntos,1/(1+exp(-puntos)),type="l")
title("plot de la curva logistica en (-10,10)")
#Notar que hay algo de similtud con el plot anterior
#Agrupando la variable pesomama en 5 intervalos para sacar estimaciones de
#las probabilidades de bajopeso
ngrupos<-5 # el metodo de scott no conviene aqui
grupos<-cut(pesobebe[,3],ngrupos)
datos1<-cbind(pesobebe[,1],pesobebe[,3],grupos)
proby<-rep(0,ngrupos)
for(i in 1:ngrupos)
{datos2<-datos1[datos1[,3]==i,]
proby[i]<-sum(datos2[,1])/length(datos2[,1])
}
cat("las probabilidades estimadas en cada grupos son:",proby,"\n")
#Desafortunadamente los grupos no son homogeneos y no se prestan para
#aplicar regresion logistica en datos agrupados
#De ahora en adelante vamos a excluir la columna peso de pesobebe
pesobebe<-pesobebe[,-10]
# Haciendo la regresion logistica simple
logis1<-glm(bajopeso~pesomama,data=pesobebe,family=binomial)
summary(logis1)
#Haciendo la regresion logistica multiple
logis2<-glm(bajopeso~.,data=pesobebe,family=binomial)
summary(logis2)

```

```

#Haciendo otra vez la regresion logistica incluyendo solo las variables mas
#significativas
logis3<-
glm(bajopeso~pesomama+raza+fuma+hipertensio,data=pesobebe,family=binomial)
summary(logis3)
#Observando el valor de la devianza y del AIC el tercer modelo seria el mejor
modelo
#sin embargo en lo que sigue vamos a considerar los resultados del segundo
modelo
#*****
#Prediciendo las clases con el segundo modelo
#Haciendo la clasificacion por el metodo mas simple comparando con p=0.5
#*****
phat<-fitted.values(logis2)
nobs<-dim(pesobebe)[1]
clases<-rep(0,nobs)
for(i in 1:nobs)
{if(phat[i]>=0.5){clases[i]<-1}
}
errores<-sum(clases!=pesobebe[,1])
rate<-errores/nobs
cat("la tasa de mala clasificacion es=",rate,"\n")
#*****
# Haciendo la clasificacion con el metodo mas complicado calculando la
sensitividad y #especificidad
#*****
p<-seq(.1,.9,length=9)
sensit<-rep(0,9)
especif<-rep(0,9)
for(j in 1:9)
{clases1<-rep(0,nobs)
for(i in 1:nobs)
{if(phat[i]>=p[j]){clases1[i]<-1}
}
pesobebe1<-cbind(pesobebe[,1],clases1)
sibajo<-pesobebe1[pesobebe1[,1]==1,]
nobajo<-pesobebe1[pesobebe1[,1]==0,]
sensit[j]<-mean(sibajo[,1]==sibajo[,2])
especif[j]<-mean(nobajo[,1]==nobajo[,2])
}
tabla<-cbind(p,sensit,especif)
cat("Sensitividad y especificidad para varios valores de p\n")
print(tabla)
#Haciendo el plot para hallar el p optimo
win.graph()
plot(p,sensit,type="l")
lines(p,especif)
text(p,sensit,labels=p)
title("Ploteando la sensitividad y especificidad para varios p")
# p=0.30 parece ser el optimo
#Ploteando la curva ROC
win.graph()
plot(1-especif,sensit,type="l")
text(1-especif,sensit,labels=p)
title("La curva ROC")
#Notar que para p=.30 la curva esta mas cerca a la esquina superior izquierda
#Clasificacion final

```

```

clasesf<-rep(0,nobs)
for(i in 1:nobs)
{if(phat[i]>=0.3){clasesf[i]<-1}
}
erroresf<-sum(clasesf!=pesobebe[,1])
ratef<-erroresf/nobs
cat("la tasa de mala clasificacion optima es=",ratef,"\n")
#*****

#*****
# Laboratorio 17:  Seleccion de variables
# a) Usando el metodo forward. Hace uso de la libreria leaps y de la
# funcion regsubsets de dicha libreria
# b) Seleccion de los mejores subconjuntos usando leaps y step con
# el criterio AIC. Usa la funcion selforw
# Edgar Acuna,
# abril 2013
#*****
#Leyendo el archivo de datos de grasa directamente de la internet
grasa<-read.table(file="http://math.uprm.edu/~edgar/grasa.txt",header=T)
#grasa<-read.table(file="c:/grasa.txt",header=F)
# El numero maximo de variables a entrar sera igual al numero de
# predictoras del conjunto original
maxvar<-dim(grasa)[2]
#
#llamando a la libreria leaps
library(leaps)
#Aplicando el metodo forward
freg<-regsubsets(grasa~., data=grasa,method="forward",nvmax=maxvar)
#Mostrando la salida de todos los pasos con la estadisticas respectiva
selforw(grasa[,2:14],grasa[,1],.15)
#
# Aplicando el metodo de los mejores subconjuntos
#matrix de predictoras
grasa.x<-grasa[,2:14]
#vector de respuesta
grasa.y<-grasa[,1]
#nombres de las variables predictoras
nombres<-colnames(grasa.x)
leaps(grasa.x,grasa.y,method="r2",nbest=2,names=nombres)
leaps(grasa.x,grasa.y,method="adjr2",nbest=2,names=nombres)
#Mejor modelo usando Cp de mallows
bcp<-leaps(grasa.x,grasa.y,method="Cp",nbest=1,names=nombres)
bcp$Cp
p<-2:maxvar
plot(p,bcp$Cp,type="l")
title("Grafica del Cp de Mallows segun el tamano del modelo")
lines(2:maxvar,2:maxvar)
#Hallando el mejor subcjunto usando stepwise y el criterio AIC
l1<-lm(grasa~.,data=grasa)
step(l1,scope=~.,direction="backward")
#Hallando primero la regresion con la variable predictora mas correlacionada
V7

l2=lm(grasa~abdomen,data=grasa)

```

```

step(l2, scope=~.+edad+peso+altura+cuello+pecho+cadera+muslo+rodilla+tobillo+b
iceps+antebrazo+muneca, direction="forward")
#####
#####
backelim=function(x, y, alpha) {
# Hace forward elimination using el paquete leaps
# x es la matriz o data frame de variables independientes
# y es el vector de respuestas
require(leaps)
m<-ncol(x)           # numero de variables independientes
n<-nrow(x)           # tamaño de la muestra
vm<-1:m
x=as.matrix(x)       # en caso de que x sea data frame
pvmin<-rep(0, m)
regsubsets(x, y, method="backward")->out.x
pv.orig<-1-pf((out.x$rss[vm]-out.x$rss[vm+1])*(n-vm-1)/out.x$rss[vm+1], 1, n-
vm-1)
# sequential min of p-values down from full model
for (i in m:1) {pvmin[i]<-min(pv.orig[m:i])}
sigma2=out.x$sserr/(n -1+ out.x$intercept - out.x$last)
cat("Eliminacion hacia atras", fill=T)
cat("", fill=T)
out=data.frame(p=c(vm, m+1), nvar=c(vm-
1, m), rem.var=c(NA, colnames(x)[out.x$vorder-1]),
pvmin=round(c(NA, pvmin), 4),
s=round(sqrt(out.x$rss/(n-c(vm, m+1))), 3),
r2=round(1-(out.x$rss/out.x$nullrss), 3),
r2adj=round(1-(out.x$rss/out.x$nullrss)*(n-1)/(n-c(vm, m+1)), 3),
Cp=round((out.x$rss/sigma2)-n+2*c(vm, m+1), 3)
)
a=length(pvmin[pvmin>alpha])
cat("p=numero de coeficientes en el modelo", fill=T)
cat("nvar=p-1=numero de variables predictoras", fill=T)
cat("rem.var=la variable a ser removida, el modelo actual no incluye", fill=T)
cat(" esta variable", fill=T)
cat("pvmin=pvalue de la F parcial correspondiente a la variable menos
importante en cada paso", fill=T)
cat("", fill=T)
out1=cbind(out[rev((m+2)-a):(m+1)], 1:4], out[rev((m+1)-a):m], 5:8])
print(out1)
}
#####
forwabic=function(x, y, alpha) {
# Hace seleccion forward usando el paquete leaps
# x es una matriz o data frame de variables independientes
# y es el vector de respuestas
#alpha es el nivel de significacion para la f-parcial
#
require(leaps)
m=ncol(x)           # numero de variables independientes
n=nrow(x)           # tamaño de muestra
vm=1:m
x=as.matrix(x)     # si x es una data frame
pvmax<-rep(0, m)
out.x=regsubsets(x, y, nbest=2, method="forward")
pv.orig<-1-pf((out.x$rss[vm]-out.x$rss[vm+1])*(n-vm-1)/out.x$rss[vm+1], 1, n-
vm-1)

```

```

for (i in 1:m){pvmax[i]<-max(pv.orig[1:i])} # sequential max of pvalues
cat("Seleccion Forward",fill=T)
cat("",fill=T)
out<-data.frame(p=c(vm,m+1),nvar=c(vm-
1,m),add.var=c(NA,colnames(x)[out.x$vorder-1]),
pvmax=round(c(NA,pvmax),4),
aic=round(n * log(out.x$rss/n) + 2*c(vm,m+1),3),
bic=round(n * log(out.x$rss/n) + c(vm,m+1)*log(n),3),
gcv=round(n*out.x$rss/((n-c(vm,m+1))^2),4)
)
a=length(pvmax[pvmax<alpha])
cat("p=numero de coeficientes en el modelo, p=1 es por el
intercepto",fill=T)
cat("nvar=p-1=numero de variables predictoras",fill=T)
cat("add.var=la variable que ha sido anadida al modelo actual",fill=T)
cat("pvmax=p-value de F-parcial correspondiente a la variable mas importante
en cada paso",fill=T)
cat("",fill=T)
print(out[2:(a+1),])
}
#####
selforw=function(x,y,alpha){
# Hace seleccion forward usando el paquete leaps
# x es una matriz o data frame de variables independientes
# y es el vector de respuestas
#alpha es el nivel de significacion para la f-parcial
#
require(leaps)
m=ncol(x) # numero de variables independientes
n=nrow(x) # tamano de muestra
vm=1:m
x=as.matrix(x) # si x es una data frame
pvmax<-rep(0,m)
out.x=regsubsets(x,y,method="forward")
pv.orig<-1-pf((out.x$rss[vm]-out.x$rss[vm+1])*(n-vm-1)/out.x$rss[vm+1],1,n-
vm-1)
for (i in 1:m){pvmax[i]<-max(pv.orig[1:i])} # sequential max of pvalues
cat("Seleccion Forward",fill=T)
cat("",fill=T)
sigma2=out.x$sserr/(n -1+ out.x$intercept - out.x$last)
out<-data.frame(p=c(vm,m+1),nvar=c(vm-
1,m),add.var=c(NA,colnames(x)[out.x$vorder-1]),
pvmax=round(c(NA,pvmax),4),
s=round(sqrt(out.x$rss/(n-c(vm,m+1))),3),
r2=round(1-(out.x$rss/out.x$nullrss),3),
r2adj=round(1-(out.x$rss/out.x$nullrss)*(n-1)/(n-c(vm,m+1)),3),
Cp=round((out.x$rss/sigma2)-n+2*c(vm,m+1),3)
)
a=length(pvmax[pvmax<alpha])
cat("p=numero de coeficientes en el modelo, p=1 es por el
intercepto",fill=T)
cat("nvar=p-1=numero de variables predictoras",fill=T)
cat("add.var=la variable que ha sido anadida al modelo actual",fill=T)
cat("pvmax=p-value de F-parcial correspondiente a la variable mas importante
en cada paso",fill=T)
cat("",fill=T)
print(out[2:(a+1),])
}

```

```

}
#*****

#*****
#Laboratorio 18:Funciones para calcular Press y Validacion cruzada para
#Regresion
#Edgar Acuna
#Mayo 2014
#*****
PRESS=function (x)
{#x es un objeto que sale de aplicar lm
  sum(resid(x)^2/(1 - lm.influence(x)$hat)^2)
}
cv10reg=function(data, folds=10,repert)
{#data: el conjunto de datos
#folds: numero de partes
#numero de repeticiones del experimento
#Suposicion: la primera columna contiene la variable de respuesta
n=dim(data)[1]
p=dim(data)[2]
nombres=colnames(data)
f1=as.formula(paste(nombres[1],".",sep="~"))
#print(f1)
EVC<-rep(0,repert)
for(i in 1:repert)
{
  resid <- matrix(0, 1, folds)
  azar <- data[rank(runif(n)), ]
  parti <- floor(n/folds)
  for(j in 1:folds) {
    cc <- ((j - 1) * parti + 1):(j * parti)
    if(j == folds) {
      cc <- ((j - 1) * parti + 1):n
    }
    datap <- azar[cc, ]
    # La muestra de prueba
    datat <- azar[ - cc, ]
    #La muestra de entrenamiento
    tempo = lm(f1, data = datat)
    tempol <- predict(tempo, datap)
    resid[j] <- sum((tempol - datap[, 1])^2)
  }
  EVC[i] <- sum(resid)/n
}
cat("Los estimados del error promedio de prediccion en cada prediccion
son:\n")
print(EVC)
cat("La estimacion del error promedio de prediccion segun el numero de
repeticiones dado sera\n")
EVC1<-mean(EVC)
EVC1
}
#*****

```

```

#*****
#Laboratorio 19: Diagnosticos de Multicolineralidad: VIF y Numero condicion y
#           Regression Ridge
# Usa la libreria MASS y nuestra funcion acunaridge
# Edgar Acuna,
# Mayo 2014
#*****
library(MASS)
# Ejemplo 1: Conjunto de datos millaje
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
# Hallando la matriz de correlacion de las predictoras
mcor<-cor(millaje[,2:5])
#Hallando los VIFs
vif<-diag(solve(mcor))
cat("los VIF's son:\n")
vif
#recalculando los VIF's sin la variable hp
diag(solve(cor(millaje[,2:4])))
#Hallando el numero condicion
ev<-eigen(cor(millaje[,2:5]))
evals<-ev$values
evals
cond<-sqrt(evals[1]/evals[4])
cond
#Hallando la regresion ridge
rr1<-acunaridge(mpg~.,data=millaje,lambda=seq(0,0.1,.01))
rr1
#Haciendo la Traza Ridge
matridge<-cbind(rr1$lambda,t(rr1$coef[-1,]))
win.graph()
plot(matridge[,1],matridge[,2],ylim=c(-
2,0.5),type="l",xlab="lamb",ylab="coefs",col=2)
title("La traza Ridge para Millaje")
lines(matridge[,1],matridge[,3],col=3)
lines(matridge[,1],matridge[,4],col=4)
lines(matridge[,1],matridge[,5],col=6)

#*****
#Ejemplo 2:Conjunto de datos pollution
pollution<-
read.table(file="http://math.uprm.edu/~edgar/pollution.dat",header=F)
# Hallando la matriz de correlacion de las predictoras
mcor2<-cor(pollution[,1:15])
#Hallando los VIFs
vif2<-diag(solve(mcor2))
cat("los VIF's son:\n")
vif2
#recalculando los VIF's sin las variable 12 y 13
diag(solve(cor(pollution[,c(1:11,14:15)])))
#Hallando el numero condicion
ev2<-eigen(cor(pollution[,1:15]))
evals2<-ev2$values
evals2
cond2<-sqrt(evals2[1]/evals2[15])
cond2
# Hallando la regresion ridge
rr2<-acunaridge(V16~.,data=pollution,lambda=seq(0,.5,.05))

```

```

rr2
#Haciendo la Traza Ridge
matridge2<-cbind(rr2$lambda,t(rr2$coef[-1,]))
win.graph()
plot(matridge2[,1],matridge2[,2],ylim=c(-
20,15),type="l",xlab="lamb",ylab="coefs",col=2)
title("La Traza Ridge para Pollution")
for(i in 3:16)
{
lines(matridge2[,1],matridge2[,i],col=i)
}
#*****

#*****
#Laboratorio 20: Regresion con Componentes Principales
# Requiere la libreria mva
# Edgar Acuna,
# Mayo 2014
#*****
library(mva)
# Ejemplo 1: Conjunto de datos millaje
millaje<-read.table(file="http://math.uprm.edu/~edgar/millaje.txt",header=T)
pcmilla<-prcomp(millaje[,2:5],retx=T,scale=T)
summary(pcmilla)
#Haciendo el screeplot para seleccionar los componentes a usar
screeplot(pcmilla)
#Haciendo regresion usando los componentes principales
milla1<-as.data.frame(cbind(millaje$mpg,pcmilla$x))
colnames(milla1)[1]="mpg"
l1<-lm(mpg~PC1,data=milla1)
summary(l1)
l2<-lm(mpg~PC1+PC2,data=milla1)
# Notar que los coeficientes de regresion anteriores no varian
summary(l2)
l3<-lm(mpg~PC1+PC2+PC3,data=milla1)
summary(l3)
l4<-lm(mpg~PC1+PC2+PC3+PC4,data=milla1)
summary(l4)

#*****
#Ejemplo 2:Conjunto de datos pollution
pollution<-read.table("http://math.uprm.edu/~edgar/pollution.dat",header=F)
pcpollu<-prcomp(pollution[,1:15],retx=T,scale=T)
summary(pcpollu)
# Haciendo el screeplot para seleccionar las componentes a usar
screeplot(pcpollu,npcs=15,type="lines")
#Haciendo regresion usando los componentes principales
pollu1<-as.data.frame(cbind(pollution$V16,pcpollu$x))
#colnames(pollu1)[1]="mort"
m1<-lm(V1~PC1,data=pollu1)
#Regresion stepwise usando los componentes principales y el criterio AIC
summary(m1)
step(m1,scope=~.+PC2+PC3+PC4+PC5+PC6+PC7+PC8+PC9+PC10+PC11+PC12+PC13+PC14+PC1
5,direction="forward")

```

```

m2<-lm(V1~., data=pollu1)
step(m2, scope=~., direction="backward")
#*****

#*****
#Laboratorio 21: Suavizadores bidimensionales: regresograma, running means,
#           regresion por kernels y lowess
#Edgar Acuna,
#Enero 2015
#*****
regresorg<-function (x,y,k)
{# *****
# Funcion que calcula el regresograma
# Input
# Edgar Acuna, Enero 2015
# *****
n<-length(x)
x<-sort(x)
y<-y[order(x)]
xpoints<-x[1]
nint<-floor(n/k)
ymeans<-rep(0,k)
for(j in 1:k)
{
ind<-((j-1)*nint+1):(j*nint)
if(j<k)
xpoints<-c(xpoints,x[j*nint])
if(j==k)
{ind<-((j-1)*nint+1):n
xpoints<-c(xpoints,x[n])
}
ymeans[j]<-mean(y[ind])
}
xpoints<-c(xpoints,xpoints[2:k])
xpoints<-sort(xpoints)
ymeans1<-rep(ymeans,each=2)
plot(x,y)
lines(xpoints,ymeans1)
title("Regresograma")
cat("\las medias de y en cada subintervalo son:\n")
ymeans
}
#Aplicando regresograma al conjunto de datos air
air<-read.table("http://math.uprm.edu/~edgar/air.txt",header=T)
#air<-read.table("c:/air.txt",header=T)
regresorg(air[,2],air[,1],5)
#
runmeans<-function (x,y,k)
{
#*****
#Funcion que calcula el suavizador por running means
#Inputs: la variable de respuesta y, la variable predictora x y
#           El numero de vecinos a cada lado de una observacion
#Output: El plot del suavizador superpuesto en el diagrama de dispersion
#Edgar Acuna, Enero 2015
#*****

```

```

n<-length(x)
rm<-rep(0,n)
for(i in 1:n)
{ind1<-max(i-k,1)
ind2<-min(i+k,n)
tempo<-y[ind1:ind2]
rm[i]<-mean(tempo)
}
plot(x,y)
lines(sort(x),rm,type="l")
title("Running means")
}
#
#Aplicando running means al conjunto de datos air
runmeans(air[,2],air[,1],3)
#
# Regression por kernels
plot(air[,2],air[,1])
lines(ksmooth(air[,2],air[,1],kernel="normal",bandwidth=5))
title("regresion usando kernels")
#
# Lowess regression
plot(air[,2],air[,1])
lines(lowess(air[,2],air[,1]))
title("lowess")
#*****

# *****
# Laboratorio 22: Regresion usando arboles de decision
# Usa la libreria tree de Ripley
# Edgar Acuna
# Enero 2015
#*****
library(tree)
arbol<-tree(ozone~radiation+temperature,data=air)
arbol
summary(arbol)
win.graph()
plot.tree(arbol, type="u")
text(arbol)
mejorarbol<-prune.tree(arbol,best=5)
mejorarbol
win.graph()
plot.tree(mejorarbol, type="u")
text(mejorarbol)
gtemp<-seq(min(air$temperature),max(air$temperature),length=50)
gradiation<-seq(min(air$radiation),max(air$radiation),length=50)
grid<-cbind(gtemp,gradiation)
grid1<-list(radiation=gradiation,temperature=gtemp)
grid1<-expand.grid(grid1)
estimado<-predict.tree(arbol,grid1)
grid2<-as.data.frame(grid1)
matest<-matrix(estimado,50,50)
persp(gradiation,gtemp,matest, theta=30, phi=45, xlab="radiation",
ylab="temperature", zlab="ozone")
#*****

```

## REFERENCIAS

1. Belsley, D., Kuh, y Welsh, R. (1980) Regression Diagnostics. John Wiley, New York.
2. Draper, N y Smith, H. (1998). Applied Regression Analysis, Third Edition. John Wiley, New York.
3. Faraway, J.J. (2014). Linear Models with R. SEcond Edition CRC Press, Boca Raton, Florida .
4. Haerdle, W. (1990). Applied Nonparametric Regression. Cambridge University Press. New York.
5. Hastie, T. y Tibshirani, R. (1990). Generalized additive models. Chapman and Hall, London.
6. Hosmer, D., Lemeshow, S. Sturdivant, R.(2013). Applied Logistic Regression. Third Edition. John Wiley, New York.
7. Montgomery, D., Peck, E. y Vining, G. (2012). Introduction to Linear Regression Analysis. Fifth Edition John Wiley and Sons, New York, .
8. Myers, R. (1990). Classical and Modern Regression with Applications. Duxbury Press, Belmont, California.
9. Neter, J., Wasserman, W., Kutner, M.H, y Nachtsheim, C. (1996). Applied Linear Statistical Models, McGraw-Hill, Boston
10. Rao, C.R. (1973). Linear Statistical Inference and its applications. John Wiley and Sons, New York.
11. Rawlings, J.O., Sastry, G.P. y Dickey D.A (1998), Applied Regression Analysis: A Research Tool, Springer-Verlag, New York.
12. Rousseeuw, P. y Leroy A. (1987). Robust Regression and outlier detection. John Wiley. New York.
13. Ryan, T.P. (2009). Modern Regression Methods. Second Edition John Wiley, New York.

14. Seber, G.A.F y Lee, A. (2003) Linear Regression Analysis. Second Edition. John Wiley, New York.

15. Shaether, S. J. (2009) A Modern approach to regression with R. Springer Verlag, New York.

16. Weisberg, S. (2005). Applied Linear Regression. Third Edition. John Wiley, New York.

17. Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer-Verlag, New York.