9. REGRESIÓN LINEAL

Dr. Edgar Acuña
http://math.uprm.edu/~edgar

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGUEZ

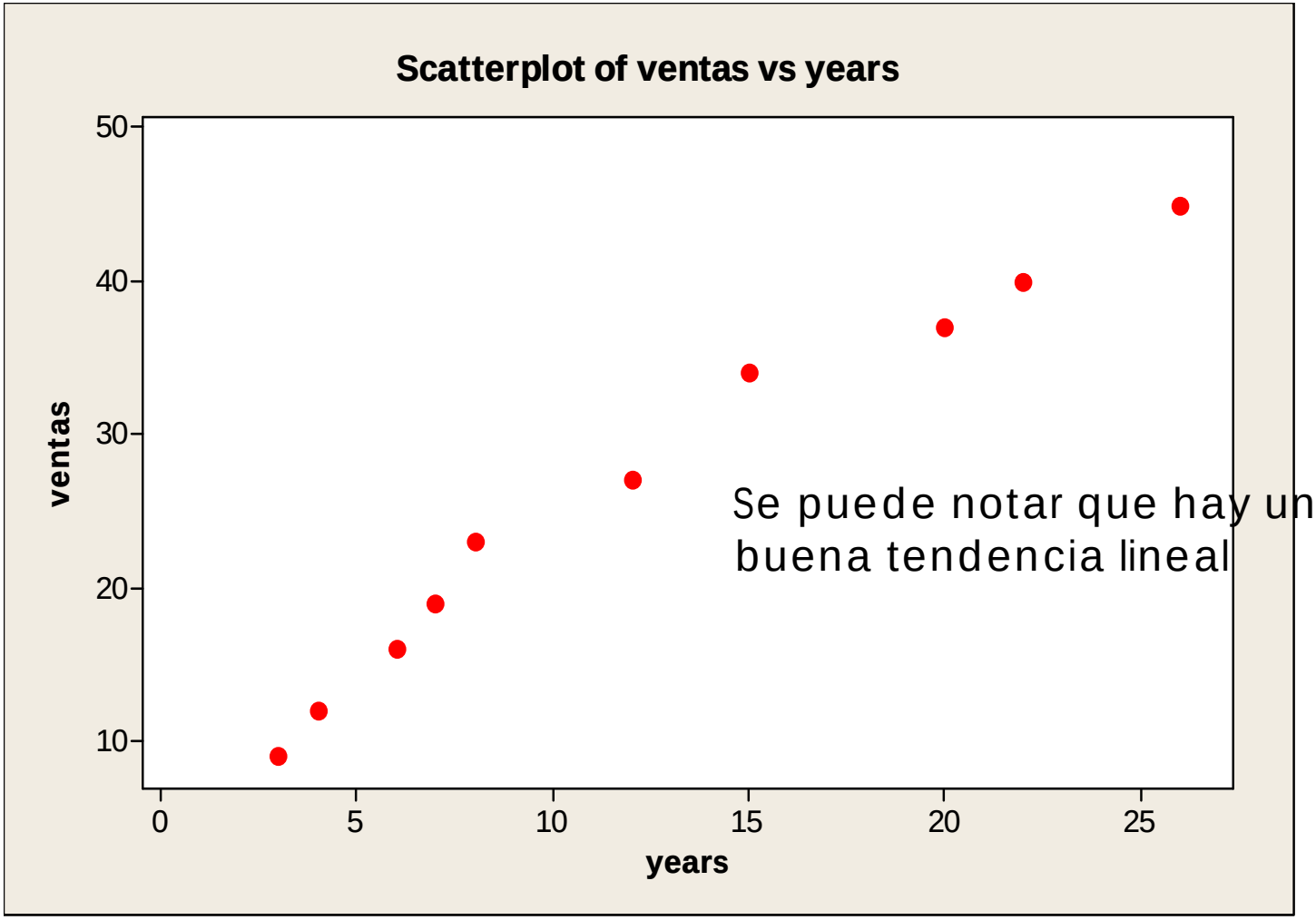
Ejemplo

Ejemplo 3.23. El dueño de una empresa que vende carros desea determinar si hay relación lineal entre los años de experiencia de sus vendedores y la cantidad de carros que venden. Los siguientes datos representan los años de experiencia (X) y las unidades de carros vendidas al año (Y), de 10 vendedores de la empresa.

X(años)	3	4	6	7	8	12	15	20	22	26
Y(ventas)	9	12	16	19	23	27	34	37	40	45

Solución:

Primero hacemos un plot considerando los años de experiencia en el eje horizontal y las ventas en el eje vertical. En **MINITAB.**
Hay que usar la secuencia GRAPH > Scatterplot


Scatterplot of ventas vs years
ventas
50
40
30
20
10
0
5
10
15
20
25
years
Se puede notar que hay una buena tendencia lineal

3.8 El Coeficiente de Correlación

Llamado también coeficiente de correlación de Pearson, se representa por **r** y es una medida que representa el grado de asociación entre dos variables cuantitativas X e Y.

Se calcula por

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Donde:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \quad , \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \quad \text{y} \quad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

Tanto S_{xx} como S_{yy} no pueden ser negativas, S_{xy} si puede ser positiva o negativa.

- La correlacion varia entre -1 y 1
- En la mayoria de los problemas, una correlacion mayor que .75 o menor que -.75 es considerada bastante aceptable. Una correlacion que cae entre -.3 y .3 es considerada muy baja.
- Si la correlacion es positiva entonces cuando X aumenta se espera que Y tambien aumente.
- Si la correlacion es negativa entonces cuando X aumenta se espera que Y disminuya.

Ejemplo (cont)

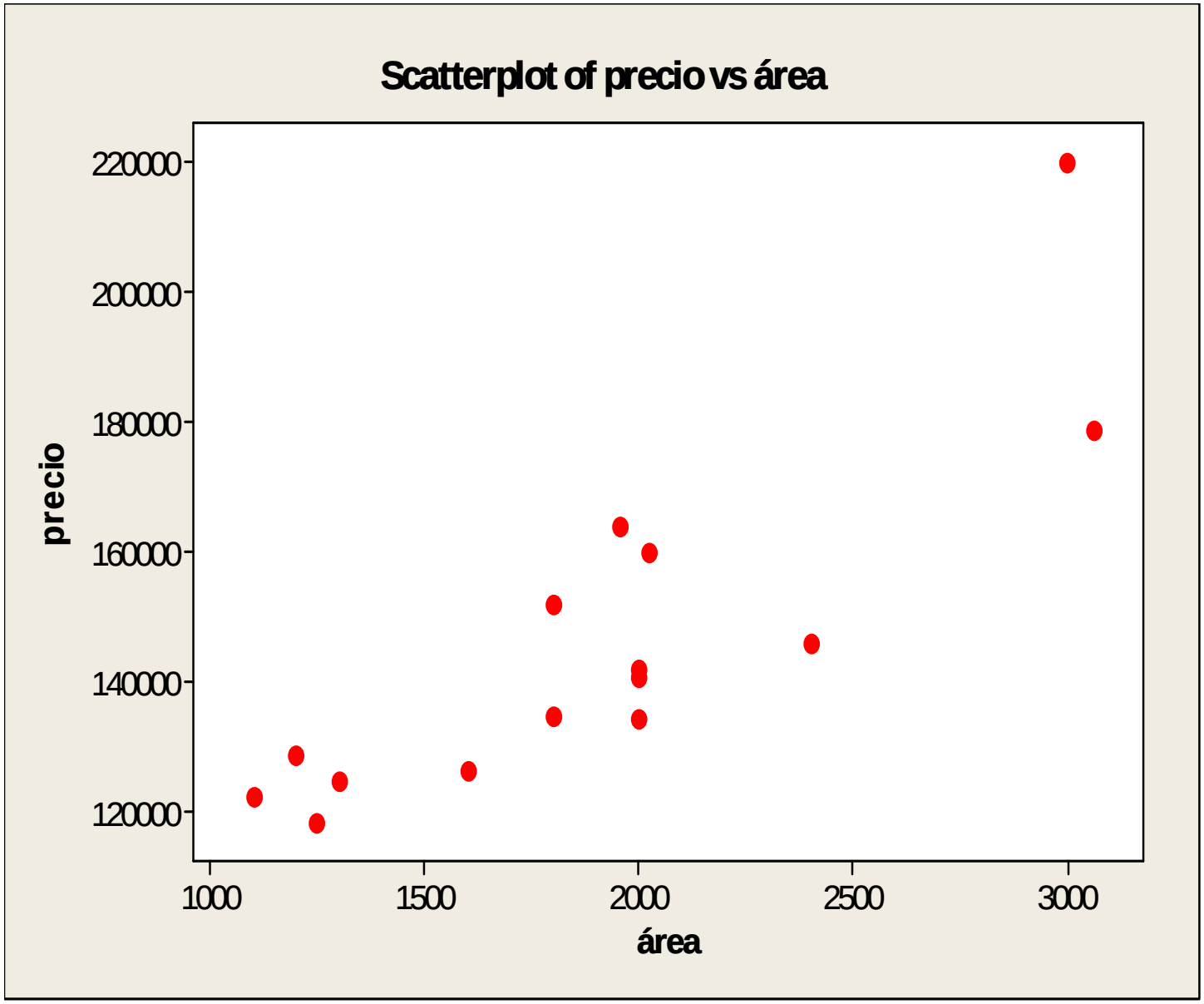
Row	years	ventas	Sxx	Syy	Sxy	r
1	3	9	590.1	1385.6	889.4	0.983593
2	4	12				
3	6	16				
4	7	19				
5	8	23				
6	12	27				
7	15	34				
8	20	37				
9	22	40				
10	26	45				

En **MINITAB**, el coeficiente de correlación se puede obtener eligiendo la opción **correlation** del submenú **Basic Statistics** del menú **Stat.**

Interpretación: *Existe una buena relación lineal entre los años de experiencia y las unidades que vende el vendedor. Además mientras más experiencia tiene el vendedor más carros venderá. Se puede usar los años de experiencia para predecir las unidades que venderá anualmente a través de una línea recta.*

Ejemplo 9.1

Casa	área(pies^2)	precio
1	3060	179000
2	1600	126500
3	2000	134500
4	1300	125000
5	2000	142000
6	1956	164000
7	2400	146000
8	1200	129000
9	1800	135000
10	1248	118500
11	2025	160000
12	1800	152000
13	1100	122500
14	3000	220000
15	2000	141000


Scatterplot of precio vs área
220000
200000
180000
160000
140000
120000
precio
1000
1500
2000
2500
3000
área

Regresión Lineal Simple

Se trata de predecir el comportamiento de Y usando X entonces el **modelo de regresión lineal simple** es de la forma:

$$Y = \alpha + \beta X + \varepsilon$$

Donde, Y es llamada la variable de respuesta o dependiente,

X es llamada la variable predictora o independiente,

α es el intercepto de la línea con el eje Y,

β es la pendiente de la línea de regresión y

ε es un error aleatorio, el cual se supone que tiene media 0 y varianza constante σ^2.

Línea de regresión estimada

El modelo de regresion lineal es estimado por la ecuacion

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

El estimado $\hat{\alpha}$ de α y el estimado $\hat{\beta}$ de β son hallados usando el método de mínimos cuadrados, que se basa en minimizar la suma de cuadrados de los errores.

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

Luego se obtienen $\hat{\beta} = \dfrac{s_{xy}}{s_{xx}}$ y $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

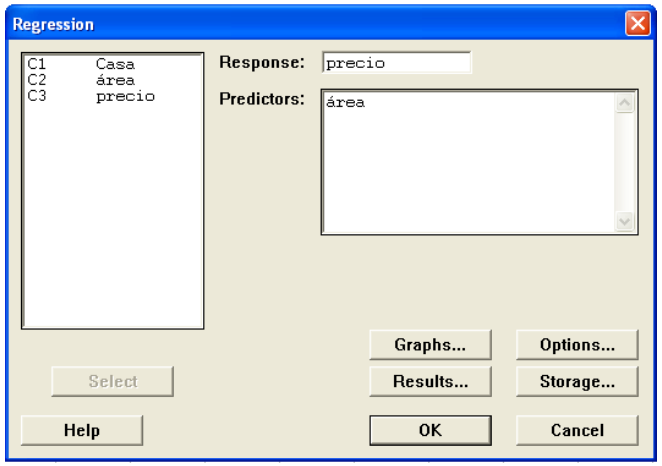
Ejemplo 9.1 (cont.)

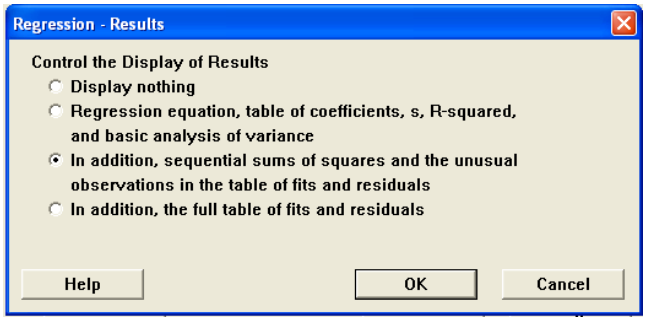
Se desea hallar una línea de regresión que permita predecir el precio de una casa (Y) basado en el *área* de la misma (X).

Solución

Para ello tenemos la Ventana de diálogo para regresión.

Luego en **results** en **regression.**





Resultados

Regression Analysis

The regression equation is
precio = 73168 + 38.5 area

Predictor	Coef	StDev	T	P
Constant	73168	12674	5.77	0.000
area	38.523	6.391	6.03	0.000

S = 14118 R-Sq = 73.6% R-Sq(adj) = 71.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	7241245891	7241245891	36.33	0.000
Residual Error	13	2591087442	199314419		
Total	14	9832333333			

Unusual Observations

Obs	area	precio	Fit	StDev Fit	Residual	St Resid
14	3000	220000	188737	7923	31263	2.68R

R denotes an observation with a large standardized residual

Interpretación de los Coeficientes de Regresión:

- **Interpretación del intercepto $\hat{\alpha}$:**

  Indica el valor promedio de la variable de respuesta Y cuando X es cero. Si se tiene certeza de que la variable predictora X no puede asumir el valor 0,

  entonces la interpretación no tiene sentido.

  En el ejemplo anterior, $\hat{\alpha}$ = 73,168 indicaría que si la casa no tiene área, su precio promedio será 73,158, lo cual no es muy razonable.

- **Interpretación de la pendiente $\hat{\beta}$:**

  Indica el cambio promedio en la variable de respuesta Y cuando X se incrementa en una unidad.

  En el ejemplo anterior $\hat{\beta}$ = 38.5 indica que por cada pie cuadrado adicional de la casa su precio aumentará en promedio en 38.5 dólares.

Inferencia en Regresión Lineal

- **Inferencia acerca de los coeficientes de regresión**

Las pruebas de hipótesis más frecuentes son, Ho: $\alpha = 0$ versus Ha: $\alpha \neq 0$ y

Ho: $\beta = 0$ versus Ha: $\beta \neq 0$.

La prueba estadística para el caso de la pendiente viene dada por:

$$t = \frac{\hat{\beta}}{s.e(\hat{\beta})} = \frac{\hat{\beta}}{\frac{s}{\sqrt{S_{xx}}}} \qquad \text{y} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}}$$

La cual se distribuye como una *t* con *n*-2 grados de libertad.

En **MINITAB** aparece el valor de la prueba estadística y el "p-value" de la prueba, el cual se puede usar para llegar a una decisión. Un "p-value" cercano a 0, digamos menor que 0.05, lleva a la conclusión de rechazar la hipótesis nula.

Si se rechaza la hipótesis nula quiere decir de que de alguna manera la variable X es importante para predecir el valor de Y usando la regresión lineal. En cambio si se acepta la hipótesis nula se llega a la conclusión de que, la variable X no es importante para predecir el comportamiento de Y usando una regresión lineal.

En el Ejemplo 9.1 el valor de la prueba estadística de *t* es 6.03 y el P-value = .0000 por lo que se rechaza la hipótesis nula. Luego hay suficiente evidencia estadística para concluir que la variable área de la casa puede ser usada para predecir el precio de la casa.

Intervalos de confianza para los parámetros de regresión.

Intervalo de confianza del 100 (1-α) % para la pendiente β.

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

MINITAB no da este intervalo de confianza.

Hay que calcular el percentil de la *t* de student

usando la secuencia: **Calc ▸ Probability Distributions ▸ t**.

En el ejemplo anterior, un intervalo del 95 % para la pendiente será:

$$38.523 \pm (2.1604)6.391$$

O sea, hay una confianza del 95 % de que la pendiente de la regresión poblacional caiga en el intervalo (24.7150, 52.3301).

El Análisis de Varianza para Regresión Lineal Simple

En el caso de regresión, la descomposición de la variación de la variable de respuesta Y es como sigue:

VAR. TOTAL DE Y = VAR. DEBIDA A LA REGRESIÓN + VAR. DEBIDA AL ERROR

Cada variación es representada por una suma de cuadrados, definidas de la siguiente manera:

Suma de Cuadrados Total = SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Suma de Cuadrados de Regresión = SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Suma de Cuadrados del Error = SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Cada una de estas sumas de cuadrados tiene una distribución Ji Cuadrado

Tabla del análisis de varianza

Fuentes de Variación	**Grados de Libertad**	**Suma de Cuadrados**	**Cuadrados Medios**	**F**
Debido a la regresión	1	SSR	MSR=SSR/1	MSR/MSE
Debido al Error	n-2	SSE	MSE=SSE/n-2	
Total	n-1	SST		

La hipótesis nula Ho: $\beta = 0$ se rechaza si el “p-value” de la prueba de F es menor que .05.

En el ejemplo 9.1 la prueba de F es 36.33 y el "P-value"=.0000, por lo que se rechaza la hipótesis nula. Notar que el valor de la prueba de $F = 36.33 = (6.03)^2$ es el cuadrado de la prueba *t*.

El Coeficiente de Determinación

Es una medida de la bondad de ajuste del modelo de regresión hallado.

Donde, $$R^2 = \frac{SSR}{SST}$$

SSR representa la suma de cuadrados debido a la regresión y
SST representa la suma de cuadrados del total.

El coeficiente de determinación es simplemente el cuadrado del coeficiente de correlación. El coeficiente de Determinación varía entre 0 y 1, aunque es bastante común expresarlo en porcentaje. Un R^2 mayor del 70 % indica una buena asociación lineal entre las variables, luego la variable X puede usarse para predecir Y. $\boldsymbol{R^2}$ indica qué porcentaje de la variabilidad de la variable de respuesta Y es explicada por su relación lineal con X.

En el ejemplo salio R^2=73.6 esto significa que solo el 73.6% de la variabilidad de los precios de las casas es explicada por su relacion lineal con el area de la misma. Se podria usar el area de la casa para predecir su precio.

Intervalos de Confianza para el valor medio de Y e Intervalo de Predicción

Se busca es establecer un intervalo de confianza para la media asumiendo que la relación entre X e Y es lineal.

$$\hat{Y}_o = \hat{\alpha} + \hat{\beta} X_o$$

Un intervalo de confianza del 100(1-α)% para el valor medio de todos los valores *Y* dado que $X = X_0$ esta dado por:

$$\hat{Y}_0 \pm t_{(\alpha/2, n-2)} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

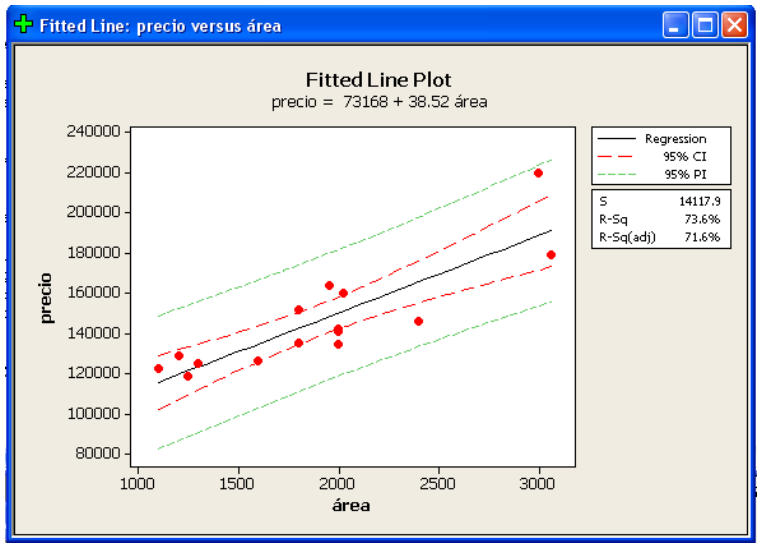
Un intervalo de confianza del 100(1-α) % para el valor predicho de *Y* dado que $X = X0$ es de la forma:

$$\hat{Y}_0 \pm t_{(\alpha/2, n-2)} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

El botón **Options** de la ventana **regression** permite hallar estos intervalos de confianza. En este ejemplo se trata de determinar el intervalo de confianza e intervalo de predicción para el precio de la casa cuando ésta tiene un área de 3,500 pies cuadrados usando un nivel de confianza del 95 %. Para ello hay que seleccionar las opciones **Confidence limits** y **Prediction limits.**

Interpretación*: Hay un 95 % de confianza de que el valor medio de todas las casas de 3,500 pies cuadrados de área caiga entre 184,536 y 231,461 dólares.*
Hay un 95 % de confianza de que el valor de una casa de 3,500 pies cuadrados caiga entre 169,518 y 2246,479 dólares.

la opción **Fitted line Plot** del menú de **Regression** permite hallar bandas de confianza tanto para el valor predicho como para el valor medio de las Y. Para esto se deben elegir las opciones **Display Confidence Interval** y **Display Prediction Interval** al oprimir el botón **Options**. Con las bandas de confianza se pueden tener intervalos de confianzas para cualquier valor dado de X. Para el presente ejemplo se obtiene:



Las bandas se van angostando cuando los valores de X que se toman están cerca del promedio

$$\overline{x}$$

Análisis de Residuales

Un residual es la diferencia entre el valor observado Y_i y el valor estimado por la línea de regresión $\hat{Y}$,

El residual puede ser considerado como el error aleatorio e_i observado.

También se acostumbra usar el **Residual estandarizado**, el cual se obtiene al dividir el residual entre la desviación estándar del residual, y el **Residual estudentizado "deleted",** que es similar al anterior pero eliminando de los cálculos la observación cuyo residual se desea hallar.

En un analisis de residuales se puede detectar:

- Si efectivamente la relación entre las variables X e Y es lineal.
- Si hay normalidad de los errores.
- Si hay valores anormales en la distribución de errores.
- Si hay varianza constante (propiedad de Homocedasticidad) y
- Si hay independencia de los errores.

Análisis de Residuales

Plot de Normalidad: Permite cotejar normalidad. Si los puntos están bien cerca de una línea recta se concluye, que hay normalidad.

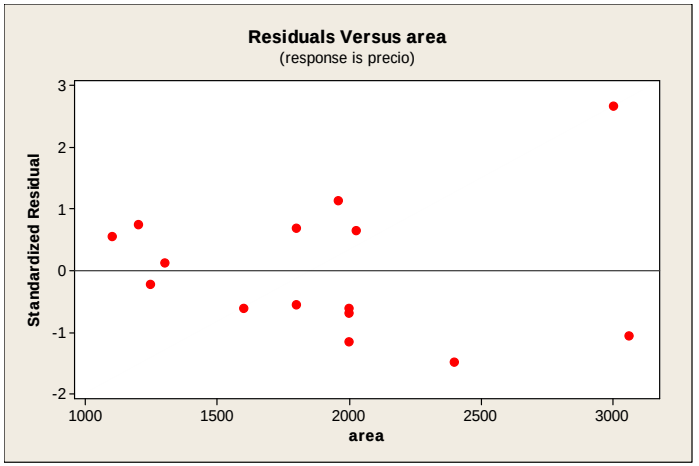
Histograma de Residuales: También permite cotejar normalidad. Cuando el histograma es simétrico, con un único pico en el centro, se concluye que hay normalidad.

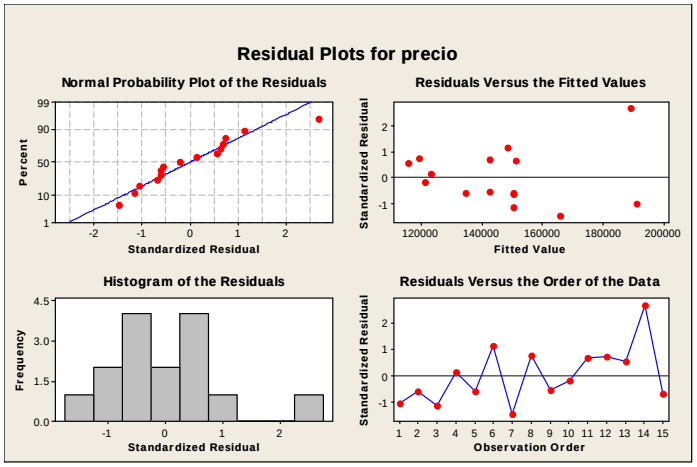
Plot de Residuales versus los valores predichos (FITS): Se usa para detectar si hay datos anormales, cuando hay datos que caen bastantes alejados, tanto en el sentido vertical como horizontal. También permite detectar si la varianza de los errores es constante con respecto a la variable de respuesta.

Plot de Residuales versus el índice de la observación: Es más específico para detectar que observación es un dato anormal. Si se usan residuales estandarizados, entonces un dato con residual más allá de 2 ó -2 es considerado un "outlier" en el sentido vertical**.**

Plot de Residuales versus la variable predictora: Es usado para detectar datos anormales así como si la varianza de los errores es constante con respecto a la variable predictora.

Gráficas





Los puntos del plot de normalidad no caen cerca de una línea recta y en el extremo superior se detecta un "outlier". Similarmente, el histograma no es simétrico con un pico central y también muestra un "outlier" en el extremo superior. En conclusión, no hay normalidad de los errores. El plot de residuales versus el índice de la observación muestra que la observación 14 es un "outlier", pues el residual estandarizado cae más allá de dos. El plot de los residuales versus los valores predichos muestra que la varianza de los errores no es constante con respecto a la variable de respuesta, pues tiende ha aumentar cuando el valor de la variable de respuesta aumenta.

Modelos No Lineales y Transformaciones

Cuando se construyen modelos de regresión el objetivo es conseguir un modelo con R^2 alto que se aproxime a 100 %, asumiendo que no hay datos atípicos presentes. Si no se desea incluir variables predictoras adicionales en el modelo, hay dos alternativas:
Tratar de usar modelos polinómicos de grado mayor o igual a dos, y
Transformando las variables tanto la predictora como la de respuesta.

Regresión Cuadrática

Un modelo cuadrático es de la forma:

$$Y = a + bX + cX^2 + \varepsilon$$

donde a, b y c son constantes a estimar. Usando la técnica de mínimos cuadrados se pueden obtener fórmulas explícitas para calcular a, b y c.

En **MINITAB,** para obtener la ecuación del modelo cuadrático, hay que elegir la opción **Quadratic** en la ventana de diálogo de **Fitted Line Plot** que es una opción del menú **Regression.**

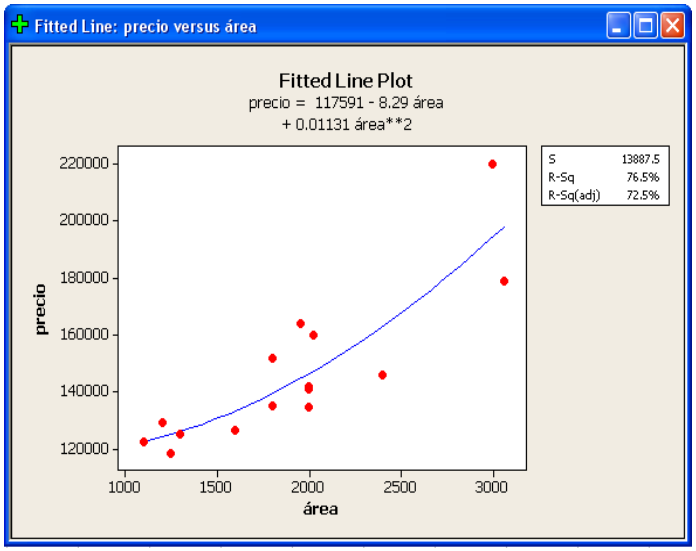
Ejemplo 9.2

Ajustar un modelo cuadrático para los datos del Ejemplo 9.1.

Polynomial Regression
precio = 117591 - 8.29281 area + 1.13E-02 area**2
R-Sq = 76.5 %
Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	7.52E+09	3.76E+09	19.4906	1.70E-04
Error	12	2.31E+09	1.93E+08		
Total	14	9.83E+09			

SOURCE	DF	Seq SS	F	P
Linear	1	7.24E+09	36.3308	4.25E-05
Quadratic	1	2.77E+08	1.43495	0.254083



Interpretación: El R^2 del modelo cuadrático es 76.5% comparado con 73.6% del modelo lineal (ver ejemplo 9.1), se ha ganado un 3% en confiabilidad, lo cual no es un aumento sustancial y se puede seguir usando un modelo lineal ya que hacer inferencias con él es mucho más simple que con un modelo cuadrático.

Modelos No lineales que pueden ser transformados en lineales

La segunda alternativa para aumentar el R^2 consiste en usar modelos no lineales que pueden ser convertidos en lineales, a través de transformaciones tanto de la variable independiente como dependiente.

Nombre del modelo	**Ecuación del Modelo**	**Transformación**	**Modelo Linealizado**
Exponencial	$Y=\alpha e^{\beta X}$	Z=Ln Y X=X	$Z=Ln\ \alpha +\beta X$
Logarítmico	$Y= \alpha +\beta Log\ X$	Y=Y W=Log X	$Y= \alpha +\beta W$
Doblemente Logarítmico	$Y=\alpha X^{\beta}$	Z=Log Y W=Log X	$Z= Log\ \alpha +\beta W$
Hiperbólico	$Y= \alpha +\beta/X$	Y=Y W=1/X	$Y= \alpha +\beta W$
Inverso	$Y=1/(\alpha +\beta X)$	Z=1/Y X=X	$Z=\alpha +\beta X$

Para predecir el valor de Y usando el modelo linealizado hay que aplicar la inversa de la transformación correspondiente al mismo.

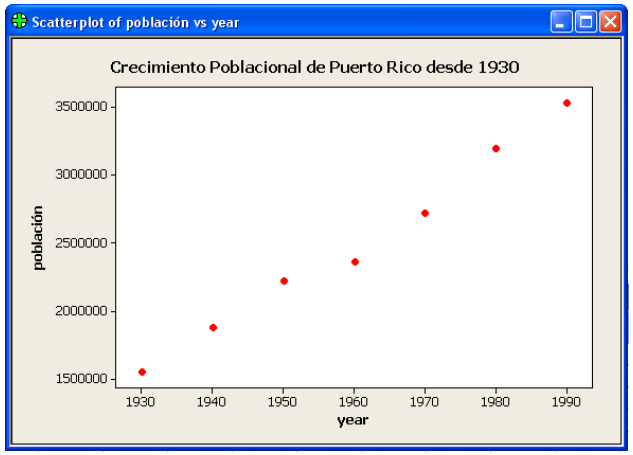
Ejemplo 9.3

Los siguientes datos representan como ha cambiado la población en Puerto Rico desde 1930 hasta 1990.

Año	Población
1930	1543913
1940	1869255
1950	2210703
1960	2349544
1970	2712033
1980	3196520
1990	3522037

Se desea establecer un modelo para predecir la población de Puerto Rico en el año 2000.

Solución:



$Poblac = \alpha e^{\beta year}$

$Ln(Poblac) = -11.4 + 0.0133\ year$

$R^2 = 98.9\%$

$$Ln(Poblac) = -11.4 + 0.0133 * 2000 = -11.4 + 26.6 = 15.2$$

$$Poblac = e^{15.2} = 3{,}992{,}787$$

Regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras $X_1,\ldots X_p$, es de la siguiente forma:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_p X_p + \varepsilon$$

Las constantes $b_0, b_1, ..., b_p$, llamadas coeficientes de regresión, se estiman usando el método de mínimos cuadrados, y usando *n* observaciones de la forma $y_i, x_{i1}, x_{i2}, ..., x_{ip}$, donde $i = 1, ..., n$. La cantidad ε es una variable aleatoria con media 0 y varianza σ^2.

Interpretación del coeficiente de regresión estimado β_j

El estimado del coeficiente de regresión poblacional b_j, con $j = 1,..., p$, se representará por β_j. Este estimado indica el cambio promedio en la variable de respuesta Y cuando la variable predictora X_j cambia en una unidad adicional asumiendo que las otras variables predictoras permanecen constantes.

Ejemplo 9.4

Se desea explicar el comportamiento de la variable de respuesta IGS (Indice General del Estudiante admitido a la Universidad de Puerto Rico) de acuerdo a X_1 (puntaje en la parte de aptitud matemática del College Board), X_2 (puntaje en la parte de aprovechamiento matemático) y $X3$ (Tipo de Escuela; 1: Pública, 2: Privada). La muestra de 50 observaciones está disponible en el archivo **igs** de la página del texto.

Solución:

Regression Analysis: igs versus aptitud, aprovech, escuela

The regression equation is

igs = 136 + 0.0569 aptitud + 0.197 aprovech + 1.93 escuela

Predictor	Coef	SE Coef	T	P
Constant	135.93	24.50	5.55	0.000
aptitud	0.05688	0.03140	1.81	0.077
aprovech	0.19698	0.03152	6.25	0.000
escuela	1.933	3.091	0.63	0.535

S = 10.8896 R-Sq = 56.0% R-Sq(adj) = 53.2%

Ejemplo 9.4 (cont.)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	6952.0	2317.3	19.54	0.000
Residual Error	46	5454.8	118.6		
Total	49	12406.9			

Interpretación: El coeficiente de una variable predictora indica el cambio promedio en la variable de respuesta igs cuando, se incrementa en una unidad la variable predictora asumiendo que las otras variables permanecen constantes.

En este ejemplo, el aumento promedio en el igs es de 0.0569 por cada punto adicional en la parte de aptitud matemática, asumiendo que las otras dos variables permanecen constantes, asímismo el aumento promedio en el igs es de 0.197 por cada punto adicional en la parte de aprovechamiento matemático asumiendo que las otras variables permanezcan constantes y hay un aumento promedio de 1.93 en el igs cuando nos movemos de escuela pública a privada asumiendo que las otras variables permanecen constantes.

Estimación de la varianza σ^2

La estimación de la varianza de los errores es crucial para hacer inferencias acerca de los coeficientes de regresión. Si en nuestro modelo hay *p* variables predictoras entonces, es estimada por:

$$s^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n-p-1} = \frac{SSE}{n-p-1} = MSE$$

Aquí, SSE representa la suma de cuadrados del error y MSE representa el cuadrado medio del error.

Inferencia en regresión lineal múltiple

Prueba de hipótesis de que cada coeficiente de regresión es cero

En este caso la hipótesis nula es $H_0 : \beta_j = 0$ ($j = 1,..., p$), o sea, la variable X_j no es importante en el modelo, versus la hipótesis alterna $H_a : \beta_j \neq 0$, que significa que la variable X_j si es importante. La prueba estadística es la prueba de *t* dada por:

$$t = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)}$$

MINITAB da el valor de la prueba estadística y de los "p-values" correspondientes

Inferencia en regresión lineal múltiple

Prueba de hipótesis de que todos los coeficientes de regresión son ceros.

En este caso la hipótesis nula es $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$ *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 sirve*

La prueba estadística es la prueba de F que se obtiene al hacer la tabla del Análisis de varianza para la regresion múltiple.

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

Se distribuye como una F con grados de libertad en el numerador y grados de libertad en el denominador.

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.

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$. En este caso al modelo que tiene las variables se le llama el **modelo completo** y al modelo que queda, asumiendo que la hipótesis nula es cierta, se le llama **modelo reducido.**

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:

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

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_o en caso contrario se acepta.

Ejemplo

Usando los datos del Ejemplo 9.4, probar la hipótesis $H_0 : \beta_1 = \beta_2 = 0$, versus H_a: al menos uno de los dos: β_1 o β_2 no es cero. Interpretar sus resultados.

Solución:

En este caso p=3, k=2, p-k = 1, y de la tabla del análisis de varianza del Ejemplo 9.4, $SSR(C)$ = 6952 y $MSE(C)$ = 118.6. Para obtener $SSR(R)$, se hace la regresión simple entre Y = *igs* y X_3 = escuela y de la tabla del análisis de Varianza se obtiene $SSR(R)$ = 52.9. Luego la prueba de ***F* parcial** será igual a:

$F = \frac{29.08}{3.19958}$ F con 2 g.l en el numerador y 46 g.l en el denominador

Como Fp> F, **se rechaza la hipótesis nula** y se concluye, que al 5% de significación hay suficiente evidencia estadística para afirmar que al menos una de las dos variables (aptitud o aprovechamiento) influye en el comportamiento de la variable de respuesta Y.

Ejemplo 9.4 (cont.)

Aún cuando el R^2 es bajo del 56%, eligiendo el botón **Options de Regression** se puede predecir el *igs* de un estudiante para hacer predicciones de la variable de respuesta Y para valores dados de las variables predictoras.

Por ejemplo el *igs* estimado de un estudiante que obtuvo 600 puntos en la prueba de aptitud y 750 en la prueba de aprovechamiento y que proviene de escuela privada será 321.66, Hay un 95% de confianza de que el valor promedio del IGS de todos los estudiantes admitidos que tienen esas caracteristicas caiga entre 313 y 329. Hay un 95% de confianza de que un estudiante cualquiera admitido que tiene esas caracteristica caiga entre 298 y 345

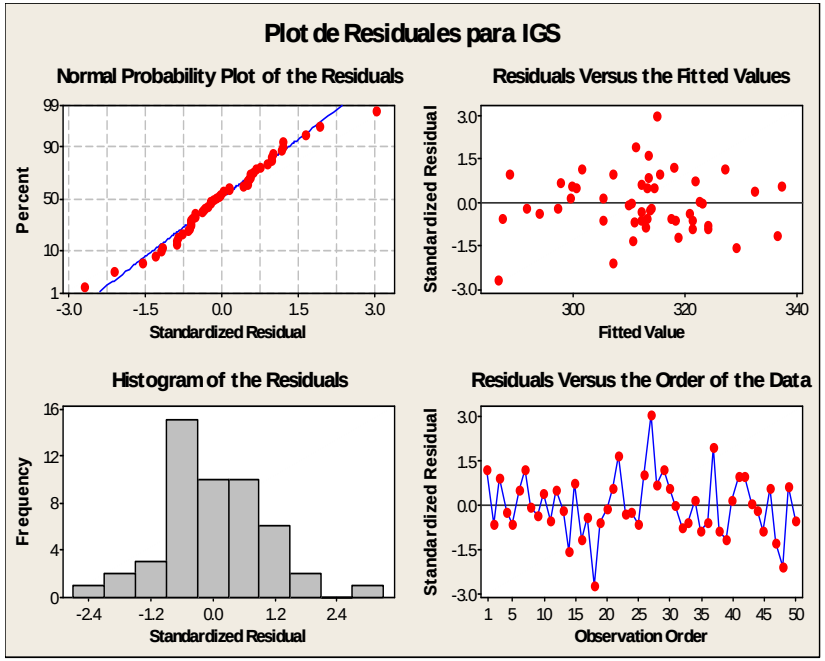
Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	321.66	4.05	(313.51, 329.81)	(298.28, 345.05)

Values of Predictors for New Observations

New Obs	escuela	aprovech	aptitud
1	2.00	750	600

Analisis de Residuales



Interpretación: *hay algo de normalidad en la distribución de los errores, debido a que los puntos no se alejan mucho de una línea recta en el primer plot y algo de simetría que se puede ver en el segundo.*

Sin embargo es clara la presencia de los "outliers" en ambos extremos, lo cual afecta la condición de normalidad.

El plot de residuales versus el orden de la observación y el plot de residuales versus valores predichos ("fits"). sugiere que las observaciones 18, 27 y 48 son "outliers" en el sentido vertical.

El plot de residuales versus valores predichos sugiere que la varianza de los errores es constante, porque no hay un patrón definido que siguen los puntos.

Selección de variables en Regresión Múltiple

Una buena propiedad de un modelo de regresión lineal es que permita explicar el comportamiento de la variable de respuesta Y lo mejor posible, haciendo uso del menor número de variables predictoras posibles, esta propiedad es llamada “parsimonía”.

Existen dos métodos generales de lograr este objetivo: los métodos “stepwise” y el método de los mejores subconjuntos.

Los métodos "stepwise"

- Método de eliminación hacia atrás (“Backward Elimination”)
- Método de Selección hacia adelante (“Forward Selection”):
- Método Paso a Paso ("Stepwise")

Método de eliminación hacia atrás

Aquí en el paso inicial se incluyen en el modelo a todas las variables predictoras y en cada paso se elimina la variable cuyo “p-value” es más grande para la prueba de *t* o cuyo valor de la prueba *t* menor que 2 en valor absoluto.

Una variable que es eliminada del modelo ya no puede volver a entrar en un paso subsiguiente.

El proceso termina cuando todos los “p-values” son menores que .05, o cuando todos los valores de la prueba t son mayores que 2 en valor absoluto.
Lo anterior también se puede hacer con una prueba *F*-parcial, puesto que $F = t^2$ (cuando el numerador tiene grados de libertad igual a 1). Luego, el método terminará cuando todas las *F* son mayores que 4.

Ejemplo 9.6. El conjunto de datos **grasa** contiene 13 variables que sirven para predecir el porcentaje de grasa en el cuerpo humano.

Columna	Nombre	
C1	grasa	VARIABLE DE RESPUESTA
C2	edad	en años
C3	peso	en libras
C4	altura	en pulgadas
C5	cuello	en cms
C6	pecho	en cms
C7	abdomen	en cms
C8	cadera	en cms
C9	muslo	en cms
C10	rodilla	en cms
C11	tobillo	en cms
C12	biceps	en cms
C13	antebrazo	en cms
C14	muñeca	en cms

Se tomaron las mediciones en 250 sujetos.

Stepwise Regression: grasa versus edad, peso, ...

Backward elimination. Alpha-to-Remove: 0.05

Response is grasa on 13 predictors, with N = 252

Step	1	2	3	4	5	6	7
Constant	-18.19	-17.93	-19.69	-26.00	-23.30	-22.66	-33.26
edad	0.062	0.063	0.062	0.065	0.063	0.066	0.068
T-Value	1.92	2.00	2.00	2.11	2.06	2.14	2.21
P-Value	0.056	0.046	0.046	0.036	0.041	0.034	0.028
peso	-0.088	-0.088	-0.093	-0.107	-0.098	-0.090	-0.119
T-Value	-1.65	-1.70	-1.96	-2.55	-2.42	-2.25	-3.51
P-Value	0.100	0.091	0.051	0.011	0.016	0.025	0.001
altura	-0.070	-0.069	-0.064				
T-Value	-0.72	-0.72	-0.69				
P-Value	0.469	0.470	0.493				
cuello	-0.47	-0.47	-0.48	-0.47	-0.49	-0.47	-0.40
T-Value	-2.02	-2.06	-2.08	-2.05	-2.18	-2.08	-1.83
P-Value	0.044	0.040	0.039	0.042	0.030	0.039	0.068
pecho	-0.024	-0.024					
T-Value	-0.24	-0.25					
P-Value	0.810	0.805					

abdomen	0.955	0.954	0.944	0.958	0.949	0.945	0.918
T-Value	11.04	11.09	12.51	13.16	13.18	13.13	13.21
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
cadera	-0.21	-0.21	-0.20	-0.18	-0.18	-0.20	
T-Value	-1.42	-1.42	-1.41	-1.29	-1.32	-1.41	
P-Value	0.156	0.156	0.161	0.199	0.189	0.159	
muslo	0.24	0.24	0.25	0.26	0.27	0.30	0.22
T-Value	1.64	1.72	1.81	1.94	1.99	2.34	1.91
P-Value	0.103	0.086	0.072	0.054	0.048	0.020	0.057
rodilla	0.02						
T-Value	0.06						
P-Value	0.950						
tobillo	0.17	0.18	0.18	0.18			
T-Value	0.79	0.81	0.82	0.85			
P-Value	0.433	0.419	0.412	0.396			
biceps	0.18	0.18	0.18	0.19	0.18		
T-Value	1.06	1.06	1.05	1.10	1.06		
P-Value	0.290	0.289	0.297	0.271	0.289		
antebraz	0.45	0.45	0.45	0.45	0.45	0.52	0.55
T-Value	2.27	2.29	2.28	2.31	2.31	2.77	2.99
P-Value	0.024	0.023	0.023	0.022	0.022	0.006	0.003
S	4.31	4.30	4.29	4.28	4.28	4.28	4.29
R-Sq	74.90	74.90	74.90	74.85	74.77	74.66	74.45
R-Sq(adj)	73.53	73.64	73.75	73.81	73.84	73.82	73.71
C-p	14.0	12.0	10.1	8.5	7.2	6.4	6.3

Step	8	9	10
Constant	-38.32	-30.97	-34.85
edad	0.063	0.041	
T-Value	2.04	1.43	
P-Value	0.042	0.154	
peso	-0.136	-0.111	-0.136
T-Value	-4.15	-3.68	-5.48
P-Value	0.000	0.000	0.000
altura			
T-Value			
P-Value			
cuello			
T-Value			
P-Value			
pecho			
T-Value			
P-Value			
abdomen	0.912	0.939	0.996
T-Value	13.07	13.69	17.76
P-Value	0.000	0.000	0.000
cadera			
T-Value			
P-Value			

muslo	0.22		
T-Value	1.89		
P-Value	0.060		
rodilla			
T-Value			
P-Value			
tobillo			
T-Value			
P-Value			
biceps			
T-Value			
P-Value			
antebraz	0.49	0.51	0.47
T-Value	2.68	2.78	2.60
P-Value	0.008	0.006	0.010
muneca	-1.78	-1.83	-1.51
T-Value	-3.60	-3.68	-3.40
P-Value	0.000	0.000	0.001
S	4.31	4.33	4.34
R-Sq	74.10	73.72	73.50
R-Sq(adj)	73.46	73.19	73.07
C-p	7.7	9.2	9.3

Interpretación: El método termina en 10 pasos.
El proceso termina, porque todos los "p-values" son menores que 0.05 o las pruebas t en valor absoluto son mayores que 2. La primera variable eliminada del modelo es rodilla, cuyo valor de la prueba t, 0.06, es el más pequeño de todos, luego se eliminan, pecho, altura, tobillo, biceps, cadera, cuello, muslo y edad en ese orden. El mejor modelo para predecir el porcentaje de grasa en el cuerpo será el que incluye a las variables:peso, circunferencia de abdomen, nuñeca y antebrazo.

El mejor modelo será:

Grasa= 34.85 -.136 peso+ .996 abdomen +0.47 antebrazo - 1.51muñeca

El cual tiene un R2 de 73.50, mientras que el modelo completo con 13 variable predictoras tiene un R2 de 74.90%, se ha perdido un 1.40% de confiablidad en las predicciones pero se ha economizado 9 variables, lo cual es más conveniente.

Método de Selección hacia adelante

Aquí en el paso inicial se considera una regresión lineal simple que incluye a la variable predictora que da la correlación más alta con la variable de respuesta.

Se incluye una segunda variable en el modelo, que es aquella variable dentro de las no incluidas aún, que da el "p-value" más bajo para la prueba *t* o el valor de la prueba de *t* más grande en valor absoluto. Y así se siguen incluyendo variables, notando que una vez que ésta es incluida ya no puede ser sacada del modelo.

El proceso termina cuando los "p-values" para la prueba *t* de todas las variables que aún no han sido incluidas son mayores que .05 ó la prueba de *t* es menor que 2 para dichas variables. Si se usa la prueba de *F*, entonces el proceso termina cuando todas las *F* son menores que 4.

Ejemplo (cont). En el primer paso se halla la regresión simple con la variable predictora más altamente correlacionada con la variable de respuesta. En este caso, es *abdomen* que tiene correlación 0.803 con *grasa.*

La segunda variable que entra al modelo es peso porque es aquella con el valor de *t* más grande en valor absoluto entre las doce variables que aún no estaban incluidas.

La salida en MINITAB es como sigue:

Stepwise Regression: grasa versus edad, peso, ...

Forward selection. Alpha-to-Enter: 0.05

Response is grasa on 13 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
antebraz				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
C-p	72.9	20.7	14.2	9.3

La variable antebrazo es la ultiam en entra al modelo porque es aquella con el valor de *t* más grande en valor absoluto entre todas las variables que aún no estaban incluidas. Aquí termina el proceso porque al hacer las regresiones de grasa con las cuatro variables consideradas hasta ahora y cada una de las 9 variables no incluidas hasta ahora se obtienen "p-values" para la prueba *t* mayores de 0.05.

Regression Analysis

The regression equation is

grasa = - 34.9 + 0.996 abdomen - 0.136 peso - 1.51 muñeca + 0.473 antebrazo

Predictor	Coef	StDev	T	P
Constant	-34.854	7.245	-4.81	0.000
abdomen	0.99575	0.05607	17.76	0.000
peso	-0.13563	0.02475	-5.48	0.000
muñeca	-1.5056	0.4427	-3.40	0.001
antebraz	0.4729	0.1817	2.60	0.010

S = 4.343 R-Sq = 73.5% R-Sq(adj) = 73.1%

Método Paso a Paso

Es una modificación del método “Forward”, donde una variable que ha sido incluida en el modelo en un paso previo puede ser eliminada posteriormente.

En cada paso se cotejan si todas las variables que están en el modelo deben permanecer alli. La mayoría de las veces, pero no siempre, los tres métodos dan el mismo resultado para el mejor modelo de regresión.

En **MINITAB**, la opción ***Stepwise*** del submenú **Regression** selecciona el mejor modelo de regresión usando los métodos "**Stepwise".**

C) Usando el método "Stepwise".

sigue la secuencia

STAT ▸ Regression ▸ Stepwise ▸

Methods y luego se elige Stepwise.

Alpha-to-Enter y **Alpha to-Remove.**

Para el conjunto de datos **grasa** el

Método "stepwise" usa

Alpha-to-Enter = 0.10 y

Alpha to-Remove = 0.15.

El alpha to remove debe ser mayor o igual que el alpha to enter

Stepwise Regression: grasa versus edad, peso, ...

Alpha-to-Enter: 0.1 Alpha-to-Remove: 0.15

Response is grasa on 13 predictors, with N = 252

Step	1	2	3	4	5
Constant	-39.28	-45.95	-27.93	-34.85	-30.65
abdomen	0.631	0.990	0.975	0.996	1.008
T-Value	22.11	17.45	17.37	17.76	17.89
P-Value	0.000	0.000	0.000	0.000	0.000
peso		-0.148	-0.114	-0.136	-0.123
T-Value		-7.11	-4.84	-5.48	-4.75
P-Value		0.000	0.000	0.000	0.000
muneca			-1.24	-1.51	-1.25
T-Value			-2.85	-3.40	-2.66
P-Value			0.005	0.001	0.008
antebraz				0.47	0.53
T-Value				2.60	2.86
P-Value				0.010	0.005
cuello					-0.37
T-Value					-1.65
P-Value					0.100
S	4.88	4.46	4.39	4.34	4.33
R-Sq	66.17	71.88	72.77	73.50	73.79
R-Sq(adj)	66.03	71.65	72.44	73.07	73.26
C-p	72.9	20.7	14.2	9.3	8.6

- La primera variable seleccionada es abdomen, porque es la que tiene la prueba estadistica de t mas grande(o p-value mas pequeno). Es decir, abdomen es la mas importante para predecir el porcentaje de grasa. Las segunda variable mas importantes es peso, la tercera, muneca, la cuarta antebrazo y la quinta cuello. El metodo para en el paso 5 porque ninguna de las variables que aun no se han escogio som importantes para predcir grasa, es decir p-values debe ser mayor que el 10%(f-to-enter).
- Ademas en cada paso no se elimino ninguna variable que ya habia sido escogida previamente

Método de los mejores subconjuntos.

La opción **Best Subsets** del submenú **Regression** del menú **Stat** se usa para seleccionar los mejores modelos para un número dado de variables de acuerdo a 3 criterios:

El coeficiente de Determinación. El mejor modelo es aquel con R^2 más alto pero con el menor número de variables posibles. $R^2 = \frac{SSR}{SST}$

El coeficiente de Determinación Ajustado. Es una variante del R^2 y que a diferencia de éste no aumenta necesariamente al incluir una variable adicional en el modelo.

$$R^2_{Ajust} = \frac{MSR}{MST} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

El **Coeficiente Cp de Mallows**. El mejor modelo es aquel para el cual se cumple aproximadamente , pero con $C_p = p+1$ el menor número de variables posibles. Notar que la igualdad anterior también se cumple cuando se usa el modelo completo.

$$C_p = \frac{SSE_p}{s^2} + 2(p+1) - n$$

Best Subsets Regression: grasa versus edad, peso, ...

Response is grasa

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	edad	peso	altura	cuello	cadera	pecho	abdomen	muslo	rodilla	biceps	antebrazo	muñeca
1	66.2	66.0	72.9	4.8775							X					
2	71.9	71.7	20.7	4.4556	X						X					
3	72.8	72.4	14.2	4.3930	X						X					X
4	73.5	73.1	9.3	4.3427	X						X					X X
5	73.8	73.3	8.6	4.3276	X	X					X					X X
6	74.1	73.5	7.7	4.3111	X X						X	X				X X
7	74.4	73.7	6.3	4.2906	X X		X		X	X	X					X X
8	74.7	73.8	6.4	4.2819	X X		X		X X X							X X
9	74.8	73.8	7.2	4.2808	X X		X		X X X							X X X
10	74.8	73.8	8.5	4.2832	X X		X		X X X		X X X X					
11	74.9	73.7	10.1	4.2879	X X 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 X X					
13	74.9	73.5	14.0	4.3053	X X X X X X X X X X X X X											

Resultados para el problema anterior

De acuerdo al R^2 el mejor modelo podría ser aquel con las dos variables predictoras peso y abdomen que aún cundo su R^2 es de 71.9 está cerca del mayor posible que es de 74.9 y además es donde el R^2 ha tenido un mayor incremento. Un resultado similar cuando se usa el R^2 ajustado. De acuerdo al C_p de Mallows, el mejor modelo es aquel que tiene las siguientes 6 variables predictoras: edad, peso, muslo, abdomen, antebrazo y muneca con un valor de C_p=7.7 muy próximo a p+1=7.