# CAPÍTULO 3
# DIAGNÓSTICOS DE REGRESIÓN

**Edgar Acuña Fernández**

**Departamento de Matemáticas**
**Universidad de Puerto Rico**
**Recinto Universitario de Mayagüez**

# 3.1 "Outliers", puntos de leverage alto y valores influenciales

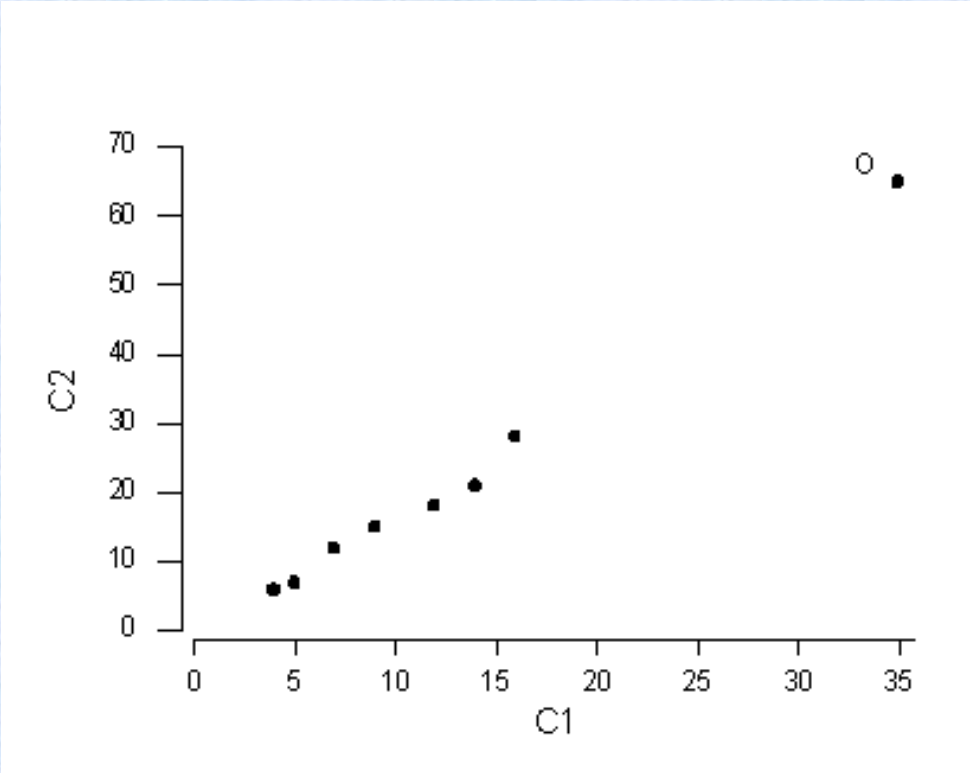
Una observación $(y^*, x^*_1, \ldots\ldots x^*_p)$ es considerado un **"outlier"** si está bastante alejado de la mayoría de los datos sea en la dirección vertical o en la horizontal.

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.

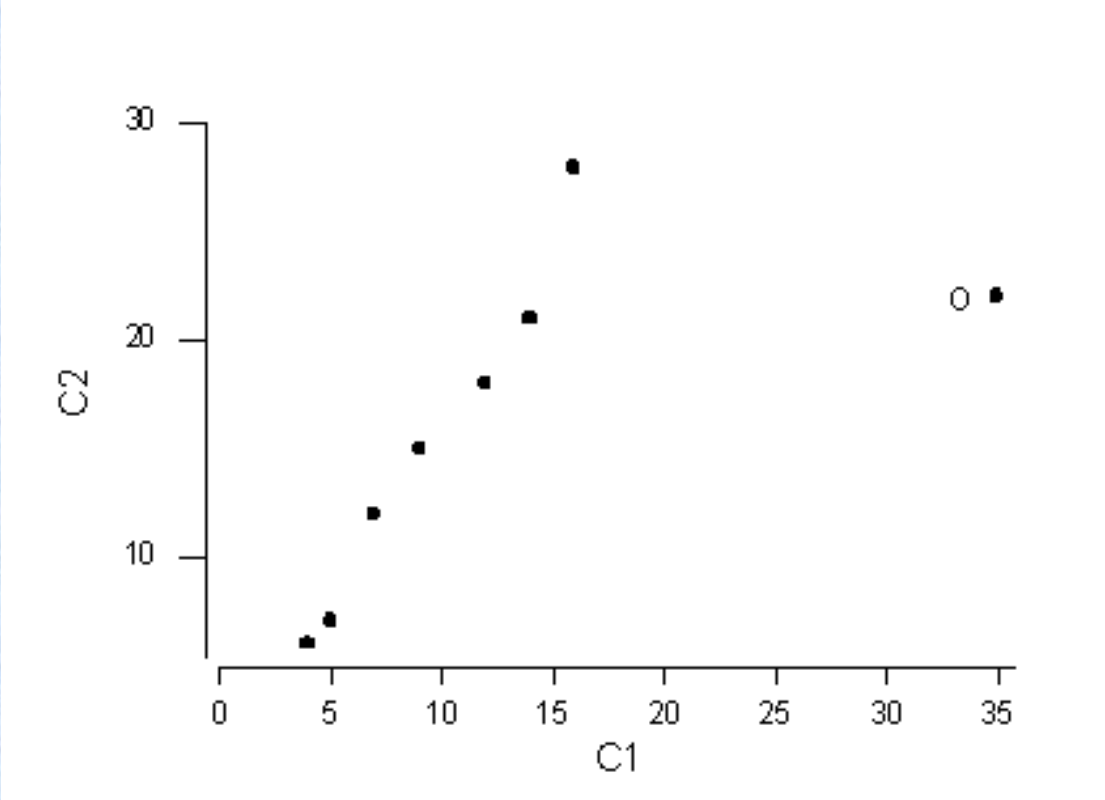
# Valor Influencial

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

# Ejemplo de una observación que es “outlier” y punto leverage alto pero que no es influencial.



# Ejemplo de una observación que es punto de leverage alto y que también es influencial.




Este punto tendrá un gran efecto sobre el $R^2$ y el cambio drástico en la pendiente.

# 3.2 Residuales y detección de "outliers".

- Consideremos el modelo

  $\mathbf{Y=XB+e}$ , donde $E(\mathbf{e})=\mathbf{0}$ y $Var(\mathbf{e})=\sigma^2\mathbf{I}$

  Luego $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$,donde $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

  la matriz HAT (sombrero) H de $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{HY}$

  actúa como una transformación de $\mathbf{Y}$ a $\hat{\mathbf{Y}}$ .

- En particular $\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$

  $h_{ij}$ es el elemento de la matríz H que está en la i-ésima fila y j-ésima columna. Así

  $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = \mathbf{(I - H)Y}$ donde $\hat{e}_i = y_i - \sum_{j=1}^{n} h_{ij} y_j$

# 3.2.1 Media y Varianza del vector de residuales

i $E(\hat{\mathbf{e}}) = (\mathbf{I} - \mathbf{H})E(\mathbf{Y}) = \mathbf{0}$

ii $Var(\hat{\mathbf{e}}) = Var[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = \sigma^2(\mathbf{I} - \mathbf{H})$ , **I-H** es simétrica e idempotente.

En particular

$Var(\hat{e}_i) = \sigma^2(1 - h_{ii})$ Se estima por $s^2(1\text{-}h_{ii})$.

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

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 si.

# 3.2.2 Residuales Estudentizados internamente

- Se define por $r_i^* = \frac{\hat{e}_i}{\sigma\sqrt{1-h_{ii}}}$

También son llamados **residuales estandarizados**.

- La covarianza de los residuales estudentizados es igual a

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

# 3.2.4 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 restantes. Luego, usando la identidad de Gauss

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

- Se obtienen las siguientes relaciones entre $\hat{\boldsymbol{\beta}}$ y $\hat{\boldsymbol{\beta}}_{(i)}$ y entre $s^2$ y $s^2_{(i)}$

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

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

# La identidad de Gauss

- Es un caso particular de la **Identidad de Sherman-Morrison-Woodburry** (1950)

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

Donde:

**A** es una matríz cuadrada nosingular n x n, y

**u** y **v** son dos vectores de dimensión n.

**En nuestro caso, A=X'X** y $\mathbf{u} = \mathbf{v} = \mathbf{x_i}$ y $\mathbf{X}'_{(i)}\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i$

Donde $\mathbf{x}'_i$ es la i-ésima fila de **X**

# Varianza del Residual $y_i - \tilde{y}_i$

Si $\tilde{y}_i$ representa el valor estimado de la variable de respuesta para la i-ésima observación

$$\tilde{y}_i = \mathbf{x'}_i \, \hat{\boldsymbol{\beta}}_{(i)}$$

$y_i$ y $\tilde{y}_i$ son independientes, (la i-ésima observación no fue usada en la estimación del modelo )

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

# Residual Estudentizado Externamente

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

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

$t_i$ 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

$$y_i - \tilde{y}_i = \frac{\hat{e}_i}{1 - h_{ii}}$$

- Relación entre los distintos tipos de residuales

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

# 3.2 Diagnósticos para detectar "outliers" y puntos 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 de leverage alto**" y pudiera ser **influencial**

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

# Otros Diagnósticos

**i) La Distancia Cook (**Cook, 1977**)**

Mide el cambio que ocurriría en el vector de coeficientes estimados de regresión si la i-ésima observación fuera omitida. Se calcula por:

$$CD_i^2 = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{ps^2} = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{ps^2} = r*_i^2 \frac{h_{ii}}{p(1-h_{ii})}$$

Un $CD_i^2 > 1$ indica que la i-ésima observación es potencialmente influencial.

Una observación con $CD_i^2 < 0.1$ no merece ninguna discusión

si $CD_i^2 < .0.5$ merece un poco de atención. En particular, una observación con $CD_i^2 >$ F(0.50,p,n-p) es considerado como un valor **influencial.**

# Otros Diagnósticos

***ii)*** **DFFITS (**Belsley, Kuh, y Welsch, 1980).

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

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

Notar que:

$$CD_i^2 = \frac{r_i^2}{pt_i^2} DFFITS_i^2$$

# Otros Diagnósticos

**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

$$(DFBETAS)_{ji} = \frac{\beta_j - \beta_{j,(i)}}{s_{(i)}\sqrt{c_{jj}}} \qquad \text{i=1,..,}n\text{, j=0,…,}p$$

Donde $c_{jj}$ es el $j$-ésimo elemento de la diagonal de $\mathbf{(X'X)^{-1}}$.

- Si $|DFBETAS|_{ji} > \frac{2}{\sqrt{n}}$ para algun j entonces la i-esima observacion es posiblemente un valor influencial.

# Otros Diagnósticos

**iv) COVRATIO (**Belsley, Kuh, y Welsch, 1980**)** Mide el efecto en la variabilidad de los coeficientes de regresión al remover la i-ésima observación.

$$COVRATIO_i = \frac{\det[s^2_{(i)}(X'_{(i)}X_{(i)})^{-1}]}{\det[s^2(X'X)^{-1}]} \qquad i = 1,\ldots,n.$$

Usando propiedades de determinantes se tiene

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

Si $(\text{COVRATIO})_i > 1+3p/n$ o si $(\text{COVRATIO})_i < 1-3p/n$ entonces la *i*-ésima observación tiene un valor influencial grande.

# 3.3 Plot de Residuales para detectar casos influenciales

- Se usan para estudiar el efecto de añadir una nueva variable predictora en un modelo.
- Permiten detectar la presencia de casos influenciales.

- Para 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 $j$

# Residuales

Definamos los siguientes residuales

i $\hat{e}_{Y/X_{-j}} = (I - H_{-j})Y$ se han considerado en el modelo todas las predictoras excepto $x_j$

ii $\hat{e}_{Y/x,X_{-j}} = (I - H)Y$ están consideradas todas las variables predictoras

iii $\hat{e}_{x_j/X_{-j}} = (I - H_{-j})X_j$ son los residuales de la regresión de $x_j$ versus las otras variables predictoras.

# Plot de residuales versus la variables predictoras.

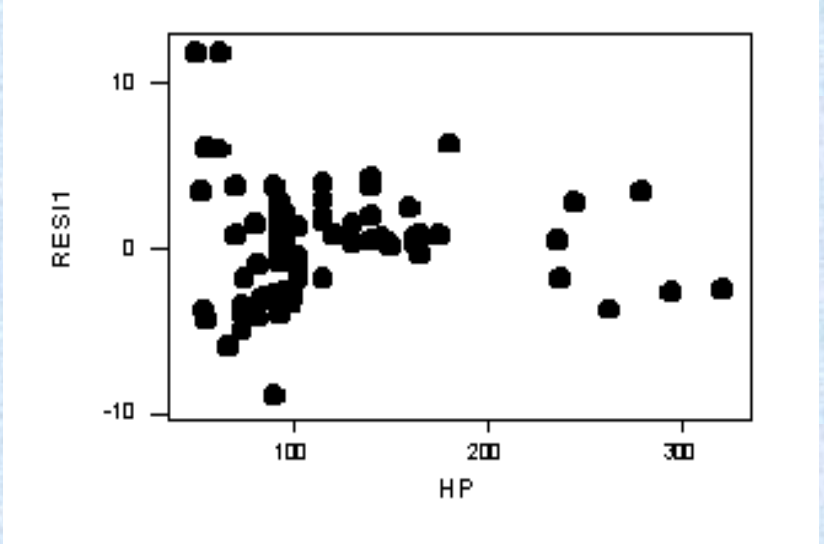
- $\widehat{e}_{Y/x_j, X_{-j}}$ versus $x_j$

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.**

Este plot no sirve para cuantificar la **importancia** de $x_j$ en el modelo.

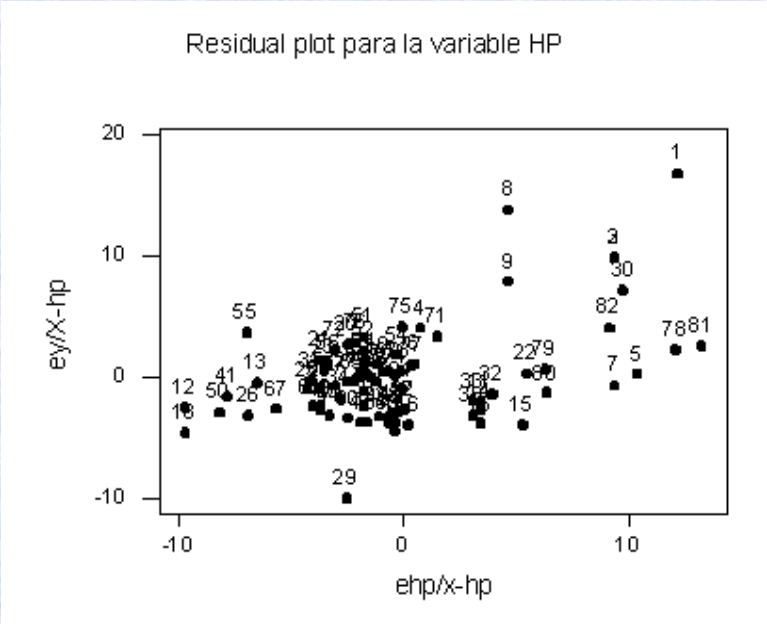
**Plot de residuales versus las predictora HP de Millaje.**



# Plots de regresión parciales (plot de variable añadida)

$\widehat{e}_{Y/X_{-j}}$ versus $\widehat{e}_{x_j/X_{-j}}$

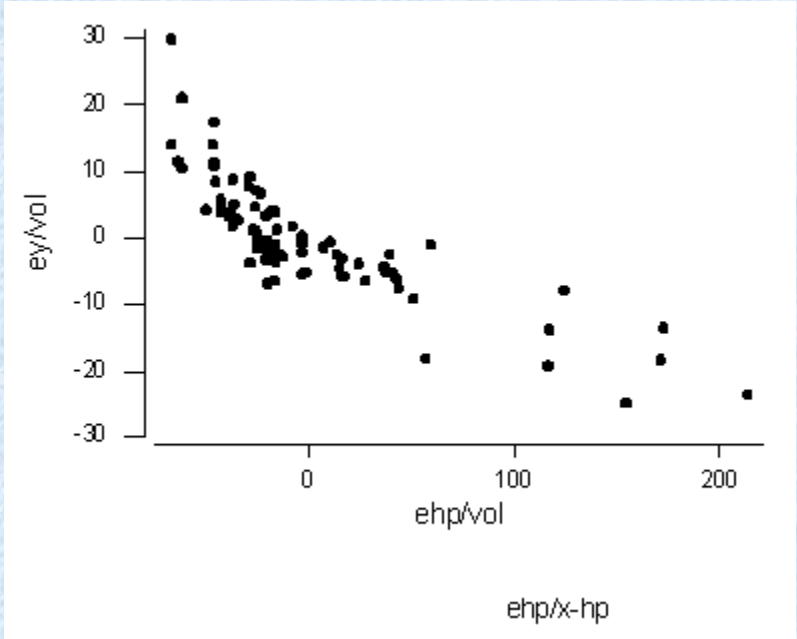
Se plotea los residuales de la regresión de y considerando todas las variables predictoras excepto $x_j$ versus los residuales de la regresión de $x_j$ contra todas las variables predictoras distintas a ella.



# Plots de regresión parciales (plot de variable añadida)

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

La variable VOL puede entrar al modelo en forma lineal



# Plot de residuales parciales o de residuales más componente

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

Es más efectivo para detectar nolinealidad que el plot de regresión parcial No es muy adecuado para detectar casos influenciales.

# Plot de residuales parciales aumentados

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

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

# 3.4 Plot de residuales para detectar Normalidad

- La suposición de la normalidad de los errores es bién importante para el proceso de hacer inferencia en regresión lineal múltiple.
- Puede ser cotejado haciendo un plot de normalidad para los errores estudentizados internamente.
- El plot de normalidad consiste en un plot de los **scores normales** (estadísticos de orden normales) versus los residuales estandarizados ordenados.

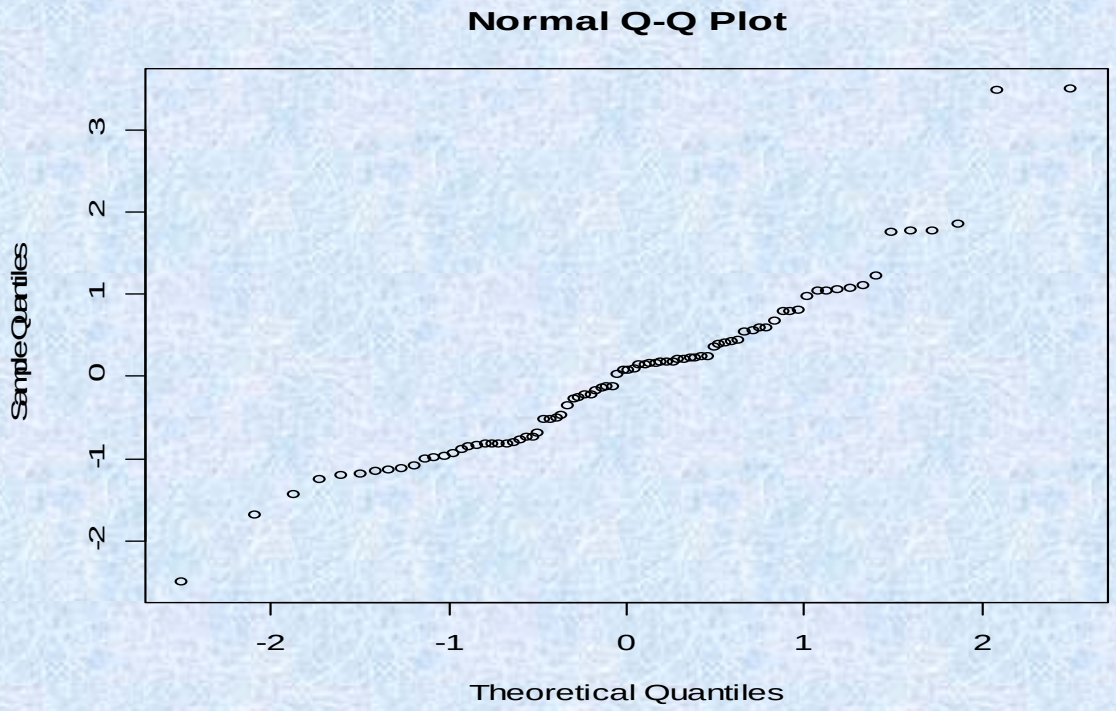
# Score Normal

- 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.

# Plot de normalidad acompañado de pruebas noparamétricas para detectar normalidad.




El “p-value” de la prueba de Kolmogrov-Smirnov es mayor que 0.05 por lo tanto se acepta la hipótesis de que hay normalidad de los residuales.

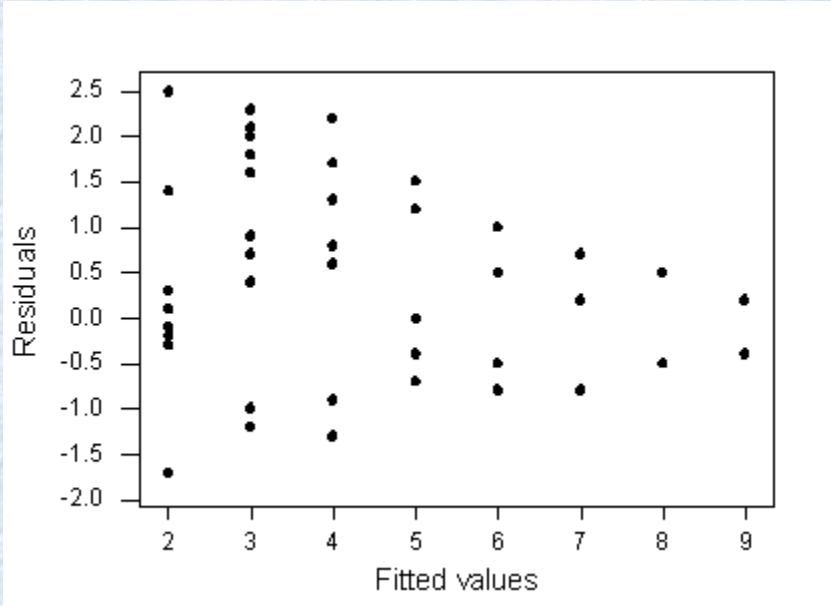
# 3.5 Detectando varianza no constante

La suposición de que en el modelo de regresión lineal múltiple, los errores tienen varianza constante es importante para que los estimadores mínimos cuadráticos sean óptimos.

- La 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.

# La varianza de los errores no es constante

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




Este plot es típico cuando los errores siguen una distribución Poisson o log-normal.

# Remedios cuando la varianza poblacional $\sigma^2$ no es constante

- Usar mínimos cuadrados ponderados donde los pesos que se usan son hallados en base a los datos tomados.

- Transformar la variable de respuesta Y usando tranfomación que estabiliza la varianza

# 3.6 Errores correlacionados en Regresión

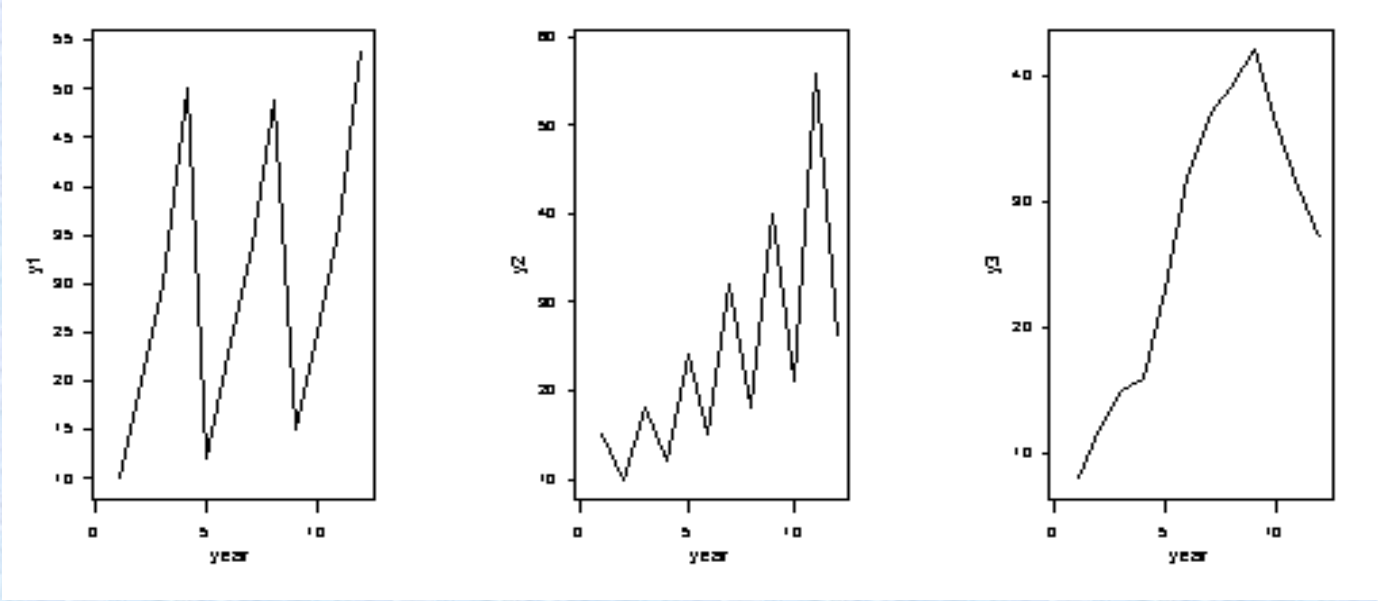
- Una de las suposiciones que se hace en regresión lineal es que los errores no se correlacionan entre si
- $\text{Cov}(e_i, e_j) = \text{E}(e_i e_j) = 0$ para $i \neq j$.

# Autocorrelación

- Cuando la variable predictora es tiempo, pudiera ocurrir que $E(e_i, e_{i+k}) \neq 0$ para un cierto $k$ en este caso se dice que los errores tiene una correlación serial y estan autocorrelacionados .

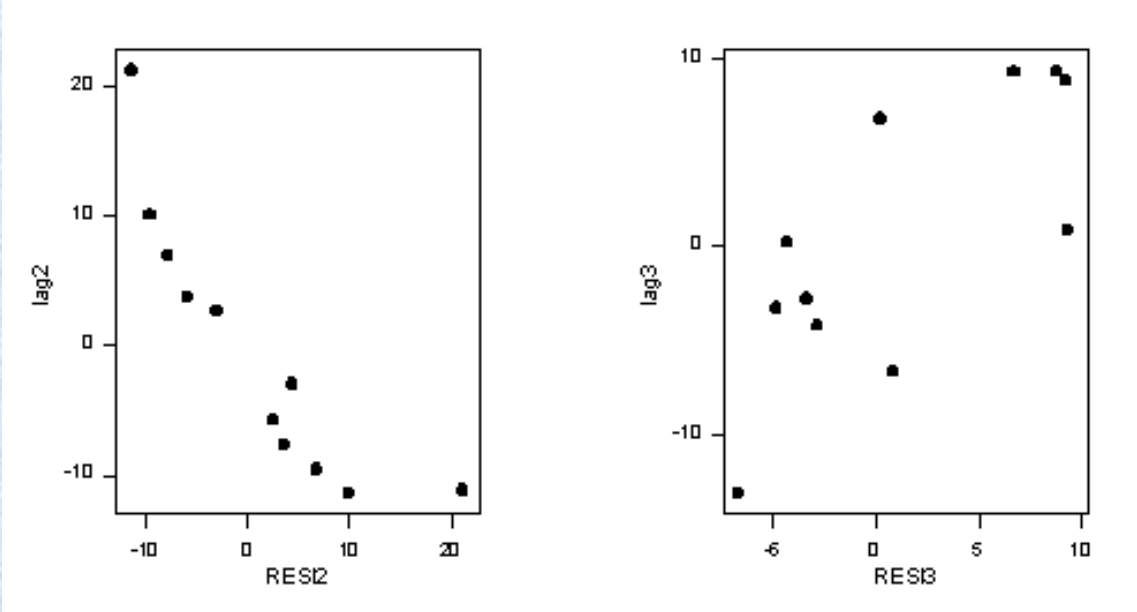
- **Gráficamente,** cuando los residuales cambian frecuentemente de signo hay autocorrelación negativa y si hay un conglomerado de residuales de un mismo signo antes de cambiar a otro entonces la autocrrelación es positiva.

# Gráfica de las 3 series de tiempo




- En los dos primeros plots la autocorrelación es negativa y en la última es positiva

# Plot de los residuales en el tiempo *t* versus los residuales en el tiempo *t*-1.



# La prueba de Durbin-Watson

- Se usa para detectar si hay una positiva correlación serial de orden uno.
- $H_o$: $\rho = 0$ vs $H_a$: $\rho > 0$.
- La prueba está dada por $D = \dfrac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$

- Se rechaza Ho si $D<D_L$
- Se acepta Ho si $D>D_U$
- La prueba no lleva a ninguna conclusión si $D_L<D<D_U$.

Los valores límites $D_L$ y $D_U$ son leidos de tabla de Durbin-Watson.

# Prueba de dos lados

Se tienen las hipótesis:

$H_o$: $\rho = 0$, versus $H_a$:$\rho \neq 0$

entonces

- Se rechaza $H_o$: si $D<D_L$ ó 4-$D<D_L$, al nivel de significación de $2\alpha$.
- No se rechaza $H_{o.}$ si $D>D_U$ y 4-$D>D_U$
- Para cualquier otro valor de $D$ la prueba no llega a ninguna conclusión.