Modelos Lineales de Regresión

Análisis de Residuos

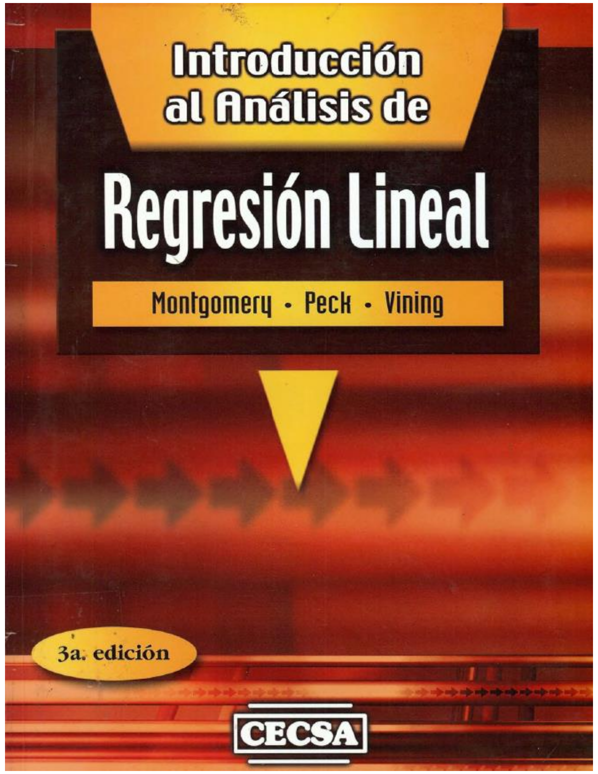
Balanceo e Influencia

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

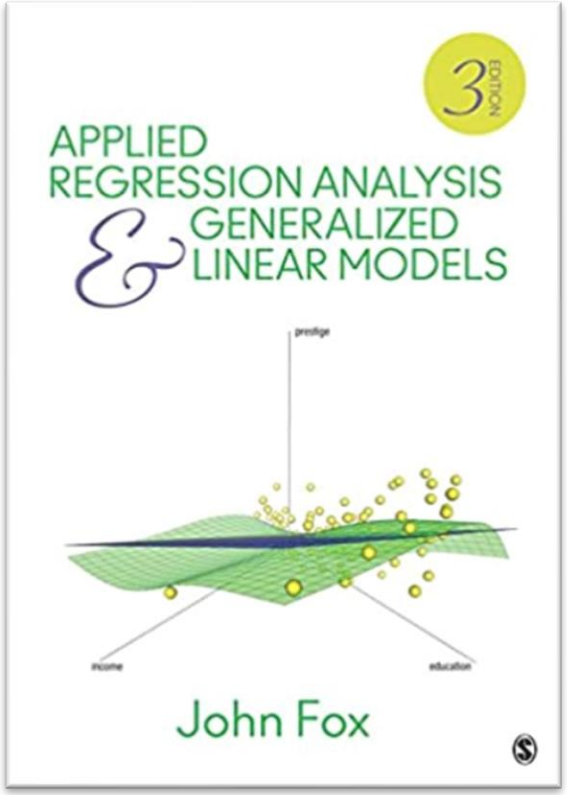
Mariano Bonoli Escobar
Javier Gil

Residuos, Balanceo e Influencia

Bibliografía



Capítulo 6



Capítulo 11

Mariano Bonoli Escobar
Javier Gil

Análisis de Residuos – Diagnóstico

Objetivos

- Análisis de Especificación
- Análisis de Outliers
- Análisis de Influencia

Mariano Bonoli Escobar
Javier Gil

Análisis de Especificación

Mariano Bonoli Escobar
Javier Gil

Análisis de especificación

Impacto de la incorrecta especificación en las perturbaciones

Modelo Verdadero

$$\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + \boxed{\beta_2 x_{2i} + \tilde{\varepsilon}_i}$$

$$\mathrm{E}(\tilde{\varepsilon}_i) = 0 \quad \mathrm{Var}(\tilde{\varepsilon}_i) = \sigma^2$$

$$\tilde{y}_i = \beta_0 + \beta_1 e^{x_i} + \tilde{\varepsilon}_i = \beta_0 + \beta_1^* x_i - \boxed{\beta_1^* x_i + \beta_1 e^{x_i} + \tilde{\varepsilon}_i}$$

$$\mathrm{E}(\tilde{\varepsilon}_i) = 0 \quad \mathrm{Var}(\tilde{\varepsilon}_i) = \sigma^2$$

Modelo Utilizado

$$\tilde{y}_i = \beta^*{}_0 + \beta_1^* x_i + \boxed{\tilde{\varepsilon}_i^*}$$

$$\mathrm{E}(\tilde{\varepsilon}_i^*) = \beta_2 x_{2i}$$

$$\tilde{y}_i = \beta^*{}_0 + \beta_1^* x_i + \boxed{\tilde{\varepsilon}_i^*}$$

$$\mathrm{E}(\tilde{\varepsilon}_i^*) = -\beta_1^* x_i + \beta_1 e^{x_i}$$

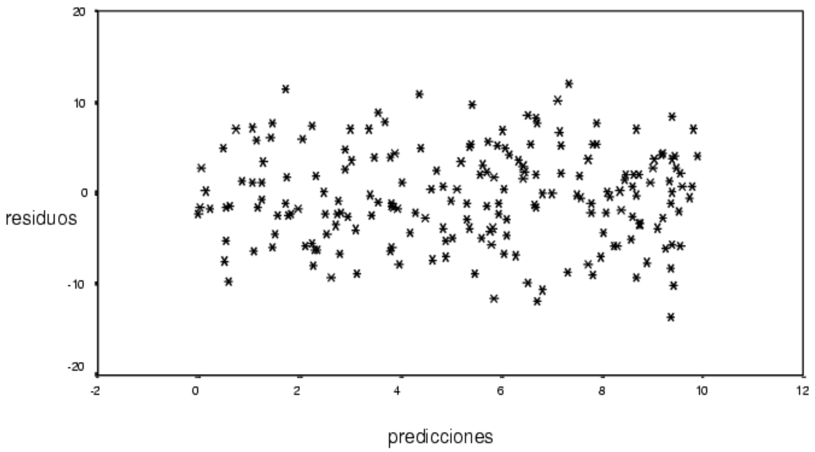
Una incorrecta especificación del modelo, implicará que las perturbaciones tengan un sesgo sistemático. Ello se verá reflejado en los residuos.

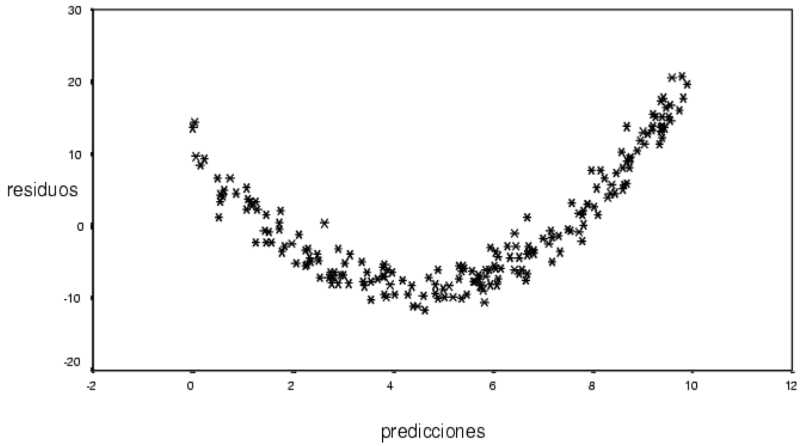
Si la varianza de las perturbaciones depende de las x_i, ello también se evidenciará en los residuos

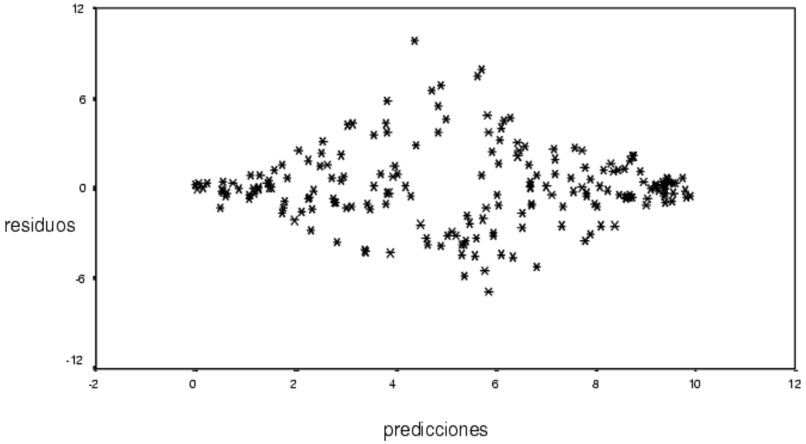
Mariano Bonoli Escobar
Javier Gil

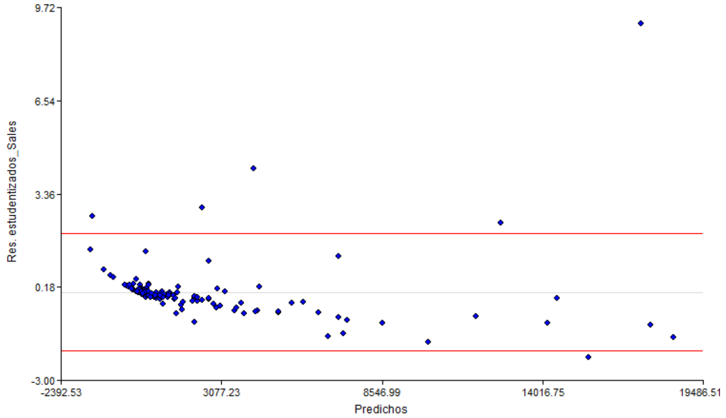
Análisis de especificación

Análisis de residuos









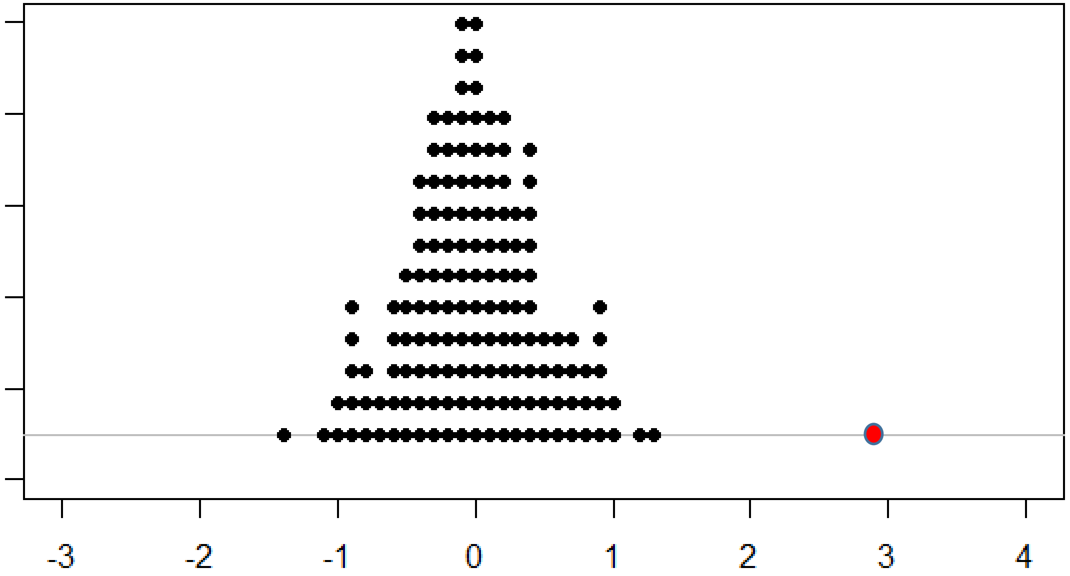
Mariano Bonoli Escobar
Javier Gil

Análisis de Outliers

FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Mariano Bonoli Escobar
Javier Gil

Outliers
Outlier Unidimensional

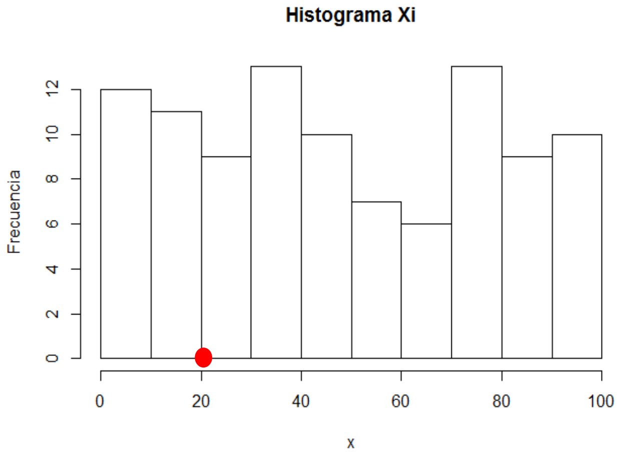


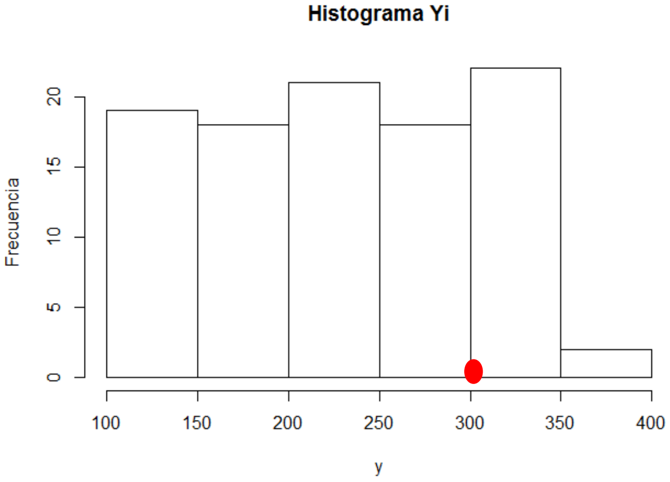
Un Outlier es un valor de una variable X o Y que es incondicionalmente inusual

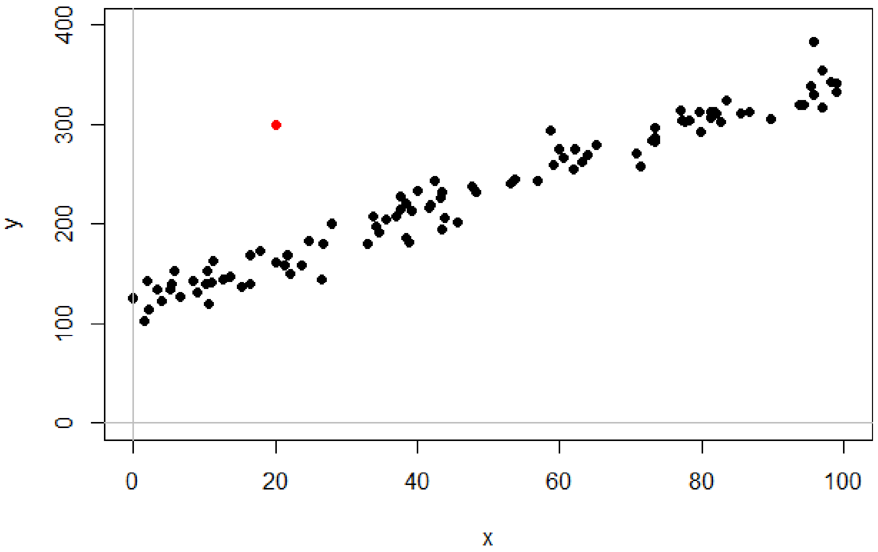
Mariano Bonoli Escobar
Javier Gil

Outliers
Outlier en el contexto de la Regresión

- Un Outlier en regresión es un valor de la variable Y que es condicionalmente inusual dados los valores de X
- Un valor inusual en regresión no necesariamente es un outlier unidimensional y viceversa







Mariano Bonoli Escobar
Javier Gil

Outliers
Definicion

Un Valor Atípico ("Outlier") debe ser investigado cuidadosamente:

- Solo puede excluirse de los datos si se tiene certeza que se trata de un error de medición o de ingreso de la información.
- Pueden dar información muy importante sobre el comportamiento del modelo.
- Pueden afectar seriamente las estimaciones, hacer que los coeficientes de regresión tengan signos contrarios a los esperados o producir pruebas no significativas para un coeficiente de regresión

Mariano Bonoli Escobar
Javier Gil

Análisis de Residuos

Residuos

$$e_i = y_i - \hat{y}_i$$

El inconveniente que presentan es que su escala depende de la escala de y. Esto dificulta determinar si son elevados o no.

Mariano Bonoli Escobar
Javier Gil

Análisis de Residuos

Residuos estandarizados

$$d_i = \frac{y_i - \hat{y}_i}{S_e} = \frac{e_i}{S_e} = \frac{e_i}{\sqrt{CM_{\mathrm{Re}\,s}}}$$

Sin embargo, a pesar de que la varianza de los ε_i es contante

$$\mathrm{D}^2(\varepsilon_i) = \sigma_e^2 = constante$$

No ocurre lo mismo con la varianza de los residuos muestrales

$$\mathrm{D}^2(e_i) \neq constante$$

Mariano Bonoli Escobar
Javier Gil

Leverage o Balanceo

$$\mathrm{V}(e_i) = \sigma^2(1-h_{ii}) \quad \rightarrow \quad \widehat{\mathrm{V}}(e_i) = s^2(1-h_{ii})$$

En Regresión lineal Simple: $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}$ $\qquad \frac{1}{n} \leq h_{ii} \leq 1$

- Llamaremos Leverage al h_{ii} (Balanceo en Montgomery - Peck)
- El leverage es la razón por la cual los Intervalos de Predicción y Confianza son mas amplios al alejarnos del baricentro ($\bar{x}$)

$$\mathrm{E}(\tilde{y}|x_o) \sim Normal\left(\beta_o + \beta_1 x_o \; ; \; \sigma\sqrt{\frac{1}{n} + \boxed{\frac{(x_o-\bar{x})^2}{S_{xx}}}}\right) \qquad \tilde{y}|x_o \sim Normal\left(\beta_o + \beta_1 x_o \; ; \; \sigma\sqrt{1 + \frac{1}{n} + \boxed{\frac{(x_o-\bar{x})^2}{S_{xx}}}}\right)$$

Mariano Bonoli Escobar
Javier Gil

Análisis de Residuos

Residuos estudentizados

$$r_i = \frac{e_i}{\mathrm{D}(e_i)} = \frac{y_i - \hat{y}_i}{S_e\sqrt{1 - h_{ii}}} \qquad\qquad \mathrm{Var}(r_i) \quad = 1$$

- Como $\frac{1}{n} \leq h_{ii} \leq 1,$ los residuos estandarizados están sub-estimados
- Dependen de la ubicación del vector x_i
- Puntos más alejados del centroide $(\bar{x})$ de los datos tendrán valores mayores de residuos estudentizados que si se calculan los estandarizados u ordinarios.

Mariano Bonoli Escobar
Javier Gil

Análisis de Residuos

Residuo PRESS

$$e_{-i} = y_i - \hat{y}_{-i}$$

Fundamento: si hay alguna observación atípica posiblemente influya mucho en la estimación del hiperplano de regresión. Lo más conservador sería calcular el residuo excluyendo esa observación.

Mariano Bonoli Escobar
Javier Gil

Análisis de Residuos

Residuo PRESS

Residuo PRESS internamente estudentizados

$$\frac{y_i - \hat{y}_{-i}}{S\sqrt{1 - h_{ii}}}$$

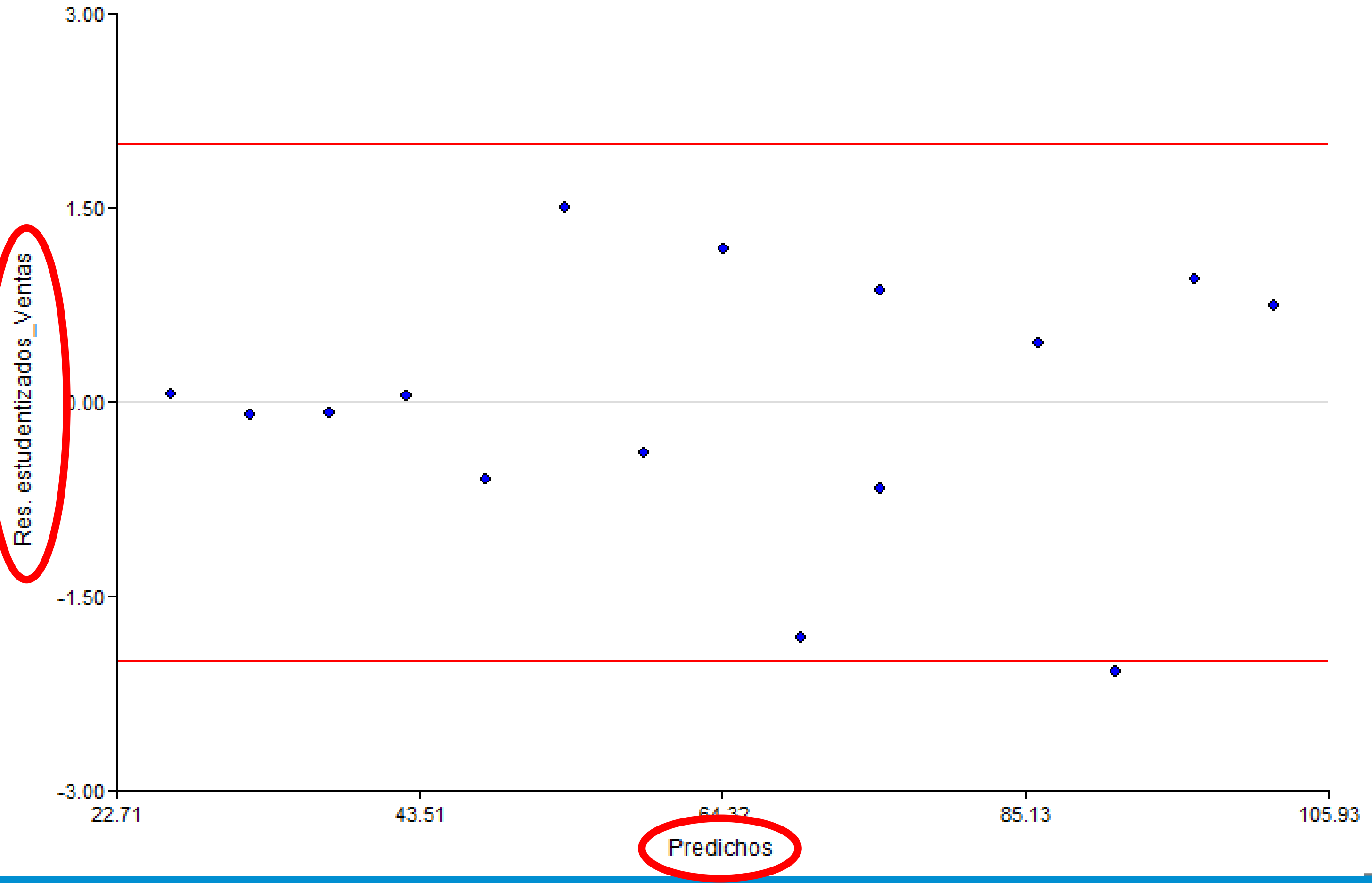
Residuo PRESS externamente estudentizados o R-Student

$$t_i = \frac{y_i - \hat{y}_{-i}}{S_{-i}\sqrt{1 - h_{ii}}} \approx t_{n-p-1}$$

$$S_{-i}^2 = \frac{1}{\nu - 1}\left(\nu s^2 - \frac{e_i^2}{1 - h_{ii}}\right)$$

La varianza residual S_{-i}^2 se estima sin tener en consideración el punto. Si la observación es atípica puede influir notablemente en la estimación de la varianza.
Si se cumplen los supuestos de la regresión tienen distribución $t\ Student$ con $\nu = n - p - 1$

Análisis de Residuos

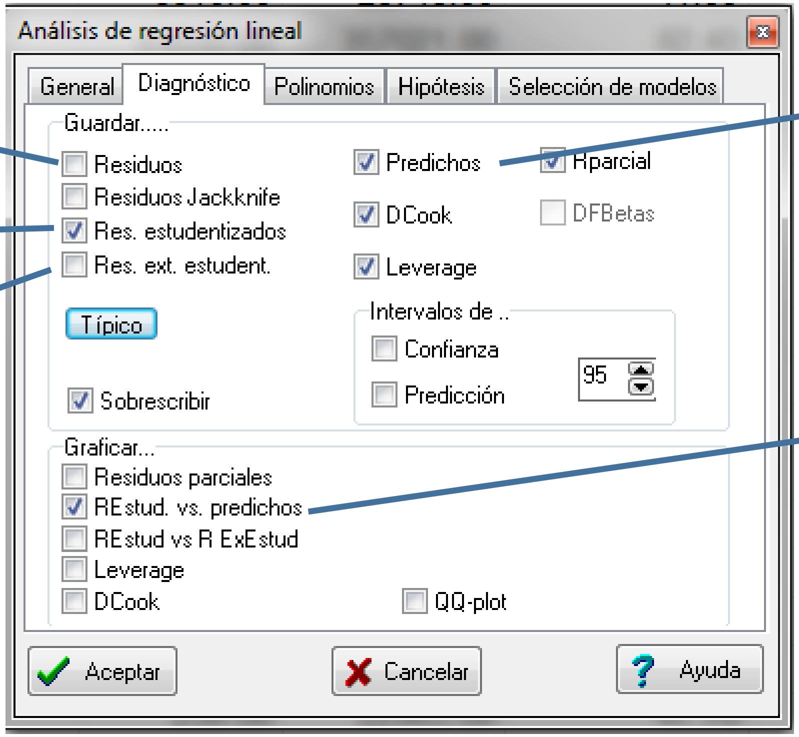


Mariano Bonoli Escobar
Javier Gil

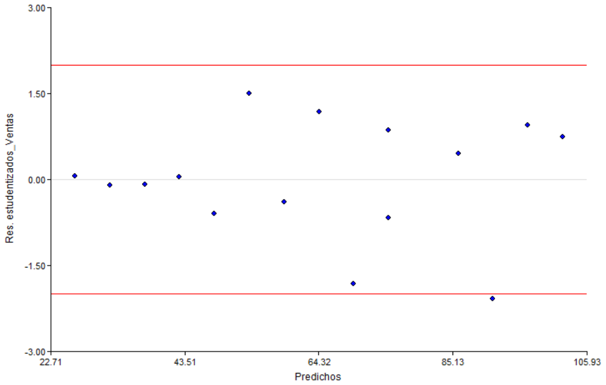
$$e_i = y_i - \hat{y}_i$$

$$\frac{y_i - \hat{y}_{-i}}{S\sqrt{1 - h_{ii}}}$$

$$t_i = \frac{y_i - \hat{y}_{-i}}{S_{-i}\sqrt{1 - h_{ii}}}$$



$$\hat{y}_i$$



Mariano Bonoli Escobar
Javier Gil

Análisis de Influencia

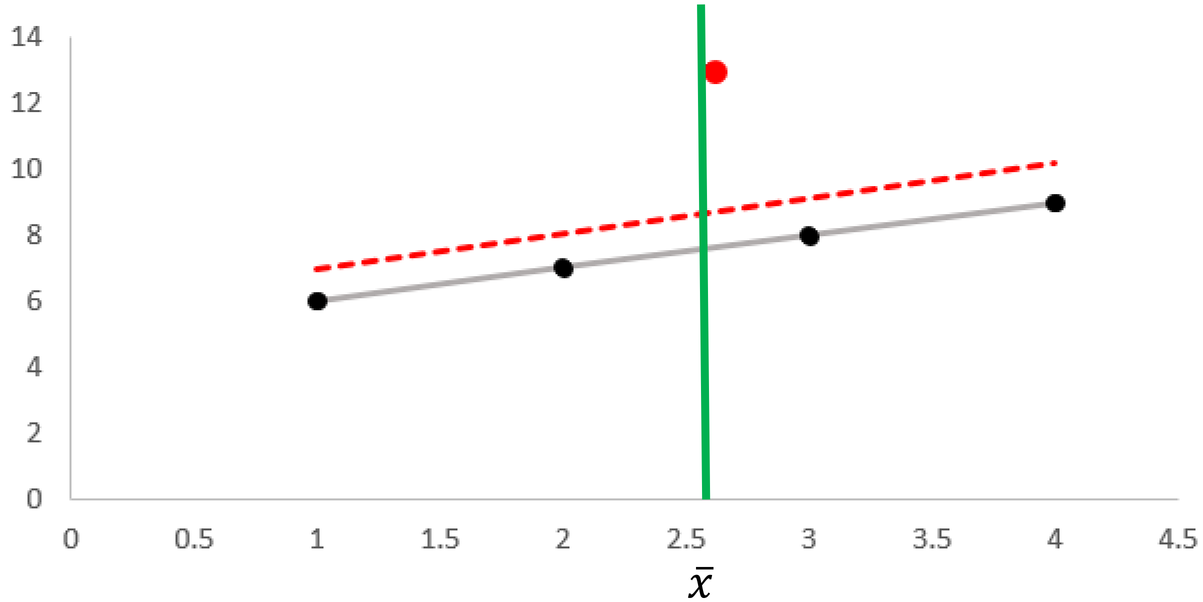
Balanceo e influencia

FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

Mariano Bonoli Escobar
Javier Gil

Leverage

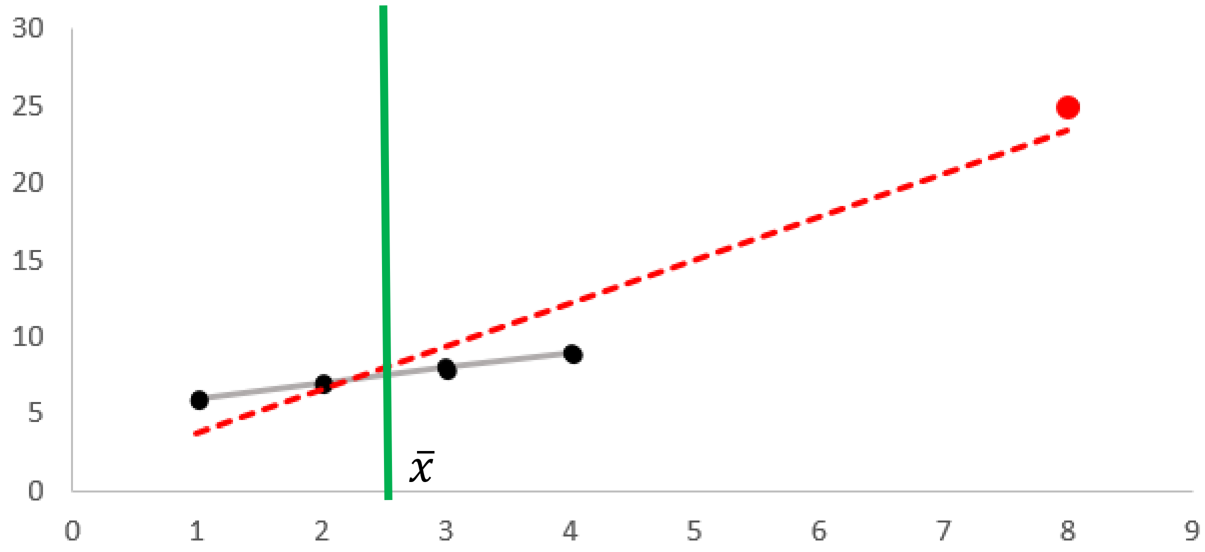
Concepto



- Punto de bajo Leverage (balanceo) → poca influencia potencial sobre los parámetros de la recta

Mariano Bonoli Escobar
Javier Gil

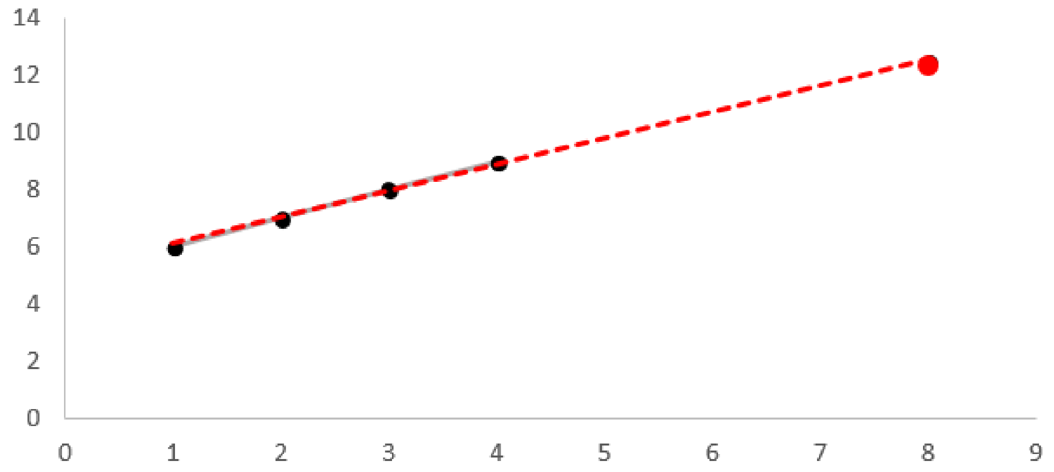
Leverage
Concepto



- Punto de alto balanceo → alta influencia potencial sobre los parámetros de la recta

Leverage
Concepto

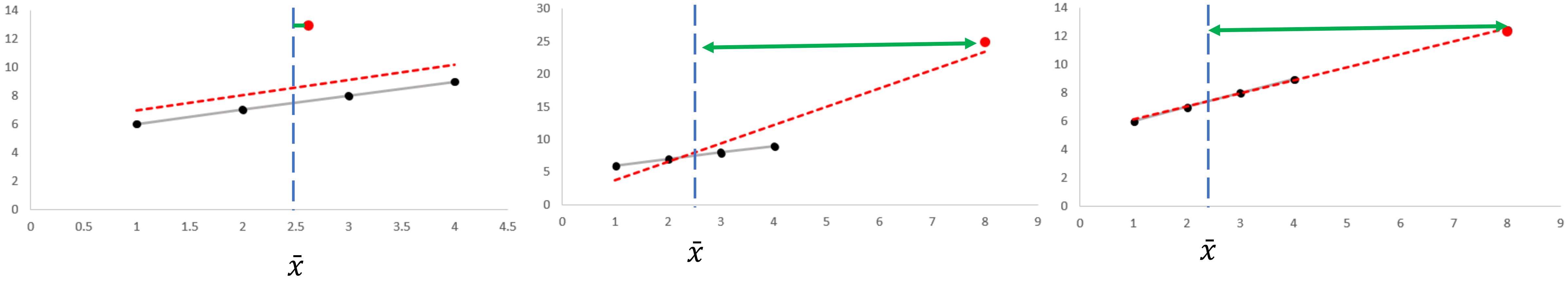
Influencia = Leverage x Discrepancia



- Alto Leverage
- Baja discrepancia

Mariano Bonoli Escobar
Javier Gil

Leverage
Concepto



El Leverage solamente depende de x_i y está asociado a la lejanía de la observación del baricentro ($\bar{x}$ en regresión lineal simple)

Criterio: $\sum_{i=1}^{n} h_{i_i} = p \quad \Rightarrow \quad \bar{h} = p/n$

Se considera de alto Leverage si supera a $h_{i_i} \geq 2p/n$

$h_{i_i} \geq 3p/n$ Muestras grandes

Mariano Bonoli Escobar
Javier Gil

Diagnósticos

Resumen

Leverage, Balanceo o Apalancamiento

- Se relaciona exclusivamente con la posición en el espacio de la observación x_i
- Métrica : h_{ii}

Outlier en regresión

- Lo clasificamos como Outlier por tener un residuo extremadamente grande

Punto influyente

- Se tiene en cuenta en cuenta conjuntamente la magnitud del residuo y la posición de la observación en el espacio.

Mariano Bonoli Escobar
Javier Gil

Diagnósticos

Resumen

- Los outliers no son necesariamente puntos influyentes
- Observaciones con alto Leverage no son necesariamente puntos influyentes
- Para determinar si un punto es influyente debe considerarse el efecto conjunto de su leverage y la magnitud de su residuo

Influencia = Leverage x Discrepancia

Mariano Bonoli Escobar
Javier Gil

Puntos Influyentes
Distancia de Cook

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(-i)}\right)^2}{p\ S^2} = \frac{r_i^2}{k+1}\ \frac{h_{ii}}{1-h_{ii}}$$

- Se calcula una distancia para cada observación. Mide la influencia en los valores predichos si se elimina la observación analizada.
- Es un indicador global de influencia que tiene en cuenta el tamaño del residuo y el Leverage.
- No hay reglas exactas, pero se deben analizarse aquellas observaciones con Distancia de Cook mayores a la unidad

Mariano Bonoli Escobar
Javier Gil

Puntos Influyentes
DFBETAS

$$DFBETAS_{i,j} = \frac{b_j - b_{j(-i)}}{S_{b_{j(-i)}}}$$

- Se calcula un DFBETAS para cada observación y parámetro estimado. Mide la influencia en los coeficiente estimado si se elimina la observación analizada.
- Deben analizarse las observaciones donde $|DFBETAS_{i,j}| > 2/\sqrt{n}$

Mariano Bonoli Escobar
Javier Gil

Puntos Influyentes
DFFITS

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(-i)}}{S_{(-i)}\sqrt{h_{i_i}}}$$

- Se calcula un DFFITS para cada observación y parámetro estimado. Mide la influencia en los valores predichos si se elimina la observación analizada.
- Deben analizarse las observaciones donde $|DFFITS_i| > 2\sqrt{p/n}$

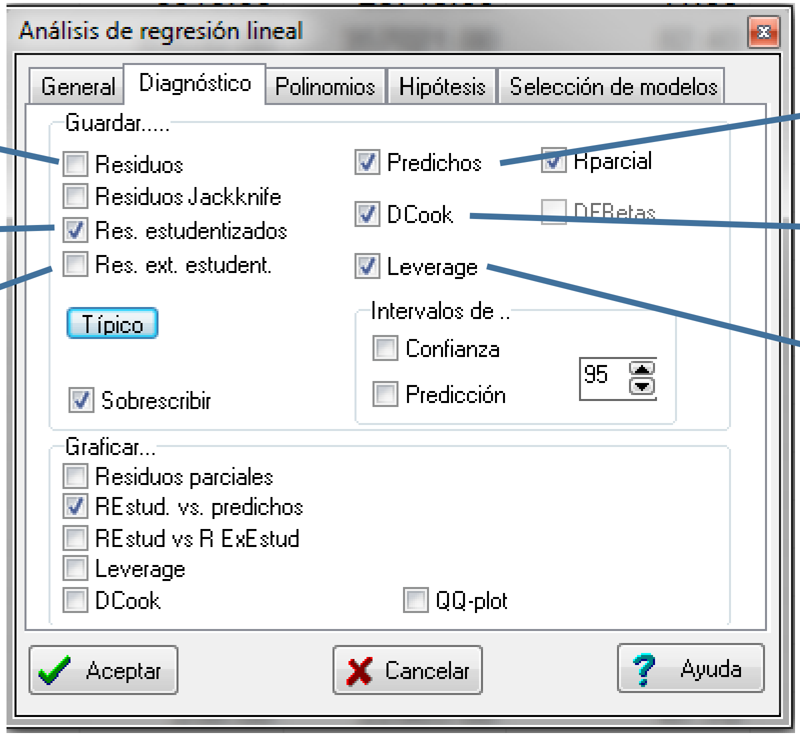
Mariano Bonoli Escobar
Javier Gil

INFOSTAT

$$e_i = y_i - \hat{y}_i$$

$$\frac{y_i - \hat{y}_{-i}}{S\sqrt{1 - h_{ii}}}$$

$$t_i = \frac{y_i - \hat{y}_{-i}}{S_{-i}\sqrt{1 - h_{ii}}}$$



$$\hat{y}_i$$

$$D_i = \frac{\sum_{j=1}^{n}(y_i - \hat{y}_{-i})^2}{p\ S^2}$$

$$h_{i_i}$$

Mariano Bonoli Escobar
Javier Gil