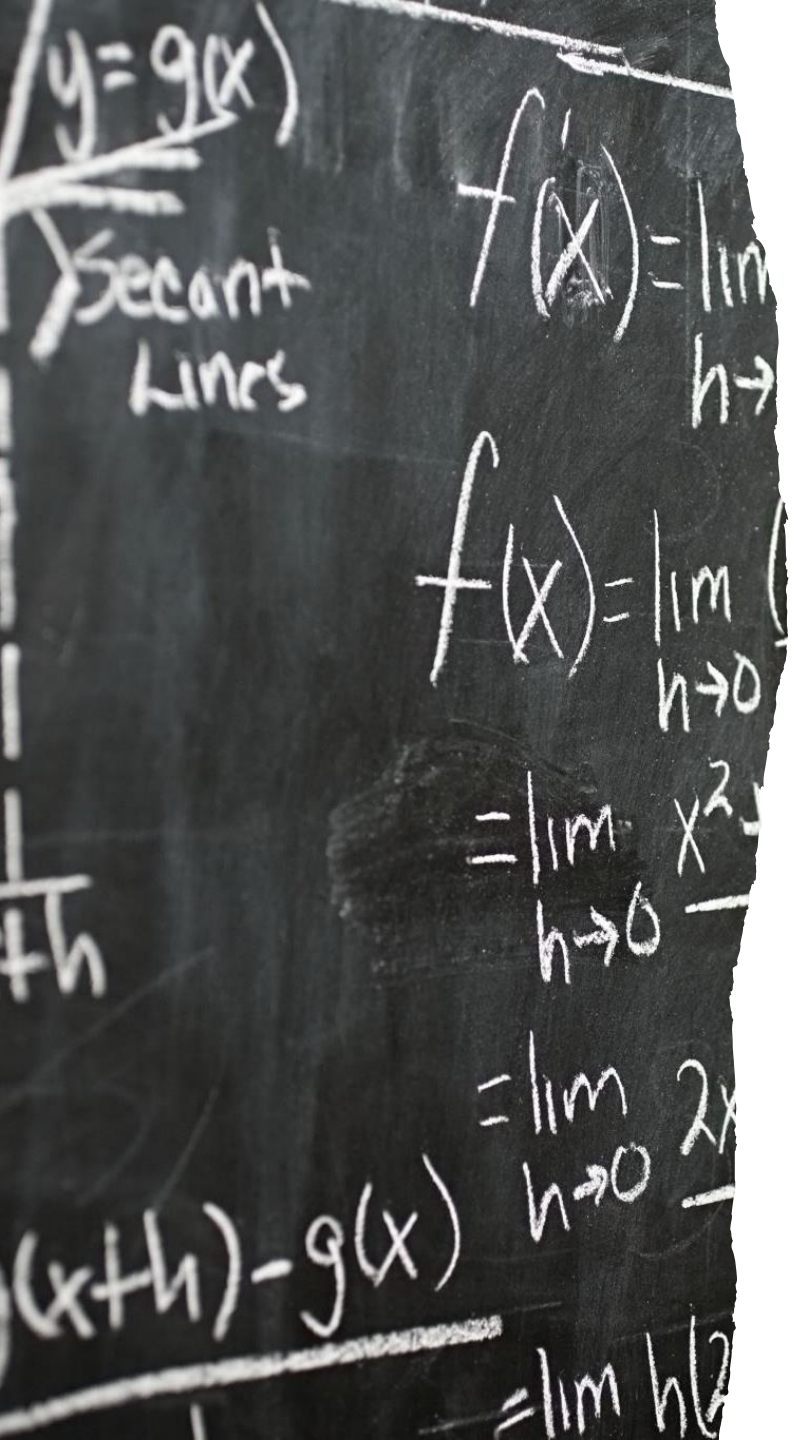
Análisis de Regresión lineal
Análisis de residuos



Ecuación de Regresión Lineal Múltiple

- *Dadas $x_1 \ldots x_k$, k variables independientes, el modelo de regresión lineal múltiple es*

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

- *Para* i=1, 2, …, n
- *Supuestos del modelo*
  - *La relación es lineal*
  - *Varianza constante*
  - *$\varepsilon_i \sim N(0, \sigma)$ para todo i*
  - *Muestra independiente*

Supuestos de la regresión lineal

- La relación entre x e y es lineal
- Los errores tienen distribución normal con media cero y desvío σ constante
- No hay valores atípicos ni en x ni en y

Una vez ajustado el modelo es posible chequear los supuestos analizando los residuos generados

Chequeo de los supuestos

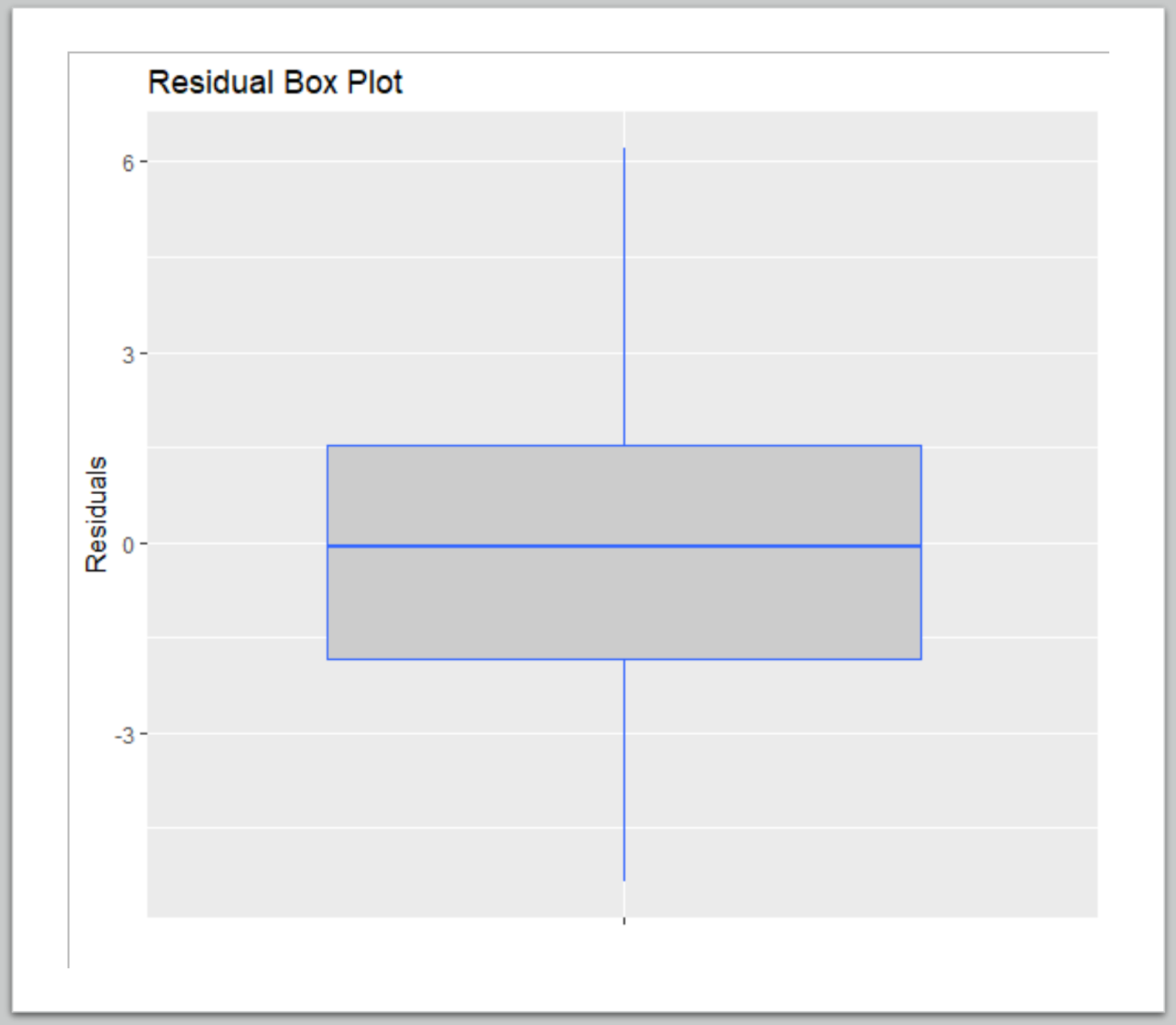
- Para chequear los supuestos del modelo se utilizan distintos gráficos de los residuos generados
- 1. gráficos de la distribución de los residuos
  - Boxplot
  - Histograma
- 2. Gráfico para ver normalidad de los residuos
  - Qqplot de los residuos
- 3. gráfico para evaluar normalidad de los residuos y la presencia de heterocedasticidad o outliers
  - Gráfico de residuos contra el valor predicho

Boxplots de los residuos

ols_plot_resid_box(model)

- No debe haber outliers
- La mediana debe pasar por el medio de la caja
- Los bigotes deben tener una longitud similar

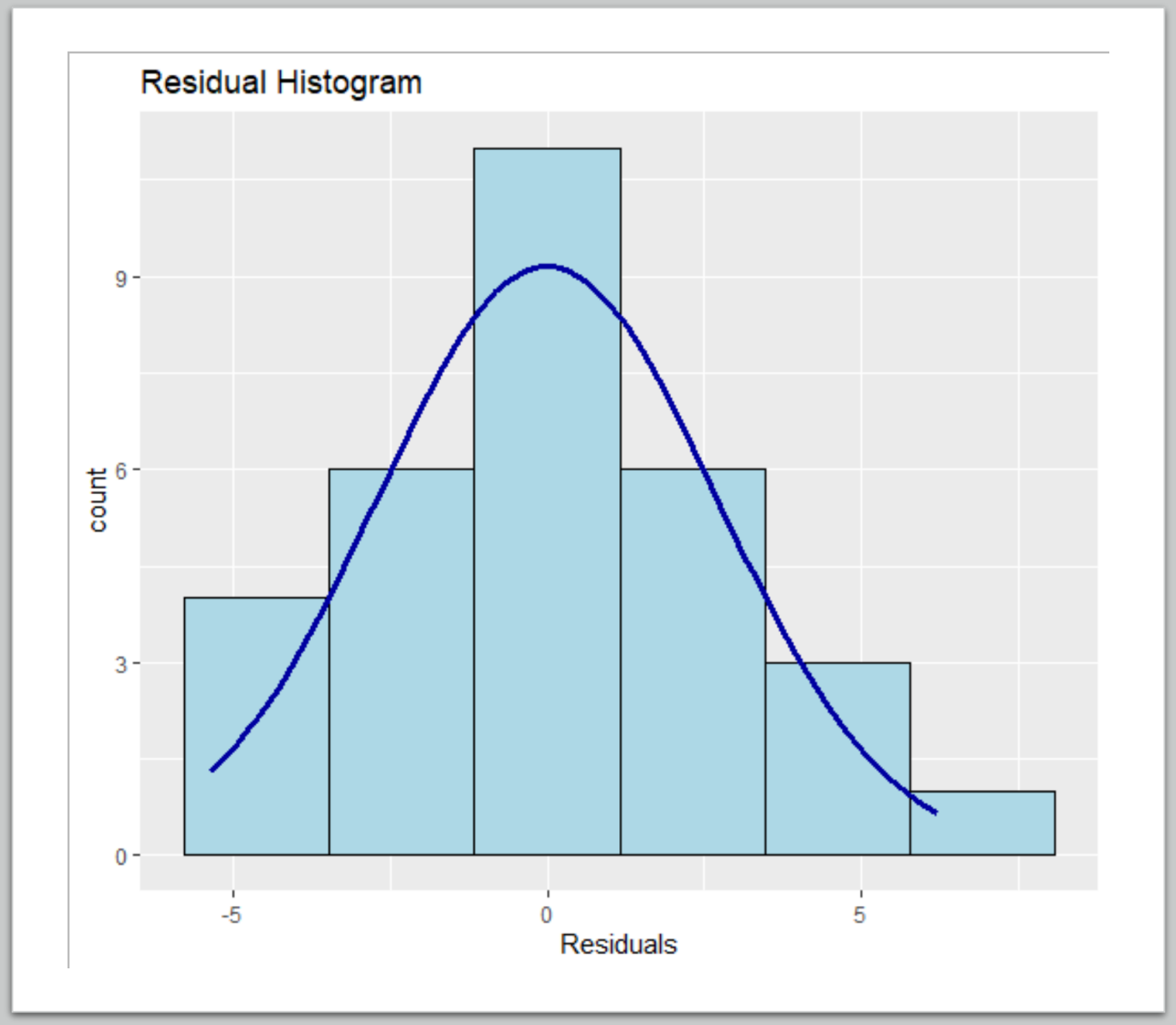
- Paquete olsrr()



Histograma de los residuos

ols_plot_resid_hist(model)

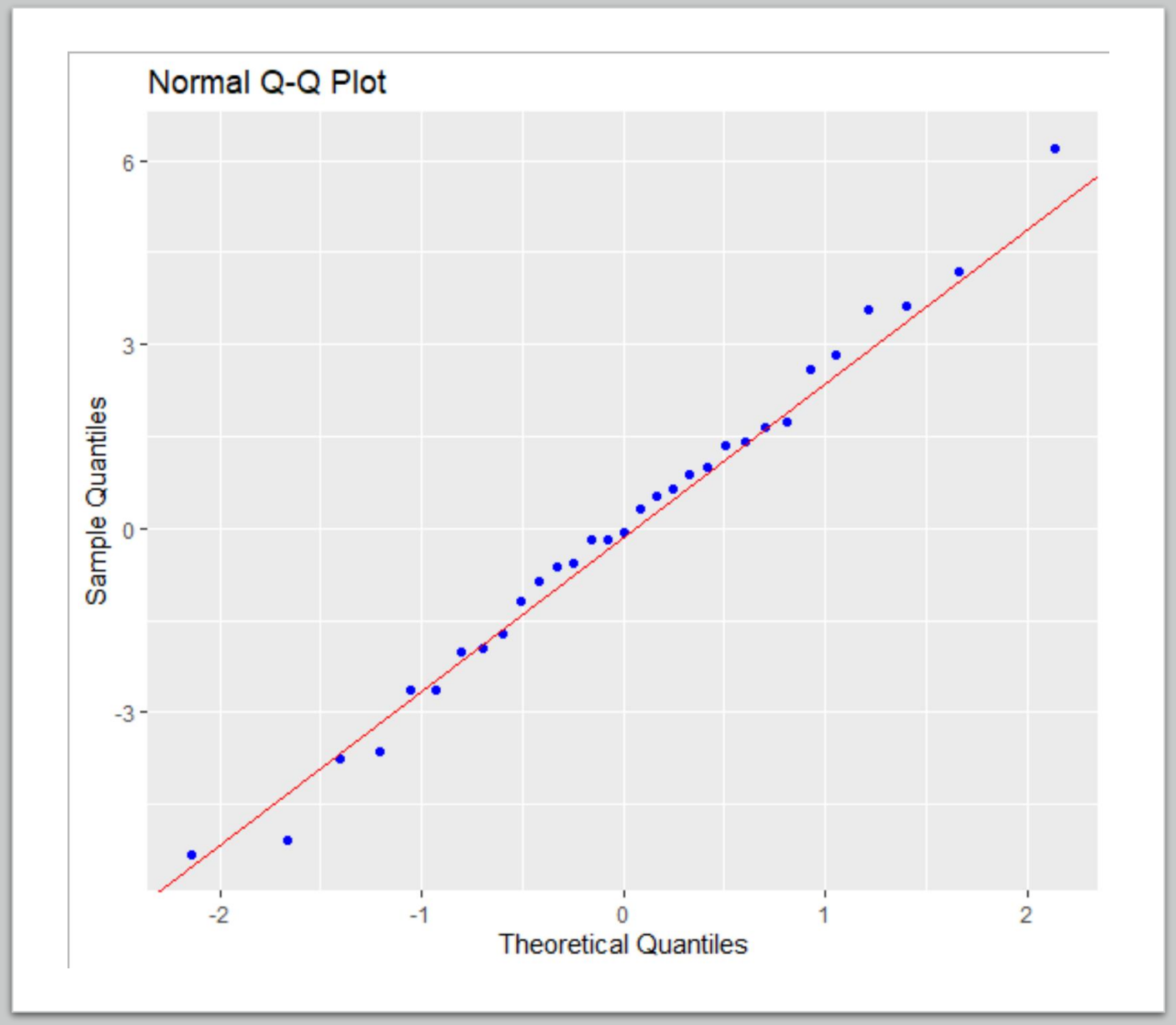
- La curva superpuesta simula una normal.
- No debe haber un apartamiento importante respect de la curvar



Qqplot de los residuos

ols_plot_resid_qq(model)

- Debe haber alineamiento de los puntos en torno a la recta para que haya normalidad



Ecuación de Regresión Lineal Múltiple

- *La recta de regresión lineal múltiple estimada o ajustada es*

$$y_i = b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} + e_i$$

- *Para i=1,2,..,n*
- $e_i = y_i - \hat{y}_i$ *es el residuo y describe el error en el ajuste del modelo en cada punto.*

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

Una manera de ver si se cumplen los supuestos es analizar los residuos

Forma matricial del modelo de Regresión

$$y = X\beta + \varepsilon$$

Donde,

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix}$$

Estimador de mínimos cuadrados

$$b = (X'X)^{-1}X'y$$

Forma matricial del modelo de Regresión

Predictor de y

$$\hat{y} = X(X'X)^{-1}X'y = Hy$$

H se llama matriz sombrero y sus elementos diagonales son h_{ii}

$$0 < h_{ii} < 1 \qquad \sum h_{ii} = k + 1$$

Residuos

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

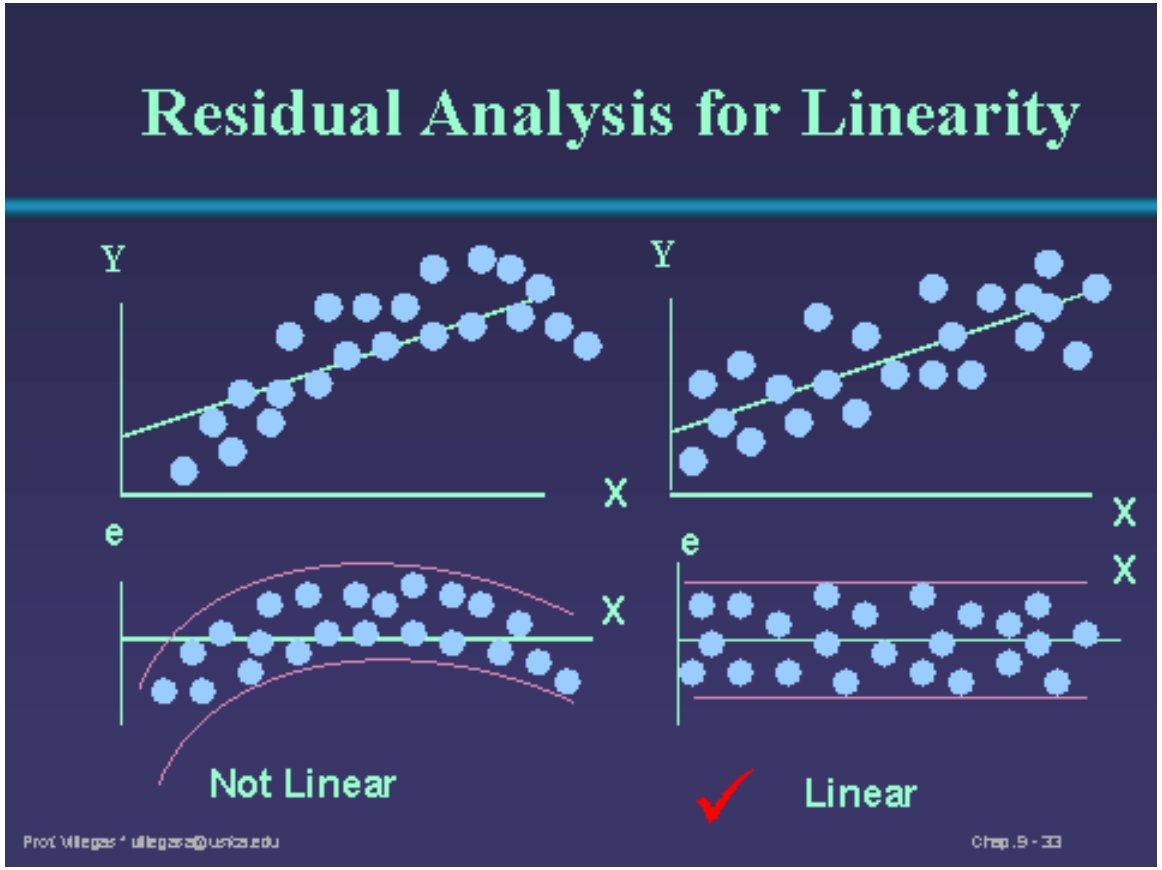
Residuo studentizado $r_i = \frac{e_i}{S\sqrt{1 - h_{ii}}}$

Residuo R Student $t_i = \frac{e_i}{S_{-i}\sqrt{1 - h_{ii}}}$

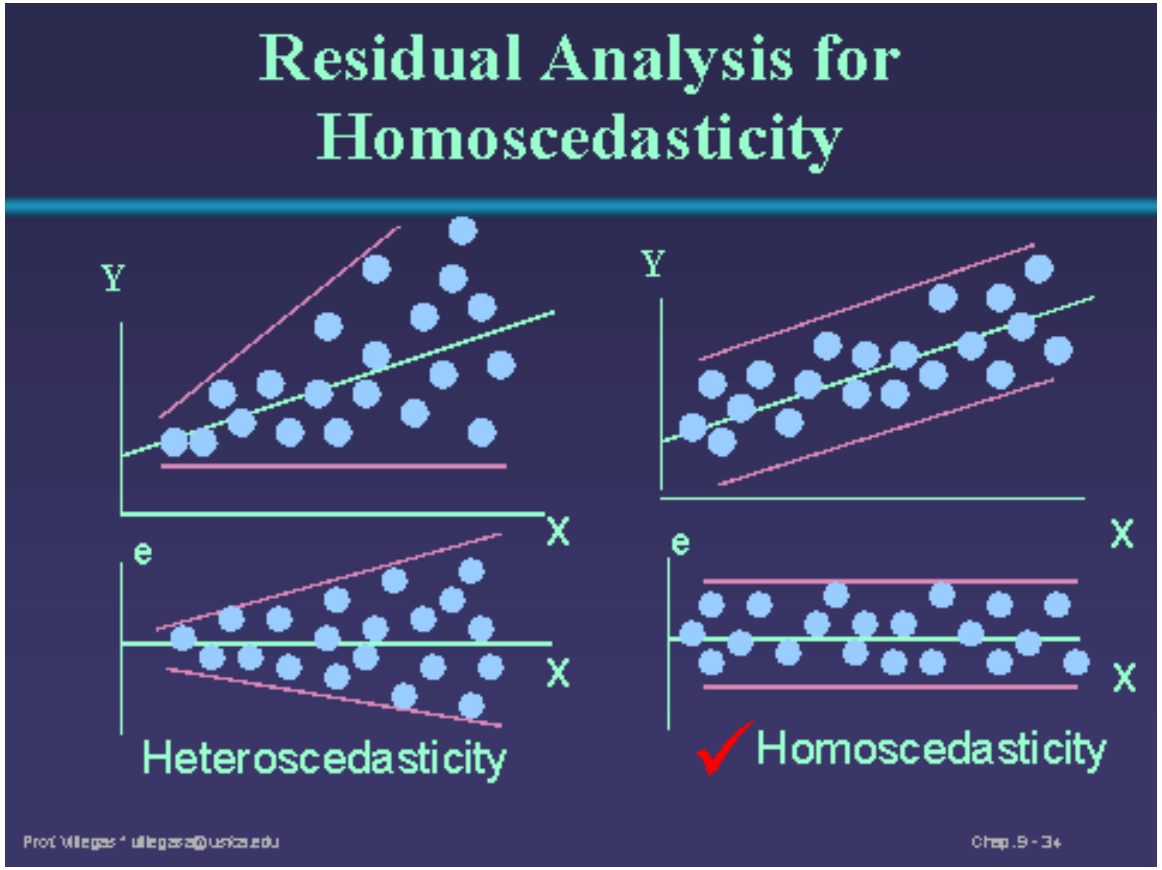
Gráficos de residuos contra el predictor de y

- Se cumplen los supuestos cuando
- No hay outliers
- No hay patrones en los residuos
- No hay variaciones en la variabilidad del gráfico

Gráficos de Residuos



Gráficos de Residuos



Gráficos de Residuos

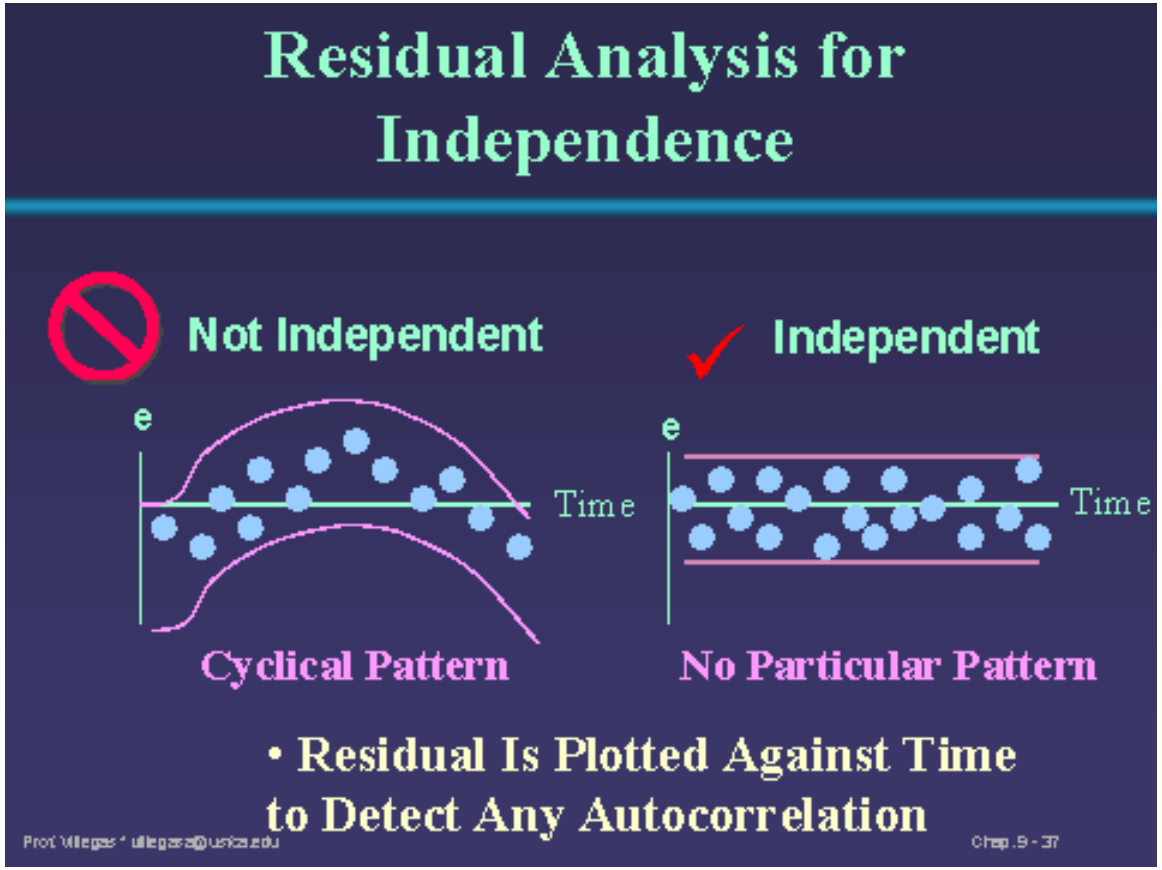
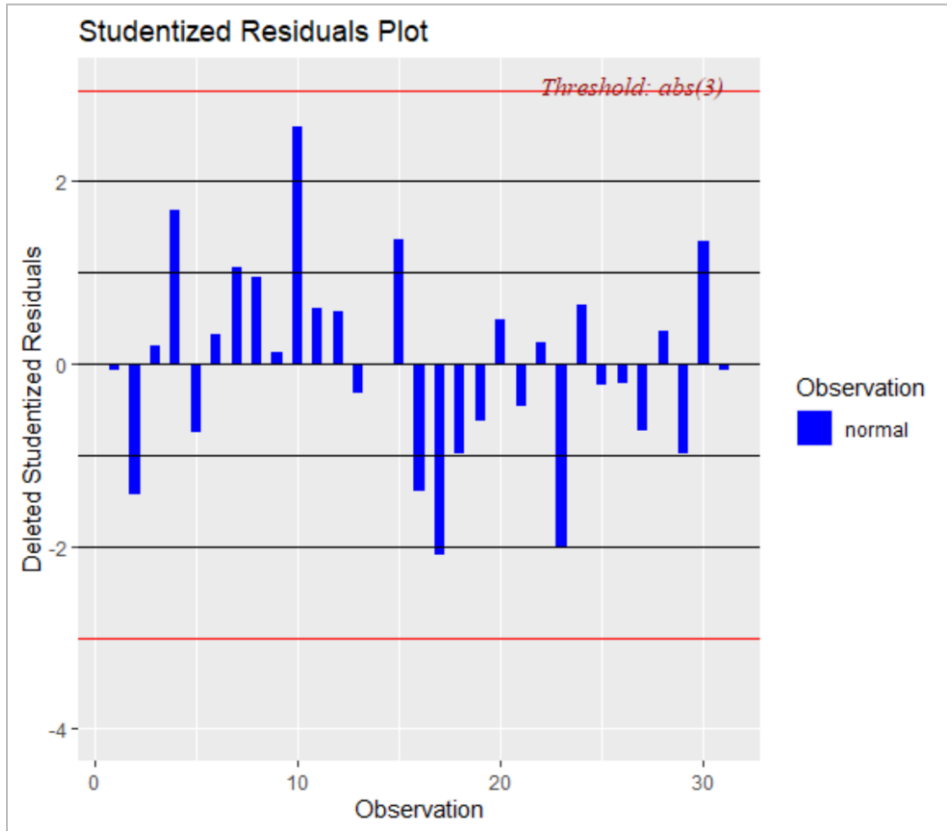


Gráfico de residuos contra predictor

ols_plot_resid_stud(model)

- Los puntos tienen que estar entre -2 y 2 o entre -3 y 3 para que se cumpla la normalidad
- No debe apreciarse ninguna forma específica
- No debe haber crecimiento o decrecimiento en el gráfico



Valores atípicos (outliers) en una variable

Un outlier es un valor de la variable muy poco probable de ser obtenido

En ocasiones puede deberse a un error en la toma de información

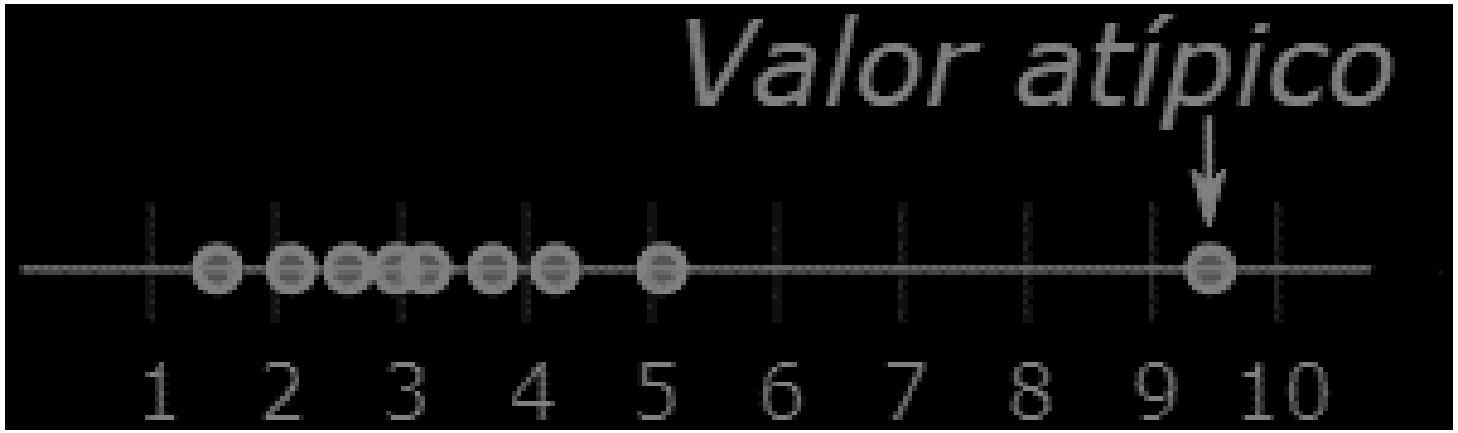
En otros indica una situación inesperada o muy poco frecuente

La estadística clásica está basada en métodos que son muy sensibles a la presencia de valores atípicos

Por eso es muy importante detectarlos

Valores atípicos (outliers) en una variable

- La media se ve muy afectada por los valores atípicos
- La varianza y el desvío estándar se ven muy afectados por los valores atípicos



Valores atípicos (outliers) en la regresión

- Como el método de regresión es un procedimiento de maximización, es muy sensible a la presencia de valores atípicos
- Pocos valores atípicos pueden producir severos trastornos en los modelos estimados

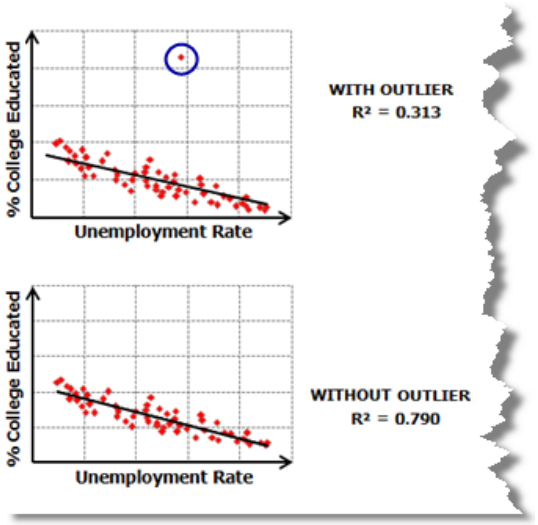
- Los valores atípicos pueden estar
  - En la variable dependiente (outliers en los residuos)
  - En las variables independientes (puntos de influencia)

Valores atípicos y puntos de influencia

- Existen dos formas de tratar el problema de los puntos de influencia y los valores atípicos
  - Detección y eliminación
  - Regresión robusta

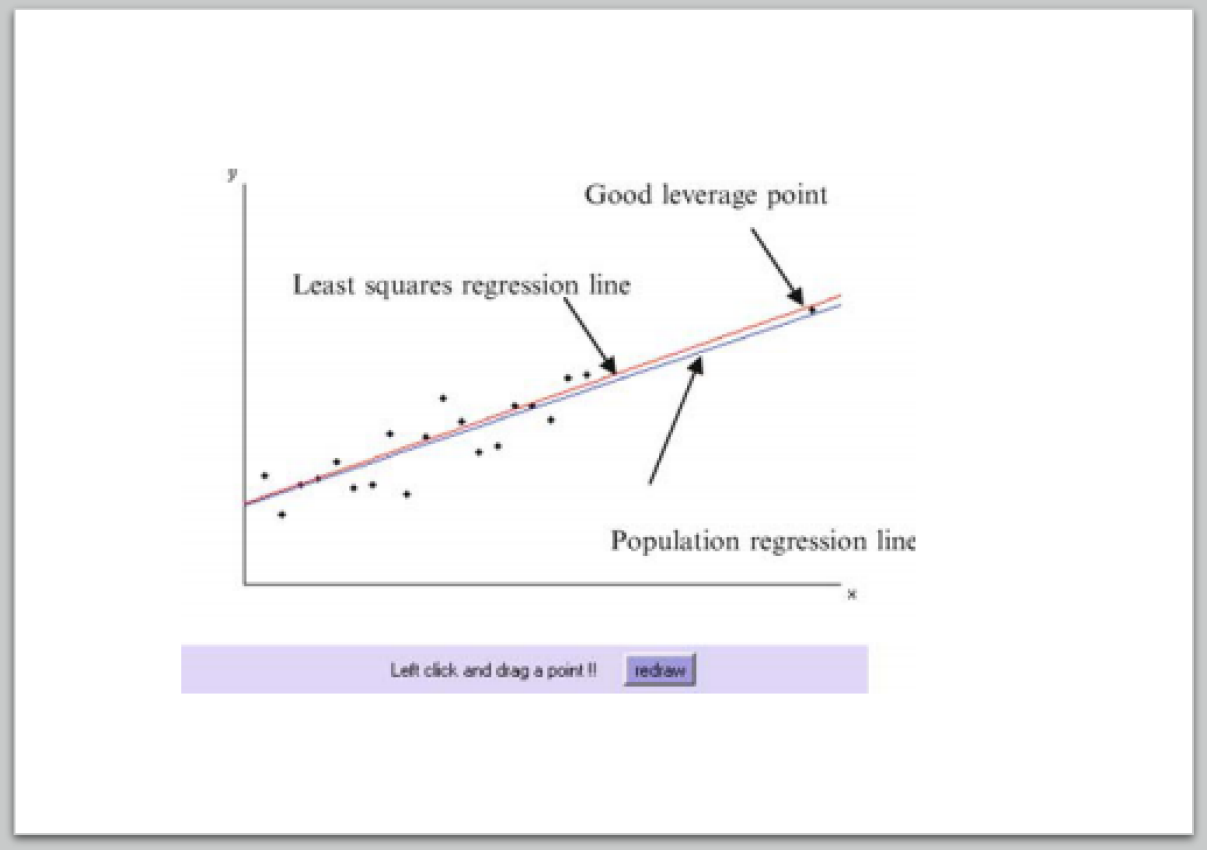
Valores atípicos en la variable dependiente

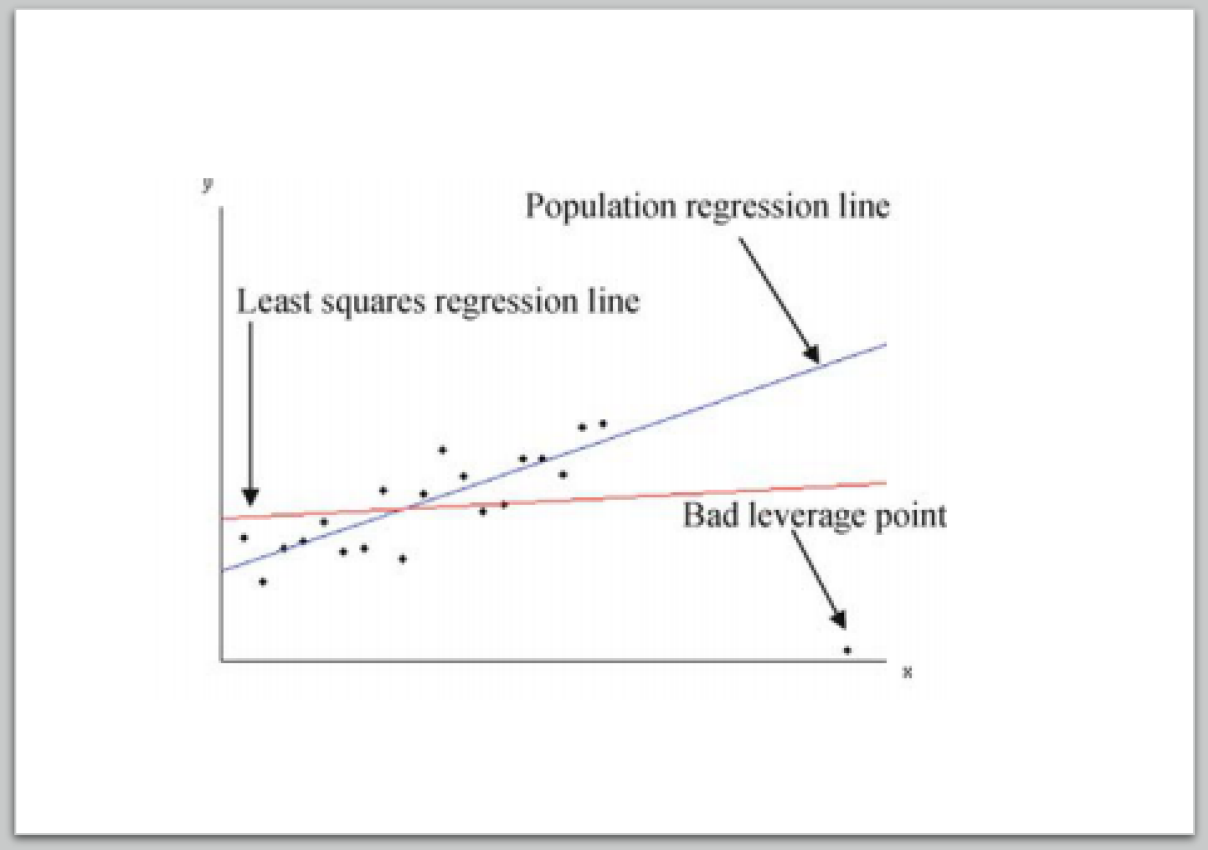
- Un valor atípico en la variable dependiente será mal ajustado por el modelo y tendrá un residuo importante
- Puede También afectar elajuste general del modelo



Puntos de influencia (Leverage points)

- Son aquellos datos que ejercen una considerable influencia en el ajuste del modelo
- En general, se asocian a observaciones atípicas en los predictores





Detección de valores atípicos y puntos de influencia

- Detección de outliers en los predictores
- Distancia de Mahalanobis

$$D_i = (x_i - \bar{x})' S^{-1} (x_i - \bar{x})$$

  - Indica cuán lejos se encuentra el caso del centroide.

Detección de valores atípicos y puntos de influencia

- Detección de puntos influyentes
- Distancia de Cook $C_i = \frac{1}{p+1} r_i^2 \frac{h_{ii}}{1-h_{ii}}$

  - Indica la influencia conjunta de un caso, sea outier de *y* o de los predictores
  - Un valor de C_i mayor que 1 es considerado grande o bien 4/n

Detección de valores atípicos y puntos de influencia

- DFFIT se utiliza para identificar datos influyentes. Cuantifica cuánto cambia el valor ajustado cuando la i-ésima observación es omitida.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i-1}}{s_{-i}\bar{h}_{ii}}$$

- Pasos para su cálculo
  - Se omite una observación por vez
  - Se ajusta el modelo con las restantes n-1 observaciones
  - Se examinacuánto se modifica el valor ajustado, respect del modelo calculado con todas las observacione.
- Criterio a utilizar

$$|DFFITS_i| > \frac{2\sqrt{p+1}}{\sqrt{n-p-1}}$$

Detección de valores atípicos y puntos de influencia

Detección de puntos influyentes (DFBETA)

- DFBETAS

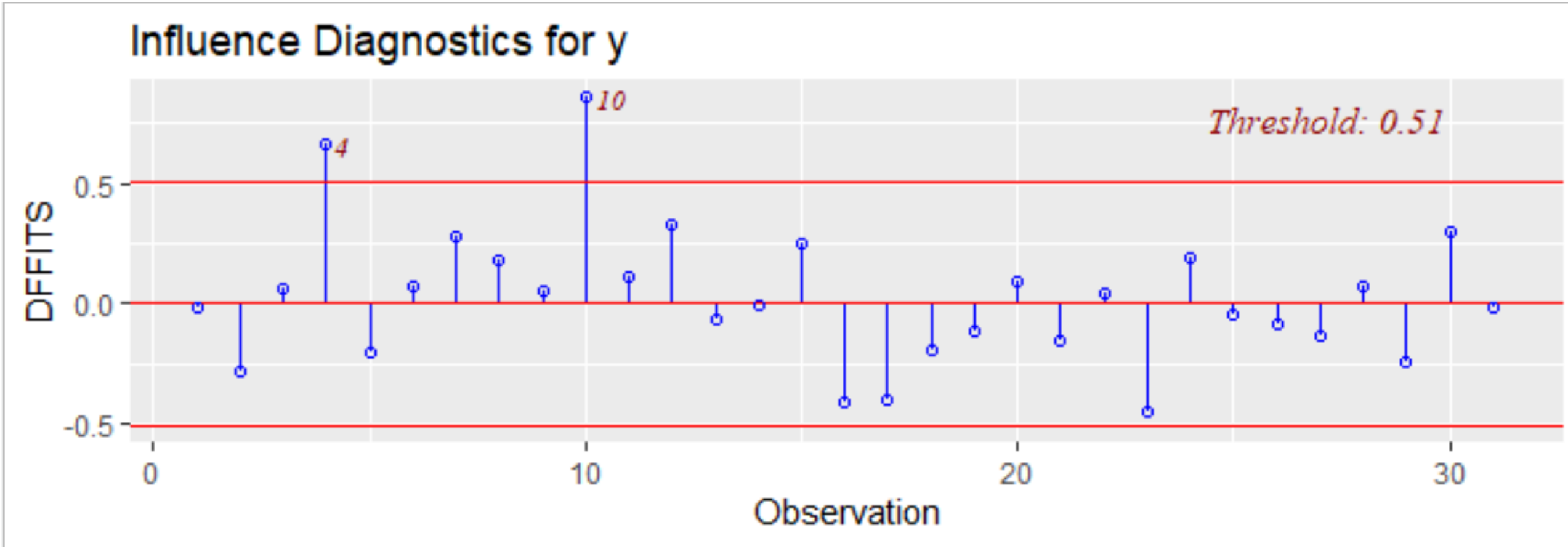
$$DFBETAS_{j,i} = \frac{b_j - b_{j,-i}}{s_{-i}c_{jj}}$$

- Miden la diferencia en la estimación de cada parámetro con y sin la observación.
- Hay tantos DFBETAs como parametros en la ecuación, y se calculan para cada individuo.

- Criterio a utilizar DFBETA > 2 en valor absoluto

- O bien DFBETA > 2 / raíz(n)

Dffits


Influence Diagnostics for y
Threshold: 0.51
10
4
0.5
0.0
-0.5
DFFITS
0
10
20
30
Observation