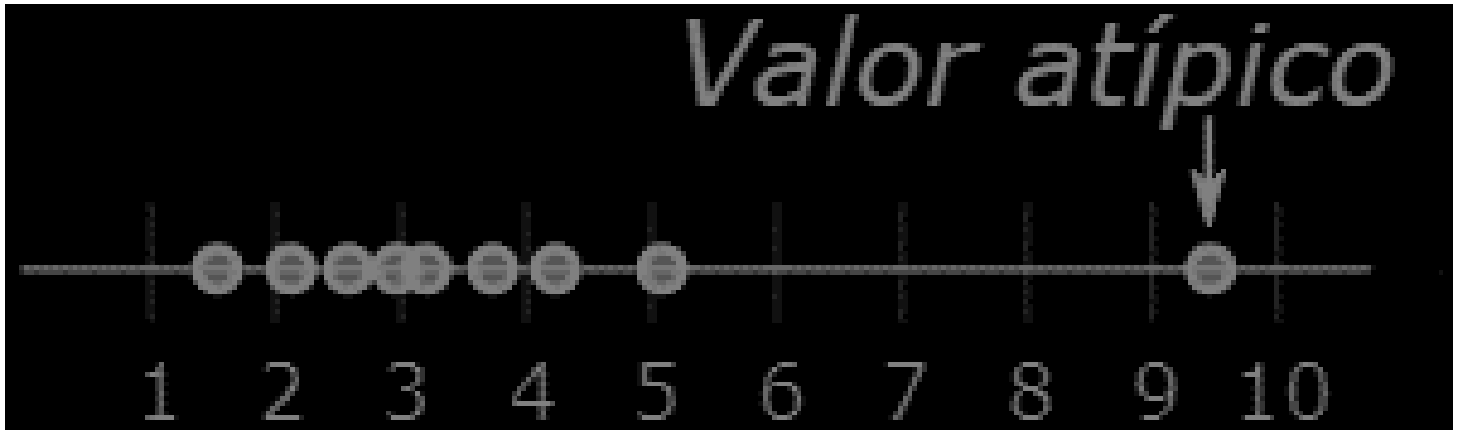
Valores atípicos y puntos de influencia

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

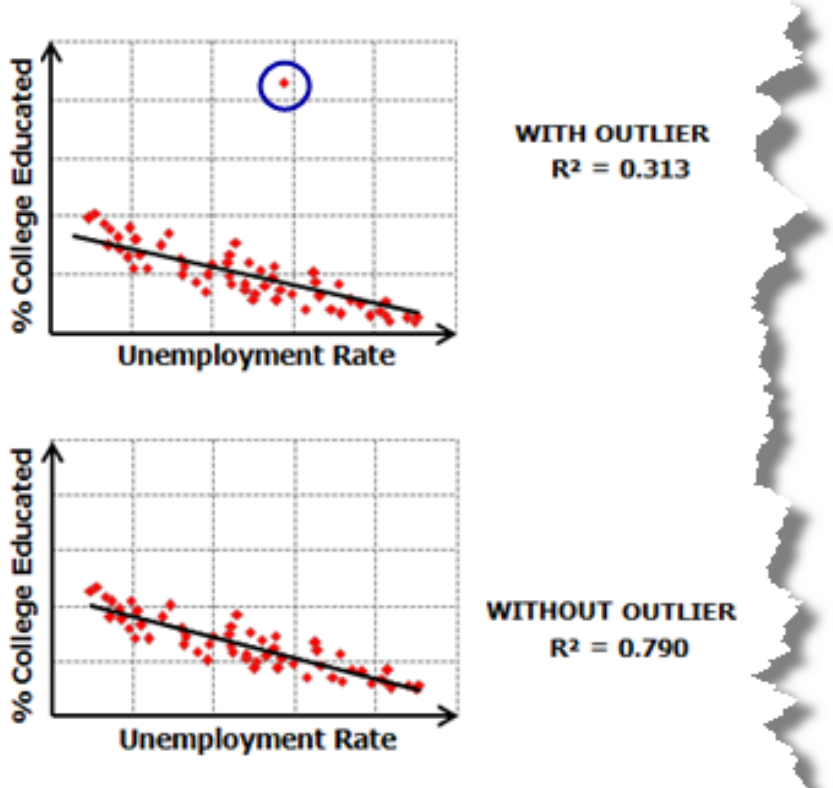
- Los valores atípicos pueden estar
  - En la variable dependiente (outliers en los residuos)
  - En las variables independientes (puntos de influencia)

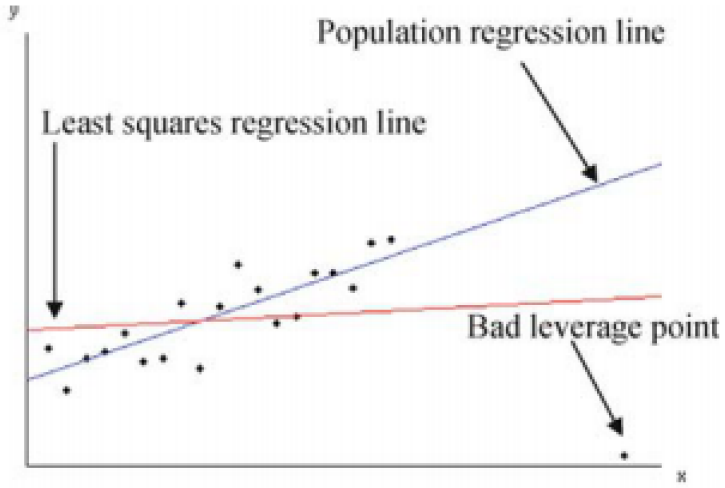
Valores atípicos y puntos de influencia

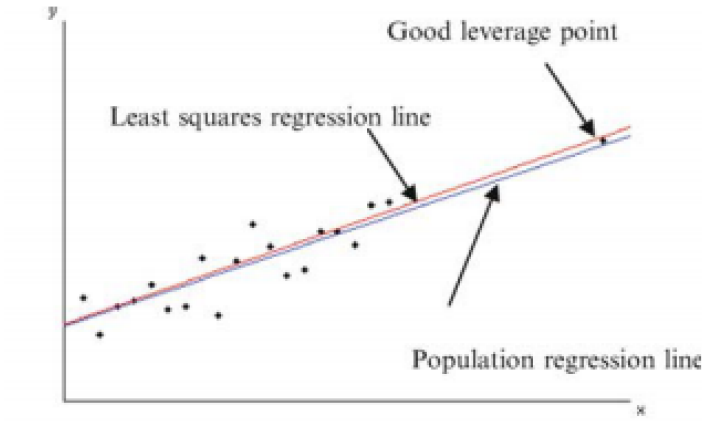
- Existen dos formas de tratar el problema de los puntos de influencia y los valores atípicos
  - Detección y eliminación
  - Regresión robusta

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







Left click and drag a point !! redraw

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

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}

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

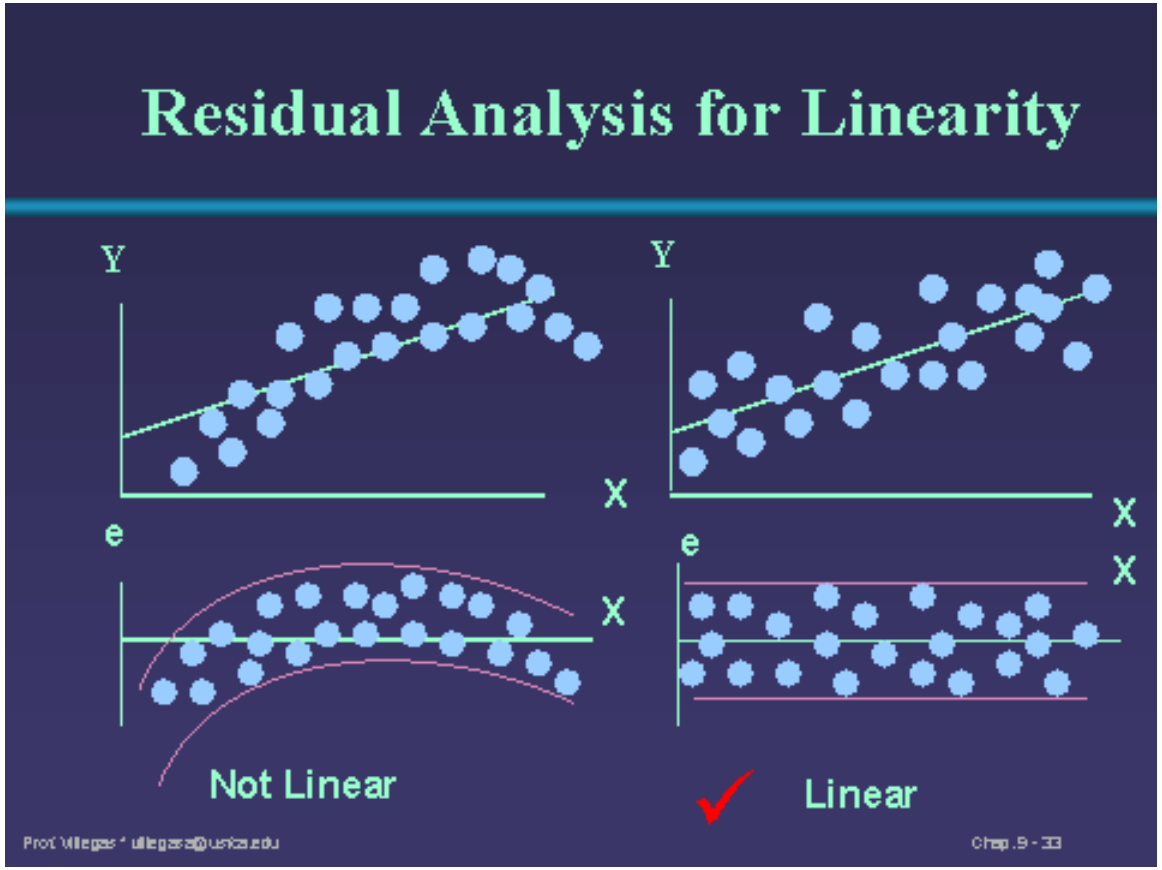
Residuos

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

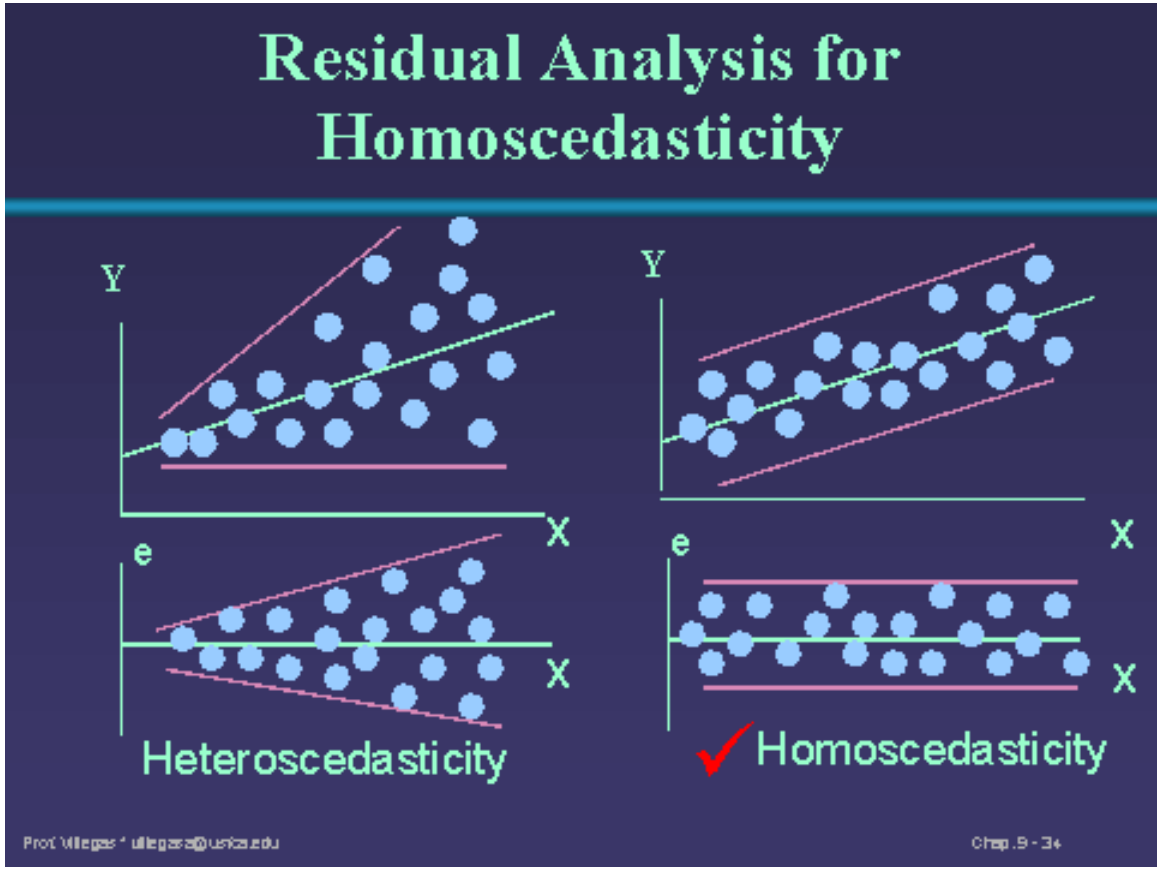
Residuo studentizado $r_i = \frac{e_i}{S\sqrt{1 - h_{ii}}}$

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

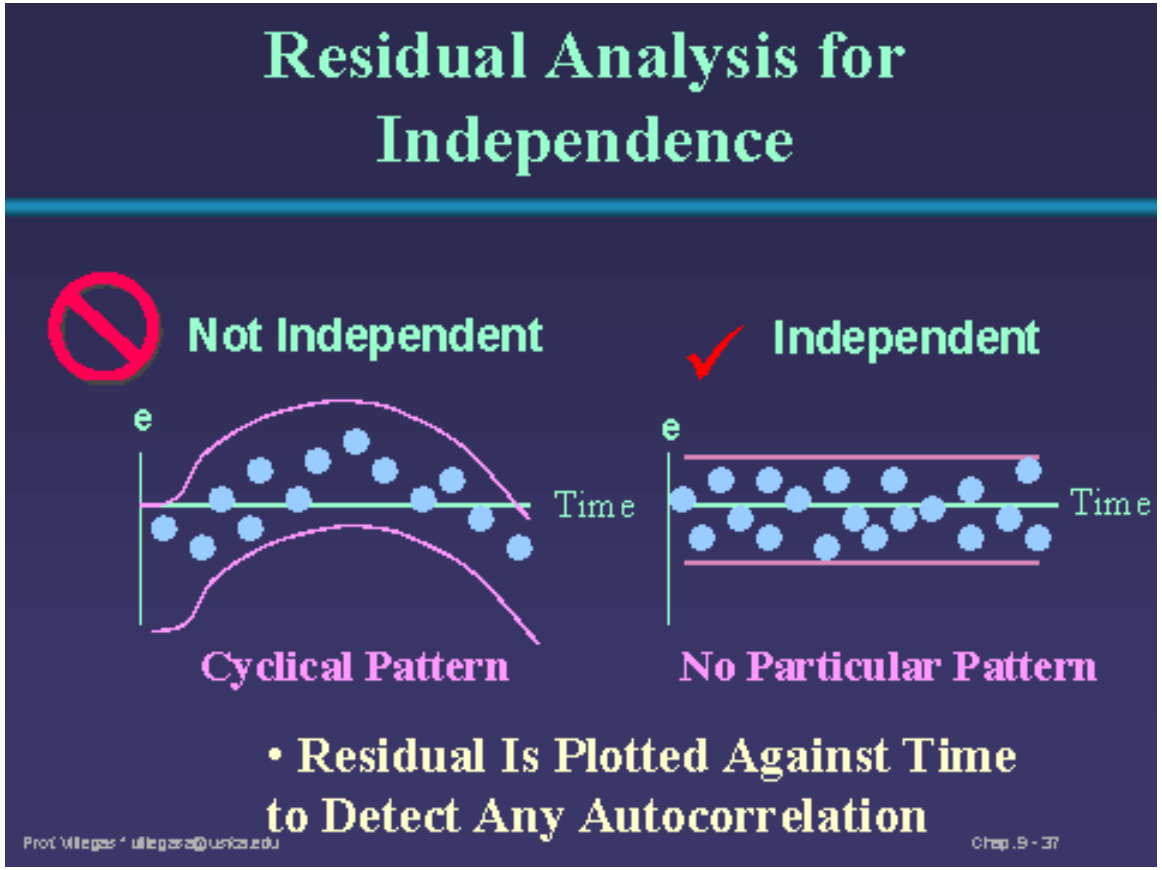
Gráficos de Residuos



Gráficos de Residuos



Gráficos de Residuos



Detección de valores atípicos y puntos de influencia

Detección de outliers en *y*

- Análisis de residuos y gráficos de los residuos

Detección de outliers en los predictores

Los h_{ii} $0 < h_{ii} < 1$ $\sum h_{ii} = p + 1$

Regla: se deben examinar los puntos con

$$h_{ii} \geq {}^{2(p+1)}/_{n}$$

Siendo n la cantidad de datos y p la cantidad de parámetros de regresión estimados

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 imite 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

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

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)