

# MODELOS LINEALES

## DIAGNÓSTICO

Fernando Massa; Bruno Bellagamba

30 de abril 2024



FACULTAD DE  
CIENCIAS ECONÓMICAS  
Y DE ADMINISTRACIÓN

**IESTA** INSTITUTO  
DE ESTADÍSTICA



UNIVERSIDAD  
DE LA REPÚBLICA  
URUGUAY



# TEMARIO

1 INTRODUCCIÓN

2 RESIDUOS

3 HOMOSCEDASTICIDAD

4 NORMALIDAD

5 OBSERVACIONES ATÍPICAS Y/O INFLUYENTES

6 PRÓXIMA CLASE



## OBJETIVOS

- Continuar con la etapa de diagnóstico.
- Definir los residuos *normalizados* y *estudentizados*.
- Describir herramientas gráficas y formales para corroborar el cumplimiento del supuesto de **homoscedasticidad**.
- Ídem para el supuesto de **normalidad**.
- Identificar observaciones atípicas y/o influyentes.

## LO QUE YA VIMOS

La clase anterior nos adentramos en la etapa de diagnóstico del modelo, indicando que su cumplimiento es importante a la hora de realizar inferencia.

## LO QUE YA VIMOS

La clase anterior nos adentramos en la etapa de diagnóstico del modelo, indicando que su cumplimiento es importante a la hora de realizar inferencia.

Realizamos una discusión del supuesto de multicolinealidad (aproximada) y como es posible estudiar esto en un momento *anterior* al ajuste del modelo. Luego llevamos a cabo un análisis meramente gráfico del supuesto de linealidad.

## LO QUE YA VIMOS

La clase anterior nos adentramos en la etapa de diagnóstico del modelo, indicando que su cumplimiento es importante a la hora de realizar inferencia.

Realizamos una discusión del supuesto de multicolinealidad (aproximada) y como es posible estudiar esto en un momento *anterior* al ajuste del modelo. Luego llevamos a cabo un análisis meramente gráfico del supuesto de linealidad.

Hoy, nos enfrentaremos a los dos supuestos que tienen que ver con la distribución de los errores: la normalidad y la homoscedasticidad.

## RESIDUOS

Anteriormente, observamos que el vector de residuos se puede obtener como:

$$\hat{\varepsilon} = (I_n - H) Y$$

A partir, de este planteamiento, pudimos calcular la matriz de covarianzas de este vector, observando que:

$$\text{Var}(\hat{\varepsilon}) = \sigma^2 (I_n - H)$$

Lo cual indica que los residuos SIEMPRE son heteroscedásticos (salvo que las variables explicativas sean ortogonales). Esto nos indica, que pueden no ser la mejor opción para emplearlos como un fascimil razonable de los errores.

## RESIDUOS

Anteriormente, observamos que el vector de residuos se puede obtener como:

$$\hat{\varepsilon} = (I_n - H) Y$$

A partir, de este planteamiento, pudimos calcular la matriz de covarianzas de este vector, observando que:

$$\text{Var}(\hat{\varepsilon}) = \sigma^2 (I_n - H)$$

Lo cual indica que los residuos SIEMPRE son heteroscedásticos (salvo que las variables explicativas sean ortogonales). Esto nos indica, que pueden no ser la mejor opción para emplearlos como un fascimil razonable de los errores.

Iratando de solucionar este problema, surgen los residuos **estandarizados** y los residuos **estudentizados**.

## RESIDUOS

El primer método para re-escalar los residuos consiste en dividirlos entre su desviación estándar, es decir.

$$r_i = \frac{\hat{\varepsilon}_i}{\sigma \sqrt{1 - h_{ii}}}$$

## RESIDUOS

El primer método para re-escalar los residuos consiste en dividirlos entre su desviación estándar, es decir.

$$r_i = \frac{\hat{\varepsilon}_i}{\sigma \sqrt{1 - h_{ii}}}$$

No obstante, este método tiene el problema de que requiere conocer el valor del parámetro  $\sigma^2$ . De esta manera surgen los residuos *studentizados*.

$$s_i = \frac{\hat{\varepsilon}_i}{s \sqrt{1 - h_{ii}}}$$

Siendo  $s = \sqrt{\frac{SCRes}{n-k-1}}$ .

## EXTERNA O INTERNAMENTE STUDENTIZADOS

En cuanto a los residuos studentizados, existen dos maneras de resolver el problema de la heteroscedasticidad. Ambas difieren en la forma de cálculo de  $s$ .

## EXTERNA O INTERNAMENTE STUDENTIZADOS

En cuanto a los residuos studentizados, existen dos maneras de resolver el problema de la heteroscedasticidad. Ambas difieren en la forma de cálculo de  $s$ .

Los residuos studentizados **internamente** son:

$$s_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_{ii}}}$$

Mientras que los residuos studentizados **externamente** son:

$$t_i = \frac{\hat{\varepsilon}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

## EXTERNA O INTERNAMENTE STUDENTIZADOS

En cuanto a los residuos studentizados, existen dos maneras de resolver el problema de la heteroscedasticidad. Ambas difieren en la forma de cálculo de  $s$ .

Los residuos studentizados **internamente** son:

$$s_i = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_{ii}}}$$

Mientras que los residuos studentizados **externamente** son:

$$t_i = \frac{\hat{\varepsilon}_i}{s_{(i)}\sqrt{1-h_{ii}}}$$

La diferencia radica en que en la studentización externa, la observación  $i$ -ésima no participa en la estimación de  $s$ . Las razones para hacer esto son:

- Poder identificar de manera más clara posibles observaciones atípicas.
- Evitar que se *inflé* la varianza de los residuos.

## ¿POR QUÉ HACER DIAGNÓSTICO?

Definimos los insumos que usaremos para llevar a cabo la etapa de diagnóstico, pero recordemos por qué nos interesan tanto los supuestos de normalidad y homoscedasticidad.

## ¿POR QUÉ HACER DIAGNÓSTICO?

Definimos los insumos que usaremos para llevar a cabo la etapa de diagnóstico, pero recordemos por qué nos interesan tanto los supuestos de normalidad y homoscedasticidad.

Estos supuestos forman la piedra angular sobre la cual descansa la distribución de los estadísticos que usamos al realizar inferencia. Las pruebas  $F$ , las pruebas  $t$ , los valores críticos, los p-valores, los intervalos de confianza, todos estos requieren el cumplimiento de estos supuestos (además de la independencia) para ser válidos.

## ¿POR QUÉ HACER DIAGNÓSTICO?

Definimos los insumos que usaremos para llevar a cabo la etapa de diagnóstico, pero recordemos por qué nos interesan tanto los supuestos de normalidad y homoscedasticidad.

Estos supuestos forman la piedra angular sobre la cual descansa la distribución de los estadísticos que usamos al realizar inferencia. Las pruebas  $F$ , las pruebas  $t$ , los valores críticos, los p-valores, los intervalos de confianza, todos estos requieren el cumplimiento de estos supuestos (además de la independencia) para ser válidos.

Cuando se cumplen, las inferencias que realizamos, son válidas. Cuando no...

## ¿POR QUÉ HACER DIAGNÓSTICO?

Definimos los insumos que usaremos para llevar a cabo la etapa de diagnóstico, pero recordemos por qué nos interesan tanto los supuestos de normalidad y homoscedasticidad.

Estos supuestos forman la piedra angular sobre la cual descansa la distribución de los estadísticos que usamos al realizar inferencia. Las pruebas  $F$ , las pruebas  $t$ , los valores críticos, los p-valores, los intervalos de confianza, todos estos requieren el cumplimiento de estos supuestos (además de la independencia) para ser válidos.

Cuando se cumplen, las inferencias que realizamos, son válidas. Cuando no... ya veremos que hacemos.

## HOMOSCEDASTICIDAD

El primer supuesto que estudiaremos es el de homoscedasticidad. Recordémoslo brevemente.

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i = 1, 2, \dots, n$$

## HOMOSCEDASTICIDAD

El primer supuesto que estudiaremos es el de homoscedasticidad. Recordémoslo brevemente.

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i = 1, 2, \dots, n$$

Una lectura directa de esta situación, indica que la varianza de los errores es **constante**, no obstante, nuestra forma de enfrentarnos a este supuesto es la siguiente:

*La varianza de los errores NO depende de ninguna de las variables explicativas*

## HOMOSCEDASTICIDAD

El primer supuesto que estudiaremos es el de homoscedasticidad. Recordémoslo brevemente.

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i = 1, 2, \dots, n$$

Una lectura directa de esta situación, indica que la varianza de los errores es **constante**, no obstante, nuestra forma de enfrentarnos a este supuesto es la siguiente:

*La varianza de los errores NO depende de ninguna de las variables explicativas*

De esta manera, si bien existen infinitas maneras en las que la varianza de los errores puede NO ser constante, solo nos enfocaremos en aquellas donde alguna (o algunas) de las variables explicativas *explica* el cambio en la varianza.

## MÉTODO GRÁFICO

Una primera aproximación al diagnóstico de la heteroscedasticidad consiste en graficar los residuos con respecto a cada variable explicativa. Dependiendo de la naturaleza de las variables explicativas, esto lo podemos llevar a cabo de 2 maneras.

- $x_j$  cuantitativa: Realizamos un diagrama de dispersión.

En este caso esperamos no detectar ningún patrón observable a lo largo del recorrido de  $x_j$ .

- $x_j$  cualitativa: Realizamos un gráfico de caja.

En este caso esperamos que las cajas presenten una variabilidad similar en todos los niveles de  $x_j$ .



## EJEMPLO

Para ilustrar lo que se planteó en la diapositiva anterior, realizaremos un análisis del peso de un conjunto de niños de entre 12 y 60 meses de edad.



## EJEMPLO

Para ilustrar lo que se planteó en la diapositiva anterior, realizaremos un análisis del peso de un conjunto de niños de entre 12 y 60 meses de edad.



Carguemos los datos en , construyamos un modelo de regresión y exploremos la presencia de heteroscedasticidad.

## TEST DE HIPÓTESIS

Existen varios estadísticos para llevar a cabo la siguiente prueba de hipótesis.

$$H_0) \quad \text{Var}(\varepsilon_i) = \sigma^2$$

$$H_1) \quad \text{no } H_0$$

## TEST DE HIPÓTESIS

Existen varios estadísticos para llevar a cabo la siguiente prueba de hipótesis.

$$H_0) \quad \text{Var}(\varepsilon_i) = \sigma^2$$

$$H_1) \quad \text{no } H_0$$

En la práctica, la hipótesis alternativa de nuestra prueba de hipótesis tiene la siguiente forma:

$$H_0) \quad \text{Var}(\varepsilon_i) = \sigma^2$$

$$H_1) \quad \text{Var}(\varepsilon_i) = \sigma^2 \times h(X_1, X_2, \dots, X_k)$$

## TEST DE HIPÓTESIS

Existen varios estadísticos para llevar a cabo la siguiente prueba de hipótesis.

$$H_0) \quad \text{Var}(\varepsilon_i) = \sigma^2$$

$$H_1) \quad \text{no } H_0$$

En la práctica, la hipótesis alternativa de nuestra prueba de hipótesis tiene la siguiente forma:

$$H_0) \quad \text{Var}(\varepsilon_i) = \sigma^2$$

$$H_1) \quad \text{Var}(\varepsilon_i) = \sigma^2 \times h(X_1, X_2, \dots, X_k)$$

Irónicamente, el estadístico de prueba para llevar a cabo esta prueba de hipótesis, requiere que ajustemos otro modelo de regresión lineal

## TEST DE BREUSCH-PAGAN

En el modelo de *RLM* se analiza el efecto de un conjunto de variables explicativas sobre el valor esperado de la variable de respuesta. Es decir:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$$

## TEST DE BREUSCH-PAGAN

En el modelo de *RLM* se analiza el efecto de un conjunto de variables explicativas sobre el valor esperado de la variable de respuesta. Es decir:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$$

Uno de los supuestos (el que nos interesa en este momento) indica que:

$$\text{Var}(Y_i) = \text{Var}(\varepsilon_i) = \sigma^2$$

## TEST DE BREUSCH-PAGAN

En el modelo de *RLM* se analiza el efecto de un conjunto de variables explicativas sobre el valor esperado de la variable de respuesta. Es decir:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$$

Uno de los supuestos (el que nos interesa en este momento) indica que:

$$\text{Var}(Y_i) = \text{Var}(\varepsilon_i) = \sigma^2$$

Pero si además tenemos en cuenta que los errores tienen media cero, podemos modificar ligeramente la ecuación anterior y plantear:

$$\text{Var}(Y_i) = \text{Var}(\varepsilon_i) = \mathbb{E}(\varepsilon_i^2) = \sigma^2$$

## TEST DE BREUSCH-PAGAN

A partir de la última observación, podríamos pensar que *si alguna variable fuese la responsable de la falta de homoscedasticidad* esto se podría detectar planteando:

$$\begin{aligned}\mathbb{E}(\varepsilon_i^2) &= \sigma^2 [\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \dots + \gamma_k X_{ik}] \\ \mathbb{E}(\varepsilon_i^2) &= \gamma_0^* + \gamma_1^* X_{i1} + \gamma_2^* X_{i2} + \dots + \gamma_k^* X_{ik}\end{aligned}$$

## TEST DE BREUSCH-PAGAN

A partir de la última observación, podríamos pensar que *si alguna variable fuese la responsable de la falta de homoscedasticidad* esto se podría detectar planteando:

$$\begin{aligned}\mathbb{E}(\varepsilon_i^2) &= \sigma^2 [\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \dots + \gamma_k X_{ik}] \\ \mathbb{E}(\varepsilon_i^2) &= \gamma_0^* + \gamma_1^* X_{i1} + \gamma_2^* X_{i2} + \dots + \gamma_k^* X_{ik}\end{aligned}$$

Entonces, el estadístico de *Breusch-Pagan* consiste en:

## TEST DE BREUSCH-PAGAN

A partir de la última observación, podríamos pensar que *si alguna variable fuese la responsable de la falta de homoscedasticidad* esto se podría detectar planteando:

$$\begin{aligned}\mathbb{E}(\varepsilon_i^2) &= \sigma^2 [\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \dots + \gamma_k X_{ik}] \\ \mathbb{E}(\varepsilon_i^2) &= \gamma_0^* + \gamma_1^* X_{i1} + \gamma_2^* X_{i2} + \dots + \gamma_k^* X_{ik}\end{aligned}$$

Entonces, el estadístico de *Breusch-Pagan* consiste en:

- 1 Obtener los residuos del modelo lineal que se quiere diagnosticar ( $Y = X\beta + \varepsilon$ )

## TEST DE BREUSCH-PAGAN

A partir de la última observación, podríamos pensar que *si alguna variable fuese la responsable de la falta de homoscedasticidad* esto se podría detectar planteando:

$$\begin{aligned}\mathbb{E}(\varepsilon_i^2) &= \sigma^2 [\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \dots + \gamma_k X_{ik}] \\ \mathbb{E}(\varepsilon_i^2) &= \gamma_0^* + \gamma_1^* X_{i1} + \gamma_2^* X_{i2} + \dots + \gamma_k^* X_{ik}\end{aligned}$$

Entonces, el estadístico de *Breusch-Pagan* consiste en:

- ① Obtener los residuos del modelo lineal que se quiere diagnosticar ( $Y = X\beta + \varepsilon$ )
- ② Elevar estos residuos al cuadrado ( $\hat{\varepsilon}^2$ ).

## TEST DE BREUSCH-PAGAN

A partir de la última observación, podríamos pensar que *si alguna variable fuese la responsable de la falta de homoscedasticidad* esto se podría detectar planteando:

$$\begin{aligned}\mathbb{E}(\varepsilon_i^2) &= \sigma^2 [\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \dots + \gamma_k X_{ik}] \\ \mathbb{E}(\varepsilon_i^2) &= \gamma_0^* + \gamma_1^* X_{i1} + \gamma_2^* X_{i2} + \dots + \gamma_k^* X_{ik}\end{aligned}$$

Entonces, el estadístico de *Breusch-Pagan* consiste en:

- ① Obtener los residuos del modelo lineal que se quiere diagnosticar ( $Y = X\beta + \varepsilon$ )
- ② Elevar estos residuos al cuadrado ( $\hat{\varepsilon}^2$ ).
- ③ Realizar la regresión auxiliar ( $\hat{\varepsilon}^2 = X\gamma + \eta$ ).

## TEST DE BREUSCH-PAGAN

A partir de la última observación, podríamos pensar que *si alguna variable fuese la responsable de la falta de homoscedasticidad* esto se podría detectar planteando:

$$\begin{aligned}\mathbb{E}(\varepsilon_i^2) &= \sigma^2 [\gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \dots + \gamma_k X_{ik}] \\ \mathbb{E}(\varepsilon_i^2) &= \gamma_0^* + \gamma_1^* X_{i1} + \gamma_2^* X_{i2} + \dots + \gamma_k^* X_{ik}\end{aligned}$$

Entonces, el estadístico de *Breusch-Pagan* consiste en:

- ① Obtener los residuos del modelo lineal que se quiere diagnosticar ( $Y = X\beta + \varepsilon$ )
- ② Elevar estos residuos al cuadrado ( $\hat{\varepsilon}^2$ ).
- ③ Realizar la regresión auxiliar ( $\hat{\varepsilon}^2 = X\gamma + \eta$ ).
- ④ Calcular  $BP = n \times R^2$  que, bajo  $H_0$  cierta, converge en distribución a  $\chi_k^2$ .

## TEST DE BREUSCH-PAGAN

Bajo el cumplimiento de  $H_0$  (los errores son homoscedásticos), las variables explicativas no deberían explicar la variabilidad de  $\hat{\epsilon}^2$  y por ende el  $R^2$  de la regresión auxiliar debería ser “*bajo*”.

## TEST DE BREUSCH-PAGAN

Bajo el cumplimiento de  $H_0$  (los errores son homoscedásticos), las variables explicativas no deberían explicar la variabilidad de  $\hat{\epsilon}^2$  y por ende el  $R^2$  de la regresión auxiliar debería ser “*bajo*”.

La prueba postula la condición de homoscedasticidad en la hipótesis nula de la siguiente manera:

$$H_0) \gamma_1 = \gamma_2 = \dots = \gamma_k = 0$$

$$H_1) \text{al menos uno de los } \gamma \neq 0$$

## TEST DE BREUSCH-PAGAN

Bajo el cumplimiento de  $H_0$  (los errores son homoscedásticos), las variables explicativas no deberían explicar la variabilidad de  $\hat{\epsilon}^2$  y por ende el  $R^2$  de la regresión auxiliar debería ser “*bajo*”.

La prueba postula la condición de homoscedasticidad en la hipótesis nula de la siguiente manera:

$$H_0) \gamma_1 = \gamma_2 = \dots = \gamma_k = 0$$

$$H_1) \text{al menos uno de los } \gamma \neq 0$$

En esta prueba nos interesa NO RECHAZAR la hipótesis nula.



## EJEMPLO

Para diagnosticar la falta de homoscedasticidad en el ejemplo del peso de los bebés de un modo mas formal, usemos la prueba de Breusch-Pagan.



Volvamos al  y llevemos a cabo la prueba.

## NORMALIDAD

Tal vez sea el supuesto que más tenemos en cuenta a la hora de llevar a cabo el diagnóstico, debido a que es fácil notar cuánto lo necesitamos para plantear la verosimilitud y construir los estadísticos de prueba.

## NORMALIDAD

Tal vez sea el supuesto que más tenemos en cuenta a la hora de llevar a cabo el diagnóstico, debido a que es fácil notar cuánto lo necesitamos para plantear la verosimilitud y construir los estadísticos de prueba.

No obstante (como veremos más adelante), cuando el tamaño muestral es lo suficientemente grande, la falta de normalidad de los errores no presenta grandes repercusiones.

## NORMALIDAD

Tal vez sea el supuesto que más tenemos en cuenta a la hora de llevar a cabo el diagnóstico, debido a que es fácil notar cuánto lo necesitamos para plantear la verosimilitud y construir los estadísticos de prueba.

No obstante (como veremos más adelante), cuando el tamaño muestral es lo suficientemente grande, la falta de normalidad de los errores no presenta grandes repercusiones.

De la misma manera que procedimos con el supuesto de homoscedasticidad, a la hora de comprobar la normalidad tendremos métodos gráficos y métodos basados en pruebas de hipótesis.

- **Métodos gráficos:** Q-Q plots e histogramas.
- **Pruebas de hipótesis:** Shapiro-Wilk, Jarque-Bera, Kolmogorov-Smirnov, etc

## Q-Q PLOT

Consiste en un diagrama de dispersión donde se comparan los cuantiles empíricos de una muestra (observaciones) ordenadas con sus correspondientes versiones teóricas bajo cierta función de densidad.

## Q-Q PLOT

Consiste en un diagrama de dispersión donde se comparan los cuantiles empíricos de una muestra (observaciones) ordenadas con sus correspondientes versiones teóricas bajo cierta función de densidad.

Para evaluar la normalidad de los residuos, los pasos serían los siguientes:

- Obtener los residuos studentizados del modelo (media 0 y varianza 1).
- Los cuantiles empíricos no son más que estos valores ordenados de mayor a menor:  
 $t_{(1)}, t_{(2)}, \dots, t_{(n)}$ .
- Obtenemos los cuantiles teóricos mediante  $z_{(i)} = \Phi^{-1} \left( \frac{i}{n+1} \right)$ .

En nuestro caso, queremos que los puntos de este diagrama estén sobre una linea recta (podemos valernos de una banda de confianza).

## HISTOGRAMA

Consiste en una estimación no paramétrica de la densidad de una muestra de observaciones.

Dos aspectos a tener en cuenta son:

- El ancho de los *bins*: Muy angostos y el histograma se vuelve inestable, muy anchos y no muestra nada claro.
- Donde comenzar el primer *bin*: Asumiendo que todos tienen el mismo ancho, la ubicación del primer *bin* determina la posición de los demás. Ligeros cambios en este detalle pueden cambiar radicalmente la apariencia del histograma.

## HISTOGRAMA

Consiste en una estimación no paramétrica de la densidad de una muestra de observaciones.

Dos aspectos a tener en cuenta son:

- El ancho de los *bins*: Muy angostos y el histograma se vuelve inestable, muy anchos y no muestra nada claro.
- Donde comenzar el primer *bin*: Asumiendo que todos tienen el mismo ancho, la ubicación del primer *bin* determina la posición de los demás. Ligeros cambios en este detalle pueden cambiar radicalmente la apariencia del histograma.

En nuestro caso, necesitamos observar una distribución que se asemeje a una campana simétrica. Es posible complementarlos con una estimación suavizada de la densidad (más de esto en el curso de Estadística no paramétrica).



## EJEMPLO

Para explorar el supuesto de normalidad de los errores (a través de los residuos del modelo) utilizaremos los datos de la tasa de ahorro en un conjunto de 50 países.



Volvamos al ...

## TEST DE HIPÓTESIS

Existe una pléthora de pruebas de normalidad, sin embargo todas parten del mismo planteamiento. Disponiendo de una muestra de observaciones  $W_1, W_2, W_3, \dots, W_n$ :

$$H_0) W \sim \text{Normal}$$

$$H_1) \text{no } H_0)$$

## TEST DE HIPÓTESIS

Existe una pléthora de pruebas de normalidad, sin embargo todas parten del mismo planteamiento. Disponiendo de una muestra de observaciones  $W_1, W_2, W_3, \dots, W_n$ :

$$H_0) W \sim \text{Normal}$$

$$H_1) \text{no } H_0)$$

Nótese como en la hipótesis nula, no se puntualiza *qué* distribución normal. Por lo general se asume que los datos están standarizados (media 0 y varianza 1).

Lo importante es que todos provengan de **la misma** distribución normal, siendo este el motivo para emplear los residuos studentizados y no los residuos brutos ( $y_i - \hat{y}_i$ ).

## TEST DE NORMALIDAD

Como se mencionaba en la diapositiva anterior, existen numerosas pruebas de normalidad (y se alienta a los estudiantes a que investiguen si así lo desean), pero en este curso se presentarán algunas.

## TEST DE NORMALIDAD

Como se mencionaba en la diapositiva anterior, existen numerosas pruebas de normalidad (y se alienta a los estudiantes a que investiguen si así lo desean), pero en este curso se presentarán algunas.

- ① Prueba de **Shapiro-Wilk**: se basa en la comparación de los cuantiles empíricos y teóricos bajo el supuesto de normalidad.
- ② Prueba de **Jarque-Beta**: se basa en la comparación de los estadísticos de asimetría y kurtosis bajo el supuesto de normalidad
- ③ Prueba de **Kolmogorov-Smirnov**: se basa en la máxima discrepancia entre la Función de distribución empírica y la teórica bajo el supuesto de normalidad.

## TEST DE NORMALIDAD

Como se mencionaba en la diapositiva anterior, existen numerosas pruebas de normalidad (y se alienta a los estudiantes a que investiguen si así lo desean), pero en este curso se presentarán algunas.

- ① Prueba de **Shapiro-Wilk**: se basa en la comparación de los cuantiles empíricos y teóricos bajo el supuesto de normalidad.
- ② Prueba de **Jarque-Beta**: se basa en la comparación de los estadísticos de asimetría y kurtosis bajo el supuesto de normalidad
- ③ Prueba de **Kolmogorov-Smirnov**: se basa en la máxima discrepancia entre la Función de distribución empírica y la teórica bajo el supuesto de normalidad.

En todos los casos, nos interesa **NO RECHAZAR** la hipótesis nula.



## EJEMPLO

Ahora que contamos con algunas herramientas un poco más formales, volvamos al ejemplo de los ahorros y llevemos a cabo las pruebas de normalidad.



Volvamos al  ...

## OBSERVACIONES ATÍPICAS Y/O INFLUYENTES

A veces el modelo tiene un comportamiento “adecuado” para la mayoría de las observaciones excepto una (unas pocas) donde el (los) residuo(s) escapa a la norma. Estos casos son los que consideramos *outliers* u **observaciones atípicas**.

## OBSERVACIONES ATÍPICAS Y/O INFLUYENTES

A veces el modelo tiene un comportamiento “adecuado” para la mayoría de las observaciones excepto una (unas pocas) donde el (los) residuo(s) escapa a la norma. Estos casos son los que consideramos *outliers* u **observaciones atípicas**.

Cuando estos residuos se deben a un error en el ingreso de los datos, se deben corregir (o descartar) y continuar con el análisis. No obstante, esto puede deberse a otros motivos. En estos casos el procedimiento consiste en analizar el modelo CON y SIN estas observaciones.

## OBSERVACIONES ATÍPICAS Y/O INFLUYENTES

A veces el modelo tiene un comportamiento “adecuado” para la mayoría de las observaciones excepto una (unas pocas) donde el (los) residuo(s) escapa a la norma. Estos casos son los que consideramos *outliers* u **observaciones atípicas**.

Cuando estos residuos se deben a un error en el ingreso de los datos, se deben corregir (o descartar) y continuar con el análisis. No obstante, esto puede deberse a otros motivos. En estos casos el procedimiento consiste en analizar el modelo CON y SIN estas observaciones.

Cuando los resultados difieren de forma significativa entre ambos análisis, se suele tomar una de las 3 alternativas que se presentan a continuación:

- Mantener ambos análisis hasta que se recolecten más datos.
- Descartar el/los dato/s atípico/s.
- Utilizar métodos de estimación robustos a la presencia de observaciones atípicas.



## EXPLORANDO OBSERVACIONES ATÍPICAS

La forma de determinar qué residuos son grandes es a través de los residuos studentizados externamente. Esto se debe a:

- De no studentizar, es posible creer que un residuo es excesivamente grande, cuando en realidad, su varianza es grande.
- Tras studentizar, es posible llevar a cabo una prueba de hipótesis que permite determinar si el residuo es o no un *outlier*

## EXPLORANDO OBSERVACIONES ATÍPICAS

La forma de determinar qué residuos son grandes es a través de los residuos studentizados externamente. Esto se debe a:

- De no studentizar, es posible creer que un residuo es excesivamente grande, cuando en realidad, su varianza es grande.
- Tras studentizar, es posible llevar a cabo una prueba de hipótesis que permite determinar si el residuo es o no un *outlier*

Recordemos a los residuos externamente studentizados:

$$t_i = \frac{\hat{\varepsilon}_i}{s_{(i)}\sqrt{1 - h_{ii}}}$$

Debido a que la estimación de la varianza NO incluye a la  $i$ -ésima observación, el numerador y el denominador son independientes y (junto con otros supuestos) se puede determinar que estos residuos se distribuyen  $t_{n-k-1}$ .

## EXPLORANDO OBSERVACIONES ATÍPICAS

De esta manera es posible llevar a cabo  $n$  pruebas de hipótesis de la forma

$H_0$ ) La observación  $i$  NO es atípica

$H_1$ ) La observación  $i$  SI es atípica

## EXPLORANDO OBSERVACIONES ATÍPICAS

De esta manera es posible llevar a cabo  $n$  pruebas de hipótesis de la forma

$H_0$ ) La observación  $i$  NO es atípica

$H_1$ ) La observación  $i$  SI es atípica

El procedimiento requiere iterar entre las  $n$  observaciones (o al menos entre las “sospechosas”) de la siguiente forma:

## EXPLORANDO OBSERVACIONES ATÍPICAS

De esta manera es posible llevar a cabo  $n$  pruebas de hipótesis de la forma

$H_0$ ) La observación  $i$  NO es atípica

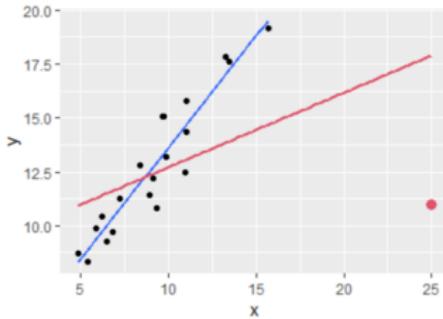
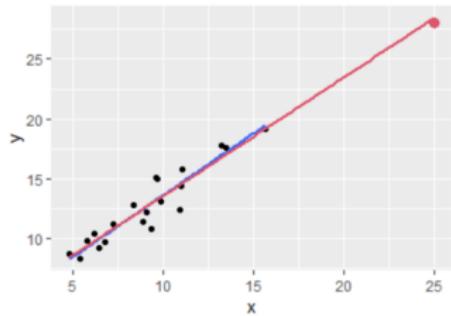
$H_1$ ) La observación  $i$  SI es atípica

El procedimiento requiere iterar entre las  $n$  observaciones (o al menos entre las "sospechosas") de la siguiente forma:

- Calcular el residuo studentizado externamente de la obsetvación  $i$ .
- Compararlo con un valor crítico de la distribución  $t_{n-k-1}$  (Típicamente con un valor de  $\alpha = 0,01$  o menor).
- De rechazar  $H_0$ , etiquetar la observación como atípica.

## OBSERVACIONES INFLUYENTES

En la práctica, las observaciones que realmente nos preocupan son las aquellas cuya presencia distorsiona los resultados del análisis. Típicamente, las que alteran el vector de estimaciones  $\hat{\beta}$  o el vector de valores ajustados  $\hat{Y}$ .



## *Leverage*

Un indicador útil para determinar la influencia de cada observación sobre los valores ajustados, se esconde en los valores de la diagonal de la matriz  $H$ .

## Leverage

Un indicador útil para determinar la influencia de cada observación sobre los valores ajustados, se esconde en los valores de la diagonal de la matriz  $H$ .

### Leverage

Cuantifica el cambio en el vector de predicciones ejercido por la presencia de cada observación.

$$\hat{Y} = HY$$

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \dots + \color{red}{h_{ii}y_i} + \dots + h_{in}y_n$$

## Leverage

Un indicador útil para determinar la influencia de cada observación sobre los valores ajustados, se esconde en los valores de la diagonal de la matriz  $H$ .

### Leverage

Cuantifica el cambio en el vector de predicciones ejercido por la presencia de cada observación.

$$\hat{Y} = HY$$

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \dots + \color{red}{h_{ii}y_i} + \dots + h_{in}y_n$$

Valores altos de  $h_i$  se dan en zonas del espacio de las  $x_j$  donde hay pocos puntos y suelen tener una gran influencia (*leverage*) sobre la recta de regresión.

## Leverage

Un indicador útil para determinar la influencia de cada observación sobre los valores ajustados, se esconde en los valores de la diagonal de la matriz  $H$ .

### Leverage

Cuantifica el cambio en el vector de predicciones ejercido por la presencia de cada observación.

$$\hat{Y} = HY$$

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \dots + \color{red}{h_{ii}y_i} + \dots + h_{in}y_n$$

Valores altos de  $h_i$  se dan en zonas del espacio de las  $x_j$  donde hay pocos puntos y suelen tener una gran influencia (*leverage*) sobre la recta de regresión.

Para las observaciones con mayor leverage existen reglas empíricas ( $h_i > \frac{2(k+1)}{n}$ ) pero resulta conveniente comparar todos los valores del vector  $h = (h_1, h_2, \dots, h_n)$ .

## DISTANCIA DE COOK

No disponemos de pruebas de hipótesis que nos permitan llevar a cabo un diagnóstico formal de la influencia. Pero contamos con algunos indicadores que pueden ser de utilidad.

## DISTANCIA DE COOK

No disponemos de pruebas de hipótesis que nos permitan llevar a cabo un diagnóstico formal de la influencia. Pero contamos con algunos indicadores que pueden ser de utilidad.

### Distancia de Cook

Cuantifica el cambio en el vector de estimaciones luego de remover la i-ésima observación.

$$\begin{aligned} D_i &= \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})}{(k+1)s^2} \\ &= \frac{(X\hat{\beta}_{(i)} - X\hat{\beta})' (X\hat{\beta}_{(i)} - X\hat{\beta})}{(k+1)s^2} \\ &= \frac{(\hat{y}_{(i)} - \hat{y})' (\hat{y}_{(i)} - \hat{y})}{(k+1)s^2} \end{aligned}$$

## DISTANCIA DE COOK

No disponemos de pruebas de hipótesis que nos permitan llevar a cabo un diagnóstico formal de la influencia. Pero contamos con algunos indicadores que pueden ser de utilidad.

### Distancia de Cook

Cuantifica el cambio en el vector de estimaciones luego de remover la i-ésima observación.

$$\begin{aligned} D_i &= \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})}{(k+1)s^2} \\ &= \frac{(X\hat{\beta}_{(i)} - X\hat{\beta})' (X\hat{\beta}_{(i)} - X\hat{\beta})}{(k+1)s^2} \\ &= \frac{(\hat{y}_{(i)} - \hat{y})' (\hat{y}_{(i)} - \hat{y})}{(k+1)s^2} \end{aligned}$$

Véase como  $D_i$  adicionalmente cuantifica el cambio en el vector de valores ajustados.

Para determinar si alguna observación tiene una influencia “significativa” existen algunas reglas empíricas ( $D_i > \frac{4}{n}$ ) pero es conveniente comparar los valores del vector  $D = (D_1, D_2, \dots, D_n)$ .



## EJEMPLO

Retomemos (por última vez) el ejemplo de la tasa de ahorro de los países para ver como obtener estos indicadores en R.



Volvamos al  y terminemos con esta clase.



## EN LA PRÓXIMA CLASE

La próxima clase veremos qué alternativas tenemos para trabajar con modelos en los que NO se cumplan los supuestos.

- Transformaciones.
- Métodos computacionales.
- Métodos robustos.



## BIBLIOGRAFÍA

-  Carmona, Francesc (2003). *Modelos Lineales (notas de curso)*. Departament d'Estadística.
-  Faraway, Julian (2014). *Linear Models with R, second edition*. Chapman Hall/CRC.
-  Rencher, Alvin y Bruce Schaalje (2008). *Linear Models in Statistics, second edition*. John Wiley Sons, Inc.

¿Preguntas?

# Muchas Gracias