

MODELOS LINEALES

DIAGNÓSTICO - ¿Y AHORA QUÉ?

Fernando Massa; Bruno Bellagamba

23 de mayo 2023



FACULTAD DE
CIENCIAS ECONÓMICAS
Y DE ADMINISTRACIÓN

IESTA INSTITUTO
DE ESTADÍSTICA



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



TEMARIO

1 INTRODUCCIÓN

2 TRANSFORMACIONES

3 ESPECIFICACIÓN

4 OBSERVACIONES ATÍPICAS

5 INFERENCIA COMPUTACIONAL

6 OTROS MÉTODOS

7 PRÓXIMA CLASE



OBJETIVOS

- Conocer distintas alternativas para “remediar” situaciones donde los supuestos no se cumplen.
- Plantear algunas alternativas para realizar inferencia en situaciones donde no es posible remediar el no cumplimiento de los supuestos.
- Diagramar una serie de pasos a seguir en la etapa de diagnóstico.

INTRODUCCIÓN

En las clases anteriores repasamos los supuestos sobre los que descansa la teoría de modelos lineales bajo la cual nos basamos para realizar inferencia.

INTRODUCCIÓN

En las clases anteriores repasamos los supuestos sobre los que descansa la teoría de modelos lineales bajo la cual nos basamos para realizar inferencia.

Se introdujeron herramientas para diagnosticar el no cumplimiento de los supuestos.

- Multicolinealidad (aproximada).
- Linealidad.
- Homoscedasticidad
- Normalidad
- Observaciones atípicas/influyentes.

INTRODUCCIÓN

En las clases anteriores repasamos los supuestos sobre los que descansa la teoría de modelos lineales bajo la cual nos basamos para realizar inferencia.

Se introdujeron herramientas para diagnosticar el no cumplimiento de los supuestos.

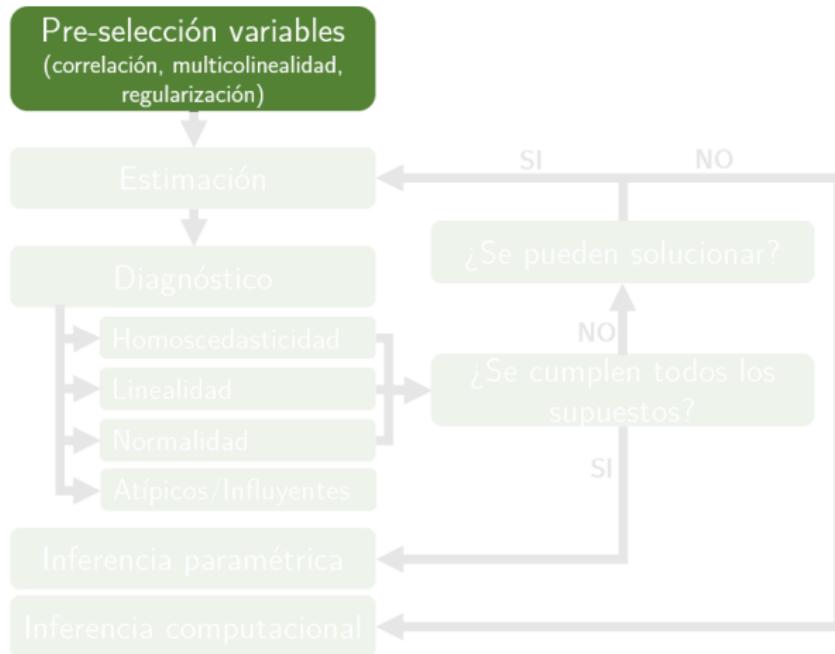
- Multicolinealidad (aproximada).
- Linealidad.
- Homoscedasticidad
- Normalidad
- Observaciones atípicas/influyentes.

Hoy plantearemos distintas estrategias a seguir con 2 cursos de acción:

- ① Intentar “corregir” el modelo para que se cumplan los supuestos.
- ② Cuando no sea posible obtener un modelo que siga los supuestos, realizar inferencias con procedimientos que se adecúen a estas situaciones.

ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



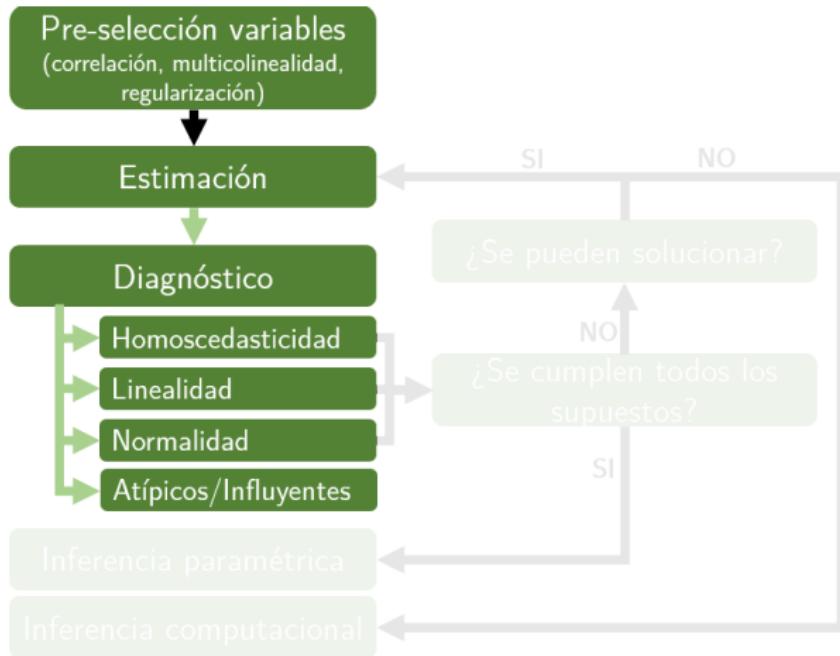
ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



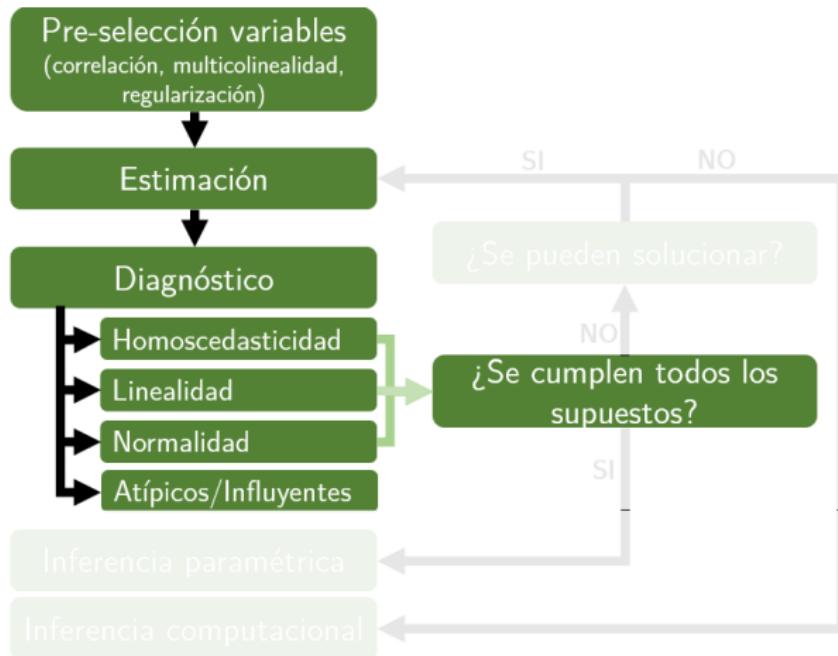
ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



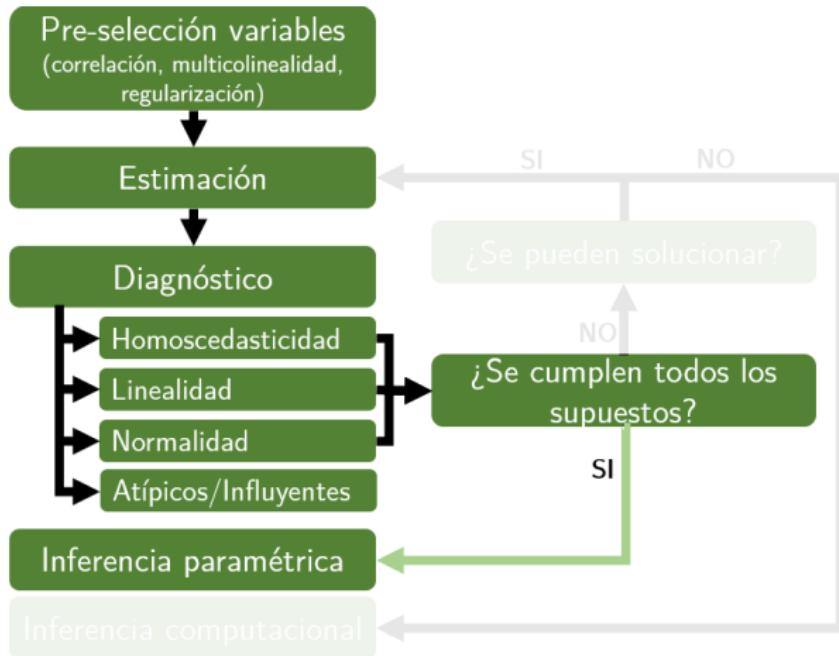
ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



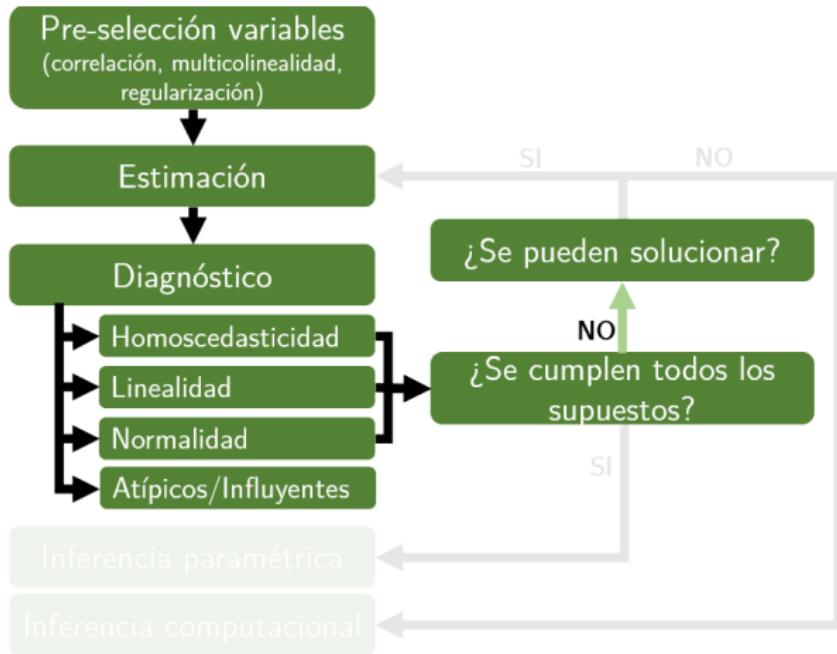
ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



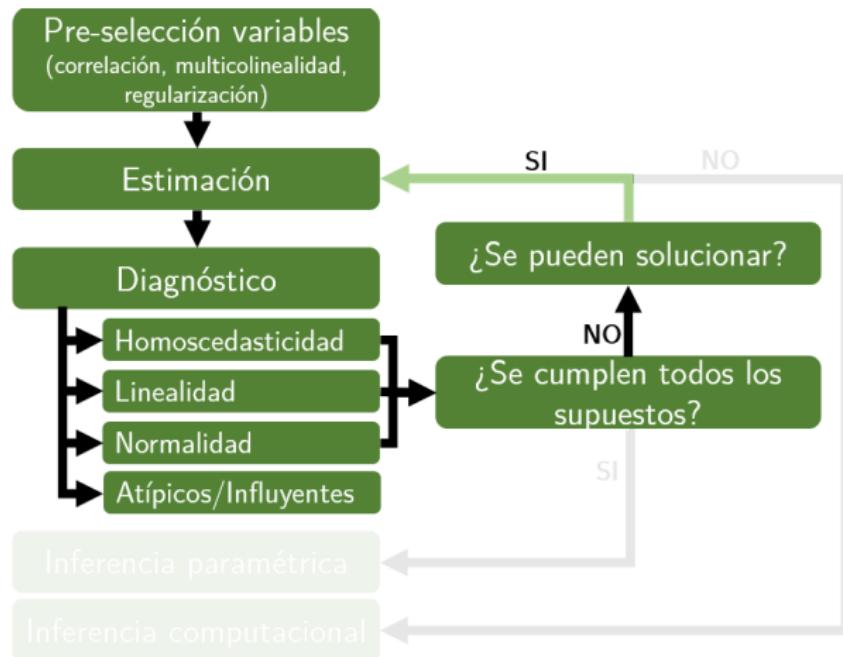
ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



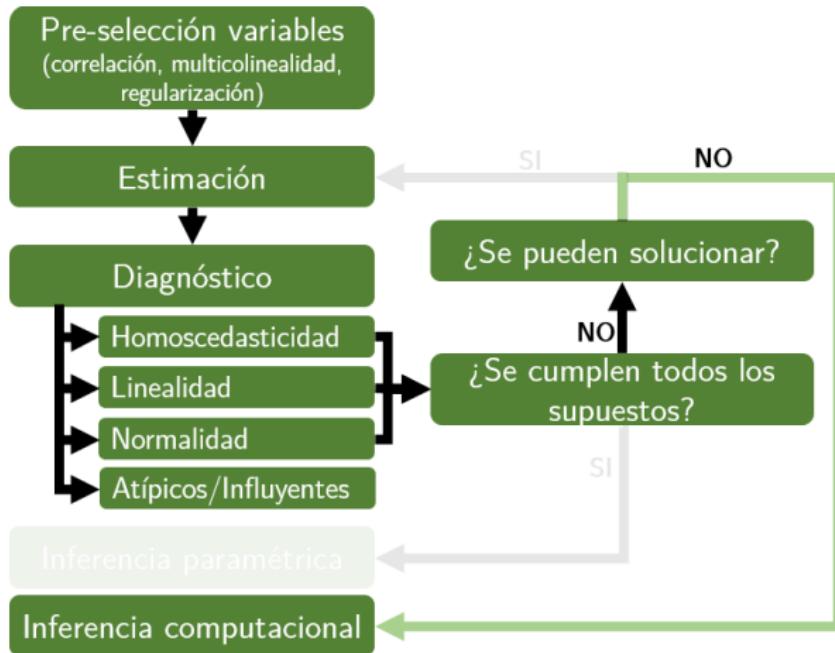
ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



ESTRATEGIA

Un posible diagrama de flujo que podemos seguir a la hora de trabajar con modelos estadísticos se ilustra en la siguiente figura:



TRANSFORMACIONES

Varios de los problemas que surgen en la etapa de diagnóstico pueden solucionarse transformando la variable de respuesta y/o las variables explicativas.

TRANSFORMACIONES

Varios de los problemas que surgen en la etapa de diagnóstico pueden solucionarse transformando la variable de respuesta y/o las variables explicativas.

Sin embargo, esto tiene sus consecuencias. Salvo la transformación logarítmica, el resto de las transformaciones dificultan la interpretación y análisis del modelo.

TRANSFORMACIONES

Varios de los problemas que surgen en la etapa de diagnóstico pueden solucionarse transformando la variable de respuesta y/o las variables explicativas.

Sin embargo, esto tiene sus consecuencias. Salvo la transformación logarítmica, el resto de las transformaciones dificultan la interpretación y análisis del modelo.

La transformación logarítmica tiene la virtud de que:

- Soluciona problemas de homoscedasticidad cuando la variabilidad de los residuos aumenta al predecir valores más grandes de Y .
- En algunas situaciones, cuando los residuos presentan asimetría positiva, puede solucionar problemas de normalidad.
- Puede solucionar casos donde se evidencia que la relación de Y con una o más variables explicativas no sea lineal.

TRANSFORMACIÓN LOGARÍTMICA

En la diapositiva anterior se argumentó que la transformación logarítmica es la única con la que es posible interpretar los resultados del modelo lineal. Existen 3 posibles situaciones:

TRANSFORMACIÓN LOGARÍTMICA

En la diapositiva anterior se argumentó que la transformación logarítmica es la única con la que es posible interpretar los resultados del modelo lineal. Existen 3 posibles situaciones:

- Transformar Y .
- Transformar alguna/todas las X .
- Transformar Y alguna/todas las X .

TRANSFORMACIÓN LOGARÍTMICA

En la diapositiva anterior se argumentó que la transformación logarítmica es la única con la que es posible interpretar los resultados del modelo lineal. Existen 3 posibles situaciones:

- Transformar Y .
- Transformar alguna/todas las X .
- Transformar Y alguna/todas las X .

A continuación se presentan estos casos en un modelo de *RLS* pero las interpretaciones se generalizan sin dificultad al caso del modelo de *RLM*.

TRANSFORMAR Y

Consideré el modelo de *RLS* pero empleando como variable de respuesta $\log(Y)$, el modelo sería

$$\log(Y) = \beta_0 + \beta_1 X + \varepsilon$$

TRANSFORMAR Y

Consideré el modelo de *RLS* pero empleando como variable de respuesta $\log(Y)$, el modelo sería

$$\log(Y) = \beta_0 + \beta_1 X + \varepsilon$$

Para interpretar el coeficiente β_1 se resta esta ecuación evaluada en X_1 a su análoga evaluada en X_2 , obteniendo:

$$\begin{aligned}\log(Y_1) - \log(Y_2) &= \beta_1 X_1 - \beta_1 X_2 \\ \log\left(\frac{Y_1}{Y_2}\right) &= \beta_1 (X_1 - X_2) \\ \frac{Y_1}{Y_2} &= e^{\beta_1(X_1 - X_2)} \\ \frac{Y_1}{Y_2} &\approx (1 + \beta_1)^{(X_1 - X_2)}\end{aligned}$$

TRANSFORMAR Y

Consideré el modelo de *RLS* pero empleando como variable de respuesta $\log(Y)$, el modelo sería

$$\log(Y) = \beta_0 + \beta_1 X + \varepsilon$$

Para interpretar el coeficiente β_1 se resta esta ecuación evaluada en X_1 a su análoga evaluada en X_2 , obteniendo:

$$\begin{aligned}\log(Y_1) - \log(Y_2) &= \beta_1 X_1 - \beta_1 X_2 \\ \log\left(\frac{Y_1}{Y_2}\right) &= \beta_1 (X_1 - X_2) \\ \frac{Y_1}{Y_2} &= e^{\beta_1(X_1 - X_2)} \\ \frac{Y_1}{Y_2} &\approx (1 + \beta_1)^{(X_1 - X_2)}\end{aligned}$$

En la última igualdad se aplicó que $e^u \approx 1 + u$.

TRANSFORMAR Y

Consideré el modelo de *RLS* pero empleando como variable de respuesta $\log(Y)$, el modelo sería

$$\log(Y) = \beta_0 + \beta_1 X + \varepsilon$$

Para interpretar el coeficiente β_1 se resta esta ecuación evaluada en X_1 a su análoga evaluada en X_2 , obteniendo:

$$\begin{aligned}\log(Y_1) - \log(Y_2) &= \beta_1 X_1 - \beta_1 X_2 \\ \log\left(\frac{Y_1}{Y_2}\right) &= \beta_1 (X_1 - X_2) \\ \frac{Y_1}{Y_2} &= e^{\beta_1(X_1 - X_2)} \\ \frac{Y_1}{Y_2} &\approx (1 + \beta_1)^{(X_1 - X_2)}\end{aligned}$$

En la última igualdad se aplicó que $e^u \approx 1 + u$.

De esta forma, si $X_1 - X_2 = 1$, podemos interpretar que, por cada unidad que aumente X esperamos que Y aumente $100 \times \beta \%$.

TRANSFORMAR X

Ahora considere el modelo de *RLS* pero usando como variable explicativa $\log(X)$ en lugar de X , es decir:

$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$

TRANSFORMAR X

Ahora considere el modelo de *RLS* pero usando como variable explicativa $\log(X)$ en lugar de X , es decir:

$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$

Con el mismo procedimiento que usamos en la diapositiva anterior...

$$Y_1 - Y_2 = \beta_1 \log(X_1) - \beta_1 \log(X_2)$$

$$Y_1 - Y_2 = \beta_1 [\log(X_1) - \log(X_2)]$$

$$Y_1 - Y_2 = \beta_1 \log\left(\frac{X_1}{X_2}\right)$$

$$Y_1 - Y_2 = \beta_1 \log\left(1 + \frac{X_1 - X_2}{X_2}\right)$$

$$Y_1 - Y_2 \approx \beta_1 \frac{X_1 - X_2}{X_2}$$

$$Y_1 - Y_2 \approx \frac{\beta_1}{100} 100 \frac{X_1 - X_2}{X_2}$$

En esta ocasión, si X aumenta 1% (suponga $X_1 = 101$ y $X_2 = 100$), Y aumenta $\beta/100$ unidades.

TRANSFORMAR X Y Y

Por último, considere el modelo de *RLS* empleando la transformación lineal tanto en X como en Y .

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$$

Al comparar esta ecuación en X_1 y X_2

TRANSFORMAR X Y Y

Por último, considere el modelo de *RLS* empleando la transformación lineal tanto en X como en Y .

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \varepsilon$$

Al comparar esta ecuación en X_1 y X_2

$$\begin{aligned}\log(Y_1) - \log(Y_2) &= \beta_1 \log(X_1) - \beta_1 \log(X_2) \\ \log\left(\frac{Y_1}{Y_2}\right) &= \beta_1 [\log(X_1) - \log(X_2)] \\ \log\left(\frac{Y_1}{Y_2}\right) &= \beta_1 \log\left(\frac{X_1}{X_2}\right) \\ \frac{Y_1}{Y_2} &= e^{\beta_1 \log\left(1 + \frac{X_1 - X_2}{X_2}\right)} \\ \frac{Y_1}{Y_2} &\approx (1 + \beta_1)^{\frac{X_1 - X_2}{X_2}}\end{aligned}$$

En este caso, interpretamos que por cada unidad porcentual que aumente X , se espera que Y aumente $100 \times \beta$ %.

- **Si solo se log-transforma la Y**

Evalue el exponencial del coeficiente. El valor obtenido es el aumento PORCENTUAL esperado en la Y al aumentar una unidad la X . Si $\hat{\beta} = 0,07$, entonces $e^{\hat{\beta}} = 1,0725$, al aumentar X en una unidad, se espera que la Y aumente 7,25 %.

- **Si solo se log-transforma la Y**

Evalue el exponencial del coeficiente. El valor obtenido es el aumento PORCENTUAL esperado en la Y al aumentar una unidad la X . Si $\hat{\beta} = 0,07$, entonces $e^{\hat{\beta}} = 1,0725$, al aumentar X en una unidad, se espera que la Y aumente 7,25 %.

- **Si solo se log-transforma la X**

Divida al coeficiente entre 100. El valor obtenido es el aumento UNITARIO esperado en la Y , ante un aumento de 1% en la X . Si $\hat{\beta} = 0,07$, entonces $\hat{\beta}/100 = 0,0007$, al aumentar X en un 1%, se espera que la Y aumente 0.0007 unidades.

- **Si solo se log-transforma la Y**

Evalue el exponencial del coeficiente. El valor obtenido es el aumento PORCENTUAL esperado en la Y al aumentar una unidad la X . Si $\hat{\beta} = 0,07$, entonces $e^{\hat{\beta}} = 1,0725$, al aumentar X en una unidad, se espera que la Y aumente 7,25 %.

- **Si solo se log-transforma la X**

Divida al coeficiente entre 100. El valor obtenido es el aumento UNITARIO esperado en la Y , ante un aumento de 1% en la X . Si $\hat{\beta} = 0,07$, entonces $\hat{\beta}/100 = 0,0007$, al aumentar X en un 1%, se espera que la Y aumente 0.0007 unidades.

- **Si se log-transforma la Y y la X**

NO HAGA NADA. El valor de $\hat{\beta}$ indica el aumento porcentual esperado en la Y al aumentar la X en 1%. Si $\hat{\beta} = 0,07$, entonces al aumentar la X en 1%, se espera que la Y aumente 0.07 %.



EJEMPLO

Veamos un ejemplo donde se quiere modelar el peso del cerebro de un conjunto de mamíferos en función del peso corporal, el período de gestación y el número promedio de crías.



Vayamos al  a hacer transformaciones.

ESPECIFICACIÓN

En la primera clase de diagnóstico, vimos como a partir de los residuos parciales, es posible indagar sobre la relación de Y con cada variable explicativa de modo de diagnosticar posibles no linealidades.

ESPECIFICACIÓN

En la primera clase de diagnóstico, vimos como a partir de los residuos parciales, es posible indagar sobre la relación de Y con cada variable explicativa de modo de diagnosticar posibles no linealidades.

Esta herramienta nos sirve para descubrir posibles transformaciones de las X de modo de “linealizar” la relación.

ESPECIFICACIÓN

En la primera clase de diagnóstico, vimos como a partir de los residuos parciales, es posible indagar sobre la relación de Y con cada variable explicativa de modo de diagnosticar posibles no linealidades.

Esta herramienta nos sirve para descubrir posibles transformaciones de las X de modo de “linealizar” la relación.

Veamos un ejemplo de como poner esto en práctica.



EJEMPLO

En el siguiente ejemplo se desea modelar el precio mediano (no medio) de las casas en un conjunto de tractos censales en la ciudad de Boston.



Volvamos al , pero para hacer transformaciones.

OBSERVACIONES ATÍPICAS

A veces, la presencia de datos atípicos puede incidir en el supuesto de normalidad, incluso sin ser influyentes. Sobre todo si no se cuenta con una gran cantidad de datos.

OBSERVACIONES ATÍPICAS

A veces, la presencia de datos atípicos puede incidir en el supuesto de normalidad, incluso sin ser influyentes. Sobre todo si no se cuenta con una gran cantidad de datos.

En la clase anterior se introdujeron herramientas para identificar observaciones atípicas. También se plantearon indicadores para detectar observaciones influyentes.

Para detectar **observaciones atípicas**:

- Gráficos de caja de los residuos (studentizados externamente).
- Pruebas de hipótesis basadas en los residuos (studentizados externamente).

Para detectar **observaciones influyentes**:

- *Leverage* (h_i).
- Distancia de Cook (D_i).

OBSERVACIONES ATÍPICAS

A la hora de corregir problemas con el supuesto de normalidad (o algunas pocas veces también el de homoscedasticidad) conviene tener en cuenta que:

- Si se logra el cumplimiento del supuesto al remover una (o pocas) observaciones atípicas que NO sean influyentes, es posible llevar a cabo su eliminación detallando esto en el informe.
- Por otro lado, en caso de detectar observaciones influyentes, removerlas para lograr el cumplimiento de los supuestos puede ser más peligroso en tanto que:

OBSERVACIONES ATÍPICAS

A la hora de corregir problemas con el supuesto de normalidad (o algunas pocas veces también el de homoscedasticidad) conviene tener en cuenta que:

- Si se logra el cumplimiento del supuesto al remover una (o pocas) observaciones atípicas que NO sean influyentes, es posible llevar a cabo su eliminación detallando esto en el informe.
- Por otro lado, en caso de detectar observaciones influyentes, removerlas para lograr el cumplimiento de los supuestos puede ser más peligroso en tanto que:
 - 1 Se lo puede acusar de “*manipular los datos*” a su conveniencia.

OBSERVACIONES ATÍPICAS

A la hora de corregir problemas con el supuesto de normalidad (o algunas pocas veces también el de homoscedasticidad) conviene tener en cuenta que:

- Si se logra el cumplimiento del supuesto al remover una (o pocas) observaciones atípicas que NO sean influyentes, es posible llevar a cabo su eliminación detallando esto en el informe.
- Por otro lado, en caso de detectar observaciones influyentes, removerlas para lograr el cumplimiento de los supuestos puede ser más peligroso en tanto que:
 - ① Se lo puede acusar de "*manipular los datos*" a su conveniencia.
 - ② Se desperdicia información solo con el fin de cumplir con un supuesto, siendo que es posible realizar inferencia sin su cumplimiento.

OBSERVACIONES ATÍPICAS

A la hora de corregir problemas con el supuesto de normalidad (o algunas pocas veces también el de homoscedasticidad) conviene tener en cuenta que:

- Si se logra el cumplimiento del supuesto al remover una (o pocas) observaciones atípicas que NO sean influyentes, es posible llevar a cabo su eliminación detallando esto en el informe.
- Por otro lado, en caso de detectar observaciones influyentes, removerlas para lograr el cumplimiento de los supuestos puede ser más peligroso en tanto que:
 - ① Se lo puede acusar de "*manipular los datos*" a su conveniencia.
 - ② Se desperdicia información solo con el fin de cumplir con un supuesto, siendo que es posible realizar inferencia sin su cumplimiento.

En cualquier caso, la eliminación de observaciones anómalas debe justificarse adecuadamente y ser reportada en el análisis.



EJEMPLO

Veamos un par de ejemplos donde el supuesto de normalidad se ve perjudicado por la presencia de outliers.



Una vez más, vayamos al .

¿QUÉ HACER CUANDO TODO LO DEMÁS FALLA?

Hasta aquí vimos algunos métodos (hay más) que nos permiten lidiar con los problemas que surgen en la etapa de diagnóstico adoptando una actitud de intentar corregir estos problemas.

¿QUÉ HACER CUANDO TODO LO DEMÁS FALLA?

Hasta aquí vimos algunos métodos (hay más) que nos permiten lidiar con los problemas que surgen en la etapa de diagnóstico adoptando una actitud de intentar corregir estos problemas.

A partir de aquí, nuestra postura será emplear métodos que en lugar de corregir el problema, sean **robustos** ante la falta del cumplimiento de los supuestos.

¿QUÉ HACER CUANDO TODO LO DEMÁS FALLA?

Hasta aquí vimos algunos métodos (hay más) que nos permiten lidiar con los problemas que surgen en la etapa de diagnóstico adoptando una actitud de intentar corregir estos problemas.

A partir de aquí, nuestra postura será emplear métodos que en lugar de corregir el problema, sean **robustos** ante la falta del cumplimiento de los supuestos.

En una de las clases sobre los aspectos inferenciales, se introdujo el procedimiento de *bootstrap* para evaluar la incertidumbre de las observaciones y eventualmente construir intervalos de confianza. Hoy veremos otro método computacional que nos permite realizar inferencia en el modelo lineal sin normalidad ni homoscedasticidad.

¿QUÉ HACER CUANDO TODO LO DEMÁS FALLA?

Hasta aquí vimos algunos métodos (hay más) que nos permiten lidiar con los problemas que surgen en la etapa de diagnóstico adoptando una actitud de intentar corregir estos problemas.

A partir de aquí, nuestra postura será emplear métodos que en lugar de corregir el problema, sean **robustos** ante la falta del cumplimiento de los supuestos.

En una de las clases sobre los aspectos inferenciales, se introdujo el procedimiento de *bootstrap* para evaluar la incertidumbre de las observaciones y eventualmente construir intervalos de confianza. Hoy veremos otro método computacional que nos permite realizar inferencia en el modelo lineal sin normalidad ni homoscedasticidad.

Son las pruebas basadas en *permutaciones*.

PRUEBAS DE PERMUTACIONES

IDEA

Este método nos permite obtener la distribución del estadístico que nos interese, bajo el cumplimiento de la hipótesis nula. Esta distribución nos interesa, ya que a partir de la misma se obtiene el *p-valor* de la prueba en cuestión.

Ejemplo

Suponga que está frente al problema de comparar la media de dos grupos de observaciones independientes: W_1, W_2, \dots, W_n y Z_1, Z_2, \dots, Z_m . La prueba consiste en contrastar la siguiente hipótesis:

$$H_0: \mu_W = \mu_Z$$

$$H_1: \mu_W \neq \mu_Z$$

Para visualizar la idea clave en las pruebas basadas en permutaciones, conviene observar la matriz de datos de este problema.

PRUEBAS DE PERMUTACIONES

La formulacion que resulta conveniente es la que convierte al problema en uno propio del ámbito de los modelos lineales, más concretamente, el siguiente modelo lineal.

$$Y = \beta_0 + \beta_1 I_Z + \varepsilon$$

PRUEBAS DE PERMUTACIONES

La formulacion que resulta conveniente es la que convierte al problema en uno propio del ámbito de los modelos lineales, más concretamente, el siguiente modelo lineal.

$$Y = \beta_0 + \beta_1 I_Z + \varepsilon$$

Siendo Y un vector que agrupa las $m+n$ observaciones y I_Z una variable *dummy* que señala las posiciones donde se ubican las observaciones del grupo de variables Z .

$$\begin{bmatrix} W_1 & 0 \\ \vdots & \vdots \\ W_n & 0 \\ Z_1 & 1 \\ \vdots & \vdots \\ Z_m & 1 \end{bmatrix}$$

PRUEBAS DE PERMUTACIONES

A partir del modelo de la diapositiva anterior se obtiene un valor para el parámetro $\hat{\beta}_1$ que se puede demostrar que equivale a $\bar{W} - \bar{Z}$.

PRUEBAS DE PERMUTACIONES

A partir del modelo de la diapositiva anterior se obtiene un valor para el parámetro $\hat{\beta}_1$ que se puede demostrar que equivale a $\bar{W} - \bar{Z}$.

También se puede obtener el estadístico t asociado a la prueba:

$$H_0) \beta_1 = 0$$

$$H_1) \beta_1 \neq 0$$

Que equivale al estadístico t de la prueba $\mu_W = \mu_Z$. Pero para obtener el p-valor se necesita conocer la distribución de estos estadísticos bajo el cumplimiento de H_0 . Hasta este punto, solo podíamos acceder a este conocimiento imponiendo el supuesto de normalidad.

PRUEBAS DE PERMUTACIONES

A partir del modelo de la diapositiva anterior se obtiene un valor para el parámetro $\hat{\beta}_1$ que se puede demostrar que equivale a $\bar{W} - \bar{Z}$.

También se puede obtener el estadístico t asociado a la prueba:

$$H_0) \beta_1 = 0$$

$$H_1) \beta_1 \neq 0$$

Que equivale al estadístico t de la prueba $\mu_W = \mu_Z$. Pero para obtener el p-valor se necesita conocer la distribución de estos estadísticos bajo el cumplimiento de H_0 . Hasta este punto, solo podíamos acceder a este conocimiento imponiendo el supuesto de normalidad.

Entonces...

PRUEBAS DE PERMUTACIONES

Ahora la idea consiste en que: si H_0 es cierta, β_1 es cero, $\mu_W = \mu_Z$ y las observaciones de la segunda columna de la matriz de datos no aportan ninguna información, por lo cual, sin ningún problema las podemos desordenar y obtener un resultado igual de válido al que obtuvimos.

PRUEBAS DE PERMUTACIONES

Ahora la idea consiste en que: si H_0 es cierta, β_1 es cero, $\mu_W = \mu_Z$ y las observaciones de la segunda columna de la matriz de datos no aportan ninguna información, por lo cual, sin ningún problema las podemos desordenar y obtener un resultado igual de válido al que obtuvimos.

De esta manera, lo único que deberíamos hacer es ordenar la columna de la variable *dummy* de todas las maneras posibles, obtener el valor de $\hat{\beta}_1$ para cada caso y a través de esa colección de valores, aproximarnos a la distribución del estadístico.

PRUEBAS DE PERMUTACIONES

Ahora la idea consiste en que: si H_0 es cierta, β_1 es cero, $\mu_W = \mu_Z$ y las observaciones de la segunda columna de la matriz de datos no aportan ninguna información, por lo cual, sin ningún problema las podemos desordenar y obtener un resultado igual de válido al que obtuvimos.

De esta manera, lo único que deberíamos hacer es ordenar la columna de la variable *dummy* de todas las maneras posibles, obtener el valor de $\hat{\beta}_1$ para cada caso y a través de esa colección de valores, aproximarnos a la distribución del estadístico.

Peeeeeeeero ...

PRUEBAS DE PERMUTACIONES

Ahora la idea consiste en que: si H_0 es cierta, β_1 es cero, $\mu_W = \mu_Z$ y las observaciones de la segunda columna de la matriz de datos no aportan ninguna información, por lo cual, sin ningún problema las podemos desordenar y obtener un resultado igual de válido al que obtuvimos.

De esta manera, lo único que deberíamos hacer es ordenar la columna de la variable *dummy* de todas las maneras posibles, obtener el valor de $\hat{\beta}_1$ para cada caso y a través de esa colección de valores, aproximarnos a la distribución del estadístico.

Peeeeeeeero ... si tenemos $n+m$ observaciones, habría que evaluar $(n+m)!$ permutaciones. Que con $n=m=5$ ya es más de 3 millones.

PRUEBAS DE PERMUTACIONES

Ahora la idea consiste en que: si H_0 es cierta, β_1 es cero, $\mu_W = \mu_Z$ y las observaciones de la segunda columna de la matriz de datos no aportan ninguna información, por lo cual, sin ningún problema las podemos desordenar y obtener un resultado igual de válido al que obtuvimos.

De esta manera, lo único que deberíamos hacer es ordenar la columna de la variable *dummy* de todas las maneras posibles, obtener el valor de $\hat{\beta}_1$ para cada caso y a través de esa colección de valores, aproximarnos a la distribución del estadístico.

Peeeeeeeero ... si tenemos $n+m$ observaciones, habría que evaluar $(n+m)!$ permutaciones. Que con $n=m=5$ ya es más de 3 millones.

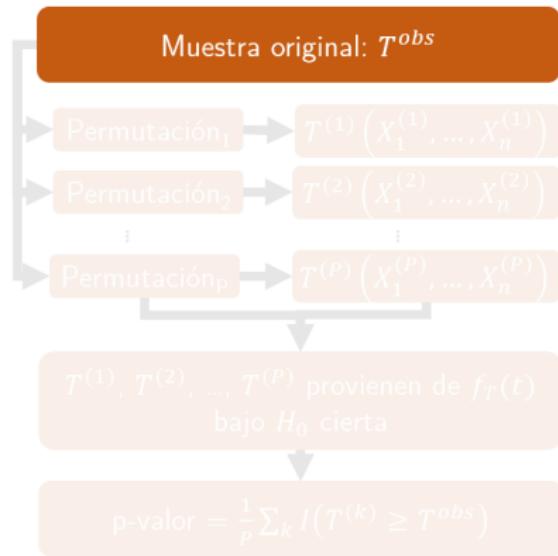
Solución

En la práctica se evalúa una muestra (tan grande como querremos) de todas esas posibles permutaciones. Estas pruebas son conocidas como *randomization tests*.



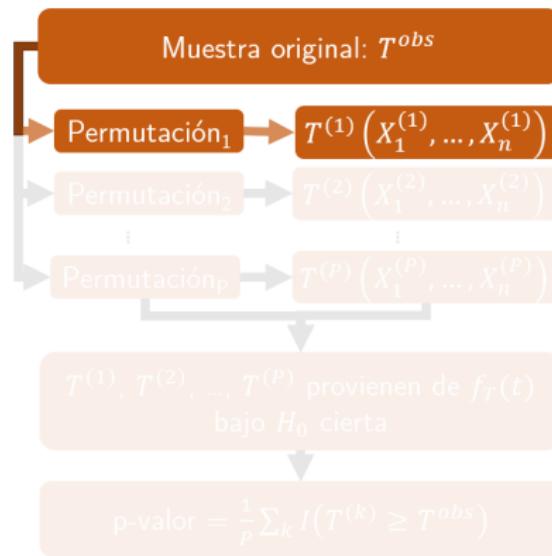
PRUEBAS DE PERMUTACIONES

El procedimiento tiene alguna similaridad al *bootstrap* en la medida en que ambos son computacionalmente intensivos, no requieren supuestos distribucionales (más allá de la independencia de las observaciones) y hacen uso de la aleatoriedad.



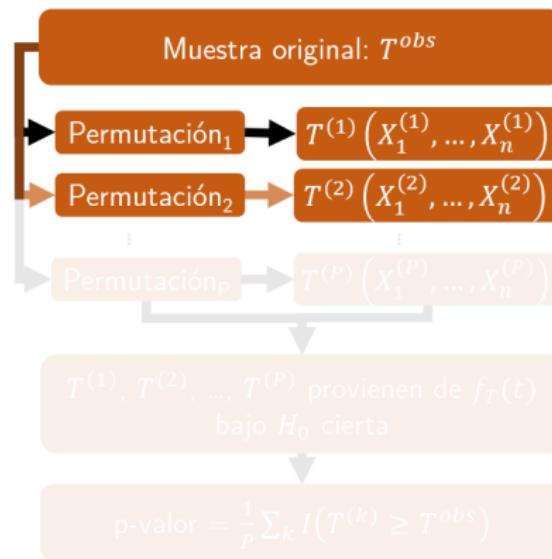
PRUEBAS DE PERMUTACIONES

En la práctica el procedimiento tiene alguna similaridad al *bootstrap* en la medida en que ambos son computacionalmente intensivos, no requieren supuestos distribucionales (más allá de la independencia de las observaciones) y hacen uso de la aleatoriedad.



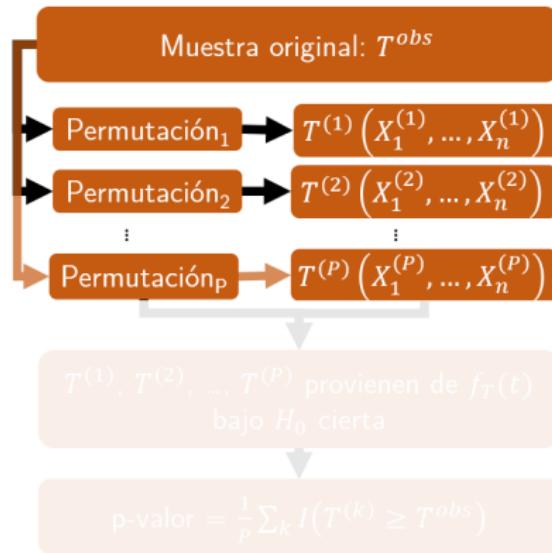
PRUEBAS DE PERMUTACIONES

En la práctica el procedimiento tiene alguna similaridad al *bootstrap* en la medida en que ambos son computacionalmente intensivos, no requieren supuestos distribucionales (más allá de la independencia de las observaciones) y hacen uso de la aleatoriedad.



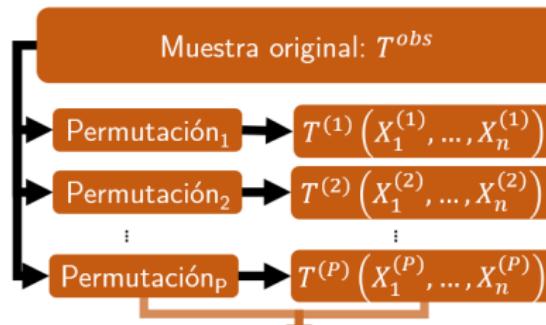
PRUEBAS DE PERMUTACIONES

En la práctica el procedimiento tiene alguna similaridad al *bootstrap* en la medida en que ambos son computacionalmente intensivos, no requieren supuestos distribucionales (más allá de la independencia de las observaciones) y hacen uso de la aleatoriedad.



PRUEBAS DE PERMUTACIONES

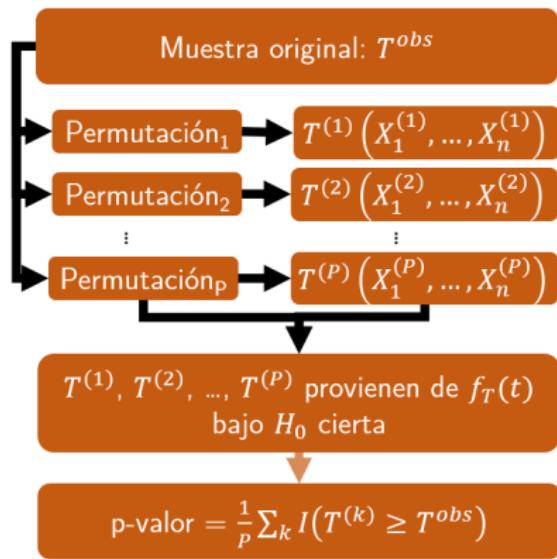
En la práctica el procedimiento tiene alguna similaridad al *bootstrap* en la medida en que ambos son computacionalmente intensivos, no requieren supuestos distribucionales (más allá de la independencia de las observaciones) y hacen uso de la aleatoriedad.



$$p\text{-valor} = \frac{1}{P} \sum_{k=1}^P I(T^{(k)} \geq T^{obs})$$

PRUEBAS DE PERMUTACIONES

En la práctica el procedimiento tiene alguna similaridad al *bootstrap* en la medida en que ambos son computacionalmente intensivos, no requieren supuestos distribucionales (más allá de la independencia de las observaciones) y hacen uso de la aleatoriedad.



PRUEBAS DE PERMUTACIONES - MODELO *RLM*

En el modelo de *RLS* el procedimiento es el que vimos en el ejemplo. No obstante, en el modelo de *RLM* este procedimiento no se aplica de manera tan inmediata.

PRUEBAS DE PERMUTACIONES - MODELO *RLM*

En el modelo de *RLS* el procedimiento es el que vimos en el ejemplo. No obstante, en el modelo de *RLM* este procedimiento no se aplica de manera tan inmediata.

Nótese que para evaluar la significación de cada variable, deben llevarse a cabo k muestras de permutaciones.

PRUEBAS DE PERMUTACIONES - MODELO *RLM*

En el modelo de *RLS* el procedimiento es el que vimos en el ejemplo. No obstante, en el modelo de *RLM* este procedimiento no se aplica de manera tan inmediata.

Nótese que para evaluar la significación de cada variable, deben llevarse a cabo k muestras de permutaciones. Además, permutar los valores de cada x_j por separado no es correcto ya que de esta manera se pierde la posible incidencia de la correlación entre las variables explicativas.

PRUEBAS DE PERMUTACIONES - MODELO *RLM*

En el modelo de *RLS* el procedimiento es el que vimos en el ejemplo. No obstante, en el modelo de *RLM* este procedimiento no se aplica de manera tan inmediata.

Nótese que para evaluar la significación de cada variable, deben llevarse a cabo k muestras de permutaciones. Además, permutar los valores de cada x_j por separado no es correcto ya que de esta manera se pierde la posible incidencia de la correlación entre las variables explicativas.

Solución

Transformar el problema de *RLM* en k problemas de *RLS* utilizando la matriz $I_{n \times n} - H_{(j)}$.

PRUEBAS DE PERMUTACIONES - MODELO *RLM*

Supongamos que queremos evaluar la significancia de la variable X_j en un modelo donde contamos con otro conjunto de variables explicativas.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

Al premultiplicar a Y y a X_j por $I_{n \times n} - H_{(j)}$ obtenemos el siguiente modelo de *RLS*

$$Y^* = \beta_j X_j^* + \varepsilon^*$$

Siendo Y^* y X_j^* las versiones de dichas variables *libres* del efecto de las demás variables.

PRUEBAS DE PERMUTACIONES - MODELO *RLM*

Supongamos que queremos evaluar la significancia de la variable X_j en un modelo donde contamos con otro conjunto de variables explicativas.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

Al premultiplicar a Y y a X_j por $I_{n \times n} - H_{(j)}$ obtenemos el siguiente modelo de *RLS*

$$Y^* = \beta_j X_j^* + \varepsilon^*$$

Siendo Y^* y X_j^* las versiones de dichas variables *libres* del efecto de las demás variables.

Y en este problema de *RLS* sí podemos utilizar el procedimiento descrito anteriormente sin problemas.



EJEMPLO

Retomemos algunos ejemplos empleando este tipo de pruebas.



Volvamos al  por última vez.

OTROS MÉTODOS

En la clase de hoy vimos algunas alternativas para lidiar ante el no cumplimiento de los supuestos. Sin embargo esta lista no es exhaustiva.

OTROS MÉTODOS

En la clase de hoy vimos algunas alternativas para lidiar ante el no cumplimiento de los supuestos. Sin embargo esta lista no es exhaustiva.

En solo una clase no es posible ofrecer un conjunto de métodos adecuados a cada situación particular (ni siquiera en un curso).

OTROS MÉTODOS

En la clase de hoy vimos algunas alternativas para lidiar ante el no cumplimiento de los supuestos. Sin embargo esta lista no es exhaustiva.

En solo una clase no es posible ofrecer un conjunto de métodos adecuados a cada situación particular (ni siquiera en un curso).

Otras alternativas a explorar pueden ser:

- Transformación de Box-Cox

OTROS MÉTODOS

En la clase de hoy vimos algunas alternativas para lidiar ante el no cumplimiento de los supuestos. Sin embargo esta lista no es exhaustiva.

En solo una clase no es posible ofrecer un conjunto de métodos adecuados a cada situación particular (ni siquiera en un curso).

Otras alternativas a explorar pueden ser:

- Transformación de Box-Cox
- Regresión robusta

OTROS MÉTODOS

En la clase de hoy vimos algunas alternativas para lidiar ante el no cumplimiento de los supuestos. Sin embargo esta lista no es exhaustiva.

En solo una clase no es posible ofrecer un conjunto de métodos adecuados a cada situación particular (ni siquiera en un curso).

Otras alternativas a explorar pueden ser:

- Transformación de Box-Cox
- Regresión robusta
- Regresión cuantil

OTROS MÉTODOS

En la clase de hoy vimos algunas alternativas para lidiar ante el no cumplimiento de los supuestos. Sin embargo esta lista no es exhaustiva.

En solo una clase no es posible ofrecer un conjunto de métodos adecuados a cada situación particular (ni siquiera en un curso).

Otras alternativas a explorar pueden ser:

- Transformación de Box-Cox
- Regresión robusta
- Regresión cuantil
- GAMLSS

OTROS MÉTODOS

En la clase de hoy vimos algunas alternativas para lidiar ante el no cumplimiento de los supuestos. Sin embargo esta lista no es exhaustiva.

En solo una clase no es posible ofrecer un conjunto de métodos adecuados a cada situación particular (ni siquiera en un curso).

Otras alternativas a explorar pueden ser:

- Transformación de Box-Cox
- Regresión robusta
- Regresión cuantil
- GAMLSS
- ...



EN LA PRÓXIMA CLASE

La próxima:

- Puede que tengamos un taller en el que veamos un ejemplo completo de diagnóstico.
- O puede que comencemos con los modelos de Análisis de Varianza (ANOVA).



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