

MODELOS LINEALES

DIAGNÓSTICO

Fernando Massa; Bruno Bellagamba

25 de mayo 2024



FACULTAD DE
CIENCIAS ECONÓMICAS
Y DE ADMINISTRACIÓN

UESTA INSTITUTO
DE ESTADÍSTICA



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



1 INTRODUCCIÓN

2 MULTICOLINEALIDAD

3 LINEALIDAD

4 PRÓXIMA CLASE



- Repasaremos los supuestos del modelo de *RLM* y discutiremos sobre la relevancia de cada uno de ellos.
- Comenzaremos evaluando el efecto de la multicolinealidad de los vectores que componen la matriz X .
- Continuaremos la etapa de diagnóstico a través de análisis gráficos sobre los residuos del modelo.
- Nos enfocaremos sobre los gráficos que vinculan \hat{y} , $\hat{\varepsilon}$ y cada x_j .

En las primeras clases del curso realizamos un conjunto de supuestos sobre el modelo

$$Y = X\beta + \varepsilon$$

La optimalidad del vector de estimaciones $\hat{\beta}$ (en el sentido de Gauss-Markov) dependía de un conjunto de supuestos realizados sobre los errores del modelo.

- $\mathbb{E}(\varepsilon) = 0$.
- $\mathbb{V}ar(\varepsilon) = \sigma^2$.
- Independencia.

En las primeras clases del curso realizamos un conjunto de supuestos sobre el modelo

$$Y = X\beta + \varepsilon$$

La optimalidad del vector de estimaciones $\hat{\beta}$ (en el sentido de Gauss-Markov) dependía de un conjunto de supuestos realizados sobre los errores del modelo.

- $\mathbb{E}(\varepsilon) = 0$.
- $\mathbb{V}ar(\varepsilon) = \sigma^2$.
- Independencia.

Adicionalmente hicimos el supuesto de que el rango de X era completo, lo cual nos aseguraba la existencia de $(X'X)^{-1}$. El otro supuesto subyacente al modelo es que la relación entre Y y las x_j es lineal.

Luego se introdujo el supuesto de normalidad que (junto con los anteriores) nos permitió realizar inferencia.

- Estimaciones máximo verosímiles.
- Pruebas de hipótesis.
- Intervalos de confianza.

Luego se introdujo el supuesto de normalidad que (junto con los anteriores) nos permitió realizar inferencia.

- Estimaciones máximo verosímiles.
- Pruebas de hipótesis.
- Intervalos de confianza.

Tener en cuenta este conjunto de supuestos es sumamente relevante ya que sin los mismos, las inferencias realizadas **NO** son válidas. Esto se traduce en que:

- Las pruebas de hipótesis no tienen el nivel nominal ni la potencia esperada.
- Los intervalos de confianza no tienen la cobertura nominal.

Luego se introdujo el supuesto de normalidad que (junto con los anteriores) nos permitió realizar inferencia.

- Estimaciones máximo verosímiles.
- Pruebas de hipótesis.
- Intervalos de confianza.

Tener en cuenta este conjunto de supuestos es sumamente relevante ya que sin los mismos, las inferencias realizadas **NO** son válidas. Esto se traduce en que:

- Las pruebas de hipótesis no tienen el nivel nominal ni la potencia esperada.
- Los intervalos de confianza no tienen la cobertura nominal.

Por este motivo, resulta de especial interés realizar una etapa de diagnóstico donde se ponga a prueba el cumplimiento de los supuestos realizados.

El primer supuesto a tener en cuenta es el que se realiza sobre $rg(X)$.

El primer supuesto a tener en cuenta es el que se realiza sobre $rg(X)$.

Este difiere de los demás en tanto que su diagnóstico no depende de Y y por ende, tampoco depende de $\hat{\varepsilon}$.

El primer supuesto a tener en cuenta es el que se realiza sobre $rg(X)$.

Este difiere de los demás en tanto que su diagnóstico no depende de Y y por ende, tampoco depende de $\hat{\varepsilon}$.

MULTICOLINEALIDAD

Nos referimos por multicolinealidad al hecho de que una o más de las columnas de la matriz X pueda obtenerse mediante una combinación lineal (CL) de otras columnas de X .


MULTICOLINEALIDAD EXACTA

El primer supuesto a tener en cuenta es el que se realiza sobre $rg(X)$.

Este difiere de los demás en tanto que su diagnóstico no depende de Y y por ende, tampoco depende de $\hat{\epsilon}$.

MULTICOLINEALIDAD

Nos referimos por multicolinealidad al hecho de que una o más de las columnas de la matriz X pueda obtenerse mediante una combinación lineal (CL) de otras columnas de X .


En la práctica esto rara vez sucede y, de darse, el  advierte la presencia de este problema y remueve la variable en cuestión.

El primer supuesto a tener en cuenta es el que se realiza sobre $rg(X)$.

Este difiere de los demás en tanto que su diagnóstico no depende de Y y por ende, tampoco depende de $\hat{\epsilon}$.

MULTICOLINEALIDAD

Nos referimos por multicolinealidad al hecho de que una o más de las columnas de la matriz X pueda obtenerse mediante una combinación lineal (CL) de otras columnas de X .

En la práctica esto rara vez sucede y, de darse, el  advierte la presencia de este problema y remueve la variable en cuestión.

El problema se torna más *interesante* cuando alguna(s) variable es “*casi*” combinación lineal de otras.

Los problemas surgen cuando una cierta variable explicativa esté cerca de ser obtenida mediante una combinación lineal de las demás variables.

$$x_j = a_0 + a_1x_1 + \dots + a_{j-1}x_{j-1} + a_{j+1}x_{j+1} + \dots + a_kx_k + \eta$$

Esta formulación, ya sugiere una forma de diagnosticar el problema...

Los problemas surgen cuando una cierta variable explicativa esté cerca de ser obtenida mediante una combinación lineal de las demás variables.

$$x_j = a_0 + a_1x_1 + \dots + a_{j-1}x_{j-1} + a_{j+1}x_{j+1} + \dots + a_kx_k + \eta$$

Esta formulación, ya sugiere una forma de diagnosticar el problema...

El detalle a tener en cuenta es que entre los coeficientes a_1, a_2, \dots, a_k puede o no haber ceros.

Los problemas surgen cuando una cierta variable explicativa esté cerca de ser obtenida mediante una combinación lineal de las demás variables.

$$x_j = a_0 + a_1x_1 + \dots + a_{j-1}x_{j-1} + a_{j+1}x_{j+1} + \dots + a_kx_k + \eta$$

Esta formulación, ya sugiere una forma de diagnosticar el problema...

El detalle a tener en cuenta es que entre los coeficientes a_1, a_2, \dots, a_k puede o no haber ceros.

EL PROBLEMA

Si al menos 1 de los coeficientes es distinto de cero y $\text{Var}(\eta)$ es “cercana” a 0, entonces x_j es “casi” CL de las demás variables. Esto hace que el número de condición de $X'X$ aumente dramáticamente, haciendo que su inversa se vuelva inestable

Cuando hablamos de una “CL de las demás variables”, incluimos situaciones donde varios coeficientes sean cero. Incluso, podemos pensar en casos donde todos los coeficientes son 0, excepto 1.

Cuando hablamos de una “CL de las demás variables”, incluimos situaciones donde varios coeficientes sean cero. Incluso, podemos pensar en casos donde todos los coeficientes son 0, excepto 1.

Este último caso, es más fácil de diagnosticar, ya que se trata de situaciones donde existe alguna variable explicativa que está muy correlacionada con otra. Esto podemos diagnosticarlo en las etapas iniciales del modelado, cuando realizamos una descripción inicial de los datos mediante la matriz de correlaciones.

Cuando hablamos de una “CL de las demás variables”, incluimos situaciones donde varios coeficientes sean cero. Incluso, podemos pensar en casos donde todos los coeficientes son 0, excepto 1.

Este último caso, es más fácil de diagnosticar, ya que se trata de situaciones donde existe alguna variable explicativa que está muy correlacionada con otra. Esto podemos diagnosticarlo en las etapas iniciales del modelado, cuando realizamos una descripción inicial de los datos mediante la matriz de correlaciones.

Pero, ¿cómo diagnosticamos estos casos en las situaciones más generales?

Una primera aproximación consiste en explorar los valores propios de $X'X$.

DIAGNÓSTICO DE LA M.A.

Una primera aproximación consiste en explorar los valores propios de $X'X$.

Esto se debe a que si $\lambda_j > 0 \quad \forall j = 1, \dots, k$, entonces existe la inversa.

Una primera aproximación consiste en explorar los valores propios de $X'X$.

Esto se debe a que si $\lambda_j > 0 \quad \forall j = 1, \dots, k$, entonces existe la inversa.

Sin embargo, esto solo diagnosticaría la multicolinealidad exacta. Para detectar la presencia de multicolinealidad aproximada debemos prestar atención al valor propio más pequeño.

DIAGNÓSTICO DE LA M.A.

Una primera aproximación consiste en explorar los valores propios de $X'X$.

Esto se debe a que si $\lambda_j > 0 \quad \forall j = 1, \dots, k$, entonces existe la inversa.

Sin embargo, esto solo diagnosticaría la multicolinealidad exacta. Para detectar la presencia de multicolinealidad aproximada debemos prestar atención al valor propio más pequeño.

NÚMERO DE CONDICIÓN

Una manera de aproximar el número de condición de una matriz A usando sus valores propios es:

$$\kappa(A) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

DIAGNÓSTICO DE LA M.A.

Una primera aproximación consiste en explorar los valores propios de $X'X$.

Esto se debe a que si $\lambda_j > 0 \quad \forall j = 1, \dots, k$, entonces existe la inversa.

Sin embargo, esto solo diagnosticaría la multicolinealidad exacta. Para detectar la presencia de multicolinealidad aproximada debemos prestar atención al valor propio más pequeño.

NÚMERO DE CONDICIÓN

Una manera de aproximar el número de condición de una matriz A usando sus valores propios es:

$$\kappa(A) = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

Una posible regla es que $\kappa(A) > 30$ suele indicar problemas de multicolinealidad.

Una segunda aproximación consiste en calcular el determinante de $X'X$.

Una segunda aproximación consiste en calcular el determinante de $X'X$.

Debido a que $|X'X| = \prod_{j=1}^p \lambda_j$, si alguno de los valores propios es “casí” cero, entonces la matriz es singular y su inversa es inestable.

Una segunda aproximación consiste en calcular el determinante de $X'X$.

Debido a que $|X'X| = \prod_{j=1}^p \lambda_j$, si alguno de los valores propios es “casí” cero, entonces la matriz es singular y su inversa es inestable.

Sin embargo, es difícil saber cuando el determinante está cerca de cero, una alternativa para evitar este problema es trabajar con el determinante de $\tilde{X}'\tilde{X}$. Siendo \tilde{X} la matriz de regresores (sin la constante) estandarizados.

Una segunda aproximación consiste en calcular el determinante de $X'X$.

Debido a que $|X'X| = \prod_{j=1}^p \lambda_j$, si alguno de los valores propios es “casi” cero, entonces la matriz es singular y su inversa es inestable.

Sin embargo, es difícil saber cuando el determinante está cerca de cero, una alternativa para evitar este problema es trabajar con el determinante de $\tilde{X}'\tilde{X}$. Siendo \tilde{X} la matriz de regresores (sin la constante) estandarizados.

La virtud de esta alternativa consiste en que:

$$0 \leq \left| \frac{\tilde{X}'\tilde{X}}{n-1} \right| \leq 1$$

Una segunda aproximación consiste en calcular el determinante de $X'X$.

Debido a que $|X'X| = \prod_{j=1}^p \lambda_j$, si alguno de los valores propios es “*casí*” cero, entonces la matriz es singular y su inversa es inestable.

Sin embargo, es difícil saber cuando el determinante está cerca de cero, una alternativa para evitar este problema es trabajar con el determinante de $\tilde{X}'\tilde{X}$. Siendo \tilde{X} la matriz de regresores (sin la constante) estandarizados.

La virtud de esta alternativa consiste en que:

$$0 \leq \left| \frac{\tilde{X}'\tilde{X}}{n-1} \right| \leq 1$$

Si bien no existe un valor que indique “*cercanía*” al cero, al contar con una cota superior, es más sencillo tomar una decisión acerca de la singularidad de $X'X$.

Luego de detectar la presencia de la multicolinealidad exacta surge el problema de detectar cual (o cuales) de las x_j es la responsable del problema.

Luego de detectar la presencia de la multicolinealidad exacta surge el problema de detectar cual (o cuales) de las x_j es la responsable del problema.

Para esto, una posible solución tiene que ver con el planteamiento de la expresión de x_j como una función lineal de las demás variables explicativas.

Luego de detectar la presencia de la multicolinealidad exacta surge el problema de detectar cual (o cuales) de las x_j es la responsable del problema.

Para esto, una posible solución tiene que ver con el planteamiento de la expresión de x_j como una función lineal de las demás variables explicativas.

VIF (VARIANCE INFLATION FACTOR)

De esta manera, el procedimiento a seguir se basa en la idea de que si x_j es casi CL de las demás, el R^2 de una regresión donde x_j es la variable de respuesta debería ser *alto*.

De esta manera, cada x_j tendrá asociado su VIF_j obtenido como:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Luego de detectar la presencia de la multicolinealidad exacta surge el problema de detectar cual (o cuales) de las x_j es la responsable del problema.

Para esto, una posible solución tiene que ver con el planteamiento de la expresión de x_j como una función lineal de las demás variables explicativas.

VIF (VARIANCE INFLATION FACTOR)

De esta manera, el procedimiento a seguir se basa en la idea de que si x_j es casi CL de las demás, el R^2 de una regresión donde x_j es la variable de respuesta debería ser *alto*.

De esta manera, cada x_j tendrá asociado su VIF_j obtenido como:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Típicamente, se suele considerar a x_j como “culpable” si $VIF_j > 5$ (o 10, depende a quién le pregunte).

El nombre de este indicador deriva de la siguiente expresión:

$$\mathbb{V}ar(\hat{\beta}_j) = \sigma^2 \frac{1}{1 - R_j^2} \frac{1}{\sum_i (x_{ij} - \bar{x}_j)^2}$$

El nombre de este indicador deriva de la siguiente expresión:

$$\mathbb{V}ar(\hat{\beta}_j) = \sigma^2 \frac{1}{1 - R_j^2} \frac{1}{\sum_i (x_{ij} - \bar{x}_j)^2}$$

Se puede observar cómo, a mayor valor del *VIF*, más se incrementa la varianza del estimador del coeficiente correspondiente a la variable x_j .

El nombre de este indicador deriva de la siguiente expresión:

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \frac{1}{1 - R_j^2} \frac{1}{\sum_i (x_{ij} - \bar{x}_j)^2}$$

Se puede observar cómo, a mayor valor del *VIF*, más se incrementa la varianza del estimador del coeficiente correspondiente a la variable x_j .

En la situación óptima, si $R_j^2 = 0$, x_j es ortogonal a las demás variables explicativas y $\text{Var}(\hat{\beta}_j)$ es inversamente proporcional a la variabilidad de x_j .

Una vez detectada la (o las) culpable de la existencia de multicolinealidad aproximada existen (al menos) 2 alternativas.


- Quitar la variable del análisis.
- Emplear los métodos de regularización vistos en la clase anterior.

Veamos estos conceptos con algunos ejemplos en datos simulados y en el conjunto de datos de los autos del 74.



Veamos estos conceptos con algunos ejemplos en datos simulados y en el conjunto de datos de los autos del 74.



Vayamos al 

El siguiente supuesto a tener en cuenta es uno que por lo general es olvidado y dado por sentado, se trata de la **linealidad** entre Y y las variables contenidas en X .

El siguiente supuesto a tener en cuenta es uno que por lo general es olvidado y dado por sentado, se trata de la **linealidad** entre Y y las variables contenidas en X .

La consecuencia del incumplimiento de este supuesto puede tener varias manifestaciones.

- Falta de ajuste en el modelo.
- Problemas de correlación entre los residuos.
- Problemas en la variabilidad de los residuos.

El siguiente supuesto a tener en cuenta es uno que por lo general es olvidado y dado por sentado, se trata de la **linealidad** entre Y y las variables contenidas en X .

La consecuencia del incumplimiento de este supuesto puede tener varias manifestaciones.

- Falta de ajuste en el modelo.
- Problemas de correlación entre los residuos.
- Problemas en la variabilidad de los residuos.

Es importante que contemos con recursos para diagnosticar este problema.

Típicamente, la manera de diagnosticar esta situación es mediante el análisis de los residuos del modelo, lo cual (a diferencia del supuesto de multicolinealidad) supone que el modelo ya está ajustado.

Típicamente, la manera de diagnosticar esta situación es mediante el análisis de los residuos del modelo, lo cual (a diferencia del supuesto de multicolinealidad) supone que el modelo ya está ajustado.

El primer gráfico a analizar es el que relaciona \hat{Y} con $\hat{\epsilon}$.

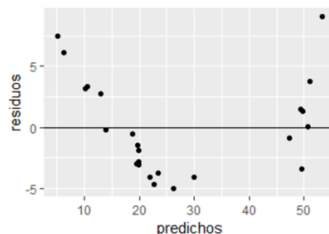
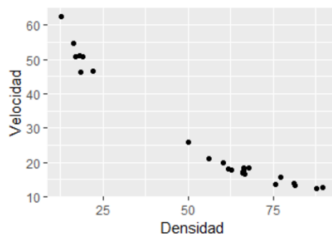
En caso de que NO existan problemas por falta de linealidad, esperamos que en este gráfico no se observe ningún patrón.

PREDICHOS VS RESIDUOS

En este ejemplo se presenta el análisis de un modelo RLS donde se explica la velocidad de los automóviles en una autopista, en función de la densidad de vehículos por kilómetro.

PREDICHOS VS RESIDUOS

En este ejemplo se presenta el análisis de un modelo *RLS* donde se explica la velocidad de los automóviles en una autopista, en función de la densidad de vehículos por kilómetro.



El modelo ajustado fue:

$$Velocidad_i = \beta_0 + \beta_1 Densidad_i + \varepsilon_i$$

El caso anterior, es un bueno ejemplo de la forma de proceder para corregir problemas de ajuste en un modelo de *RLS*, pero en un modelo múltiple no nos dice cual (o cuales) de las variables explicativas fueron especificadas incorrectamente.

RESIDUOS PARCIALES

El caso anterior, es un bueno ejemplo de la forma de proceder para corregir problemas de ajuste en un modelo de *RLS*, pero en un modelo múltiple no nos dice cual (o cuales) de las variables explicativas fueron especificadas incorrectamente.

Una forma de realizar el análisis *variable a variable* es a través de los gráficos de residuos parciales.

El caso anterior, es un buen ejemplo de la forma de proceder para corregir problemas de ajuste en un modelo de *RLS*, pero en un modelo múltiple no nos dice cual (o cuales) de las variables explicativas fueron especificadas incorrectamente.

Una forma de realizar el análisis *variable a variable* es a través de los gráficos de residuos parciales.

Supongamos que se desea explicar una cierta variable Y a partir de X_1 y X_2 , pero donde la especificación correcta es:

$$y_i = \beta_0 + \beta_1 x_{i1} + g(x_{i2}) + \varepsilon_i$$

Siendo $g(\cdot)$ alguna función no lineal.

El caso anterior, es un buen ejemplo de la forma de proceder para corregir problemas de ajuste en un modelo de *RLS*, pero en un modelo múltiple no nos dice cual (o cuales) de las variables explicativas fueron especificadas incorrectamente.

Una forma de realizar el análisis *variable a variable* es a través de los gráficos de residuos parciales.

Supongamos que se desea explicar una cierta variable Y a partir de X_1 y X_2 , pero donde la especificación correcta es:

$$y_i = \beta_0 + \beta_1 x_{i1} + g(x_{i2}) + \varepsilon_i$$

Siendo $g(\cdot)$ alguna función no lineal.

El problema surge, al ajustar un modelo de la forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \eta_i$$

Supongamos ahora que se estiman los coeficientes del modelo incorrecto, es decir:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

Supongamos ahora que se estiman los coeficientes del modelo incorrecto, es decir:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

Luego, los residuos se definen como:

$$\hat{\eta}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}$$

Supongamos ahora que se estiman los coeficientes del modelo incorrecto, es decir:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

Luego, los residuos se definen como:

$$\hat{\eta}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}$$

Si sustituimos y_i por su definición en el modelo correcto:

$$\begin{aligned}\hat{\eta}_i &= \beta_0 + \beta_1 x_{i1} + g(x_{i2}) + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \\ \underbrace{\hat{\eta}_i + \hat{\beta}_2 x_{i2}}_{\text{residuo parcial}} &= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_{i1} + g(x_{i2}) + \varepsilon_i \\ \hat{\eta}_i + \hat{\beta}_2 x_{i2} &\approx g(x_{i2})\end{aligned}$$

Supongamos ahora que se estiman los coeficientes del modelo incorrecto, es decir:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

Luego, los residuos se definen como:

$$\hat{\eta}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}$$

Si sustituimos y_i por su definición en el modelo correcto:

$$\begin{aligned}\hat{\eta}_i &= \beta_0 + \beta_1 x_{i1} + g(x_{i2}) + \varepsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} \\ \underbrace{\hat{\eta}_i + \hat{\beta}_2 x_{i2}}_{\text{residuo parcial}} &= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_{i1} + g(x_{i2}) + \varepsilon_i \\ \hat{\eta}_i + \hat{\beta}_2 x_{i2} &\approx g(x_{i2})\end{aligned}$$

Si las estimaciones de los coeficientes asociados al resto de las variables estuvieron


“cerca” de los verdaderos valores, entonces el residuo parcial debería asemejarse a $g(x_{i2})$.

Veamos ahora un ejemplo donde se quiere determinar el impacto del salario y la educación sobre el “*prestigio*” de un conjunto de profesiones



Veamos ahora un ejemplo donde se quiere determinar el impacto del salario y la educación sobre el “*prestigio*” de un conjunto de profesiones



Vayamos al 



La próxima hablaremos de:

- Continuaremos la etapa de diagnóstico del modelo.
- Nos adentraremos sobre el análisis formal de los supuestos de normalidad y homoscedasticidad.
- Veremos distintas alternativas a tener en cuenta a la hora de lidiar con el NO cumplimiento de estos supuestos.
- Y (si nos da el tiempo) comenzaremos a charlar sobre el efecto de observaciones atípicas y/o influyentes.



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