TEMA 2: EXTENSIONES DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE: MULTICOLINEALIDAD Y TRANSFORMACIONES LINEALES.

Wooldridge: Capítulos 6 (apartado 6.1) y 7

Gujarati: Capítulos 9 (apartado 9.8), 10 y 12

1. MULTICOLINEALIDAD

- **En el contexto del MLG**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K + \varepsilon$$

con los supuestos clásicos, uno de ellos es que ninguna de las variables X_1, X_2, …, X_K es constante ni existe una relación lineal exacta entre ellas → <u>Ausencia de multicolinealidad exacta.</u>

Econometría. Tema 2. Curso 2015-16. Pilar Abad

- **Recuérdese que si existe multicolinealidad exacta el sistema de ecuaciones normales tiene infinitas soluciones aunque se pueden estimar de forma única combinaciones lineales de parámetros.**

- **Vamos a analizar lo que ocurre cuando NO HAY MULTICOLINEALIDAD EXACTA pero sí que HAY MULTICOLINEALIDAD APROXIMADA.**

  <u>Multicolinealidad aproximada:</u> El problema es que una o varias de las variables X_1, X_2, …, X_K es "casi" una constante o existe una "casi" relación lineal entre ellas → No es ningún incumplimiento de los supuestos clásicos, lo que supone que todos los resultados vistos para el estimador MCO e inferencia son válidos.

● **Sin embargo un "elevado grado" de colinealidad aproximada tiene consecuencias empíricas:**

1ª) Pequeñas variaciones muestrales pueden suponer grandes variaciones en las estimaciones MCO.

2ª) La varianza del estimador MCO de las variables X´s colineales ($V(\hat{\beta}_j)$) aumenta, con lo cuál la inferencia sobre ese parámetro se ve afectada → La estimación de ese parámetro es muy "imprecisa" (intervalo de confianza amplio).

● **Veamos este último punto.**

En el contexto del MLG y del estimador MCO se cumple que:

$$V(\hat{\beta}_j \mid X) = \frac{\sigma^2}{\left(1 - R_j^2\right) n S_{X_j}^2}$$

en donde $V(Y / X_1, X_2, ..., X_k) = \sigma^2$ **y** R_j^2 **es el R**^{2} **de la regresión de** X_j **sobre el resto de variables explicativas incluyendo el término constante.**

● **Si** R_j^2 **está próximo a la unidad:**

$\hat{V}(\hat{\beta}_j) = \dfrac{\hat{\sigma}^2}{\left(1 - R_j^2\right) n S_{X_j}^2}$ **será "grande" →** $\hat{\beta}_j \pm z_{\alpha/2} \times s(\hat{\beta}_j)$ **será amplio, reflejando que hay mucha incertidumbre sobre cuál puede ser el verdadero valor de** β_j**.**

- **Lo que ocurre tiene una interpretación intuitiva clara → Como varios regresores mantienen una relación lineal estrecha, es muy difícil "aislar" el efecto que sobre la variable a explicar tienen cada uno de esos regresores.**

- **La MULTICOLINEALIDAD APROXIMADA se puede detectar mediante el análisis de la posible existencia de relaciones lineales entre los regresores.**

- **La presencia de altas correlaciones entre las variables explicativas es una señal de multicolinealidad. Pero es posible que exista una relación lineal muy alta entre varias variables y los coeficientes de correlación simple sean bajos.**

Econometría. Tema 2. Curso 2015-16. Pilar Abad

- **La multicolinealidad puede afectar mucho a unos parámetros y nada a otros. Los parámetros asociados a variables independientes o con poca relación lineal con el resto, no se ven afectados.**

- **La alta multicolinealidad no afecta a la predicción.**

- **Una característica típica de la multicolinealidad es encontrar que los coeficientes del modelo de regresión no sean individualmente significativamente distintos de cero (por las altas varianzas) pero sí son significativamente distintos de cero de forma conjunta.**

- **La MULTICOLINEALIDAD APROXIMADA es muy difícil de resolver en las Ciencias Sociales ya que no se puede experimentar →**

  Muchos autores sugieren eliminar variables colineales pero eso puede genera sesgos por omitir variables relevantes.

  Autores como Wooldridge sugieren "exigir" menos a los datos, es decir, ser conscientes de que si tenemos mucha colinelidad aproximada es difícil interpretar por separado las estimaciones de los parámetros de los regresores colineales entre sí.

<u>Ejemplo:</u>

Consideremos el siguiente modelo poblacional

$$Y = 0{,}5 + 0{,}4X_1 + 0{,}3X_2 - 0{,}8X_3 + \varepsilon$$

con

$$V(Y / X_1, X_2, X_3) = 1$$

Consideremos tres muestras de tamaño 500:

a) **En la primera X_1 , X_2 y X_3 son independientes**

b) **En la segunda X_1 y X_2 mantienen una correlación lineal de 0,7. X_3 es independiente.**

c) **En la tercera X_1 y X_2 mantienen una correlación lineal de 0,95. X_3 es independiente.**

Se repite en ejercicio eliminando 50 datos.

Caso 1: Independencia total

Dependent Variable: Y
Method: Least Squares
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.557725	0.045785	12.18150	0.0000
X1	0.446357	0.047487	9.399617	0.0000
X2	0.267752	0.046599	5.745875	0.0000
X3	-0.796338	0.046108	-17.27132	0.0000

R-squared	0.462783	Mean dependent var	0.539872
Adjusted R-squared	0.459533	S.D. dependent var	1.389877
S.E. of regression	1.021788	Akaike info criterion	2.888953
Sum squared resid	517.8493	Schwarz criterion	2.922670
Log likelihood	-718.2383	F-statistic	142.4254
Durbin-Watson stat	1.931620	Prob(F-statistic)	0.000000

Dependent Variable: Y
Method: Least Squares
Sample: 51 500
Included observations: 450

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.573571	0.048897	11.73028	0.0000
X1	0.453967	0.050594	8.972698	0.0000
X2	0.251355	0.049112	5.117982	0.0000
X3	-0.791560	0.049197	-16.08969	0.0000

R-squared	0.461001	Mean dependent var	0.538858
Adjusted R-squared	0.457375	S.D. dependent var	1.404859
S.E. of regression	1.034862	Akaike info criterion	2.915263
Sum squared resid	477.6393	Schwarz criterion	2.951790
Log likelihood	-651.9342	F-statistic	127.1530
Durbin-Watson stat	1.937107	Prob(F-statistic)	0.000000

Caso 2: X_1 y X_2 mantienen una correlación lineal de 0,7. X_3 es independiente.

Dependent Variable: Y
Method: Least Squares
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.508131	0.044678	11.37320	0.0000
X1	0.499978	0.064621	7.737042	0.0000
X2	0.264658	0.054326	4.871662	0.0000
X3	-0.831860	0.044936	-18.51205	0.0000
R-squared	0.569431	Mean dependent var		0.531683
Adjusted R-squared	0.566827	S.D. dependent var		1.512957
S.E. of regression	0.995766	Akaike info criterion		2.837358
Sum squared resid	491.8084	Schwarz criterion		2.871075
Log likelihood	-705.3396	F-statistic		218.6549
Durbin-Watson stat	1.990629	Prob(F-statistic)		0.000000

Dependent Variable: Y
Method: Least Squares
Sample: 51 500
Included observations: 450

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.527775	0.046797	11.27803	0.0000
X1	0.486549	0.067106	7.250431	0.0000
X2	0.289291	0.056460	5.123822	0.0000
X3	-0.849571	0.047042	-18.05969	0.0000
R-squared	0.589041	Mean dependent var		0.528560
Adjusted R-squared	0.586277	S.D. dependent var		1.538062
S.E. of regression	0.989302	Akaike info criterion		2.825215
Sum squared resid	436.5083	Schwarz criterion		2.861741
Log likelihood	-631.6733	F-statistic		213.0890
Durbin-Watson stat	1.923667	Prob(F-statistic)		0.000000

Caso 3: X_1 y X_2 mantienen una correlación lineal de 0,95. X_3 es independiente.

Dependent Variable: Y
Method: Least Squares
Sample: 1 500
Included observations: 500

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.444755	0.044198	10.06278	0.0000
X1	0.689852	0.158781	4.344684	0.0000
X2	-0.020511	0.150236	-0.136528	0.8915
X3	-0.752525	0.044530	-16.89936	0.0000
R-squared	0.504476	Mean dependent var		0.447055
Adjusted R-squared	0.501479	S.D. dependent var		1.397274
S.E. of regression	0.986559	Akaike info criterion		2.818781
Sum squared resid	482.7565	Schwarz criterion		2.852498
Log likelihood	-700.6953	F-statistic		168.3204
Durbin-Watson stat	1.912957	Prob(F-statistic)		0.000000

Dependent Variable: Y
Method: Least Squares
Sample: 51 500
Included observations: 450

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.449265	0.046283	9.706979	0.0000
X1	0.486027	0.170175	2.856039	0.0045
X2	0.151620	0.159971	0.947796	0.3437
X3	-0.759169	0.046625	-16.28236	0.0000
R-squared	0.510427	Mean dependent var		0.427518
Adjusted R-squared	0.507133	S.D. dependent var		1.396903
S.E. of regression	0.980688	Akaike info criterion		2.807724
Sum squared resid	428.9399	Schwarz criterion		2.844251
Log likelihood	-627.7379	F-statistic		154.9990
Durbin-Watson stat	1.990739	Prob(F-statistic)		0.000000

2. CAMBIOS DE ESCALA Y DE ORIGEN

- **En este apartado vamos a analizar cómo se modifican los resultados básicos obtenidos para el Modelo Lineal General ante cambios de escala y/o origen en las variables del modelo.**

- **Los cambios de escala son cambios en las unidades de medidas de las variables.**

- **Los cambios de origen son cambios en el origen de medida de las variables. El más frecuente es considerar las variables en desviaciones respecto al valor medio.**

CAMBIOS DE ESCALA

Consideremos el modelo original

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

Consideremos el modelo con cambio de escala en *Y* y *X*

$Y^* = \beta_0^* + \beta_1^* X^* + \varepsilon^*$ **con**

$$Y^* = \lambda_0 Y \; con \; \lambda_0 \neq 0$$

$$X^* = \lambda_1 X \; con \; \lambda_1 \neq 0$$

¿Qué relación existe entre los dos modelos?

$$\lambda_0 Y = \beta_0^* + \beta_1^* \lambda_1 X + \varepsilon^* \quad \rightarrow \quad Y = \frac{\beta_0^*}{\lambda_0} + \beta_1^* \frac{\lambda_1}{\lambda_0} X + \frac{\varepsilon^*}{\lambda_0}$$

Luego $\dfrac{\beta_0^*}{\lambda_0} = \beta_0 \rightarrow \beta_0^* = \lambda_0 \beta_0 \quad y \quad \beta_1^* \dfrac{\lambda_1}{\lambda_0} = \beta_1 \rightarrow \beta_1^* = \dfrac{\lambda_0}{\lambda_1} \beta_1$

- **Los cambios de escala en la variable endógena afectan a los valores de todos los parámetros del modelo.**
- **Un cambio de escala en una variable exógena sólo afecta al parámetro ligado a dicha variable.**
- **Si se realiza el mismo cambio de escala en la variable endógena y exógena no afecta al valor del parámetro ligado a dicha variable exógena.**
- **Estas relaciones se mantienen entre los parámetros estimados.**

Ejemplo: $Y = \beta_0 + \beta_1 X + \varepsilon$

***Y* = salario/hora (EN PESETAS) y *X* = años de estudio**

Dependent Variable: Y
Method: Least Squares
Sample: 1 5050
Included observations: 5050

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	309.4686	12.68597	24.39455	0.0000
X	50.48059	1.253275	40.27893	0.0000

R-squared	0.243223	Mean dependent var	771.8208
Adjusted R-squared	0.243073	S.D. dependent var	441.1647
S.E. of regression	383.8202	Akaike info criterion	14.73862
Sum squared resid	7.44E+08	Schwarz criterion	14.74121
Log likelihood	-37213.02	F-statistic	1622.392
Durbin-Watson stat	1.809545	Prob(F-statistic)	0.000000

Y = salario/hora (EN EUROS)

Dependent Variable: Y/166.386
Method: Least Squares
Sample: 1 5050
Included observations: 5050

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.859944	0.076244	24.39455	0.0000
X	0.303394	0.007532	40.27893	0.0000

R-squared	0.243223	Mean dependent var	4.638736
Adjusted R-squared	0.243073	S.D. dependent var	2.651453
S.E. of regression	2.306806	Akaike info criterion	4.510000
Sum squared resid	26862.19	Schwarz criterion	4.512585
Log likelihood	-11385.75	F-statistic	1622.392
Durbin-Watson stat	1.809545	Prob(F-statistic)	0.000000

Nótese:

$\hat{\beta}_0^* = \lambda_0 \hat{\beta}_0 \quad \rightarrow$ **1,859944=(1/166,386)*309,4686**

$\hat{\beta}_1^* = \lambda_0 \hat{\beta}_1 \quad \rightarrow$ **0,303394=(1/166,386)*50,48059**

Además:

- **El coeficiente de determinación no varía.**
- **El t-statistic no varía.**

CAMBIOS DE ORIGEN

Consideremos el modelo original

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

Consideremos el modelo con cambio de origen en *Y* y *X*

$$Y^* = \beta_0^* + \beta_1^* X^* + \varepsilon^* \text{ con}$$

$$Y^* = Y + \gamma_0 \; con \; \gamma_0 \neq 0$$

$$X^* = X + \gamma_1 \; con \; \gamma_1 \neq 0$$

¿Qué relación existe entre los dos modelos?

$$Y + \gamma_0 = \beta_0^* + \beta_1^*(X + \gamma_1) + \varepsilon^* \rightarrow Y = \beta_0^* + \beta_1^* \gamma_1 - \gamma_0 + \beta_1^* X + \varepsilon^*$$

Luego: $\beta_0 = \beta_0^* + \beta_1^* \gamma_1 - \gamma_0 \quad y \quad \beta_1 = \beta_1^* \rightarrow \beta_0^* = \beta_0 - \beta_1 \gamma_1 + \gamma_0$

- **Los cambios de origen sólo afectan al término constante del modelo.**

- **El cambio de origen más común es “tomar desviaciones respecto al valor medio” → Modelo en desviaciones respecto a la media**

Modelo original $Y = \beta_0 + \beta_1 X + \varepsilon$

Modelo en desviaciones $Y^* = \beta_0^* + \beta_1^* X^* + \varepsilon^*$

con

$Y^* = Y - E(Y)$ **y** $X^* = X - E(X)$

Entonces

$$\beta_1 = \beta_1^* \quad y \quad \beta_0^* = \beta_0 - \left(E(Y) - \beta_1 E(X)\right) = 0$$

Luego

$$Y^* = \beta_1 X^* + \varepsilon$$

- **Estas relaciones se mantienen entre los parámetros estimados.**

Ejemplo $Y = \beta_0 + \beta_1 X + \varepsilon$

Y **= salario/hora en ptas**
X **= años de estudio**

Dependent Variable: Y
Method: Least Squares
Sample: 1 5050
Included observations: 5050

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	309.4686	12.68597	24.39455	0.0000
X	50.48059	1.253275	40.27893	0.0000

R-squared	0.243223	Mean dependent var	771.8208
Adjusted R-squared	0.243073	S.D. dependent var	441.1647
S.E. of regression	383.8202	Akaike info criterion	14.73862
Sum squared resid	7.44E+08	Schwarz criterion	14.74121
Log likelihood	-37213.02	F-statistic	1622.392
Durbin-Watson stat	1.809545	Prob(F-statistic)	0.000000

Desviaciones respecto a la media

$$Y^* = \beta_1 X^* + \varepsilon$$

$$Y^* = Y - \overline{Y} \text{ y } X^* = X - \overline{X}$$

Dependent Variable: Y^*

Method: Least Squares

Sample: 1 5050

Included observations: 5050

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X^*	50.48059	1.253151	40.28292	0.0000

R-squared	0.243223	Mean dependent var	-7.82E-14
Adjusted R-squared	0.243223	S.D. dependent var	441.1647
S.E. of regression	383.7822	Akaike info criterion	14.73823
Sum squared resid	7.44E+08	Schwarz criterion	14.73952
Log likelihood	-37213.02	Durbin-Watson stat	1.809545

Nótese:

- $\hat{\beta}_1^* = \hat{\beta}_1$ **y** $\hat{\beta}_0$ **podría recuperarse cómo** $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

- **El coeficiente de determinación no varía.**

- **El t-statistic no varía.**