Maestría en Economía

Economía Aplicada

Prof. Martin A. Rossi

Tutores: Paola Llamas y Tomas Pacheco

Problem Set 3: Fuentes de sesgo e imprecisión

Garcia Ojeda, Juan
Hausvirth, Martina
Hayduk, Gaspar
Salvatierra, Elias Lucas D.

Fecha de entrega: 6 de septiembre de 2024

PROBLEM SET 3: FUENTES DE SESGO E IMPRECISIÓN

GARCIA OJEDA - HAUSVIRTH - HAYDUK - SALVATIERRA

Ejercicio 1

Se repite la simulación hecha en clase, incluyendo modificaciones menores para mostrar diferentes puntos.

1. Los errores estándar de los regresores disminuyen si aumenta el tamaño muestral. El error estándar se define como la siguiente ecuación:

$$SE = \frac{s}{\sqrt{n(1 - R_j^2)V(X_j)}}$$

donde s denota el desvío estándar muestral, n el número de observaciones, R_j^2 el r cuadrado que se obtiene de regresar X_j contra los demas regresores y $V(X_j)$ la varianza de X_j. Por lo tanto, es evidente que un aumento del tamaño muestral n implica una disminucion del error estándar.

En este sentido, la Table 1 muestra la variación en los errores estandar ante un aumento del tamaño muestral, siendo la primer columna la que refiere al modelo estimado a partir de una muestra mas pequeña. A partir de dicha tabla, se evidencia que los errores estandár del segundo modelo, es decir, del estimado a partir de una muestra mas grande, son menores.

Table 1. Regresión original vs Regresión con mayor observaciones

	(1) belleza	(2) belleza
alegria	10.000***	9.996***
	(0.0198)	(0.00696)
altura	2.001***	2.001***
	(0.00359)	(0.00111)
Constant	-0.215	-0.193
	(0.611)	(0.190)
Number of observations	100	1000
R-Squared	1.000	1.000

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2. Los errores estándar de los regresores aumentan si aumenta la varianza del termino de error, μ. Como puede observarse en la ecuación planteada en el inciso anterior, el aumento de la varianza del término de error aumenta el desvío estándar muestral.

La Table 2 evidencia que los errores estándar de los regresores del segundo modelo, cuyo término de error tiene mayor varianza respecto al modelo restante, son mayores a los del primer modelo.

Table 2. Regresión original vs Regresión con mayor varianza del término de error

	(1) belleza	(2) belleza
alegria	10.000***	9.673***
	(0.0198)	(0.279)
altura	2.001***	1.947***
	(0.00359)	(0.0507)
Constant	-0.215	11.37
	(0.611)	(8.629)
Number of observations	100	100
R-Squared	1.000	0.961

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3. Los errores estándar de un regresor disminuyen si aumenta la varianza de X. Una mayor varianza de un regresor implica una menor magnitud del error estándar. Los resultados expuestos en la Table 3 muestran que la varianza de la variable del segundo modelo es mayor a la varianza de la variable del primer modelo. Por lo tanto, el error estándar del coeficiente estimado en el segundo modelo es menor al del primer modelo.

Table 3. Regresión original vs Regresión con mayor varianza de la variable alegria

	(1) belleza	(2) belleza
alegria	10.000***	9.991***
	(0.0198)	(0.00535)
altura	2.001***	2.001***
	(0.00359)	(0.00352)
Constant	-0.215	-0.0856
	(0.611)	(0.557)
Number of observations	100	100
R-Squared	1.000	1.000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4. De la salida de Stata se puede observar que la suma de los residuos es un numero aproximadamente igual a cero. Tiene sentido, ya que vimos que el R cuadrado era 1.

5. Los residuos son ortogonales a los regresores. Se observa a partir de la Table 4 que la correlacion entre los residuos y los regresores es nula, lo que significa que los residuos no estan linealmente relacionados con la variable explicativa, es decir, los residuos son ortogonales a los regresores.

Table 4. Matriz de Correlación

	residuos	alegria	altura
residuos	1.0000	-0.0000	0.0000
alegria	-0.0000	1.0000	-0.1259
altura	0.0000	-0.1259	1.0000

6. Observando la Table 5, se puede decir que los valores predichos de la variable dependiente a partir del modelo con multicolinealidad son similaresa los predichos a partir del modelo sin multicolinealidad. Se concluye entonces que la multicolinealidad no afecta los valores de Y y que las predicciones no difieren significativamente de los verdaderos valores de las variables.

Belleza	$\widehat{Belleza}$	$\widehat{Belleza}_{multicol}$
374	373.9688	373.9476
360	359.9665	360.0013
437	436.0079	436.0615
443	441.9724	441.9215
414	413.9679	413.8671

Table 5. Comparación de predicciones para las primeras 5 observaciones

7. Como puede observarse en la Table 6, el valor de los coeficientes estimados son los mismo en los primeros dos modelos. Por lo que se puede decir que el error no aleatorio incluido en el regresor de interés no incide sobre la magnitud de los coeficeintes estimados. Lo que si se ve afectado es el valor del intercepto, dado que al tratarse de un error constante, el efecto es totalmente capturado por el intercepto. Por su parte, el valor de los coeficientes estimados en el tercer modelo difieren de los valores estimados en el primer modelo. Por lo tanto, puede decirse que un error aleatorio en un regresor si modifica la magnitud de los coeficientes. El coeficiente estimado del regresor de interes en el tercer modelo es mas chico en terminos absolutos que estimado en el primer modelo.

8. Tal como lo muestra la Table 8, un error no aleatorio en la variable explicada no afecta a la magnitud de los coeficientes estimados. Al igual que en el inciso anterior, el efecto del error es capturado en su totalidad por el intercepto. Por su parte, un error aleatorio si afecta el valor de los coeficientes estimados dado que la presencia de dicho error en la variable explicada genera estimadores mas imprecisos.

TABLE 6

	(1) belleza	(2) belleza	(3) belleza
alegria	10.000***		
	(0.0198)		
alegria1		10.000***	
		(0.0198)	
alegria2			1.299***
			(0.318)
altura	2.001***	2.001***	1.879***
	(0.00359)	(0.00359)	(0.171)
Constant	-0.215	-50.21***	97.61***
	(0.611)	(0.657)	(27.31)
Number of observations	100	100	100
R-Squared	1.000	1.000	0.567

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 7

	(1) belleza	(2) belleza1	(3) belleza2
alegria	10.000***	10.000***	10.14***
	(0.0198)	(0.0198)	(0.276)
altura	2.001***	2.001***	1.946***
	(0.00359)	(0.00359)	(0.0500)
Constant	-0.215	4.785***	6.305
	(0.611)	(0.611)	(8.522)
Number of observations	100	100	100
R-Squared	1.000	1.000	0.963

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

EJERCICIO 2

Se asume que estamos interesados en estimar el efecto causal de X_1 en Y. Para ello, supondremos que Y es la nota en un exámen de matemática, X_1 la nasistencia a clases, X_2 el promedio del alumno, y X_3 la cantidad de horas que estudia dicho alumno por semana. Haremos dos regresiones distintas, la primera

$$score_i = \beta_0 + \beta_1 attend_i + \mu_i$$

en donde llamaremos $\tilde{\beta}_1$ al estimador de la regresión de Y en X_1

$$score_i = \beta_0 + \beta_1 attend_i + \beta_2 cgpa_i + \beta_3 study_i + \mu_i$$

Para esta última regresión, llamaremos $\hat{\beta}_1$ al coeficiente de la asistencia de la regresión de y en X_1, X_2, X_3.

1. Si X_1 está altamente correlacionada con X_2 y X_3, y a su vez estas últimas dos tienen una correlación alta con Y, esperamos que los coeficientes de ambas estimaciones $\tilde{\beta}_1$ y $\hat{\beta}_1$ sean muy distintos. Esto es porque dado que sabemos que las variables X_2 y X_3 son relevantes (es decir, los coeficientes son distintos de cero β_2 y $\beta_3 \neq 0$, y además estan correlacionadas con X_1 ($E(X_1 X_s) \neq 0$ y $E(X_1 X_3) \neq 0$), tendremos un problema de sesgo por variables omitidas. Lo que significa, que probablemente $\tilde{\beta}_1$ esté estimando también el efecto de β_2 y β_3. Con lo cual Si X_1 está altamente correlacionada con X_2 y X_3, y a su vez estas últimas dos tienen una correlación alta con Y, esperamos que los coeficientes de ambas estimaciones $\tilde{\beta}_1$ y $\hat{\beta}_1$ sean muy distintos. Esto es porque dado que sabemos que las variables X_2 y X_3 son relevantes (es decir, los coeficientes son distintos de cero β_2 y $\beta_3 \neq 0$, y además estan correlacionadas con X_1 ($E(X_1 X_s) \neq 0$ y $E(X_1 X_3) \neq 0$), tendremos un problema de sesgo por variables omitidas. Lo que significa, que probablemente $\tilde{\beta}_1$ esté estimando también el efecto de β_2 y β_3. Con lo cual $\tilde{\beta}_1 > \hat{\beta}_1$ ya que cuando estamos frente a un problema de omisión de variables:

$$E(\tilde{\beta}_1) = \beta_1 + \underbrace{(X_1 X_1)^{-1} X_1 X_2}_{\neq 0} * \underbrace{\beta_2}_{\neq 0} + \underbrace{(X_1 X_1)^{-1} X_1 X_3}_{\neq 0} * \underbrace{\beta_3}_{\neq 0}$$

2. En el caso en que X_1 no esté correlacionada con X_2 y X_3, por mas de que entre ellas estén correlacionadas, o incluso sean relevantes para el modelo, esto no producirá sesgos por variable omitida. La esperanza del coeficiente estimado entonces será igual al verdadero valor en ambas estimaciones, lo que significa en conclusión, que se espera que $\tilde{\beta}_1$ y $\hat{\beta}_1$ sean similares.

$$E(\tilde{\beta}_1) = \beta_1 + \underbrace{(X_1 X_1)^{-1} X_1 X_2}_{=0} * \underbrace{\beta_2}_{\neq 0} + \underbrace{(X_1 X_1)^{-1} X_1 X_3}_{=0} * \underbrace{\beta_3}_{\neq 0}$$

$$E(\tilde{\beta}_1) = E(\hat{\beta}_1) = \beta_1$$

3. En econometría siempre enfrentamos un trade off entre *sesgo y varianza.* El primero se produce por omisión de variables relevantes, mientras que el segundo se produce por inclusión de variables irrelevantes. Si bien en los problemas a resolver nunca conocemos el *verdadero modelo*, podríamos sospechar en este caso, que al incluir una variable adicional X_4, que intenta explicar cual es el efecto de comer chocolate sobre la nota en un exámen de matemática, sea irrelevante, y nos genere un problema de varianza.

Si estuvieramos frente a este problema, lo que ocurriría es que si bien los coeficientes de ambas regresiones serían iguales (es decir $\dot{\beta}_1 = \hat{\beta}_1$), la varianza del estimador de la regresión que incluye la variable X_4 sería mucho mas alta que la que no la inclye. Es decir:

$$Var(\dot{\beta}_1) > Var(\hat{\beta}_1)$$

4. En el hipotético caso en el que X_1 estuviera altamente correlacionado con X_2 y X_3, pero que entre ellos hubiera correlación con Y, aunque fuera poca, esperamos que la estimación de los coeficientes $\hat{\beta}_1$ y $\tilde{\beta}_1$ sean distintos. Como fue mencionado en el punto 1, asumiento que X_2 y X_3 son variables relevantes y que β_2 y $\beta_3 \neq 0$ estamos frente a un problema de variables omitidas. Sin embargo, si la relevancia de estas dos variables para explicar a Y fuera muy baja, cercana a 0, este problema se soluciona, ya que aunque estén correlacionadas con X_1, su inclusión no es importante ya que no son variables relevantes, por lo que:

$$E(\tilde{\beta}_1) = \beta_1 + \underbrace{(X_1X_1)^{-1}X_1X_2}_{\neq 0} * \underbrace{\beta_2}_{=0} + \underbrace{(X_1X_1)^{-1}X_1X_3}_{\neq 0} * \underbrace{\beta_3}_{=0}$$

$$E(\tilde{\beta}_1) = E(\hat{\beta}_1) = \beta_1$$

5. Si bien incluir las variables X_2 y X_3 en este caso no afecta al error estándar del coeficiente estimado de X_1, dado que esta última no está correlacionada con las demás; esperamos que los errores estándar de $\hat{\beta}_1$ y $\tilde{\beta}_1$ sean distintos. La inclusión de X_2 y X_3, que tienen grandes efectos marginales sobre Y, reduce la varianza del término de error en el modelo. Una menor varianza del término de error implica que los estimadores obtenidos mediante Mínimos Cuadrados Ordinarios (MCO) tendrán varianzas más pequeñas. En consecuencia, el error estándar de $\hat{\beta}_1$ será menor que el de $\tilde{\beta}_1$.

6. Dado que $\dot{\beta}_1$ es el coeficiente de X_1 de la regresión de Y en X_1, X_2, X_3, X_4, y que $\hat{\beta}_1$ es el coeficiente de X_1 pero de la regresión que **no** incluye la variable X_4, la cual asumimos irrelevante, los errores estándar de ambos coeficientes serán distintos.

A partir de la ecuación descripta en el inciso 1 del ejercicio 1, vemos que los errores estándar de una variable dependen negativamente del R^2, y positivamente de la varianza del término de error (s).

Como supusimos en el ejercicio anterior que X_4 no es una variable relevante, su inclusión aumenta la varianza del término de error. Asimismo, como X_1 y X_4 están correlacionados entre si, el R^2 aumenta con la regresión que inlcuye a X_4. Por lo cual, por dos vías distintas sucede que el error estándar de $\dot{\beta}_1$ será mayor que el error estándar de $\hat{\beta}_1$.

```
/********************************************************************************
                   Semana 4: Fuentes de sesgo e imprecisión
                          Universidad de San Andrés
                              Economía Aplicada
*******************************************************************************/

/********************************************************************************
Este archivo sigue la siguiente estructura:

0) Set up environment

1) Ejercicio 1

2) Ejercicio 2
*******************************************************************************/

* 0) Set up environment
*==============================================================================*

global main "/Users/gasparhayduk/Desktop/Economía Aplicada/PS3"
global input "$main/input"
global output "$main/output"

cd "$main"

* Ejercicio 1

* Realizamos una simulación similar a la realizada en la tutorial
* Vamos a generar un modelo que relaciona la belleza de las personas con la altura y un indice de alegria.

*generamos 100 observaciones
clear
set obs 100
set seed 1233

*Generamos el primer regresor: la altura. Le asignamos una distribución normal con media en 160cm y desvío estandar de 20cm.
gen altura = int(rnormal(160,20))

*Generamos la variable de peso y vamos a suponer que existe una relación positiva cercana a 1 entre la altura y el peso de las personas, de manera tal que cada individuo pesa en kilos aproximadamente la mitad de su altura menos un valor aleatorio que tiene distribución normal con media 10 y desvío estandar 5.
gen peso = int(altura/2 - rnormal(10,5))
```

```
*chequeamos la correlacion que existe entre la altura y el peso. Resulta ser 0.9
corr peso altura

*Generamos la variable "alegria" que es ortogonal con la altura y el peso, es decir, la alegria esta presente en los gorditos, flaquitos, altos y enanos.
gen alegria = int(rnormal(10,3))

*Chequeamos la correlación entre los tres regresores. Efectivamente correlaciona muy poco con el resto de regresores
corr peso altura alegria

*Generamos el termino de error
gen u = int(rnormal(0,1))

*Definimos el verdadero modelo. Como somos fieles creyentes que en la vida no todo entra por los ojos, y basandonos en la frase de la canción de Riki Maravilla "De nada sirve la pinta cuando no tienes el fuego", es que la alegría tiene mayor ponderación en explicar la belleza. Además, en nuestro mundo ideal nadie tiene en cuenta el peso de la persona para determinar su belleza, por lo que esta es irrelevante.
gen belleza =  10*alegria + 2*altura + u


*** Consigna a) ¿Que sucede con los errores estandar de los regresores si aumenta el tamaño muestral? ***

*Generamos la regresión con los regresores correctos
reg belleza alegria altura

*Guardamos la salida
predict y_hat
est store ols1

*Aumentamos el tamaño de la muestra
preserve
set obs 1000
replace altura = int(rnormal(160,20)) in 101/1000
replace alegria = int(rnormal(10,3)) in 101/1000
replace u = int(rnormal(0,1)) in 101/1000
replace belleza = 10*alegria + 2*altura + u in 101/1000
*Generamos la regresión con muestra mayor
reg belleza alegria altura
predict y_hat2
est store ols2
*Exportamos tablas
esttab ols1 ols2 using "$output/tables/Table 1.tex", replace label se ///
stats(N r2, fmt(0 3) labels("Number of observations" "R-Squared"
```

```
))
restore


*** Consigna b) ¿Que sucede con los errores estandar de los regresores si aumenta la varianza del termino de error? ***
preserve
replace u = int(rnormal(0,10))
replace belleza =  10*alegria + 2*altura + u

*Generamos la regresión con mayor varianza del termino de error
reg belleza alegria altura
predict y_hat2
est store ols2

*Exportamos tablas
esttab ols1 ols2 using "$output/tables/Table 2.tex", replace label se ///
stats(N r2, fmt(0 3) labels("Number of observations" "R-Squared" ))
restore

*** Consigna c) ¿Que sucede con los errores estandar de un regresor si aumenta la varianza de X?
preserve
replace alegria = int(rnormal(10,15))
replace belleza = 10*alegria + 2*altura + u

*Generamos regresión con mayor varianza de X=alegria
reg belleza alegria altura
predict y_hat2
est store ols2

*Exportamos tablas
esttab ols1 ols2 using "$output/tables/Table 3.tex", replace label se ///
stats(N r2, fmt(0 3) labels("Number of observations" "R-Squared" ))
restore

*** Consigna d) ¿Cuanto vale la suma de residuos?
*DE ESTA CONSIGNA NO ENTIENDO SI PIDE QUE COMPAREMOS LA SUMA DE RESIDUOS ENTRE MODELOS O SOLAMENTE LE DEMOS EL NUMERO DE LA SUMA DE RESIDUOS DE UN MODELO

*Primero obtengo la suma de residuos del modelo original
reg belleza alegria altura
predict residuos, residuals
egen suma_residuos = total(residuos)

di "La suma de residuos es: " suma_residuos
```

```
*Practicamente dan cero

**** Consigna e) ¿Son los residuos ortogonales a los regresores?

*Para ver si son ortogonales se puede observar la relación entre los mismos. Tomamos los residuos de la regresión llevada a cabo en el anterior inciso

estpost corr residuos alegria altura
eststo correlation
esttab using "$output/tables/Tabla 4.tex", replace

**** Consigna f) ¿Como afecta la alta multicolinealidad a la estimación de Y?

*Hacemos la regresión incluyendo peso que tiene una relación fuerte con altura y guardamos la prediccion

reg belleza alegria altura peso
predict y_hat_mult

*estpost list belleza y_hat y_hat_mult in 1/5
*esttab using "$output/tables/Tabla 5.tex", replace


*armarmos una lista con la variable original, la prediccion del modelo sin multicolinealidad y la prediccion del modelo con multicolinealidad
list belleza y_hat y_hat_mult in 1/5

*No se bien como exportar una tabla decente para este caso, asi que la arme manual REVISAR!!!!!!!


**** Consigna g) ¿Que sucede si corren una regresion con un error no aleatorio en X?¿Y si ese error fuera aleatorio?

*Comenzamos generando un error no aleatorio en alegria
gen alegria1 = alegria + 5
*Regresamos belleza con sus regresores y reemplazamos alegria por alegria1
reg belleza alegria1 altura
predict y_hat_noaleat
est store ols3


*Generamos un error aleatorio en alegria
gen error = int(rnormal(0,10))
gen alegria2 = alegria + error
*Regresamos belleza con sus regresores y reemplazamos alegria por alegria2
```

```
reg belleza alegria2 altura
predict y_hat_aleat
est store ols4

*Exportamos las salidas de las regresiones
esttab ols1 ols3 ols4 using "$output/tables/Table 6.tex", replace
 label se ///
stats(N r2, fmt(0 3) labels("Number of observations" "R-Squared"
))

**** Consigna h) ¿Que sucede si corren una regresión con un
error no aleatorio en Y?¿Y si ese error fuera aleatorio?

*Comenzamos generando un error no aleatorio en belleza
gen belleza1 = belleza + 5
*Regresamos belleza con sus regresores y reemplazamos alegria
por alegria1
reg belleza1 alegria altura
predict y_hat_noaleat1
est store ols5


*Generamos un error aleatorio en belleza, utilizamos la misma
variable aleatoria que en el inciso anterior
gen belleza2 = belleza + error
*Regresamos belleza con sus regresores y reemplazamos alegria
por alegria2
reg belleza2 alegria altura
predict y_hat_aleat1
est store ols6

*Exportamos las salidas de las regresiones
esttab ols1 ols5 ols6 using "$output/tables/Table 7.tex", replace
 label se ///
```