Problem Set 3:
Fuentes de sesgo e imprecisión

Economía Aplicada

Universidad de
SanAndrés

Alumnos:
Juan Diego Barnes
Fransisco Legaspe
Rodrigo Martin
Diego Fasan

Profesor: Martín Rossi
Tutores: Gastón García Zavaleta
Tomás Pacheco

1. Ejercicio 1: Fuentes de imprecisión

En este punto realizaremos simulaciones, para demostrar diferentes puntos acerca propiedades y otros aspectos del estimador de MCO. Principalmente queremos ver como afectan determinadas caracteristicas de los datos, sobre el nivel de presicion del estimador de MCO, para esto es bueno recordar la formula de la varianza del estimador,

$$V(\hat{\beta}_j) = \frac{\sigma^2}{n(1 - R_j^2)V(X_j)}, \tag{1}$$

esta excreción nos permitirá ser mas claros al exponer como afectas diferentes aspectos a la precisión de del estimador, donde n es el tamaño muestral, σ^2 la varianza del error, $V(X_j)$ es la variabilidad del regresor j y R_j^2 da cuenta de como la multicolinealidad afecta a la varianza del estimador.

Para todas las simulaciones generaremos las mismas variables que en en el ejemplo provisto: *wage*, *education*, *intelligence*, *a* y *b*. Para alguno casos generaremos todas, en otros sola las necesarias para exponer nuestra idea, para mas detalle sobre como fueron creadas las diferentes variables se puede ver el código en el apéndice.

1.1. Tamaño Muestral

Para ver el efecto del tamaño muestral sobre la precisión del estimador, generamos dos muestras aleatorias una de 50 y otra de 200 observaciones de las variables *wage*, *education* y *intelligence*, las cuales provienen del mismo DGP (*data generating process*).

Tabla 1: Resultados de Regresión

Variables	*wage*			
	DGP 1		DGP 2	
	Coeficiente	EE	Coeficiente	EE
Education	1.97	(0.098)	1.99	(0.067)
Intelligence	3.00	(0.004)	3.00	(0.003)
R^2	1.00		1.00	
N	100		200	

Nota: Los errores estándar presentados son robustos a heterocedasticidad. Todos los coeficientes son significativos al 1 % de significatividad.

Fuente: Elaboración propia.

Como podemos observar en la ecuación (1) la varianza del estimador es una función inversa del tamaño muestral, por lo que podemos observar que la varianza del estimador, y por lo tanto sus errores estándar, es menores al aumentar el tamaño de la muestra.

1.2. Varianza del error

En segundo lugar, lo que queremos ver es como afecta una mayor varianza del error al estimador de MCO, para esta generamos dos DGPs con diferentes términos de error, los cuales solo difieren en el grado de variabilidad, el primero u_1 tiene una varianza, σ_1^2, de 0.5 y el segundo tiene una varianza, σ_2^2, de 2.

Tabla 2: Resultados de Regresión

Variables	*wage*			
	DGP 1		DGP 2	
	Coeficiente	EE	Coeficiente	EE
Education	2.00	(0.041)	2.00	(0.129)
Intelligence	3.00	(0.002)	3.00	(0.005)
R^2	1.00		1.00	
N	200		200	

Nota: Los errores estándar presentados son robustos a heterocedasticidad. Todos los coeficientes son significativos al 1 % de significatividad.

Fuente: Elaboración propia.

Como ya sabemos de la formula de la varianza del estimador, en (1), podemos observar que esta depende positivamente de la varianza del error. De modo que los errores estándar, aumenta con la varianza del error.

1.3. Varianza del regresor

Ahora queremos ver como la varianza del regresor afecta la precisión del estimador, para esto generamos dos DGPs donde la única diferencia es la variabilidad del regresor, *intelligence*.

Tabla 3: Resultados de Regresión

Variables	*wage*			
	DGP 1		DGP 2	
	Coeficiente	EE	Coeficiente	EE
Education	1.98	(0.129)	1.98	(0.129)
Intelligence	3.00	(0.129)	3.00	(0.002)
R^2	1.00		1.00	
N	200		200	

Nota: Los errores estándar presentados son robustos a heterocedasticidad. Todos los coeficientes son significativos al 1 % de significatividad.

Fuente: Elaboración propia.

Podemos observar que los errores estándar, y por lo tanto la varianza, del estimador del efecto de ese regresor disminuyen con la variabilidad del mismo.

1.4. Suma de residuos

Queremos ver como se comporta la suma de residuos de la estimación de MCO, para esto tomamos el DGP utilizamos en el primer ítem, con 200 observaciones. Podemos ver que la suma de residuos es prácticamente igual a 0, esto ocurre por definición del MCO. Si esto no ocurriera MCO no minimizaría esta suma de cuadrados de los residuos, lo cual es su definición.

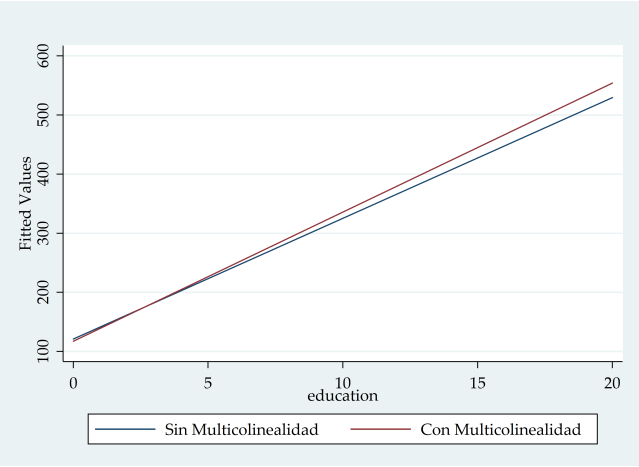
1.5. Ortogonalidad de los residuos

Esto es necesariamente así, geométricamente, la minimización de la residuos al cuadrado, implica un proyección ortogonal del vector que representa el regresando, sobre el espacio que que forman los regresores. Esto se puede ver como una interpretación geométrica del punto anterior, ocurre por definición del estimador de MCO, los residuos siempre serán ortogonales a los regresores. Para chequear esto realizamos un regresión de los residuos sobre los regresores del modelo y como es esperado encontramos que todos los coeficientes son ceros.

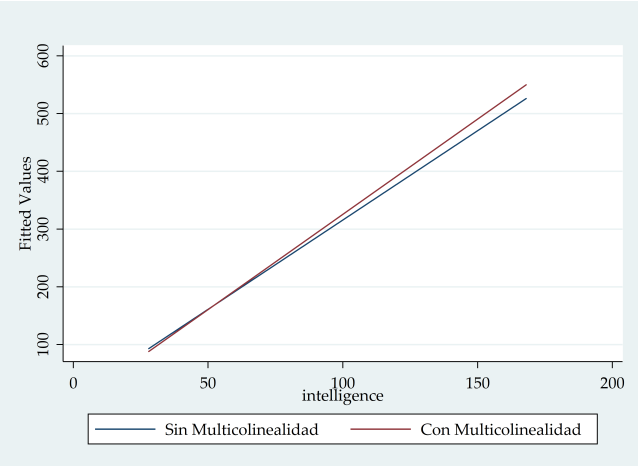
1.6. Multicolinealidad

La multicolinealidad no afecta las estimaciones, solo afecta la varianza del estimador como vemos en la formula (1). Para demostrar este punto generamos dos DGPs, en el primero supondremos que las variables *education*, *intelligence* no están correlacionadas, y luego lo compararemos con un segundo donde estas se encuentran correlacionadas (no perfectamente). Podemos ver que las estimaciones de los parámetros, son las mismas en ambos casos, la multicolinealidad solo afecta el tamaño de los errores estándar de los coeficientes. Por lo tanto, las predicciones de *wâge* son las mismas ante la presencia de multicolinealidad. Y podemos ver que el R^2 no se ve afectado por esta.

Figura 1: Efecto de la Multicolinealidad sobre las precicciones



wâge sobre *education*



wâge sobre *intelligence*

La multicolinealidad no es un problema para las predicciones. En muestras muy pequeñas puede generar un sesgo en los estimadores al contarse con poca variabilidad entre regresores que permitan reconocer los efectos, cuando existen variables relevantes excluidas, aunque estas variables no se correlacionen con las incluidas. Pero esto no es un problema para la predicción, solo para los estimadores de los β. En nuestra simulación no incluimos variables relevantes, no correlacionadas con *education*, *intelligence*, para wage. Incluirlas, nos podrías llevar a pensar erróneamente que la multicolinealidad genera un sesgo, cuando la verdadera razón se relaciona con el tamaño de la muestra.

1.7. Error no aleatorio en X

Para ver el efecto de un error no aleatorio un uno de los regresores, generamos la variable *intelligence*, con un error que definimos como una secuencia que suma a cada valor el numero de su posesión, es decir a la primera observación le suma 1, a la segunda 2 y así sucesivamente. Los resultados de la regresión se pueden observar en la tabla 4, donde podemos observar que esto no genera un sesgo, y además por la forma en la que definimos el error no aleatorio este genera mayor variabilidad en X por lo que la varianza del estimador disminuye.

Tabla 4: Resultados de Regresión

Variables	**Score**					
	Baseline		**Error no aleatorio en X**		**Error no aleatorio en Y**	
	Coeficiente	EE	Coeficiente	EE	Coeficiente	EE
education	1.98	(0.129)	1.98	(0.129)	1.94	(0.180)
intelligence	3.01	(0.005)	3.01	(0.004)	3.01	(0.007)
R^2	0.99		0.99		0.51	
N	200		200		200	

Nota: Los errores estándar presentados son robustos a heterocedasticidad. Todos los coeficientes son significativos al 1 % de significatividad.

Fuente: Elaboración propia.

1.8. Error no aleatorio en Y

Para este caso, generamos el error tal que este dependa de la variable *education*, lo que seria un caso de heterocedasticidad, esto aumenta la varianza del estimador pero no afecta su insesgades o inconsistencia, como podemos observar en la columna 3 de la Tabla 4. Esto solo genera que al estar utilizando errores robustos a la heterocedasticidad estos sean mayores en este caso.

2. Ejercicio 2

En este ejercico lo que nos interesa es ver el efecto causal de X_1 sobre Y, bajo diferentes DGPs. Donde podemos suponer que Y es la nota en un examen de matemática, X_1 la asistencia a clases, X_2 el promedio del alumno y X_3 la cantidad de horas que estudia el alumno por semana. Ahora llamaremos $\tilde{\beta}_1$ al estimador de la regresión:

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

y $\hat{\beta}_1$ al coeficiente de la regresión del modelo:

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

A partir de estos modelos podemos ejemplificar las respuestas a diferentes preguntas que pueden surgir respecto al problema de variable omitida y la relación entre las variables:

1. Si *attend* está altamente correlacionada con *cgpa* y con *study*, y *cgpa* y *study* tienen grandes efectos paciales en el *score*, ¿esperan que $\hat{\beta}_1$ y $\tilde{\beta}_1$ sean similares o distintos?

   Si *attend* esta altamente correlacionado con *cgpa* y *study*, el no incluir estas variables generara un sesgo en el coeficiente $\tilde{\beta}_1$. El cual dependerá del signo y el tamaño de las correlaciones, como del efecto parcial de las variables omitidas. Al incluir *cgpa* y *study* en la regresion, se elimina este sesgo en el estimador de β_1, $\hat{\beta}_1$ no sera sesgaedo. La multicolinealidad de *attend* con *cgpa* y *study* no genera un problema de sesgo al incluir estas variables, solo genera una mayor varianza del estimador.

2. Si *attend* no está correlacionada con *cgpa* y con *study*, pero *cgpa* y con *study* están altamente correlacionadas entre ellas, ¿esperan que $\hat{\beta}_1$ y $\tilde{\beta}_1$ sean similares o distintos? Expliquen.

Si *attend* no esta correlacionado con *cgpa* y *study*, el no incluir estas variables no generara un sesgo en el coeficiente $\tilde{\beta}_1$. Al incluir *cgpa* y *study* en la regresion, lo que se obtendra es una menor varianza del estimador de β_1. La multicolinealidad entre *cgpa* y *study* no genera un problema de sesgo, solo aumenta la varinza del estimador del efectos de estos.

3. Si $\dot{\beta}_1$ es el coeficiente de asistencia a clase en la regresión de Y en *attend*, *cgpa*, *study* y *chocolate*, en donde *chocolate* es el consumo de chocolate del alumno, ¿esperan que $\dot{\beta}_1$ y $\hat{\beta}_1$ sean similares o distintos?

   Podemos considerar que si el alumno consume mucho chocolate podría tener problemas estomacales, o insomnio por el azúcar lo que podria generar una correlación negativa con nuestra variable de interes, *attend*. Bajo este razonamiento, $\hat{\beta}_1$ seria sesgado al no considerar el consumo de chocolate en el modelo, pero no sabemos la dirección del sesgo al menos que también asumamos el signo del coeficiente del consumo de chocolate sobre el *score*, en el caso de que el efecto parcial del consumo de chocolate sobre el *score* sea cero, no existiría sesgo.

   Si consideramos que el efecto del consumo de chocolate es distinto de cero, $\dot{\beta}_1$ seria insesgado y $\hat{\beta}_1$ sesgado, en cambio si es cero $\hat{\beta}_1$ seria insesgdo.

4. Si *attend* está altamente correlacionada con *cgpa* y *study* pero *cgpa* y *study* tienen pequeños efectos parciales en Y, ¿esperan que $\hat{\beta}_1$ y $\tilde{\beta}_1$ sean similares o distintos?

   Mientras menores sean los efectos parciales sobre menor sera el sesgo que genere no incluir las variables en el modelo, como discutimos en el punto anterior cuando consideramos que el efecto parcial del consumo de chocolate podría ser cero. De modo que, mientras mas cercano a cero sea el efecto parcial, mas similares deberían ser $\hat{\beta}_1$ y $\tilde{\beta}_1$ a pesar de estar correlacionados.

5. ¿Cómo esperan que sea la relación entre los errores estándar de los coeficientes $\hat{\beta}_1$ y $\tilde{\beta}_1$ si: *attend* esta incorrelacionada con *cgpa* y *study*, *cgpa* y *study* tienen grandes efectos marginales en el *score* y están altamente correlacionadas entre si?

   Como vimos en la ecuacion (1), sabemos que la varianza de los coeficientes dependen positivamente del grado de multicolinealidad, de modo que los errores de $\hat{\beta}_1$ y $\tilde{\beta}_1$ no se verán afectados al no estar correlacionados con *cgpa* y *study*, R_1^2 no se vera afectado por lo tanto tampoco $V(\beta_1)$. Pero el error estándar de $\hat{\beta}_1$ si podría ser menor al de $\tilde{\beta}_1$, al incluir mas regresores.

6. ¿Cómo esperan que sea la relación entre los errores estándar de $\dot{\beta}_1$ y $\hat{\beta}_1$?

   Como consideramos que si el alumno consume mucho chocolate podría tener problemas estomacales, o insomnio por el azúcar lo que podría generar una correlación negativa con nuestra variable de interés, *attend*. Al existir una relación entre *chocolate* y *attend*, el error estándar sera mayor en $\dot{\beta}_1$ que en $\hat{\beta}_1$.

Apéndice. Codigo PS3 (STATA)

2023-08-17

```
/*******************************************************************************
                    Semana 4: Fuentes de sesgo e imprecisión

                            Universidad de San Andrés
                                Economía Aplicada
/*******************************************************************************/
                     Barnes, Fasan, Legaspe y Martin
/*******************************************************************************
Este archivo sigue la siguiente estructura:

0) Set up environment

1) Simulaciones:

1.1) Un ejemplo para diferencias en el Tamaño Muestral

1.2) Un ejemplo para diferencias en la varianza del error

1.3) Un ejemplo para diferencias en la varianza del regresor

1.4) El valor de la suma de los residuos

1.5) ¿Los residuos son ortogonales a los regresores?

1.6) Un ejemplo para Multicolinealidad

1.7) Error no aleatorio en X

1.8) Erro no aleatorio en Y


2) Fuentes de Sesgo e Inconsistencia

********************************************************************************/

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

global main "C:\Users\Usuario\Desktop\MAESTRIA\Economia Aplicada\TPs\PS3"
global input "$main/input"
global output "$main/output"

cd "$main"
*================================================================================*
```

```
* 1) Simulaciones
* Para todas las simulaciones generaremos las variables que en en el
*   ejemplo provisto: wage, education, intelligence, a y b.
*   Para alguno casos generaremos todas, en otros sola las necesarias para
*   exponer nuestra idea.   0
*================================================================================*
* 1.1) Un ejemplo para diferencias en el Tamaño Muestral
* N = 100
clear
* GPD:
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*1+7)
gen b=int(invnormal(uniform())*1+5)
gen wage=3*intelligence+ 2*education +6 + u

reg wage education intelligence, robust


* N = 200
clear
* GPD:
set obs 200
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)
gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*1+7)

gen wage=3*intelligence+ 2*education +6 + u

reg wage education intelligence, robust



*================================================================================*
*1.2) Un ejemplo para diferencias en la varianza del error
* sigma_1 = 0.5
clear
* GPD:
set obs 200
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*0.5+7)

gen wage=3*intelligence+ 2*education +6 + u
```

```
reg wage education intelligence, robust

* sigma_2 = 2
clear
* GPD:
set obs 200
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*2+7)

gen wage=3*intelligence+ 2*education +6 + u

reg wage education intelligence, robust



*================================================================================*
*1.3) Un ejemplo para diferencias en la varianza del regresor
* caso 1
clear
* GPD:
set obs 200
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*2+7)

gen wage=3*intelligence+ 2*education +6 + u

reg wage education intelligence, robust


* Caso 2
clear
* GPD:
set obs 200
set seed 1234
gen intelligence=int(invnormal(uniform())*50+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*2+7)

gen wage=3*intelligence+ 2*education +6 + u

reg wage education intelligence, robust
```

```
*================================================================================*
*1.4) El valor de la suma de los residuos
* DGP
* Aca si agregamos una varaible mas para que no parezca que los resiudos son 0,
* por que el modelo ajsuta con R^2 = 1.
clear
* GPD:
set obs 200
set seed 1234
gen intelligence=int(invnormal(uniform())*50+100)

gen education= int(invnormal(uniform())*70+5)
corr education intelligence
gen u=int(invnormal(uniform())*2+7)
gen a = int(invnormal(uniform())*1+7)
gen b = int(invnormal(uniform())*2+15)

gen wage=3*intelligence+ 2*education + 6*b + 20*a +6 + u

reg wage education intelligence, robust


predict residuos, residuals
sum residuos
* Podemos ver que dan cero, lo que ocurre por definicion
*================================================================================*
*1.5) ¿Los residuos son ortogonales a los regresores?
* Una forma sencilla de ver esto es regresando los residuos sobre las variables
reg residuos education intelligence, robust
* Podemos ver que todos los coeficientes son 0

*================================================================================*
*1.6) Un ejemplo para Multicolinealidad
* Sin Multicolinealidad
clear
set obs 5000
set seed 1233
gen intelligence=int(invnormal(uniform())*20+100)

gen education= int(invnormal(uniform())*50+5)
corr education intelligence
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+ 2*education + u

reg wage education intelligence, robust
predict y_hat_1, xb

* Con mutlicolinealidad
set seed 1233
replace education=int(intelligence/10+invnormal(uniform())*1) // multicolinealidad
corr education intelligence
replace wage=3*intelligence+ 2*education + u
```

```
reg wage education intelligence, robust
predict y_hat_2, xb

sort intelligence
twoway (lfit y_hat_1 intelligence) ///
       (lfit y_hat_2 intelligence), ///
       title(" ") ///
       legend(label(1 "Sin Multicolinealidad") label(2 "Con Multicolinealidad")) ///
       xtitle("intelligence") ytitle("Fitted Values")
graph export "$output/Multicol_intel.png", width(4000)
sort education
twoway (lfit y_hat_1 education) ///
       (lfit y_hat_2 education), ///
       title(" ") ///
       legend(label(1 "Sin Multicolinealidad") label(2 "Con Multicolinealidad")) ///
       xtitle("education") ytitle("Fitted Values")
graph export "$output/Multicol_educ.png", width(4000)

*================================================================================*
* BASELINE PARA 1.7 y 1.8
clear
* GPD:
set obs 200
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*2+7)

gen wage=3*intelligence+ 2*education +6 + u

reg wage education intelligence, robust

*================================================================================*

*================================================================================*
*1.7) Error no aleatorio en X
clear
* GPD con error no aleatorio en X:
set obs 200
set seed 1234
* Generamos un error no aleatorio para intelligence
gen v = _n // error no aleatorio que aumenta con cada observacion
gen intelligence =int(invnormal(uniform())*20*0.1*v+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*2+7)

gen wage= 3*intelligence+ 2*education + 6 + u

reg wage education intelligence, robust
```

```
*================================================================================*
*1.8) Erro no aleatorio en Y
clear
* GPD con error no aleatorio en X:
set obs 200
set seed 1234

gen intelligence=int(invnormal(uniform())*20+100)

gen education= int(invnormal(uniform())*1+5)
corr education intelligence
gen u=int(invnormal(uniform())*2+7) // error no aleatorio
* Generamos un error no aleatorio para intelligence como una secuancia
gen v = _n
gen wage= 3*intelligence+ 2*education + 6 + u + v

reg wage education intelligence, robust


*================================================================================*
*2) Fuentes de sesgo (ESTO NO ERA NECESARIO PERO COMO LO HICIOS LO DEJAMOS)
*================================================================================*
* GDP
clear
set obs 200
set seed 1233
gen attend = rnormal(25, 4)
*================================================================================*
* X1 altamente correlacionada con X2 y X3
gen cgpa = int(attend/1.6+invnormal(uniform())*2)
corr attend cgpa

gen study = int(attend/0.7+invnormal(uniform())*3)
corr attend study

gen u=int(invnormal(uniform())*2+5)


gen score = 1.5*attend + 3*cgpa + 4*study + u

reg score attend


reg score attend cgpa study, robust
*================================================================================*
* X1 no correlacionada con X2 y X3, X2 y X3 altamente correlacionadas
replace cgpa = int(invnormal(uniform())*2)
corr attend cgpa

replace study = int(cgpa/0.3+invnormal(uniform())*3)
corr attend study
corr cgpa study
```

```
replace score = 1.5*attend + 3*cgpa + 4*study + u

reg score attend

reg score attend cgpa study, robust
```