Universidad de SanAndrés

PS3 Economía Aplicada

Elaborado por:

Casiano,Denys
Quispe, Anzony
Rigirozzi, Gonzalo
Sambrana, Gerónimo

Ejercicio 1

1. ¿Qué sucede con los errores estándar de los regresores si aumenta el tamaño muestral?

Cuadro 1: Resultado de las regresiones aumentando el tamaño muestral

	(1)	(2)
	wage	wage
education	0.250*	0.0151
	(0.0960)	(0.0314)
intelligence	2.982***	3.000***
	(0.0107)	(0.00348)
a	0.967***	1.004***
	(0.0514)	(0.0165)
b	2.096***	2.026***
	(0.0940)	(0.0312)
Constant	5.855***	6.279***
	(0.838)	(0.267)
Observations	100	1000

Standard errors in parentheses

* $p < 0{,}05$, ** $p < 0{,}01$, *** $p < 0{,}001$

Del Cuadro 1 se observa que el mayor tamaño muestral generó una reducción en el valor de los errores estándar. Adicionalmente, vemos que se estima mejor los coeficientes, puntualmente el de educación se aproxima mas a su verdadero valor, 0.

2. ¿Qué sucede con los errores estándar de los regresores si aumenta la varianza de μ (término de error)?

Cuadro 2: Resultado de las regresiones aumentando la varianza del término de error

	(1)	(2)
	wage	wage
education	0.250*	7.156*
	(0.0960)	(2.733)
intelligence	2.982***	2.516***
	(0.0107)	(0.304)
a	0.967***	-0.0147
	(0.0514)	(1.461)
b	2.096***	5.978*
	(0.0940)	(2.675)
Constant	5.855***	-19.42
	(0.838)	(23.85)
Observations	100	100

Standard errors in parentheses

* $p < 0{,}05$, ** $p < 0{,}01$, *** $p < 0{,}001$

Del Cuadro 2 se puede observar que con el aumento en la varianza del termino de error, los valores de los errores estándar aumentan de forma significativa. Otra consecuencia es que se estima los coeficientes de manera muy sesgada con respecto a su verdadero valor.

3. ¿Qué sucede con los errores estándar de los regresores si aumenta la varianza de X=Inteligencia?

Cuadro 3: Resultado de las regresiones aumentando la varianza de 20 a 50 en la variable X=Inteligencia

	(1)	(2)
	wage	wage
education	0.250*	0.219*
	(0.0960)	(0.0978)
intelligence	2.982***	2.981***
	(0.0107)	(0.00996)
a	0.967***	0.957***
	(0.0514)	(0.0520)
b	2.096***	2.084***
	(0.0940)	(0.0946)
Constant	5.855***	6.397***
	(0.838)	(0.706)
Observations	100	100

Standard errors in parentheses

* $p < 0{,}05$, ** $p < 0{,}01$, *** $p < 0{,}001$

Del Cuadro 3 vemos que los errores estándar del regresor intelligence se reducen significativamente al contar con mayor variabilidad. Esto se explica, en parte porque al tener mayor variabilidad en el regresor,se puede medir su efecto sobre y de una manera mas precisa.

4. El valor de la suma de los residuos.

Cuadro 4: Suma de los residuos

Variables	Sum of Residuals
Model 1	-1.30e-07
Model 2	9.31e-09

Lo que observamos es que, al aumentar la variabilidad en el regresor, al mismo tiempo se reduce la suma de los residuos.

5. ¿Son los residuos ortogonales a los regresores?

Cuadro 5: Cross-correlation table

Variables	Residuals	intelligence	a	b	education
Residuals	1.000				
intelligence	-0.000	1.000			
a	-0.000	-0.067	1.000		
b	-0.000	-0.108	0.029	1.000	
education	0.000	0.902	-0.062	-0.129	1.000

En el Cuadro 5 podemos ver la ortogonalidad vía la correlación de los errores con cada uno de los regresores. Efectivamente, al ser 0 para cada uno, se da ortogonalidad.

6. ¿Cómo afecta la alta multicolinealidad a la estimación de Y?

Cuadro 6: Cross-correlation table

Variables	Fitted values	Fitted values
Fitted values	1.000	
Fitted values	1.000	1.000

Del Cuadro 6 se puede observar que la correlación entre los $\hat{Y}$ estimados es perfecta, la introducción de un regresor altamente correlacionado con otro, no produce problemas al momento de estimar Y.

7. ¿Qué sucede si corren una regresión con un error no aleatorio en X? ¿Y si ese error fuera aleatorio?

   Del Cuadro 7 se puede observar que al incluir la variable educación que no está en el Proceso Generador de Datos y está altamente correlacionada con la inteligencia, los coeficientes y el error estándar de las variables a y b no cambian, pero el error estándar de la variable inteligencia cambia, y mucho. Vemos un sesgo extremadamente grande para la variable inteligencia, que se contagia a la variable educación, que también está exageradamente sesgada, al estar altamente correlacionada. Vemos un sesgo extremadamente grande para la variable inteligencia(aunque mayor que con un error no aleatorio), que se contagia a la variable educación, que también esta exageradamente sesgada, al estar altamente correlacionada.

8. ¿Qué sucede si corren una regresión con un error no aleatorio en Y? ¿Y si ese error fuera aleatorio?

   Del Cuadro 8 se puede observar que al agregar aleatoriedad a los errores en Y la variable educación dejar de ser significativa, mientras que el resto de variables explicativas mantienen su significancia estadística aunque modifican ligeramente el valor de los coeficientes.

Cuadro 7: Resultados de la regresión incorporando un no aleatoriedad y aleatoriedad en el error de X

	(1)	(2)	(3)
	wage	wage	wage
education	0.250*	24.68***	25.29***
	(0.0960)	(1.261)	(1.357)
intelligence	2.982***		
	(0.0107)		
a	0.967***	0.224	-0.630
	(0.0514)	(1.416)	(1.600)
b	2.096***	5.974*	-1.615
	(0.0940)	(2.964)	(3.290)
intelligencemed		-0.0842	-0.117
		(0.152)	(0.167)
Constant	5.855***	66.73**	112.4***
	(0.838)	(25.16)	(32.14)
Observations	100	100	100

Standard errors in parentheses

* $p < 0{,}05$, ** $p < 0{,}01$, *** $p < 0{,}001$

Cuadro 8: Resultados de la regresión incorporando un no aleatoriedad y aleatoriedad en el error de Y

	(1)	(2)	(3)
	wage	wagemed	wagemed
education	0.250*	0.250*	0.260
	(0.0960)	(0.0960)	(0.140)
intelligence	2.982***	2.982***	2.978***
	(0.0107)	(0.0107)	(0.0156)
a	0.967***	0.967***	0.905***
	(0.0514)	(0.0514)	(0.0750)
b	2.096***	2.096***	2.127***
	(0.0940)	(0.0940)	(0.137)
Constant	5.855***	13.86***	108.1***
	(0.838)	(0.838)	(1.224)
Observations	100	100	100

Standard errors in parentheses

* $p < 0{,}05$, ** $p < 0{,}01$, *** $p < 0{,}001$

Ejercicio 2

1. Si X_1 esta altamente correlacionada con X_2 y con X_3, y X_2 y X_3 tienen grandes efectos parciales en Y, Esperan que $\widehat{\beta}_1$ y $\widetilde{\beta}_1$ sean similares o distintos? Expliquen.
   En la primer regresión, al no incluir a las variables X_2 y X_3, estaríamos omitiendo controles importantes para estimar insesgada y consistentemente nuestro parámetro de interés. Por ende, si sabemos que existe correlación entre las variables, lo óptimo seria incluir ambas como en el segundo modelo propuesto, ya que de esa manera se estaría estimando de forma consistente el efecto de X_1, mas allá de generar mayor ineficiencia debida a un potencial problema de multicolinealidad, ya que esto ultimo es second order. Finalmente, por estas razones, esperamos que las dos estimaciones del parámetro de interés difieran.
   Adicionalmente, al no incluir estas variables y correr el primer modelo, estaríamos violando el supuesto de identificación de OLS, que es el de exogeneidad, por lo que ya no se puede decir que hayan propiedades atractivas en el estimador. Es importante notar que, como las variables X_2 y X_3 tienen grandes efectos en y, están en el verdadero DGP.

2. En el caso de que no exista correlación entre las variables que se omiten y la que estamos utilizando para estimar el parámetro, el sesgo de variable omitida ya no parece un gran problema para nosotros, no se genera sesgo. Sin embargo, las variables que no se incluyen van al error y, de esa forma, aumenta la varianza en el error, lo que lleva a una mayor ineficiencia.
   En este caso no esperaríamos que difieran mucho los parámetros en cuestión, si habría una diferencia en la dispersión de sus distribuciones.

3. Como en este caso la variable adicional no estaría en el verdadero DGP (o eso podemos pensar), al incluirla no deberíamos estar generando una inconsistencia, por lo que los parámetros en cuestión no deberían diferir mucho, lo que si puede estar ocurriendo es que se aumente la ineficiencia ya que el coeficiente para ese parámetro debería ser 0 pero el modelo lo estima como algo distinto de 0, y eso hace que se mueva junto con el error.

4. Al existir una gran correlación entre nuestra variable de interés y las omitidas, podríamos tener un problema de sesgo de variable omitida al no incluirlas por lo que, lo conveniente seria incluirlas. Sin embargo, al incluir dos controles muy correlacionados con el que ya teníamos, podríamos tener un problema

de multicolinealidad potencial, con lo cual se gana ineficiencia.
Los parámetros en cuestión van a diferir ya que, la regresión reducida estima un parámetro que es sesgado pero con menor varianza en comparación a la regresión mas completa que estima uno insesgado pero de mayor varianza. En este caso, a diferencia del primero, X_2 y X_3 no estarían en el verdadero DGP al explicar poco de Y.

5. Esperaríamos que $\widehat{\beta_1}$ tenga errores estándar mayores que en $\widetilde{\beta_1}$. A pesar de que ambas variables están en el verdadero DGP, al no estar correlacionadas con la variable de mi interés, no corro riesgo de sesgo al omitirlas, a pesar de que si aumente la variabilidad del error.
   Sin embargo, también se aumenta el error estándar de los estimadores al incluirlas, debido al potencial problema de multicolinealidad. En este caso tengo que, puedo incluirlas para reducir la varianza en el error pero, al mismo tiempo, estoy aumentando mi variabilidad al incluirlas.
   Suponiendo que el efecto de la multicolinealidad es mayor, los ee de $\widehat{\beta_1}$ serán mucho mayores.

6. En este caso, suponiendo que el consumo de chocolate no esta en el verdadero DGP y que no esta correlacionado con las variables de interés dentro del modelo, esperaríamos que los errores estándares del parámetro obtenido con la regresión que añade X_4 sean mucho mayores que los errores estándares del mismo parámetro en la regresión en la que X_4 se excluye, ya que a pesar de que no esta en el verdadero DGP, al estimar el parámetro de X_4 no obtendremos 0, y eso va a variar junto con el termino de error.

Statistics/Data analysis

```
*ECONOMIA APLICADA PROBLEM SET 3
*QUISPE, CASIANO, SAMBRANA, RIGIROZZI


************PUNTO 1***************
****AUMENTAMOS EL TAMAÑO MUESTRAL PARA VER QUE SUCEDE CON LOS ERRORES STANDARD
*Primero seteamos el modelo de base mostrado en clases.
clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

*notar la correlacion entre educacion e inteligencia.

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage education intelligence a b
eststo OLS1


**Definimos un nuevo tamanio muestral mayor, de 1000
clear
set obs 1000
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)


gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u

reg wage education intelligence a b
eststo OLS2
*Lo que se observa con este mayor tamanio muestral es que se da una reduccion en el va
> lor de los correspondientes errores estandar. Adicionalmente, vemos que se estima me
> jor los coeficientes, puntualmente el de educacion se aproxima aun mas a su verdader
> o valor, 0.

esttab OLS1 OLS2 using "ejercicio1.tex",  replace se label


*******************************************************************************
************PUNTO 2**************
***AUMENTAMOS LA VARIANZA DEL TERMINO DE ERROR

*Seteamos nuevamente el modelo original:
clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

*notar la correlacion entre educacion e inteligencia.

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u
```

```
reg wage education intelligence a b
eststo OLS3
*Luego proponemos la modificacion en la varianza del termino de error
clear

set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

gen education=int(intelligence/10+invnormal(uniform())*1)


gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*30+7)
gen wage=3*intelligence+a+2*b+u

reg wage education intelligence a b
eststo OLS4
*Con este aumento en la varianza del termino de error, vemos claramente que los valore
> s de los errores estandar aumentan de forma significativa. Otra consecuencia es que
> se estima los coeficientes de manera muy sesgada con respecto a su verdadero valor.

esttab OLS3 OLS4 using "ejercicio2.tex",  replace se label

*****************************************************************************
************PUNTO 3*******************

*Ahora tomamos un modelo que solo considere a la variable inteligencia como regreso
* Con varianza de X=20

clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

*notar la correlacion entre educacion e inteligencia.

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage education intelligence a b
eststo OLS5
*************************************
*CON VARIANZA X=50


clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*50+100)

*notar la correlacion entre educacion e inteligencia.

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u
```

```
reg wage education intelligence a b
eststo OLS6
*Vemos que los errores estandar del regresor intelligence se reducen significativament
> e al contar con mayor variabilidad. Esto se explica, en parte porque al tener mayor
> variabilidad en el regresor,se puede medir su efecto sobre y de una manera mas preci
> sa.

esttab OLS5 OLS6 using "ejercicio3.tex",  replace se label

********************************************************************************
******PUNTO 4**********
clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

*notar la correlacion entre educacion e inteligencia.

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage education intelligence a b

predict residuals, res
tabstat residuals, s (sum)

****veamos la suma de residuos con mayor variabilidad del regresor intelligence

clear

set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*50+100)

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage education intelligence a b


predict residuals, res
tabstat residuals, s (sum)
*Lo que observamos es que, al aumentar la variabilidad en el regresor, al mismo tiempo
>  se reduce la suma de los residuos.
*********************************************************************************
******PUNTO 5**********

clear all

set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

/* We set the standard error of this variable so the correlation between education and
>  intelligence is high (0.90 approximate).*/

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence
```

```
gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*30+7)
gen wage=3*intelligence+a+2*b+u

reg wage education intelligence a b


predict residuals, res
tabstat residuals, s (sum)

// instalar corrtex con el siguiente comando: ssc install corrtex
corrtex residuals intelligence a b education, file(ejercicio5)
*Vemos la ortogonalidad via la correlacion de los errores con cada uno de los regresor
> es. Efectivamente, al ser 0 para cada uno, se da ortogonalidad.

*******************************************************************************
******PUNTO 6**********

*probemos predecir sin la variable que genera multicolinealidad y con ella.
clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)


gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage intelligence a b
predict y_hat_1

reg wage education intelligence a b
predict y_hat_2
corrtex y_hat_1 y_hat_2, file(ejercicio6)
*Como la correlacion entre los y estimados es perfecta, la introduccion de un regresor
>  altamente correlacionado con otro, no produce problemas al momento de estimar y.


********************************************************************************
******PUNTO 7**********
clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

/* We set the standard error of this variable so the correlation between education and
>  intelligence is high (0.90 approximate).*/

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u
```

```
* Include education that is not in the Data Generating Process and it is highly correl
> ated with intelligence. Note that coefficients and SE for a and b do not change, but
>  the SE for the coefficient of intelligence changes, and a lot.
reg wage education intelligence a b
eststo OLS7
*planteamos el modelo en el que tenemos un regresor, inteligencia, con un error de med
> icion no aleatorio.

clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)
gen intelligencemed=int(invnormal(uniform())*20+100) + 4
/* We set the standard error of this variable so the correlation between education and
>  intelligence is high (0.90 approximate).*/

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage education intelligencemed a b
eststo OLS8
*Vemos un sesgo extremadamente grande para la variable inteligencia, que se contagia a
>  la variable educacion, que tambien esta exageradamente sesgada, al estar altamente
> correlacionada.



*planteamos el modelo en el que tenemos un regresor, inteligencia, con un error de med
> icion aleatorio.

clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)
gen intelligencemed=int(invnormal(uniform())*20+100) + int(invnormal(uniform())*1+2)
/* We set the standard error of this variable so the correlation between education and
>  intelligence is high (0.90 approximate).*/

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage education intelligencemed a b
eststo OLS9
*Vemos un sesgo extremadamente grande para la variable inteligencia(aunque mayor que c
> on un error no aleatorio), que se contagia a la variable educacion, que tambien esta
>  exageradamente sesgada, al estar altamente correlacionada.

esttab OLS7 OLS8 OLS9 using "ejercicio7.tex",  replace se label

********************************************************************************
******PUNTO 8**********
clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

/* We set the standard error of this variable so the correlation between education and
>  intelligence is high (0.90 approximate).*/
```

```
gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u


reg wage education intelligence a b
eststo OLS10

*suest ols11 ols12, robust
predict residuals
tabstat residuals,s(v mean)



*******************************
*errores en Y (no aleatorio)

clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)

/* We set the standard error of this variable so the correlation between education and
>  intelligence is high (0.90 approximate).*/

gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u
gen wagemed=3*intelligence+a+2*b+u + 8


reg wagemed education intelligence a b
eststo OLS11

predict residuals
tabstat residuals,s(v mean)

************************************

*errores en Y (aleatorio)
clear
set obs 100
set seed 1234
gen intelligence=int(invnormal(uniform())*20+100)


gen education=int(intelligence/10+invnormal(uniform())*1)
corr education intelligence

gen a=int(invnormal(uniform())*2+10)
gen b=int(invnormal(uniform())*1+5)
gen u=int(invnormal(uniform())*1+7)
gen wage=3*intelligence+a+2*b+u
gen wagemed=3*intelligence+a+2*b+u + int(uniform()*4+100)


reg wagemed education intelligence a b
eststo OLS12
predict residuals
tabstat residuals,s(v mean)

*Notar el aumento en la media de los residuos.
```

```
esttab OLS10 OLS11 OLS12 using "ejercicio8.tex",  replace se label
```