Universidad de SanAndrés

Economía Aplicada

Problem Set 3: Sesgo e Imprecisión

Ignacio Anchorena, Rodrigo Braga, Federico Lopez, Joaquin Musich

Fecha de entrega: 6 de septiembre de 2024

Ejercicio 1:

1. Si aumenta el tamaño de la muestra los errores estándar van a tender a disminuir. Esto lo podemos ver precisamente en la fórmula de la varianza del estimador beta, que a su vez es el error estándar al cuadrado:

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

   Aquí podemos observar la relación inversa entre la varianza del estimador con la cantidad de valores en la muestra. Ante un aumento de n, la varianza va a bajar y así también el arroz estándar.
   Otra explicación puede venir por la ley de los grandes números. Esto nos dice que a medida que el tamaño de la muestra va aumentando, los estimadores se aproximan cada vez más al verdadero valor, por lo que un aumento en la muestra tendería a reducir el error estándar.

	(1) wage	(2) wage
intelligence	3.004^{***}	3.006^{***}
	(0.00563)	(0.00400)
w	1.048^{***}	0.972^{***}
	(0.0308)	(0.0250)
z	1.953^{***}	1.881^{***}
	(0.101)	(0.0706)
_cons	5.962^{***}	6.688^{***}
	(0.736)	(0.540)
N	100	200

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

2. Partiendo de la misma fórmula del ejercicio anterior, inicialmente definimos la varianza del término de error (μ) como σ^2. Ante un aumento de la varianza del término de error, los errores estándar deberían de aumentar también.

	(1) wage	(2) wage
intelligence	3.004^{***}	3.033^{***}
	(0.00563)	(0.0368)
w	1.048^{***}	1.309^{***}
	(0.0308)	(0.202)
z	1.953^{***}	1.900^{**}
	(0.101)	(0.663)
_cons	5.962^{***}	1.762
	(0.736)	(4.806)
N	100	100

Standard errors in parentheses

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

3. Un aumento en la varianza de X lleva a que los errores estándar disminuyan. Analíticamente lo podemos ver en la fórmula de la varianza del estimador. Además, intuitivamente se puede pensar en que hay más variabilidad de la variable que intenta explicar a la variable dependiente, por lo que puede capturar más la variabilidad del Y.

	(1) wage	(2) wage
intelligence	3.004^{***}	3.002^{***}
	(0.00563)	(0.00282)
w	1.048^{***}	1.048^{***}
	(0.0308)	(0.0309)
z	1.953^{***}	1.953^{***}
	(0.101)	(0.102)
_cons	5.962^{***}	6.173^{***}
	(0.736)	(0.573)
N	100	100

Standard errors in parentheses

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

4. La suma de los residuos es casi cero.

5. Para saber si son ortogonales, la correlación entre ellos debería ser cero (o muy cercano a cero). Corriendo la regresión entre los residuos y los regresores obtenemos:

	(1) residuos
intelligence	-1.57e-10 (0.00282)
w	7.29e-10 (0.0309)
z	-1.57e-10 (0.102)
_cons	7.28e-09 (0.573)
N	100

Standard errors in parentheses

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

Como podemos ver la correlación es muy cercana a cero, por lo que los residuos son ortogonales.

6. Para que haya multicolinealidad tiene que haber alta correlación entre dos variables independientes. En la muestra generamos una variable que está compuesta por otra variable independiente, vamos a usarla para ver el efecto de la multicolinealidad en la estimación de Y.

	(1) wage	(2) wage
intelligence	3.004^{***} (0.00563)	2.994^{***} (0.0119)
w	1.048^{***} (0.0308)	1.045^{***} (0.0310)
z	1.953^{***} (0.101)	1.960^{***} (0.102)
education		0.109 (0.110)
_cons	5.962^{***}	5.966^{***}

	(0.736)	(0.736)
N	100	100

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

Lo que sucede es que, si bien no hay sesgo, el estimador puede eficiencia porque aumenta la varianza.

7. En el caso de que el error no sea aleatorio, es decir que es el mismo para todos los X, no hay un problema de consistencia del estimador. Esto sucede porque el error es siempre igual y no tiene varianza. Por otro lado, si el error es aleatorio vamos a tener un estimador con mayor varianza, lo que nos lleva a un problema de sesgo.

	(1) wage	(2) wage	(3) wage
intelligence	3.004^{***}		
	(0.00563)		
w	1.048^{***}	1.077^{***}	1.048^{***}
	(0.0308)	(0.0829)	(0.0308)
z	1.953^{***}	1.518^{***}	1.953^{***}
	(0.101)	(0.273)	(0.101)
intelligence_al_err		2.997^{***}	
		(0.0151)	
intelligence_esp_err			3.004^{***}
			(0.00563)
_cons	5.962^{***}	6.531^{**}	3.258^{***}
	(0.736)	(1.975)	(0.739)
N	100	100	100

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

8. Cuando el error en la medición de Y es aleatorio, es absorbido por el término de error, por lo que no vamos a tener un problema de sesgo en el estimador pero si aumenta su dispersión (varianza). Que aumente la varianza nos lleva a pensar que el estimador es más ineficiente.

   Por otro lado, si el error no es aleatorio y es generalizado en toda la muestra no nos va a generar ningún problema en nuestro estimador. Como podemos ver en la regresión, lo único que se ve afectado es la constante. Lo podemos pensar como un desplazamiento de todos los datos con la misma magnitud.

	(1) wage	(2) wage_al_err	(3) wage_esp_err
intelligence	3.004^{***}	3.000^{***}	3.004^{***}
	(0.00563)	(0.00718)	(0.00563)
w	1.048^{***}	1.042^{***}	1.048^{***}
	(0.0308)	(0.0394)	(0.0308)
z	1.953^{***}	1.894^{***}	1.953^{***}
	(0.101)	(0.130)	(0.101)
_cons	5.962^{***}	7.284^{***}	6.862^{***}
	(0.736)	(0.939)	(0.736)
N	100	100	100

Standard errors in parentheses

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

Ejercicio 2

1. Esperamos que los estimadores sean distintos porque vamos a tener un sesgo por omisión de variables relevantes. Al estimar $\tilde{\beta}_1$, solo estamos considerando la asistencia a clases como variable explicativa de la nota del examen y estamos omitiendo las variables promedio del alumno y cantidad de horas de estudio. Como éstas dos variables que estamos omitiendo son relevantes para explicar la nota (tienen mucha relación con Y) y, al mismo tiempo, la asistencia a clases está altamente correlacionada con ambas variables, se dan las condiciones para que nuestro estimador $\tilde{\beta}_1$ sea sesgado y, por ende, será distinto a $\hat{\beta}_1$.

2. En este caso, como las variables omitidas no están correlacionadas con mi regresor de interés (la asistencia a clases), la estimación $\tilde{\beta}_1$ resultará insesgada y $\tilde{\beta}_1$ y $\hat{\beta}_1$ serán similares. Es decir, no se cumple una de las dos condiciones que deben de cumplirse en simultáneo para que se produzca un sesgo por omisión de variables relevantes.
3. Al incluir una variable irrelevante (el consumo de chocolate), en términos de sesgo no habrá inconvenientes porque no estaremos omitiendo nada y, por ende, $\dot{\beta}_1$ será similar a $\hat{\beta}_1$. Sin embargo, la inclusión de variables irrelevantes puede provocar que aumente la varianza de nuestro estimador si esta variable está relacionada con el regreso de interés. Esto se puede ver en la siguiente fórmula:

$$V(\dot{\beta}_1) = \frac{\sigma^2}{n(1 - R_j^2)V(X_j)}$$

El R_j^2 mide la multicolinealidad de las variables explicativas y es creciente en el número de ellas. Por lo tanto, al incorporar el consumo de chocolate del alumno, el R_j^2 aumenta si el consumo de chocolate está relacionado con la asistencia a clases y por esto $V(\dot{\beta}_1)$ aumenta.
4. Ocurre algo parecido a lo que ocurre en el apartado 1; esperamos que $\tilde{\beta}_1$ y $\hat{\beta}_1$ sean distintos, pero no tanto como antes. Lo que sucede ahora es que las variables omitidas en este caso no son tan relevantes como antes para explicar la nota (tienen poca relación), por lo que la magnitud del sesgo será mucho menor.
5. En primer lugar, cuando estimamos $\tilde{\beta}_1$ solo estamos incluyendo un solo regresor, por lo que su varianza será:

$$V(\tilde{\beta}_1) = \frac{\sigma^2}{nV(X_1)}$$

En cambio, como $\hat{\beta}_1$ es el coeficiente de la asistencia de la regresión de Y en X1, X2 y X3 su varianza será:

$$V(\hat{\beta}_1) = \frac{\sigma^2}{n(1 - R_1^2)V(X_1)}$$

Entonces, es de esperar que mientras más correlacionadas estén las variables X_2 y X_3 con X_1, mayor sea la varianza y el desvío estándar de $\hat{\beta}_1$ por sobre $\tilde{\beta}_1$ (ya que el R_1^2 será mayor). Ahora bien, como la consigna nos dice que X1 está incorrelacionada con X2 y X3, el desvío estándar de $\tilde{\beta}_1$ y $\hat{\beta}_1$ será similar.

6. Para ver la relación entre los errores estándar de $\dot{\beta}_1$ y $\hat{\beta}_1$ podemos hacer 2 supuestos. El primero es que el consumo de chocolate no esté correlacionado con la asistencia a clases; si esto es así el R_1^2 de $\dot{\beta}_1$ y $\hat{\beta}_1$ será igual y por ende los errores estándar serán similares. En cambio, si suponemos que el consumo de chocolate si está relacionado con la asistencia a clases, entonces el R_1^2 de $\dot{\beta}_1$ será mayor que el de $\hat{\beta}_1$ y obtendremos como resultado un aumento de la varianza y del error estándar del estimador $\dot{\beta}_1$. Es decir, al incluir esta variable correlacionada, estamos incorporando una fuente de imprecisión que hace que el error estándar de $\dot{\beta}_1$ sea mayor al de $\hat{\beta}_1$.

```
/****************************************************************************
                          Semana 3: Problem Set 2

                          Universidad de San Andrés
                              Economía Aplicada
                                     2024
                        Integrantes: Federico Ariel Lopez
                                     Rodrigo Braga
                                     Joaquin Musich
                                     Ignacio Anchorena
****************************************************************************/
clear all
gl main "/Users/ignacioanchorena/Desktop/Eco Aplicada/PS3" // No vamos a usar ninguna base de datos para esta parte del trabajo

gl input "$main/input"
gl output "$main/output"

clear

set obs 100
set seed 1233 // seteamos la semilla asi las 100 obs son siempre iguales

*Ahora genero las variables que voy a usar para la parte 1 del problem set

gen intelligence=int(rnormal(100,20)) // Primero la variable intelligence

gen education=int(intelligence/10+rnormal(0,1)) // Depues ponemos la variable education que va a depender de intelligence

gen w=int(rnormal(10,3)) // Generamos una variable que se llama w

gen z=int(rnormal(5,1)) // Generamos una variable que se llama z

gen u=int(rnormal(7,1)) // Por último, generamos la variable u que será el termino error

gen wage=3*intelligence+w+2*z+u // Wage va a ser nuestar variable explicada

* Ahora corremos una regresion con wage como variable dependiente de w y z para guardar y comparar con los futuros resultados

reg wage intelligence w z
predict wage_hat_1 // Guardamos wage estiamdo
est store ols1

////// Inciso 1 //////

//Para este inciso vamos a hacer lo mismo que antes pero esta vez con el doble de observaciones. Para esto vamos a utilizar la misma seed y las variables estarán conformadas como antes. Corremos la misma regresión y vemos las diferencias en los desviós estandar

clear
set obs 200
set seed 1233

gen intelligence=int(rnormal(100,20))
gen education=int(intelligence/10+rnormal(0,1))
corr education intelligence

gen w=int(rnormal(10,3))
gen z=int(rnormal(5,1))
gen u=int(rnormal(7,1))
gen wage=3*intelligence+w+2*z+u
```

```

reg wage intelligence w z
predict y_hat_2
est store ols2

esttab ols1 ols2 using "$output/Tabla1.rtf", se replace

////// Inciso 2 //////

// Para este inciso vamos a contruir la mismsa regresion que al principio pero a la variable "u" le vamos a aumentar su varianza.
clear
set obs 100
set seed 1233

gen intelligence=int(rnormal(100,20))
gen education=int(intelligence/10+rnormal(0,1))
corr education intelligence

gen w=int(rnormal(10,3))
gen z=int(rnormal(5,1))
gen u=int(rnormal(7,7))
gen wage=3*intelligence+w+2*z+u

reg wage intelligence w z
predict mu_hat, residuals // Guardamos los residuos
generate sumatoria_mu_hat = round(sum(mu_hat),0)

est store ols3

esttab ols1 ols3 using "$output/Tabla2.rtf", se replace

////// Inciso 3 //////

// Para este inciso vamos a contruir la misma regresion que al principio pero a la variable intelligence la vamos a multiplicar por 40 en vez de 20 para aumentar su varianza

clear
set obs 100
set seed 1233
gen intelligence=int(rnormal(100,40))
gen education=int(intelligence/10+rnormal(0,1))
corr education intelligence

gen w=int(rnormal(10,3))
gen z=int(rnormal(5,1))
gen u=int(rnormal(7,1))

gen wage=3*intelligence+w+2*z+u
reg wage intelligence w z
predict y_hat_3

est store ols4

esttab ols1 ols4 using "$output/Tabla3.rtf", se replace

////// Inciso 4 //////

// Usando el modelo del inciso 3, calculamos los residuos
predict residuos, residuals
total residuos

////// Inciso 5 //////
```

```
// Para saber si los residuos son ortogonales a los regresores, corremos una regresión entre los residuos y los regresores. Si fuesen ortogonales, la correlación debería dar cero.

reg residuos intelligence w z
est store ols5

esttab ols5 using "$output/Tabla5.rtf", se replace


////// Inciso 6 //////
clear

set obs 100
set seed 1233

gen intelligence=int(rnormal(100,20))
gen education=int(intelligence/10+rnormal(0,1))
gen w=int(rnormal(10,3))
gen z=int(rnormal(5,1))
gen u=int(rnormal(7,1))
gen wage=3*intelligence+w+2*z+u


reg wage intelligence w z
predict wage_hat_1
est store ols1

reg wage education intelligence w z
predict y_hat_6
est store ols6
esttab ols1 ols6 using "$output/Tabla6.rtf", se replace

////// Inciso 7 //////

// Vamos a generar una variable donde tenga error aleatorio y otra donde el error esté en un dato en especifico.

gen intelligence_al_err = intelligence+int(invnormal(uniform())*1+1) // variable con error aleatorio

gen intelligence_esp_err = intelligence+0.9 // Variable con error especifico

eststo: reg wage intelligence w z
eststo: reg wage intelligence_al_err w z
eststo: reg wage intelligence_esp_err w z

esttab using "$output/Tabla7.rtf", se replace

eststo clear
////// Inciso 8 //////

// Ahora el problema va a estar en la variable Y, el salario. Vamos a usar el mismo procedimiento que el inciso 7, pero vamos a correr la regresión con la variable intelligence correcta.

gen wage_al_err=wage+int(invnormal(uniform())*1+1)

gen wage_esp_err=wage+0.9

eststo: reg wage intelligence w z
eststo: reg wage_al_err intelligence w z
eststo: reg wage_esp_err intelligence w z

esttab using "$output/Tabla8.rtf", se replace
```