ECONOMETRIA

II Semestre 2019

Profesores: Fernando Díaz - Víctor Macías - Benjamín Villena -
Ayudantes: Catalina Pérez-García - Sebastián González - Juan Felipe Ly- Mauricio Vásquez

> "The fundamental cause of the trouble is that in the modern world the stupid are cocksure while the intelligent are full of doubt."
>
> (Bertrand Russell)

Pregunta 1, 25 puntos

Un investigador esta interesado en determinar el efecto de una madre fumadora en la salud de los niños recién nacidos. Dado que no tiene un indicador de la salud de los niños, decide utilizar el peso en gramos del niño al nacer (bweight) como variable dependiente. Como variable explicativa, utiliza la cantidad diaria promedio de cigarillos que fumó la madre durante su embarazo (cigs). Como controles, utiliza la educación de la madre (meduc), la educación del padre (feduc), una variable dummy (male) que toma el valor de 1 si el bebé es hombre, la edad de la madre (mage), y la edad del padre (fage). La regresión que estima el investigador es la siguiente:

$$bwght_i = \beta_1 + \beta_2 cigs_i + \beta_3 feduc_i + \beta_4 meduc_i + \beta_5 male_i + \beta_6 fage_i + \beta_7 mage_i + \beta_8 mage_i^2 + \mu_i \quad (1)$$

donde $mage_i^2$ corresponde al cuadrado de los años de educación de la madre. La estadística de descriptiva de estas variables se presenta en la tabla 1.

Table 1: Estadística Descriptiva

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
bwght	1,832	3,401.122	576.544	360	3,076	3,770	5,204
cigs	1,722	1.089	4.222	0.000	0.000	0.000	40.000
feduc	1,785	13.915	2.266	3.000	12.000	16.000	17.000
meduc	1,802	13.718	2.092	3.000	12.000	16.000	17.000
male	1,832	0.514	0.500	0	0	1	1
fage	1,826	31.919	5.713	18.000	28.000	35.000	64.000
mage	1,832	29.558	4.771	16	26	33	44

1. (5 puntos) Explique la intuición detrás de las variables explicativas del modelo. ¿Cuál es la intuición respecto de los controles incluidos? ¿Cuáles son los signos esperados de los coeficientes de los regresores? En particular, refiérase a la lógica de la inclusión de la edad de los padres y del cuadrado de la edad de la madre?
2. (5 puntos) ¿Cuál es la lógica de incluir un término cuadrático para la edad de la madre? Por otro lado, ¿no le parece a usted que debería incluirse también el cuadrado de la edad del padre como variable explicativa. Explique sus respuestas.

Los resultados de estimar la regresión (1) por OLS se presentan en las tablas 2 y 3.

Table 2: Estimación por OLS

	Dependent variable:
	bwght
cigs	-9.109***
	(3.334)
feduc	9.017
	(7.572)
meduc	-4.014
	(8.450)
male	89.780***
	(27.618)
fage	6.016*
	(3.397)
mage	64.052**
	(27.397)
magesq	-1.101**
	(0.454)
Constant	2,203.053***
	(396.049)
Observations	1,672
R^2	0.020
Adjusted R^2	0.016
Residual Std. Error	562.998 (df = 1664)
F Statistic	4.874*** (df = 7; 1664)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Matriz de Varianzas y Covarianzas

	(Intercept)	cigs	feduc	meduc	male	fage	mage	magesq
(Intercept)	156,855.200	-93.206	-177.227	-59.219	-347.676	-29.611	-10,343.080	171.352
cigs	-93.206	11.116	2.525	1.587	0.576	-0.164	1.919	-0.031
feduc	-177.227	2.525	57.338	-34.441	4.805	-1.538	-6.442	0.098
meduc	-59.219	1.587	-34.441	71.406	-12.399	0.859	-25.077	0.308
male	-347.676	0.576	4.805	-12.399	762.773	-3.026	4.153	0.035
fage	-29.611	-0.164	-1.538	0.859	-3.026	11.540	-12.748	0.056
mage	-10,343.080	1.919	-6.442	-25.077	4.153	-12.748	750.578	-12.283
magesq	171.352	-0.031	0.098	0.308	0.035	0.056	-12.283	0.206

3. (10 puntos) Respecto de la estimación de la ecuación (1), ¿tienen los coeficiente estimados los signos esperados? ¿Esperaba usted que la edad de la madre tuviese un efecto positivo en el peso del niño al nacer? ¿Esperaba usted que el coeficiente del cuadrado de la edad de la madre tuviese signo negativo? Intente explicar este resultado. Apoye su respuesta con la información presentada en la tabla 1.

4. (5 puntos) Construya un intervalo de confianza al 95% de certeza para el coeficiente estimado de la variable *cigs*.

Solo alumnos de la sección del profesor F. Díaz, 25 puntos

El investigador está preocupado por la falta de significancia de la educación de los padres. Desde un punto de vista teórico, padres más educados deberían tender a preocuparse más por sus hijos durante el embarazo, acudir con mayor frecuencia a controles, etc..Un colega le sugiere que la falta de signficancia de las variables relacionadas con la educación se puede deber a un problema de multicolinealidad. De acuerdo a los argumentos de su colega, hombres y mujeres más educados tienden a relacionarse con hombres y mujeres más educados.

5. (5 puntos) De los resultados presentados en la tabla 2, ¿le parece a usted que pueda existir un problema de multicolinealidad en la estimación de la ecuación (1)? Explique su respuesta.

Para analizar más en detalle si existe un problema de multicolinealidad en sus estimaciones, el investigador re-estima el modelo, pero excluyendo la educación de la madre. Además, estima una regresión auxiliar donde utiliza como variable dependiente la educación del padre y el resto de las variables utilizadas en la estimación de la ecuación (1) como variables explicativas. Los resultados de ambas estimaciones se presentan en la tabla 4.

6. (10 puntos) ¿Cuál es la lógica, en términos del análisis de la posible existencia de multicolinealidad, de las especificaciones presentadas la tabla 4? Si efectivamente existiese un problema derivado de una alta correlación de las variables feduc y meduc, ¿qué resultados esperaría usted observar en la tabla 4?

7. (10 puntos) De acuerdo a los resultados de la tabla 4, ¿qué se puede concluir respecto de la existencia de multicolinealidad en el modelo? Explique claramente su respuesta.

Table 4: Regresiones Auxiliares

	Dependent variable:	
	bwght	feduc
	(1)	(2)
meduc		0.601***
		(0.023)
cigs	-9.030***	-0.044***
	(3.328)	(0.011)
feduc	6.997	
	(6.376)	
male	88.166***	-0.084
	(27.542)	(0.089)
fage	6.059*	0.027**
	(3.395)	(0.011)
mage	62.803**	0.112
	(27.222)	(0.089)
magesq	-1.087**	-0.002
	(0.452)	(0.001)
Constant	2,199.728***	3.091**
	(395.798)	(1.280)
Observations	1,676	1,672
R^2	0.020	0.360
Adjusted R^2	0.016	0.358
Residual Std. Error	562.923 (df = 1669)	1.822 (df = 1665)
F Statistic	5.617*** (df = 6; 1669)	156.106*** (df = 6; 1665)
Note:		*p<0.1; **p<0.05; ***p<0.01

Pregunta 2, 25 puntos

Un economista estima ecuaciones de Mincer que explican el logaritmo natural del salario por hora, **log(sal/hora)**, en función de los años de **escolaridad** y la experiencia potencial (**exper**). A su vez, la experiencia potencial se define como exper = edad − escolaridad − 6. Para distintos modelos de regresión, los resultados que se obtienen a partir de la encuesta de Caracterización Socioeconómica Nacional 2017, CASEN 2017, son los siguientes:

	(1)	(2)	(3)
Variables	log(sal/hora)	log(sal/hora)	log(sal/hora)
escolaridad	0.0898	0.110	-0.0986
	(0.000632)	(0.000733)	(0.00267)
escolaridad^2			0.00903
			(0.000112)
exper		0.0136	0.0220
		(0.000525)	(0.000510)
exper^2		-6.64e-05	-0.000281
		(9.74e-06)	(9.63e-06)
Intercepto	8.054	7.545	8.585
	(0.00801)	(0.0120)	(0.0172)
Observaciones	61,421	61,421	61,421
R^2	0.248	0.285	0.353

Nota: Desviaciones estándar entre paréntesis bajo los coeficientes correspondientes.

Respecto a estos antecedentes, responda las siguientes preguntas

1. (6 p) Para los modelos (1) y (2) interprete el significado económico del coeficiente asociado a la escolaridad. Considere la presencia o ausenceia de otros regresores en la ecuación.

   Respuesta: En el modelo (1), el coeficiente de escolaridad es el retorno marginal de la escolaridad, es decir, el incremento porcentual esperado que tendría alguien que aumenta su escolaridad en un año incondicional, es decir, promediando cualquier otra característica de la persona. Formalmente esto es $\frac{\partial E[\log(sal/hora)]}{\partial \text{escolaridad}}$.

   En el modelo (2), la interpretación es el incremento porcentual que tendría alguien que aumenta su escolaridad en un año, manteniendo su nivel de experiencia potencial constante. Formalmente esto es $\frac{\partial E[\log(sal/hora)|exper]}{\partial \text{escolaridad}}$

2. (7 p) En el modelo (1), construya un test para la hipótesis nula de que el retorno marginal de la escolaridad es 0.1 al 5% de significancia. Escriba la regla de decisión usada, distribución relevante y grados de libertad asociado. Indique cómo calcularía el valor de significancia exacto de este test o valor p.

   Respuesta:

   Bajo supuesto de normalidad de los errores (o alternativamente, como la muestra es grande, los errores son aproximadamente normales), el estadístico de prueba tiene distribución t-Student con $N - K = 61421 - 2 = 61419$ grados de libertad. Esta distribución es prácticamente igual a una normal estándar. La hipótesis nula sería $H_0 : \beta_1 = 0.1$ y la alternativa sería $H_1 : \beta_1 \neq 0.1$. El enunciado no especifica una hipótesis alternativa más específica, lo que da origen a un test de dos colas.

$$t_c = \frac{|\hat{\beta}_1 - 0.1|}{\hat{DS}(\hat{\beta}_1)} = \frac{|0.0898 - 0.1|}{\sqrt{0.000632}} = 0.405$$

Este valor es menor que el valor crítico $t_{N-K}(0.975) \approx \Phi(0.975) = 1.96$, donde $\Phi(z)$ es la función de distribución acumulada de una normal estándar. Por esto, la hipótesis nula NO se puede rechazar.

Para calcular el valor p del test, es decir, el nivel de significancia que nos haría estar indiferentes entre rechazar o no rechazar H_0, se debe considerar que éste es un test de dos colas, y por ello el valor debe ser el área bajo la curva de la función de densidad de probabilidad (fdp) de la normal desde $-\infty$ hasta -0.405 más el área desde 0.405 hasta $+\infty$. Si $\Phi(z)$ es la función de distribución acumulada de la normal estándar (o la t-Student con $N - K = 61419$), esto se puede calcular como $\Phi(-0.405) + 1 - \Phi(0.405)$, lo que también se puede calcular como $1 - 2\Phi(-0.405)$, o bien, $2\Phi(0.405) - 1$, debido a la simetría de la distribución normal (o t-Student).

3. (7 p) Construya un test para la hipótesis de que la experiencia potencial es una variable relevante para explicar a los salarios por hora, a partir del modelo (2). Escriba la regla de decisión usada, distribución relevante y grados de libertad asociado. (Pista: la experiencia afecta a través de dos regresores).

Respuesta: Con la información disponible la única forma de responder esto es a través del test de R^2 libre versus restringido. El modelo (2) considera experiencia potencial lineal y cuadrática, por lo que esta variable es irrelevante bajo una $H_0 : \beta_2 = \beta_3 = 0$, donde β_2 es el coeficiente asociado a exper y β_3 es el ecoeficiente asociado a exper^2 en el modelo (2), que es el modelo libre o no restringido. El modelo restringido, que no considera la variable exper, es el modelo (1). Por lo tanto, se formula un estadístico de prueba

$$F_c = \frac{\frac{R^2_{libre} - R^2_{restr}}{\text{Num restricciones}}}{\frac{R^2_{libre}}{N-K}} = \frac{\frac{0.285-0.248}{2}}{\frac{0.285}{61421-4}} = 3986.7$$

Este valor se compara con el valor crítico de una tabla F de Fischer con grados de libertad (2, 61417) para numerador y denominador, respectivamente. Como el valor calculado es muy alto, la hipótesis nula se rechaza a casi cualquier nivel de significancia.

4. (5 p) En función de los resultados del modelo (3), determine el efecto esperado de un año de escolaridad adicional si usted se plantea hacer un postgrado (pasando del año 17 al 18, por ejemplo) suponiendo que su experiencia potencial NO se ve afectada por esta decisión.

Respuesta: El efecto marginal de un año de escolaridad adicional en el modelo (3) corresponde a

$$\frac{\partial E[\log(sal/hora)|exper]}{\partial \text{escolaridad}} = -0.0986 + 2 \times 0.00903 \times \text{escol}$$

Para $escol = 17$, el efecto esperado de incremental un año de escolaridad, manteniendo experiencia potencial constante sería $-0.0986+2\times0.00903\times17 = 0.2084$, vale decir, un año extra de escolaridad al pasar del año 17 al 18 eleva el salario por hora en 20.84%.

Sólo alumnos profesor Benjamín Villena

(25 p) El Ministerio de Educación está estudiando en muestras de cursos de colegios $c = 1, 2, ..., C$, en dos regiones distintas, si el porcentaje de apoderados desempleados de un curso de del colegio, D, afecta negativamente el rendimiento escolar de sus alumnos, R lo que es medido a través de un puntaje de la prueba SIMCE. Naturalmente existen factores desconocidos U que pueden afectar el desempeño escolar de los alumnos, que no son incluidos en el siguiente modelo:

$$R_c = \alpha_0 + \alpha_1 D_c + U_c$$

1. (6 p) Si el modelo anterior se estima por MCO para cada uno de los cursos de colegios de cada región (A y B) separadamente, ¿Cuál será la suma de los errores muestrales? ¿Qué puede decir de la suma de errores poblacionales?

   Respuesta: Dado que el modelo tiene intercepto, la suma de los errores muestrales en ambas regiones será exactamente cero, es decir $\sum_{n=1}^{N} \hat{U}_c = 0$ Por otro lado, la suma de errores poblacionales, $\sum_{n=1}^{C} U_c$ no tiene por qué ser exactamente cero. Sin embargo, se supone usualmente que $E[U|D] = 0$.

2. (6 p) Suponga que en la región A, la varianza de puntaje SIMCE entre cursos de los colegios es alta, y la varianza del porcentaje de apoderados desempleados entre cursos de colegios, es baja. En la región B, se observa exactamente el patrón contrario. Si los tamaños de muestra son muy similares, ¿En qué región posiblemente se obtendrá una estimación más precisa de α_1?

   Respuesta: La varianza de los estimadores en cada región serían los siguientes:

   $$Var(\hat{\alpha}_{1A}) = \frac{\sigma_A^2}{CVar(D_A)}$$

   y

   $$Var(\hat{\alpha}_{1B}) = \frac{\sigma_B^2}{CVar(D_B)}$$

   Si la varianza de la variable dependiente es alta en la región A, entonces $Var(R_A) > Var(R_B)$ lo que sugiere que $\sigma_A^2 > \sigma_B^2$. Por otro lado, se nos indica que $Var(D_A) < Var(D_B)$ por lo que la varianza del estimador MCO será más baja en la región B.

3. (7 p) Un asesor de un parlamentario, viendo resultados de $\hat{\alpha}_1$ negativos en ambas regiones, señala que "es indudable que el desempleo de los padres es la causa del bajo desempeño escolar. Si el gobierno interviniera para reducir el desempleo, los resultados del SIMCE subirían". ¿Qué opina usted de estas afirmaciones?

   Respuesta: La relación estadística negativa indica correlación, y no necesariamente causalidad. Podría ser que exista una asociación a una causa común, por ejemplo, el nivel educacional de los padres, que puede estar causando simultáneamente menor rendimiento escolar y mayor desempleo. Esto implicaría que el supuesto $E[U|D] = 0$ no sería correcto. Por lo tanto, una forma de decir que es posible que este modelo no describa una relación causal es decir que $E[U|D] \neq 0$.

4. (6 p) Otro asesor cree que los valores obtenidos no son suficientemente negativos y se prepone mostrar que la evidencia está de su lado. Para ello, separa la muestra por comunas dentro de las regiones, y estima el modelo descrito separadamente para colegios municipales y subvencionados, y por cada nivel escolar (1, 2, 3 básico, etc). Finalmente el analista encuentra que el efecto del desempleo de apoderados es tres veces más negativo al considerar colegios de comunas cercanas al borde

costero, en establecimientos municipales, para apoderados de 6 básico, ratificando aparentemente sus planteamientos. ¿Qué tan creíbles son estos resultados? Explique su respuesta.

Respuesta: El asesor incurre en la práctica de manipulación de valor p, o "p-hacking", visto en clases, que consiste en buscar submuestras para las cuales se rechaza alguna hipótesis de interés. Como los tests de hipótesis siempre admiten un porcentaje de error tipo I, si hay múltiples submuestras, es natural que encontremos algún rechazo. Por esto, la evidencia no es fiable.

Pregunta 3, 25 puntos

Considere el siguiente modelo de regresión lineal:

$$y_i = \beta_1 + u_i$$

$\forall \quad i = 1, 2, ..., n$
donde $E(u_i) = 0$, $E(u_i^2) = \sigma_u^2$, $E(u_i u_j) = 0 \quad \forall \quad i \neq j$

1. **(10 puntos)** Muestre que el estimador de MCO de β_1 es insesgado
2. **(5 puntos)** Considere el siguiente estimador alternativo de β_1:

$$\hat{\beta}_1 = \frac{n\overline{Y}}{n+1}$$

   donde n es el tamaño de la muestra y $\overline{Y}$ es la media muestral de Y. ¿Es este estimador insesgado?
3. **(10 puntos)** Calcule la varianza del estimador de MCO y del estimador alternativo. ¿Cuál tiene menor varianza?

Solución:

1. $\hat{\beta}_1^{MCO} = (X'X)^{-1}X'Y = \frac{1}{n}\sum_{i=1}^n y_i$
$E(\hat{\beta}_1^{MCO}) = \frac{1}{n}E(\sum_{i=1}^n(\beta_1+u_i)) = \beta_1+\frac{1}{n}\sum_{i=1}^n E(u_i) = \beta_1$, porque $E(u_i) = 0$. Por lo tanto, el estimador de MCO de β_1 es insesgado.

2.

$$E(\hat{\beta}_1) = E(\frac{n\overline{Y}}{n+1}) = \frac{n}{n+1}\frac{1}{n}E(\sum_{i=1}^n y_i) = \frac{1}{n+1}E(\sum_{i=1}^n(\beta_1+u_i)) = \frac{n\beta_1}{n+1} + \frac{1}{n+1}\sum_{i=1}^n E(u_i) = \frac{n\beta_1}{n+1}$$

porque $E(u_i) = 0$ y, por lo tanto, el estimador alternativo es sesgado, ya que $E(\hat{\beta}_1) \neq \beta_1$

3. $var(\hat{\beta}_1^{MCO}) = E(\hat{\beta}_1^{MCO} - E(\hat{\beta}_1^{MCO}))^2 = E(\hat{\beta}_1^{MCO} - \beta_1)^2 = E(\frac{1}{n}\sum_{i=1}^n u_i)^2 = \frac{\sigma_u^2}{n}$, porque $E(u_i u_j) = 0$ $\forall \quad i \neq j$
$var(\hat{\beta}_1) = E(\hat{\beta}_1 - E(\hat{\beta}_1))^2 = E(\frac{\sum_{i=1}^n y_i}{n+1} - \frac{n\beta_1}{n+1})^2 = (\frac{1}{n+1})^2 E(\sum_{i=1}^n(y_i - \beta_1))^2 = (\frac{1}{n+1})^2 E(\sum_{i=1}^n u_i)^2 = \frac{n\sigma_u^2}{(n+1)^2}$, porque $E(u_i u_j) = 0 \quad \forall \quad i \neq j$
Dado que $\frac{n}{(n+1)^2} < \frac{1}{n}$, entonces $var(\hat{\beta}_1^{MCO}) > var(\hat{\beta}_1)$ y, por lo tanto, el estimador alternativo tiene una varianza menor que el estimador de MCO.