DEPARTAMENTO DE INGENIERÍA INDUSTRIAL Y DE SISTEMAS
FACULTAD DE INGENIERÍA
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

ECONOMETRÍA APLICADA

Profesora: Javiera Vásquez

Pauta Interrogación Nº1

Jueves 30 de marzo de 2017

TIEMPO: 90 minutos

I. COMENTES (35 puntos)

Para cada una de las siguientes afirmaciones indique si es Falsa, Verdadera o Incierta. Siempre debe justificar su respuesta.

1. Los modelos de regresión lineal sólo pueden ser estimados por Mínimos Cuadrados Ordinarios. Comente. (5 puntos)

Falso, un estimador es simplemente un método o fórmula que nos dice como aproximar un parámetro poblacional a partir de una muestra, podríamos plantear muchos estimadores distintos para encontrar los coeficientes estimados en un modelo de regresión lineal, en efecto, existen otros métodos como GMM y Máxima Verosimilitud que son aplicados en modelos de regresión lineal. Pero, MCO sólo se puede utilizar en modelos de regresión lineal.

2. El estimador de MCO es una variable aleatoria, por lo tanto, tiene una distribución de probabilidad, media y varianza. Comente. (5 puntos).

Verdadero, el estimador MCO se puede expresar de la siguiente manera:

$$\hat{\beta} = \beta + (X'X)^{-1}X'u$$

Donde podemos ver claramente que depende del error, que es una variable aleatoria. Así, si por ejemplo, el error tiene una distribución normal, $\hat{\beta}$ también tendrá una distribución normal. La media de $\hat{\beta}$ es β (bajo los 10 supuestos) y la varianza es $\sigma^2(X'X)^{-1}$.

3. Cuando utilizamos el modelo de regresión para estudiar la relación entre dos variables, estamos haciendo lo mismo que estudiar la correlación entre estas dos variables. La única diferencia es que en el modelo de regresión podemos "controlar" por otras variables, incorporándolas como variables explicativas en el modelo. Comente. (5 puntos).

Falso, en el modelo de regresión no estudiamos una correlación si no la relación de dependencia de la variable Y con respecto a la(s) variable(s) X. En efecto, lo que nos

interesa predecir con un modelo de regresión lineal es el efecto causal de una variable explicativa sobre la variable dependiente, todo lo demás constante. Para esto estimamos el efecto marginal de una variable explicativa sobre el valor esperado de la variable dependiente, todo lo demás constante, lo que está representado a través de los β.

4. En un modelo de regresión lineal donde no existen variables explicativas, sino sólo la constante del modelo:

$$y_i = \beta_0 + u_i$$

El estimador MCO de β_0 es igual a la media muestral de y. Comente. (5 puntos).

Verdadero, el rol de la constante es que la recta de regresión pase por los promedios, sino hay variables explicativas, la recta no tendrá pendiente, y para pasar por los promedios, la constante debe ser igual al promedio de y.

En efecto, el problema de minimizar la suma de los errores al cuadrado en este caso es:

$$\min_{\hat{\beta}_o} \sum (y_i - \hat{\beta}_o)^2$$

De la CPO:

$$-2\sum (y_i - \hat{\beta}_o) = 0$$

Se obtiene que $\hat{\beta}_o = \bar{y}$.

5. Un modelo de regresión donde se incluyen variables explicativas en potencias (x^2, x^3, etc.) no puede ser estimado por MCO, ya que no es lineal. Comente. (5 puntos).

Falso, para que el modelo de regresión sea estimable por MCO, tiene que ser lineal en los coeficientes, pero las variables explicativas pueden estar transformadas de manera de capturar una relación no lineal entre la variable dependiente y la variable explicativa.

Así, un modelo de la forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + u_i$$

Se puede estimar por MCO, y permitirá capturar una relación no lineal entre y y x_1.

6. Suponga un modelo donde la variable dependiente es el logaritmo del salario por hora (***lyph***), y tenemos dos variables explicativas: ***esc*** (años de escolaridad) y ***test*** (test de habilidades con escala de 0 a 100). Si el coeficiente estimado para años de escolaridad es 0.11 y el coeficiente estimado para la variable test es 0.03, podemos afirmar entonces que la variable escolaridad es más relevante que la variable test. Comente. (5 puntos).

Falso, la única forma de poder comparar entre los coeficientes estimados, es obtener los coeficientes estandarizados, de esta manera el efecto marginal está expresado en una misma unidad (desviación estándar) y serán comparables entre ellos.

7. Mientras mayor es el R^2 el modelo es mejor, por lo tanto, siempre me quedaré con un modelo con un R^2 alto. Comente. (5 puntos).

Falso, el R^2 siempre aumenta con la cantidad de variables que se incorporen al modelo, pero esto no significa que el modelo sea mejor, el indicador correcto es el R^2-ajustado, el penaliza la pérdida de grados de libertad. Así, siempre me quedará con el modelo con mayor R^2-ajustado.

II. EJERCICIOS CORTOS

1. Suponga el siguiente modelo de regresión lineal (sin constante):

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

Donde:

$$y = \begin{bmatrix} 10 \\ 20 \\ 5 \\ 5 \\ 25 \\ 35 \\ 5 \\ 15 \end{bmatrix}; \qquad X = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

y: corresponde al número de meses que la persona ha estado desempleada
x_1: es una variable dicotómica que toma valor 1 para las mujeres y 0 para los hombres
x_2: es una variable dicotómica que toma valor 1 para los hombres y 0 para las mujeres

[Ayuda: recuerde que la forma general del estimador MCO es $\hat{\beta} = (X'X)^{-1}X'y$]

a) ¿Cuántas observaciones se disponen para la estimación? (1 punto)

N=8

b) Con los datos presentados, ¿Puede decir cuál es el promedio de meses desempleados de los hombres? ¿y de las mujeres? (4 puntos)

Si, al observar la matriz X podemos notar claramente que las primeras 4 observaciones corresponden a hombres y las últimas 4 observaciones a mujeres. De esta forma, el promedio de meses desempleado de los hombres es (10+20+5+5)/4=10, y el de las mujeres (25+35+5+15)/4=20.

c) Encuentre el estimador MCO de β_1 y β_2 (7 puntos)

El estimador MCO de $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ es igual a:

$$\hat{\beta} = (X'X)^{-1}X'y$$

De esta forma,

$$\hat{\beta} = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 80 \\ 40 \end{bmatrix} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/4 \end{bmatrix} \begin{bmatrix} 80 \\ 40 \end{bmatrix} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}$$

d) Interprete los resultados obtenidos, en particular, ¿que representan $\hat{\beta}_1$ y $\hat{\beta}_2$? (3 puntos).

$\hat{\beta}_1$ representa el valor promedio de la variable dependiente, meses desempleado, para las mujeres, y $\hat{\beta}_2$ representa el valor promedio de la variable dependiente, meses desempleado, para los hombres.

2. Suponga que se estima el siguiente modelo:

$$lyph = \beta_0 + \beta_1 esc + \beta_2 edad + \beta_3 edad^2 + u$$

Con 10.000 observaciones. Se obtienen los siguientes resultados:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 2.56 \\ 0.11 \\ 0.05 \\ -0.0005 \end{bmatrix}$$

$$\sqrt{\hat{V}(\hat{\beta}_1)} = 0.0027855;\ \sqrt{\hat{V}(\hat{\beta}_1)} = 0.003759;\quad \sqrt{\hat{V}(\hat{\beta}_1)} = 0.0000379$$

a) Sin necesidad de hacer los cálculos precisos, puede afirmar que las tres variables son estadísticamente significativas, ¿por qué si o por qué no? (5 puntos)

Para testear la hipótesis nula de que un coeficiente del modelo de regresión es estadísticamente significativo, se debe calcular el estadístico t asociado a la hipótesis nula, el que corresponde a la razón entre el valor estimado del coeficiente y su desviación estándar, el que debe ser comparado con el valor de tabla de la distribución t, que para muchas observaciones al 5% de significancia es 1.96. Así, mientras el estadístico calculado sea menor a -1.96 o mayor a 1.96 se rechaza la hipótesis nula que el coeficiente es igual a cero y se puede afirmar que la variable e estadísticamente significativa.

Así, el estadístico t para la hipótesis nula de que el coeficiente que acompaña a esc es igual a cero es igual a 0.11/0.0027855, claramente mayor a 1.96, por lo tanto, esta variable es estadísticamente significativa. Lo mismo sucede para edad donde 0.05/0.003759 es claramente mayor a 1.96, y para la edad al cuadrado donde -0.0005/0.0000379 es claramente menor a -1.96. Las tres variables son significativas.

b) Según el modelo estimado, ¿existe alguna edad (razonable) para la cual comienza a disminuir el efecto de la edad sobre el logaritmo del salario por hora? (5 puntos)

Del modelo estimado tenemos que el efecto marginal de la edad sobre el logaritmo del salario por hora es:

$$\frac{\partial \, \mathrm{E}[\widehat{\ln yph}]}{\partial edad} = \hat{\beta}_2 + 2\hat{\beta}_3 edad$$

Tomando los valores estimados se tiene que:

$$\frac{\partial \, \mathrm{E}[\widehat{\ln yph}]}{\partial edad} = 0.05 - 2 \cdot 0.0005 \cdot edad$$

Para una persona con 20 años de edad, el efecto marginal de la edad sobre el logaritmo de salario por hora es 0.03. Pero para una persona de 60 años es -0.01. Para encontrar la edad donde este efecto marginal pasa de ser positivo a negativo debemos igual la ecuación anterior a cero y despejar la edad:

$$0.05 - 2 \cdot 0.0005 \cdot edad = 0$$

Por lo cual, a los 50 años de edad, el efecto marginal de la edad sobre logaritmo del salario por hora pasa a ser negativo.

3. La alcaldesa de la comuna de Renca hizo una encuesta antes de la elección pasada para ver cuál era la probabilidad de que fuese elegida nuevamente. Se entrevistaron a 400 vecinos de la comuna, de los cuales 215 dijeron que votarían por la actual alcaldesa, y los restantes 185 personas por el nuevo candidato (solo hubo dos candidatos por los cuales se preguntó).

La variable observada y_i es una variable aleatoria tipo Bernoulli, que toma valor 1 si la persona entrevistada dice que votaría a la alcaldesa, lo cual sucede con probabilidad p y 0 si la persona votaría al contrincante, lo que sucede con probabilidad $1 - p$.

a) Plantee un estimador para la probabilidad de que se vote a la alcaldesa, es decir, la $\Pr(y = 1) = p$. (3 puntos)

Notemos que la probabilidad de que y sea igual a 1 se puede aproximar muestralmente con la proporción de unos en la muestra, lo que corresponde a la media muestral de la variable y. Entonces, un estimador de p es:

$$\hat{p} = \bar{y} = \frac{\sum y}{n}$$

b) Encuentre el valor para dicho estimador. (2 puntos)

El valor del estimador, es:

$$\hat{p} = \bar{y} = \frac{215}{400}$$

c) Demuestre que el estimador es insesgado (5 puntos)

Para que el estimador sea insesgado, debemos demostrar que $E(\hat{p}) = p$:

$$E(\hat{p}) = E\left(\frac{\sum y}{n}\right) = \frac{\sum E(y)}{n} = \frac{\sum p}{n} = \frac{np}{n} = p$$

d) Calcule la varianza del estimador [Ayuda: recuerde que si la varianza de una variable Bernoulli es $p(1-p)$] (5 puntos)

$$V(\hat{p}) = V\left(\frac{\sum y}{n}\right) = \frac{\sum v(y)}{n^2} = \frac{\sum p(1-p)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

III. EJERCICIO APLICADO (24 puntos)

Para el siguiente ejercicio se utilizan datos de una muestra aleatoria de egresados de educación secundaria en Estados Unidos, los que fueron entrevistados en 1980 y luego en 1986.

El objetivo es poder investigar la relación entre los años de educación completados y la distancia a las universidades, la hipótesis es que mientras más cerca residan de una universidad mayor es la probabilidad de que asistan a la universidad y tengan mayor nivel educacional, ya que los costos son menores.

La siguiente imagen muestra las estadísticas descriptivas de los años de escolaridad completados:

```
. sum ed, detail
```

ed

Percentiles		Smallest		
1%	12	12		
5%	12	12		
10%	12	12	Obs	3796
25%	12	12	Sum of Wgt.	3796
50%	13		Mean	13.82929
		Largest	Std. Dev.	1.813969
75%	16	18		
90%	16	18	Variance	3.290483
95%	17	18	Skewness	.4129986
99%	18	18	Kurtosis	1.719979

a) ¿Cuántas observaciones hay en la base de datos? (1 punto)

N=3.796

b) ¿Cuál es el valor promedio de los años de escolaridad? (1 punto)

La escolaridad promedio es de 13,83.

c) ¿Pareciera ser que esta variable tiene una distribución normal? Justifique su respuesta (2 puntos)

Si, la media es muy parecida a la mediana y además el coeficiente de asimetría (skewness) es cercano a cero.

La siguiente figura nos muestra las estadísticas descriptivas de la variable distancia a una universidad, medida en kilómetros:

```
. sum distk, detail

                          distk
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0              0
 5%        1.61              0
10%        1.61              0       Obs                3796
25%        6.44              0       Sum of Wgt.        3796

50%        16.1                      Mean           27.77123
                        Largest       Std. Dev.      34.35476
75%       40.25          257.6
90%        64.4          257.6       Variance       1180.249
95%    83.71999          257.6       Skewness       2.904585
99%       177.1          257.6       Kurtosis       15.48146
```

d) ¿Cuál es la distancia mediana a una universidad? (1 punto)

La distancia mediana a una universidad es de 16,1 kms.

e) ¿Pareciera ser que esta variable tiene una distribución normal? Justifique su respuesta (2 puntos)

No, la media es muy diferente a la mediana (superior) además el coeficiente de asimetría es mayor a cero.

La siguiente figura muestra la estimación por MCO de la relación entre años de escolaridad (variable dependiente) y la distancia a una universidad (variable explicativa).

```
. reg ed distk

      Source |       SS       df       MS              Number of obs =    3796
-------------+------------------------------           F(  1,  3794) =   28.48
       Model |  93.0256748     1  93.0256748            Prob > F      =  0.0000
    Residual |  12394.3568  3794  3.26683101           R-squared     =  0.0074
-------------+------------------------------           Adj R-squared =  0.0072
       Total |  12487.3825  3795  3.29048287           Root MSE      =  1.8074

------------------------------------------------------------------------------
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       distk |  -.0045573    .000854    -5.34   0.000    -.0062317   -.0028829
       _cons |   13.95586   .0377241   369.95   0.000     13.88189    14.02982
------------------------------------------------------------------------------
```

f) Interprete el coeficiente estimado de distancia (distk) (2 puntos)

El coeficiente estimado para la variable distancia nos indica que el un kilómetro adicional de distancia a una universidad disminuye en promedio los años de escolaridad en 0.0046.

g) ¿Es este coeficiente estadísticamente significativo? Justifique su respuesta (2 puntos)

Si, ya que el estadístico t-calculado es menor a -1.96.

La siguiente tabla muestra la estimación de un segundo modelo, donde se han incorporado las siguientes variables:

- bytest: test aplicado al salir de la educación secundaria (puntos entre 0 y 100)
- cue80: tasa de desempleo en el condado en 1980 (tasa 0 y 25)
- stwmfg80: salario por hora estatal en manufactura en 1980 (dólares)

```
. reg ed distk bytest cue80 stwmfg80
```

Source	SS	df	MS		
Model	2891.0467	4	722.761675	Number of obs	= 3796
Residual	9596.33581	3791	2.53134682	F(4, 3791)	= 285.52
				Prob > F	= 0.0000
Total	12487.3825	3795	3.29048287	R-squared	= 0.2315
				Adj R-squared	= 0.2307
				Root MSE	= 1.591

ed	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
distk	-.0033926	.0007806	-4.35	0.000	-.0049231	-.0018621
bytest	.0981493	.0029553	33.21	0.000	.0923551	.1039435
cue80	.0176413	.0098448	1.79	0.073	-.0016603	.0369429
stwmfg80	-.0545766	.0201521	-2.71	0.007	-.0940866	-.0150666
_cons	9.304227	.2265778	41.06	0.000	8.86	9.748453

h) Interprete el coeficiente estimado para stwmfg80 (2 puntos)

El coeficiente estimado para la variable stwmfg80, nos indica que por cada dólar adicional en el salario promedio por hora estatal en la manufactura en 1980 la escolaridad disminuye en promedio 0.055 años.

i) Interprete el coeficiente estimado para cue80. ¿Le parece razonable el resultado obtenido? (3 puntos)

Por cada punto adicional de tasa de desempleo en el condado, la escolaridad aumenta en promedio 0,018 años. Es razonable el resultado ya que al ser mayor la tasa de desempleo, es menor probable encontrar el trabajo y menor el costo alternativo de estudiar.

j) ¿Se mantiene la significancia de la variable distancia? ¿Son las otras variables estadísticamente significativas? (3 puntos)

Si, la variable distancia sigue siendo significativa (el estadístico calculado es menor a -1.96). Con respecto a las otras variables, bytest y stwfmg80, son significativas ya que en el primer caso el estadístico calculado es mayor a 1.96 y en el segundo caso es menor a -1.96. Sin embargo, cue80 no es estadísticamente significativa ya que el estadístico calculado es menor a 1.96 y mayor a -1.96.

k) ¿Que puede decir sobre la bondad de ajuste del modelo, como cambia con respecto al primer modelo? (3 puntos)

En el primer modelo el R2-ajustado era 0.0072, indicando que la variable distancia explica menos de un 1% de la varianza de los años de escolaridad. Al incorporar las otras variables el R2-ajustado aumenta de manera importante a 0.2307, indicando que estas variables en conjunto son capaces de explicar 23% del comportamiento de la varianza de los años de escolaridad.

Finalmente, la siguiente imagen nos muestra la estimación de los coeficientes estandarizados:

```
. reg ed distk bytest cue80 stwmfg80, beta
```

Source	SS	df	MS		
Model	2891.0467	4	722.761675	Number of obs	= 3796
Residual	9596.33581	3791	2.53134682	F(4, 3791)	= 285.52
				Prob > F	= 0.0000
Total	12487.3825	3795	3.29048287	R-squared	= 0.2315
				Adj R-squared	= 0.2307
				Root MSE	= 1.591

ed	Coef.	Std. Err.	t	P>|t|	Beta
distk	-.0033926	.0007806	-4.35	0.000	-.0642529
bytest	.0981493	.0029553	33.21	0.000	.4771875
cue80	.0176413	.0098448	1.79	0.073	.0278704
stwmfg80	-.0545766	.0201521	-2.71	0.007	-.0410508
_cons	9.304227	.2265778	41.06	0.000	.

l) ¿Cuál de las 4 variables explicativas tiene mayor relevancia relativa? (2 puntos)

Los coeficientes estandarizados nos permiten comparar los efectos marginales de las distintas variables y ver cual es más relevantes, ya que todos los efectos marginales se miden en desviaciones estándar que es una mediada estandarizada libre de la unidad de las variables. En este caso, bytest es la variable que afecta más los años de escolaridad.