MÉTODOS ESTADÍSTICOS APLICADOS A LOS NEGOCIOS

TRABAJO PRÁCTICO 1 - 06/09/2017

GRUPO
ANTÓN MARÍA PAZ
FIGLIOLIA JULIETA ALDANA
PONCIO FEDERICO
SABATER ANNA
WANG JIA QI

PUNTO 1

Ejercicio 1a

Se esperaría que el signo de β_2 fuera positivo: ante un mayor ingreso familiar de los padres, aumentaría el peso del bebé recién nacido ya que -a priori- el efecto de un mayor nivel de salario permitirá satisfacer las necesidades básicas y nutricionales del bebé.

Ejercicio 1b

Argumentos a favor de una relación positiva: se supone que ante un mayor ingreso familiar, mayor consumo de cigarrillo debido a:

→Que las familias viven un mayor stress en su trabajo y fumar es una forma de liberar sus preocupaciones.

→Mayor propensión a consumir y satisfacer el deseo de fumar.

Argumentos a favor de una relación negativa: se supone que ante un mayor ingreso familiar, menor consumo de cigarrillo debido a:

→Mayor posibilidad de acceder a recursos médicos e información

→Mayor posibilidad de sustituir el cigarrillo por otro producto alternativa, como pipa o habano.

Como se observa en la salida de STATA, el modelo estima una correlación negativa. Esta estimación resultaría ser la más intuitiva para el actual contexto histórico, ya que, hoy en día, fumar durante el embarazo no está bien visto socialmente. Más allá de otros posibles factores, esta correlación negativa podría estar pasando por alguna de las causas mencionadas anteriormente.

```
             |   faminc     cigs
-------------+------------------
      faminc |   1.0000
        cigs |  -0.1730   1.0000
```

Ejercicio 1c

El modelo de regresión simple, que considera solamente la variable *cigs,* estima que por cada cigarrillo extra que fume la madre durante el embarazo, el bebé nacido pierde 0.5137721 onzas de peso. En cambio, el modelo de regresión múltiple, que incluye las variables *cigs* y *faminc*, computa un coeficiente β_1 menor y estima que, si se mantienen constantes el resto de las variables, por cada cigarrillo extra que fume la madre durante el embarazo, se reduce el peso del bebé nacido en 0.46 onzas.

Este cambio en el coeficiente del regresor β_1 podría ocurrir porque hay variables que están siendo omitidas en el modelo de regresión simple, por lo que, al agregarlas, el coeficiente de la regresión múltiple se ajusta y se explica mejor.

Al comparar los R^2 ajustados, se observa que el modelo simple explica en un 2.2% la variabilidad de la variable dependiente, es decir, el peso del bebé nacido. Mientras que, el modelo de regresión múltiple lo explica en un 2.8%. Notamos que, aunque el modelo de múltiples variables explica en un mayor porcentaje la variabilidad del peso del bebé, la diferencia entre los R^2 ajustados es pequeña, sólo de 0.06%.

```
. reg bwght cigs, robust
Linear regression                               Number of obs     =      1,388
                                                F(1, 1386)        =      34.29
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0227
                                                Root MSE          =     20.129
------------------------------------------------------------------------------
             |               Robust
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5137721   .0877334    -5.86   0.000    -.6858767   -.3416675
       _cons |   119.7719   .5745494   208.46   0.000     118.6448    120.899
------------------------------------------------------------------------------
. display _result(8)
.02202402

. reg bwght cigs faminc, robust
Linear regression                               Number of obs     =      1,388
                                                F(2, 1385)        =      22.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0298
```

```
                                             Root MSE          =     20.063
------------------------------------------------------------------------------
             |               Robust
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.4634075   .0887594    -5.22   0.000     -.637525   -.2892901
      faminc |   .0927647   .0285864     3.25   0.001     .0366875     .148842
       _cons |   116.9741   1.037207   112.78   0.000     114.9395    119.0088
------------------------------------------------------------------------------
. display _result(8)
.02840383
```

Ejercicio 1d

Se evalúa la significatividad del modelo de regresión múltiple, que incluye la variable *faminc:*

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u_i$$

Observamos las siguientes tres cuestiones: el p-value de la variable *faminc*, el p-value del test F de significatividad conjunta del modelo y los R^2 ajustados de las estimaciones de la regresión simple y múltiple.

→La variable *faminc* tiene signo positivo (β_3=.092), su coeficiente es estadísticamente distinto de cero y significativo al 1% (p-value=.001). Su intervalo de confianza es (0.03;0.14).

→El modelo de regresión múltiple presenta un R^2 ajustado mayor (=.0284) que el modelo de regresión simple, por lo que esto se considera como un primer indicio a favor de incluir la variable *faminc* en la estimación.

→El test F de significatividad conjunta evalúa la hipótesis nula de no significatividad conjunta (β_0+β_1+β_2=0) y la hipótesis alternativa de sí significatividad conjunta (Algún β_j≠0). El p-value de F (.000) permite rechazar la hipótesis nula y por lo tanto existe evidencia a favor de la significatividad conjunta del modelo a un nivel de significancia del 1%.

A partir de esto, se concluye que es conveniente incluir la variable *faminc* en la estimación del modelo.

El coeficiente β_1=-.046 implica que ante un aumento en una unidad en el consumo de cigarrillo por día durante el embarazo se reduce el peso del bebé en 0.046 onzas. La variable *cigs* presenta un p.value de .000, por lo que, su coeficiente es estadísticamente distinto de cero y significativo individualmente a un nivel de significancia del 1%.

Por lo tanto, el modelo propuesto por los investigadores estima que existiría algún efecto dañino del consumo de cigarrillo en la salud del bebé.

El coeficiente β_2=.092 implica que ante un aumento de $1000 en el salario familiar se incrementa el peso del bebé en 0.092 onzas.

```
. test cigs faminc
 ( 1)  cigs = 0
 ( 2)  faminc = 0
       F(  2,  1385) =   22.11
            Prob > F =    0.0000
```

Ejercicio 1e.a

Etapa 1 del teorema Frisch-Waugh: Se corre una regresión de *cigs* contra *faminc* y se computan los residuos $\hat{u}_1$ los cuales incluyen todos los efectos del modelo que no fueron capturado por las variables *faminc* y la constante.

```
. reg cigs faminc
      Source |       SS           df       MS      Number of obs   =     1,388
-------------+----------------------------------   F(1, 1386)      =     42.78
       Model |  1481.60979         1  1481.60979   Prob > F        =    0.0000
    Residual |  47996.8419     1,386  34.6297561   R-squared       =    0.0299
-------------+----------------------------------   Adj R-squared   =    0.0292
       Total |  49478.4517     1,387  35.6730005   Root MSE        =    5.8847

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      faminc |  -.0551538   .0084321    -6.54   0.000    -.0716948   -.0386129
       _cons |   3.688107   .2912973    12.66   0.000     3.116676    4.259538
------------------------------------------------------------------------------
. predict residU1, residual
. sum residU1, detail
                Residuos de la variable cigs
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -3.66053       -3.66053
 5%    -3.439915      -3.66053
10%      -3.2193      -3.66053       Obs               1,388
25%    -2.667761      -3.66053       Sum of Wgt.       1,388

50%    -1.895608                     Mean          -4.71e-11
                      Largest       Std. Dev.      5.882583
75%    -.1031081      37.55285
90%     6.946162      37.82862       Variance       34.60479
95%     16.56009      42.33947       Skewness       3.452781
99%     19.89689      47.82862       Kurtosis       17.59686
```

Ejercicio 1e.b

Etapa 2 del teorema Frisch-Waugh: se corre una regresión de *bwght* contra *faminc* y se computan los residuos $\hat{u}_2$ los cuales representan todos los efectos del modelo que no fueron capturados por las variables *faminc* y la constante.

```
. reg bwght faminc
------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      faminc |   .1183234   .0290016     4.08   0.000     .0614317    .1752152
       _cons |    115.265   1.001901   115.05   0.000     113.2996    117.2304
```

```
------------------------------------------------------------------------------
. predict residU2, residual
. sum residU2, detail

                Residuos de la variable bwght
-------------------------------------------------------------
      Percentiles      Smallest
 1%    -55.11055      -94.92731
 5%    -32.95606      -87.92731
10%    -24.15246      -83.51893       Obs               1,388
25%    -11.92156      -81.11054       Sum of Wgt.       1,388

50%     .8894544                      Mean          -9.17e-09
                       Largest       Std. Dev.      20.23283
75%     13.09365       50.29784
90%     24.04394       58.07269       Variance       409.3675
95%     30.07269       72.88946       Skewness      -.1436457
99%     43.32084       150.7062       Kurtosis       6.112592
```

Ejercicio 1e.c

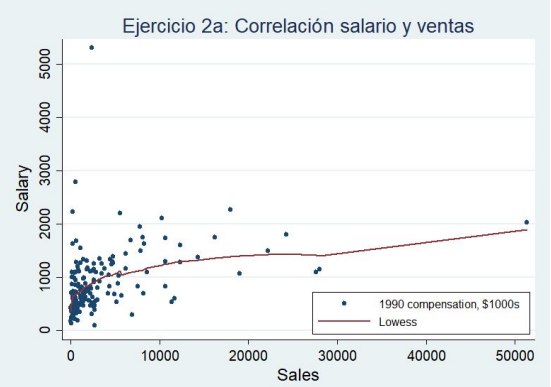
Etapa 3 del teorema Frisch-Waught: se corre una regresión de $\hat{u}_2$ contra $\hat{u}_1$, la cual da una formulación matemática de la forma en que el coeficiente de regresión múltiple β_1 estima el efecto sobre *bwght* de *cigs*, controlado por las otras variables.

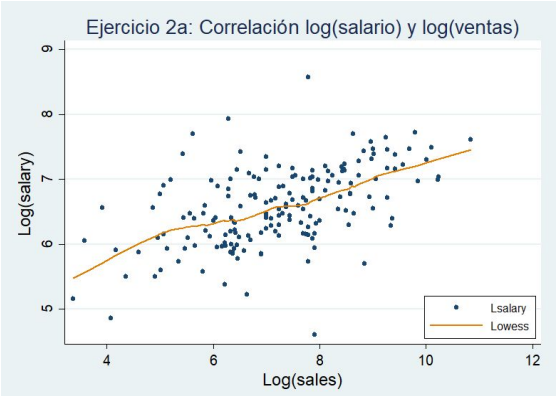
Como las dos primeras regresiones (etapas 1 y 2) eliminan de *bwght* y *cigs* su variabilidad asociada a las otras variables, la tercera regresión estima el efecto sobre *bwght* de *cigs* utilizando lo que queda después de eliminar el efecto de las otras variables.

Por eso, el coeficiente MCO de *residU1* de la etapa 3 del teorema Frisch-Waugh es igual al coeficiente MCO de *cigs* del modelo planteado por los investigadores.

PUNTO 2

Ejercicio 2a





En el gráfico de la izquierda se observa que la dispersión de Sales/Salary está acumulada en los primeros cuartiles de cada variable y, además, se ve parecido a una forma logarítmica. Esta intuición inicial se apoya en la línea trazada por el comando *lowess*.

En el gráfico de la derecha se observa la dispersión de los logaritmos de ambas variables, los cuales ahora asemejan una distribución lineal. Esto apoyaría la hipótesis que la relación entre el salario y las ventas adopta una forma logarítmica y no lineal. Además de mejorar el ajuste del modelo, esta transformación tiene la ventaja de acercar a los outliers: aquellos que en el gráfico de la izquierda están aproximadamente en los valores (100, 5000) y (50000, 2000) pasan a estar en el gráfico de la derecha en los intervalos (8, 8.5) y (11, 7.5), respectivamente.

Ejercicio 2b

Se plantea el siguiente modelo: $salary_i = \beta_0 + \beta_1 sales_i + \beta_2 mktval_i + u_i$

En la salida de STATA se observa que:

→El coeficiente β_1 de la variable *sales* es 0.0165. Por lo que, el modelo estima que ante un aumento en un millón de dólares en las ventas de la compañía, se incrementa el salario de los ejecutivos en 0.0165 mil dólares (U$S16.500). Sin embargo, este coeficiente no es significativo a un nivel del 10% (su p-value es 0.105) y por lo tanto es estadísticamente igual a cero. Su intervalo a un 90% de nivel de confianza es [-0.0002, 0.03322]

→El coeficiente β_2 de la variable *mktval* es 0.0253. Es decir, se estima que ante un aumento de un millón de dólares en el valor de la empresa en el mercado accionario, se incrementa el salario de los ejecutivos en 0.0253 millones de dólares (U$S25.300). Este coeficiente sí es significativo a un nivel del 10%, por lo que se podría afirmar con un 90% de seguridad el efecto predicho.

→Con respecto a la constante del modelo, el coeficiente β_0 es significativo a un nivel del 10%. La constante se interpreta diciendo que, en el caso hipotético que tanto las ventas de la compañía como el valor de mercado fueran cero, se esperaría ver que el salario de los ejecutivos es de U$S 716.000.

Para una empresa cuyas ventas ascienden a U$S4.000 millones y que cuyo valor de mercado es U$S3.900 millones, el modelo estima un salario esperado del ejecutivo de 881.20714 miles de dólares. Entonces se espera un salario de U$S 881.207,14.

El cálculo realizado fue:

$$salary_i = 716.000 + 0.0164994 * 4000 + 0.0252906 * 3900 = 881,20714 * 1000 = 881.207,14$$

```
. reg salary sales mktval, l(90)
      Source |       SS           df       MS      Number of obs   =       177
-------------+----------------------------------   F(2, 174)       =     18.80
       Model |  10796207.1         2  5398103.55   Prob > F        =    0.0000
    Residual |  49969757.7       174  287182.515   R-squared       =    0.1777
-------------+----------------------------------   Adj R-squared   =    0.1682
       Total |  60765964.7       176  345261.163   Root MSE        =    535.89
------------------------------------------------------------------------------
      salary |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
       sales |   .0164994   .0101116     1.63   0.105    -.0002218    .0332206
      mktval |   .0252906   .0095566     2.65   0.009     .0094872     .041094
       _cons |   716.5762   47.18752    15.19   0.000     638.5442    794.6082
------------------------------------------------------------------------------
```

Ejercicio 2c

Se estima un modelo de forma funcional log-log donde se aplica logaritmo tanto a la variable dependiente, *lsalary*, como a *lsales* y *lmktval* para asumir que las elasticidades de estas dos variables explicativas son constantes.

El coeficiente β_1 de *lsales* estima que un aumento del 1% en las ventas está asociado con un aumento en el salario del ejecutivo de 0.16%, manteniendo constante el resto de las variables.

El coeficiente β_2 de *lmktval* estima que ante un aumento del 1% en el valor de la empresa en el mercado accionario, se espera un incremento del 0.10% en el salario de los ejecutivos, manteniendo constante el resto de las variables.

El coeficiente β_0 de la constante computa que si tanto el monto de las ventas como el valor de mercado de la empresa son nulos, el salario de los ejecutivos se estima en 4.6209 miles de dólares (U$S 4.620,91)

No se puede utilizar el R^2 ni el R^2 *ajustado* para comparar estas dos regresiones debido a que las variables dependientes son diferentes, en uno es *salary* y en el otro es *ln(salary)*. Debido a que el R^2*ajustado* mide la proporción de la varianza de la variable dependiente que está explicada por los regresores, como las variables dependientes de los modelos log-log y lineal-log son diferentes, no tiene sentido comparar los R^2*ajustados*.

```
     Lsalary |      Coef.   Std. Err.      t    P>|t|     [90% Conf. Interval]
-------------+----------------------------------------------------------------
      Lsales |   .1621283   .0396703     4.09   0.000     .0965273    .2277293
     Lmktval |    .106708    .050124     2.13   0.035     .0238201    .1895959
       _cons |   4.620917   .2544083    18.16   0.000     4.200213    5.041622
------------------------------------------------------------------------------
```

Ejercicio 2d

Al modelo estimado se le incorpora el regresor *profits* pero no se le aplica logaritmo a la variable debido a que los ingresos de las compañías pueden llegar a ser cero o negativos, por lo que, la transformación logarítmica no captaría dichos valores**.** El p-value del test de significatividad individual del coeficiente no es estadísticamente distinto de cero para ninguno de los niveles evaluados por Stata (0.1%, 1% y 5%).

A su vez, el R^2 *ajustado* de este segundo modelo explica en una menor proporción (28.7%) la variabilidad de *lsalary*, mientras que el del inciso anterior lo explica en un 29.1%.

Se concluye que el modelo del inciso anterior resultaría mejor porque su R^2 *ajustado* es mayor, explica mejor el salario de los ejecutivos y no incorpora la variable *profits*, la cual vimos que no resulta estadísticamente significativa.

La correlación computada entre *mktval* y *profits* es 0.9181. Ante esto, vemos que incorporar la variable profits en el modelo podría ocasionar problemas de multicolinealidad imperfecta, ya que dos de los regresores están altamente correlacionados. Entonces, el coeficiente de al menos un regresor individual se estimará de forma imprecisa. Cuanto mayor sea la correlación entre las dos variables explicativas, *mktval* y *profits*, mayor es la varianza estimada de *mktval*. Esto ocurre porque la varianza estimada de β_{mktval} es inversamente proporcional a 1-$\rho_{mktval,profits}$ donde $\rho_{mktval,profit}$ es la correlación entre *mktval* y *profits*.

```
. sum profits
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     profits |        177    207.8305    404.4543       -463       2700

. corr mktval profits
(obs=177)
             |   mktval  profits
-------------+------------------
      mktval |   1.0000
     profits |   0.9181   1.0000
```

```
                    (1)             (2)
                 Lsalary         Lsalary
------------------------------------------------
Lsales             0.162***        0.161***
                  (4.09)          (4.04)
Lmktval            0.107*          0.0975
                  (2.13)          (1.53)
profits                         0.0000357
                                  (0.23)
_cons              4.621***        4.687***
                  (18.16)         (12.34)
------------------------------------------------
N                    177             177
adj. R-sq          0.291           0.287
------------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

Ejercicio 2e

El modelo a evaluar es:

$$log(salary_i) = \beta_0 + \beta_1 log(sales_i) + \beta_2 log(mktval_i) + \beta_3 ceoten_i + u_i$$

Se observa que la variable *ceoten* es estadísticamente distinta de cero y significativa a un nivel del 5%. También, se observa en esta tercer regresión que tanto el R^2 como el R^2 ajustado son mayores a diferencia de los modelos anteriores. Por estas razones, se concluye que es conveniente incorporar la variable *ceoten* en la estimación.

	(1)	(2)	(3)
	Lsalary	Lsalary	Lsalary
Lsales	0.162***	0.161***	0.163***
	(4.09)	(4.04)	(4.15)
Lmktval	0.107*	0.0975	0.109*
	(2.13)	(1.53)	(2.20)
profits		0.0000357	
		(0.23)	
ceoten			0.0117*
			(2.20)
_cons	4.621***	4.687***	4.504***
	(18.16)	(12.34)	(17.51)
N	177	177	177
R-sq	0.299	0.299	0.318
adj. R-sq	0.291	0.287	0.306

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Ejercicio 2f

El test F plantea, por un lado, la hipótesis nula de que los coeficientes de los regresores de *lnsales*, *lnmktval* y *ceoten* son iguales a cero. Por otro lado, la hipótesis alternativa plantea que algún coeficiente sea distinto de cero.

Se observa que el p-value 0.0000 del test F arroja suficiente evidencia a favor que el efecto global de las variables es estadísticamente distinto de cero, es decir, el modelo es conjuntamente significativo. Por esto, se concluye que hay un efecto significativo del rendimiento de la empresa sobre el salario de los ejecutivos.

```
. test lnsales lnmktval ceoten
 ( 1)  lnsales = 0
 ( 2)  lnmktval = 0
 ( 3)  ceoten = 0
       F(  3,   173) =   26.91
            Prob > F =    0.0000
```

Ejercicio 2g

Sobre el modelo planteado en el punto 2e se quiere testear una prueba F para combinaciones lineales de parámetros y una versión modificada de la prueba "T de Student".

Recordamos el modelo a evaluar: $log(salary_i) = \beta_0 + \beta_1 log(sales_i) + \beta_2 log(mktval_i) + \beta_3 ceoten_i + u_i$

Por un lado, en la prueba F se testea las siguientes hipótesis:

$$H_0 : \beta_1 - \beta_2 = 0;\ H_1 : \beta_1 - \beta_2 \neq 0$$

Es decir, la hipótesis nula plantea que la elasticidad de las ventas de la compañía es igual a la elasticidad del valor de la empresa sobre el salario de los ejecutivos. Mientras que la hipótesis alternativa propone que las elasticidades son distintas.

Se observa que el p-value 0.5184 de la prueba F no arroja suficiente evidencia estadística para poder rechazar la hipótesis nula a un nivel del 10%. Por lo que se concluye que las elasticidades de las ventas de la compañía y del valor de la empresa son iguales.

```
 ( 1)  lsales - lmktval = 0
       F(  1,   173) =    0.42
            Prob > F =    0.5184
```

Por otro lado, se plantea una versión modificada de la prueba "T de Student". Se testean las mismas hipótesis mencionadas anteriores:

$$H_0 : \beta_1 - \beta_2 = 0;\ H_1 : \beta_1 - \beta_2 \neq 0$$

Utilizamos $V(\beta_1 - \beta_2) = V(\beta_1) + V(\beta_2) - 2Cov(\beta_1, \beta_2)$ para obtener la desviación estándar de la diferencia y así calcular el siguiente estadístico T:

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)}{SE(\hat{\beta}_1 - \hat{\beta}_2)}$$

El valor del estadístico es 0.64716991, a comparar con un valor de tabla de 1.984, de una distribución t con 176 grados de libertad y un 5% de nivel de significación a dos colas.

Como el estadístico está en el intervalo [-1.984,1.984], se concluye que no existe suficiente evidencia estadística para poder rechazar la hipótesis nula que las elasticidades son iguales con un 95% nivel de confianza. Por lo que, la elasticidades de las ventas es igual a la elasticidad del valor de la empresa sobre el salario de los ejecutivos.