Actividad 2. Regresión Lineal Multiple (1)

Econometría LECO

Dr. Francisco Cabrera

Entregable: Un archivo de Word o PDF (e.g. LaTex, R-Markdown) contestando correctamente a las preguntas abajo presentadas y gravado con su nombre, así como el archivo "Script" de R.

1. Griliches (1977) en una investigación muy influyente intenta determinar los "Retornos a la educación" en EE. UU. Es decir, cuánto contribuye un año más de educación formal al sueldo en dólares.

Algunas de las variables que utilizó son las siguientes:

Nombre de la variable	
wage	Sueldo mensual en dólares.
lwage	Logaritmo del sueldo mensual en dólares.
educ	Años de educación.
exper	Años de experiencia laboral.
IQ	Coeficiente intelectual en puntos (Media de 100 puntos, DS de 15 puntos).
age	Edad del individuo en años.
Married = 1	Si se encuentra casado.
Black = 1	Si el individuo es de raza negra.
meduc	Educación de la madre en años.
Feduc	Educación del padre en años.

El autor, primero obtiene la siguiente regresión:

Source	SS	df	MS		
Model	16.1377042	1	16.1377042	Number of obs	= 935
Residual	149.518579	933	.160255712	F(1, 933)	= 100.70
				Prob > F	= 0.0000
				R-squared	= 0.0974
				Adj R-squared	= 0.0964
Total	165.656283	934	.177362188	Root MSE	= .40032

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0598392	.0059631	10.03	0.000	.0481366	.0715418
_cons	5.973063	.0813737	73.40	0.000	5.813366	6.132759

a. Interprete el coeficiente de *educ*.
b. Interprete el coeficiente ajustado de determinación (R^2 ajustada).

c. Interprete la constante en el modelo.

Posteriormente, el autor estima la siguiente regresión:

```
. reg lwage educ IQ
```

Source	SS	df	MS	Number of obs	=	935
				F(2, 932)	=	69.42
Model	21.4779447	2	10.7389723	Prob > F	=	0.0000
Residual	144.178339	932	.154697788	R-squared	=	
				Adj R-squared	=	
Total	165.656283	934	.177362188	Root MSE	=	.39332

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0391199	.0068382	5.72	0.000	.0256998	.05254
IQ	.0058631	.0009979	5.88	0.000	.0039047	.0078215
_cons	5.658288	.0962408	58.79	0.000	5.469414	5.847162

d. Interprete el coeficiente de *IQ*.
e. ¿En cuánto aumenta el ingreso promedio si el *IQ* aumenta en una Desviación Estándar (DS)?
f. Demuestre formalmente/matemáticamente, por qué el coeficiente de *educ* es menor que el obtenido en la primera regresión.

Ahora se presenta el modelo integrando todas las variables explicativas.

Source	SS	df	MS	Number of obs	=	722
				F(8, 713)	=	23.49
Model	26.4478349	8	3.30597936	Prob > F	=	0.0000
Residual	100.364081	713	.140763087	R-squared	=	
				Adj R-squared	=	
Total	126.811916	721	.175883378	Root MSE	=	.37518

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0449289	.0088533	5.07	0.000	.0275473	.0623106
IQ	.0044401	.0011973	3.71	0.000	.0020894	.0067909
exper	.0154783	.0044473	3.48	0.001	.006747	.0242095
age	.0121436	.0055212	2.20	0.028	.0013039	.0229834
married	.1848073	.0447525	4.13	0.000	.0969449	.2726696
black	-.0735403	.0528706	-1.39	0.165	-.177341	.0302603
feduc	.0097169	.0054483	1.78	0.075	-.0009797	.0204135
meduc	.0083396	.0061964	1.35	0.179	-.0038257	.0205049
_cons	4.810451	.1920529	25.05	0.000	4.433394	5.187508

---

i. ¿Es este modelo mejor describiendo la variación en sueldos que el modelo anterior? Estime manualmente R^2 ajustada en ambos modelos.

ii. Interprete el coeficiente de *exper* ¿Es posible saber si el sesgo en *educ* por omisión de *exper* es "hacia arriba" o "hacia abajo" con los resultados de esta regresión? ¿Muéstrelo formalmente?

iii. Explique, intuitivamente, por qué al autor le interesaría controlar por la educación de los padres del individuo i?

2. Copie el siguiente Código en R:

```
#Clear the environment
rm(list = ls())

library(tidyverse)

#help
help(rnorm)

#Lets create some variables!
set.seed(1234567) #good practice for when working with random vars (reproducibility)

x <- rnorm(1000) #1000 random obs
y <- matrix((5000 + 100*x) + rnorm(1000 * 500, mean=0, sd=1), ncol = 500) #Y = bo + b1x + u
y <- data.frame(y)

#this loop renames i column names of Y matrix
for (i in 1:ncol(y)){
  colnames(y)[i] <- paste0("y",i)
}

#Run 500 regressions!
betas <- 1:500 #create an empty object with 500 entries to be filled with the B1s

for (i in 1:ncol(y)){    #This loop runs 500 regressions Yi on X for i=1 to 500
  betas[i] <- summary(
  lm(y[,i]~x))$coefficients[2,1] #extracts the coefficient beta 1 from the matrix of results provided b
}
```

a. Pega la media de todas las β_1 estimadas

b. Pega el histograma de las β_1 estimadas

c. El Teorema de Gauss-Markov establece que, bajo los supuestos clásicos, el estimador MCO es BLUE (Best Linear Unbiased Estimator/ Mejor Estimador Lineal Insesgado). ¿Sugieren los estadísticos de resumen y el histograma de las partes a. y b. que el estimador MCO es insesgado? Explique por qué.

d. Explique qué partes del código anterior garantizan que se cumplan los supuestos de Gauss-Markov.

3. Utilice el código siguiente y cambie el número de observaciones de 100, a 10.000 y a 100.000.

```
set.seed(1234567)
x <- rnorm(100) #100 random obs
resid <- rnorm(100, mean=0, sd=10) #random error
y <- (20 + 2*x + resid)
model <-  lm (y ~ x)
summary(model)
```

a. Muestre los tres resultados.
b. Explique por qué, formalmente, (¡utilice una fórmula!) el error estándar de *beta* converge a cero cuando n tiende a infinito.
c. Transforme el código anterior para mostrar que una mayor varianza de x reduce el error estándar de *beta*.
d. Por construcción, en el código anterior el residuo está centrado alrededor de cero. Cambiemos esto artificialmente. Mantenga n=100 y cambie el término residual en el código anterior a una media de 20. ¿Cuál es la constante ahora? ¿Está sesgada $\hat{\beta}_1$?
e. Dibuja un diagrama de dispersión (con una línea ajustada) para las regresiones de la parte IV.a. (n=100) y de la parte IV.d. ¿Cuáles son tus conclusiones sobre la estimación de β_1?

4. Ejecute el siguiente código:

```
#Clear the environment
rm(list = ls())

repet <- 1000
n <- 1000
beta <- NULL

set.seed(1234567)

for (i in 1:repet){
  x1 <- rnorm(n, mean=50, sd=10)
  x2 <- (rnorm(n, mean=5, sd=30)+.1*x1)
  u <- (rnorm(n, mean=0, sd=1))
  y=2+2*x1+10*x2+u # we define y, so that beta1=2 and beta2=10.
  beta[i] <- lm(y~x1)$coef[2]
  }

hist(beta, main="suit yourself, n=1000", xlim = c(0,8) )
abline(v = mean(beta), col="red", lwd=3, lty=2 )
abline(v = 2, col="blue", lwd=3, lty=2)
```

a. Ligue las líneas de código que considere relevantes con los supuestos MLR pertinentes. ¿Se cumple MLR4?

b. Describa el supuesto clave que se está analizando y demuestre mediante el uso de alguna fórmula cómo afecta la estimación de β_1 y β_2

c. Modifique el código arriba para estimar β_1 de una manera más eficiente.

5. Wooldridge Data en R.

a. Utilice la base de datos "*bwght*". Obtenga la regresión de *birth weight* (peso al nacer) en onzas sobre el consumo de cigarrillos al día. Utilice como controles *log(cigprice)* y *mothereduc*. Interprete los coeficientes β_0 y β_1.

b. Ejecute la misma regresión utilizando el logaritmo del peso al nacer como dependiente. Interprete los coeficientes β_0 y β_1.

Ejercicios Opcionales:

1 Bis) Resuelva los ejercicios 7 y 16 correspondientes al capítulo 2 de Wooldridge ed. 7.

2 Bis) Demuestre matemáticamente lo siguiente:

a. Que en un modelo Log – Level: $\mathbf{\%\Delta Y = \Delta X(}\beta_1 * 100\mathbf{)}$

b. Que un modelo Level – Log: $\mathbf{\Delta Y = \%\Delta X(}\beta_1/100\mathbf{)}$