Actividad 3. Regresión Lineal Multiple (2)

Econometría LECO

Dr. Francisco Cabrera

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

Entorno de trabajo en R

1. Sesgo por variables omitidas

Ejecute el siguiente código:

```
#Load data
library(wooldridge)
data("bwght")
?bwght
bwght$cigs2<-bwght$cigs*bwght$cigs #adding a column with squares (NO LINEALIDAD)

#Run regressions
est_0=feols(bwght~cigs,data=bwght)
est_1=feols(bwght~cigs+motheduc,data=bwght)
est_2=feols(bwght~cigs+motheduc+faminc,data=bwght)
est_3=feols(bwght~cigs+cigs2+motheduc+faminc,data=bwght)

#Nicer tables
##Standard errors (using fixest package)
etable(est_0,est_1,est_2,est_3, se="standard")  #Under MLR5
```

a. Usando *est_2* prediga el valor en onzas del peso de un niño cuya madre fumó 2 cigarros por día durante el embarazo, con todo lo demás en su promedio.
b. ¿Cuál es el signo de la correlación entre educación de la madre y número de cigarros fumados al día durante el embarazo? ¿Es lo que esperaba? Explique. (Vea Tabla 3.2. en Wooldridge ed. 7)
c. ¿Cuál es el efecto de fumar un cigarro extra en el embarazo en el peso en onzas considerando una relación no lineal? ¿Por qué utilizaría una relación no lineal? Explique.
d. Interprete R^2 y R^2adj en *est_3*. Formalmente ¿Por qué R^2adj es menor?

2. Varianza de los estimadores

```
## Standard errors (using fixest package)
#Table 1:
etable(est_0, est_1, est_2, est_3, se = "standard")# Under MLR5
#Table 2:
etable(est_0, est_1, est_2, est_3, se = "hetero") #Fixes Heteroskedasticity (HC1 type)
```

a. Demuestre matemáticamente por qué el SE en la segunda tabla es distinto al SE la Tabla 1, al menos para el estimador asociado a *cigs*. (i.e derive la varianza del error homoscedastica y heteroscedástica).
b. ¿Qué supuesto asumimos como vulnerado en la estimación de Tabla 2?
c. Dado: `summary(lm(bwght$cigs ~ bwght$motheduc))` y considerando *est_0* (en ejercicio 1) ¿aumenta la varianza del estimador *cigs* al incluirse *mothereduc* en *est_1*? Discuta lo anterior utilizando las fórmulas de la varianza de beta para un modelo simple y para uno múltiple.
d. ¿Cuál es el VIF tras agregar *motheduc*? Discuta, considerando el trade-off sesgo varianza si debemos incluir *motheduc*.
e. ¿A qué concepto nos referimos en el inciso c?

3. Los datos BEAUTY(biblioteca Wooldridge en R) contienen información sobre salarios y un índice de la belleza de las personas.

a. Explore las variables con `?beauty` y agregue una columna a los datos que incluya una aproximación de edad para cada i: `beauty$age <- beauty$educ+beauty$exper+6`

b. Estime una regresión simple: de *log(wage)* con *looks* como explicativa con errores estándar robustos a heteroscedasticidad (HC1): pegue el resultado abajo e interprete $\hat{\beta}_1$

c. Agregue como controles a la regresión en b. la experiencia, la experiencia al cuadrado, la educación en años y la edad ¿Qué pasa con su regresión y por qué? ¿Qué supuesto se puede estar violando?

d. Solucione el problema en c. y vuelva a estimar su regresión. Pegue el resultado abajo e interprete $\hat{\beta}_1$

4. Partialling-out.

Confirme la interpretación del partialling-out de las estimaciones OLS realizando explícitamente la separación parcial para el ejemplo 3.2 (ver Wooldridge 7ª Edición, Sección 3.2f). Esto primero requiere realizar una regresión de *educ* sobre *exper* y *tenure* y guardar los residuos (nombrelos `r1`). Luego, realice una regresión de *log(wage)* sobre *r1*. Compare el coeficiente de *r1* con el coeficiente de *educ* en la regresión de *log(wage)* sobre *educ*, *exper*, y *tenure*.

a. Pegue abajo el resultado de las dos regresiones de interés y explique intuitivamente por qué se observa que el coeficiente de *r1* es igual al de *educ* cuando controla por *exper* y *tenure*.

5. Siga el siguiente *script* y responda a las preguntas relacionadas:

```r
#Libreria
library(mvtnorm)
# set seed
set.seed(123)

#Example before computation:
#we define a matrix X with two vectors (regressors x1 and x2) from random normal values with mean [5,10
#and a matrix "sigma" of variances [3,6], and covariance 2

mu <- c(5,10)
sigma <- matrix(c(3,2,2,6), 2, 2)
X <- rmvnorm(1000, mean = mu, sigma = sigma)

?rmvnorm
```

```r
head(X)
summary(X)
var(X)
#plot vectors 1 and 2 in our matrix X (with covariance 2!)
plot(X[,1], X[,2])
##end of example.

# Set number of observations
n <- 500

#Another example of what is going to happen... Here we create a matrix X
#with two vectors (regressors x1 and x2) with median 50 and 100 each, and cov(X_1,X_2) = 0.25

X <-rmvnorm(n, c(50, 100), sigma = cbind(c(10, 2.5), c(2.5, 10)))
summary (X)
plot(X[,1], X[,2])

#Now the loop to do this 10000 times, fasten your seat belts:

# Empty vectors of coefficients to be filled
coefs1 <- cbind("beta_hat_1" = numeric(10000), "beta_hat_2" = numeric(10000))
coefs2 <- coefs1

###From here###

# Loop sampling and estimation (be patient it takes a while)
for (i in 1:10000) {

  # for median 50 and 100 and cov(X_1,X_2) = 0.25
  X <- rmvnorm(n, c(50, 100), sigma = cbind(c(10, 2.5), c(2.5, 10)))
  u <- rnorm(n, sd = 5)
  #We define PRF
  Y <- 5 + 2.5 * X[, 1] + 3.5 * X[, 2] + u
  #we compute SRF 10000 times using our n values in matrix X
  coefs1[i, ] <- lm(Y ~ X[, 1] + X[, 2])$coefficients[-1]

  # for cov(X_1,X_2) = 0.85
  X <- rmvnorm(n, c(50, 100), sigma = cbind(c(10, 8.5), c(8.5, 10)))
  Y <- 5 + 2.5 * X[, 1] + 3.5 * X[, 2] + u
  coefs2[i, ] <- lm(Y ~ X[, 1] + X[, 2])$coefficients[-1]

}

#Histogram of coefficients x1 and x2 for the two 10000s loops of regressions.
hist(coefs1)
hist(coefs2)

# Obtain variance estimates
diag(var(coefs1))
diag(var(coefs2))
```

a. A partir de la marca `###From here###`. Explique cada línea pertinente del código
b. Escriba la matriz de varianza – covarianza de los estimadores 1 y 2.

c. ¿Por qué los histogramas de coefs1 y coefs2 son distintos?
d. ¿Por qué las varianzas de coefs1 y coefs2 son distintas?
e. ¿qué parte del código genera estas diferencias?
f. ¿Qué concepto/supuesto central estamos simulando en este ejercicio?
g. Ahora cambie n a 1000, ¿qué sucede con los histogramas y las varianzas? Explique intuitivamente, pero con el uso de alguna fórmula ¿por qué se da este cambio?
h. Aún bajo el supuesto MRL4 ¿Se puede obtener un estimador inconsistente en una muestra aleatoria de individuos? De un ejemplo con el uso de los histogramas.