Sesgo, Varianza del estimador y Multicolinearidad

José Eduardo Valencia Espinosa

29/9/2022

Entorno

```r
library(pacman)
p_load(data.table, fixest, magrittr, sandwich, mvtnorm, MASS, sanwich, lmtest, haven, wooldrige)
set.seed(123)
```

1. Sesgo por variables omitidas

```r
# Load data
library(wooldridge)
data("bwght")
bwght$cigs2 <- bwght$cigs*bwght$cigs #adding a column with squares
# 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
```

	est_0 <chr>	est_1 <chr>	est_2 <chr>
Dependent Var.:	bwght	bwght	bwght
(Intercept)	119.8*** (0.5723)	115.4*** (3.107)	116.8*** (3.138)
cigs	-0.5138*** (0.0905)	-0.4862*** (0.0926)	-0.4633*** (0.0927)
motheduc		0.3308 (0.2328)	0.0143 (0.2580)
faminc			0.0915** (0.0325)
cigs2			
___________	_______________	_______________	_______________
S.E. type	IID	IID	IID
Observations	1,388	1,387	1,387

1-10 of 12 rows | 1-4 of 5 columns Previous 1 2 Next

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.

```r
values1 <- c(1, 2, mean(bwght$motheduc, na.rm=T), mean(bwght$faminc)) #omitiendo la fila con valores NA
coeff1 <- t(est_2$coefficients)
oz1 <- sum(coeff1*values1)
oz1
```

```
## [1] 118.7477
```

El modelo predice que el peso del bebé al nacer será de 118.7477 onzas (con 2 ciggaros diarios durante el embarazo y todos los demás valores en su promedio).

b. ¿Cuál es el signo de la covarianza 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)

```r
cov(na.omit(bwght)$motheduc, na.omit(bwght)$cigs)
```

```
## [1] -2.799661
```

El signo de la covarianza es negativo, esto quiere decir que su relación es inversa. La relación anterior puede tener sentido debido a factores socioeconómicos, por ejemplo, que una mujer embarazada con mayor educación conozca los efectos negativos en la salud de su bebé si fuma durante el embarazo. Sin embargo, esta solo es una afirmación premiliminar e intuitiva.

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.

Considerando *est_3*

```r
est_3$coefficients[2:3]
```

```
##        cigs       cigs2
## -0.77874420  0.01229368
```

$\frac{\delta \text{bwght}}{\delta \text{cigs}} = \beta_{\text{cigs}} + 2\beta_{\text{cigs2}}(\text{cigs}) =$ -0.7787 $+$ 0.0246 (cigs)

Utilizar una relación no lineal es útil para considerar que un cigarro más diario no afecta lo mismo que el anterior. Sin embargo, en este caso, esta relación no es estadísticamente significativa (al menos bajo el planteamiento de la regresión *est_3*).

d. Interprete R^2 y R^2-adj en est_3. Intuitivamente ¿Por qué R^2-adj es menor?

```r
r1 <- data.frame(t(c(summary(lm(bwght ~ cigs + cigs2 + motheduc + faminc, data = bwght))$r.squared,
summary(lm(bwght ~ cigs + cigs2 + motheduc + faminc, data = bwght))$adj.r.squared)))
colnames(r1) <- c("r.squared", "adj.r.squared")
r1
```

r.squared <dbl>	adj.r.squared <dbl>
0.03173806	0.02893557

1 row

Ambas *R's* describen que tanto las variables del modelo explican las variaciones en el peso del bebé. Así, la regresión *est_3* explica el 3.17% de las variaciones en el peso. La interpretación de la *R* ajustada es igual, sin embargo, está ajustada por el número de variables explicativas.

La *R* ajustada siempre será menor o igual a la *R* no ajustada pues la primera es una ponderación de la segunda en cuanto al número de regresores del modelo.

2. Varianza de los estimadores

```r
## 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") # Robust Standard Errors (HC1: Not White)
#Table 3:
etable(est_0, est_1, est_2, se = "cluster", cluster = c("white")) # One way Clustered
## Standard Errors using sandwich
# White's original SE (HC0) #This is with the classic White-Huber correction
etable(est_2, .vcov = sandwich::vcovHC, .vcov_args = list(type = "HC0"))
```

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.

La primera tabla asume homocedasticidad y la segunda ajusta la desviación estándar no constante por sus grados de libertad:

$$\text{se}(\hat{\beta}_j) = \frac{\sigma}{\sqrt{\text{SST}(1-R^2}}, \ \text{se}(\hat{\beta}_j) = \frac{\hat{u}_i}{\sqrt{\text{SST}(1-R^2}} \frac{1}{n-k-1}$$

b. Demuestre matemáticamente por qué SE en la tercera tabla es distinto de SE en la primera tabla. ¿Qué supuesto asumimos como vulnerado en la estimación de Tabla 3?

La tercera tabla hace el ajuste de White por clusters. Así asume que el supuesto de muestreo aleatorio fue vulnerado. Lo anterior porque asume que el muestreo aleatorio fue realizado de subgrupos y no de la población:

$$\text{se}(\hat{\beta}_j) = \frac{\hat{u}_i}{\sqrt{\text{SST}(1-R^2}}$$

c. Dado: cor(bwght$cigs, bwght$motheduc, use = "complete.obs") y considerando est_0 (en ejercicio 1) ¿aumenta la varianza del estimador cigs al incluirse mothereduc en est_1? Demuestre lo anterior utilizando las fórmulas de la varianza de beta para un modelo simple y para uno múltiple.

```r
cor(bwght$cigs, bwght$motheduc, use = "complete.obs")
```

```
## [1] -0.2138651
```

```r
variance_comparison <- data.frame(t(c((est_0$se[2])^2,(est_1$se[2])^2)))
variance_comparison
```

cigs <dbl>	cigs.1 <dbl>
0.008188608	0.008579242

1 row

Las varianzas son estadísticamente iguales.

d. ¿Cuál es el VIF Tras agregar motheduc? Discuta, considerando el trade-off sesgo varianza si debemos incluir motheduc.

```r
VIF <- 1/(1-(cor(bwght$motheduc,bwght$cigs, use="complete"))^2)
VIF
```

```
## [1] 1.047931
```

Sí debemos incluir *motheduc* en la regresión. En primer lugar, su VIF es pequeño (<10). En segundo lugar, y en el mismo sentido, la varianza de *educ* no cambiaba al agregar *motheduc*, por lo que no añade "ruido" a nuestra regresión (las variables no son multicolineares). En tercer lugar, agregar esta variable de control ayuda a disminuir el sesgo de las β's.

e. ¿A qué concepto nos referimos en el inciso c?

Multicolinearidad.

3. BEAUTY

```r
beauty_age <- beauty$educ+beauty$exper+6
coeftest(lm(beauty$lwage ~ beauty$looks + beauty$exper + beauty$expersq + beauty$educ + beauty_age),
vcov. = vcovHC(lm(beauty$lwage ~ beauty$looks + beauty$exper + beauty$expersq + beauty$educ + beauty_age),
type = 'HC1'))
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     0.04627736  0.10839716  0.4269  0.669508
## beauty$looks    0.06206811  0.02218320  2.7980  0.005221 **
## beauty$exper    0.04873264  0.00461949 10.5493 < 2.2e-16 ***
## beauty$expersq -0.00069846  0.00010411 -6.7088 2.961e-11 ***
## beauty$educ     0.06836892  0.00593334 11.5228 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```r
beauty$age <- beauty$educ+beauty$exper+6
```

b. Estime una regresión simple: de log(wage) con looks como explicativa con errores estándar robustos (HC1): pegue el resultado abajo e interprete β_1

```r
logwage1 <- lm(log(wage) ~ looks,data = beauty)
coeftest(logwage1, type = 'HC1')
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.496483   0.079610 18.7978  < 2e-16 ***
## looks       0.050951   0.024432  2.0855  0.03723 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Por cada *rating* adicional en *apariencia* aumenta en 5.09% el salario.

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?

```r
exper2 <- beauty$exper^2
logwage2 <- lm(log(wage) ~ looks + exper + (exper2) + educ + age, data = beauty)
coeftest(logwage2, type = 'HC1')
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  0.04627735  0.10533299  0.4393  0.660488
## looks        0.06206811  0.02192318  2.8312  0.004712 **
## exper        0.04873264  0.00478836 10.1773 < 2.2e-16 ***
## exper2      -0.00069846  0.00010767 -6.4870 1.257e-10 ***
## educ         0.06836892  0.00581266 11.7621 < 2.2e-16 ***
## age                  NA          NA      NA        NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No calcula la β para la variable *age*. El problema es que esta variable es una combinación lineal de *educ* y de *exper*, es decir, existe colinearidad perfecta.

d. Solucione el problema en c. y vuelva a estimar su regresión. Pegue el resultado abajo e interprete β_1.

```r
logwage3 <- lm(log(wage) ~ looks + exper + (exper2) + age, data = beauty)
coeftest(logwage3, type = 'HC1')
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -0.36393619  0.12920133 -2.8168  0.004926 **
## looks        0.06206811  0.02192318  2.8312  0.004712 **
## exper       -0.01963628  0.00788322 -2.4909  0.012871 *
## exper2      -0.00069846  0.00010767 -6.4870 1.257e-10 ***
## age          0.06836892  0.00581266 11.7621 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Eliminamos una de las variables colineares perfectamente. Controlando con las demás variables, ahora la β_1 cambia. Por cada *rating* adicional en *apariencia* aumenta en 6.2% el salario.

4. SESGO EN β

b. Ommiting an important variable: sesga a la β, ya que su covarianza con el error no es igual a 0.

5. SESGO EN VARIANZA

a. Heteroskedasticity: que la varianza no sea constante en toda la regresión la sesga.

b. A sample correlation coefficient of 0.95 between two independent variables both included in the model: por multicolinearidad (ya fue explicado).

6. PARTIALLING OUT

```r
educ <- lm(educ ~ exper + tenure, data=wage1)
r1 <- resid(educ)
logwage4r1 <- lm(lwage ~ r1, data=wage1)
logwage5 <- lm(lwage ~ educ + exper + tenure, data=wage1)
```

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.

```r
beta_comparison <- c(summary(logwage4r1)$coefficients[2,1] ,summary(logwage5)$coefficients[2,1])
beta_comparison
```

```
## [1] 0.09202899 0.09202899
```

Lo anterior lo explica el teorema de Frisch-Waugh-Lowell. Intuitivamente, los residuos de la primera regresión lineal terminan de explicar la variación en educación. Así, al proyectar esto en la tercera regresión (controlando por las mismas variables *exper* y *tenure*), obtendras el efecto de la educación en el salario.

7. ECONMATH

a. How many students received a perfect score for the course? What was the average score? Find the means and standard deviations of actmth and acteng, and discuss how they compare.

Número de personas con calificación perfecta:

```r
sum(econmath$score == 100)
```

```
## [1] 0
```

Calificación media:

```r
mean(econmath$score)
```

```
## [1] 72.59981
```

Media y desviación estandar de *actmth* y *acteng*

```r
act_comparison <- data.frame(c(mean(econmath$actmth, na.rm=T), mean(econmath$acteng, na.rm=T)),
c(sd(econmath$actmth, na.rm=T), sd(econmath$acteng, na.rm=T)))
colnames(act_comparison) <- c("mean", "sd")
rownames(act_comparison) <- c("actmth", "acteng")
act_comparison
```

	mean <dbl>	sd <dbl>
actmth	23.21130	3.773354
acteng	22.59459	3.788735

2 rows

Ambas son muy similares en su media y desviación estándar.

b. Estimate a linear equation relating score to colgpa, actmth, and acteng, where colgpa is measured at the beginning of the term. Report the results in a table.

```r
scorelm <- feols(score ~ colgpa + actmth + acteng, data=econmath)
etable(scorelm)
```

	scorelm <chr>
Dependent Var.:	score
(Intercept)	16.17*** (2.800)
colgpa	12.37*** (0.7151)
actmth	0.8834*** (0.1122)
acteng	0.0518 (0.1111)
___________	_______________
S.E. type	IID
Observations	814
R2	0.39723

1-10 of 11 rows Previous 1 2 Next

c. Would you say the math or English ACT score is a better predictor of performance in the economics course? Explain.

actmth es un mejor predictor debido a que tiene mayor efecto sobre la calificación (además, su p-value es mayor).

d. Discuss the size of the R-squared in the regression.

El modelo explica el 39.5% de las variaciones en las calificaciones.

8. MVNORM

```r
# set seed
set.seed(123)
# Set number of observations
n <- 500
#Another example of what is going to happen... Here we create a matrix X
#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
 X1 <- 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 * X1[, 1] + 3.5 * X1[, 2] + u
 #we compute SRF 10000 times using our n values in matrix X
 coefs1[i, ] <- lm(Y ~ X1[, 1] + X1[, 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]

}
```

a. A partir de la marca "From here". Explique cada línea pertinente del código.

El código establece un *loop* que realiza una regresión lineal 10,000 veces para dos distribuciones normales multivariadas. Lo que diferencia a estas dos es su covarianza (en *sigma*). Los coeficientes de cada regresión lineal los guarda en un vector vacío definido anteriormente (mediante coordenadas).

b. Escriba la matriz de varianza – covarianza de los estimadores 1 y 2.

Para la estimación 1:

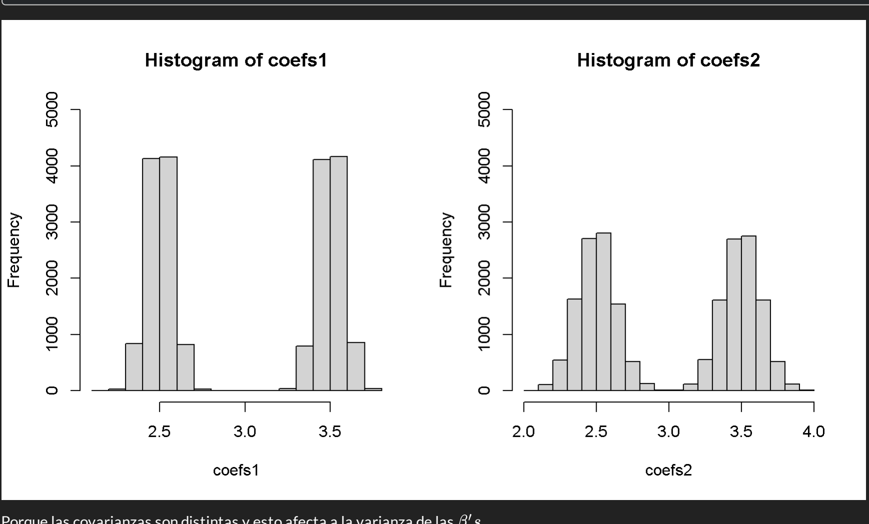
$$\begin{bmatrix} 10.3966129 & 3.2817823 \\ 3.2817823 & 9.9553062 \end{bmatrix}$$

Para la estimación 2:

$$\begin{bmatrix} 10.4933833 & 9.2995875 \\ 9.2995875 & 11.0155659 \end{bmatrix}$$

c. ¿Por qué los histogramas de coefs1 y coefs2 son distintos?

```r
#Histogram of coefficients x1 and x2 for the two 10000s loops of regressions.
par(mfrow=c(1,2))
hist(coefs1, ylim=c(0,5000))
hist(coefs2, ylim=c(0,5000))
```



Porque las covarianzas son distintas y esto afecta a la varianza de las $\beta's$

d. ¿Por qué las varianzas de coefs1 y coefs2 son distintas?

Porque entre más correlacionados (coeficiente de Pearson) estén dos regresores (cercano a $\text{cor}() = \pm 1$), aumenta la varianza de las $\beta's$ (trade-off varianza y sesgo). "*Aumenta el ruido*"

e. ¿Qué parte del código genera estas diferencias?

En *sigma*, define las covarianzas de manera distinta.

f. ¿Qué concepto/supuesto central estamos simulando en este ejercicio?

Multicolinearidad.

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?

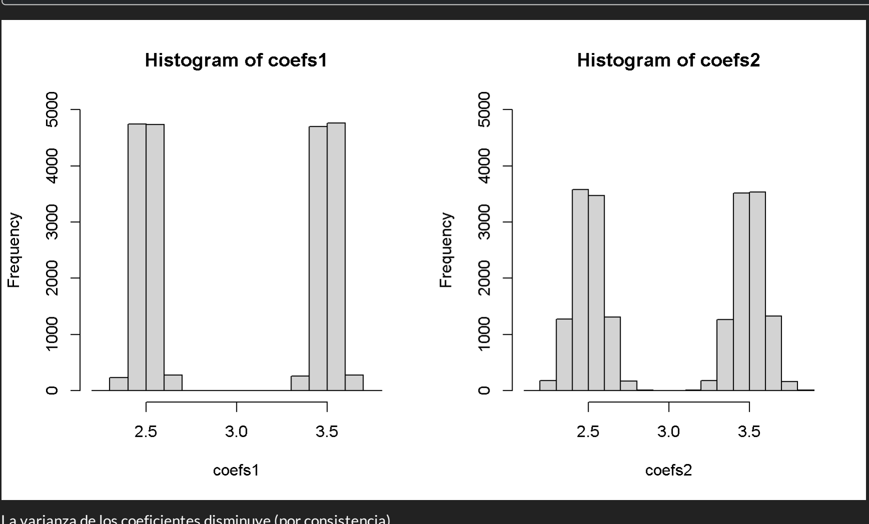
```r
n <- 1000

coefs1 <- cbind("beta_hat_1" = numeric(10000), "beta_hat_2" = numeric(10000))
coefs2 <- coefs1
###From here###

for (i in 1:10000) {

 # for median 50 and 100 and cor(X_1,X_2) = 0.25
 X1 <- 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 * X1[, 1] + 3.5 * X1[, 2] + u
 #we compute SRF 10000 times using our n values in matrix X
 coefs1[i, ] <- lm(Y ~ X1[, 1] + X1[, 2])$coefficients[-1]

 # for cor(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]
}
par(mfrow=c(1,2))
hist(coefs1, ylim=c(0,5000))
hist(coefs2, ylim=c(0,5000))
```



La varianza de los coeficientes disminuye (por consistencia).

$$Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\text{SST}_j(1-R_j^2)}, \ \hat{\sigma}^2 = \frac{\Sigma u_i^2}{n-k-1}$$

Al aumentar *n*, disminuye $\hat{\sigma}^2$. Esto provoca que la varianza de la β disminuya.

h. Aun 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.

Entre mayor sea la varianza de β, es más probable que los muestreos puedan ocurrir en los bordes de la distribución de β. De esta forma, aunque no esté sesgado, puede ser inconsistente.

Sin embargo, mientras mayor sea el número de observaciones, la $\hat{\beta}$ tenderá a β y este problema ya no será relevante:

$$\hat{\beta}_j = \beta_j + \frac{\text{cov}(u_j, x_j)}{\text{var}(x_j)} = \beta_j$$

Por ejemplo, dos histogramas de la distribución de β_1 y β_2 con correlación de 0.85 pero la primera con 100 observaciones y la segunda con 100,000 observaciones.