QUAERERE VERUM

Universidad de San Andrés

Econometría Avanzada

Walter Sosa Escudero
Gastón García Zavaleta

Trabajo Práctico 1

García Vassallo, Hedemann, Iolster, Sury

2022

Trabajo Practico 1 Econometría Avanzada

GARCIA VASSALLO, Felipe Jose; HEDEMANN, Malena; IOLSTER, Madeleine; SURY, Camila

1

1.1

La resolución de este punto se encuentra en el script.

1.2

Obtenemos la matriz de correlaciones entre las variables independientes

	x1	x2	x3	u
x1	1.00	0.02	0.14	-0.05
x2	0.02	1.00	0.01	-0.10
x3	0.14	0.01	1.00	0.00
u	-0.05	-0.10	0.00	1.00

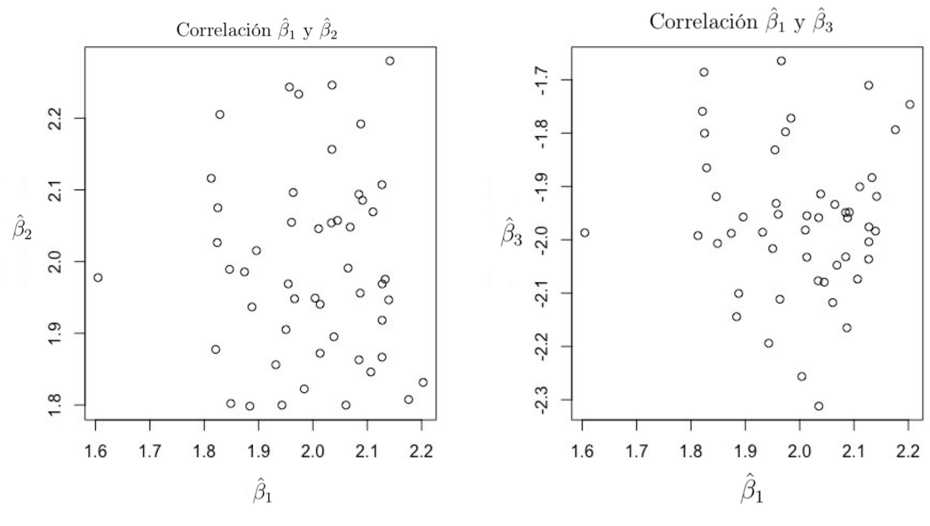
Podemos observar que no existe una correlación alta entre las variables independientes. Por lo tanto, no tenemos problemas de multicolinealidad.

1.3

La resolución de este punto se encuentra en el script.

1.4

A continuación presentamos los gráficos de la correlación entre $\hat{\beta}_1$ y $\hat{\beta}_2$ y entre $\hat{\beta}_1$ y $\hat{\beta}_3$.



En ambos gráficos puede observarse que las observaciones se encuentran lo suficientemente dispersas como para inferir que no tenemos un problema de multicolinealidad entre las variables explicativas. Esto es, no se ve una relación lineal fuerte entre las variables explicativas. Además, concluimos que las estimaciones son buenas ya que sus valores se encuentran cercanos a los verdaderos valores de cada uno de los parámetros β_1 β_2 β_3. Para complementar la intuición gráfica, a continuación calculamos los errores estándares y promedios de las estimaciones.

	Mean	SE
β_0	398,93	11,54
β_1	1,99	0,12
β_2	1,99	0,13
β_3	-1,96	0,14

Puede observarse que los promedios son muy cercanos a los β verdaderos y, además, los estimadores poseen errores estándares relativamente pequeños.

1.5

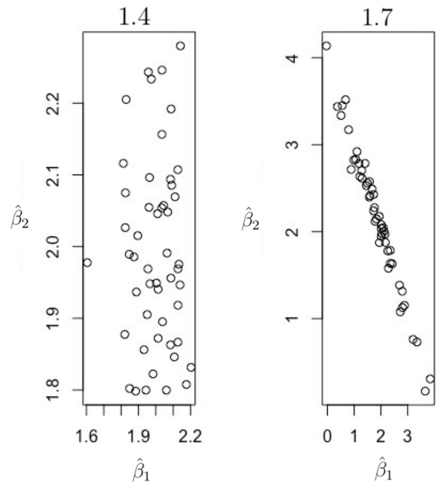
La resolución de este punto se encuentra en el script.

1.6

La resolución de este punto se encuentra en el script.

1.7

A continuación incluimos el gráfico de la correlación entre $\hat{\beta}_1$ y $\hat{\beta}_2$ para el caso en el que x_1 y x_2 correlacionan altamente en comparación al gráfico del punto 1.4, en el que x_1 y x_2 no correlacionan.



Podemos observar que ante la alta correlación entre x_1 y x_2, las estimaciones se aproximan en gran medida a la recta que representa una relación lineal exacta entre estas variables. En relación al gráfico del punto 1.4 en el que no se puede identificar una relación clara entre las variables, en este se ve una clara relación lineal. Además, si separamos al gráfico en 4 cuadrantes a partir de las medias de β_1 y β_2, en el primer gráfico tenemos estimaciones en todos los cuadrantes mientras que en este gráfico solo hay observaciones en los cuadrantes 2 y 4, donde la correlación entre las variables es positiva.

El teorema de Gauss Markov dice que si se cumplen todos los supuestos clásicos, el estimador $\hat{\beta}$ obtenido por el método de MCO es el mejor estimador lineal insesgado (MELI), donde "mejor"hace referencia a que es el más preciso. Uno de los supuestos clásicos es el de no multicolinealidad exacta entre las variables explicativas. Este es el supuesto que garantiza la existencia del estimador de MCO. Ahora, este supuesto clásico (ni ninguno otro) se viola en este ejercicio dado que a pesar de que hay multicolinealidad alta entre x_1 y x_2, la relacion lineal no es exacta. Por lo tanto, sigue aplicando el Teorema de Gauss Markov y los estimadores de β_1 y β_2 siguen siendo MELI.

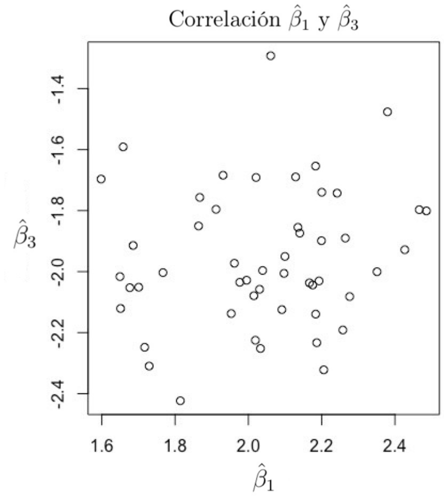
1.8

El problema con el que se esta lidiando es el de multicolinealidad alta entre x_1 y x_2. La única consecuencia es un aumento en la varianza de los estimadores de β_1 y β_2 debido a que aumentan los R_1^2 R_2^2, de los cuales la varianza de los estimadores depende positivamente.

2

2.1

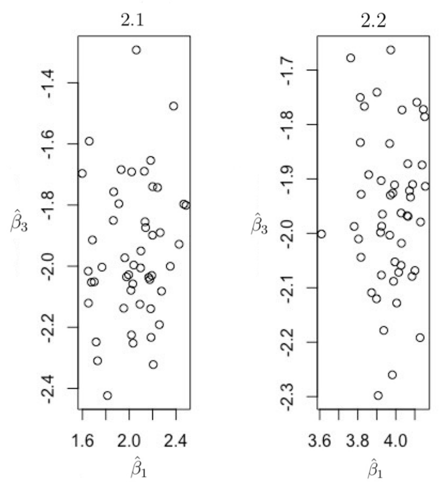
A continuación presentamos el gráfico de correlación entre $\hat{\beta}_1$ y $\hat{\beta}_3$ para el caso en el que X_2 se omite y no correlaciona con X_1.



En primer lugar, no tenemos problema de sesgo. Esto se debe a que la omisión de variables relevantes conduce a sesgos en la estimación solo si estas variables están correlacionadas con los regresores del modelo. Es por esto que, a pesar de que x_2 es una variable relevante -ayuda a explicar a Y- omitirla no genera sesgo porque no correlaciona con las variables explicativas x_1 y x_3 que están en el modelo. En otras palabras, no se viola el supuesto de exogeneidad. Esto puede observarse en el gráfico, ya que no hay un cambio significativo en los valores estimados de β_1 y β_3 con respecto al modelo en el que se incluye a x_2 como regresor (punto 1.4). Además, el promedio de $\hat{\beta}_1$ en este caso es 2.037, en comparación al 1,99 del modelo completo (punto 1.4), lo que confirma la intuición gráfica. Algo similar pasa con el promedio de $\hat{\beta}_3$ que pasa de -1.9642 en el modelo completo a -1.9557 en el modelo en que se omite a x_2. En segundo lugar, tenemos un problema de varianza. Dado que x_2 contribuye a explicar a Y, al omitirla aumenta la varianza del término de error y, en consecuencia, la varianza de los estimadores β_1 y β_3. Al calcular el error estándar de $\hat{\beta}_1$ este aumenta a 0.231 con respecto al 0.12 del modelo con x_2. Lo mismo pasa con $\hat{\beta}_3$ cuyo error estándar pasa de 0.1395 a 0.225. De este modo, si incluyéramos a x_2 en la regresión obtendríamos estimadores más eficientes dado que sus errores estándares disminuirían.

2.2

A continuación contrastamos los gráficos de correlación entre $\hat{\beta}_1$ y $\hat{\beta}_3$ en el caso en que x_1 no tiene alta correlación con x_2 (2.1) y el caso en el que sí (2.2).



En este caso tenemos un problema de sesgo por variable omitida. En particular, el estimador $\hat{\beta}_1$ será sesgado dado que x_2 es una variable relevante para explicar a Y y además esta altamente correlacionada con x_1. De este modo, $\hat{\beta}_1$ no solo da cuenta del efecto de x_1 sobre Y sino que también captura el efecto de x_2 que se encuentra en el término de error. En otras palabras, no se satisface la condición de ortogonalidad entre x_1 y el término de error. Esto puede verse en el gráfico, ya que todas las estimaciones están corridas hacia la derecha en relación al gráfico de la regresión en el que x_2 esta incluida como regresor. Además, esto se confirma al calcular la media de $\hat{\beta}_1$ que pasa de 2.037 a 3.97, lo que da cuenta que el sesgo es positivo. Por otro lado, $\hat{\beta}_3$ no estará sesgado dado que, por construcción, x_2 no correlaciona con x_3.

Con respecto a la varianza del $\hat{\beta}_1$, operan dos efectos contrapuestos. Por un lado, al omitir una variable relevante aumenta la varianza del término de error pero no aumenta tanto en comparación al caso en el que x_2 no correlaciona con x_1, dado que ahora $\hat{\beta}_1$ capta parte del efecto de x_2 sobre Y. Por lo tanto, como la varianza del estimador depende positivamente de la varianza del termino de error, la varianza de $\hat{\beta}_1$ aumenta en relación al modelo en el que x_2 esta incluida. Por otro lado, con respecto a la varianza del estimador en el modelo verdadero, el R_1^2, es decir, el R^2 de regresar x_1 contra las demás variables explicativas, se reduce en gran medida al omitir a x_2. Esto se debe a la alta correlación entre x_2 y x_1. En consecuencia, dependiendo de cual de los efectos sea mas fuerte, la varianza de $\hat{\beta}_1$ aumentará o disminuirá. En este caso al calcular el error estándar de $\hat{\beta}_1$ vemos que la varianza del estimador disminuye ya que pasa de 0.83 en el modelo con x_2 a 0.12 en el caso en que la omitimos. Por lo tanto, gana el efecto de multicolinealidad por sobre el de varianza del termino de error. Por su parte, la varianza de $\hat{\beta}_3$ aumentará dado que en este solo opera el aumento de la varianza del termino de error. Esto se confirma al calcular el error estándar de $\hat{\beta}_3$ que pasa de 0.1382 en el modelo con x_2 a 0.1423 en el caso en el que la omitimos.

3

El estimador de β que se obtiene minimizando la suma de los residuos al cuadrado es el siguiente

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Tomamos esperanza asumiendo que el regresor X es no aleatorio

$$E(\hat{\beta}) = E\left(\frac{\sum x_i y_i}{\sum x_i^2}\right)$$

Definimos $\frac{\sum x_i}{\sum x_i^2} = \sum w_i$. Dado que w_i solo depende de valores de X, por el supuesto de regresor no aleatorio, podemos sacar w_i fuera de la esperanza. Además, como la esperanza es un operador lineal, podemos escribir a la esperanza de la suma como la suma de las esperanzas. Por lo tanto,

$$E(\hat{\beta}) = \sum w_i E(y_i)$$

Luego, sabemos que $y_i = Y_i - \bar{Y}$. Reemplazando por las definiciones de Y_i e$\bar{Y}$ tenemos que $y_i = (\alpha + \beta X_i + \mu_i) - \frac{\sum(\alpha+\beta X_i+\mu_i)}{n}$ Distribuyendo el segundo paréntesis obtenemos $y_i = (\alpha + \beta X_i + \mu_i) - [\frac{n\alpha}{n} + \beta\frac{\sum X_i}{n} + \frac{\mu_i}{n}]$. Lo que a su vez es equivalente a $y_i = (\alpha + \beta X_i + \mu_i) - (\alpha + \beta\bar{X} + \bar{\mu})$ Operando tenemos que $y_i = \beta(X_i - \bar{X}) + \mu_i - \bar{\mu}$ Como $(X_i - \bar{X}) = x_i$ llegamos a que $y_i = \beta x_i + \mu_i - \bar{\mu}$.
De este modo, $E(y_i) = E(x_i\beta + \mu_i - \bar{\mu})$ lo cual, dado que la esperanza es un operador lineal, podemos escribir como: $E(x_i\beta) + E(\mu_i) - E(\bar{\mu})$. Entonces nos queda:

$$E(\hat{\beta}) = \sum w_i[E(x_i\beta) + E(\mu_i) - E(\bar{\mu})]$$

Luego, por definición, $E(\bar{\mu}) = E(\frac{\sum \mu_i}{n})$. Por otro lado, como asumimos que X no es una variable aleatoria podemos sacar a x_i fuera de la esperanza. Y dado que β es un parámetro $E(\beta) = \beta$. Así nos quedaría $x_i E(\beta) = x_i\beta$. Entonces llegamos a:

$$E(\hat{\beta}) = \sum w_i \left[x_i\beta + E(\mu_i) - E\left(\frac{\sum \mu_i}{n}\right)\right]$$

Asumimos que $E(\mu_i)$ es constante e igual a k. En caso de que $k = 0$ se cumple el supuesto de exogeneidad. Entonces, $E\left(\frac{\sum \mu_i}{n}\right) = \frac{nk}{n}$

$$E(\hat{\beta}) = \sum w_i \left(x_i\beta + k - \frac{nk}{n} \right)$$

Dado que las n se cancelan, nos queda:

$$E(\hat{\beta}) = \sum w_i(x_i\beta + k - k)$$

Luego, las k se cancelan y nos queda:

$$E(\hat{\beta}) = \sum w_i x_i \beta$$

Dado que β es constante a través de las observaciones lo podemos sacar de la sumatoria:

$$E(\hat{\beta}) = \beta \sum w_i x_i$$

Dado que habíamos definido que $\frac{\sum x_i}{\sum x_i^2} = \sum w_i$, tenemos $\frac{\sum x_i}{\sum x_i^2} x_i = \frac{x_i^2}{x_i^2}$. Por lo tanto,

$$E(\hat{\beta}) = \beta \frac{x_i^2}{x_i^2}$$

Como $\frac{x_i^2}{x_i^2} = 1$ nos queda:

$$E(\hat{\beta}) = \beta$$

A continuación se encuentran los comandos utilizados para cada ejercicio.

```
library("writexl")
library("readxl")

#Ejercicio 1

#1

#Creo lista con dataframes, cada uno con 4 columnas: x1, x2, x3 y u.
Con las características que pide el tp.
set.seed(1)
lista_data = replicate(n = 50,
                       expr = {data.frame(x1 = runif(100,0,100),
                                          x2 = runif(100,0,100),
                                          x3 = runif(100,0,100),
                                          u =
rnorm(100,0,sqrt(1400)))},
                       simplify = F)

#Hago un for loop para agregarle a cada dataframe la columna de Y
for (i in 1:50) {
  lista_data[[i]]["Y"] = 400 + 2*lista_data[[i]]["x1"]+
2*lista_data[[i]]["x2"] -2*lista_data[[i]]["x3"] + lista_data[[i]]
["u"]
}
rm(i)

#2)

#Hago la matriz de correlaciones entre las variables independientes.
round(cor(lista_data[[1]][1:4]),2)

#3)
#Hago una lista con todas las regresiones
lista_reg = list()
for (i in 1:50) {
  lista_reg[[i]] = lm(lista_data[[i]]$Y~lista_data[[i]]
$x1+lista_data[[i]]$x2+lista_data[[i]]$x3)
}
rm(i)

#Hago una lista con solo los coeficientes de cada regresión
lista_coef = list()
for (i in 1:50) {
  lista_coef[[i]] = coef(lista_reg[[i]])
}
rm(i)

#Paso los coeficientes a dataframe para poder pasarlo a excel
df_regresiones <- data.frame(matrix(unlist(lista_coef),
nrow=length(lista_coef), byrow=TRUE))

#Le cambio las columnas para que tengan sentido en el excel
colnames(df_regresiones) = c("B0","B1","B2","B3")
```

```
#Creo el excel, fijense que ustedes tienen que cambiarle la ruta del
archivo para que tenga sentido
write_xlsx(df_regresiones,"primera_estimación.xlsx")

#4)
#Cargo el excel
coefs <- read_excel("primera_estimación.xlsx")

#Hago gráficos con los coeficientes
plot(coefs$B1,coefs$B2,xlab = "Estimador X1", ylab = "Estimador X2")
plot(coefs$B1,coefs$B3, xlab = "Estimador X1", ylab = "Estimador
X3")

mean(coefs$B0)
sd(coefs$B0)
mean(coefs$B1)
sd(coefs$B1)
mean(coefs$B2)
sd(coefs$B2)
mean(coefs$B3)
sd(coefs$B3)

#5)
set.seed(1)
lista_data2 = replicate(n = 50,
                        expr = {data.frame(x1 = runif(100,0,100),
                                           x2 = runif(100,0,100),
                                           x3 = runif(100,0,100),
                                           u =
rnorm(100,0,sqrt(1400)))},
                        simplify = F)

for (i in 1:50) {
  set.seed(1)
  x2 <- scale(matrix( rnorm(100), ncol=1 ))
  x1 = lista_data2[[i]]["x1"]
  x1 = unlist(x1)
  xs <- cbind(scale(x1),x2)
  c1 <- var(xs)
  chol1 <- solve(chol(c1))
  newx <- xs
  newc <- matrix(
    c(1 , 0.987,
      0.987, 1 ), ncol=2)
  eigen(newc)
  chol2 <- chol(newc)
  xm2 <- newx %*% chol2 * sd(x1) + mean(x1)
  x2 <- xm2[, 2]
  lista_data2[[i]]["x2"] = x2
}
rm(i)

#6)
for (i in 1:50) {
```

```
  lista_data2[[i]]["Y"] = 400 + 2*lista_data2[[i]]["x1"]+
2*lista_data2[[i]]["x2"] -2*lista_data2[[i]]["x3"] +
lista_data2[[i]]["u"]
}
rm(i)

#Hago una lista con todas las regresiones
lista_reg2 = list()
for (i in 1:50) {
  lista_reg2[[i]] = lm(lista_data2[[i]]$Y~lista_data2[[i]]
$x1+lista_data2[[i]]$x2+lista_data2[[i]]$x3)
}
rm(i)

#Hago una lista con solo los coeficientes de cada regresión
lista_coef2 = list()
for (i in 1:50) {
  lista_coef2[[i]] = coef(lista_reg2[[i]])
}
rm(i)

#Paso los coeficientes a dataframe para poder pasarlo a excel
df_regresiones2 <- data.frame(matrix(unlist(lista_coef2),
nrow=length(lista_coef2), byrow=TRUE))

#Le cambio las columnas para que tengan sentido en el excel
colnames(df_regresiones2) = c("B0","B1","B2","B3")

#Creo el excel, fijense que ustedes tienen que cambiarle la ruta del
archivo para que tenga sentido
write_xlsx(df_regresiones2,"segunda_estimación.xlsx")

#7)

#Cargo el excel
coefs <- read_excel("primera_estimación.xlsx")
coefs2 <- read_excel("segunda_estimación.xlsx")

#Hago gráficos con los coeficientes
par(mfrow=c(1,2))
plot(coefs$B1,coefs$B2,xlab = "Estimador X1", ylab = "Estimador X2",
main = "Normal")
plot(coefs2$B1,coefs2$B2,xlab = "Estimador X1", ylab = "Estimador
X2", main = "Corr alta")

#Ejercicio 2
#1

#Hago una lista con todas las regresiones
lista_reg3 = list()
for (i in 1:50) {
  lista_reg3[[i]] = lm(lista_data[[i]]$Y~lista_data[[i]]
$x1+lista_data[[i]]$x3)
}
```

```
rm(i)

#Hago una lista con solo los coeficientes de cada regresión
lista_coef3 = list()
for (i in 1:50) {
  lista_coef3[[i]] = coef(lista_reg3[[i]])
}
rm(i)

#Paso los coeficientes a dataframe para poder pasarlo a excel
df_regresiones3 <- data.frame(matrix(unlist(lista_coef3),
nrow=length(lista_coef3), byrow=TRUE))

#Le cambio las columnas para que tengan sentido en el excel
colnames(df_regresiones3) = c("B0","B1","B3")

#Creo el excel, fijense que ustedes tienen que cambiarle la ruta del
archivo para que tenga sentido
write_xlsx(df_regresiones3,"tercera_estimación.xlsx")

coefs3 = read_excel("tercera_estimación.xlsx")

par(mfrow=c(1,1))
plot(coefs3$B1,coefs3$B3, xlab = "Estimador X1", ylab = "Estimador
X3")

#Hago una lista con todas las regresiones
lista_reg4 = list()
for (i in 1:50) {
  lista_reg4[[i]] = lm(lista_data2[[i]]$Y~lista_data2[[i]]
$x1+lista_data2[[i]]$x3)
}
rm(i)

#Hago una lista con solo los coeficientes de cada regresión
lista_coef4 = list()
for (i in 1:50) {
  lista_coef4[[i]] = coef(lista_reg4[[i]])
}
rm(i)

#Paso los coeficientes a dataframe para poder pasarlo a excel
df_regresiones4 <- data.frame(matrix(unlist(lista_coef4),
nrow=length(lista_coef4), byrow=TRUE))

#Le cambio las columnas para que tengan sentido en el excel
colnames(df_regresiones4) = c("B0","B1","B3")

#Creo el excel, fijense que ustedes tienen que cambiarle la ruta del
archivo para que tenga sentido
write_xlsx(df_regresiones4,"cuarta_estimación.xlsx")

coefs4 = read_excel("cuarta_estimación.xlsx")
```

```
par(mfrow=c(1,2))
plot(coefs3$B1,coefs3$B3, xlab = "Estimador X1", ylab = "Estimador
X3", main = "Normal")
plot(coefs4$B1,coefs4$B3, xlab = "Estimador X1", ylab = "Estimador
X3", main = "Corr alta")

mean(coefs3$B1)
mean(coefs4$B1)
sd(coefs3$B1)
sd(coefs$B1)
sd(coefs$B3)
sd(coefs3$B3)
sd(coefs4$B1)
sd(coefs2$B1)
sd(coefs4$B3)
sd(coefs2$B3)
```