Evaluación de procesos 3

Miguel Roca, Jesús Yuri, Gerson Julca, Fabrizzio Arce, Rodrigo Huamán

Contents

Carga de paquetes

```
library(dplyr)
```

Carga de los datos

```
data_Koop_Tobias <- read.csv(file = "Koop-Tobias.csv", header = T)
```

Ejercicio 1:

Considere la regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, donde $E[\epsilon|x] = 0$ y $E[\epsilon^2|x] = \sigma^2$. Demuestre que $\hat{\beta}_1 = \beta_1 + \frac{\sum_i^n (x_i - \bar{x})\epsilon_i}{\sum_i^n (x_i - \bar{x})^2}$

Sabemos que:

- $\hat{\beta}_1 = \frac{Sxy}{Sxx}$
- $Sxy = \sum_i^n (y_i - \bar{x})y_i$
- $Sxx = \sum_i^n (x_i - \bar{x})^2$
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
- $\sum_i^n (x_i - \bar{x}) = 0$
- $\sum_i^n (x_i - \bar{x})^2 = \sum_i^n (x_i - \bar{x})x_i$

De lo anterior tenemos:

$$\hat{\beta}_1 = \frac{Sxy}{Sxx}$$

$$\hat{\beta}_1 = \frac{\sum_i^n (y_i - \bar{x}) y_i}{\sum_i^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_i^n (y_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum_i^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\sum_i^n (y_i - \bar{x})\beta_0 + \sum_i^n (y_i - \bar{x})\beta_1 x_i + \sum_i^n (y_i - \bar{x})\epsilon}{\sum_i^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \frac{\beta_0 \sum_i^n (y_i - \bar{x}) + \beta_1 \sum_i^n (y_i - \bar{x}) x_i + \sum_i^n (y_i - \bar{x})\epsilon}{\sum_i^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \beta_1 \frac{\sum_i^n (y_i - \bar{x}) x_i}{\sum_i^n (x_i - \bar{x})^2} + \frac{\sum_i^n (y_i - \bar{x})\epsilon}{\sum_i^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i^n (y_i - \bar{x})\epsilon}{\sum_i^n (x_i - \bar{x})^2}$$

Ejercicio 2: La regresión de y respecto a una constante x_1 y x_2 produce el siguiente resultado:

$$\hat{y} = 4 + 0.4x_1 + 0.9x_2, \quad R^2 = \frac{8}{60}, \quad e'e = 520, \quad n = 29,$$

$$X'X = \begin{pmatrix} 29 & 0 & 0 \\ 0 & 50 & 10 \\ 0 & 10 & 80 \end{pmatrix}$$

Calcule la matriz de varianzas y covarianzas estimada (σ^2)

Sabemos que la matriz de covarianzas de $\hat{\beta}$ es $\sum_{\hat{\beta}} = \sigma^2 (X'X)^{-1}$, mientras que $(X'X)$ es conocida, σ^2 necesita ser estimado.

- Una estimación insesgada de la varianza es:

$$\sigma^2 = \frac{SSE}{N - p}$$

donde:
- SSE: Suma de los cuadrados de los residuos
- N: Tamaño de la población
- p: número de variables que intervienen

Luego su estimador viene dado por: $s^2 = \frac{e'e}{n-p}$, de lo expuesto en lo anterior procedemos a realizar el calculo:

```
# Ingresamos el párametro S2
s2 <- (520/26)

# Ingresamos la matriz transpuesta de X
XX <- matrix(data = c(29,0,0,0,50,10,0,10,80), nrow = 3, ncol = 3)

# Calculamos la Inversa de la matriz
XX_inv <- solve(XX)
```

```r
# Determinamos el cálculo de la matriz de varianzas y covarianzas
cov_matriz <- s2 * XX_inv

cov_matriz
```

```
##           [,1]        [,2]        [,3]
## [1,] 0.6896552  0.00000000  0.00000000
## [2,] 0.0000000  0.41025641 -0.05128205
## [3,] 0.0000000 -0.05128205  0.25641026
```

Pruebe la hipótesis que las dos pendientes suman 1.

Hipotesis:

- $H_0 : \beta_1 + \beta_2 = 1$
- $H_0 : \beta_1 + \beta_2 \neq 1$

La hipotesis planteada, se puede escribir de la siguiente manera:

$$H_0 : \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 \end{pmatrix}$$

Luego, bajo los supuestos del modelo clásico, si $H_0 : R\beta = r$, el estadístico de comprobación es:

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)/q}{e'e/(n-k)}$$

```r
R <- matrix(data = c(0, 1, 1), nrow = 1, ncol = 3)
beta_hat <- matrix(data = c(4, 0.4, 0.9), nrow = 3, ncol = 1)
r <- matrix(data = c(1), nrow = 1, ncol = 1)
q <- 1 # Número de filas de R
n <- 29 # Tamaño de la muestra
k <- 3 # Variables intervinientes
ee <- 520

F_fisher_calculado <- ((t(((R %*% beta_hat) - r)) %*%
                          solve((R %*% XX_inv %*% t(R))) %*%
                          ((R %*% beta_hat) - r))/(q)) / ((ee)/(n-k))

F_fisher_calculado <- as.numeric(F_fisher_calculado)

F_fisher_teorico <- qf(0.05, 1, 26, lower.tail = F)

msg_salida <- list(paste("F calculado: ", F_fisher_calculado),
                   paste("F teorico: ", F_fisher_teorico),
                   ifelse(F_fisher_calculado > F_fisher_teorico,
       "Se rechaza H0",
       "No existe suficiente evidencia estadística para rechazar H0"))

msg_salida
```

```
## [[1]]
## [1] "F calculado:  0.159545454545455"
##
```

```
## [[2]]
## [1] "F teorico:  4.22520127312749"
##
## [[3]]
## [1] "No existe suficiente evidencia estadística para rechazar H0"
```

Programe sus cálculos en R

Los programas se desarrollaron en cada ejercicio.

Ejercico 3: El conjunto de datos ExpMercadoLaboral.csv estudia las relaciones entre ingresos, educación, habilidad, y características familiares.

- PERSONID = Persona id (Ordenado de 1 a 2,178),
- EDUC = Educación,
- LOGWAGE = Logaritmo del salario por hora,
- POTEXPER = Experiencia potencial,
- TIMETRND = Tiempo tendencia.
- ABILITY = Habilidad invariante en el tiempo,
- MOTHERED = Educación de la madre,
- FATHERED = Educación del padre,
- BRKHOME = Dummy que indica si tiene padres divorciados,
- SIBLINGS = Número de hermanos

Los datos se estructuran en formato panel de 2,178 individuos con un total de 17,919 observaciones.

a) Sea X_1 una matriz de datos que contiene una constante, educación, experiencia y la habilidad. Sea X_2 una matriz de datos que contiene los años de educación de educación de la madre, los años de educación del padre y el número de hermanos. Sea y el logaritmo del salario por hora.

```r
X1 <- data_Koop_Tobias %>%
  select(EDUC, POTEXPER, ABILITY)

X <- as.matrix(data.frame(Intercept = rep(1, dim(data_Koop_Tobias)[1]),
                          X1))

Y <- as.matrix(data_Koop_Tobias$LOGWAGE)

(Betas <- solve(t(X) %*% X) %*% (t(X) %*% Y))
```

1) Estime la regresión de y sobre X_1, reporte y analice sus coeficientes.

```
##                 [,1]
## Intercept 1.02722913
## EDUC      0.07376210
## POTEXPER  0.03948955
```

```
## ABILITY   0.08289072
```

Por lo tanto la ecuación de la recta ajustada quedaría de la siguiente manera:

$$LogaritmoSalario = 1.0272 + (0.0737)educación + (0.0394)experiencia + (0.0828)habilidad$$

- El logaritmo del salario por hora se incrementa en promedio 0.0737 por cada unidad de incremento en educación, siempre y cuando la experiencia y la habilidad permanezcan constantes.
- El logaritmo del salario por hora se incrementa en promedio 0.0394 por cada unidad de incremento en experiencia, siempre y cuando la educación y la habilidad permanezcan constantes.
- El logaritmo del salario por hora se incrementa en promedio 0.0828 por cada unidad de incremento en la habilidad, siempre y cuando la educación y la experiencia permanezcan constantes.

```
X2 <- data_Koop_Tobias %>%
  select(MOTHERED, FATHERED, SIBLINGS)

X <- as.matrix(data.frame(Intercept = rep(1, dim(data_Koop_Tobias)[1]),
                          X1,
                          X2))

(Betas <- solve(t(X) %*% X) %*% (t(X) %*% Y))
```

2) Estime la regresión de y sobre X_1 y X_2, reporte y analice sus coeficientes.

```
##                     [,1]
## Intercept  0.9695095604
## EDUC       0.0722035023
## POTEXPER   0.0395092803
## ABILITY    0.0774678070
## MOTHERED  -0.0001170215
## FATHERED   0.0054569497
## SIBLINGS   0.0047655699
```

Por lo tanto la ecuación de la recta ajustada quedaría de la siguiente manera:

$$LogaritmoSalario = 0.9695 + (0.0722)educación + (0.0395)experiencia + (0.0774)habilidad - \\ (0.0001)AñosEduMadre + (0.0054)AñosEduPadre + (0.0047)NúmHermanos$$

- El logaritmo del salario por hora se incrementa en promedio 0.0722 por cada unidad de incremento en educación, siempre y cuando las demas varías permanezcan constantes.
- El logaritmo del salario por hora se incrementa en promedio 0.0395 por cada unidad de incremento en experiencia, siempre y cuando las demas varías permanezcan constantes.
- El logaritmo del salario por hora se incrementa en promedio 0.0774 por cada unidad de incremento en habilidad, siempre y cuando las demas varías permanezcan constantes.
- El logaritmo del salario por hora disminuye en promedio 0.0001 por cada Año de educación de la madre, siempre y cuando las demas varías permanezcan constantes.
- El logaritmo del salario por hora se incrementa en promedio 0.0054 por cada Año de educación del padre, siempre y cuando las demas varías permanezcan constantes.
- El logaritmo del salario por hora se incrementa en promedio 0.0047 por cada número de hermano que tenga, siempre y cuando las demas varías permanezcan constantes.

```r
# Regresión de Años de educación de la madre sobre X1
ajuste1 <- lsfit(x = as.matrix(X1),
                 y = as.matrix(X2$MOTHERED),
                 intercept = T)

resid1 <- ajuste1$residuals

# Regresión de Años de educación del padre sobre X1
ajuste2 <- lsfit(x = as.matrix(X1),
                 y = as.matrix(X2$FATHERED),
                 intercept = T)

resid2 <- ajuste2$residuals

# Regresión del número de hermanos sobre X1
ajuste3 <- lsfit(x = as.matrix(X1),
                 y = as.matrix(X2$SIBLINGS),
                 intercept = T)

resid3 <- ajuste3$residuals

# Generamos un data frama con los residuos generados a partir de cada regresión
Residuos <- data.frame(resid1,
                       resid2,
                       resid3)

# Tabla con los 10 primeros residuos
head(Residuos)
```

3) Realice una regresión de cada variable de X_2 sobre X_1 y estime los residuos de cada regresión. Cree una matriz de datos X_2^* de 17,919 filas y 3 columnas ¿Cuál es el promedio de cada residuo?

```
##       resid1    resid2    resid3
## 1 -0.8063414 -1.358467 -1.515250
## 2 -1.9263156 -3.115969 -1.157261
## 3 -1.8879076 -3.060867 -1.211135
## 4 -1.8687035 -3.033316 -1.238072
## 5 -1.8019739 -2.702046 -1.211852
## 6 -1.7827698 -2.674495 -1.238789
```

```r
# Mostramos los promedios de cada residuo
colMeans(Residuos)
```

```
##        resid1        resid2        resid3
##  3.508829e-17  1.054131e-15 -4.124613e-17
```

Como se puede observar, el promedio de los 3 residuos se aproxima a cero, que es lo que se busca.

4) Estime el R^2 en la regresión de y sobre X_1 y X_2. Repita el cálculo en el que se omite el intercepto en X_1. Qué sucede con el R_2?

- Cálculo de R^2 sin omitir el intercepto

```r
Y <- as.matrix(data_Koop_Tobias$LOGWAGE)
X <- as.matrix(data.frame(X1,X2))
```

```r
ajuste <- lsfit(x = X, y = Y, intercept = T)
resid <- ajuste$residuals

SST <- sum((Y-mean(Y))**2)
SSE <- sum(resid**2)

(R2 <- 1-(SSE/SST))
```

```
## [1] 0.1747197
```

```r
n <- dim(X)[1]
k <- dim(X)[2] - 1
R2_Ajust <- 1-(((n-1)/(n-k-1))*(1-R2))
```

- Cálculo de R^2 omitiendo el intercepto

```r
ajuste <- lsfit(x = X, y = Y, intercept = F)
resid <- ajuste$residuals

SST <- sum((Y-mean(Y))**2)
SSE <- sum(resid**2)

(R2 <- 1-(SSE/SST))
```

```
## [1] 0.1365989
```

El R^2 sin considerar el intercepto es menor, esto se debe a que estamos calculando un modelo diferente, ya que estamos asumiendo que el intercepto pasa exactamente por el origen, lo cual en la práctica es practicamente improbable.

```r

ajuste <- lsfit(x = X, y = Y, intercept = T)
resid <- ajuste$residuals

SST <- sum((Y-mean(Y))**2)
SSE <- sum(resid**2)

R2 <- 1-(SSE/SST)
n <- dim(X)[1]
k <- dim(X)[2]

(R2_Ajust <- 1-(((n-1)/(n-k-1))*(1-R2)))
```

5) Calcule el R^2 ajustado para la regresión de y sobre X_1 y X_2, considerando el inter-cepto. reporte y analice sus resultados.

```
## [1] 0.1744433
```

- $R^2_{Ajust} = 0.174$, esto indica que las variable endógenas solo explican en un 17.4% el logaritmo del salario por hora, mientras que el 82.6% se exlican por otras variables que no estamos considerando en el modelo.
- En conclusión el modelo es un mal predictor, pero aún podría ser usado para identificar que variables resultan significaativas (aportan valor) para el modelo.