Solución a la práctica 2 con R

En esta práctica vamos a utilizar los operadores cbind, cov, head, lm, mean, sd, sqrt, str, sum y summary.

El siguiente modelo de regresión simple relaciona el PIB por habitante (variable *pibpc*) con la educación (variable *educ*) y otros factores no observados:

$$pibpc_i = \beta_0 + \beta_1 educ_i + \varepsilon_i \qquad \textbf{(1)}$$

A partir del *Informe sobre el Desarrollo Humano* se han obtenido datos de *pibpc* (en miles de dólares PPA de 2005) y *educ* (en años de educación promedio) correspondientes a una muestra de 180 países para el año 2012 (fichero de datos: *practica2.RData*). Responda a las siguientes cuestiones:

Antes de responder a las preguntas, vamos a ver el tipo de datos que tenemos. Con el código *str(practica2)* obtenemos información sobre el nombre de las variables contenidas en *practica2*, su tipo y sus primeras observaciones, como podemos ver en la Figura 2.1.

Figura 2.1: str(practica2)

```
Classes 'tbl_df', 'tbl' and 'data.frame':       180 obs. of  4 variables:
 $ EDUC  : num  12.6 12 13.3 11.6 12.2 12.5 11.6 11.7 11 11.6 ...
 $ MORTAL: num  3 5 8 4 4 6 4 3 5 3 ...
 $ PIBPC : num  47 34.5 42.5 37.3 34.4 ...
 $ TC    : num  1.1 1.7 0.9 0.4 -0.1 1.1 1.4 0.8 0.7 0 ...
```

Notar que el nombre de las variables aparece en mayúsculas y que son numéricas. Además, con el código *head(practica2)* obtenemos las seis primeras observaciones de todas las variables contenidas en *practica2*, como se ve en la siguiente figura:

Figura 2.2: head(practica2)

	EDUC	MORTAL	PIBPC	TC
1	12.6	3	46.982	1.1
2	12.0	5	34.548	1.7
3	13.3	8	42.486	0.9
4	11.6	4	37.251	0.4
5	12.2	4	34.437	-0.1
6	12.5	6	24.818	1.1

1) Obtenga la recta de regresión MCO e interprete los parámetros estimados.

Estimamos el modelo (1) con el código: *lm(PIBPC ~ EDUC, data = practica2)*. El cual dice que se estime, con los datos contenidos en *practica2*, una ecuación en la que la variable dependiente es *pibpc* y la independiente es *educ*. Por defecto este código indica que se estime una ecuación que incluya término constante. Si ejecutamos el código, obtenemos la estimación del modelo (1) como podemos ver en la Figura 2.3.

Figura 2.3: lm(PIBPC ~ EDUC, data = practica2)

```
Call:
lm(formula = PIBPC ~ EDUC, data = practica2)

Coefficients:
(Intercept)         EDUC
     -7.473        2.614
```

$$\hat{pibpc} = -7.473 + 2.614 \cdot educ$$

Interpretación de $\hat{\beta}_0$: el valor estimado promedio del PIB por habitante cuando los años de educación valen cero es de -7.473 miles de dólares. Como vemos en este ejemplo, $\hat{\beta}_0$ no siempre tiene una interpretación económica razonable.

Interpretación de $\hat{\beta}_1$: el aumento estimado del PIB por habitante, en promedio, cuando los años de formación aumentan en un año es de 2.614 miles dólares.

En la Figura 2.3 observamos que el código ejecutado únicamente nos proporciona las estimaciones de los parámetros. Para ampliar la información de la estimación del modelo (1), utilizamos el operador *summary* y ejecutamos el siguiente código:

summary(lm(PIBPC ~ EDUC, data = practica2))

La información obtenida es la que aparece en la Figura 2.4.

Figura 2.4: summary(lm(PIBPC ~ EDUC, data = practica2))

```
Call:
lm(formula = PIBPC ~ EDUC, data = practica2)

Residuals:
    Min      1Q  Median      3Q     Max
-19.325  -5.822  -1.972   3.316  66.381

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.4732     2.2972  -3.253  0.00137 **
EDUC          2.6136     0.2795   9.349  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.23 on 178 degrees of freedom
Multiple R-squared:  0.3293,    Adjusted R-squared:  0.3256
F-statistic: 87.41 on 1 and 178 DF,  p-value: < 2.2e-16
```

Como podemos observar, ahora obtenemos mucha más información sobre la estimación del modelo (1). Esta información la iremos desgranando en la presente práctica y las futuras. Por otro lado, estos resultados los vamos a necesitar en futuras operaciones de esta práctica. De modo que se hace aconsejable y necesario guardarlas en un objeto. Por tanto, para guardar los resultados en el objeto *mod.1* ejecutamos el siguiente código:

mod.1 <- lm(PIBPC ~ EDUC, data = practica2)

Esto nos permite simplificar el código para obtener los resultados de la Figura 2.4. De modo que el código simplificado es: *summary(mod.1)*.

2) Calcule e interprete la media y la desviación típica del PIB por habitante

El PIB por habitante medio es de 12.52792 miles de dólares y se puede obtener con el código: *mean(practica2$PIBPC)*. Con este código estamos ordenando a RStudio que calcule la media de la variable *pibpc* que está contenida en data.frame *practica2*.

La cuasi-desviación típica del *pibpc*, S_c, se puede obtener con el código *sd(practica2$PIBPC)* y vale 13.67396 miles de dólares. Si llamamos *y* al PIB por habitante, tenemos:

$$S_c = \sqrt{\frac{\sum_{i=1}^{180}(y_i - \bar{y})^2}{179}} = 13.67396.$$

Para obtener la desviación típica, S, basta con multiplicar S_c por $\sqrt{179/180}$.

$$S = \sqrt{\frac{\sum_{i=1}^{180}(y_i - \bar{y})^2}{180}} = S_c \frac{\sqrt{N-1}}{\sqrt{N}} = 13.67396 \frac{\sqrt{179}}{\sqrt{180}} = 13.63593 \text{ miles de dólares}$$

Por tanto, ejecutando *sd(practica2$PIBPC)*sqrt(179/180)* obtenemos que la desviación típica es 13.63593 miles de dólares. Notar que su valor es muy similar a la media del PIB por habitante, lo que indica un elevado grado de dispersión entre los niveles de PIB por habitante de los países de la muestra.

3) ¿En cuánto se estima que aumente el PIB por habitante promedio si los años de educación se incrementan en 1 año?

En 2.614 miles de dólares. Este valor corresponde a la estimación del parámetro que multiplica a la variable *educ*, β_1.

¿Y si se incrementan en 4 años? En $2.614 \cdot 4 = 10.456$ miles de dólares.

4) ¿Cuál es el PIB por habitante estimado cuando *educ* = 8?

Es de 13.439 miles de dólares.

$$\hat{pibpc} = -7.473 + 2.614 \cdot 8 = 13.439$$

5) Calcule e interprete la suma total de cuadrados (STC), la suma de cuadrados de los residuos (SCE) y la suma de cuadrados de la regresión (SCR). Calcule e interprete el coeficiente de determinación.

El numerador de la cuasi-desviación típica del PIB por habitante es la raíz cuadrada de STC. De modo que despejando STC se obtiene su valor.

$$STC = \sum_{i=1}^{180}(y_i - \bar{y})^2 = S_C^2 \cdot (N-1) = (13.67396)^2 \cdot 179 = 33468.91 (\text{miles de dólares})^2.$$

El código para obtener STC es *var(practica2$PIBPC)*179*, o alternativamente *sd(practica2$PIBPC)^2*179*.

Podemos calcular la suma del cuadrado de los residuos ejecutando el código: *sum(mod.1$residuals^2)*. Y obtenemos que la suma del cuadrado de los residuos, SCE, es 22446.13. El código *mod.1$residuals* hace referencia al vector de residuos del modelo (1) estimado. El código *mod.1$residuals^2* nos permite obtener el cuadrado de cada observación del vector de residuos. Notar que el código *^2* ordena elevar al cuadrado. Finalmente, el operador *sum* nos permite sumar el cuadrado del vector de residuos.

Para obtener la suma de cuadrados de la regresión (SCR), es importante tener en cuenta que como el modelo tiene constante se cumple STC = SCE + SCR. Por tanto,

$$SCR = STC - SCE = 33468.91 - 22446.13 = 11022.78$$

El coeficiente de determinación se puede calcular a partir de los datos anteriores.

$$R^2 = 1 - \frac{SCE}{STC} = 1 - \frac{22446.13}{33468.91} = 0.3293$$

*1 - sum(mod.1$residuals^2)/(var(practica2$PIBPC)*179)*

No obstante, no es necesario calcular el R^2 porque aparece en la información sobre la estimación de la ecuación (1) de la Figura 2.4. El ítem *Multiple R-squared* de la Figura 2.4 nos indica el coeficiente de determinación, el cual es 0.3293. Por tanto, R^2 nos indica que la fracción de la variación muestral del PIB por habitante explicada por la función de regresión muestral es 32.93%.

6) Considere ahora los siguientes modelos de regresión múltiple que relacionan el PIB por habitante con la educación y otros factores:

$$pibpc_i = \beta_0 + \beta_1 educ_i + \beta_2 tc_i + \varepsilon_i \quad (2)$$

$$pibpc_i = \beta_0 + \beta_1 educ_i + \beta_2 tc_i + \beta_3 mortal_i + \varepsilon_i \quad (3)$$

donde *tc* es la tasa de crecimiento anual del país en 2012 y *mortal* es la tasa de mortalidad de niños menores de cinco años (por cada 1.000 nacidos vivos).

a. Estime ambos modelos por MCO y compare las estimaciones y la bondad del ajuste de los tres modelos.

A continuación estimamos los modelos (2) y (3). Los resultados de la estimación del modelo (2) los guardamos en el objeto *mod.2* y los del modelo (3) en *mod.3*. Para ello ejecutamos los códigos:

```
mod.2 <- lm(PIBPC ~ EDUC + TC, data = practica2)
mod.3 <- lm(PIBPC ~ EDUC + TC + MORTAL, data = practica2)
```

En la figuras 2.5 y 2.6 aparecen los resultados de la estimación de los modelos (2) y (3) respectivamente. Observamos que el coeficiente de *educ* cambia algo al introducir *tc*

pero recupera su valor inicial cuando además se añade *mortal*. El coeficiente de *tc* no cambia mucho cuando se incluye *mortal*. Por otro lado, al tratarse de modelos con diferente número de regresores (en concreto, cada uno de ellos contiene un regresor añadido a los regresores del modelo previo), para compararlos entre sí en términos de la bondad del ajuste, no es adecuado el uso del coeficiente de determinación R^2. En su lugar debe emplearse el coeficiente de determinación ajustado, R_a^2, cuyo valor aparece en el ítem *Adjusted R-squared* en la información que aparece en las figuras 2.5 y 2.6. El coeficiente de determinación ajustado pasa de 0.3256 en la primera especificación a 0.4923 en la segunda y a 0.5086 en la tercera. En términos únicamente de la bondad del ajuste, esta última especificación sería por tanto la preferida.

Figura 2.5: *summary(mod.2)*

```
Call:
lm(formula = PIBPC ~ EDUC + TC, data = practica2)

Residuals:
    Min      1Q  Median      3Q     Max
-28.726  -5.442  -1.097   4.668  45.946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.4838     2.4519  -7.539 2.37e-12 ***
EDUC          3.3680     0.2615  12.878  < 2e-16 ***
TC            3.3239     0.4311   7.711 8.69e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.743 on 177 degrees of freedom
Multiple R-squared:  0.498,     Adjusted R-squared:  0.4923
F-statistic: 87.79 on 2 and 177 DF,  p-value: < 2.2e-16
```

Figura 2.6: *summary(mod.3)*

```
Call:
lm(formula = PIBPC ~ EDUC + TC + MORTAL, data = practica2)

Residuals:
    Min      1Q  Median      3Q     Max
-29.271  -5.788  -1.778   4.330  45.305

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.01502    4.03123  -2.484  0.01391 *
EDUC          2.61586    0.38535   6.788 1.67e-10 ***
TC            3.30363    0.42417   7.789 5.62e-13 ***
MORTAL       -0.06282    0.02396  -2.622  0.00951 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.585 on 176 degrees of freedom
Multiple R-squared:  0.5168,    Adjusted R-squared:  0.5086
F-statistic: 62.76 on 3 and 176 DF,  p-value: < 2.2e-16
```

b. Calcule el vector de PIB por habitante estimado para cada uno de los tres modelos y compárelos ¿Tienen las mismas medias aritméticas? ¿Por qué? Compárela con la media de la variable dependiente.

Para obtener el vector del PIB por habitante ajustado o estimado ($\hat{y}$) en el primer modelo ejecutamos el código: *mod.1$fitted.values*. Y para obtener la media de todas las observaciones de dicho vector ejecutamos: *mean(mod.1$fitted.values)*.

Del mismo modo, obtenemos el vector del PIB por habitante ajustado en el segundo y tercer modelo. Para obtener la media de las observaciones de los vectores anteriores

ejecutamos respectivamente los códigos: *mean(mod.2$fitted.values)* y *mean(mod.3$fitted.values)*.

¿Tienen las mismas medias aritméticas? Sí, valen 12.52 miles de dólares.

¿Por qué? Compárela con la media de la variable dependiente. Porque estamos estimando modelos por MCO con constante, y hemos demostrado que en estas estimaciones las medias muestrales de las variables estimadas coinciden con la media muestral de la variable dependiente.

c. Calcule el vector de residuos de cada uno de los tres modelos, así como la media muestral de cada una de las series de residuos ¿Es la misma? ¿Por qué? Anteriormente hemos obtenido los residuos del primer modelo con el código *mod.1$residuals*. Del mismo modo obtenemos los residuos de las estimaciones del segundo y tercer modelo. Para obtener la media muestral de los residuos de cada modelo ejecutamos los códigos: *mean(mod.1$residuals)*, *mean(mod.2$residuals)* y *mean(mod.3$residuals)*. Los resultados aparecen en la Figura 2.7.

Figura 2.7: Media de los residuos

```
> mean(mod.1$residuals)
[1] 4.143887e-16
> mean(mod.2$residuals)
[1] -7.20462e-16
> mean(mod.3$residuals)
[1] -3.081531e-16
```

La media muestral debe ser matemáticamente cero en los tres casos porque hemos demostrado teóricamente que cuando estimamos un modelo con constante por MCO, sus residuos suman cero y en consecuencia su media es cero. En la práctica, como R hace muchos cálculos sin incluir todos los decimales, vemos que la media estimada es casi cero en los tres casos. En concreto, R trabaja internamente aproximadamente con 16 decimales.

d. Calcule las covarianzas de los residuos con la variable dependiente PIB por habitante ¿Le parece lógico el resultado? ¿Por qué?

Para obtener la matriz de varianzas-covarianzas primero generamos una matriz que contenga los residuos de cada modelo y la variable *pibpc*. A esta matriz la llamaremos *matriz*. Para este fin ejecutamos:

matriz <- cbind(mod.1$residuals,mod.2$residuals, mod.3$residuals, practica2$PIBPC)

A continuación ordenamos a RStudio que calcule la matriz de varianzas-covarianzas con el código: *cov(matriz)*. Y obtenemos:

Figura 2.8: Matriz de varianzas-covarianzas

```
          [,1]     [,2]     [,3]      [,4]
[1,] 125.39739 93.86738 90.33862 125.39739
[2,]  93.86738 93.86738 90.33862  93.86738
[3,]  90.33862 90.33862 90.33862  90.33862
[4,] 125.39739 93.86738 90.33862 186.97725
```

Notar que el número 1 en [1,] y [,1] se refiere a la variable que está en la primera columna de la matriz. En este caso, los residuos del modelo (1). Del mismo modo 2, 3 y 4 hacen referencia a la variable que está en la segunda, tercera y cuarta columna de la matriz, respectivamente. En este caso, los residuos del segundo y tercer modelo y el PIB por habitante, respectivamente.

Como vimos en teoría, los residuos no tienen covarianza nula con la variable dependiente sino con su valor ajustado o estimado. Según el método MCO, tratamos de descomponer el PIB por habitante en dos componentes, una parte que podemos modelizar a través de las variables explicativas y otra que no podremos modelizar que se queda como un ruido. Cuanto más seamos capaces de modelizar, menos se relacionará el ruido con la variable dependiente. En este caso, aunque el ruido no es observable, sí que podemos ver como la covarianza entre los residuos y el PIB por habitante disminuye conforme el modelo es más completo y la parte que podemos modelizar del PIB por habitante está mejor controlada a través de variables explicativas.