# Solución a la práctica 3 con R

En esta práctica vamos a utilizar los operadores head, I, lenght, lm, str, sum y summary.

**Ejercicio 1. Utilizando datos de 1569 empresas españolas del sector Industria para el año 2014, que se encuentran en el objeto *practica31* del fichero *practica3.RData*, se quiere explicar el coste de la producción vendida (*prod*) en euros utilizando como variables explicativas el número de empleados a tiempo completo (*emp*) y el inmovilizado material neto (*inm*) en euros. El modelo propuesto es el siguiente:**

$$prod_i = \alpha \cdot emp_i^{\beta_1} \cdot inm_i^{\beta_2} \cdot u_i \qquad \textbf{(1)}$$

Antes de responder a las preguntas examinamos el tipo de datos del data.frame *practica31*. Para ello ejecutamos el código *str(practica31)* y obtenemos la salida de la Figura 3.1.

Figura 3.1: str(practica31)

```
Classes 'tbl_df', 'tbl' and 'data.frame':       1569 obs. of  3 variables:
 $ EMP : num  354 1 6 47 401 37 38 116 97 6 ...
 $ INM : num  1236403 1048 51041 568501 167601 ...
 $ PROD: num  10746 27597 63419 71022 153243 ...
```

Observamos que el nombre de las variables aparece en mayúsculas, que todas las variables son numéricas y también las primeras observaciones de cada variable. Además, con el código *head(practica31)* obtenemos las seis primeras observaciones de nuestra muestra como podemos ver en la Figura 3.2.

Figura 3.2: head(practica31)

| | EMP | INM | PROD |
|---|---|---|---|
| 1 | 354 | 1236403 | 10746 |
| 2 | 1 | 1048 | 27597 |
| 3 | 6 | 51041 | 63419 |
| 4 | 47 | 568501 | 71022 |
| 5 | 401 | 167601 | 153243 |
| 6 | 37 | 1291237 | 160968 |

**a) ¿Es un modelo lineal en los parámetros? ¿Cómo se puede transformar el modelo para estimarlo por MCO?**

No es un modelo lineal en los parámetros por lo que es necesario realizar un proceso de linealización. Aplicando logaritmos al modelo (1), resulta el siguiente modelo transformado,

$$\log(prod_i) = \beta_0 + \beta_1 \log(emp_i) + \beta_2 \log(inm_i) + \varepsilon_i \qquad (2)$$

que es un modelo log-log y a diferencia del modelo original ya es lineal en parámetros.

**b) Estime por MCO el modelo transformado e interprete la estimación de los coeficientes $\beta_1$ y $\beta_2$.**

A continuación estimamos el modelo (2) con el operador *lm*, y guardamos los resultados en el objeto *mod.1*. Para ello ejecutamos el siguiente código:

```
mod.1 <- lm(log(PROD) ~ log(EMP) + log(INM), data = practica31)
```

Para obtener información detallada de la estimación del modelo (2), Figura 3.3, ejecutamos el código: *summary(mod.1)*.

Figura 3.3: summary(mod.1)

```
Call:
lm(formula = log(PROD) ~ log(EMP) + log(INM), data = practica31)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2723 -0.5088 -0.0109  0.5352  3.0021

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.99791    0.16290  79.790  < 2e-16 ***
log(EMP)     0.43273    0.02738  15.807  < 2e-16 ***
log(INM)     0.07249    0.01473   4.922 9.48e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8815 on 1566 degrees of freedom
Multiple R-squared:  0.2641,    Adjusted R-squared:  0.2632
F-statistic:   281 on 2 and 1566 DF,  p-value: < 2.2e-16
```

La ecuación del coste de la producción que hemos estimado es:

$$\log(\hat{prod}_i) = 12.99791 + 0.43273 \cdot \log(emp_i) + 0.07249 \cdot \log(inm_i)$$

Dado que es un modelo log-log, $\hat{\beta}_1$ mide la elasticidad del coste estimado de la producción respecto al trabajo. Es decir, la variación porcentual del coste estimado de la producción ante una variación de un 1% del número de empleados a tiempo completo. Como $\hat{\beta}_1$=0.43273, si el número de empleados se incrementa en un 1%, el coste estimado de la producción aumentará aproximadamente en un 0.43273%.

$$\hat{\beta}_1 \approx \frac{\Delta \hat{prod} \,/\, \hat{prod}}{\Delta emp \,/\, emp} = \frac{100 \cdot \Delta \hat{prod} \,/\, \hat{prod}}{100 \cdot \Delta emp \,/\, emp} = \frac{\%\Delta \hat{prod}}{\%\Delta emp}$$

Análogamente, la elasticidad del coste estimado de la producción respecto del inmovilizado neto es $\hat{\beta}_2 = 0.07249$, de manera que si el inmovilizado neto aumenta un 1%, el coste estimado de la producción aumentará aproximadamente 0.07249%.

$$\hat{\beta}_2 \approx \frac{\Delta \hat{prod} \,/\, \hat{prod}}{\Delta inm \,/\, inm} = \frac{100 \cdot \Delta \hat{prod} \,/\, \hat{prod}}{100 \cdot \Delta inm \,/\, inm} = \frac{\%\Delta \hat{prod}}{\%\Delta inm}$$

**c) ¿Cuál es el valor estimado del logaritmo del coste de la producción vendida para una empresa que tiene 80 empleados a tiempo completo y un inmovilizado material neto de 9 millones de euros? ¿Cuál es el coste estimado de la producción vendida para dicha empresa?**

El valor estimado del logaritmo del coste de la producción vendida es:

$$\log(\hat{prod}) = 12.99791 + 0.43273 \cdot \log(80) + 0.07279 \cdot \log(9000000) = 16.05491$$

Se obtiene ejecutando el código: *12.99791 + 0.43273*log(80) + 0.07249*log(9000000).*

El valor estimado del coste de la producción vendida se calcula como la exponencial del valor estimado del logaritmo de dicho coste:

$$\hat{prod} \approx \exp(\log(\hat{prod})) = e^{\log(\hat{prod})} = \exp(16.05491) = e^{16.05491} = 9387669 \text{ euros}$$

Código: *exp(12.99791 + 0.43273*log(80) + 0.07249*log(9000000))*

**d) ¿Qué porcentaje de la variación muestral del logaritmo del coste de la producción es explicado por la función de regresión muestral?**

Este porcentaje viene dado por el coeficiente de determinación, $R^2$. En este caso, el 26.41%, como se puede comprobar en el ítem *Multiple R-squared* de la Figura 3.3. Por tanto, la función de regresión estimada explica el 26.41% de la variación muestral del logaritmo del coste de la producción vendida.

## Ejercicio 2. El siguiente modelo de regresión relaciona el logaritmo del gasto en consumo de los hogares (*lgasto*), el logaritmo de la renta del hogar (*lrenta*), el número de miembros del hogar (*miembros*) y la edad del sustentador principal (*edad*):

$$lgasto_i = \beta_0 + \beta_1 lrenta_i + \beta_2 miembros_i + \beta_3 edad_i + \beta_4 edad_i^2 + \varepsilon_i \qquad \textbf{(3)}$$

**Usando datos de la *Encuesta de Presupuestos Familiares* del año 2013 (base 2006) se han obtenido datos de una muestra de 539 hogares que se encuentran en el objeto *practica32* del fichero *practica3.RData*. Responda a las siguientes cuestiones:**

Al igual que en el Ejercicio 1 examinamos primero el tipo de datos del data.frame *practica32*. Para ello ejecutamos *str(practica32)* y obtenemos Figura 3.4:

Figura 3.4: str(practica32)

```
Classes 'tbl_df', 'tbl' and 'data.frame':       539 obs. of  4 variables:
 $ EDAD    : num  48 37 62 66 74 45 30 55 71 53 ...
 $ LGASTO  : num  11.25 11.44 9.97 10.69 9.78 ...
 $ LRENTA  : num  10.22 9.52 9.33 9.43 9.36 ...
 $ MIEMBROS: num  3 3 2 3 2 2 4 4 2 2 ...
```

Podemos ver que el objeto *practica32* contiene 539 observaciones de cuatro variables numéricas. En la Figura 3.4 también aparecen las primeras observaciones de cada variable. Además, con el código *head(practica32)* vemos las seis primeras observaciones.

Figura 3.5: head(practica32)

| | EDAD | LGASTO | LRENTA | MIEMBROS |
|---|---|---|---|---|
| 1 | 48 | 11.247680 | 10.216400 | 3 |
| 2 | 37 | 11.439470 | 9.522813 | 3 |
| 3 | 62 | 9.965591 | 9.330787 | 2 |
| 4 | 66 | 10.694860 | 9.430921 | 3 |
| 5 | 74 | 9.783515 | 9.362203 | 2 |
| 6 | 45 | 10.193140 | 10.457720 | 2 |

**a) Interprete los efectos parciales o efectos *ceteris paribus* estimados de las variables explicativas.**

A continuación estimamos el modelo (3) y guardamos los resultados en el objeto *mod.2*. Para ello ejecutamos el siguiente código:

*mod.2 <- lm(LGASTO ~ LRENTA + MIEMBROS + EDAD + I(EDAD^2), data = practica32)*

donde el código *I()* permite que los términos del modelo incluyan símbolos matemáticos normales. Para obtener información detallada de la estimación del modelo (3), Figura 3.5, ejecutamos el código: *summary(mod.2)*.

Figura 3.5: summary(mod.2)

```
Call:
lm(formula = LGASTO ~ LRENTA + MIEMBROS + EDAD + I(EDAD^2), data = practica32)

Residuals:
     Min       1Q   Median       3Q      Max
-1.17524 -0.27680 -0.00786  0.26050  1.68454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7175655  0.3893794  12.116  < 2e-16 ***
LRENTA       0.4928541  0.0376045  13.106  < 2e-16 ***
MIEMBROS     0.0789848  0.0168394   4.690 3.47e-06 ***
EDAD         0.0225295  0.0093753   2.403  0.01660 *
I(EDAD^2)   -0.0002471  0.0000902  -2.740  0.00635 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4192 on 534 degrees of freedom
Multiple R-squared:  0.4183,    Adjusted R-squared:  0.414
F-statistic: 96.01 on 4 and 534 DF,  p-value: < 2.2e-16
```

Por tanto, la ecuación del gasto en consumo de los hogares que hemos estimado es:

$$\widehat{lgasto}_i = 4.718 + 0.493 \cdot lrenta_i + 0.079 \cdot miembros_i + 0.0225 \cdot edad_i - 0.0002 \cdot edad_i^2$$

La estimación del efecto *ceteris paribus* del logaritmo de la renta sobre el logaritmo del consumo de los hogares es $\hat{\beta}_1 = 0.493$. Como estas variables están en logaritmos, $\hat{\beta}_1$ se interpreta como la elasticidad del gasto en consumo de los hogares respecto de su renta. Esto es, si la renta de un hogar aumenta 1%, el gasto esperado en consumo de dicho hogar aumentará aproximadamente en 0.493%.

$$\hat{\beta}_1 \approx \frac{\Delta \widehat{gasto} / \widehat{gasto}}{\Delta renta / renta} = \frac{100 \cdot \Delta \widehat{gasto} / \widehat{gasto}}{100 \cdot \Delta renta / renta} = \frac{\%\Delta \widehat{gasto}}{\%\Delta renta}$$

En cuanto a $\hat{\beta}_2$,

$$\hat{\beta}_2 \approx \frac{\Delta \widehat{gasto} / \widehat{gasto}}{\Delta miembros}$$

$\hat{\beta}_2$ representa la semielasticidad del gasto estimado en consumo de los hogares respecto del número de miembros. Multiplicando por 100 para expresar la variación del gasto estimado en consumo de los hogares en porcentaje, se tiene que

$$100 \cdot \hat{\beta}_2 \approx \frac{100 \cdot \Delta \widehat{gasto} / \widehat{gasto}}{\Delta miembros} \Rightarrow 100 \cdot \hat{\beta}_2 \approx \frac{\%\Delta \widehat{gasto}}{\Delta miembros}$$

A partir de los resultados de la estimación obtenemos $100 \cdot \hat{\beta}_2 = 7.9$. Se interpreta como que, *ceteris paribus*, un hogar con un miembro más que otro gasta en promedio un 7.9% más, aproximadamente.

El efecto *ceteris paribus* de la *edad* sobre el logaritmo del gasto estimado en consumo de los hogares es aproximadamente:

$$\frac{\Delta \widehat{lgasto}}{\Delta edad} \approx \hat{\beta}_3 + 2\hat{\beta}_4 \cdot edad_i = 0.0225295 - 2 \cdot 0.0002471 \cdot edad_i$$

Dado que el coeficiente estimado de *edad* es positivo y el de $edad^2$ es negativo, el efecto marginal de *edad* es decreciente, y positivo únicamente hasta cierto valor de *edad*. En concreto, cuando aumenta la edad del sustentador principal en un año, el gasto esperado en consumo varía aproximadamente en:

$$\%\Delta \hat{gasto} \approx 100 \cdot (0.0225295 - 2 \cdot 0.0002471 \cdot edad_i)$$

Nótese que $\hat{\beta}_3$ y $\hat{\beta}_4$ no tienen interpretación por separado.

**b) Para un hogar cuyo sustentador principal tiene 25 años, ¿cuál es la variación esperada en el gasto en consumo si cumple un año más? ¿Y para un hogar que tiene un sustentador principal de 40 años?**

Para un hogar cuyo sustentador principal tiene 25 años, la variación aproximada del logaritmo del gasto esperado en consumo por cumplir un año más es de $0.0225295 - 2 \cdot 0.0002471 \cdot 25 = 0.0101745$. Esto supone un incremento del gasto en consumo esperado del 1.02%, aproximadamente.

De la misma manera, para un hogar que tiene un sustentador principal de 40 años, un año adicional le reportará un incremento aproximado del gasto esperado en consumo de un 0.28%, aproximadamente ( $(0.0225295 - 2 \cdot 0.0002471 \cdot 40) \cdot 100 = 0.27615$ ).

**c) ¿A partir de qué edad del sustentador principal, un año más produce una disminución del gasto en consumo de los hogares? ¿Cuántos individuos de la muestra tienen o superan esa edad? Comente los resultados.**

Tenemos que encontrar el valor de *edad* para el que el efecto marginal de *edad* sobre *lgasto* sea cero. Este es el punto donde

$$0.0225295 - 2 \cdot 0.0002471 \cdot edad = 0 \rightarrow edad = \frac{0.0225295}{2 \cdot 0.0002471} = 45.58782$$

Despejando obtenemos que aproximadamente dicha edad es de 46 años. Por tanto, si el sustentador principal de un hogar tiene 46 o más años, el efecto parcial de la edad cambia de signo y será negativo.

Para calcular el número de hogares cuyos sustentadores principales tienen 46 o más años, ejecutamos el código: *sum(practica32$EDAD >= 46)*. Obtenemos que hay 314 hogares que tienen un sustentador principal de 46 o más años. Estos hogares suponen aproximadamente un 58.26% del total (*sum(practica32$EDAD >= 46)/length(practica32$EDAD)*). Por tanto, en un 58.26% de los hogares el efecto parcial de la variable *edad* es negativo.