

# Prácticas Tema 4: Modelo con variables cualitativas

Ana J. López y Rigoberto Pérez

Departamento de Economía Aplicada. Universidad de Oviedo

**PRACTICA 4.1-** Se dispone de información sobre 16 familias sobre las que se han analizado la renta (en miles de euros), el gasto de alimentación (en miles de euros) y el ámbito de residencia

|    | Y, Gasto en alimentación | X, Renta | Ambito de residencia |
|----|--------------------------|----------|----------------------|
| 1  | 10                       | 30       | Urbano               |
| 2  | 6                        | 18       | Rural                |
| 3  | 9                        | 32       | Rural                |
| 4  | 10                       | 35       | Urbano               |
| 5  | 8                        | 20       | Urbano               |
| 6  | 4                        | 16       | Rural                |
| 7  | 8                        | 24       | Rural                |
| 8  | 11                       | 32       | Urbano               |
| 9  | 12                       | 30       | Urbano               |
| 10 | 13                       | 22       | Urbano               |
| 11 | 9,5                      | 25       | Rural                |
| 12 | 1,5                      | 5        | Rural                |
| 13 | 7                        | 10       | Urbano               |
| 14 | 10                       | 20       | Urbano               |
| 15 | 11                       | 25       | Urbano               |
| 16 | 12                       | 40       | Urbano               |

- a) A partir de esta información estimar un modelo econométrico para explicar el gasto familiar en alimentación
- b) Analizar cuál es el efecto del ámbito de residencia sobre el gasto en alimentación
- c) Se ha propuesto un modelo logit para explicar el ámbito de la residencia familiar en función de los ingresos, estimándose los parámetros:  $\hat{\beta}_1 = -1,54$   $\hat{\beta}_2 = 0,088$ . Estudiar la proporción de aciertos de este modelo.

**SOLUCIÓN:**

a) A partir de esta información estimar un modelo econométrico para explicar el gasto familiar en alimentación

Para estimar un modelo mínimo cuadrático del gasto en alimentación se define una variable dicotómica con valor unitario para las familias con residencia en ámbito urbano y nula en los restantes casos. Con la información disponible se construyen las matrices y vectores siguientes:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 16 & 384 & 10 \\ 384 & 10488 & 264 \\ 10 & 264 & 10 \end{pmatrix}$$

con determinante  $|\mathbf{X}'\mathbf{X}| = 67104$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0,52432 & -0,01788 & -0,0522 \\ -0,01788 & 0,00089 & -0,00572 \\ -0,0522 & -0,00572 & 0,30329 \end{pmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 142 \\ 3730 \\ 104 \end{pmatrix}$$

obteniéndose así los estimadores mínimo-cuadráticos

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 2,32046 \\ 0,20064 \\ 2,78255 \end{pmatrix}$$

según las cuales el efecto marginal del ingreso familiar sobre el gasto familiar en alimentación es de 0,2006, estimándose que las familias residentes en ámbito urbano gastan 2.782 euros más que las ubicadas en zona rural.

Sobre este modelo es posible llevar a cabo un análisis de la varianza (ANOVA) calculando la Variación total, explicada y no explicada

$$\begin{cases} VT = \mathbf{y}'\mathbf{y} - n\bar{Y}^2 = 141,25 \\ VE = \hat{\beta}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2 = 107,0411 \\ VNE = \hat{\mathbf{u}}'\hat{\mathbf{u}} = 34,2089 \end{cases}$$

obteniéndose el coeficiente de determinación

$$R^2 = 1 - \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2} = 0,7578$$

Además, a partir de la variación no explicada se obtiene la varianza residual

$$S^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n - k} = 2,6315$$

Para analizar la significación de este modelo se construye la matriz de varianzas covarianzas que, a partir de los resultados anteriores, se obtiene como

$$S_{\hat{\beta}}^2 = S^2 \mathbf{X}'\mathbf{X}^{-1} = 2,6315 \begin{pmatrix} 0,52432 & -0,01788 & -0,0522 \\ -0,01788 & 0,00089 & -0,00572 \\ -0,0522 & -0,00572 & 0,30329 \end{pmatrix}$$

$$= \begin{pmatrix} 1,3797 & -0,0471 & -0,1374 \\ -0,0471 & 0,0024 & -0,0151 \\ -0,1374 & -0,0151 & 0,7981 \end{pmatrix}$$

Teniendo en cuenta que la diagonal principal de esta matriz contiene las varianzas de los estimadores, sus raíces cuadradas proporcionan las correspondientes típicas, y así pues, el modelo estimado vendría dado por la expresión:

$$\hat{Y} = 2,32046 + 0,2006X + 2,78255D$$

(1,1746)      (0,0485)      (0,89336)

Estos resultados permiten también contrastar directamente la significación de las dos variables explicativas

Así, en el caso del ingreso la hipótesis nula sería  $H_0 : \beta_2 = 0$  frente a la alternativa  $H_1 : \beta_2 \neq 0$  y el contraste se llevaría a través de la expresión  $\frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \simeq t_{n-k}$  cuyo valor muestras es en este caso  $\frac{\hat{\beta}_2}{S_{\hat{\beta}_2}} = 4,1364$

Como consecuencia se obtiene el nivel critico

$$p = P(|t_{13}| > 4,1364) = 0,00117$$

que permite rechazar la nulidad del coeficiente y por tanto ratifica la validez de la variable ingresos familiares para explicar el gasto en alimentación.

De modo análogo, para la variable dicotómica ligada al ámbito urbano se tendría el contraste de significación  $H_0 : \beta_3 = 0$  frente a la alternativa  $H_1 : \beta_3 \neq 0$  y el valor muestral del estadístico t de student seria en este caso  $\frac{\hat{\beta}_3}{S_{\hat{\beta}_3}} = 3,1147$ , cuyo nivel critico es

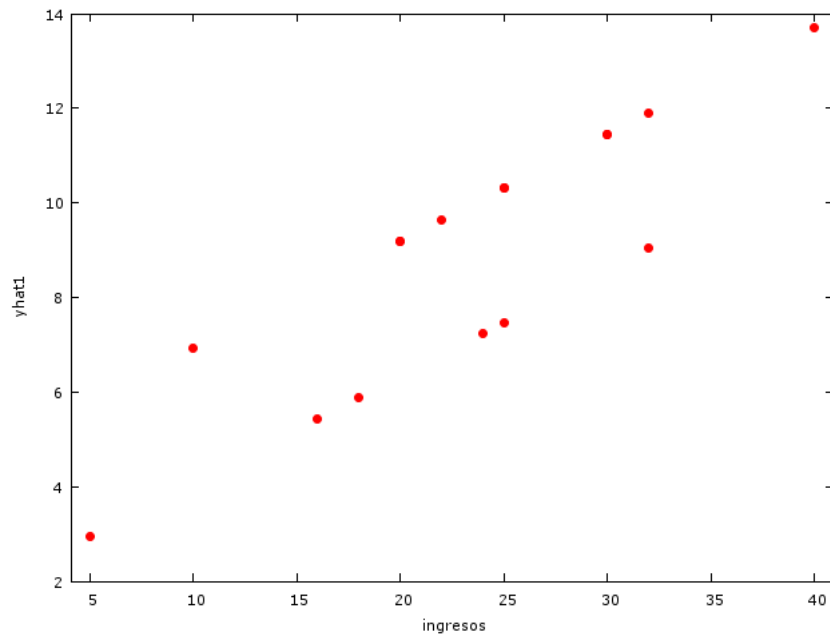
$$p = P(|t_{13}| > 3,1147) = 0,0082$$

Así pues, se rechaza la nulidad del parámetro  $\beta_3$  y se confirma que el ámbito de residencia también es significativo para explicar el gasto familiar aunque no tanto como los ingresos.

Como resultado concluiremos que el gasto estimado en alimentación vendría dado por las siguientes expresiones:

$$\begin{cases} \text{Hogares en ámbito rural (D=0)} & \hat{Y} = 2,32046 + 0,20064X \\ \text{Hogares en ámbito urbano (D=1)} & \hat{Y} = 5,1030 + 0,20064X \end{cases}$$

cuya representacion gráfica es la siguiente



**b)** Analizar cuál es el efecto del ámbito de residencia sobre el gasto en alimentación

En el modelo anteriormente estimado se ha estimado el efecto del ámbito de residencia sobre el consumo fijo, pero sería posible ampliar el planteamiento mediante una variable de interacción, para permitir así estudiar el efecto del ámbito sobre el efecto marginal del ingreso.

Así pues, generamos una variable de interacción que viene dada por el producto de la dummy ligada al ámbito urbano y los ingresos familiares, estimándose así un nuevo modelo mínimo cuadrático en el que las tres variables resultan significativas:

$$\hat{Y} = -0,03875 + 0,318605X + 7,08699D - 0,191644$$

(1,49613)
(0,06888)
(2,1233)
(0,0878)

para el cual se obtendría una nueva descomposición ANOVA, donde lógicamente la variación total es la misma pero ha aumentado la variación explicada

$$\begin{cases} VT = 141,25 \\ VE = 116,762 \\ VNE = 24,4881 \end{cases}$$

Como consecuencia se obtiene un coeficiente de determinación más elevado que el anterior

$$R^2 = 1 - \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2} = 0,826633$$

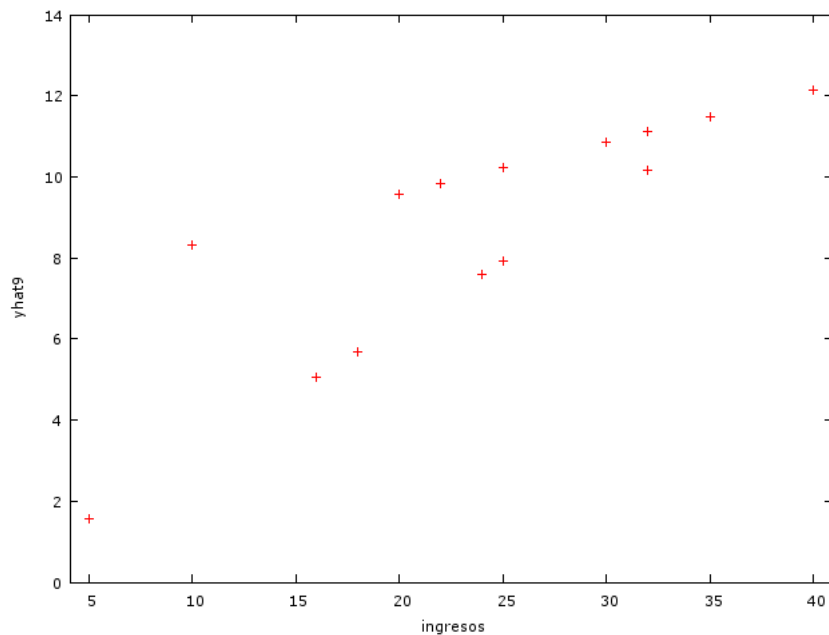
Dado que este modelo tiene un parámetro más la comparación debería realizarse a través del coeficiente de determinación corregido o ajustado, que en

este caso adopta resultado  $\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k} = 0,7832$ , superior al del modelo anterior (que con la misma expresión conduce al coeficiente 0,7205).

Este nuevo modelo proporciona las siguientes expresiones para los hogares en ámbito rural y urbano:

$$\begin{cases} \text{Hogares en ámbito rural (D=0)} & \hat{Y} = -0,0387 + 0,3186X \\ \text{Hogares en ámbito urbano (D=1)} & \hat{Y} = 7,0482 + 0,1269X \end{cases}$$

cuya representación gráfica es la siguiente



c) Se ha propuesto un modelo logit para explicar el ámbito de la residencia familiar en función de los ingresos, estimándose los parámetros:  $\hat{\beta}_1 = -1,54$ ;  $\hat{\beta}_2 = 0,088$ . Estudiar la proporción de aciertos de este modelo.

Con los parámetros estimados se llega a la expresión logit

$$\text{Log} \frac{P_i}{1 - P_i} = -1,54 + 0,088X$$

pudiendo despejarse las probabilidades mediante la expresión  $P_i = \frac{e^{L_i}}{1 + e^{L_i}}$ , que proporciona los resultados de la tabla siguiente, donde la comparación entre el valor observado de D y la probabilidad estimada permite identificar los 11 casos correctamente clasificados (68,8 % del total), así como 4 falsos positivos y 1 falso negativo.

|    | X, Renta | D | Estimación $P_i$ |                |
|----|----------|---|------------------|----------------|
| 1  | 30       | 1 | 0,7503           | OK             |
| 2  | 18       | 0 | 0,51099          | Falso positivo |
| 3  | 32       | 0 | 0,78181          | Falso positivo |
| 4  | 35       | 1 | 0,82351          | OK             |
| 5  | 20       | 1 | 0,55478          | OK             |
| 6  | 16       | 0 | 0,46702          | OK             |
| 7  | 24       | 0 | 0,63924          | Falso positivo |
| 8  | 32       | 1 | 0,78181          | OK             |
| 9  | 30       | 1 | 0,7503           | OK             |
| 10 | 22       | 1 | 0,59774          | OK             |
| 11 | 25       | 0 | 0,65928          | Falso positivo |
| 12 | 5        | 0 | 0,24968          | OK             |
| 13 | 10       | 1 | 0,34069          | Falso negativo |
| 14 | 20       | 1 | 0,55478          | OK             |
| 15 | 25       | 1 | 0,65929          | OK             |
| 16 | 40       | 1 | 0,8787           | OK             |

**PRACTICA 4.2-** Sobre una muestra de 58 países se ha estimado un modelo para la proporción de población universitaria ( $Y$ , en %) en función de la tasa de paro ( $X_2$ , en %), la renta per cápita ( $X_3$ , en miles de euros) y una variable dicotómica con valor unitario para países de la Unión Europea ( $D^{UE}$ ):

$$\hat{Y}_i = -1,443 + 2,74 X_{2i} + 11,084 D_i^{UE} - 3,245 X_{2i} D_i^{UE} + 0,121 X_{3i} D_i^{UE}$$

$(2,93) \quad (0,891) \quad (3,294) \quad (1,024) \quad (0,014)$

$$\hat{u}'\hat{u} = 68,13 \quad ; \quad R^2 = 0,7331$$

- a) ¿Se puede afirmar que la pertenencia a la UE afecta a la tasa de universitarios? ¿Cómo?
- b) Calcular razonadamente el valor del estadístico F de Snedecor interpretando su resultado
- c) ¿Cuál será la proporción de universitarios prevista para un país de la UE con tasa de paro del 5 % y renta per cápita de 26000 euros?

### SOLUCIÓN:

- a) ¿Se puede afirmar que la pertenencia a la UE afecta a la tasa de universitarios? ¿Cómo?

Tal y como se observa en el modelo estimado, la pertenencia a la UE afecta a la tasa de universitarios por sí misma y al interaccionar con las variables Paro y Renta. Más concretamente, podemos expresar los modelos diferenciados para la proporción de universitarios según que los países pertenezcan o no a la UE:

$$\begin{cases} D_i^{UE} = 0 \Rightarrow \hat{Y}_i = -1,443 + 2,74 X_{2i} \\ D_i^{UE} = 1 \Rightarrow \hat{Y}_i = 9,641 - 0,505 X_{2i} + 0,121 X_{3i} \end{cases}$$

- b) Calcular razonadamente el valor del estadístico F de Snedecor interpretando su resultado

El estadístico F permite analizar la validez global del modelo, contrastando la hipótesis nula  $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$  mediante las expresiones:

$$\frac{\frac{\hat{\beta}' \mathbf{X}' \mathbf{y} - n \bar{Y}^2}{k-1}}{\frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{n-k}} \simeq F_{n-k}^{k-1}$$

o alternativamente  $\frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}} \simeq F_{n-k}^{k-1}$

Con la información muestral disponible en nuestro caso se obtiene el resultado  $F^* = 42,68$  que lleva asociado un nivel crítico

$$p = P(F_{53}^4 > 42,68) \approx 0$$

y por tanto permite rechazar la hipótesis nula, confirmando que el modelo propuesto tiene sentido.

**c)** ¿Cuál será la proporción de universitarios prevista para un país de la UE con tasa de paro del 5 % y renta per cápita de 26000 euros?

Dado que el país pertenece a la UE, la variable dummy tomaría valor 1, y con los valores asumidos para las explicativas tasa de paro y renta per cápita se obtendría la predicción:  $\hat{Y}_i = 9,641 - 0,505(5) + 0,121(26) = 10,262$ .