Notas de Clase Sobre Regresión Lineal
Regresión Lineal Múltiple - Parte V

Nelfi González Alvarez

Profesora Asociada Escuela de Estadística

e-mail: ngonzale@unal.edu.co

Facultad de Ciencias, Universidad Nacional de Colombia Sede Medellín

UNIVERSIDAD NACIONAL DE COLOMBIA

Escuela de Estadística

2022

Contenido I

Contenido

Predictor cualitativo y variables indicadoras

Definición 1.1

Una variable indicadora o variable dummy es una variable binaria que toma el valor de 1 cuando un evento de interés es observado y 0 cuando éste no es observado.

En los modelos de regresión con predictores $X's$ cualitativos (nominales u ordinales) se hace necesario el uso de variables indicadoras, con el fin de representar niveles específicos de estas variables observados en una unidad experimental (U.E), ya que en la expresión del MRL es imposible operar directamente con los valores no numéricos del predictor cualitativo.

Contenido

MRL en presencia de una variable explicatoria cualitativa

Sea Y la respuesta de naturaleza numérica y X la variable de tipo categórica, con c categorías o niveles. Definimos las variables I_j, $j = 1, 2, \cdots, c$, tales que

Variable indicadora del nivel j

$$I_j = \begin{cases} 1 & \text{si en la U.E es observada la categoría } j \\ 0 & \text{si en la U.E no es observada la categoría } j. \end{cases} \tag{1}$$

Es decir, I_j es la variable indicadora de la categoría j de X. Tenemos que

- En la i-ésima U.E solo una de las c categorías es observada, por tanto:

Restricción lineal sobre las c indicadoras

$$\sum_{j=1}^{c} I_{ij} = 1, \text{ en cada } i = 1, 2, \ldots, n, \tag{2}$$

donde I_{ij} es el valor de I_j en la i-ésima U.E.

- Por tanto, solo son necesarias $c - 1$ de las I_j para representar a una X cualitativa.

MRL cuando solo hay un predictor y de tipo cualitativo

Considere inicialmente el MRL de Y vs. X, esta última con c categorías,

$$Y_i = \beta_0 + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_c I_{ic} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \tag{3}$$

Sin ninguna restricción lineal sobre los β_j, $j = 1, 2, \ldots, c$, este modelo no es estimable, pues las columnas de su matriz de diseño $\boldsymbol{X}$ son linealmente dependientes (LD), y por tanto, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ no existe y en consecuencia tampoco el estimador MCO, $\widehat{\boldsymbol{\beta}}$. Ejemplo, $c = 6$, n=12, con dos obs. en cada categoría:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E} \Rightarrow \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \\ E_5 \\ E_6 \\ E_7 \\ E_8 \\ E_9 \\ E_{10} \\ E_{11} \\ E_{12} \end{bmatrix} \tag{4}$$

Modelos alternativos:

1. **Eliminar el intercepto β_0:**

$$Y_i = \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_c I_{ic} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{5}$$

entonces $\boldsymbol{\beta_j}$ es la media de Y en la categoría j,

$$\beta_j = E[Y|I_j = 1]$$

2. **Eliminar una de las I_j, por ejemplo, I_c:**

$$Y_i = \beta_0 + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_{c-1} I_{i,c-1} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{6}$$

entonces $\boldsymbol{\beta_j}$, $j \neq c$, es la diferencia de la media de Y en la categoría j con relación a la media de Y en la categória c,

$$\beta_j = E[Y|I_j = 1] - E[Y|I_c = 1],\ j \neq c.$$

3. Introducir la restricción lineal $\sum\limits_{j=1}^{c} \beta_j = 0$:

$$Y_i = \beta_0 + \beta_1 I_{i1} + \beta_2 I_{i2} + \cdots + \beta_c I_{ic} + E_i, \text{ sujeto a } \sum_{j=1}^{c} \beta_j = 0, \text{ con } E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right), \tag{7}$$

entonces $\boldsymbol{\beta_j}$, $j = 1, 2, \ldots, c$, representa el efecto de la categoría j con respecto a la media general de Y representada por el intercepto,

$$\beta_j = E[Y|I_j = 1] - \beta_0$$

Nota 2.1

Recuerde que la función de regresión, es decir, la parte no aleatoria y que es función de las X's, representa la media de la respuesta dado los valores de las X's, de ahí que en las interpretaciones de los párametros β_j previamente presentadas se haga alusión a medias.

MRL con una X_1 cuantitativa y una X_2 cualitativa

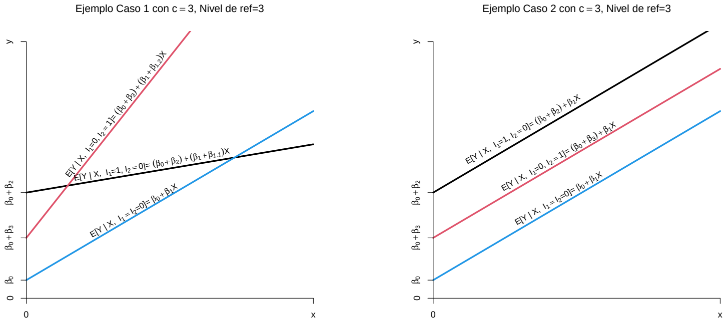
Objetivo: Modelar la relación lineal de Y vs. X_1 (variable predictora cuantitativa), en presencia de X_2, ésta última una variable cualitativa con c categorías (usaremos las indicadoras de las primeras $c-1$ categorías):

Caso 1. *El efecto promedio de X_1 sobre la respuesta Y cambia según la categoría en que X_2 sea observada.* ir a ecuación (10) ir a ecuación (13) ir a ecuación (16)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \cdots + \beta_c I_{i,c-1} + \beta_{1,1} X_{i1} * I_{i1} + \beta_{1,2} X_{i1} * I_{i2} + \cdots + \beta_{1,c-1} X_{i1} * I_{i,c-1} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \quad (8)$$

Caso 2. *El efecto promedio de X_1 sobre la respuesta Y es el mismo en todas las categorías de X_2 pero la media general de Y no es igual en al menos dos de las categorías.* ir a ecuación (11) ir a ecuación (19)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \cdots + \beta_c I_{i,c-1} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \quad (9)$$



(a) (b)

Figura 1: (a) Ilustración caso 1, con $c=3$ y nivel de referencia el 3ro; (b) Ilustración caso 2, con $c=3$ y nivel de referencia el 3ro.

Respuesta media:

1. **En el Caso 1:** Tomando esperanza en la ecuación (8) volver a ecuación (8),

$$\begin{aligned}\mathrm{E}[Y|X_1,X_2] = \beta_0 + \beta_1 X_1 + \beta_2 I_1 + \beta_3 I_2 + \cdots + \beta_c I_{c-1} + \\ \beta_{1,1} X_1 * I_1 + \beta_{1,2} X_1 * I_2 + \cdots + \beta_{1,c-1} X_1 * I_{c-1}\end{aligned} \tag{10}$$

2. **En el Caso 2:** Tomando esperanza en ecuación (9) volver a ecuación (9),

$$\mathrm{E}[Y|X_1,X_2] = \beta_0 + \beta_1 X_1 + \beta_2 I_1 + \beta_3 I_2 + \cdots + \beta_c I_{c-1} \tag{11}$$

Tabla 1: Respuestas medias según niveles de X_2.

Nivel	Valor Indicadoras	En el caso 1	En el caso 2
1	$I_1 = 1,\ I_j = 0,\ \forall\, j \neq 1$	$(\beta_0 + \beta_2) + \left(\beta_1 + \beta_{1,1}\right) X_1$	$(\beta_0 + \beta_2) + \beta_1 X_1$
2	$I_2 = 1,\ I_j = 0,\ \forall\, j \neq 2$	$(\beta_0 + \beta_3) + \left(\beta_1 + \beta_{1,2}\right) X_1$	$(\beta_0 + \beta_3) + \beta_1 X_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$c-1$	$I_{c-1} = 1,\ I_j = 0,\ \forall\, j \neq c-1$	$(\beta_0 + \beta_c) + \left(\beta_1 + \beta_{1,c-1}\right) X_1$	$(\beta_0 + \beta_c) + \beta_1 X_1$
c	$I_j = 0,\ \forall\, j = 1,2,\ldots,c-1$	$\beta_0 + \beta_1 X_1$	$\beta_0 + \beta_1 X_1$

Algunas pruebas de interés

En el Caso 1:

1. En el modelo dado en la ecuación (8), probar si la relación lineal de Y vs. X_1 no difiere según categoría o nivel de X_2 **(las c rectas de regresión son coincidentes, es decir, tienen mismo intercepto y misma pendiente)**:

$$\begin{aligned} &H_0: \beta_2 = \beta_3 = \cdots = \beta_c = \beta_{1,1} = \beta_{1,2} = \cdots = \beta_{1,c-1} = 0 \\ &H_1: \text{algún } \beta_j \neq 0,\ j = 2, \ldots, c, \text{ y/o algún } \beta_{1k} \neq 0,\ k = 1, \ldots, c-1. \end{aligned} \tag{12}$$

El modelo reducido (MR) bajo H_0 es: volver a ecuación (8)

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i,\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right). \tag{13}$$

El estadístico de prueba es:

$$F_0 = \frac{\left[\text{SSE}_{(\text{MR})} - \text{SSE}_{(\text{MF})}\right]/\nu}{\text{MSE}_{(\text{MF})}} = \frac{\text{SSR}\left(I_1, \ldots, I_{c-1}, X_1 * I_1, \ldots, X_1 * I_{c-1} | X_1\right)/[2(c-1)]}{\text{MSE}\left(X_1, I_1, \ldots, I_{c-1}, X_1 * I_1, \ldots, X_1 * I_{c-1}\right)} \tag{14}$$

Criterio de rechazo: Como bajo H_0 y $E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$, $\boldsymbol{F_0 \sim f_{2(c-1),n-2c}}$, se rechaza H_0 con VP si se cumple que $\boldsymbol{P\left(f_{2(c-1),n-2c} > F_0\right)}$ es pequeño.

2. En el modelo dado en la ecuación (8), probar si el efecto medio o cambio medio en la respuesta por unidad de cambio en X_1 es igual para las c categorías de X_2 **(las c rectas de regresión son paralelas, es decir, tienen misma pendiente)**:

$$\begin{aligned} H_0 &: \beta_{1,1} = \beta_{1,2} = \cdots = \beta_{1,c-1} = 0 \\ H_1 &: \text{algún } \beta_{1k} \neq 0,\ k = 1, \ldots, c-1. \end{aligned} \tag{15}$$

El modelo reducido (MR) bajo H_0 es: volver a ecuación (8)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 I_1 + \beta_3 I_2 + \cdots + \beta_c I_{,c-1} + E_i \; E-i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right). \tag{16}$$

El estadístico de prueba es:

$$F_0 = \frac{\left[\text{SSE}_{(\text{MR})} - \text{SSE}_{(\text{MF})}\right]/\nu}{\text{MSE}_{(\text{MF})}} = \frac{\text{SSR}\left(X_1 * I_1, \ldots, X_1 * I_{c-1} | X_1, I_1, \ldots, I_{c-1}\right)/(c-1)}{\text{MSE}\left(X_1, I_1, \ldots, I_{c-1}, X_1 * I_1, \ldots, X_1 * I_{c-1}\right)} \tag{17}$$

Criterio de rechazo: Como bajo H_0 y $E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$, $\boldsymbol{F_0 \sim f_{(c-1),n-2c}}$, se rechaza H_0 con VP si se cumple que $\boldsymbol{P\left(f_{(c-1),n-2c} > F_0\right)}$ es pequeño.

En el Caso 2: En el modelo dado en la ecuación (9) (donde las rectas son paralelas), probar si la respuesta media no difiere según niveles de X_2, en presencia de X_1 **(las rectas son coincidentes, tienen mismo intercepto)**.

$$\begin{aligned} &H_0 : \beta_2 = \beta_3 = \cdots = \beta_c = 0 \\ &H_1 : \text{algún } \beta_j \neq 0,\ j = 2, \ldots, c. \end{aligned} \tag{18}$$

El modelo reducido (MR) bajo H_0 es: volver a ecuación (9)

$$Y_i = \beta_0 + \beta_1 X_1 + E_i\ E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right). \tag{19}$$

El estadístico de prueba es:

$$F_0 = \frac{\left[\mathrm{SSE}_{(\mathrm{MR})} - \mathrm{SSE}_{(\mathrm{MF})}\right]/\nu}{\mathrm{MSE}_{(\mathrm{MF})}} = \frac{\mathrm{SSR}\left(I_1, \ldots, I_{c-1} | X_1\right)/(c-1)}{\mathrm{MSE}\left(X_1, I_1, \ldots, I_{c-1}\right)} \tag{20}$$

El criterio de rechazo es: Como bajo H_0 y $E_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$, $\boldsymbol{F_0 \sim f_{c-1,n-c-1}}$, se rechaza H_0 con VP si se cumple que $\boldsymbol{P\left(f_{c-1,n-c-1} > F_0\right)}$ es pequeño.

Nota 2.2

Ver problema de aplicación Sección 6.4 en Capítulo 6, Notas de Clase.

Kutner, M. H., Natchtsheim, C. J., Neter, J. and Li, W., (2005). *Applied Linear Statistical Models, 5th. ed.*. McGraw-Hill Irwing, New York.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, 5th ed. Wiley, New Jersey.