Apuntes de modelos lineales

Clase N° 4

Ernesto Ponsot Balaguer
Dr. en Estadística, MSc. en Estadística Aplicada, Ing. de Sistemas
http://webdelprofesor.ula.ve/economia/ernesto
E-mail: ernesto@ula.ve

Departamento de Estadística. Universidad de Los Andes. Mérida, Venezuela

Mayo de 2012

Contenido

El modelo lineal general (LM)

Al postulado:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \tag{1}$$

donde

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

tal que $\boldsymbol{X}_{n\times p}$ es una matriz de constantes conocidas (**variables independientes o explicativas**), $\beta_{p\times 1}$ es un vector columna de **parámetros** desconocidos (a estimar), $\boldsymbol{\epsilon}_{n\times 1}$ es un vector columna aleatorio no observable (**error**), con $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ y $\mathrm{V}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}$, y $\boldsymbol{Y}_{n\times 1}$ es un vector columna aleatorio observable (**variable respuesta**), se le conoce como **Modelo Lineal General**.

LM...

Algunas precisiones:

1. La idea es que se cuenta con n observaciones y se postulan p parámetros que deben ser estimados. Se supone que combinaciones lineales de los parámetros ponderados, más un cierto error, pueden reproducir apropiadamente la respuesta observada.
2. La ecuación (1) es **"lineal"** en los parámetros. No hay restricciones de linealidad sobre los elementos de $\boldsymbol{X}$, luego, la expresión $\sum_{j=1}^{p} x_{ij}\beta_j$ con por ejemplo $x_{ij} = z_i z_j$ o bien $x_{ij} = z_{ij}^2$, con z_i, z_j, z_{ij} constantes conocidas, es perfectamente válida en el contexto del LM.
3. Sin embargo el postulado $\sum_{j=1}^{p} x_{ij}\beta_j^2$ no es válido, ni tampoco lo es, por ejemplo, el postulado $\mathrm{E}[y_{ijk}] = \sum_i \sum_j x_{ijk}\beta_i\beta_j$.

LM...

...

4. $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{X}\beta$ y $\mathrm{V}[\boldsymbol{Y}] = \boldsymbol{\Sigma}$
5. En general, $\boldsymbol{\Sigma} = [\sigma_{ij}]$ para $i, j = 1, \cdots, n$, es una matriz llamada de **varianzas y covarianzas**, simétrica. $\mathrm{Cov}[\epsilon_i, \epsilon_j] = \sigma_{ij}$ y $\mathrm{Cov}[\epsilon_i, \epsilon_i] = \sigma_{ii} = \sigma_i^2 = \mathrm{V}[\epsilon_i]$.
6. Entonces, una versión univariada equivalente del modelo lineal general, puede ser expuesta como:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, 2, \cdots, n$$

con $\mathrm{E}[\epsilon_i] = 0$ y $\mathrm{V}[\epsilon_i] = \sigma_i^2$.

LM...

...

7. En (1) la situación varía dependiendo de:
   - La distribución de probabilidades de ϵ.
   - La estructura de $\boldsymbol{\Sigma}$.
   - La estructura y rango de $\boldsymbol{X}$.
8. Si $n = p$ el modelo se dice **saturado**, se obtiene el mejor ajuste posible de los parámetros, pero al no dejar lugar a la variabilidad, no es posible estimar el error que se comete y por lo tanto el caso tiene muy poca utilidad estadística. Los mismos inconvenientes y algunos otros se producen si $n < p$ (el modelo está **sobreparametrizado**). En consecuencia, por lo general se prefiere $n > p$.
9. La matriz $\boldsymbol{X} = [x_{ij}]$, con $x_{ij} \in \mathbb{R}$ para $i = 1, \cdots, n$ y $j = 1, \cdots, p$ caracteriza el modelo que ha decidido el investigador.

LM...

Algunas palabras sobre $\boldsymbol{X}$

1. Cuando $x_{ij} \in \{0, 1\}\ \forall j$, indica la ausencia (0) o presencia (1) de un factor en el modelo, postulado para explicar la respuesta. Se habla entonces del **modelo de diseño de experimentos** y de factores explicativos en lugar de variables.
2. Por otra parte, cuando x_{ij} no está restringido a tomar valores en el conjunto $\{0, 1\}$, al menos para algún j, se habla entonces del **modelo de regresión** y de variables explicativas o independientes, en lugar de factores.
3. En general, el modelo de diseño es tal que $r(\boldsymbol{X}) < p$, mientras que en el de regresión, $r(\boldsymbol{X}) = p$ (esto es, $\boldsymbol{X}$ es de rango completo por columnas). Esto marca la **diferencia** entre ambos tipos de modelos lineales y condiciona el tratamiento matemático que será necesario para su desarrollo.

LM...

Ejemplo 1

Un investigador está interesado en estudiar tres variedades de maíz (digamos $1, 2, 3$) y dos tipos de fertilizantes (digamos $1, 2$), en términos de la productividad, medida con el número de mazorcas obtenidas al momento de la cosecha. Para ello, propone el siguiente experimento: un campo homogéneo, en el sentido de que toda el área tiene las mismas condiciones de suelo, humedad, acceso a la luz solar y riego, entre otras, es dividido en dos parcelas de igual área, cada una de las cuales se fertiliza con los productos 1 y 2 respectivamente. A su vez, cada parcela se subdivide en tres sub-parcelas, también de igual área, a cada una de las cuales se les siembra una variedad distinta de maíz ($1, 2, 3$ respectivamente), empleando el mísmo número de semillas en cada caso.

LM...

Ejemplo 1

La siguiente tabla contiene el número de mazorcas cosechadas finalmente.

Tabla : N° de mazorcas cosechadas

	Variedades			
Fertilizantes	1	2	3	Total
1	42	36	38	116
2	44	38	41	123
Total	86	74	79	239

Ahora bien, los modelos...

Debe ser claro que las observaciones se corresponden, en este caso, con las sub-parcelas definidas. Luego, el número de mazorcas de maíz en cada una de ellas será la variable a modelar de forma lineal. Entonces:

Modelos...

1. Supongamos que el investigador se olvida de su preocupación por variedades y fertilizantes y sólo se preocupa por la producción. Entonces tiene un experimento sin factores y con 6 réplicas. Luego, un modelo podría ser:

$$y_i = \mu + \epsilon_i, \quad i = 1, 2, \cdots, 6$$

Nótese que el modelo implica la idea de que la producción de maíz depende simplemente de una constante (la media teórica), más un cierto error.

Ahora bien, los modelos...

Modelos...

2. Ahora, supongamos que se olvida de las variedades y quiere concentrarse en los fertilizantes. Tiene un experimento con un solo factor en dos niveles y 3 réplicas para cada uno. Luego, otro modelo podría ser:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, 2; j = 1, 2, 3$$

En este caso, α_i representa el efecto del i-ésimo fertilizante, que se asume el mismo en cada sub-parcela que haya recibido el mismo producto.

Ahora bien, los modelos...

Modelos...

3. En esta oportunidad supongamos que se olvida de los fertilizantes y se concentra en las variedades. Tiene, como antes, un experimento con un solo factor, pero ahora en tres niveles con 2 réplicas para cada uno. Así, otro modelo podría ser:

$$y_{ij} = \mu + \gamma_j + \epsilon_{ij}, \quad i = 1, 2; j = 1, 2, 3$$

En este caso, γ_j representa el efecto de la j-ésima variedad, asumido como el mismo en cada sub-parcela que la tenga sembrada.

Ahora bien, los modelos...

Modelos...

4. Ahora, supongamos que no se olvida de los fertilizantes ni de las variedades. Tiene entonces un experimento de dos factores con una réplica para cada combinación. Luego, otro modelo podría ser:

$$y_{ij} = \mu + \alpha_i + \gamma_j + \epsilon_{ij}, \quad i = 1, 2; j = 1, 2, 3$$

5. Y si además del efecto de los fertilizantes y de las variedades, por separado, el investigador sospecha que puede haber un efecto por la **interacción** entre ellos, un último modelo sería:

$$y_{ij} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{ij}, \quad i = 1, 2; j = 1, 2, 3$$

Ahora bien, los modelos...

¿Es $y_i = \mu + \epsilon_i,\ i = 1, 2, \cdots, 6$ un LM?

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 42 \\ 36 \\ 38 \\ 44 \\ 38 \\ 41 \end{bmatrix}, \quad \boldsymbol{X} = \boldsymbol{J} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \end{bmatrix} = \begin{bmatrix} \beta_1 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

Modelos...

¿Es $y_{ij} = \mu + \alpha_i + \epsilon_{ij},\ i = 1, 2; j = 1, 2, 3$ un LM?

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 42 \\ 36 \\ 38 \\ 44 \\ 38 \\ 41 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}$$

Modelos...

¿Es $y_{ij} = \mu + \gamma_j + \epsilon_{ij},\ i = 1, 2; j = 1, 2, 3$ un LM?

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 42 \\ 36 \\ 38 \\ 44 \\ 38 \\ 41 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}$$

Modelos...

¿Es $y_{ij} = \mu + \alpha_i + \gamma_j + \epsilon_{ij},\ i = 1, 2; j = 1, 2, 3$ un LM?

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 42 \\ 36 \\ 38 \\ 44 \\ 38 \\ 41 \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}$$

Modelos...

¿Es $y_{ij} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{ij},\ i = 1, 2; j = 1, 2, 3$ un LM?

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$\boldsymbol{\beta} =$
$[\mu, \alpha_1, \alpha_2, \gamma_1, \gamma_2, \gamma_3, (\alpha\gamma)_{11}, (\alpha\gamma)_{12}, (\alpha\gamma)_{13}, (\alpha\gamma)_{21}, (\alpha\gamma)_{22}, (\alpha\gamma)_{23}]'$.