

# Modelos para la media

Guillermo Ayala Gallego

## Modelos para la media

Guillermo Ayala Gallego

Invalid Date

## Datos y problema

### ¿Qué información tenemos?

- Tenemos información (numérica, categórica) sobre una serie de muestras u observaciones.
- Para la  $i$ -ésima observación tenemos  $p$  variables que recogemos en el vector  $(\mathbf{x}_i \in \mathbb{R}^p)$  donde  $(\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix})$ .
- $(\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip}))$ .
- Estas variables pueden ser numéricas o categóricas.
- Pueden valores fijados por nosotros (dosis de medicación por ejemplo) o bien observados y por lo tanto realizaciones de variables aleatorias.

## Variable respuesta

- En lo que sigue consideramos que hay una variable **importante**.
- La variable importante (para el experimentador) se le llama **variable respuesta**.
- En denominaciones más clásicas, **variable dependiente**.
- El resto de variables del estudio (las que recogemos en  $(\mathbf{x}_i)$ ) son las variables predictoras o independientes.
- Nuestra información está formada por  $((\mathbf{x}_i, y_i))$  con  $(i=1, \dots, n)$ .

## ¿Qué problema queremos resolver?

- Una respuesta fácil es decir que queremos conocer el valor de la variable respuesta utilizando las variables predictoras.
  - La respuesta anterior es **falsamente** simple.
  - ¿Solamente queremos conocer el valor de la respuesta?
  - Quizás estamos pensando en un futuro en donde conozcamos las variables predictoras y nos interese saber cuál será la respuesta correspondiente.
  - ¿El valor exacto de la respuesta?
  - ¿La media de la respuesta?
  - ¿Un valor numérico que aproxime cada una de estas cantidades o bien un intervalo que las contenga?

## Media condicionada

- Los valores observados  $(y_i)$  consideraremos que son realizaciones de una variable aleatoria  $(Y)$ .
  - Realmente en lo que sigue modelizaremos el comportamiento aleatorio de la variable  $(Y_i)$  condicionada a los valores observados  $(\mathbf{x}_i)$ .
  - Es decir, nuestro interés estará en **la distribución condicionada**.
  - Los valores  $(Y_i)$  serán independientes entre sí.
  - El interés se centrará (fundamentalmente) en la dependencia de la media de  $(Y_i)$  respecto de las covariables.
  - Nuestro interés fundamental (pero no único) estará en conocer las medias condicionadas  $(\mu_i = E[Y_i | \mathbf{x}_i])$ .

## Datos de Galton

- Históricamente podemos considerar el origen de este tipo de modelos.
- El problema que se plantea Galton es estudiar la posible relación que puede tener la estatura de un hijo o hija en edad adulta con las estaturas de sus padres.

## Datos de Galton $(\dots)$

### Modelos sobre la media

- Nos planteamos conocer la media de la respuesta aleatoria.
- Pretendemos conocer la media de la respuesta aleatoria **condicionada** a los valores de las variables predictoras.
- Denotamos  $(\mu_i = E[Y_i | \mathbf{x}_i])$ .

## Dependencia lineal

### ¿Qué tipos de dependencia vamos a considerar?

- En la mayor parte de los casos dependencias de tipo lineal.
- Suponemos dos predictores  $(\mathbf{x} = (x_1, x_2)^T)$  tales que  $(x_1)$  es numérico y el segundo es una variable categórica binaria codificada con 1 y 0.
- ¿Cómo modelizamos la dependencia de la media condicionada respecto de  $(x_1)$ ?
- Una dependencia lineal vendría dada como  $[\mu_i = \beta_0 + \beta_1 x_{i1}]$ .

$(\dots)$

### ¿Y la dependencia de las $(\mu_i)$ respecto de la variable binaria?

- Obviamente simplemente tenemos dos valores.
- Un modo simple es  $[\mu_i = \beta_0 + \beta_2 x_{i2}]$ .
- Cuando  $(x_{i2} = 0)$   $[\mu_i = \beta_0]$
- Cuando  $(x_{i2} = 1)$   $[\mu_i = \beta_0 + \beta_2]$

### ¿Y las dos variables predictoras conjuntamente consideradas?

- El modelo más sencillo sería:  $[\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}]$ .
- Cuando  $(x_{i2} = 0)$  entonces  $[\mu_i = \beta_0 + \beta_1 x_{i1}]$ .
- Cuando  $(x_{i2} = 1)$  entonces  $[\mu_i = \beta_0 + \beta_2 + \beta_1 x_{i1}]$ .

### ¿Como podemos expresar la dependencia de ambas covariables?

- Otra vez recurrimos a la opción de expresarlo de un modo lineal.  $[\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}x_{i2}]$
- Cuando  $(x_{i2} = 0)$  tenemos  $[\mu_i = \beta_0 + \beta_1 x_{i1}]$ .
- Cuando  $(x_{i2} = 1)$  tenemos  $[\mu_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1}]$ .

## Predictor categórico y variables dummy

- Cuando consideramos una variable predictora categórica con más de dos categorías entonces es habitual codificarla utilizando variables tontas.
- Si la variable predictora categórica tiene  $(I)$  categorías entonces se elige una categoría de referencia (por ejemplo, la primera).
- Las variables binarias asociadas a la variable original  $(x)$  serían:  $(v_1 = 1)$  si  $(x=2)$  y cero en otro caso;  $(v_2 = 1)$  si  $(x=3)$  y cero en otro caso;  $(\dots)$ ;  $(v_{I-1} = 1)$  si  $(x=I)$  y cero en otro caso.
- Obviamente cuando todas las variables  $(v)$  son nulas estamos en la primera categoría.

### $(\dots)$

- Si solamente tenemos la variable categórica como predictora entonces la media sería función lineal de  $(x)$  del siguiente modo:  $(\mu_i = \beta_0 + \beta_1 v_1 + \dots + \beta_{I-1} v_{I-1})$ .

### $(\dots)$

- Por ejemplo, supongamos una numérica  $(x)$  y una categórica  $(v)$  con  $(I)$  categorías. Construimos las variables tontas siendo  $(I)$  la de referencia. Podemos considerar modelos como  $(\mu_i = \beta_0 + \beta_1 x_i)$ , que lo expresamos como  $(y \sim x)$
- Un modelo que contiene solamente a  $(v)$  sería el dado previamente  $(\mu_i = \beta_0 + \beta_1 v_{i1} + \dots + \beta_{I-1} v_{i,I-1})$ . y lo expresamos como  $(y \sim v)$

### $(\dots)$

- Un modelo que contempla ambas variables puede ser  $(\mu_i = \beta_0 + \beta_1 x_i + \beta_2 v_{i1} + \dots + \beta_{I-1} v_{i,I-1})$ . Este modelo lo podemos abreviar como  $(y \sim x + v)$
- Un modelo más completo que contempla la posible interacción sería  $(\begin{multline} \mu_i = \beta_0 + \beta_1 x_i + \beta_2 v_{i1} + \dots + \beta_{I-1} v_{i,I-1} + \beta_{I+1} x_i v_{i1} + \dots + \beta_{2I-1} x_i v_{i,I-1} \end{multline})$ . Esto lo indicaremos como  $(y \sim x * v)$   $(y \sim x + v + x:v)$

## Matriz modelo y espacio modelo

### Matriz modelo

- Tenemos el vector de medias  $(\mathbf{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix})$

- $$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$
 -  $(\mathbf{X})$  es la **matriz modelo**.

### Matriz modelo $(\dots)$

- Asumimos  $\mu_i = \sum_{j=1}^p \beta_j x_{ij}$ , - En modo matricial,  $\mathbf{\mu} = \mathbf{X} \mathbf{\beta}$ , siendo el vector de coeficientes  $\mathbf{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ .