Modelos Lineales

Regresión Lineal Mútiple

Fernando Massa; Bruno Bellagamba

martes 19 de marzo 2024

FCEA Facultad de Ciencias Económicas y de Administración
IESTA Instituto de Estadística
Universidad de la República Uruguay

Temario

- Comenzaremos con el modelo de regresión Lineal Múltiple.
- Veremos la interpretación geométrica del método de MCO.
- Veremos el teorema de Gauss-Markov.
- Generalizaremos algunos resultados del modelo de RLS.
- Definiremos el vector de predichos y de residuos.

El análisis de Regresión Lineal Múltiple (RLM) se adecúa a un marco de trabajo *ceteris paribus*. Esto significa que permite controlar de manera **simultánea** el efecto de un conjunto de variables explicativas sobre la variable de respueta.

Es un modelo más completo que el de RLS, ya que lo incluye como caso particular y que permite mayor flexibilidad para descrbir la forma funcional de la relación de las variables explicativas y Y.

Veremos que el modelo de RLM no solo permite mejorar las predicciones del modelo RLS sino que además permite identificar qué variables tienen mayor importancia al explicar Y.

El modelo de RLM (en su forma escalar) se caracteriza por la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

- Y_i continua siendo la variable de respuesta.
- La principal diferencia con el modelo RLS es la presencia de k variables explicativas $X_1, X_2, \ldots, X_k$.
- Debido al punto anterior, el vector de coeficientes β contiene $k+1$ componentes, 1 para la ordenada en el origen y 1 para cada variable explicativa.
- La aleatoriedad del modelo viene dada por ε.

Una cierta empresa utiliza 3 medios de publicidad para publicitar sus productos (youtube, facebook y periódicos). Para un conjunto de 200 prdocutos cuenta con las siguientes variables:

- Ventas de cada artículo (miles de pesos).
- Presupuesto de publicidad en youtube (miles de pesos).
- Presupuesto de publicidad en facebook (miles de pesos).
- Presupuesto de publicidad en periódicos (miles de pesos).

La empresa desea saber cuál de estos medios tiene un mayor impacto en las ventas. El modelo de RLM que permite contestar esta pregunta es:

Una cierta empresa utiliza 3 medios de publicidad para publicitar sus productos (youtube, facebook y periódicos). Para un conjunto de 200 prdocutos cuenta con las siguientes variables:

- Ventas de cada artículo (miles de pesos).
- Presupuesto de publicidad en youtube (miles de pesos).
- Presupuesto de publicidad en facebook (miles de pesos).
- Presupuesto de publicidad en periódicos (miles de pesos).

La empresa desea saber cuál de estos medios tiene un mayor impacto en las ventas. El modelo de RLM que permite contestar esta pregunta es:

$$Ventas_i = \beta_0 + \beta_1 Youtube_i + \beta_2 Facebook_i + \beta_3 Periódicos_i + \varepsilon_i$$

En el modelo de RLS la interpretación de los coeficientes del modelo eran:

- β_0 es el valor promedio de Y cuando X vale 0.
- β_1 es el incremento promedio de Y por cada unidad de aumento en X.

En el modelo de RLS la interpretación de los coeficientes del modelo eran:

- β_0 es el valor promedio de Y cuando X vale 0.
- β_1 es el incremento promedio de Y por cada unidad de aumento en X.

En el modelo de RLM esto cambia un poco. Para verlo, planteemos la esperanza condicional de Y según el modelo RLM en forma escalar para la observación i.

$$\mathbb{E}(Y_i|X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

En el modelo de RLS la interpretación de los coeficientes del modelo eran:

- β_0 es el valor promedio de Y cuando X vale 0.
- β_1 es el incremento promedio de Y por cada unidad de aumento en X.

En el modelo de RLM esto cambia un poco. Para verlo, planteemos la esperanza condicional de Y según el modelo RLM en forma escalar para la observación i.

$$\mathbb{E}(Y_i|X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

Pero cuando X_1 adopta el valor $x_{i1} + 1$ este valor esperado se transforma en:

$$\mathbb{E}(Y_i|X_1 = x_{i1} + 1, X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_0 + \beta_1(x_{i1} + 1) + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

Si restamos las dos expresiones de la diapositiva anterior:

$$\mathbb{E}(Y_i|X_1 = x_{i1} + 1, X_2 = x_{i2}, \ldots, X_k = x_{ik}) - \mathbb{E}(Y_i|X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_1$$

Interpretación

Esto quiere decir que la interpretación de β_1 es *el incremento promedio en Y por cada unidad que aumenta X_1* **DEJANDO LAS DEMÁS VARIABLES CONSTANTES**.

Si restamos las dos expresiones de la diapositiva anterior:

$$\mathbb{E}(Y_i|X_1 = x_{i1} + 1, X_2 = x_{i2}, \ldots, X_k = x_{ik}) - \mathbb{E}(Y_i|X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_1$$

Interpretación

Esto quiere decir que la interpretación de β_1 es *el incremento promedio en* Y *por cada unidad que aumenta* X_1 **DEJANDO LAS DEMÁS VARIABLES CONSTANTES**.

Con un procedimiento análogo podríamos interpretar los demás coeficientes asociados a las variables eplicativas.

Si restamos las dos expresiones de la diapositiva anterior:

$$\mathbb{E}(Y_i|X_1 = x_{i1} + 1, X_2 = x_{i2}, \ldots, X_k = x_{ik}) - \mathbb{E}(Y_i|X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_1$$

INTERPRETACIÓN

Esto quiere decir que la interpretación de β_1 es *el incremento promedio en* Y *por cada unidad que aumenta* X_1 **DEJANDO LAS DEMÁS VARIABLES CONSTANTES**.

Con un procedimiento análogo podríamos interpretar los demás coeficientes asociados a las variables eplicativas.

En cuanto al término independiente (β_0), continúa siendo el valor esperado de Y cuando **TODAS** las variables explicativas valen 0.

En forma matricial, el modelo continúa siendo:

$$Y = X\beta + \varepsilon$$

En forma matricial, el modelo continúa siendo:

$$Y = X\beta + \varepsilon$$

- $\mathbb{E}(\varepsilon) = \mathbf{0} \Rightarrow \mathbb{E}(Y) = X\beta$.
- $\mathbb{C}ov(\varepsilon) = \sigma^2 I_n \Rightarrow \mathbb{C}ov(Y) = \sigma^2 I_n$.
- Las variables $x_1, x_2, \ldots, x_k$ son fijas. Es decir, la matriz de diseño X es **no estocástica**.
- La matriz X es de rango completo, $rg(X) = k + 1$

  Esto garantiza la existencia de la inversa de la matriz $X^{'}X$.

MCO

El estimador *MCO* continúa siendo el visto en clases anteriores:

$$\hat{\beta}_{MCO} = \left(X^{'}X\right)^{-1} X^{'}Y$$

El estimador *MCO* continúa siendo el visto en clases anteriores:

$$\hat{\beta}_{MCO} = \left(X^{'}X\right)^{-1} X^{'}Y$$

Con la única particularidad de que la matriz de diseño X cuenta con $k+1$ columnas (k variables explicativas y 1 columna extra para el vector de unos).

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

El estimador *MCO* continúa siendo el visto en clases anteriores:

$$\hat{\beta}_{MCO} = \left(X^{'}X\right)^{-1} X^{'}Y$$

Con la única particularidad de que la matriz de diseño X cuenta con $k+1$ columnas (k variables explicativas y 1 columna extra para el vector de unos).

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Consecuentemente, el vector β tiene $k+1$ parámetros. El primero sigue siendo β_0 y los restantes $(\beta_1, \beta_2, \dots, \beta_k)$ están asociados a cada una de las variables explicativas.

Retomemos el ejemplo de las ventas en R. Estimemos los valores de los coeficientes e interpretémoslos.

Antes de comenzar con este apartado puede resultar útil recordar algunos resultados de álgebra lineal.

Repaso de álgegra lineal

1. Dos vectores u y v pertenecientes a $\mathbb{R}^n$ son ortogonales si $u^{'}v = 0$.
2. Siendo W un subespacio de $\mathbb{R}^n$, el subespacio de los vectores v, tales que para cualquier $u \in W$ se cumple que $u^{'}v = 0$, es llamado **complemento ortogonal** de W y es denotado por $W^{\perp}$.
3. Los subespaciones complementarios W y $W^{\perp}$ cumplen que $\dim(W) + \dim(W^{\perp}) = n$.
4. Cualquier vector $Y \in \mathbb{R}^n$ tiene una **descomposición ortogonal** única de la forma $Y = Y_1 + Y_2$ de forma tal que $Y_1 \in W$ y $Y_2 \in W^{\perp}$

El objetivo del estimador *MCO* es aproximarse al vector *n-dimensional* Y a partir de una combinación lineal de las columnas de la matriz X. Es decir, de un elemento perteneciente al subespacio que generan sus columnas.

Representación geometrica del estimador *MCO*

El objetivo del estimador *MCO* es aproximarse al vector *n-dimensional* Y a partir de una combinación lineal de las columnas de la matriz X. Es decir, de un elemento perteneciente al subespacio que generan sus columnas.

Las combinaciones lineales de las columnas de una matriz dan lugar a subespacios. En el caso de la matriz X hablamos de $\mathbb{C}(X)$, cuya dimensión es $k+1$.

El objetivo del estimador *MCO* es aproximarse al vector *n-dimensional* Y a partir de una combinación lineal de las columnas de la matriz X. Es decir, de un elemento perteneciente al subespacio que generan sus columnas.

Las combinaciones lineales de las columnas de una matriz dan lugar a subespacios. En el caso de la matriz X hablamos de $\mathbb{C}(X)$, cuya dimensión es $k+1$.

En el modelo de RLS, cuando X contaba con una columna de unos y otra columna con la variable x_1, $\mathbb{C}(X)$ eran las **rectas** determinadas por los coeficientes de esa combinación lineal.

Representación geometrica del estimador *MCO*

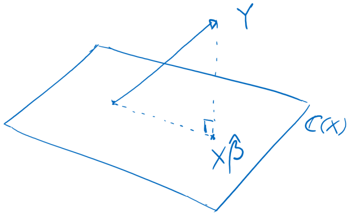
El objetivo del estimador *MCO* es aproximarse al vector *n-dimensional* Y a partir de una combinación lineal de las columnas de la matriz X. Es decir, de un elemento perteneciente al subespacio que generan sus columnas.

Las combinaciones lineales de las columnas de una matriz dan lugar a subespacios. En el caso de la matriz X hablamos de $\mathbb{C}(X)$, cuya dimensión es $k+1$.

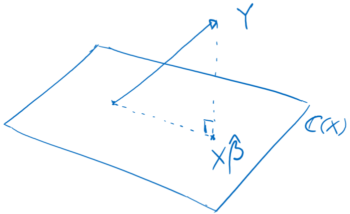
En el modelo de RLS, cuando X contaba con una columna de unos y otra columna con la variable x_1, $\mathbb{C}(X)$ eran las **rectas** determinadas por los coeficientes de esa combinación lineal.

Si adicionalmente X contiene otra variable explicativa x_2 (que no sea múltiplo de x_1), entonces $\mathbb{C}(X)$ son los **planos** determinados por los 3 coeficientes de las combinaciones lineales de las columnas de X.

En una dimensión mayor, esto se podría visualizar de la siguiente manera:

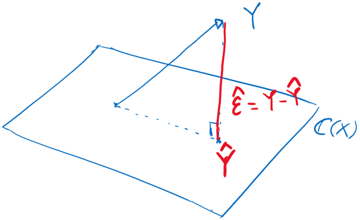


En una dimensión mayor, esto se podría visualizar de la siguiente manera:

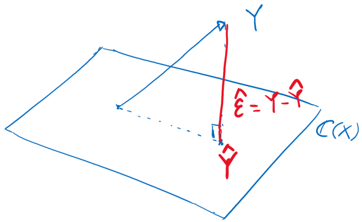


- $Y \in \mathbb{R}^n$.
- $\hat{\beta}_{MCO}$ minimiza la distancia de Y a $\mathbb{C}(X)$.
  $\hat{Y} = X\hat{\beta}$ es la **proyección ortogonal** de Y en $\mathbb{C}(X)$.
- Como $X\hat{\beta} \in \mathbb{C}(X)$ entonces tiene dimensión $k+1$.

Luego, el vector que va de Y a $\mathbb{C}(X)$ es el vector de **residuos** $\hat{\varepsilon}$.



Luego, el vector que va de Y a $\mathbb{C}(X)$ es el vector de **residuos** $\hat{\varepsilon}$.



- $Y \in \mathbb{R}^n$
- Como $\hat{Y} \in \mathbb{C}(X)$ entonces $\hat{Y} \in \mathbb{R}^{k+1}$.
- Como $\hat{\varepsilon}$ es ortogonal al subespacio $\mathbb{C}(X)$, entonces pertenece a su complemento ortogonal $\mathbb{C}^{\perp}(X)$ el cual tiene dimensión $n - k - 1$.

A partir de esta observación ($\hat{\varepsilon} \perp \mathbb{C}(X)$) es posible obtener el estimador MCO de otra manera.

A partir de esta observación ($\hat{\varepsilon} \perp \mathbb{C}(X)$) es posible obtener el estimador *MCO* de otra manera.

Al definir $\lambda_j = (0, \ldots, 0, 1, 0, \ldots, 0)$, la combinación lineal $X\lambda_j \in \mathbb{C}(X)$ no es más que la j-ésima columna de X. Por lo tanto se cumple que:

$$\left\langle \hat{\varepsilon}, (X\lambda_j) \right\rangle = 0.$$

A partir de esta observación ($\hat{\varepsilon} \perp \mathbb{C}(X)$) es posible obtener el estimador *MCO* de otra manera.

Al definir $\lambda_j = (0, \ldots, 0, 1, 0, \ldots, 0)$, la combinación lineal $X\lambda_j \in \mathbb{C}(X)$ no es más que la j-ésima columna de X. Por lo tanto se cumple que:

$$\langle \hat{\varepsilon}, (X\lambda_j) \rangle = 0.$$

De esta manera, el vector $\hat{\varepsilon}$ es ortogonal a todas las columnas de X.

$$\begin{aligned} X'\hat{\varepsilon} &= 0 \\ X'\left(Y - X\hat{\beta}\right) &= 0 \\ X'Y - X'X\hat{\beta} &= 0 \\ X'X\hat{\beta} &= X'Y \\ \hat{\beta} &= \left(X'X\right)^{-1} X'Y \end{aligned}$$

Este teorema indica que:

En el modelo de regresión lineal de la forma $Y = X\beta + \varepsilon$ donde:

1. $\mathbb{E}(\varepsilon) = 0$.
2. $\mathbb{C}ov(\varepsilon) = \sigma^2 I_n$.
3. X es de rango completo.

Este teorema indica que:

En el modelo de regresión lineal de la forma $Y = X\beta + \varepsilon$ donde:

1. $\mathbb{E}(\varepsilon) = 0$.
2. $\mathbb{C}ov(\varepsilon) = \sigma^2 I_n$.
3. X es de rango completo.

El estimador MCO es el **Mejor Estimador Lineal Insesgado (MELI)** de β. Esto significa que el estimador de MCO es el de mínima varianza entre todos los estimadores lineales insesgados.

Este teorema indica que:

En el modelo de regresión lineal de la forma $Y = X\beta + \varepsilon$ donde:

1. $\mathbb{E}(\varepsilon) = 0$.
2. $\mathbb{C}ov(\varepsilon) = \sigma^2 I_n$.
3. X es de rango completo.

El estimador MCO es el **Mejor Estimador Lineal Insesgado (MELI)** de β. Esto significa que el estimador de MCO es el de mínima varianza entre todos los estimadores lineales insesgados.

Veamos la demostración:

En clases anteriores, vimos que $\hat{\beta}$ es insesgado, veamos ahora qué significa que sea lineal.

$$\begin{aligned} \hat{\beta}_{MCO} &= \left(X^{'}X\right)^{-1}X^{'}Y \\ \hat{\beta}_{MCO} &= BY \end{aligned}$$

Siendo B una matriz de $k+1\times n$. De esta manera notamos que cada componente de $\hat{\beta}_{MCO}$ es una combinación lineal de los valores del vector Y, donde los coeficientes de dicha combinación vienen dados por la matriz $\left(X^{'}X\right)^{-1}X^{'}$.

Ahora demostremos la parte *interesante* del teorema, la de la mínima varianza.

Ahora demostremos la parte *interesante* del teorema, la de la mínima varianza.

La idea será construir un estimador de la forma AY (que es lineal) y ver que si cumple la condición de insesgadez, los elementos de la diagonal de su matriz de covarianzas se minimizan cuando $A = \left(X^{'}X\right)^{-1}X^{'}$.

Paso 1) La condición de insesgadez.

Si tenemos en cuenta la hipótesis de que $\mathbb{E}(Y) = X\beta$, para que AY sea insesgado al estimar a β debe cumplirse que:

$$\begin{aligned} \mathbb{E}(AY) &= \beta \\ A\mathbb{E}(Y) &= \beta \\ AX\beta &= \beta \end{aligned}$$

De esta forma, la elección de la matriz A debe ser tal que se cumpla la ecuación

$AX = I_{k+1}$

Paso 2) La matriz de covarianzas de AY.

$$\mathbb{C}ov(AY) = A\mathbb{C}ov(Y)A^{'} = A\sigma^2 I_n A^{'} = \sigma^2 AA^{'}$$

Las varianzas de $\hat{\beta}$ están en la diagonal de esta matriz, asi que debemos elegir A (sujeta a que $AX = I_{k+1}$) de forma de minimizar los elementos de esa diagonal.

Paso 2) La matriz de covarianzas de AY.

$$\mathbb{C}ov(AY) = A\mathbb{C}ov(Y)A^{'} = A\sigma^2 I_n A^{'} = \sigma^2 AA^{'}$$

Las varianzas de $\hat{\beta}$ están en la diagonal de esta matriz, asi que debemos elegir A (sujeta a que $AX = I_{k+1}$) de forma de minimizar los elementos de esa diagonal.

Paso 3) El *truquito*.

Para relacionar A con $\hat{\beta}_{MCO}$ sumamos y restamos $\left(X^{'}X\right)^{-1}X^{'}$ en el producto $AA^{'}$ para obtener:

$$\left[A - \left(X^{'}X\right)^{-1}X^{'} + \left(X^{'}X\right)^{-1}X^{'}\right]\left[A - \left(X^{'}X\right)^{-1}X^{'} + \left(X^{'}X\right)^{-1}X^{'}\right]^{'}$$

Y realizamos la distributiva empleando los términos $A - \left(X^{'}X\right)^{-1}X^{'}$ y $\left(X^{'}X\right)^{-1}X^{'}$.

Al llevar a cabo la distributiva obtenemos 4 términos:

1. $\left[A - \left(X'X\right)^{-1} X'\right] \left[A - \left(X'X\right)^{-1} X'\right]'$
2. $\left[A - \left(X'X\right)^{-1} X'\right] X \left(X'X\right)^{-1}$
3. $\left(X'X\right)^{-1} X' \left[A' - X \left(X'X\right)^{-1}\right]$
4. $\left(X'X\right)^{-1}$

Al llevar a cabo la distributiva obtenemos 4 términos:

1. $\left[A-\left(X'X\right)^{-1}X'\right]\left[A-\left(X'X\right)^{-1}X'\right]'$
2. $\left[A-\left(X'X\right)^{-1}X'\right]X\left(X'X\right)^{-1}$
3. $\left(X'X\right)^{-1}X'\left[A'-X\left(X'X\right)^{-1}\right]$
4. $\left(X'X\right)^{-1}$

En el segundo se puede obtener que:

$$\begin{aligned}\left[A-\left(X'X\right)^{-1}X'\right]X\left(X'X\right)^{-1} &= AX\left(X'X\right)^{-1}-\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}\\ &= I_{k+1}\left(X'X\right)^{-1}-\left(X'X\right)^{-1}\\ &= 0\end{aligned}$$

Y lo mismo ocurre en el tercer término.

Asi que luego de hacer la distributiva obtenemos:

$$AA^{'} = \left[A - \left(X^{'}X\right)^{-1}X^{'}\right]\left[A - \left(X^{'}X\right)^{-1}X^{'}\right]^{'} + \left(X^{'}X\right)^{-1}$$

El interés recae sobre el primero de estos dos sumandos, ya que es el que depende de A. Se trata de una matriz semidefinida-positiva. Por ende los elementos de su diagonal son todos mayores o iguales a cero. La forma de anular los elementos de su diagonal (y todos los demás) es forzando $A = \left(X^{'}X\right)^{-1}X^{'}$.

Asi que luego de hacer la distributiva obtenemos:

$$AA^{'} = \left[A - \left(X^{'}X\right)^{-1}X^{'}\right]\left[A - \left(X^{'}X\right)^{-1}X^{'}\right]^{'} + \left(X^{'}X\right)^{-1}$$

El interés recae sobre el primero de estos dos sumandos, ya que es el que depende de A. Se trata de una matriz semidefinida-positiva. Por ende los elementos de su diagonal son todos mayores o iguales a cero. La forma de anular los elementos de su diagonal (y todos los demás) es forzando $A = \left(X^{'}X\right)^{-1}X^{'}$.

Si adicionalmente comprobamos que con esta elección de A se cumple que $AX = I_{k+1}$ ya que $\left(X^{'}X\right)^{-1}X^{'}X = I_{k+1}$ entonces el estimador que minimiza los elementos de la diagonal de la matriz de covarianzas del estimador es:

$$AY = \left(X^{'}X\right)^{-1}X^{'}Y = \hat{\beta}_{MCO}$$

Según lo observado en la interpretación geométrica, es posible descomponer al vector Y de la siguiente manera:

$$\begin{aligned} Y &= Y + \hat{Y} - \hat{Y} \\ &= {\color{blue}\hat{Y}} + {\color{red}Y - \hat{Y}} \\ &= {\color{blue}\hat{Y}} + {\color{red}\hat{\varepsilon}} \end{aligned}$$

Según lo observado en la interpretación geométrica, es posible descomponer al vector Y de la siguiente manera:

$$\begin{aligned} Y &= Y + \hat{Y} - \hat{Y} \\ &= \hat{Y} + Y - \hat{Y} \\ &= \hat{Y} + \hat{\varepsilon} \end{aligned}$$

Donde:

- $\hat{Y}$ es el vector de valores **predichos** (proyección de Y en $\mathbb{C}(x)$).
- $\hat{\varepsilon}$ es el vector de **residuos** (proyección de Y en $\mathbb{C}^{\perp}(x)$).

Predichos y Residuos

Según lo observado en la interpretación geométrica, es posible descomponer al vector Y de la siguiente manera:

$$\begin{aligned} Y &= Y + \hat{Y} - \hat{Y} \\ &= \hat{Y} + Y - \hat{Y} \\ &= \hat{Y} + \hat{\varepsilon} \end{aligned}$$

Donde:

- $\hat{Y}$ es el vector de valores **predichos** (proyección de Y en $\mathbb{C}(x)$).
- $\hat{\varepsilon}$ es el vector de **residuos** (proyección de Y en $\mathbb{C}^{\perp}(x)$).

¡NO DEBEMOS CONFUNDIR ERRORES CON RESIDUOS!

La próxima clase:

- Retomaremos la definición de predichos y residuos
- y la vincularemos con la matriz de proyección (H).
- Veremos las propiedades de esta última.
- Y hablaremos de la *parcialidad* de los coeficientes del modelo de RLM.

Carmona, Francesc (2003). *Modelos Lineales (notas de curso).* Departament d'Estadística.

Farraway, Julian (2014). *Linear Models with R, second edition.* Chapman Hall/CRC.

Rencher, Alvin y Bruce Schaalje (2008). *Linear Models in Statistics, second edition.* John Wiley Sons, Inc.

¿Preguntas?

Muchas Gracias