

MODELOS LINEALES

REGRESIÓN LINEAL SIMPLE

Fernando Massa; Bruno Bellagamba

Jueves 7 de marzo 2024



FACULTAD DE
CIENCIAS ECONÓMICAS
Y DE ADMINISTRACIÓN

IESTA INSTITUTO
DE ESTADÍSTICA



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



1 ¿QUÉ ES UN MODELO LINEAL?

2 COMPONENTES

3 SUPUESTOS

4 MOMENTOS

5 ESTIMACIÓN

6 VÍNCULO CON ρ

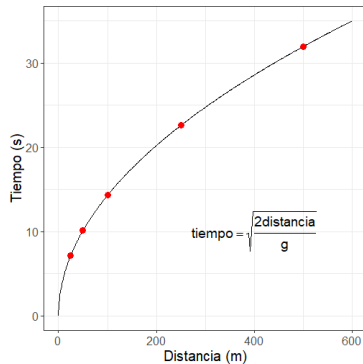
7 PRÓXIMA CLASE



- Habremos definido nuestro primer modelo lineal.
- Sabremos reconocer sus componentes.
- En qué situaciones puede ser utilizado.
- Habremos indagado en su relación con el coeficiente de correlación lineal de Pearson.
- Habremos definido el método de mínimos cuadrados ordinarios (MCO).

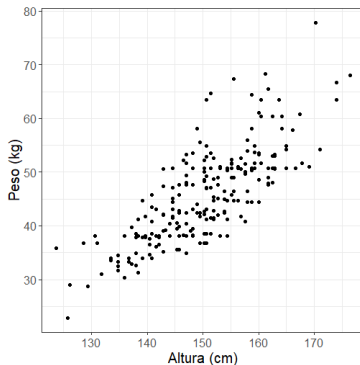
MODELADO DETERMINÍSTICO

Permiten establecer relaciones con **completa certeza** sin incorporar elementos de caracter aleatorio en la descripción del fenómeno estudiado



MODELADO ESTOCÁSTICO

Un modelo es de caracter *estocástico* si incorpora elementos aleatorios para describir un fenómeno. Estos modelos pueden ser analizados empleando la estadística y las predicciones que realizan incorporan un grado de **incertidumbre**.



MODELO ESTADÍSTICO

Se trata de una representación matemática que recoge ciertas características de un fenómeno de interés permitiendo analizar sus propiedades probabilísticas y realizar predicciones.

Una frase muy conocida de George Box dice:

"All models are wrong, but some are usefull"

Esta frase hace referencia a que todos los modelos son incapaces de representar toda la complejidad del fenómeno que intentan representar, pero aún así pueden ser útiles.

MODELO LINEAL

Modelo que relaciona una variable de interés respecto de una o más variables que se cree que están relacionadas con ella. Cuando dicha relación es lineal en los parámetros, se está frente a un modelo lineal.

El modelo de **regresión lineal simple** (RLS), parte de la siguiente especificación:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Siendo Y la variable a predecir (o explicada, o dependiente) y X la variable predictora (o explicativa, o independiente).

¿PARA QUÉ SIRVE?

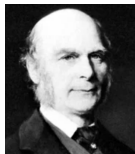
El análisis de los modelos de regresión suele responder a (al menos) uno de los siguientes objetivos:

OBJETIVOS


- Predicción de observaciones (futuras) a partir de ciertos valores en la/s variable/s explicativa/s.
- Evaluar el *efecto* de una o más variables explicativas sobre Y .

El análisis de regresión también puede ser empleado con fines meramente descriptivos, pero en la mayoría de las aplicaciones es empleado con propósitos inferenciales.

¿POR QUÉ SE LLAMA ASÍ?



Término acuñado por Sir Francis Galton en su trabajo de 1875 *Regression towards mediocrity in hereditary stature*. Observó que características extremas en los padres no solían ser heredadas en los hijos, sino que más bien, *regresaban* hacia un punto *mediocre*.

A continuación se presentan algunos datos (disponibles en ) sobre los cuales trabajó Galton.

| family | father | mother | midparentHeight | childHeight |
|--------|--------|--------|-----------------|-------------|
| 97 | 69 | 68,5 | 71,49 | 65 |
| 44 | 71,5 | 65 | 70,85 | 68,5 |
| 62 | 70 | 69 | 72,26 | 68 |
| 53 | 71 | 63 | 69,52 | 63 |

¿POR QUÉ SE LLAMA ASÍ?

Al realizar un diagrama de dispersión entre la altura promedio de los padres y la altura de los hijos, se puede observar una correlación positiva que indica que padres altos, suelen tener hijos altos y que padres bajos suelen tener hijos más bajos.

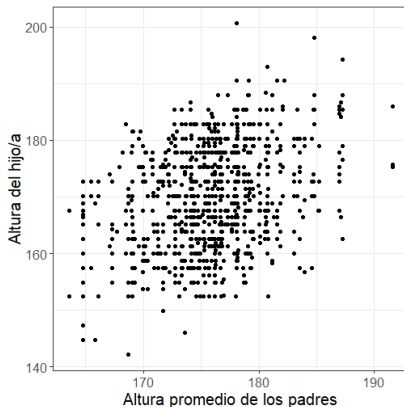


FIGURA: Alturas de padres e hijos

¿POR QUÉ SE LLAMA ASÍ?

La observación de Galton (en términos coloquiales) consistía en qué los hijos de padres altos también son altos (pero no tanto) y que los hijos de padres bajos, también son bajos (pero no tanto).

¿Cómo creen que puso a prueba esta hipótesis?

¿POR QUÉ SE LLAMA ASÍ?

La observación de Galton (en términos coloquiales) consistía en qué los hijos de padres altos también son altos (pero no tanto) y que los hijos de padres bajos, también son bajos (pero no tanto).

¿Cómo creen que puso a prueba esta hipótesis?

En primer lugar postuló el siguiente modelo lineal:

$$Altura \text{ hijos} = \beta_0 + \beta_1 altura \text{ padres} + \varepsilon$$

¿POR QUÉ SE LLAMA ASÍ?

La observación de Galton (en términos coloquiales) consistía en qué los hijos de padres altos también son altos (pero no tanto) y que los hijos de padres bajos, también son bajos (pero no tanto).

¿Cómo creen que puso a prueba esta hipótesis?

En primer lugar postuló el siguiente modelo lineal:

$$Altura \text{ hijos} = \beta_0 + \beta_1 altura \text{ padres} + \varepsilon$$

Si la hipótesis de Galton era correcta, el valor de β_1 debería ser positivo y menor a 1.

El modelo de RLS parte del supuesto de que la relación entre X e Y puede ser modelada “en promedio” por una recta, donde:

- La ordenada en el origen vale β_0 .
- El valor del coeficiente angular vale β_1 .

El modelo de RLS parte del supuesto de que la relación entre X e Y puede ser modelada “en promedio” por una recta, donde:

- La ordenada en el origen vale β_0 .
- El valor del coeficiente angular vale β_1 .

No obstante, parte fundamental del modelo es reconocer que el mismo NO provee de una explicación completa de la variabilidad de Y . Por este motivo, se incluye un término de error (ε), al cuál suponemos aleatorio, siendo sus principales características:

- $\mathbb{E}(\varepsilon_i) = 0 \quad \forall i = 1, \dots, n.$
- $\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i = 1, \dots, n.$
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j.$

Recapitulando, el modelo de RLS, expresado de manera escalar (porque luego lo veremos en forma matricial) es:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \forall i = 1, \dots, n$$

En resumen, los componentes del modelo de RLS son de 3 tipos:

- Variables aleatorias (ε y en teoría la Y).
- Datos (X_i y en la práctica Y_i).
- Constantes desconocidas (β_0 y β_1 que son los parámetros a estimar).

Al estar inmersos en un marco de trabajo estadístico, es importante que seamos conscientes de que el proceso de modelado (ya sea lineal o no) se basa en la siguiente igualdad.

$$\textit{observación} = \textit{modelo} + \textit{error aleatorio}$$

En esta relación, el término de error dota de aleatoriedad al modelo matemático y además contiene la variabilidad de las observaciones que no ha sido considerada en el modelo. En el caso de los modelos de regresión, el término de error contiene todas las discrepancias entre lo observado y lo predicho por el modelo, por lo general se debe a:

- Variables no incluidas en el modelo.
- Errores en la especificación de la relación entre las X y la Y .
- Otras fuentes de variación.

¿EN QUÉ CONDICIONES PUEDE UTILIZARSE?

Es un modelo sumamente flexible como herramienta inferencial en muchos problemas. Sin embargo es importante conocer los supuestos en los que se basa.

SUPUESTOS

- La variable de respuesta debe ser de caracter **cuantitativo** (y en lo posible continua).
- La/s variables explicativas pueden ser de cualquier naturaleza. Pero de ser cualitativas deben codificarse de alguna forma numérica.
- La relación entre X e Y debe ser lineal, o al menos poder ser razonablemente aproximada por una recta a través de alguna transformación de X y/o Y .
- Las observaciones que serán analizadas por este modelo deben cumplir con el supuesto de independencia.

Para establecer que el modelo es **lineal en los parámetros** es necesario poder escribir la parte **sistémica** del modelo de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \langle \beta, \mathbf{X}_i \rangle + \varepsilon_i$$

Donde $\mathbf{X}_i = [1, X_i]$ y $\beta = [\beta_0, \beta_1]$.

LINEALIZABLE

De forma más general, diremos que un modelo es lineal si puede ser planteado de la siguiente manera:

$$Y_i = \langle \beta, \mathbf{g}(\mathbf{X}_i) \rangle + \varepsilon_i$$

Siendo $\mathbf{g}(\mathbf{X}_i)$ un vector de la misma dimensión de β , pudiendo cada una de ellas ser distintas transformaciones de la/s variable/s explicativa/s.

¿LINEAL, LINEALIZABLE O NO LINEAL?

Según lo visto en la diapositiva anterior, el siguiente modelo, NO es lineal:

$$Y_i = \beta_0 e^{\beta_1 X_i} \varepsilon_i$$

¿LINEAL, LINEALIZABLE O NO LINEAL?

Según lo visto en la diapositiva anterior, el siguiente modelo, NO es lineal:

$$Y_i = \beta_0 e^{\beta_1 X_i} \varepsilon_i$$

Sin embargo, es un modelo **linealizable**, ya que al aplicar logaritmo en ambos lados de la igualdad:

$$\log(Y_i) = \log(\beta_0) + \beta_1 X_i + \log(\varepsilon_i)$$

Al renombrar $Z_i = \log(Y_i)$, $\beta_0^* = \log(\beta_0)$ y $\eta_i = \log(\varepsilon_i)$ se obtiene

$$Z_i = \beta_0^* + \beta_1 X_i + \eta_i$$



Indique si los siguientes modelos son lineales, linealizables o no lineales.

- $Y_i = \delta_0 + \delta_1 X_i + \delta_2 X_i^2 + \delta_3 X_i^3 + \varepsilon_i$
- $Y_i = \lambda_0 \lambda_1^{X_{i1}} \lambda_2^{X_{i2}} \varepsilon_i$.
- $Y_i = \alpha_0 \sin(\alpha_1 X_{i1}) + \varepsilon_i$.
- $Y_i = \gamma_0 + \gamma_1 X_i^{\gamma_2} + \varepsilon_i$.

Si *regresamos* al modelo de RLS en su forma escalar:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

y recordamos las características del error aleatorio, es posible deducir que:

- Como Y es una función de una variable aleatoria (VA), entonces la propia Y es una VA. En la práctica, contamos con **realizaciones** de dicha VA.
- Como $\mathbb{E}(\varepsilon) = 0$, entonces $\mathbb{E}(Y) = \beta_0 + \beta_1 X$.
- Como $\mathbb{V}ar(\varepsilon) = \sigma^2$, entonces $\mathbb{V}ar(Y) = \sigma^2$.

Nótese que en los últimos 2 puntos se aplicó la linealidad de la esperanza y que a la derecha del signo de igual, el único término aleatorio es ε .

A partir de esta caracterización de los momentos de Y podemos plantear que el centro de atención del modelo de RLS está en:

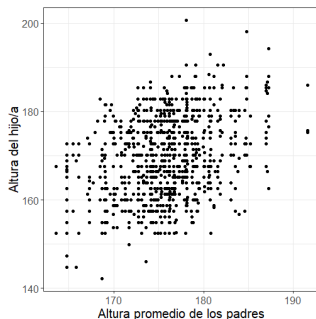
$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$$

- A partir de esta observación notamos que en este curso ampliamos los objetivos de Inferencia I (la media que queremos estimar depende de otras variables).
- Podemos interpretar con mayor precisión los parámetros del modelo.

INTERPRETACIÓN DE LOS PARÁMETROS (RLS)

- 1 β_0 es el valor promedio de Y cuando X vale 0.
- 2 β_1 es el incremento promedio de Y por cada unidad que aumente X .

Nuestra primera tarea al trabajar con el modelo de RLS es la de estimar los valores de los parámetros a partir de una muestra $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$. Si recordamos el ejemplo de las alturas de los padres e hijos...



... nuestra tarea consiste en determinar cual es la recta que *mejor* se ajusta a los datos.

La idea de este método de estimación consiste en que las estimaciones de los parámetros ($\hat{\beta}_0$ y $\hat{\beta}_1$) minimicen la suma de los cuadrados de los residuos, es decir:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^n \varepsilon_i^2$$

Si al último término lo llamamos $\phi(\beta_0, \beta_1)$, la solución surge de derivar e igualar a cero:

$$\begin{aligned} \frac{\partial}{\partial \beta_0} \phi(\beta_0, \beta_1) &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \\ \frac{\partial}{\partial \beta_1} \phi(\beta_0, \beta_1) &= -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) \end{aligned}$$

A estas ecuaciones se las conoce como **Ecuaciones Normales**.

A partir de la primera ecuación es fácil obtener que:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Dos aspectos a tener en cuenta:

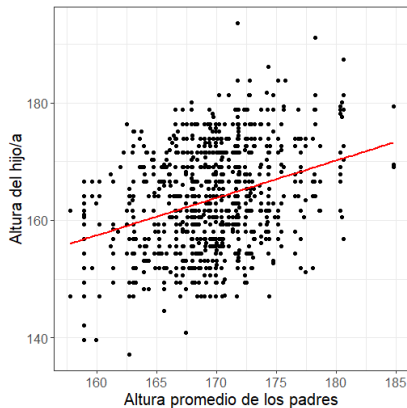
- Nótese como los “gorros” surgen luego de igualar a cero.
- El valor estimado de β_0 es la media de Y con una corrección que depende de la media de X , salvo que ...

La segunda ecuación lleva un poco más de trabajo y requiere substituir la estimación de β_0 en la primera ecuación. Al final se obtiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

A partir de las estimaciones, podemos calcular *predicciones* de la forma

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$



El método de mínimos cuadrados no nos provee de una estimación de σ^2 . No obstante, apelando a resultado del curso de Inferencia I, podemos construir un estimador teniendo en cuenta que:

$$\text{Var}(Y) = \mathbb{E}[Y - \mathbb{E}(Y)]^2 = \sigma^2$$

En una diapositiva anterior vimos que:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X$$

Por lo que un estimador natural podría ser:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n - 2}$$

RELACIÓN ENTRE $\hat{\beta}_1$ Y r

Si retomamos al estimador del coeficiente angular:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_1 &= \frac{s_{xy}}{s_x^2}\end{aligned}$$

Si multiplicamos y dividimos por s_y y reescribimos la varianza de X en términos del desvío standard.

$$\begin{aligned}\hat{\beta}_1 &= \frac{s_{xy}}{s_x^2} \\ &= \frac{s_y}{s_x} \frac{s_{xy}}{s_x s_y} \\ &= \frac{s_y}{s_x} r\end{aligned}$$

El valor estimado de β_1 no es más que r sometido a un cambio de escala, salvo que...



Continuaremos con el modelo de regresión simple.

- Retomaremos el estimador MCO en un contexto matricial.
- Exploraremos algunas de sus propiedades estadísticas (\mathbb{E} y \mathbb{Var}).
- Definiremos la partición de *sumas de cuadrados* y el coeficiente de determinación.
- Veremos como incorporar variables explicativas cualitativas en el modelo de RLS.
- Compararemos los resultados obtenidos con este modelo respecto a la prueba t de Inferencia I.



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



Faraway, 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