Unidad 9: Regresión Lineal simple

Probabilidad y Estadística

Año 2020

- Dadas las observaciones: (x_1, y_1) , (x_2, y_2) , …, (x_n, y_n)
- Hacer un Diagrama de Dispersión
- Calcular r_{XY}
- Si se sospecha que hay una relación lineal entre las variables se propone un modelo:
- $Y_i = \alpha + x_i \beta + \varepsilon_i$

Modelo de Regresión Lineal Simple

- $y_i = \alpha + x_i \beta + \varepsilon_i$
- y es la variable dependiente
- x es la variable independiente
- Los ε_i son los errores, que se supone tienen media nula, varianza constante y que son no correlacionados
- α y β son los parámetros a estimar

Modelo de regresión lineal simple (formal)

$$\begin{cases} \mathbf{Y_i = \alpha + \beta x_i + \varepsilon_i \quad 1 \leq i \leq n} \\ \mathrm{x_i} \text{ fijos} \\ \varepsilon_\mathrm{i} \text{ aleatorios con:} \\ \qquad \mathrm{E}(\varepsilon_i) = 0 \\ \qquad \text{Varianza}(\varepsilon_i) = \sigma^2 \\ \qquad \mathrm{Cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = 0 \text{ para } \mathrm{i} \neq \mathrm{j} \\ \alpha, \beta \text{ y } \sigma^2 \text{son los parámetros a estimar} \end{cases}$$

$$y_i = \alpha + x_i \beta + \varepsilon_i$$

- ¿Por qué lineal?
- ¿Por qué simple?
- ¿Cuál es el parámetro más importante?
- ¿Qué representa β?
- ¿Qué representa α?
- ¿Qué es ε_i ?

$$y_i = \alpha + x_i \beta + \varepsilon_i$$

- ¿Por qué lineal?
- ¿Por qué simple?
- El parámetro más importante es β, nos indica cuánto aumenta la media de la variable Y cuando aumenta una unidad la variable X
- El parámetro α es la ordenada al origen y representa el valor de la media de Y cuando la variable independiente toma el valor cero.
- ε_i es el error y en base a la minimización de este error se estima α y β

Métodos de Mínimos Cuadrados

- Se trata de minimizar la suma de cuadrados de los residuos:

$$h(\hat{\alpha},\hat{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\alpha} + \hat{\beta} x_i) \right]^2$$

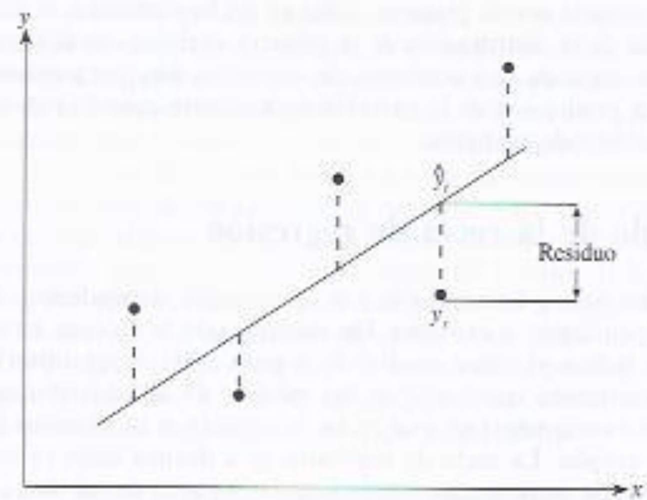


Figura 9.2. *El ajuste de la recta de regresión.*

Métodos de Mínimos Cuadrados

- Como resultado de la aplicación de este criterio se obtiene que:

$$\hat{\beta} = \frac{Cov(X,Y)}{\hat{\sigma}_x^2} = r_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

- Otras formas de $\hat{\beta}$:

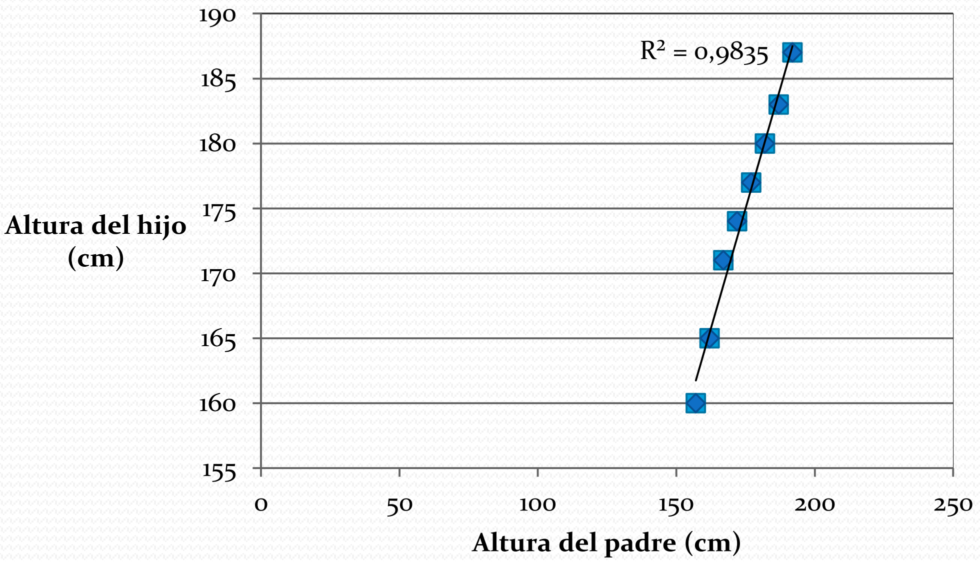
$$\hat{\beta} = \frac{Cov(X,Y)}{\hat{\sigma}_x^2} = r_{XY}\frac{\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}{\sqrt{(x_i-\overline{x})^2}} = r_{XY}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X} = \frac{\sum_{i=1}^{n}x_i y_i - n\overline{xy}}{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2}$$

Fin primera parte

Ejemplo

- Se desea predecir la estatura del hijo en base a la estatura del padre. Los datos observados , en cm, son:

Padre (X)	Hijo (Y)
157	160
162	165
167	171
172	174
177	177
182	180
187	183
192	187



r=0,992

Modelo

$$\begin{cases} \mathbf{Y_i = \alpha + \beta x_i + \varepsilon_i \quad 1 \leq i \leq n} \\ \mathrm{x_i} \text{ fijos} \\ \varepsilon_\mathrm{i} \text{ aleatorios con:} \\ \qquad \text{media cero} \qquad (\mathrm{E}(\varepsilon_i) = 0) \\ \qquad \text{Varianza cte } \left(\sigma^2_{\varepsilon_i} = \text{cte}\right) \\ \qquad \mathrm{Cov}\left(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}\right) = 0 \text{ para } \mathrm{i} \neq \mathrm{j} \\ \alpha, \beta \text{ y } \sigma^2 \text{son los parámetros a estimar} \end{cases}$$

Uso de calculadora

Calculadora símil fx82

- https://www.youtube.com/watch?v=NN_eoHgJQQk

Calculadora símil fx570

- https://www.youtube.com/watch?v=qyHjD2Ia6Dw

Uso de calculadora

Calculadora símil fx82

- Link cátedra

Calculadora símil fx570

- Link cátedra

Padre (X)	Hijo (Y)	xy	x^2
157	160	25120	24649
162	165	26730	26244
167	171	28557	27889
172	174	29928	29584
177	177	31329	31329
182	180	32760	33124
187	183	34221	34969
192	187	35904	36864
1396	**1397**	**244549**	**244652**

$$\hat{\beta} = 0.74$$

$$\hat{\alpha} = 46.24$$

$$\hat{y} = 46.24 + 0.74x$$

Predicciones

- Predicción para x=158

$$\hat{Y}_{158} = 46.24 + 0.74\text{x}158 = 161.99$$

$$\hat{Y}_{150} = No\ se\ puede\ calcular$$

- **No se puede extrapolar**

Medida de variabilidad

- La recta de regresión se complementa con una medida de variabilidad de los datos.
- La diferencia entre la recta de regresión y la observación es lo que llamamos residuos. Se puede mostrar que la suma de los residuos es cero.
- La medida de variabilidad de la recta es igual a:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Ejemplo

- $\hat{\sigma}_{\varepsilon}^{2}$=1.64,   luego
- $\hat{\sigma}_{\varepsilon}$=1.28

Fin Segunda parte

Diagnóstico del Modelo

- Validez
- Bondad

Validez del modelo

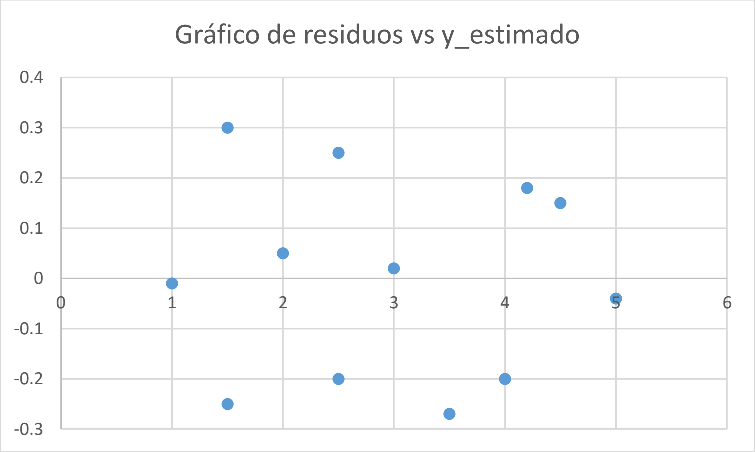
- Se debe validar las hipótesis del modelo

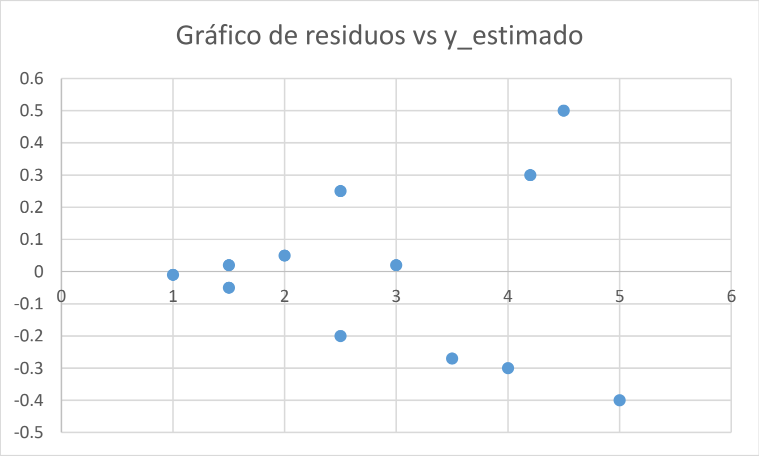
$$\begin{cases} Linealidad \\ E(\varepsilon_i) = 0 \\ \sigma^2_{\varepsilon_i} = cte \\ Cov(\varepsilon_i, \varepsilon_j) = 0 \; para \; i \neq j \end{cases}$$

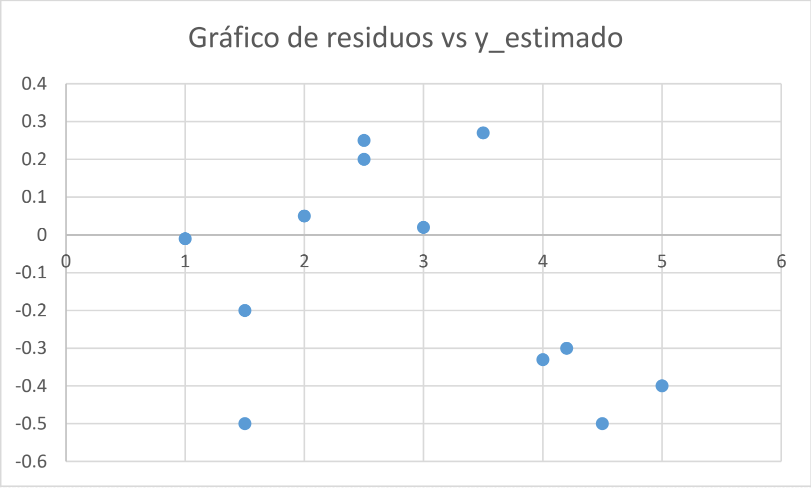
- Mediante un diagrama de dispersión de

$$e_i \quad vs \quad \hat{y}_i$$

Validez del modelo







Bondad del Modelo

- Se debe determinar qué proporción de variabilidad total de y (variable dependiente) es explicada por el modelo.

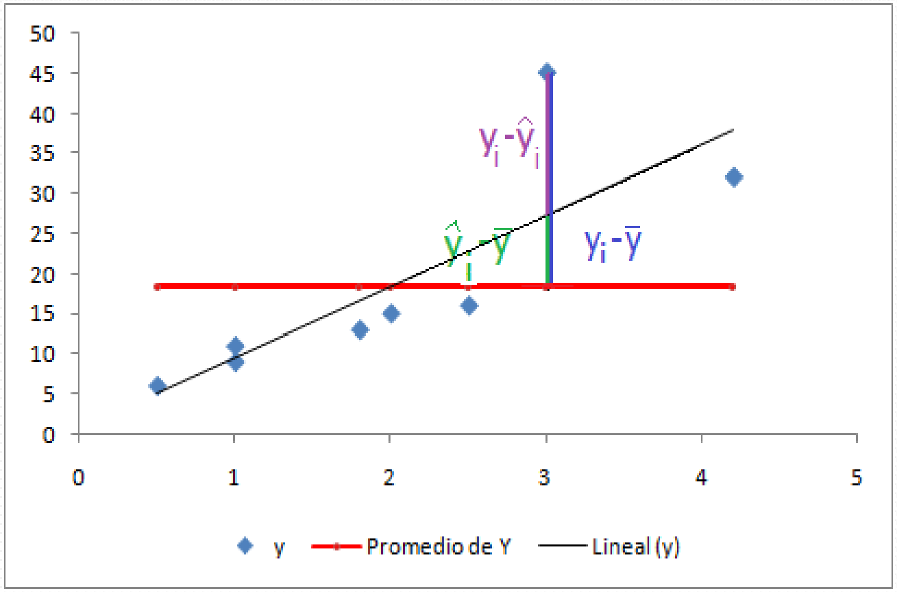
$$\sum (y_i - \bar{y})^2 \quad = \quad \sum (y_i - \hat{y}_i)^2 \quad + \quad \sum (\hat{y}_i - \bar{y})^2$$

Variabilidad total de Y	Variabilidad “no explicada” por la recta	Variabilidad “explicada” por la recta

Bondad del Modelo

$$1 = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} + \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Proporción de la variabilidad total <u>no explicada</u> por la recta	Proporción de la variabilidad total <u>explicada</u> por la recta



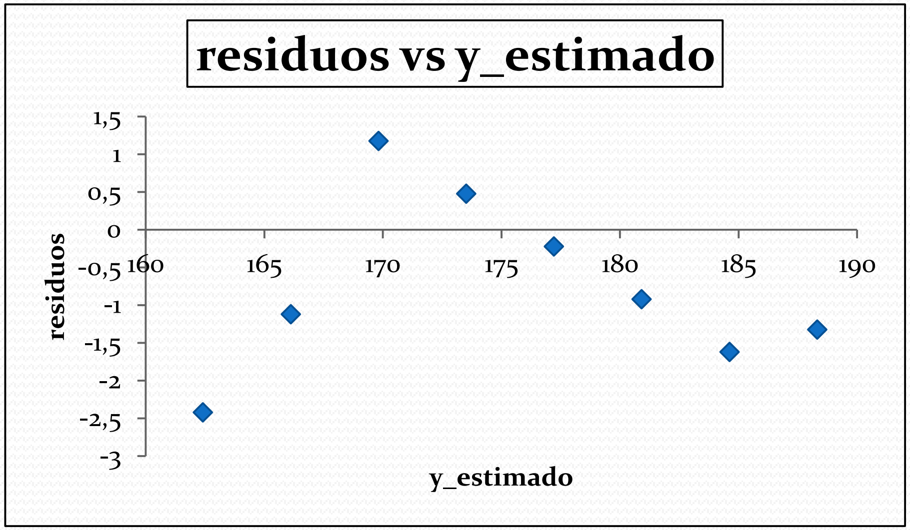
$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

Cuando R^2 es más próximo a 1 es mejor el ajuste

R^2 Se denomina Coeficiente de determinación y es igual a r^2

Ejemplo

Padre (X)	Hijo (Y)	y_estimado	residuos
157	160	162,42	-2,42
162	165	166,12	-1,12
167	171	169,82	1,18
172	174	173,52	0,48
177	177	177,22	-0,22
182	180	180,92	-0,92
187	183	184,62	-1,62
192	187	188,32	-1,32
1396	**1397**		



$$R^2 = 0.98$$

Conclusiones:

El modelo es bueno, pues $R^2 = 0.98$

El modelo no es válido, pues los residuos están correlacionados