



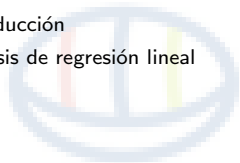
Regresiones Lineales



Instructor: Juan Luis Palacios Soto

palacios.s.j.l@gmail.com

- 1 Introducción
- 2 Análisis de regresión lineal



scidata
matemáticas para
la ciencia de datos

Las regresiones son maneras de *modelar* una característica medida utilizando otras, también medidas, de la misma muestra con el objetivo de crear predicciones. Esto es: si $X_1, X_2, \dots, X_n, X_{n+1}$ son algunas de las columnas de la tabla, encontrar una función f tal que

$$X_{n+1} = f(X_1, X_2, \dots, X_n).$$

En términos más sencillos: ¿será posible explicar el comportamiento de una de las características a través del conocimiento de otras?

Bajo la idea anterior, decimos que las características X_1, X_2, \dots, X_n son explicativas y la característica X_{n+1} es la variable objetivo. *(respuesta)*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

\uparrow $(x_1, x_2, \dots, x_n, y)$

←

RL Simple

RL Multiple

RL Univarada

RL Multivarada

$$Y_1 = \beta_0 + \beta_1 X$$

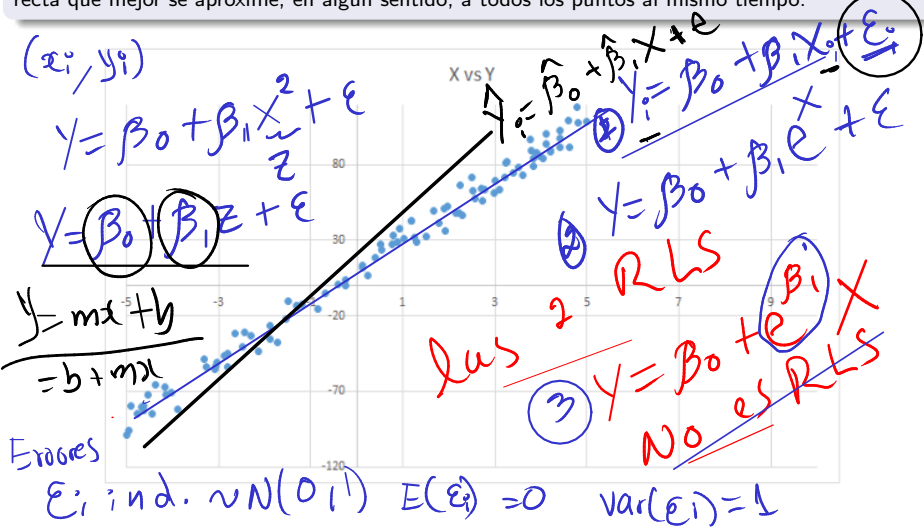
$$Y_2 = \gamma_0 + \gamma_1 X$$

Respuestas Predictores	Y	Y_1, \dots, Y_m
X	RL simple univariada	RL simple multivariada
X_1, X_2, \dots, X_n	RL multiple univariada	RL multiple multivariada.

so $Y = \{0, 1\}$ RL. Logística

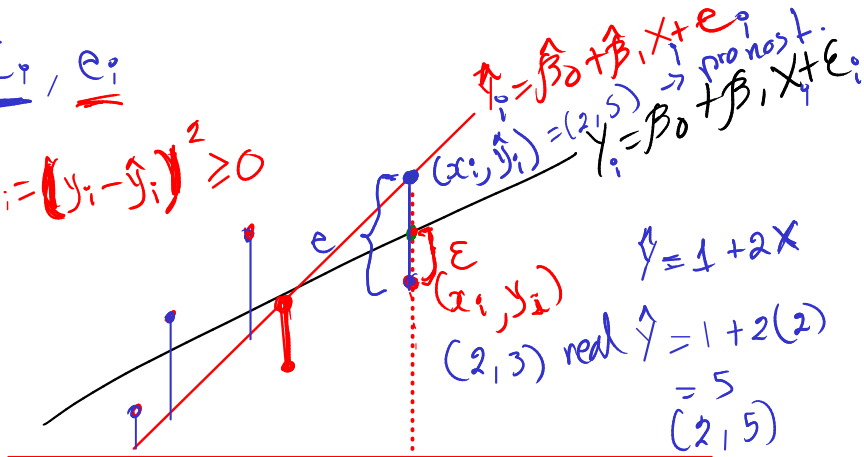
Regresión lineal y nube de puntos

En este capítulo platicaremos de un problema de regresión muy sencillo conocido como regresión lineal. Observemos la siguiente nube de puntos. Debido a su forma, vale preguntarse cuál será la recta que mejor se aproxime, en algún sentido, a todos los puntos al mismo tiempo.



$$\underline{\Sigma_i}, \underline{e_i}$$

$$e_i = (y_i - \hat{y}_i)^2 \geq 0$$



$$SCE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad \text{campo escalar}$$

$$\frac{\partial f}{\partial \hat{\beta}_0} = 0, \quad \frac{\partial f}{\partial \hat{\beta}_1} = 0$$

$$\frac{\partial f}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

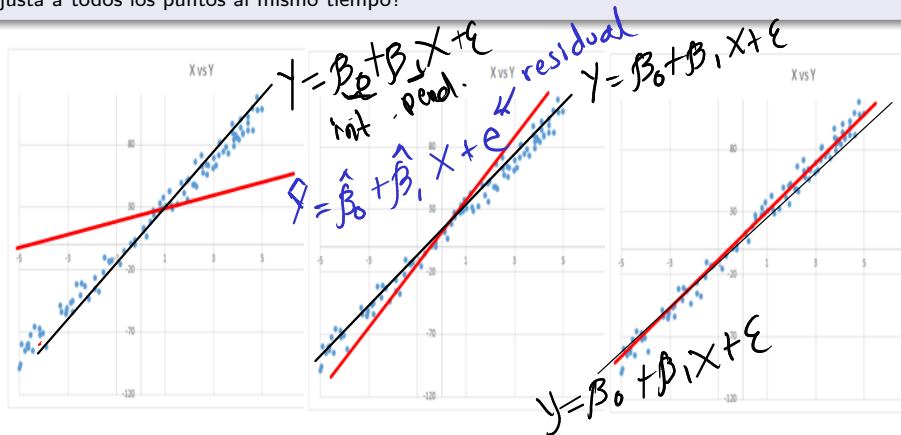
$$\Rightarrow \frac{n}{n} \sum_{i=1}^n y_i - n \hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i = 0$$

$$\Rightarrow \cancel{n} \bar{y} - \cancel{n} \hat{\beta}_0 - \hat{\beta}_1 \cancel{n} \bar{x} = 0$$

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

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \leftarrow \}$$

Observemos varias rectas graficadas con la nube de puntos. ¿Cuál dirías que es la que más se ajusta a todos los puntos al mismo tiempo?




$e \rightarrow \text{error}$

$e \rightarrow \text{residual}$

Definición (Recta)

Antes de avanzar, recordemos nuestros cursos de Geometría Analítica: toda recta en el plano es una ecuación de la forma

$$y = mx + b,$$


*donde b se conoce como ordenada al origen, pero en el contexto de la regresión lineal se conoce como **intercepto**.* 

Mientras que m , tanto para regresión lineal como matemáticas, es la pendiente de la recta. Si $m > 0$ se dice que la pendiente es positiva, mientras que si $m < 0$, diremos que la pendiente es negativa y finalmente si $m = 0$, la pendiente es horizontal.

Por lo tanto, hallar la ecuación de una recta equivale a hallar los valores de m y b .

Objetivo y procedimiento de la regresión lineal

Dada la siguiente lista de pareja de puntos: $(x_1, y_1), (x_2, y_2), \dots, (x_d, y_d)$, los podemos ver como una nube de puntos en el plano cartesiano y deseamos determinar la recta que mejor aproxime, en cierto sentido, a todos los puntos al mismo tiempo. Digamos que $y = mx + b$ es una recta que tomamos como aproximación. Por lo tanto, a cada x_i se le asignan dos números: el y_i (que es un valor que conocemos) y el $\hat{y}_i = mx_i + b$, que es el valor que nos da la recta para ese número x_i .

▷ **Predicciones:** Las predicciones son los valores $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_d$. Es decir, los valores que la recta predice.

▷ **Residuos:** “¿Qué tanto se equivocó la recta?”. Como la recta de aproximación, le asigna a x_i el valor \hat{y}_i , pero el valor verdadero que corresponde a x_i es y_i , los residuos son los errores que la recta “cometió”. Dichos errores los denotamos como

$$e_i = (y_i - \hat{y}_i)^2$$

- ▷ **Error estándar:** Es la desviación estándar de los residuos. También se conoce como error cuadrático medio y se interpreta como el error cometido conjuntamente por la recta en todos los puntos.
- ▷ **Coefficiente de determinación.** Es la proporción de la varianza total de la variable explicada por la regresión. El coeficiente de determinación, denotado por R^2 , refleja la bondad del ajuste (grado de acoplamiento) de un modelo a la variable que pretender explicar.

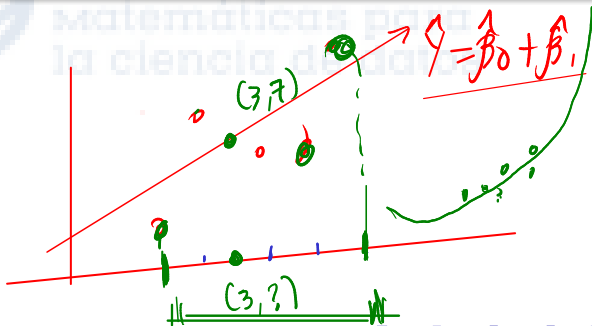


¿Cómo se interpreta?

- 1 Tomar una tabla de datos que contiene al menos dos columnas y d renglones. Además, hay dos columnas, llamadas X y Y , de las cuales obtenemos la regresión lineal (vía Excel o cualquier otro software), lo que significa que obtenemos números m y b tales que la recta $\hat{y} = mx + b$ es la que mejor aproxima a la nube de puntos.
- 2 Si tenemos evidencia para afirmar que el modelo es bueno (lo que significa que los errores cometidos fueron suficientemente pequeños), entonces al recibir un nuevo dato del cual desconocemos su característica Y (y_{d+1}), pero conocemos su valor en la característica X (x_{d+1}), entonces el valor en Y deberá ser aproximadamente

$$\hat{y}_{d+1} = mx_{d+1} + b$$

X	Y
x_1	y_1
x_2	y_2
\vdots	
x_d	y_d



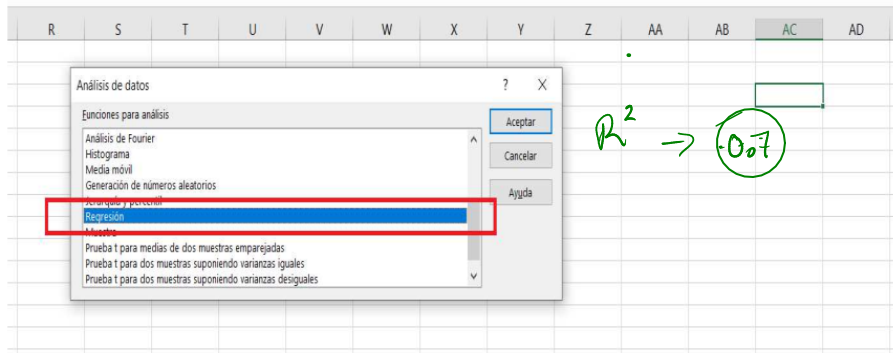
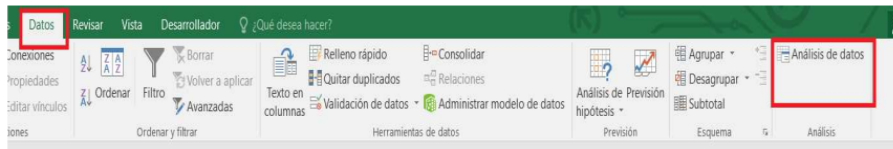
¿Qué significa que la recta sea la que mejor aproxime a todos los puntos al mismo tiempo? Bueno, cada recta que das te genera una lista de residuos. La suma de todos los residuos te da el error cometido globalmente. Pero eso es precisamente el error estándar. O sea que buscas la recta que te consiga el menor error estándar posible de todos.

Analíticamente, el problema anterior se escribe como: Hallar números m y b tales que la función

$C(m, b) = \sqrt{\frac{1}{d} \sum_{i=1}^d (mx_i + b - y_i)^2}$ sea mínima. Esto es lo que se conoce como Problema de los **mínimos cuadrados**. Para el caso de la **regresión lineal simple**, se sabe que si tomamos a la pendiente como $m = \frac{\sum (x_i - \bar{X}) \sum (y_i - \bar{Y})}{\sum (x_i - \bar{X})^2}$ y $b = \bar{Y} - m\bar{X}$, entonces la recta $y = mx + b$ es la que mejor aproxima.

Regresión lineal simple en Excel

En Excel, se usa la aplicación **Regresión** del panel especial de **Análisis de datos** dentro del menú **Datos**.



$(2, 1), (3, 4), (3, 5), (4, 7), (5, 6)$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^5 (x_i - \bar{x})^2}$$

x	y
2	1
3	4
3	5
4	7
5	6

