Introducción a los Modelos Lineales

Adrián Rey, Valentina Cardona & Juan Diego Baez
GitHub https://github.com/vcardonas/intro-modelos-lineales
Rpubs https://rpubs.com/vcardonas/intro-modelos-lineales

Contents

1 Introducción

Los modelos lineales son ampliamente utilizados en estadística y aprendizaje automático debido a su simplicidad y facilidad de interpretación.

1.1 Asociación vs Causalidad

Asociación	Causalidad
* Cuantifica la **fuerza de relación** entre dos variables.	* Relación en la que un cambio en una variable (la **causa**) produce un cambio en otra variable (el **efecto**).

Asociación	Causalidad
* El análisis de **correlación** mide la asociación, no significa causalidad. * El coeficiente de correlación indica el grado de relación simultáneo entre dos variables. $\hat{\rho}_{x,y} = \hat{\rho}_{y,x}$	* El análisis de **regresión** puede ayudar a confirmar una relación de causa y efecto, pero no puede ser la única base de tal afirmación. * En la regresión me interesa construir una función que arroje una **predicción**. $Y = f(X) + e$

Por lo que en la regresión, se pueden hacer dos diferentes análisis:

- **Análisis explicativos:** Desentrañar la estructura y la forma de función.
- **Análisis predictivos:** Aprender la función para meterle valores de X y obtener predictores de Y.

Advertencias:

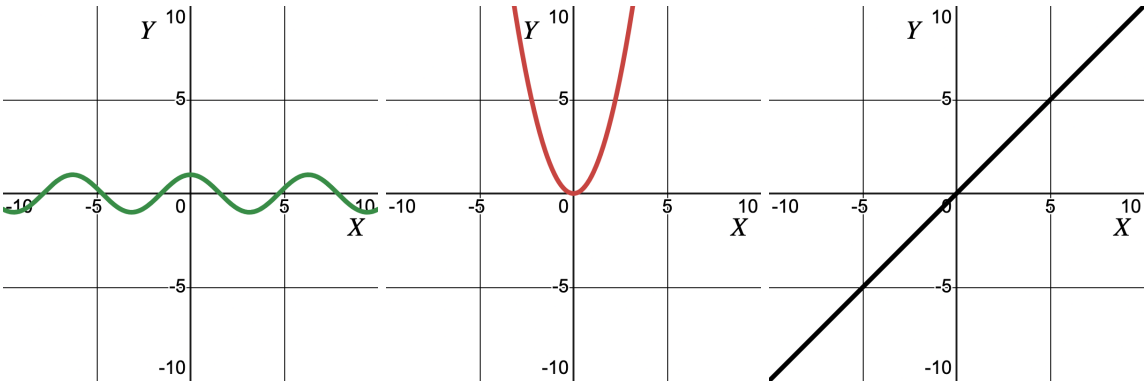
- La causalidad implica una correlación necesaria.
- Se puede predecir si la relación **NO** es causal.
- La causalidad puede ser **bidireccional**.
- La dependencia (escogencia de variable independiente) puede ser **pragmática**.
- Puede existir causalidad así hayan datos atípicos, esto debido a **errores de medición**.

1.2 ¿Qué es un Modelo Lineal?

Un modelo proporciona **un marco teórico para comprender mejor un fenómeno de interés**, brindando aproximaciones útiles sobre las relaciones entre las variables.

1.2.1 Una Ecuación Lineal

En un modelo, se asume que **la relación entre** X y Y se puede capturar por medio de una función matemática.



Le llamamos modelos **lineales** porque esta función es una **ecuación lineal**, la cual describe una línea recta en un plano cartesiano (espacio bidimensional).

1.2.2 Ejemplo gráfico

Una ecuación lineal tiene la forma:

$$y = a + bx$$

donde

- a: Es el valor inicial o **intercepto**.
- b: Es la **pendiente**, que determina cómo cambia y por cada unidad de cambio en x.

Una ecuación lineal expresa que la **combinación lineal** de las variables es igual a un valor constante b.

```
knitr::include_url("https://www.desmos.com/calculator/tohv4xqb0k")
```

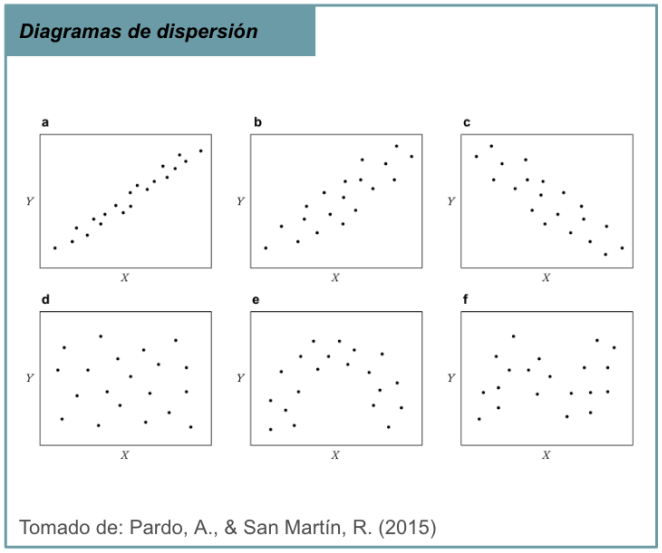
Explorar: https://www.desmos.com/calculator/tohv4xqb0k

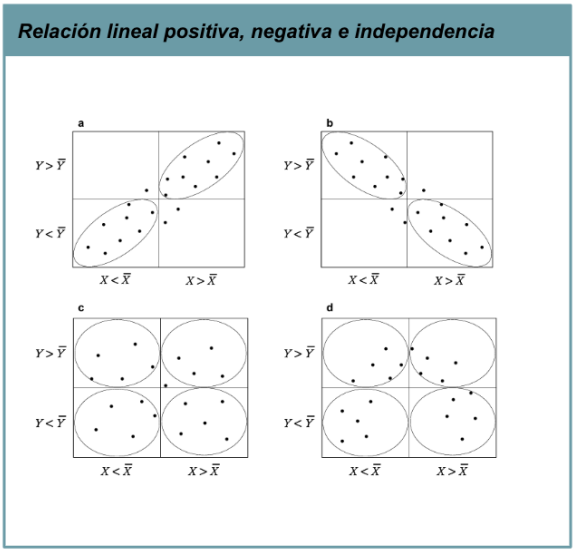
```
# Vectores
x1 <- c(1, 1)
x2 <- c(0, 2)

# Combinaciones lineales
(2 * x1) - (3 * x2)

(-10 * x1) + (2000 * x2)

((2/3) * x1) + (sqrt(35) * x2)
```

```
## [1]  2 -4
## [1]  -10 3990
## [1]  0.6666667 12.4988262
```

2 Correlación

2.1 Visualización entre variables cuantitativas





2.2 Cuantificar la intensidad de la relación

Covarianza

Sujetos	*X*	*Y*	*Z*	*V*	*W*	*x*	*y*	*z*	*v*	*w*	*xy*	*xz*	*xv*	*xw*
1	1	1	6	3	3	−4	−4	1	−2	−2	16	−4	8	8
2	2	2	9	1	8	−3	−3	4	−4	3	9	−12	12	−9
3	3	3	7	7	6	−2	−2	2	2	1	4	−4	−4	−2
4	4	4	8	4	1	−1	−1	3	−1	−4	1	−3	1	4
5	6	6	4	8	7	1	1	−1	3	2	1	−1	3	2
6	7	7	2	2	4	2	2	−3	−3	−1	4	−6	−6	−2
7	8	8	1	9	9	3	3	−4	4	4	9	−12	12	12
8	9	9	3	6	2	4	4	−2	1	−3	16	−8	4	−12
Sumas											60	−50	30	1

Media 5

2.3 Otras medidas de asociación

Covarianza

$$Cov_{XY} = S_{XY} = \frac{\sum_i x_i y_i}{n-1}$$

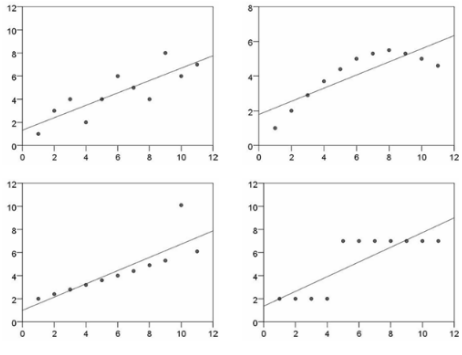
Covarianza

$$S_{XY} = \frac{60}{7} = 8{,}57, \quad S_{XZ} = \frac{-50}{7} = -7{,}14, \quad S_{XV} = \frac{30}{7} = 4{,}29, \quad S_{XW} = \frac{1}{7} = 0{,}14$$

2.4 Correlación de Pearson

$$R_{XY} = S_{XY}/(S_X S_Y)$$

Negativa	Independiente	Positiva
-1	0	1



$R_{XY} = 0{,}84$

Tomado de: Pardo, A., & San Martín, R. (2015)

Relaciones espurias

Correlación de Spearman

https://rpsychologist.com/correlation/

https://www.guessthecorrelation.com/

3 Regresión lineal

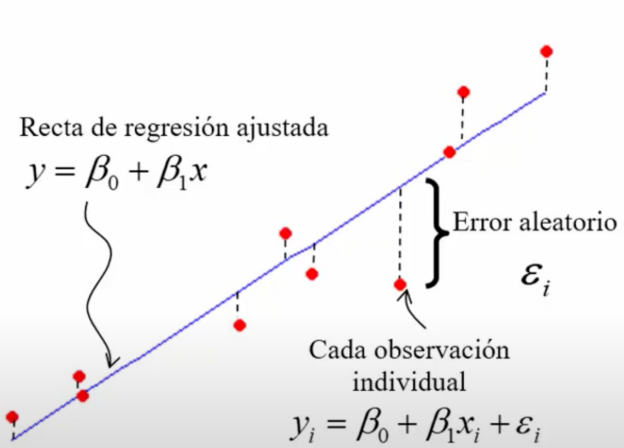
La ecuación de una línea recta que relaciona dos variables es

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}$$

¡Pero el mundo NO es estrictamente lineal!

Aunque los datos pueden seguir un patrón aproximado lineal, *la línea generalmente no pasará por encima de todos los puntos.*

Esto se debe a los **errores**, que representan la distancia entre los datos observados y la línea recta.



Tomado de: Regresión Lineal Simple. Conceptos básicos

El objetivo de la regresión lineal es encontrar los coeficientes β que minimicen estos errores.

Esto se hace generalmente utilizando el **método de mínimos cuadrados**.

Jugar a minimizar el error: https://huggingface.co/spaces/FreddyHernandez/linear_regression_game

3.1 Regresión Lineal Simple

La regresión lineal simple tiene la siguiente forma:

$$\mathbf{y} = \beta_\mathbf{0} + \beta_1 \mathbf{x} + \epsilon$$

donde

- $\mathbf{y}$ es la variable dependiente o respuesta.
- $\mathbf{x}$ es la variable independiente o predictora.
- $\beta_\mathbf{0}$ es el intercepto (el valor de y cuando $x = 0$).
- β_1 es la pendiente de la línea o el coeficiente que indica el cambio en y por cada unidad de cambio en x.
- ϵ es el término de error, que captura la variabilidad en y no explicada por x.

Aclaración: El modelo es lineal en los parámetros (β) no en las variables independientes ($\mathbf{X}$).

3.2 Regresión Lineal Múltiple

Cuando la respuesta $\mathbf{y}$ está influenciada por **más de una variable predictora**, se extiende a la forma:

$$\mathbf{y} = \beta_\mathbf{0} + \beta_1 \mathbf{x_1} + \beta_2 \mathbf{x_2} + \cdots + \beta_{p-1} \mathbf{x}_{p-1} + \epsilon$$

o de forma matricial:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}_{n\times p} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}_{p\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1} \tag{2}$$

3.2.1 Representación geométrica

Se puede observar que $\mathbf{X}\beta$ es una combinación lineal de los vectores $\mathbf{X}$.

En regresión lineal, buscamos **modelar la relación entre las variables predictoras X y la respuesta y usando un vector de coeficientes** β.

El **espacio columna de X** son todas las combinaciones lineales posibles.

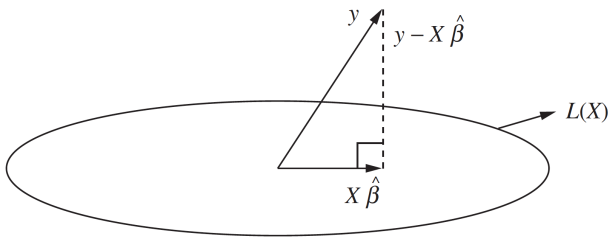
Pero **la variable respuesta y no hace parte de ese espacio columna**.

¿Qué se debe hacer?

Proyectar a $\mathbf{y}$ sobre el espacio columna de $\mathbf{X}$ y encontrar la que más minimice el error.

De forma que las predicciones serán:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$



Tomado de: Abraham & Ledolter (2004)

3.3 Ejemplo en R

Los datos a analizar corresponden a una serie de variables sociales, educativas, deportivas y de seguridad en las zonas rurales de Antioquia en Colombia para el año 2016.

Los datos son tomados de:

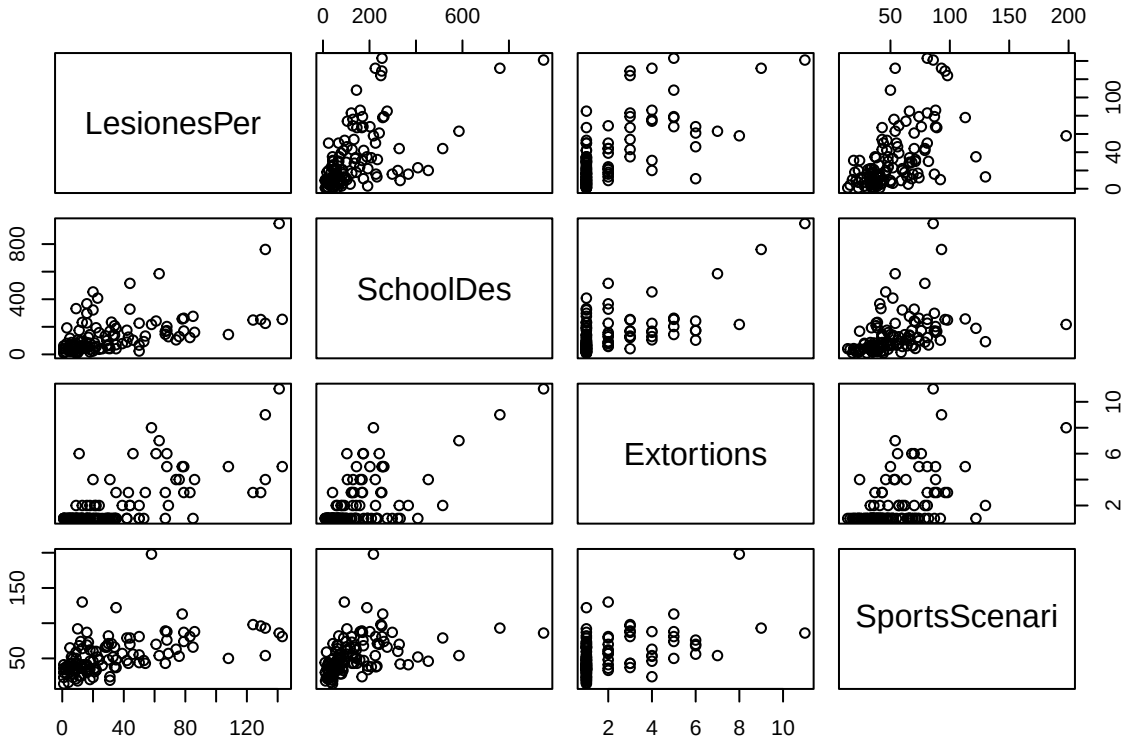
- Shiny App URL
- RStudio Cloud Project

que a su vez, se encuentran disponibles públicamente en la página del Anuario Estadístico de Antioquia del Departamento Administrativo de Planeación.

```
datos <- read.csv("./data/safety_data.csv", dec = ",")
str(datos)
```

```
## 'data.frame':    118 obs. of  10 variables:
##  $ Subregion          : chr  "Valle de Aburra" "Valle de Aburra" "Valle de Aburra" "Valle de Aburra"
##  $ Municipality       : chr  "Barbosa" "Caldas" "Copacabana" "Girardota" ...
##  $ ProjectedPopulation: num  50.8 78.8 71 55.5 63.3 ...
##  $ Thefts             : int  115 141 385 174 264 506 17 212 17 5 ...
##  $ TrafAccid          : int  19 12 18 15 6 12 10 22 4 1 ...
##  $ Homicides          : int  7 12 19 8 9 7 22 33 40 1 ...
##  $ SchoolDes          : int  105 276 254 129 172 225 228 761 515 332 ...
##  $ SportsScenari      : int  63 66 81 53 76 54 38 93 79 42 ...
##  $ Extortions         : int  4 1 5 4 6 4 1 9 2 1 ...
##  $ LesionesPer        : int  74 85 143 76 68 132 16 132 44 9 ...
```

```
# Diagrama de dispersión
plot(datos[c("LesionesPer", "SchoolDes", "Extortions", "SportsScenari")])
```




```r
# Pearson
cor(datos[c("LesionesPer", "SchoolDes", "Extortions", "SportsScenari")])
```

```
##                LesionesPer SchoolDes Extortions SportsScenari
## LesionesPer      1.0000000 0.5664804  0.6961702     0.5075335
## SchoolDes        0.5664804 1.0000000  0.6619255     0.4072873
## Extortions       0.6961702 0.6619255  1.0000000     0.4922734
## SportsScenari    0.5075335 0.4072873  0.4922734     1.0000000
```

```r
# Regresión Lineal Simple
mls <- lm(LesionesPer ~ Extortions, data = datos)

# Resumen del modelo
summary(mls)
```

```
##
## Call:
## lm(formula = LesionesPer ~ Extortions, data = datos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -71.52 -14.68  -4.73  10.02  83.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    8.372      3.229   2.593   0.0108 *
## Extortions    12.358      1.183  10.445   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.99 on 116 degrees of freedom
## Multiple R-squared:  0.4847, Adjusted R-squared:  0.4802
## F-statistic: 109.1 on 1 and 116 DF,  p-value: < 2.2e-16
```

Interpretación:

- **Intercepto (8.372):** Cuando no hay extorsiones (Extortions = 0), el modelo predice un promedio de 8.372 lesiones personales.
- **Pendiente (12.358):** Por cada aumento de una unidad en extorsiones, se predice un incremento promedio de 12.358 en las lesiones personales. Este coeficiente es estadísticamente significativo (**p < 0.001**).
- **Coeficiente de determinación (R-squared):** expresa el porcentaje de varianza explicado por el modelo, en este caso, es del 40.5%
- **Coeficiente de determinación (Adjusted R-squared):** penaliza a R^2 dependiendo el número de parámetros. **No se puede interpretar como porcentaje de varianza explicado**. Tiene valores entre 0 y 1, y cuánto más cercano a 1, mejor.

4 Bibliografía

- Abraham, B. & Ledolter J. (2004). Introduction to Regression Modeling.
- Rencher, A. C. & Schaalje, G. B. (2008). Linear models in statistics. John Wiley & Sons.