Regresión lineal

Introducción a R para Ciencias Sociales

Pablo Aguirre Hörmann

Regresión lineal

En una regresión lineal el objetivo es poder modelar la relación entre una variable dependiente Y y una o varias variables explicativas/independientes $X_1, X_2, ..., X_k$.

Por ejemplo, si un colegio reduce el número de alumnos en una sala de clases al contratar nuevos profesores podríamos pensar en términos de como la reducción de una variable X_1, el *ratio estudiante-profesor*, afecta una variable Y, el *resultado* de los estudiantes en una prueba estandarizada. A través de la regresión lineal no solo podremos saber si es que el *ratio estudiante-profesor* tiene un impacto en los *resultados* de la prueba si no que también la **dirección** y **fuerza** de este efecto.

Regresión lineal simple

Para empezar con un ejemplo simple, crearemos dos vectores siguiendo el ejemplo recién mencionado: `STR` (*ratio estudiante-profesor*) y `Resultados` (*resultados en la prueba*).

```
# crear vectores de datos
STR <- c(15, 17, 19, 20, 22, 23.5, 25)
Resultados <- c(680, 640, 670, 660, 630, 660, 635)

# juntar ambos vectores en un data frame
datos_colegio <- data.frame(STR, Resultados)
```

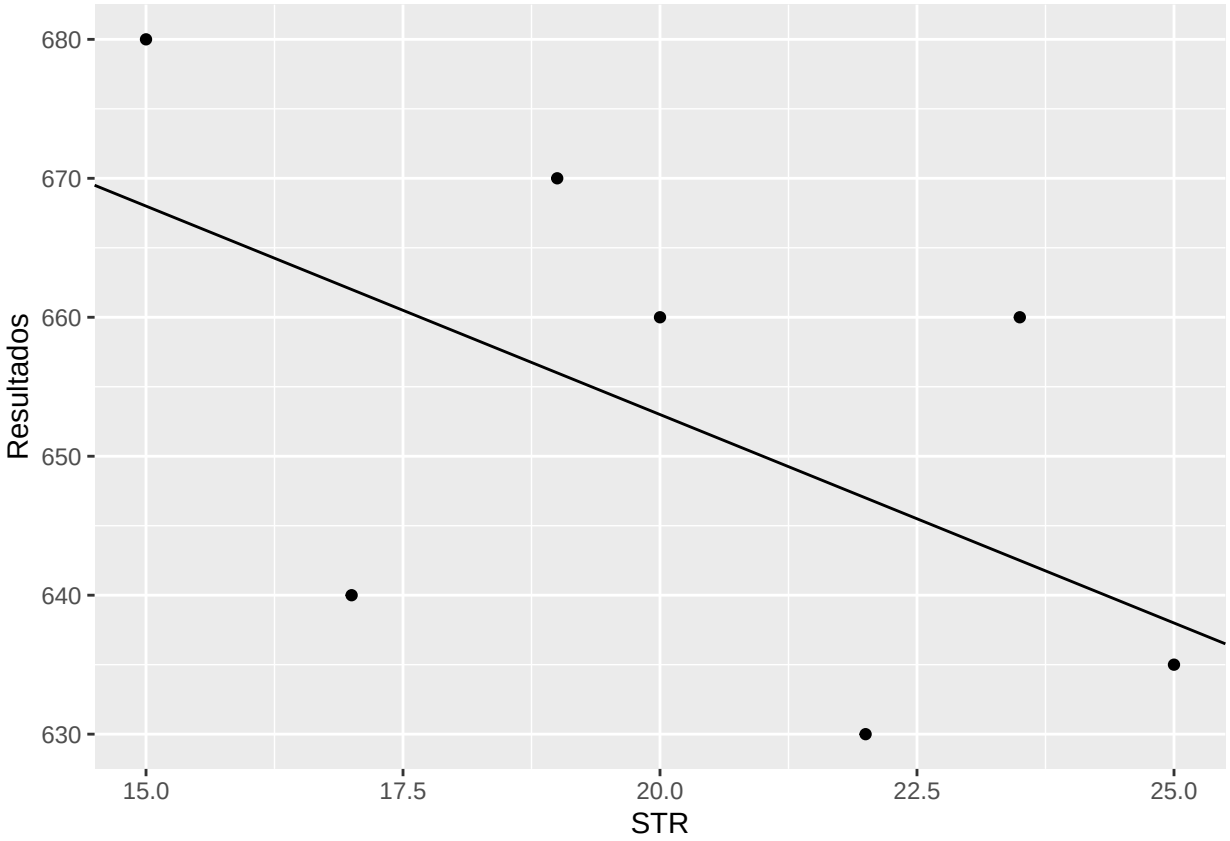
En una regresión lineal simple modelamos la relación entre dos variables a tavés de una linea recta:

$$Y = b * X + a$$

Por ahora, supongamos que la función que relaciona ambas variables es

$$Resultados = 713 - 3 * REP$$

Los datos creados y la **recta** que acabamos de describir se grafican a continuación.



La linea no toca ninguno de los puntos pero podríamos decir que más o menos captura la relación del conjunto de datos. La mayoría de las veces existen muchos otros factores de influencia que implican que no hayan relaciones perfectas entre dos variables.

Con el fin de tomar en cuenta la diferencia entre los datos observados (**puntos**) y la relación intrínsica (**linea**) que pueden presentar dos variable, ampliamos el modelo descrito anteriormente incluyéndo un componente de *error* (u) que capture efectos de aleatoriedad. En otras palabras, u considera todas las diferences entre la linea de regresión y los datos observados.

Entonces, extendemos nuestro modelo de la siguiente manera:

$$Resultados = \beta_0 + \beta_1 * REP + otros factores$$

Y podemos generalizarlo como se describe a continuación:

$$Y_i = \beta_0 + \beta_1 * X_i + u_i$$

Estimación de los coeficientes en un modelo de regresión simple

En la práctica, el intercepto β_0 y la pendiente β_1 son desconocidos. Entonces, lo que debemos hacer es usar datos para tratar de estimar ambos parámetros desconocidos. Lo que haremos ahora es utilizar datos reales para poder hacer esta estimación.

Para ejemplificar esto utilzaremos datos correspondientes a una serie de variables sobre educación. *Estos datos se obtienen de una librería llamada* ***AER*** *(en el módulo 3 se les explicará sobre como instalar y cargar librerías).*

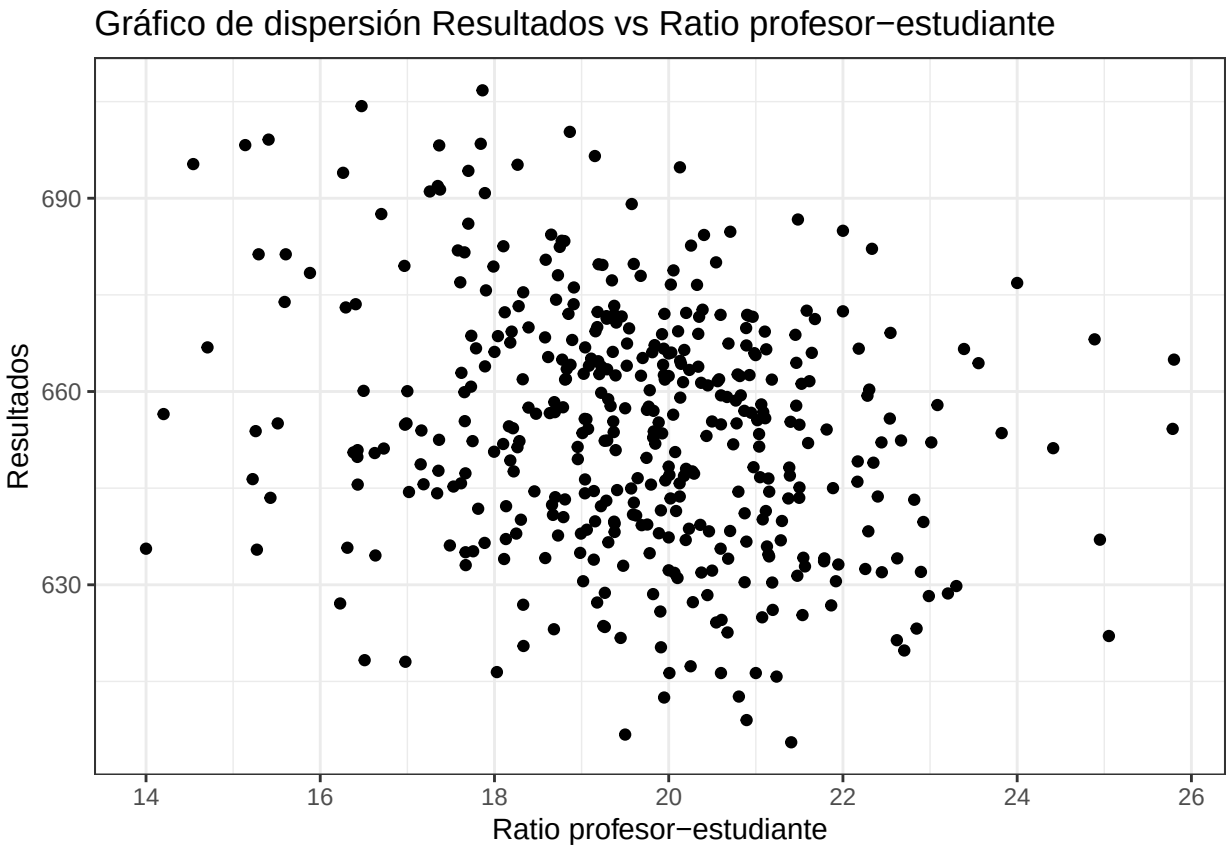
```r
library(AER) #install.packages("AER")
library(tidyverse)  #install.packages("tidyverse")
data("CASchools") #Cargamos datos (estos aparecerán en el "global environment")

# cómo usar estas funciones se verá en el módulo 4
CASchools <- mutate(CASchools,STR = students/teachers,score = (read + math)/2)
```

```r
summary(CASchools)
```

```
##    district            school                  county      grades
##  Length:420         Length:420         Sonoma     : 29   KK-06: 61
##  Class :character   Class :character   Kern       : 27   KK-08:359
##  Mode  :character   Mode  :character   Los Angeles: 27
##                                        Tulare     : 24
##                                        San Diego  : 21
##                                        Santa Clara: 20
##                                        (Other)    :272
##     students          teachers          calworks          lunch
##  Min.   :   81.0   Min.   :   4.85   Min.   : 0.000   Min.   :  0.00
##  1st Qu.:  379.0   1st Qu.:  19.66   1st Qu.: 4.395   1st Qu.: 23.28
##  Median :  950.5   Median :  48.56   Median :10.520   Median : 41.75
##  Mean   : 2628.8   Mean   : 129.07   Mean   :13.246   Mean   : 44.71
##  3rd Qu.: 3008.0   3rd Qu.: 146.35   3rd Qu.:18.981   3rd Qu.: 66.86
##  Max.   :27176.0   Max.   :1429.00   Max.   :78.994   Max.   :100.00
##
##     computer       expenditure       income          english
##  Min.   :   0.0   Min.   :3926   Min.   : 5.335   Min.   : 0.000
##  1st Qu.:  46.0   1st Qu.:4906   1st Qu.:10.639   1st Qu.: 1.941
##  Median : 117.5   Median :5215   Median :13.728   Median : 8.778
##  Mean   : 303.4   Mean   :5312   Mean   :15.317   Mean   :15.768
##  3rd Qu.: 375.2   3rd Qu.:5601   3rd Qu.:17.629   3rd Qu.:22.970
##  Max.   :3324.0   Max.   :7712   Max.   :55.328   Max.   :85.540
##
##       read            math            STR            score
##  Min.   :604.5   Min.   :605.4   Min.   :14.00   Min.   :605.5
##  1st Qu.:640.4   1st Qu.:639.4   1st Qu.:18.58   1st Qu.:640.0
##  Median :655.8   Median :652.5   Median :19.72   Median :654.5
##  Mean   :655.0   Mean   :653.3   Mean   :19.64   Mean   :654.2
##  3rd Qu.:668.7   3rd Qu.:665.9   3rd Qu.:20.87   3rd Qu.:666.7
##  Max.   :704.0   Max.   :709.5   Max.   :25.80   Max.   :706.8
##
```

Las variables que nos interesan (según lo que hemos visto) son *resultados de pruebas* (score) y *ratio estudiante-profesor* (STR). A continuación las graficaremos para ver como se relacionan ambas:



Pareciera haber una relación negativa entre ambas variables pero la verdad es dificil verlo claramente con este gráfico. Para estar seguros, podemos calcular la correlación de ambas variables utilizando la función `cor()`.

```
cor(CASchools$STR, CASchools$score)
```

```
## [1] -0.2263627
```

Efectivamente ambas variables tienen una correlación negativa. Ahora bien, ¿cuál es la mejor linea que describe esta relación? Podriamos tener en cuenta la correlación ya calculada y tratar de elegir la mejor linea según lo que nuestros ojos nos digan, pero esto sería bastante subjetivo y se producirían tantas lineas como observadores hayan. Entonces, a continuación definiremos un método para poder estimar esta linea conocido como Mínimos Cuadrados Ordinarios (MCO u OLS por sus siglas en inglés).

Estimador por Mínimos Cuadrados Ordinarios (MCO)

El estimador de MCO elige los coeficiente que produzcan la curva de regresión más cercana posible a los puntos. La cercanía se define a través de la suma del cuadrado de los errores hechos al estimar Y a partir de X.

Tomemos b_0 y b_1 como estimadores de β_0 y β_1. Entonces, podemos describir la suma del cuadrado de los errores hechos como:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{1i})^2$$

El estimador de MCO en un modelo de regresón simple es el par de coeficientes para el intercepto y pendiente que minimizan la expresión recién descrita. Esto se puede describir como:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

Por su parte, los valores estimados ($\hat{Y}_i$) y los residuales ($\hat{u}_i$) se definen como:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

$$\hat{u}_i = Y_i - \hat{Y}_1$$

Teniendo en cuenta las formulas recién descritas podemos calcular $\hat{\beta}_0$ y $\hat{\beta}_1$ para los datos de `CASchools` de la siguiente manera

```
beta_1 <- sum((CASchools$STR - mean(CASchools$STR)) *
                (CASchools$score - mean(CASchools$score))) /
  sum((CASchools$STR - mean(CASchools$STR))^2)

beta_0 <- mean(CASchools$score) - beta_1 * mean(CASchools$STR)

c(beta0 = beta_0, beta1 = beta_1)
```

```
##      beta0      beta1
## 698.932949  -2.279808
```

Lo bueno es que en `R` ya existe una función (`lm()`) que nos permiten calcular estos coeficientes de forma más simple. La función sigue la forma `lm(VarDependiente ~ VarIndependiente, data = Datos)`:

```
modelo_lineal <- lm(score ~ STR, data = CASchools)
modelo_lineal
```

```
##
## Call:
## lm(formula = score ~ STR, data = CASchools)
##
## Coefficients:
## (Intercept)          STR
##      698.93        -2.28
```

Con estos calculos podemos ahora visualizar como se ve esta curva de regresión estimada.

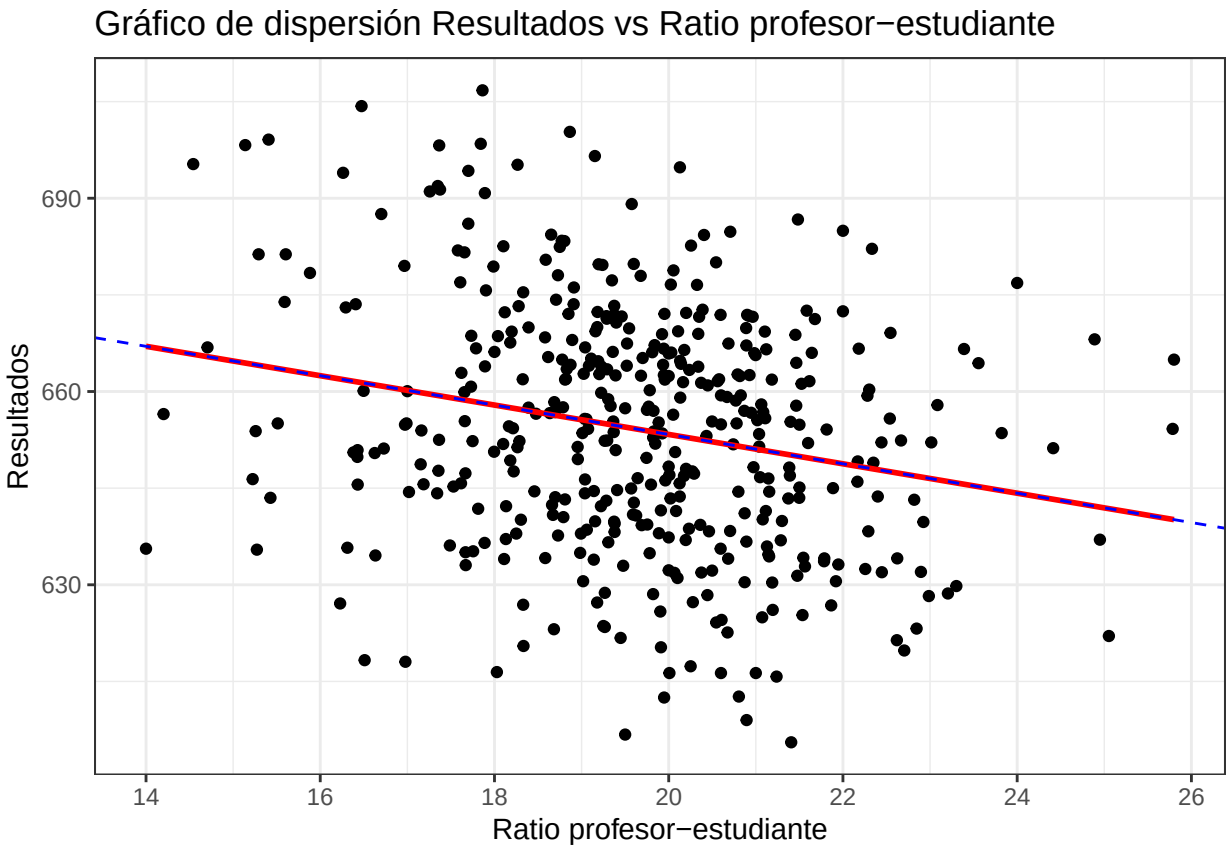


Gráfico de dispersión Resultados vs Ratio profesor−estudiante
Resultados
690
660
630
14
16
18
20
22
24
26
Ratio profesor−estudiante