

# Trabajo Práctico N°1

## Estimadores de Máxima Verosimilitud

## Trabajo Práctico N°1

### Estimadores de Máxima Verosimilitud

Nro. de grupo:

Integrantes:

- De Luca, Francisco. 109794
- Salluzzi, Luca. 108088

### Condiciones de entrega

El trabajo práctico debe realizarse en grupos de 2 personas. Para la entrega del TP deben subir:

- Este notebook
- En caso de resolverlo por fuera del notebook, un archivo en formato **pdf** con los cálculos analíticos solicitados.
- Una versión en pdf del notebook (para poder hacer correcciones)

### Enunciado: Estimación de coeficientes de una regresión lineal

Un problema muy común en la estadística es el de estimación. En este caso, tendremos una variable objetivo,  $Y$ , que se desea estimar a partir de  $r$  variables observables,  $X_1, X_2, \dots, X_r$ . Un modelo de regresión lineal para este problema será de la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_r X_r + \varepsilon,$$

donde  $\varepsilon$  es un término de error que podemos suponer sigue una distribución Normal de media 0 y varianza desconocida  $\sigma_\varepsilon^2$ .

El objetivo del problema de regresión lineal resulta hallar los valores óptimos para  $\beta_0, \beta_1, \dots, \beta_r$  a partir de  $n$  observaciones del vector aleatorio  $(Y, X_1, \dots, X_r)$ . Una forma de hacerlo es a partir del estimador de máxima verosimilitud.

Para ello debemos observar que dada una observación  $\mathbf{x} = (x_1, \dots, x_r)$ , la distribución condicional de  $Y$  para  $\mathbf{x}$  es también Normal:

$$Y|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \dots + \beta_r x_r, \sigma_\varepsilon^2)$$

**1. Ejercicio 1** Consideren un modelo de regresión lineal a partir 1 variable predictora, que tiene la forma

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

**a. Estimador de máxima verosimilitud** Encontrar analíticamente el estimador de máxima verosimilitud para  $\beta_0, \beta_1$ , a partir de una muestra aleatoria de tamaño  $n$ .

Observación: Si te sentís cómodo usando LaTeX, podés hacer el desarrollo acá abajo.

$$L(\beta_0, \beta_1) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i + \beta_0 + \beta_1 x_i)^2}$$

$$\ln(L(\beta_0, \beta_1)) = \frac{n}{\sqrt{2\pi}\sigma} + \sum_{i=1}^n \frac{1}{2\sigma^2} (-y_i + \beta_0 + \beta_1 x_i)^2$$

I)

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_0} = \frac{1}{2\sigma^2} \sum_{i=1}^n (-y_i + \beta_0 + \beta_1 x_i) = 0$$

$$\sum_{i=1}^n (-y_i + \beta_0 + \beta_1 x_i) = 0$$

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

II)

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_1} = \frac{1}{2\sigma^2} \sum_{i=1}^n (-y_i + \beta_0 + \beta_1 x_i) x_i = 0$$

utilizando I)

$$\sum_{i=1}^n \left[ \left( y_i - \frac{1}{n} \sum_{j=1}^n (y_j - \hat{\beta}_1 x_j) - \beta_1 x_i \right) x_i \right] = 0$$

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

##### b. (Opcional) ¿El resultado les resulta conocido? Una vez finalizado el trabajo nos dimos cuenta que la obtención del EMV es similar a un problema de cuadrados mínimos, ya que si en vez de despejar  $\beta_0$  y  $\beta_1$  de las ecuaciones como hicimos, se puede plantear un sistema de la forma:  $Ax = b$  Donde  $A$  es una matriz de dimensiones  $n \times 2$ ,  $x$  es de dimensiones  $2 \times 1$  y  $b$  es de  $n \times 1$ . Una vez obtenido esto como el sistema no tiene solución se puede multiplicar por  $A^t$  en ambos lados para así obtener una solución de  $\hat{b}$ .

**2. Ejercicio 2** Probemos ahora cómo funciona esto con un conjunto de datos. Vamos a usar el dataset `Income`, que mide la felicidad en función del salario. Si vas a descargar el archivo desde el link, recordá mover el archivo csv dentro de la misma carpeta que el notebook. Para cargarlo, vamos a usar la función `read_csv`.

```
XY <- read.csv("income.data.csv", row.names = 1)
```

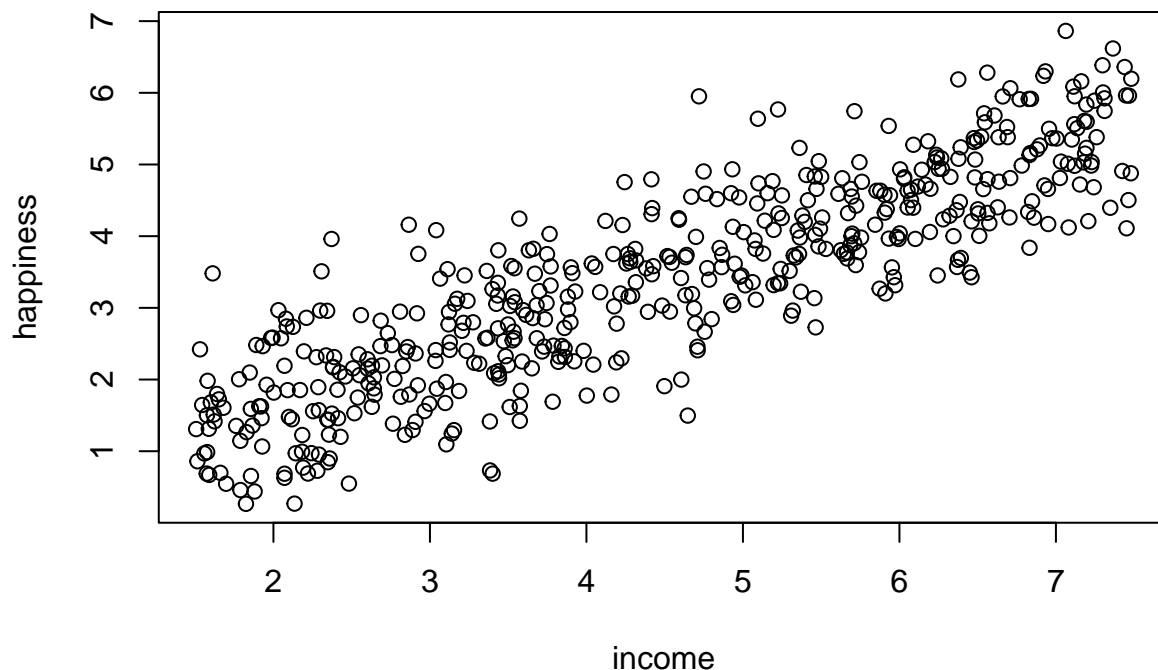
A continuación vamos a visualizar un resumen de los mismos usando la función `summary`, donde podemos ver algunos valores representativos de cada variable, como valor mínimo, máximo y los cuatro cuartiles.

```
summary(XY)
```

```
##      income      happiness
##  Min.   :1.506   Min.     :0.266
## 1st Qu.:3.006   1st Qu.:2.266
##  Median :4.424   Median :3.473
##   Mean  :4.467   Mean    :3.393
## 3rd Qu.:5.992   3rd Qu.:4.503
##   Max.  :7.482   Max.    :6.863
```

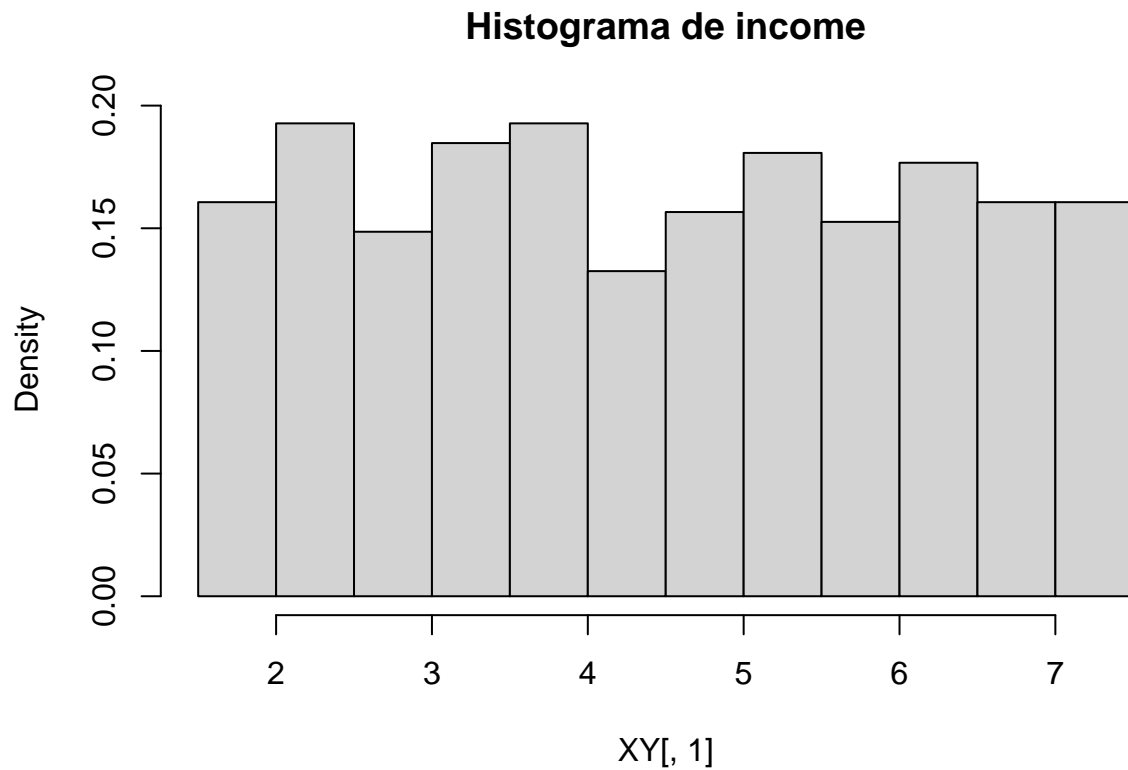
Grafiquemos los datos en un *scatter plot*:

```
plot(XY)
```



Podemos graficar también un histograma de cada una de las variables, para conocer un poco mejor su distribución. Esto lo podemos hacer con la función `hist`

```
hist(XY[, 1], main = "Histograma de income", freq = FALSE) # Graficamos la columna income
```



```
hist(XY[, 2], main = "Histograma de happiness", freq = FALSE) # Graficamos la columna happiness
```



a. **Análisis de los gráficos** ¿Qué pueden decir a partir de estos gráficos? Vincular con el modelo propuesto

Se puede decir, observando el scatter plot, que la variable `income` y `happiness` tienen una relación lineal positiva. Es decir, a medida que aumenta el ingreso, aumenta la felicidad. Al tener únicamente la variable objetivo (`happiness`) y una sola predictora (`income`), equivale al sistema lineal de dos ecuaciones con dos incógnitas. Por lo tanto, se puede resolver el sistema y obtener los valores de  $\beta_0$  y  $\beta_1$ .

Luego, en el histograma de `income` se puede observar que la distribución de los datos aparenta tener una distribución uniforme, con algunas perturbaciones.

Finalmente, en el histograma de `happiness` se puede observar que los datos distribuyen de forma similar a los datos de una distribución normal. ya que posee una mayor densidad en el centro y en los extremos la densidad disminuye.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

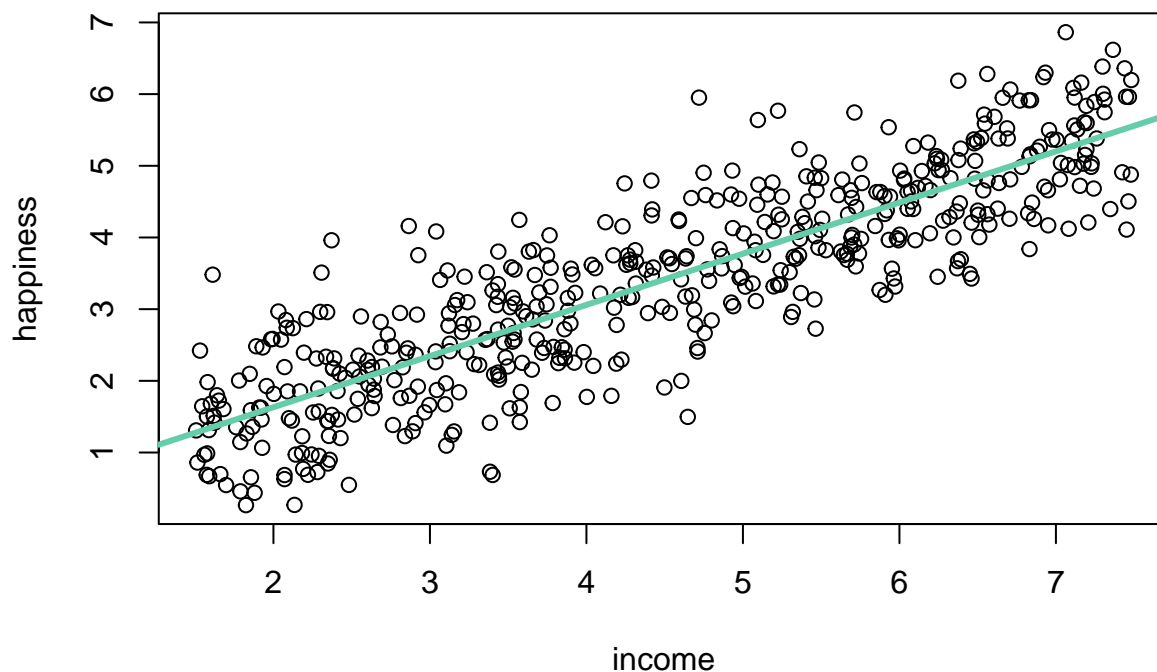
**b. Estimación de  $\beta_0$  y  $\beta_1$  usando R** Usar las herramientas de R para calcular una estimación de los valores de  $\beta_0$  y  $\beta_1$  basada en los datos cargados. Podés hacer esto en la siguiente celda. Remplazá los valores de NULL por las expresiones correspondientes, también podés agregar cálculos auxiliares en líneas anteriores.

```
beta1 <- sum((XY$income - mean(XY$income)) * (XY$happiness - mean(XY$happiness))) / sum((XY$income - mean(XY$income))^2)

beta0 <- sum(XY$happiness) / nrow(XY) - beta1 * sum(XY$income) / nrow(XY)
```

Graficar nuevamente los datos y superponer la recta estimada a partir de los valores de  $\hat{\beta}_0$  y  $\hat{\beta}_1$  hallados.

```
plot(XY)
abline(a = beta0, b = beta1, col = "aquamarine3", lw = 3)
```



## Conclusiones

Usen este espacio para finalizar el trabajo práctico con sus comentarios y conclusiones finales.