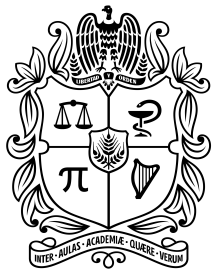
# Estadística II - 3006915
# Regresión Lineal Simple

Mateo Ochoa Medina

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín

Periodo académico 2023-2S



UNIVERSIDAD
NACIONAL
DE COLOMBIA

# Contenido

# Contenido

# Estimación por mínimos cuadrados ordinarios de los parámetros de regresión

Dados los pares de observaciones $(x_1, y_1), \ldots, (x_n, y_n)$, donde, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, los respectivos valores de $\beta_0$ y $\beta_1$ que minimizan a $S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$ son:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{1}$$

y

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \\ &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned} \tag{2}$$

Por consiguiente, $\hat{\beta}_0$ y $\hat{\beta}_1$ en las ecuaciones (1) y (2) son las estimaciones por mínimos cuadrados de $\beta_0$ y $\beta_1$, respectivamente.

# Estimación de la respuesta media en $X = x_i$ y estimación de la varianza

Una estimación de la respuesta media (o respuesta ajustada), en $X = x_i$, es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} + (x_1 - \bar{x})\,\hat{\beta}_1. \tag{3}$$

La diferencia entre el valor observado $y_i$ y el respectivo valor ajustado $\hat{y}_i$ se llama *residual*. Así, el $i$-ésimo residual se expresa como

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \tag{4}$$

De esta manera, una estimación de la varianza $\sigma^2$ a partir de la ecuación (4) está dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{n-2}. \tag{5}$$

# Tipo de sumas

Las principales sumas en el ajuste por mínimos cuadrados son:

1. Suma corregida de cuadrados de las $x_i$:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} (x_i - \bar{x})\, x_i. \quad (3)$$

2. Suma corregida de los productos cruzados de $x_i$ y $y_i$:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})\, y_i = \sum_{i=1}^{n} (x_i - \bar{x})\, (y_i - \bar{y})\,. \quad (4)$$

3. Suma de cuadrados de residuales:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1\, S_{xy}. \quad (5)$$

4. Suma de cuadrados corregida de las $y_i$. También es conocida como suma de cuadrados totales o $SST$:

$$S_{yy} = SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})\, y_i. \quad (6)$$

# Consideraciones de la estimación de los parámetros de regresión

## Notas relacionadas con $\hat{\beta}_0 \left(= \bar{y} - \hat{\beta}_1 \bar{x}\right)$ y $\hat{\beta}_1 \left(= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$:

- $\hat{\beta}_1$ puede ser expresado en función de la suma corregida de cuadrados de las $x_i$ $(S_{xx})$ y de la suma corregida de los productos cruzados de $x_i$ y $y_i$ $(S_{xy})$ así:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}. \tag{7}$$

- Bajo el modelo normal, los estimadores de máxima verosimilitud de los parámetros de regresión $\left(\tilde{\beta}_0 \text{ y } \tilde{\beta}_1\right)$ son iguales a los respectivos estimadores de mínimos cuadrados $\left(\hat{\beta}_0 \text{ y } \hat{\beta}_1\right)$.

# Consideraciones de la estimación de la varianza

## Notas relacionadas con $\hat{\sigma}^2 \left( = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \right)$:

- $\hat{\sigma}^2$ puede ser expresado en función de la suma de cuadrados de residuales ($SSE$) así:

$$\hat{\sigma}^2 = \frac{SSE}{n-2}. \tag{8}$$

- Bajo los supuestos del modelo en relación a los errores, el estimador de $\sigma^2$ es insesgado y es tal que, $E\left(\hat{\sigma}^2\right) = \sigma^2$. Este estimador también recibe el nombre de error cuadrático medio y es denotado por $MSE$.
- Se puede escribir el estimador de máxima verosimilitud de la varianza, $\tilde{\sigma}^2$, de la siguiente forma

$$\tilde{\sigma}^2 = \left(\frac{n-2}{n}\right) \hat{\sigma}^2.$$

# Contenido

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

Bajo la validez de los supuestos considerados sobre los errores, tenemos que:

1. $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{Y}_i$ (respuesta estimada para $X = x_i$) son combinaciones lineales de las variables aleatorias $Y_1, \ldots, Y_n$, por tanto, son variables aleatorias normales. Así:

$$\hat{\beta}_0 = \sum_{i=1}^{n} m_i Y_i, \text{ donde } m_i = \frac{1}{n} - \bar{x}\, c_i,$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i, \text{ donde } c_i = \frac{x_i - \bar{x}}{S_{xx}},$$

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j, \text{ donde } h_{ij} = m_j + c_j x_i = \frac{1}{n} + \frac{(x_i - \bar{x})\,(x_j - \bar{x})}{S_{xx}}.$$

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

2. $\hat{\beta}_0$ y $\hat{\beta}_1$, son los mejores estimadores lineales insesgados de $\beta_0$ y $\beta_1$, respectivamente. Por tanto, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, es un estimador insesgado para $E(Y \mid X) = \beta_0 + \beta_1 X$. Luego:

$$E\left(\hat{\beta}_0\right) = E\left(\sum_{i=1}^{n} m_i Y_i\right) = \beta_0 \sum_{i=1}^{n} m_i + \beta_1 \sum_{i=1}^{n} m_i x_i = \beta_0.$$

Note que $\sum_{i=1}^{n} m_i = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x}\, c_i\right) = 1 - \bar{x} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} = 1$, $\sum_{i=1}^{n} m_i x_i = \sum_{i=1}^{n} \left[\left(\frac{1}{n} - \bar{x}\, c_i\right) x_i\right] = \bar{x} - \bar{x} \sum_{i=1}^{n} \frac{(x_i - \bar{x}) x_i}{S_{xx}} = 0.$

$$E\left(\hat{\beta}_1\right) = E\left(\sum_{i=1}^{n} c_i Y_i\right) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1.$$

Note que $\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} = 0$, $\sum_{i=1}^{n} m_i x_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) x_i}{S_{xx}} = 1$.

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

3. $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{Y}_i$ tienen varianza dada por, respectivamente:

$$V\left(\hat{\beta}_0\right) = V\left(\sum_{i=1}^{n} m_i Y_i\right) = \sum_{i=1}^{n} m_i^2 \sigma^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2.$$

Note que $\sum_{i=1}^{n} m_i^2 = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x}\, c_i\right)^2 = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}\,(x_i - \bar{x})}{S_{xx}}\right)^2$
$= \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}$.

$$V\left(\hat{\beta}_1\right) = V\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 \sigma^2 = \frac{\sigma^2}{S_{xx}}.$$

Note que $\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{S_{xx}}$.

$$V\left(\hat{Y}_i\right) = V\left(\sum_{j=1}^{n} h_{ij} Y_j\right) = \sum_{j=1}^{n} h_{ij}^2 \sigma^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right].$$

Note que $\sum_{j=1}^{n} h_{ij}^2 = \sum_{j=1}^{n} \left[\frac{1}{n} + (x_i - \bar{x})\, c_j\right]^2 = \left[\frac{1}{n} + \frac{1}{S_{xx}}(x_i - \bar{x})^2\right]$.

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

4. La covarianza entre los estimadores de los parámetros es:

$$
\begin{aligned}
COV\left(\hat{\beta}_0, \hat{\beta}_1\right) &= COV\left(\sum_{i=1}^{n} m_i Y_i, \sum_{i=1}^{n} c_i Y_i\right) \\
&= \sum_{i=1}^{n} m_i c_i \, COV\left(Y_i, Y_i\right) + \sum_{i=1}^{n} \sum_{j \neq i} m_i c_j COV\left(Y_i, Y_j\right) \\
&= \sum_{i=1}^{n} m_i c_i V\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} m_i c_i = -\frac{\bar{x}}{S_{xx}} \sigma^2.
\end{aligned}
$$

Note que $\sum_{i=1}^{n} m_i c_i = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} c_i\right) c_i = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right) - \bar{x} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 = -\frac{\bar{x}}{S_{xx}^2} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 = -\frac{\bar{x}}{S_{xx}}$.

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

5. La covarianza entre la variable respuesta y su correspondiente estimador, en un valor dado $x_i$, es:

$$\begin{aligned} COV\left(Y_i,\ \hat{Y}_i\right) &= COV\left(Y_i,\ \sum_{j=1}^{n} h_{ij} Y_j\right) \\ &= h_{ii}\, COV\left(Y_i,\ Y_i\right) + \sum_{j \neq i}^{n} h_{ij}\, COV\left(Y_i,\ Y_j\right) \\ &= h_{ii}\sigma^2 = \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \sigma^2. \end{aligned}$$

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

6. La covarianza entre la variable respuesta ajustada en $x_i$ y la ajustada en $x_k$, con $i, k \in \{1, 2, \ldots, n\}$, $i \neq k$, es,

$$
\begin{aligned}
COV\left(\hat{Y}_i, \hat{Y}_k\right) &= COV\left(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\beta}_0 + \hat{\beta}_1 x_k\right) \\
&= COV\left(\hat{\beta}_0, \hat{\beta}_0\right) + x_i\, COV\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
&\quad + x_k\, COV\left(\hat{\beta}_0, \hat{\beta}_1\right) + x_i x_k\, COV\left(\hat{\beta}_1, \hat{\beta}_1\right) \\
&= V\left(\hat{\beta}_0\right) + (x_i + x_k)\, COV\left(\hat{\beta}_0, \hat{\beta}_1\right) + x_i x_k\, V\left(\hat{\beta}_1\right) \\
&= \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2 - \frac{\bar{x}\,(x_i + x_k)}{S_{xx}}\sigma^2 + \frac{x_i x_k}{S_{xx}}\sigma^2 \\
&= \left[\frac{1}{n} + \frac{(x_i - \bar{x})\,(x_k - \bar{x})}{S_{xx}}\right]\sigma^2.
\end{aligned}
$$

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

7. La suma de los residuales del modelo de regresión con intercepto es siempre cero:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0.$$

8. La suma de los valores observados $y_i$ es igual a la suma de los valores ajustados $\hat{y}_i$:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$$

9. La recta de regresión de mínimos cuadrados siempre pasa por el centroide de los datos $(\bar{x}, \bar{y})$.

# Propiedades de los estimadores de mínimos cuadrados bajo el modelo normal y el modelo ajustado de regresión

10. La suma de los residuales ponderados por el correspondiente valor de la variable predictora es cero:

$$\sum_{i=1}^{n} x_i e_i = 0.$$

11. La suma de los residuales ponderados por el correspondiente valor ajustado es siempre igual a cero:

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0.$$

# Contenido

# Intervalos de confianza para los parámetros de regresión

Bajo los supuestos del modelo de regresión, se cumple que:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{V\left(\hat{\beta}_0\right)}}} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \sim t_{n-2}. \quad (9)$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{V\left(\hat{\beta}_1\right)}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}. \quad (11)$$

Por tanto un intervalo de confianza del $(1-\alpha)\,\%$ para $\beta_0$ es:

$$\hat{\beta}_0 \pm t_{\alpha/2,\,n-2} \times \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}. \quad (10)$$

Por tanto un intervalo de confianza del $(1-\alpha)\,\%$ para $\beta_1$ es:

$$\hat{\beta}_1 \pm t_{\alpha/2,\,n-2} \times \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}. \quad (12)$$

**Nota:**

$t_{n-2}$ es la variables aleatoria $t$-Student con $n-2$ grados de libertad, en tanto que $t_{\alpha/2,\,n-2}$ es un percentil de la distribución $t$-Student con $n-2$ grados de libertad, tal que, $P\left(t_{n-2} > t_{\alpha/2,\,n-2}\right) = \alpha/2$.

# Pruebas de hipótesis para los parámetros de regresión

Para probar si $\beta_0$ es significativamente distinto de cero se plantea el test:

$$H_0 : \beta_0 = 0 \quad \text{vs.} \quad H_1 : \beta_0 \neq 0. \quad (13)$$

El estadístico de prueba está dado por:

$$\begin{aligned} T &= \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \\ &= \frac{\hat{\beta}_0}{\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \sim t_{n-2}, \end{aligned} \quad (14)$$

Se define $T_0$ como el valor observado de $T$, luego, se rechaza $H_0$ si $|T_0| > t_{\alpha/2,\, n-2}$ o con valor P si $P\left(|t_{n-2}| > |T_0|\right)$ es pequeña.

Para probar si $\beta_1$ es significativamente distinto de cero se plantea el test:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0. \quad (15)$$

El estadístico de prueba está dado por:

$$\begin{aligned} T &= \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2} \\ &= \frac{\hat{\beta}_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}, \end{aligned} \quad (16)$$

Se define $T_0$ como el valor observado de $T$, luego, se rechaza $H_0$ si $|T_0| > t_{\alpha/2,\, n-2}$ o con valor P si $P\left(|t_{n-2}| > |T_0|\right)$ es pequeña.

# Consideración sobre el test de significancia de la pendiente

**Nota sobre $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ (los cuales se relacionan también con la significancia de la regresión):**

Si la pendiente de la recta de regresión es significativa (es decir, cuando se rechaza $H_0$), entonces el modelo de regresión lineal simple también lo es, es decir, la variabilidad en la variable respuesta explicada por la regresión en $X$ es significativa respecto a la variabilidad total observada. Mientras que, el no rechazar $H_0$ implica que no hay relación lineal entre $X$ e $Y$. Note que eso puede implicar que $X$ tiene muy poco valor para explicar la variación de $Y$ y que el mejor estimador para cualquier $X$ es $\hat{Y} = \bar{Y}$, o que la verdadera relación entre $X$ e $Y$ no es lineal.

# Contenido

# Referencias


Montgomery, D. C., Peck, E. A., y Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley, New Jersey, quinta edición.

Álvarez, N. G. (2022). Notas de Clase Análisis de Regresión - 3006918, Capítulo 2: Regresión Lineal Simple. Notas no publicadas.

Álvarez, N. G. y Gómez, C. M. L. (2018). Notas de Clase - Estadística II (3006918): Análisis de Regresión Lineal e Introducción al Muestreo.