Facultad de Ciencias
UNAM

Estadística II

Tarea 1B

Regresión Lineal Simple

Enríquez Hernández Leobardo
Huitrón Zambrano Victor Manuel
Suárez López David

12 de abril de 2024

Índice

1. Regresión a través del origen.

$$y_i = \beta x_i + \xi_i \quad i = 1 \dots n$$

donde $\xi_1, \dots, \xi_n$ son v.a.i. talque $\xi_i \sim N\left(0, \frac{\sigma^2}{w_i}\right)$

$$\forall i = 1 \dots n$$

Suponiendo σ^2 conocida y $\omega_i = \frac{1}{x_i^2} \quad i = 1, \dots, n$ I) Como las ξ_i son normales, entonces $y_i \sim N\left(\beta x_i, x_i^2 \sigma^2\right)$ y son independientes, entonces la funcion de verosimilitud nos queda:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} x_i \sigma} e^{-\frac{(y_i - \beta x_i)^2}{2x_i^2 \sigma^2}}$$

es decir:

$$\frac{1}{(2\pi)^{n/2} x_i^n \sigma^n} e^{\sum_{i=1}^{n} \frac{-(y_i - \beta x_i)^2}{2x_i^2 \sigma^2}}$$

Aplicando logaritmo:

$$\ln(1) - \frac{n}{2}\ln(2\pi) - n\ln(x_i) - n\ln(\sigma) + \sum_{i=1}^{n} \frac{-(y_i - \beta x_i)^2}{2x_i^2 \sigma^2}$$

derivando e igualando a cero obtenemos:

$$\frac{d}{d\beta}\ln(f) = -\sum_{i=1}^{n} \frac{(y_i - \beta x_i)}{x_i^2 \sigma^2}(-x_i) = 0$$

$$\rightarrow \sum_{i=1}^{n} \frac{(y_i - \beta x_i)}{x_i \sigma^2} = 0 \rightarrow \sum_{i=1}^{n} \frac{y_i}{x_i \sigma^2} - \sum_{i=1}^{n} \frac{\beta}{\sigma^2} = 0$$

Asi

$$\sum_{i=1}^{n} \frac{y_i}{x_i \sigma^2} = \frac{n\beta}{\sigma^2} \rightarrow \hat{\beta} = \sum_{i=1}^{n} \frac{y_i}{x_i n}$$

II)

$$\begin{aligned}
&\mathrm{Var}(\hat{\beta}) \\
&= \mathrm{Var}\left(\sum_{i=1}^{n} \frac{y_i}{x_i n}\right) = \frac{1}{n^2} \cdot \mathrm{Var}\left(\sum_{i=1}^{n} \frac{y_i}{x_i}\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\left(\frac{y_i}{x_i}\right) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{{x_i}^2} \mathrm{Var}\left(y_i\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \frac{1^2}{x_i^2}\left(x_i^2 \sigma^2\right) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n} \\
&\therefore \mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{n}
\end{aligned}$$

III) Como $\hat{\beta} = \sum_{i=1}^{n} \frac{y_i}{x_i n}$, Sea $c_i = \frac{1}{x_i n}$ entonces

$$\hat{\beta} = \sum_{i=1}^{n} c_i y_i \quad \therefore \text{ es estimador lineal}$$

Ademas

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}\left(\sum_{i=1}^{n} \frac{y_i}{x_i n}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\frac{y_i}{x_i}\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i}\mathbb{E}\left(y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \beta = \beta$$

De esta forma $\hat{\beta}$ es estimador lineal y ademas es insesgado, por el teorema Gauss - Markoy $\hat{\beta}$ es el UMVUE

2. Regresión lineal simple.

Considere el modelo de regresión $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, donde $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$ y $Cov(\epsilon_i, \epsilon_j) = 0$, $\forall i \neq j$; $i, j = 1, ..., n$.

Calcular $V(e_i)$, donde $e_i = y_i - \hat{y}_i$ y $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, con $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de los parámetros del modelo.

Hint: Se puede usar que $V(A - B) = V(A) + V(B) - 2Cov(A, B)$ y que $\hat{y}_i$ se puede escribir como una combinación lineal de las $y_{i's}$.

SOLUCIÓN

Como $V(y_i) = V(\beta_0 + \beta_1 x_i + \epsilon_i) = V(\epsilon_i) = \sigma^2$ por ser β_0, β_1 y x_i constantes.

Como $V(\hat{y}_i) = V(\hat{\beta}_0 + \hat{\beta}_1 x_i) = V(\beta_0) + V(\beta_1 x_i) + 2Cov(\beta_0, \beta_1 x_i) = V(\hat{\beta}_0) + x_i^2 V(\hat{\beta}_1) + 2x_i Cov(\hat{\beta}_0, \hat{\beta}_1) = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{SSx}) + x_i^2(\frac{\sigma^2}{SSx}) + 2x_i(\frac{-\bar{X}\sigma^2}{SSx}) = \sigma^2(\frac{SSx + n\bar{X}^2}{nSSx} + \frac{x_i^2}{SSx} - \frac{2x_i\bar{X}}{SSx}) = \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{X})^2}{SSx})$, con $SS_x = \sum_{i=1}^n (x_i - \bar{X})^2$.

Como $Cov(y_i, \hat{y}_i) = Cov(y_i, \bar{Y} + \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{X}) = Cov(y_i, \bar{Y}) + Cov(y_i, \hat{\beta}_1 x_i) + Cov(y_i, -\hat{\beta}_1 \bar{X}) = Cov(y_i, \hat{\beta}_0 + \hat{\beta}_1 \bar{X}) + x_i Cov(y_i, \hat{\beta}_1) - \bar{X} Cov(y_i, \hat{\beta}_1) = Cov(y_i, \hat{\beta}_0) + \bar{X} Cov(y_i, \hat{\beta}_1) + x_i Cov(y_i, \hat{\beta}_1) - \bar{X} Cov(y_i, \hat{\beta}_1) = Cov(y_i, \hat{\beta}_0) + x_i Cov(y_i, \hat{\beta}_1) = (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})\sigma^2 + x_i(\frac{x_i - \bar{X}}{SSx})\sigma^2 = \sigma^2(\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx} + x_i(\frac{x_i - \bar{X}}{SSx})$.

Entonces:

$$V(e_i) = V(y_i - \hat{y}_i) = V(y_i) + V(\hat{y}_i) - 2Cov(y_i, \hat{y}_i) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_i - \bar{X})^2}{SS_x}) - 2\sigma^2(\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx} + \frac{x_i(x_i - \bar{X})}{SSx})$$

$$= \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2(x_i - \bar{X})^2}{SSx} - \frac{-2\sigma^2}{n} + \frac{2\sigma^2\bar{X}(x_i - \bar{X})}{SSx} - \frac{2\sigma^2 x_i(x_i - \bar{X})}{SSx} = \sigma^2 + \frac{\sigma^2}{n} + (\frac{-\sigma^2 x_i^2 - \sigma^2\bar{X}^2 + 2\sigma^2\bar{X}x_i}{SSx})$$

$$= \sigma^2 + \frac{\sigma^2}{n} - \frac{\sigma^2}{SSx}(x_i - \bar{X})^2 = \sigma^2 + \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = \sigma^2$$

También se usaron los siguientes resultados:

$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{Y} - \bar{X}\hat{\beta}_1) + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 x_i - \hat{\beta}_1 \bar{X}$

$V(\hat{\beta}_0) = Cov(\hat{\beta}_0, \hat{\beta}_0) = Cov(\sum_{i=1}^n k_{i_0} y_i, \sum_{j=1}^n k_{j_0} y_j) = \sigma^2 \sum_{i=1}^n k_{i_0}^2 = \sigma^2 \sum_{i=1}^n (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})^2 = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{SSx})$,

$V(\hat{\beta}_1) = Cov(\hat{\beta}_1, \hat{\beta}_1) = Cov(\sum_{i=1}^n k_{i_1} y_i, \sum_{j=1}^n k_{j_1} y_j) = \sigma^2 \sum_{i=1}^n k_{i_1}^2 = \sigma^2 \sum_{i=1}^n (\frac{x_i - \bar{X}}{SSx})^2 = \frac{\sigma^2}{(SSx)^2} \sum_{i=1}^n (x_i - \bar{X})^2 = \frac{\sigma^2}{SSx}$,

$Cov(\hat{\beta}_0, \hat{\beta}_0) = Cov(\sum_{i=1}^n k_{i_0} y_i, \sum_{j=1}^n k_{j_1} y_j) = \sigma^2 \sum_{i=1}^n k_{i_0} k_{i_1} = \sigma^2 \sum_{i=1}^n (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})(\frac{x_i - \bar{X}}{SSx}) = -\frac{\bar{X}\sigma^2}{SSx}$,

$Cov(y_i, \hat{\beta}_0) = Cov(y_i, \sum_{i=1}^n k_{i_0} y_i) = k_{i_0} Cov(y_i, y_i) = k_{i_0} V(y_i) = k_{i_0}\sigma^2 = (\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SSx})\sigma^2$,

$Cov(y_i, \hat{\beta}_1) = Cov(y_i, \sum_{i=1}^n k_{i_1} y_i) = k_{i_1} Cov(y_i, y_i) = k_{i_1} V(y_i) = k_{i_1}\sigma^2 = (\frac{x_i - \bar{X}}{SSx})\sigma^2$,

$\frac{SSx}{(x_i - \bar{X})^2} = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{(x_i - \bar{X})^2} = \sum_{i=1}^n 1 = n$

3. Expresión alternativa para R^2

Considere el coeficiente de correlación muestral o de Pearson para dos variables X y Y :

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)}{\left(\sum_{i=1}^{n} \left(x_i - \bar{X}\right)^2 \sum_{i=1}^{n} \left(y_i - \bar{Y}\right)^2\right)^{1/2}}$$

.

Considere el modelo de regresión

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

.

a. Demuestre que:

$$R^2 = r_{xy}^2$$

b. Demuestre que $t^* = t$, donde t es la estadística usada para contrastar " $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ ":

$$t = \frac{\widehat{\beta}_1}{\sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2}}}.$$

Por otra parte, $t^* = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ es la estadística usada para contrastar " $H_0 : \rho = 0$ vs $H_a : \rho \neq 0$ " cuando (X, Y) sigue una distribución normal bivariada con coeficiente de correlación $\rho = \rho_{xy}$.

SOLUCIÓN a.

Recordemos que: $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SCR}{SCT}$

Y por la expresión (77) de las notas sabemos que:

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

Entonces tenemos:

$$\begin{aligned} R^2 &= \frac{SCR}{SCT} \\ &= \frac{\frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \\ &= \frac{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \end{aligned}$$

Por tanto tenemos que:

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \cdot \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

Ahora, notemos que:

$$\begin{aligned} r_{xy}^2 &= \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \right)^2 \\ &= \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= R^2 \end{aligned}$$

$$\therefore R^2 = r_{xy}^2$$

SOLUCIÓN b.

Primero notemos lo siguiente:

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

Así: $t = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \cdot \sqrt{n-2}}{\left(\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \widehat{y_i})^2\right)^{\frac{1}{2}}}$

Por otro lado tenemos que:

$$
\begin{aligned}
t^* &= \frac{r_{xy}\cdot\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \\
&= \frac{\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\left(\sum_{i=1}^n (x_i-\bar{x})^2\cdot\sum_{i=1}^n (y_i-\bar{y_i})^2\right)^{\frac{1}{2}}}\cdot\sqrt{n-2}}{\sqrt{1-\frac{\sum_{i=1}^n (\widehat{y_i}-\bar{y_i})^2}{\sum_{i=1}^n (y_i-\bar{y_i})^2}}} \\
&= \frac{\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\cdot\sqrt{n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^n (y_i-\bar{y_i})^2}}}{\sqrt{\frac{\sum_{i=1}^n (y_I-\bar{y_i})^2-\sum_{i=1}^n (\widehat{y_i}-\bar{y_i})^2}{\sum_{i=1}^n (y_i-\bar{y_i})^2}}}
\end{aligned}
$$

Ahora, por la expresión (68) de las notas de clase sabemos que:

$$
\sum_{i=1}^n (y_i-\bar{y_i})^2 = \sum_{i=1}^n (\widehat{y_i}-\bar{y_i})^2 + \sum_{i=1}^n (y_i-\widehat{y_i})^2 \Longrightarrow \sum_{i=1}^n (y_i-\bar{y_i})^2 - \sum_{i=1}^n (\widehat{y_i}-\bar{y_i})^2 = \sum_{i=1}^n (y_i-\widehat{y_i})^2
$$

Así:

$$
\begin{aligned}
t^* &= \frac{\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\cdot\sqrt{n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^n (y_i-\bar{y_i})^2}}}{\sqrt{\frac{\sum_{i=1}^n (y_i-\widehat{y_i})^2}{\sum_{i=1}^n (y_i-\bar{y_i})^2}}} \\
&= \frac{\frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\cdot\sqrt{n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^n (y_i-\bar{y_i})^2}}}{\frac{\sqrt{\sum_{i=1}^n (y_i-\widehat{y_i})^2}}{\sqrt{\sum_{i=1}^n (y_i-\bar{y_i})^2}}} \\
&= \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\cdot\sqrt{n-2}}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^n (y_i-\widehat{y_i})^2}} \\
&= \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})\cdot\sqrt{n-2}}{\left(\sum_{i=1}^n (x_i-\bar{x})^2\cdot\sum_{i=1}^n (y_i-\widehat{y_i})^2\right)^{\frac{1}{2}}} \\
&= t
\end{aligned}
$$

$$
\therefore t^* = t
$$

4. Problema Anova. Equivalencia con la estimación considerando dos poblaciones normales.

Sea $X_1, \ldots, X_n$ una m.a de la distribución $N\left(\mu_x, \sigma^2\right)$ y $Y_1, \ldots, Y_m$ una m.a de la distribución $N\left(\mu_y, \sigma^2\right)$ independientes entre si, sea $Z = 1$ si la observacion es de la poblacion con distribucion $N\left(\mu_x, \sigma^2\right)$ y $Z = -1$ si la poblacion es de la poblacion con distribucion $N\left(\mu_y, \sigma^2\right)$

I. Consideramos el modelo de regresion lineal simple:

$$w_j = \beta_0 + \beta_1 z_j + \varepsilon_j$$

con $\varepsilon_1, \ldots, \varepsilon_{n+m}$ variables independientes talque $\varepsilon_j \sim N\left(0, \sigma^2\right) \quad \forall j = 1, \ldots, n+m$

Entonces:

$$\mathbb{E}(w; z = 1) = \mathbb{E}(x; z = 1) = \beta_0 + \beta_1$$

observamos que $\mathbb{E}(x; z = 1) = \mathbb{E}(x) = \mu_x$

Por otro lado tenemos:

$$\mathbb{E}(w; z = -1) = \mathbb{E}(y; z = -1) = \beta_0 - \beta_1$$

observamos que $\mathbb{E}(Y; z = -1) = \mathbb{E}(y) = \mu_y$

Es decir, $\mu_x = \beta_0 + \beta_1$ y $\mu_y = \beta_0 - \beta_1$,

II. Conocemos los estimadores:

$$\hat{\beta}_0 = \bar{w} - \hat{\beta}_1 \bar{z} \text{ y } \hat{\beta}_1 = \frac{\sum_{i=1}^{n+m} (z_i - \bar{z})(w_i - \bar{w})}{\sum_{i=1}^{n+m} (z_i - \bar{z})^2}$$

Con esto podemos obtener: a) $\hat{E}(w; z = 1)$ como $Z = 1$ entonces $w = x$ y $\bar{z} = 1$, asi:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}{\sum_{i=1}^{n} (z_i - \bar{z})^2} \to \hat{\beta}_1 = 0$$

$$\text{y } \hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{z} = \bar{x} - \hat{\beta}_1 = \bar{x}$$
$$\therefore \hat{\mathbb{E}}(W; z = 1) = \hat{\beta}_0 + \hat{\beta}_1 = \bar{x}$$

b) Como $Z = -1$ entonces $w = y$ y $\bar{z} = -1$, asi:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{m} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{m} (z_i - \bar{z})^2} \longrightarrow \hat{\beta}_1 = 0$$
$$y, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 = \bar{y} \quad \therefore \hat{\mathbb{E}}(W; z = -1) = \hat{\beta}_0 - \hat{\beta}_1 = \bar{y}$$

5. Problema ANOVA. Medicamentos.

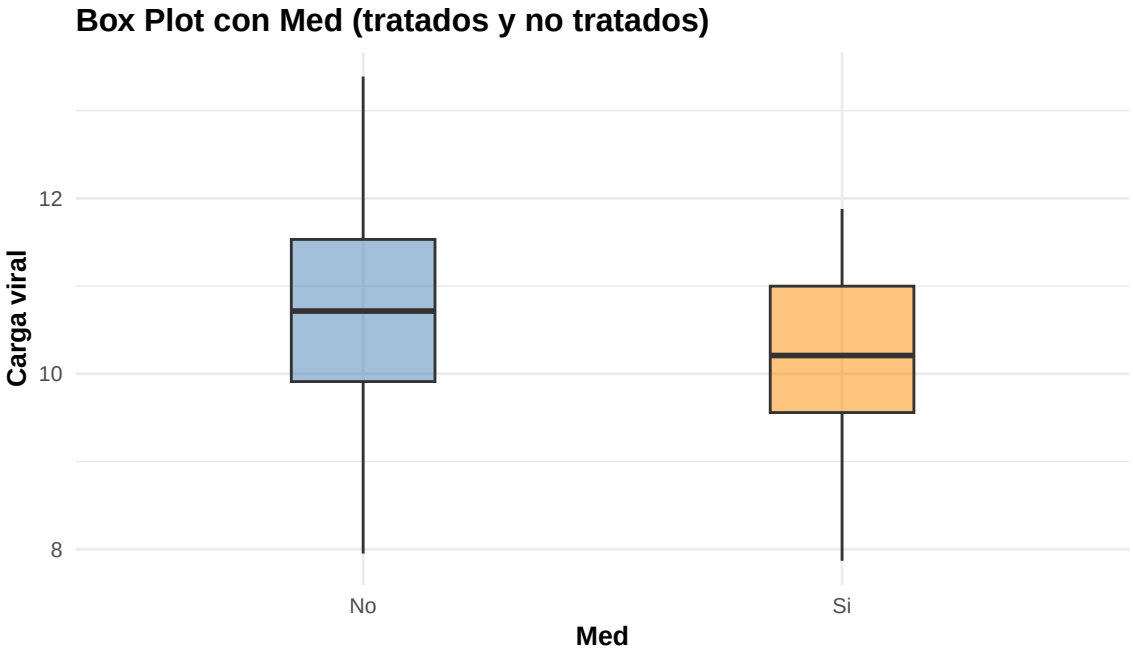
I. Análisis descriptivo y/o visualización de datos.

En la base de datos `Ejercicio5B` se tiene información del índice de carga viral (Y) y si se aplicó o no el medicamento contra Covid (Med) para un grupo de 100 personas. A 50 personas se le aplicó el medicamento.

En el siguiente Cuadro se muestra la estadística descriptiva del dato numérico, que es el índice de carga viral, es posible observar los pacientes presentaron un mínimo de 7.8681037 y un maximo de 13.3899626, con una media de 10.4800674. La desviación estándar de 1.1477774 es pequeña, por lo que parece que los datos no son tan dispersos.

Statistic	N	Mean	St. Dev.	Min	Max
datos$Y	100	10.480	1.148	7.868	13.390

En la siguiente Gráfica de caja y bigotes (brazos), podemos observar que hay una mediana mayor de carga viral para los individuos no tratados, con respecto a los tratados, además de una mayor dispersión de los datos para los no tratados. No se observaron outliers o valores atípicos.



II. Planteamiento y estimación del modelo.

Para ver si la menor carga viral está asociada con la aplicación del medicamento planteamos un modelo de regresión donde la variable dependiente es la carga viral y_i y la variable independiente x_i se puede ver como categórica, donde $x_i = 1$ si el paciente es tratado y $x_i = 0$ si no.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Verificaremos que el modelo presente linealidad, además de homocedasticidad y no autocorrelación en los errores, para poder interpretar los resultados del ajuste del modelo que se muestra en la columna (1) del Cuadro al final de esta sección.

III. Validación de supuestos y pruebas de hipótesis.

Haremos una prueba de hipótesis, para responder a la pregunta de si existe una relación entre la aplicación del medicamento y la disminución de la carga viral. Es decir, planteamos la hipótesis nula $H =: \beta_1 > 0$ y la alternativa $H_a : \beta_1 < 0$. En la prueba `Simultaneous Tests for General Linear Hypotheses` se rechaza H_0 con un nivel de confianza del 95 %, pues el p-value asociado es de 0,0204 y t-value de $-2{,}074$. Sin embargo, para que este resultado tenga validez y para que también las pruebas de hipótesis individuales y global del modelo de la columna (1) del cuadro al final de esta sección tengan validez y podamos interpretar los coeficientes estimados, debemos de hacer las pruebas de cumplimiento de los supuestos del modelo de regresión lineal.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = Y ~ Med, data = datos)
##
## Linear Hypotheses:
##        Estimate Std. Error t value Pr(<t)
## 1 >= 0  -0.4682     0.2258  -2.074 0.0204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Simultaneous Tests for General Linear Hypotheses				
Fit: lm(formula = Y ∼Med, data = datos)				
Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(<t)
1 >= 0	-0.4682	0.2258	-2.074	0.0204*
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			
(Adjusted p values reported – single-step method)				

En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk, Breusch-Pagan y Durbin-Watson para el Modelo 1, que plantean la hipótesis nulas de normalidad, homoscedasticidad y no autocorrelación, respectivamente. Se concluye que el Modelo 1 presenta normalidad de los errores, sin embargo presenta autocorrelación y heteroscedasticidad. Por lo que tendríamos que hacer algunos ajustes al modelo, con algunos tratamientos a las variables, si quisiéramos usarlo para inferencia.

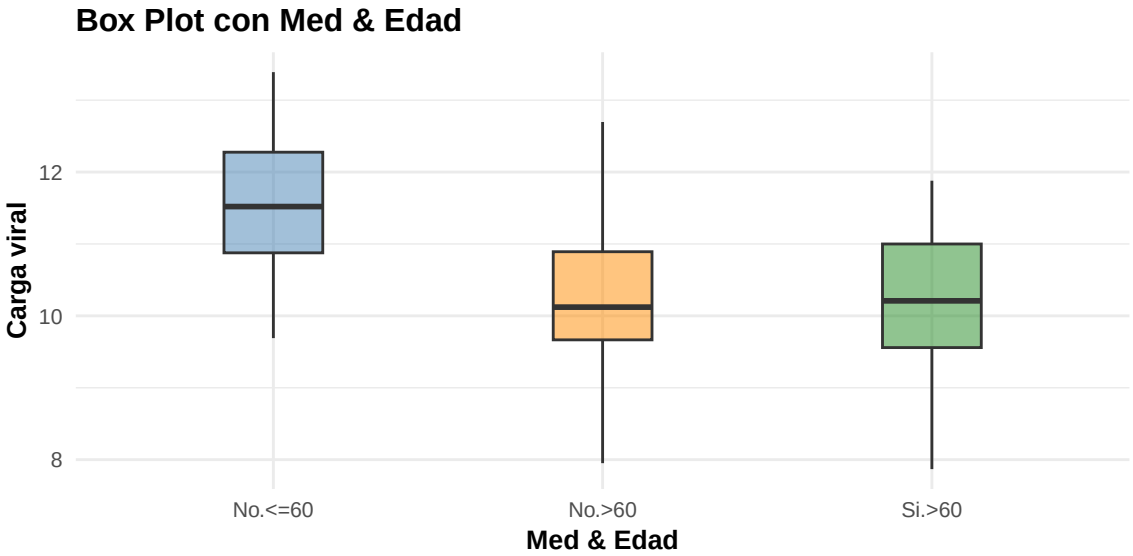
	1
Normality (Shapiro-Wilk)	0.532
Homoscedasticity (Breusch-Pagan)	0.048
Autocorrelation of residuals (Durbin-Watson)	0.021

IV. Consideración de la variable Edad.

Tenemos un total de 100 pacientes, de los cuales 80 tienen más de 60 años y a la mitad de todos los pacientes se le aplicó el medicamento (tratados). En la base de datos tenemos 20 observaciones de personas no tratadas menores de 60 años y no tenemos personas de ese grupo de edades que sean tratadas por el medicamento, lo que podría sesgar los resultados.

A continuación podemos observar en la gráfica de caja y bigotes que a los pacientes menores de 60 años a quienes no se les aplicó la vacuna (no fueron tratadas) tienen una mayor carga viral, sin embargo no disponemos de datos para personas menores a 60 años a quienes se les aplicó el medicamento, para una mejor comparación. Por otro lado, para mayores de 60 años, parece no haber una diferencia clara entre los pacientes a los que se le aplicó la vacuna y a las personas a las que no se le aplicó, pues ambos grupos tienen cargas virales menores pero muy parecidas. Finalmente, no se detectaron datos atípicos (outliers) en los datos. Por

lo tanto, consideramos que los rsultados anteriores no son contundentes, por lo que convendría controlar por la variable de Edad.



V. Planteamiento del nuevo modelo y pruebas de hipótesis.

En este caso planteamos el mismo modelo, pero como no tenemos individuos tratados y no tratados menores de 60 años en la base de datos (solamente tenemos a los no tratados), tomaremos en cuenta a los grupos homogéneos de tratados y no tratados mayores o iguales a 60 años, por lo que nuestra base de datos se reduce de 100 a 80 observaciones.

En la columna (2) del Cuadro a final de esta sección se muestran los resultados correspondientes a la regresión lineal para los pacientes tratados y no tratados mayores a 60 años. La prueba global F muestra que no es posible rechazar la hipótesis nula de que los coeficientes asociados a las variables explicativas son cero, pues el p-value asociado es grande, incluso mayor a 0.1. Al parecer, las conclusiones obtenidas al tomar toda la muestra estaban sesgadas por el grupo de edad, pues cuando quitamos a los no tratados menores de 60 años que tenían una carga viral bastante importante, los resultados cambiaron. Como en la prueba global no es posible rechazar la hipítesis nula, no es posible continuar el análisis con este modelo, y esto está relacionado con el no rechazo de la hipótesis nula de la prueba individual t-student para β_1, es decir, que la variable explicativa no es estadísticamente significativa en el modelo de regresión lineal.

NOTA: Si tomáramos todos los datos, sin considerar los grupos heterogéneos, para el coeficiente estimado asociado a la aplicación o no al medicamento, no se puede rechazar la hipótesis nula de $\beta_1 = 0$ contra la alternativa de $\beta_1 \neq 0$. Esto se muestra en la columna (3) del cuadro al final de esta sección. Además, la variable estadísticamente significativa es la edad, la prueba global F también rechaza H_0 de que todos los coeficientes estimados son cero. Adicionalmente, para este modelo se cumplen los tres supuestos más importantes, como la normalidad, no autocorrelación y homocedasticidad, como se muestra en el Cuadro correspondiente a las pruebas Shapiro-Wilk, Breusch-Pagan y Durbin-Watson. Sin embargo, no podemos continuar por este camino, en primera porque estamos analizando datos heterogéneos, tal vez si tuviéramos a los individuos menores de 60 años con tratamiento, podríamos analizarlo, y en segundo lugar, la respuesta a que si hay un efecto del medicamento parecería ser negativa y solamente dependería de la variable edad.

	1
Normality (Shapiro-Wilk)	0.818
Homoscedasticity (Breusch-Pagan)	0.253
Autocorrelation of residuals (Durbin-Watson)	0.206

	Dependent variable:		
		Y	
	(1)	(2)	(3)
MedSi	-0.468^{**}	0.029	0.029
	(0.226)	(0.245)	(0.242)
Edad>60			-1.244^{***}
			(0.302)
Constant	10.714^{***}	10.217^{***}	11.460^{***}
	(0.160)	(0.194)	(0.234)
Observations	100	80	100
R^2	0.042	0.0002	0.184
Adjusted R^2	0.032	-0.013	0.168
Residual Std. Error	1.129 (df = 98)	1.062 (df = 78)	1.047 (df = 97)
F Statistic	4.299^{**} (df = 1; 98)	0.014 (df = 1; 78)	10.960^{***} (df = 2; 97)
Note:			*p<0.1; **p<0.05; ***p<0.01

6. Uso del modelo de regresión lineal simple.

A continuación se presentan los datos de los pesos de los huevos de 11 nidadas de pingüinos Macaroni, cada nidada tiene dos huevos, uno más pequeño (x) que el otro (y).

x	79	93	100	105	101	96	96	109	70	71	87
y	123	138	154	161	155	149	152	160	117	123	138

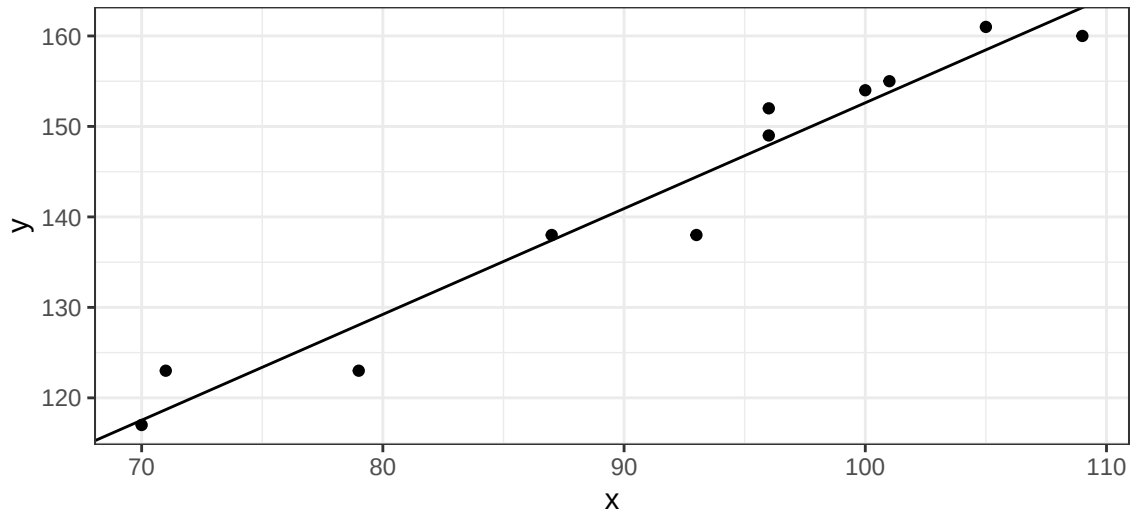
I. Ajuste del modelo de regresión.

Ajustaremos una recta de regresión para estimar el peso promedio del huevo mayor (y) dado el peso del huevo menor (x), es decir, la variable dependiente es el peso del huevo más grande y_i y la variable independiente es el peso del huevo menor.

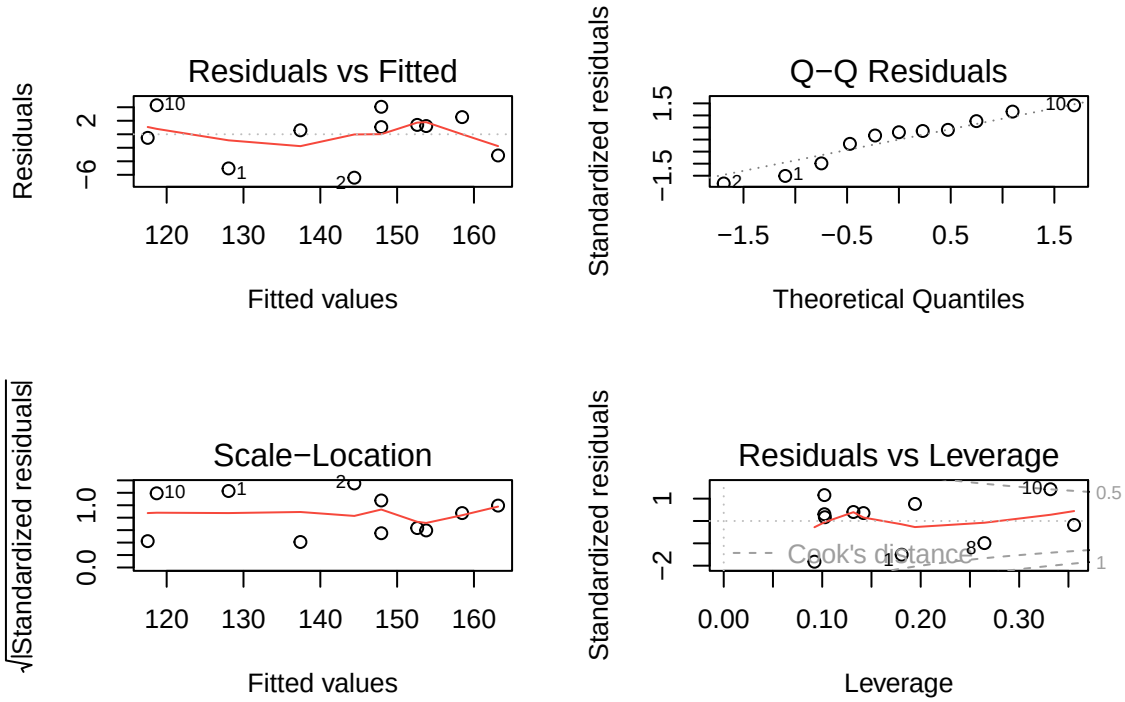
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

En el siguiente Cuadro podemos observar que el p-valor asociado a la prueba F es de menor a 0.05, por lo que se rechaza la hipótesis nula de que los coeficientes asociados a las variables explicativas son cero contra la alternativa de que al menos un coeficiente estimado es distinto de cero. En este caso, como hay una solo variable explicativa, esta prueba coincide con la prueba $t - student$ individual para la β_1= 1.1693983, que también rechaza la hipótesis nula de que $\beta_1 = 0$ contra la alternativa de que $\beta_1 \neq 0$.

	Dependent variable:
	y
x	1.169^{***}
	s.e.(0.088)
	t-value: 13.225
	Pr(>|t|): 3.35e-07
Constant	35.674^{***}
	s.e.(8.171)
	t-value: 4.366
	Pr(>|t|): 0.00181
Observations	11
R^2	0.951
Adjusted R^2	0.946
Residual Std. Error	3.702 (df = 9)
F Statistic	174.895^{***} (df = 1; 9); p-value: 3.351e-07



En las siguientes Gráficas podemos observar las pruebas gráficas para el cumplimiento de los supuestos del modelo de regresión lineal. La Gráfica **Residuals vs Fitted**, se utiliza para comprobar los supuestos de relación lineal, una línea horizontal, sin patrones distintos, es indicación de una relación lineal, lo que es bueno en nuestro caso. La Gráfica **Normal Q-Q**, se utiliza para examinar si los residuos se distribuyen normalmente, es bueno que los puntos residuales sigan la línea recta discontinua, en nuestro caso parece que todo se ajusta bien. La Gráfica **Scale-Location**, se utiliza para comprobar la homogeneidad de la varianza de los residuos (homoscedasticidad), la línea horizontal con puntos igualmente distribuidos es una buena indicación de homocedasticidad, este es el caso en nuestro ejemplo, donde no tenemos un problema de heterocedasticidad. La Gráfica **Residuals vs Leverage**, se utiliza para identificar casos de valores influyentes, es decir, valores extremos que podrían influir en los resultados de la regresión cuando se incluyen o excluyen del análisis, al parecer ningún valor sale de la distancia de Cook.



En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk, Breusch-Pagan y Durbin-Watson, en todos los casos el p-value asociado es mayor a 0.05, por lo que no hay evidencia para rechzar las hipótesis nulas de normalidad, homoscedasticidad y no autocorrelación, respectivamente.

	1
Normality (Shapiro-Wilk)	0.291
Homoscedasticity (Breusch-Pagan)	0.594
Autocorrelation of residuals (Durbin-Watson)	0.107

II. Prueba de hipótesis. Diferencia entre peso mayor y menor como constante.

Ante la sospecha de que en promedio la diferencia entre el peso mayor y el peso menor es constante (es decir, no depende del peso del huevo menor observado), haremos una prueba de hipótesis. En primer lugar notemos que esto implicaría $H_0 : \beta_1 = 1$, contra la alternativa $\beta_1 \neq 1$, a continuación se muestra la aprueba `Simultaneous Tests for General Linear Hypotheses`, donde no se rechaza esta hipótesis nula, pues el p-valor asociado es mayor a 0.05, considerando un nivel del confianza del 95 %.

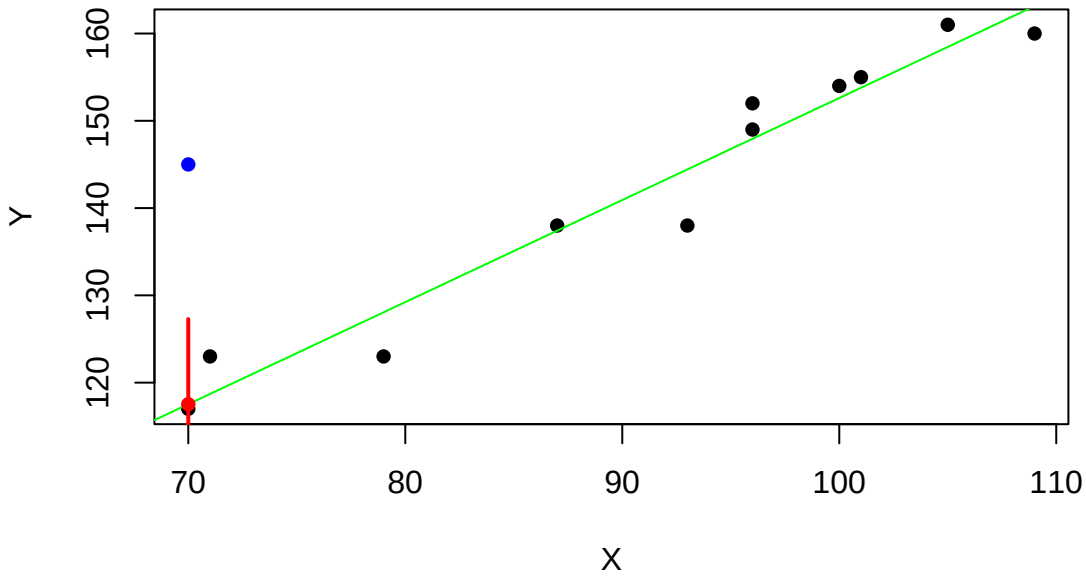
Simultaneous Tests for General Linear Hypotheses Fit: lm(formula = Y ∼x, data = Datos6) Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(<t)
1 == 1	1.16940	0.08842	1.916	0.0877
Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			
(Adjusted p values reported – single-step method)				

III. Nueva observación (nidada). ¿Los huevos provienes de pinguinos Macaroni?

Se observa el peso de los huevos de una nueva nidada, observándose un peso de 70 y 145 gramos. Usando un intervalo de confianza del 95 %, veremos si la nidada de huevos sí proviene de pinguinos Macaroni.

Si tomamos en cuenta la recta de regresión anterior, podemos ver que $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ = 35.6741701 + 1.1693983 *(70)= 117.5320539. Éste valor se ve alejado de los 145 gramos del huevo más grande encontrado.

Podemos predecir que el valor del huevo más grande debería estar entre el valor 107.7811413 y el valor 127.2829666 de acuerdo con un intervalo de predicción. El valor de 145 gramos no cae dentro del intervalo, como también podemos observar en la siguiente Gráfica, por lo que podríamos concluir que la nueva observación no corresponde a los huevos de los pingüinos Macaroni.


Y
120
130
140
150
160
70
80
90
100
110
X

7. Regresión lineal simple con datos de "performance".

Consideraremos los datos en la base `performance.csv` y las variables: y = academic performance of the school (api00) y x = percentage of students receiving free meals (meals). Estos datos corresponden a una muestra aleatoria de 400 escuelas primarias en California, en donde por escuela se realizaron mediciones que tienen que ver con su desempeño en el año 2000.

i) Regresión lineal simple y verificación de supuestos.

Ajustaremos un modelo de regresión lineal simple del desempeño escolar (api00) en función del procentaje de estudiantes que recibieron desayunos gratuitos en las escuelas (meals).

MODELO 1

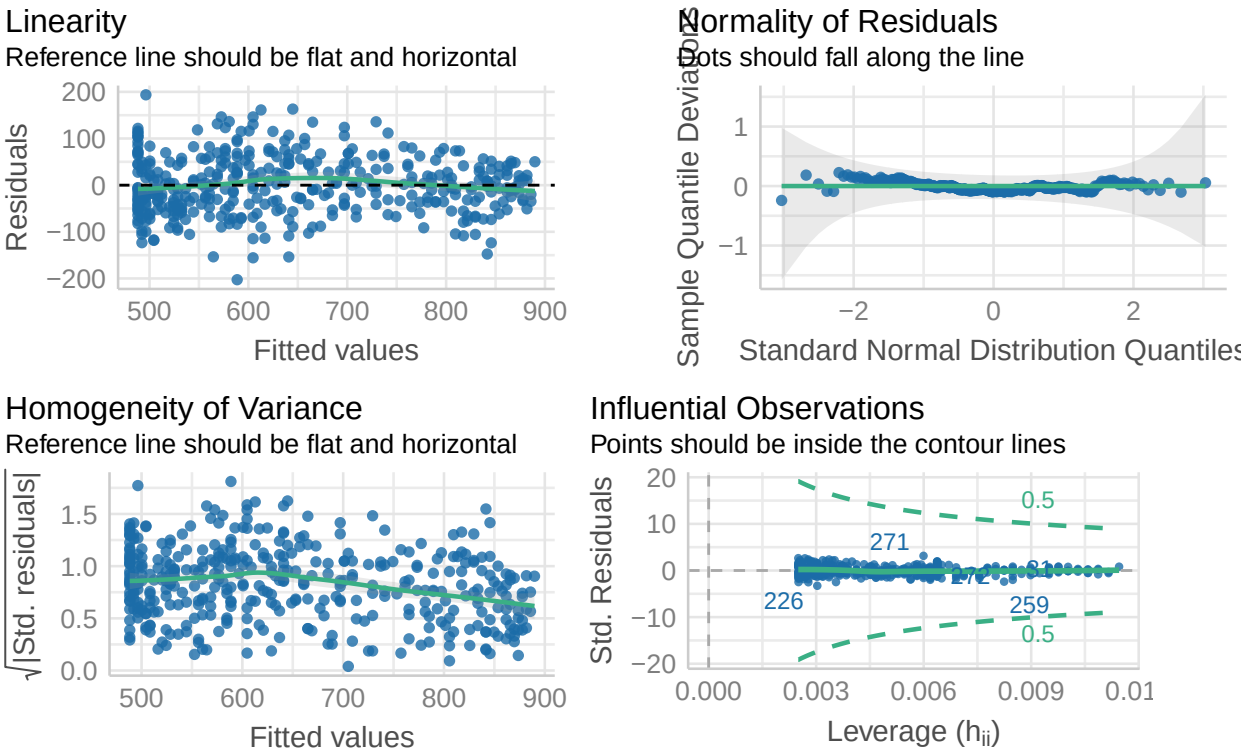
	Dependent variable:
	api00
meals	−4.015***
	s.e.(0.097)
	p-value: <2e-16
Constant	889.783***
	(6.622)
	p-value: <2e-16
Observations	400
R^2	0.811
Adjusted R^2	0.811
Residual Std. Error	61.877 (df = 398)
F Statistic	1,710.691*** (df = 1; 398); p-value: <2.2e-16
Note:	*p<0.1; **p<0.05; ***p<0.01

En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk y Breusch-Pagan y Durbin-Watson, en el primer caso de la normalidad el p-value asociado es mayor a 0.05, por lo que no hay evidencia para rechzar las hipótesis nulas de normalidad, sin embargo hay problemas de heterocedasticidad. Como la muestra se generó aleatoriamente, podemos asumir que no tenemos problemas de autocorrelación de los errores.

	1
Normality (Shapiro-Wilk)	0.618
Homoscedasticity (Breusch-Pagan)	0.002

La normalidad de los errores se confirma con la prueba `Anderson-Darling normality test` con la función `ad.test` que muestra un p-value de 0.276497297460581. La heterocedasticidad se confirma con la prueba `Non-constant Variance Score Test` con la función `ncvTest`, que muestra un p-value de 0.00171467461271069. Finalmente, con la función de `residualPlots` obtenemos para la prueba de `Tukey test` un p value de 0,01174, por lo que se rechaza la hipótesis nula de linealidad.

También podemos observar de forma gráfica estos resultados. Observemos la gráfica de **Fitted values** contra **Residuals**, parece haber un problema de linealidad. En la gráfica **Standard Normal distribution Quantiles** contra **Sample Quantile Deviations** tenemos que la normalidad sí se preserva. En la gráfica de **Fitted Values** contra $\sqrt{|Std.Residuals|}$ parece no haber homogeneidad de la varianza. Y finalmente, en la gráfica de **Leverage(hii)** contra **Std. Residuals** parece no haber valores atípicos influyentes. Entonces podemos concluir que nuestro modelo no cumple con dos supuestos importantes, la linealidad y la homocedasticidad.



ii) Ajuste de un mejor modelo que cumple los supuestos.

Presentamos pruebas para ver qué transformación es adecuada para la variable dependiente e independiente.

```
## Estimated transformation parameter
##       Y1
## 1.588293
```

```
##  MLE of lambda Score Statistic (t) Pr(>|t|)
##        0.93552              0.7774   0.4374
##
## iterations =  2
```

El resultado de la prueba `Estimated transformation parameter` con la función `powerTransform` para transformación de tipo BoxCox, para conocer el exponente λ de la variable dependiente, muestra que el valor es de c(Y1 = 1.58829320628201). Esto sugiere elevar a un exponente de 1,6 a la variable dependiente, por simplicidad en la interpretación consideraremos un exponente de 2. Por otra parte, la prueba con la función `BoxTidwell` para la tranformación de la variable independiente (modificada al sumarle +1 y tomando en cuenta la variable dependiente al cuadrado) muestra un valor λ de 0,93552 con un p-value asociado de 0,4374, lo que implica que la hipótesis nula de que $\lambda = 1$ no se rechaza, i.e., no hay evidencia suficiente para rechazar la linealidad de la variable independiente. Entonces ajustamos el MODELO 2.

En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk y Breusch-Pagan, el p-value asociado es mayor a 0.05 para ambos casos, por lo que no hay evidencia para rechzar las hipótesis nulas de normalidad y homocedasticidad. Como se mencionó anteriormente, la muestra se generó aleatoriamente, por lo que podemos asumir que no tenemos problemas de autocorrelación de los errores.

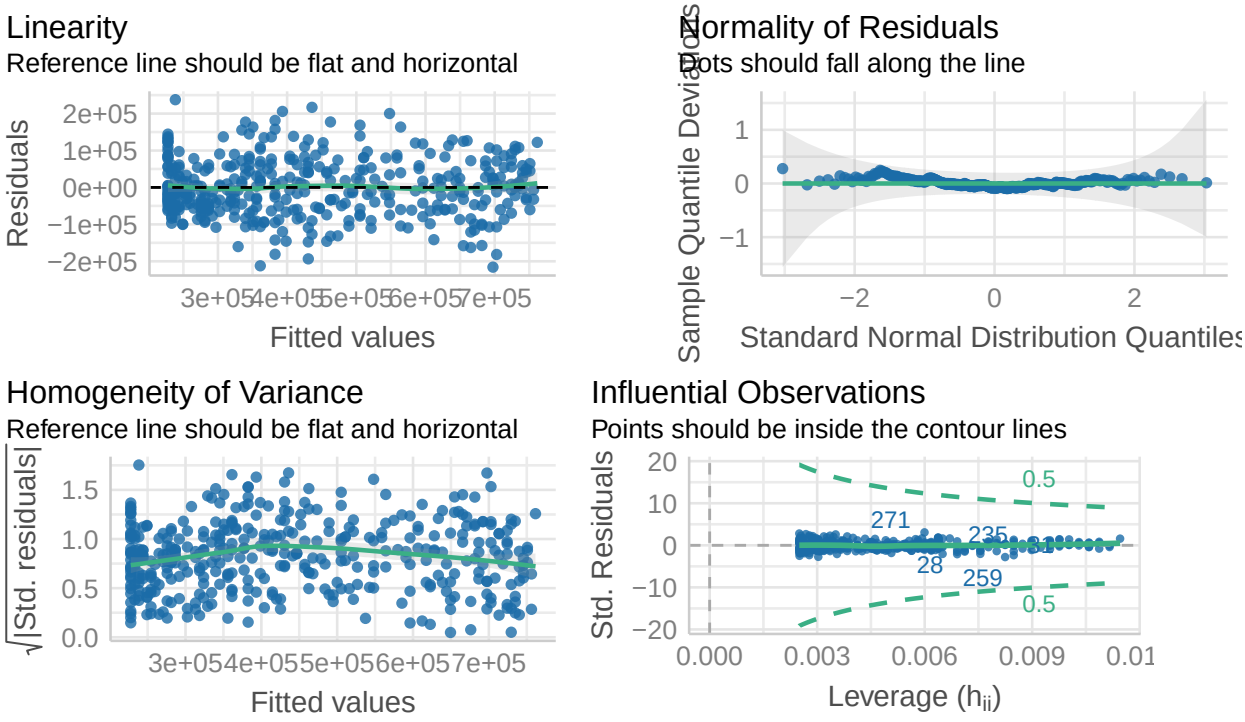
	1
Normality (Shapiro-Wilk)	0.384
Homoscedasticity (Breusch-Pagan)	0.419

MODELO 2

	Dependent variable:
	I(api00^2)
meals	−5,337.734***
	s.e.(121.696)
	p-value: <2e-16
Constant	761,544.500***
	s.e. (8,301.866)
	p-value: <2e-16
Observations	400
R^2	0.829
Adjusted R^2	0.828
Residual Std. Error	77,573.340 (df = 398)
F Statistic	1,923.808*** (df = 1; 398); p-value: <2.2e-16
Note:	*p<0.1; **p<0.05; ***p<0.01

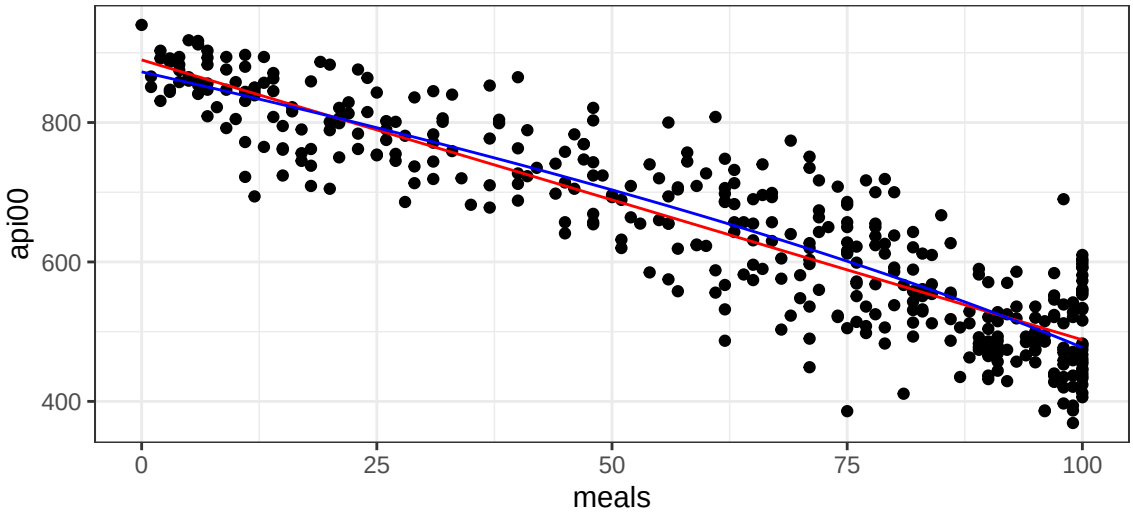
La normalidad de los errores se confirma con la prueba `Anderson-Darling normality test` con la función `ad.test` que muestra un p-value de 0.129178020851007. La homocedasticidad se confirma con la prueba `Non-constant Variance Score Test` con la función `ncvTest`, que muestra un p-value de 0.418222362546048. Finalmente, con la función de `residualPlots` obtenemos para la prueba de `Tukey test` un p value de 0,5969, por lo que no se rechaza la hipótesis nula de linealidad.

También podemos confirmar de forma gráfica estos resultados.



iii)Gráfica de datos originales y las curvas ajustadas de ambos modelos.

A continuación se muestra la Gráfica de los datos originales y las curvas ajustadas tanto para el primero modelo sin tratamiento de las variables (recta roja) y la curva ajustada del segundo modelo con la variable dependiente cuadrática (curva azul).



iv) Interpretación de la prueba ANOVA y la R^2.

En el Modelo 2, se tiene un R^2 de 0.82, el cual es el coeficiente de determinación que en este caso se interpreta como que el 82 % de la variabilidad del rendimiento acedémico en la escuela `api00` se explica por el modelo que incluye la variable del porcentaje de estudiantes que reciben desayuno en la esceula `meal`. Por otra parte, la prueba F asociada a la tabla ANOVA, contrasta en este caso de la regresión lineal simple las hipótesis nula $H_0 : \beta_1 = 0$ contra la alternativa $H_a : \beta_1 \neq 0$. Como el p-value asociado es menor a $2e-16$ se rechaza H_0 con una significancia estadística del 5 %, podemos concluir que la inclusión de la variable explicativa `meal` ayuda a modelar $E(api00; meal)$. Es decir, el rendimiento acedémico en la escuela `api00` se relaciona linealmente con la variable del porcentaje de estudiantes que reciben desayuno en la esceula `meal`.

v) Prueba de hipótesis de investigación.

Para verificar el argumento de que "A mayor porcentaje de comidas gratis en la escuela es menor el desempeño de la escuela", plantearemos una prueba de hipótesis. Planteamos la hipótesis nula $H_0 : \beta_1 \geq 0$ contra la alternativa $H_a : \beta_1 < 0$, donde β_1 es el parámetro estimado asociado a la variable independiente `meal`. A continuación se muestra la aprueba `Simultaneous Tests for General Linear Hypotheses`, donde se rechaza esta hipótesis nula, pues el p-valor asociado es menor a 0.05, con un nivel del confianza de 95 %.

Simultaneous Tests for General Linear Hypotheses Fit:lm(formula = I(api00^2) ~ meals, data = datos7) Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(<t)
1 >= 0	-5337.7	121.7	-43.86	<2e-16 ***
Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
(Adjusted p values reported – single-step method)				