Modelos lineales

Hortensia J. Reyes Cervantes
colab. Alonso Nahir Ramírez

Septiembre 2023

1. Regresión lineal simple

Un modelo de regresión lineal simple es aquel que tiene un único regresor x que tiene una relación con la variable de respuesta y en una línea recta. Este modelo lineal simple es

$$y = \mu_{y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

donde el intercepto β_0 y la pendiente β_1 son constantes desconocidas y ε es el error aleatorio. A partir de una muestra $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ de las variables x e y, se trata de obtener la ecuación 1. $\mu_{y|x}$ es el componente determinista y ε es el componente aleatorio. Así, esta relación no es determinista, por lo cual encontraremos a la recta que mejor represente la tendencia de los datos encontrando los estimadores.

1.1. Suposiciones del modelo de regresión lineal simple

Es posible hacer una estimación de las constantes en 1 si los siguientes **supuestos** son ciertos:

1. y es una variable aleatoria con función de densidad que depende de x.

$$E(y) = \mu_{y|x} \ , \ V(y) = \sigma^2_{y|x}$$

2. Modelo de línea recta $E(y|x) = \mu_{y|x} = \beta_0 + \beta_1 x$

3. Homogeneidad de varianzas:

$$\sigma^2_{y|x_1} = \sigma^2_{y|x_2} = ... = \sigma^2_{y|x_n}$$

4. Independencia entre los valores de la variable dependiente y. Los valores de y deberán ser estadísticamente independientes.
   Ejemplos:
   (No independencia) Se registró dos veces el peso de un individuo en un tiempo menor de una hora.
   (No independencia) Grado de agresividad de dos hermanos que viven en el mismo ambiente familiar.

5. Normalidad en las observaciones de x (variable manipulable), por lo cual la variable aleatoria no observable ε es normal.

$$\varepsilon_i \sim N(0, \sigma^2) \Rightarrow V(\varepsilon_i) = \sigma^2 = E(\varepsilon_i) \Rightarrow cov(\varepsilon_i, \varepsilon_j) = 0 \ si \ i \neq j \Rightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

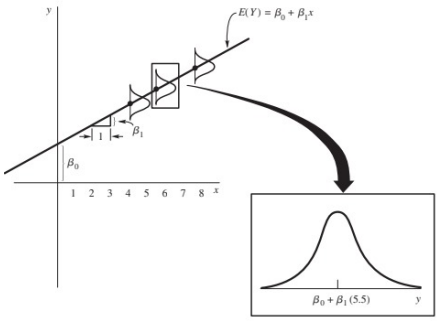


Figura 1: Gráfica del modelo probabilístico $y = \beta_0 + \beta_1 x + \varepsilon$

Si X_i es una variable observada, entonces y_i es una variable que obtiene la aleatoriedad por el error aleatorio ε_i. Es sencillo ver que

$$E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i) = \beta_0 + \beta_1 x_i$$

y

$$V(y_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V(\beta_0 + \beta_1 x_i) + V(\varepsilon_i) = \sigma^2$$

1.2. Estimación de la ecuación de regresión por mínimos cuadrados

Podemos dibujar un diagrama de puntos para ver la tendencia de los datos, aunque no tiene validez estadística por ser subjetivo e impreciso.
El método de mínimos cuadrados, usado por Gauss (1777-1855), produce estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan la suma de cuadrados de las distancias entre los valores observados y_i y los valores estimados $\hat{y}_i$. El modelo muestral de regresión se puede explicar por la ecuación

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, ..., n \tag{2}$$

Donde nos interesa minimizar la suma positiva de los errores $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$. Sea $f(\beta_0, \beta_1)$ una función sobre los parámetros

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \tag{3}$$

Por simplicidad, vamos a denotar $\sum_{i=1}^{n} = \sum$. Los estimadores por mínimos cuadrados que denotados por $\hat{\beta}_0$ y $\hat{\beta}_1$ deben satisfacer

$$\begin{aligned} \frac{\partial f}{\partial \beta_0}|_{\hat{\beta}_0, \hat{\beta}_1} &= 2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0 \\ &\Rightarrow n y_i - n\hat{\beta}_0 - \sum \hat{\beta}_1 x_i = 0 \end{aligned} \tag{4}$$

y

$$\begin{aligned} \frac{\partial f}{\partial \beta_1}|_{\hat{\beta}_0, \hat{\beta}_1} &= 2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \\ &\Rightarrow \sum x_i y_i - \hat{\beta}_0 \sum x_i - \sum \hat{\beta}_1 x_i^2 = 0 \end{aligned} \tag{5}$$

De aquí, se obtiene el siguiente sistema llamado **sistema de ecuaciones normales**:

$$\sum y_i = n\hat{\beta}_0 + \sum \hat{\beta}_1 x_i$$
$$\sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2$$

Dividiendo (4) entre n se obtiene

$$\frac{\sum y_i}{n} - \hat{\beta}_0 - \frac{\sum \hat{\beta}_1 x_i}{n} = 0 \Rightarrow \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Sustituyendo esto en (5),

$$\sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i = \sum \hat{\beta}_1 x_i^2$$
$$\sum x_i y_i - \bar{y} \sum x_i + \hat{\beta}_1 \bar{x} \sum x_i = \sum \hat{\beta}_1 x_i^2$$
$$\sum x_i y_i - \bar{y} \sum x_i = \sum \hat{\beta}_1 x_i^2 - \hat{\beta}_1 \bar{x} \sum x_i$$

Desarrollando el lado derecho

$$\begin{aligned}
\sum \hat{\beta}_1 x_i^2 - \hat{\beta}_1 \bar{x} \sum x_i &= \hat{\beta}_1 (\sum x_i^2 - \bar{x} \sum x_i (\frac{n}{n})) \\
&= \hat{\beta}_1 (\sum x_i^2 - n\bar{x}^2) \\
&= \hat{\beta}_1 (\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2) \\
&= \hat{\beta}_1 (\sum x_i^2 - 2\sum x_i \bar{x} + n\bar{x}^2) \\
&= \hat{\beta}_1 \sum (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\
&= \hat{\beta}_1 \sum (x_i - \bar{x})^2
\end{aligned}$$

y el lado izquierdo

$$\begin{aligned}
\sum x_i y_i - \bar{y} \sum x_i &= \sum (x_i y_i - \bar{y} x_i - \bar{x} y_i + \bar{x}\bar{y}) \\
&= \sum (x_i y_i - \bar{y} x_i) \\
&= \sum (x_i y_i - \bar{y} x_i + 0) \\
&= \sum (x_i y_i - \bar{y} x_i) - \bar{x} \sum y_i + n\bar{x}\bar{y} \qquad (*) \\
&= \sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y}) \\
&= \sum (x_i - \bar{x})(y_i - \bar{y})
\end{aligned}$$

Veamos que la expresión en $(*)$ es cero:

$$-\bar{x} \sum y_i + n\bar{x}\bar{y} = \bar{x}(-\sum y_i + \frac{n \sum y_i}{n}) = \bar{x}(-\sum y_i + \sum y_i) = \bar{x}(0) = 0$$

Ahora, igualando el término izquierdo y derecho

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum (x_i - \bar{x})^2$$
$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Por lo tanto, los estimadores para β_0 y β_1 están dados por

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x}_i)^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x}_i)^2} = \frac{\sum y_i x_i - \frac{(\sum y_i)(\sum x_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

y

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

El modelo ajustado de regresión lineal simple en (2) es, entonces,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6}$$

Notación:

$$\bar{x} = \frac{\sum x_i}{n}$$
$$\bar{y} = \frac{\sum y_i}{n}$$
$$S_{xx} = \sum(x_i - \bar{x})^2$$
$$S_{yy} = \sum(y_i - \bar{y})^2$$
$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$$

Ejemplo: Los siguientes datos corresponden a la presión barométrica y la temperatura del punto de ebullición para 17 lugares en los Alpes y en Escocia.

Número de caso	Temp(°F)	Presión(In Hg)	Lpres=100*log(Presión)
1	194.5	20.79	131.79
2	194.3	20.79	131.79
3	197.9	22.40	135.02
4	198.4	22.67	135.55
5	199.4	23.15	136.46
6	199.9	23.35	136.83
7	200.9	23.89	137.82
8	201.1	23.99	138.00
9	201.4	24.02	138.06
10	201.3	24.01	138.04
11	203.6	25.14	140.04
12	204.6	26.57	142.44
13	209.5	28.49	145.47
14	208.6	27.76	144.34
15	210.7	29.04	146.30
16	211.9	29.88	147.54
17	212.2	30.06	147.80

Si nos interesa hacer un modelo de regresión lineal simple sobre las observaciones de temperatura y la transformación logarítmica de la presión podemos realizarlo con los estimadores anteriores. Sea X la temperatura

y Y como Lpres, entonces

$$\bar{x} = 202.95294 \qquad S_{xx} = 530.78235$$
$$\bar{y} = 139.60529 \qquad S_{xy} = 427.79402$$

De esta manera, los estimadores son

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.895 \qquad \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \text{-42.138}$$

Dado que este fenómeno tiene un comportamiento físico determinista, la línea estimada (Fig. 2) tiene un gran ajuste con las observaciones. Los errores se pueden deber, por ejemplo, a los instrumentos de medición. La ecuación es

$$\hat{y} = -42.138 + 0.895 Temp$$

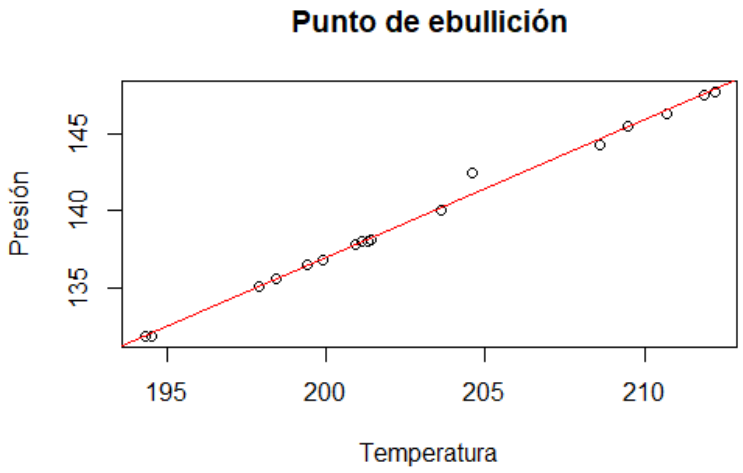


Figura 2: Línea estimada Temp Lpress

Teorema 1. *Los estimadores obtenidos por el método de mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1$ son los mejores estimadores lineales insesgados y de mínima varianza.*

Demostración. Sea $\hat{\hat{\beta}}_1 = \sum c_i y_i$ donde los c_i son constantes

$$\begin{aligned} E(\hat{\hat{\beta}}_1) = \beta_1 \iff E(\hat{\hat{\beta}}_1) &= E(\sum c_i y_i) = E(\sum c_i(\beta_0 + \beta_1 x_i + \varepsilon_i)) \\ &= E(\beta_0 \sum c_i + \beta_1 \sum c_i x_i + \sum c_i \varepsilon_i) \\ &= \beta_0 \sum c_i + \beta_1 \sum c_i x_i + \sum c_i E(\varepsilon_i) \\ &= \beta_0 \sum c_i + \beta_1 \sum c_i x_i \\ &= \beta_1 \iff \sum c_i = 0 \wedge \sum c_i x_i = 1 \\ &\therefore E(\hat{\hat{\beta}}_1) = E(\sum c_i y_i) = \beta_1 \end{aligned}$$

Por otro lado

$$\begin{aligned} V(\hat{\beta}_1) = V(\sum c_i y_i) &= E(\sum c_i y_i - E(\sum c_i y_i))^2 \\ &= E(\sum c_i y_i - \beta_1)^2 \\ &= E(\sum c_i(\beta_0 + \beta_1 x_i + \varepsilon_i) - \beta_1)^2 \\ &= E(\beta_0 \sum c_i + \beta_1 \sum c_i x_i + \sum c_i \varepsilon_i - \beta_1)^2 \\ &= E(\beta_1 + \sum c_i \varepsilon_i - \beta_1)^2 \\ &= E(\sum c_i \varepsilon_i)^2 \\ &= E(\sum c_i \varepsilon_i + \sum \sum_{i \neq j} c_i \varepsilon_i c_j \varepsilon_j) \\ &= \sum c_i^2 E(\varepsilon_i^2) + \sum \sum_{i \neq j} c_i c_j E(\varepsilon_i \varepsilon_j) = \sum c_i^2 \sigma^2 \\ &\therefore V(\hat{\beta}_1) = \sum c_i^2 \sigma^2 = \sigma^2 \sum c_i^2 \end{aligned}$$

Se desea minimizar esta varianza sujeto a $\sum c_i = 0 \wedge \sum c_i x_i - 1 = 0$. Esto es equivalente a minimizar $\sum c_i^2$, pues $\sigma^2 > 0$, con las mismas condiciones.

Utilizando multiplicadores de Lagrange. Sea $\varphi = \sum c_i^2 - 2\lambda \sum c_i - 2\delta(\sum c_i x_i - 1) = \varphi(c_i, \lambda, \delta)$. Entonces

$$\begin{aligned} \frac{\partial \varphi}{\partial c_i} &= 2\sum c_i - 2\lambda - 2\delta x_i = 0 && \Rightarrow && \lambda = -\delta \bar{x} && (*) \\ \frac{\partial \varphi}{\partial \lambda} &= -2\sum c_i = 0 && \Rightarrow && \sum c_i = 0 \\ \frac{\partial \varphi}{\partial \delta} &= -2(\sum c_i x_i - 1) = 0 && \Rightarrow && \sum c_i x_i = 1 \end{aligned}$$

veamos que $(*)$ se cumple

$$\begin{aligned} 2c_i = 2\lambda + 2\delta x_i \Rightarrow c_i &= \lambda + \delta x_i,\ i = \overline{1,n} \qquad \diamond \\ \Rightarrow c_i &= n\lambda + \delta \sum x_i \\ \lambda &= \frac{\sum c_i - \delta \sum x_i}{n} \\ \lambda &= \frac{\sum c_i}{n} - \delta \bar{x} \\ \lambda &= -\delta \bar{x} \end{aligned}$$

Si sustituimos este último resultado en $\diamond$, entonces

$$\begin{aligned} c_i = -\delta\bar{x} + \delta x_i \Rightarrow c_i &= \delta(x_i - \bar{x}) \\ \Rightarrow c_i x_i &= \delta(x_i - \bar{x})x_i \end{aligned}$$

Sumando

$$\begin{aligned} 1 = \sum c_i x_i &= \delta \sum (x_i - \bar{x})x_i \\ &= \delta \sum (x_i - \bar{x})^2 \\ \therefore 1 &= \delta \sum (x_i - \bar{x})^2 \\ \Rightarrow \delta &= \frac{1}{\sum (x_i - \bar{x})^2} \end{aligned}$$

De donde el valor de c_i es

$$c_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

Por lo tanto

$$\hat{\hat{\beta}}_1 = \sum c_i y_i = \sum \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} y_i = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \hat{\beta}_1$$

Se queda como ejercicio para el lector hacer el resto de la demostración para $\hat{\hat{\beta}}_0$. □

Aunque ya hemos dado un sustento estadístico a la estimación de los parámetros por mínimos cuadrados, también es posible obtener los mismos resultados **maximizando la función de verosimilitud**.

$$\begin{aligned} L(y_i, x_i, \beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} exp[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2] \\ &= (2\pi\sigma^2)^{-n/2} exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2] \end{aligned}$$

Llamemos $\ddot{\beta}_0$, $\ddot{\beta}_1$ y $\ddot{\sigma}^2$ a los estimadores por máxima verosimilitud. La máxima verosimilitud de L es equivalente a maximizar lnL. de tal modo que

$$lnL(y_i, x_i, \beta_0, \beta_1, \sigma^2) = \frac{-n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \beta_0 - \beta_1 x_i)^2$$

y los estimadores deben satisfacer

$$\begin{aligned} \frac{\partial ln(L)}{\partial \beta_0} &= \frac{1}{\ddot{\sigma}^2} \sum (y_i - \ddot{\beta}_0 - \ddot{\beta}_1 x_i) = 0 \\ \frac{\partial ln(L)}{\partial \beta_1} &= \frac{1}{\ddot{\sigma}^2} \sum (y_i - \ddot{\beta}_0 - \ddot{\beta}_1 x_i) x_i = 0 \\ \frac{\partial ln(L)}{\partial \sigma 2} &= \frac{-n}{2\ddot{\sigma}^2} + \frac{1}{2\ddot{\sigma}^2} \sum (y_i - \ddot{\beta}_0 - \ddot{\beta}_1 x_i)^2 = 0 \end{aligned}$$

de aquí llegamos a las ecuaciones normales

$$\begin{aligned} \sum y_i - n\ddot{\beta}_0 - \ddot{\beta}_1 \sum x_i &= 0 \\ \sum x_i y_i - \ddot{\beta}_0 \sum x_i - \ddot{\beta}_1 \sum x_i^2 &= 0 \\ \frac{1}{2\ddot{\sigma}^2} \sum (y_i - \ddot{\beta}_0 - \ddot{\beta}_1 x_i)^2 &= \frac{n}{2\ddot{\sigma}^2} \end{aligned}$$

de donde

$$\bar{y} = \ddot{\beta}_0 + \ddot{\beta}_1 \bar{x}$$
$$\sum x_i y_i = \ddot{\beta}_0 \sum x_i + \ddot{\beta}_1 \sum x_i^2$$

Entonces

$$\ddot{\beta}_0 = \bar{y} - \ddot{\beta}_1 \bar{x}$$

y

$$\begin{aligned}\sum y_i x_i &= (\bar{y} - \ddot{\beta}_1 \bar{x}) \sum x_i + \ddot{\beta}_1 \sum x_i^2 \\ &= \bar{y} \sum x_i - \ddot{\beta}_1 \sum x_i \bar{x} + \ddot{\beta}_1 \sum x_i^2 \\ &= \bar{y} \sum x_i + \ddot{\beta}_1 (\sum x_i^2 + \sum x_i \bar{x})\end{aligned}$$

Por lo tanto

$$\ddot{\beta}_1 = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \sum x_i \bar{x}} = \hat{\beta}_1$$

de la última ecuación podemos obtener un estimador de máxima verosimilitud sesgado para σ^2

$$\ddot{\sigma^2} = \frac{\sum (y_i - \ddot{\beta}_0 - \ddot{\beta}_1 x_1)^2}{n} = \frac{\sum e_i^2}{n}$$

1.3. Propiedades de los estimadores

Definamos al residual como la diferencia del valor observado y_i y el valor ajustado correspondiente $\hat{y}_i$. Es decir

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i),\ i = 1, 2, ..., n$$

De esta manera,

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

El lector puede verificar que se cumplen las siguientes identidades:

$$\begin{aligned}&\sum y_i = \sum \hat{y}_i \\ &\sum x_i e_i = 0 \\ &\sum \hat{y}_i e_i = 0\end{aligned} \tag{7}$$

Teorema 2. *Los estimadores cumplen las siguientes propiedades:*

I. $E(\hat{\beta}_1) = \beta_1$

II. $E(\hat{\beta}_0) = \beta_0$

III. $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

IV. $V(\hat{\beta}_0) = \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$

V. $Cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{x}}{S_{xx}}$

Demostración. I. Sea $w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$, entonces

$$\sum w_i = \frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i - n\bar{x}}{\sum (x_i - \bar{x})^2} = \frac{n\bar{x} - n\bar{x}}{\sum (x_i - \bar{x})^2} = 0$$

$$\sum x_i w_i = \frac{\sum x_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i^2 - x_i \bar{x})}{\sum (x_i^2 - 2x_i \bar{x} + \bar{x}^2)} = \frac{\sum x_i^2 - \bar{x} \frac{n \sum x_i}{n}}{\sum x_i^2 - 2\bar{x} \frac{n \sum x_i}{n} + n\bar{x}^2} = \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2} = \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i^2 - n\bar{x}^2} = 1$$

Ahora

$$\begin{aligned} E(\hat{\beta}_1) &= E(\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}) = E(\sum w_i y_i) = \sum w_i E(y_i) = \sum w_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i x_i = \beta_0 (0) + \beta_1 (1) = \beta_1 \end{aligned}$$

II.

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{y} - \bar{x}\hat{\beta}_1) = E(\bar{y}) - \bar{x} E(\hat{\beta}_1) = \frac{\sum E(y)}{n} - \bar{x}\beta_1 \\ &= \frac{\sum (\beta_0 + \beta_1 x_i)}{n} - \bar{x}\beta_1 = \frac{n\beta_0}{n} + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0 \end{aligned}$$

III.

$$\begin{aligned} V(\hat{\beta}_1) &= E[\hat{\beta}_1 - E(\hat{\beta}_1)]^2 = E[\beta_1 + \frac{\sum (x_i - \bar{x}) \varepsilon_i}{\sum (x_i - \bar{x})^2} - \beta_1]^2 = [\frac{1}{\sum (x_i - \bar{x})^2}]^2 E[\sum (x_i - \bar{x}) \varepsilon_i]^2 \\ &= [\frac{1}{\sum (x_i - \bar{x})^2}]^2 E[\sum (x_i - \bar{x})^2 \varepsilon_i^2 + \sum_i \sum_j (x_i - \bar{x}) \varepsilon_i (x_j - \bar{x}) \varepsilon_j] \\ &= [\frac{1}{\sum (x_i - \bar{x})^2}]^2 \sum (x_i - \bar{x})^2 E[\varepsilon_i^2] + \sum_i \sum_j (x_i - \bar{x})(x_j - \bar{x}) E[\varepsilon_i \varepsilon_j] \qquad (*) \\ &= \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

Verifiquemos (*). Como los errores son independientes, entonces

$$\begin{aligned} E(\varepsilon_i \varepsilon_j) &= E(\varepsilon_i) E(\varepsilon_j) = 0 \qquad \forall\, i \neq j \\ V(\varepsilon_i) &= \sigma^2 = E(\varepsilon_i^2) - E(\varepsilon_i)^2 = E(\varepsilon_i^2) \end{aligned}$$

IV. Antes, algunos resultados útiles

$$\begin{aligned} V(\bar{y}) &= V(\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}) = V(\bar{\varepsilon}) = V(\frac{\sum \varepsilon_i}{n}) = \frac{1}{n} V(\varepsilon_i) = \frac{\sigma^2}{n} \\ cov(\bar{y}, \hat{\beta}_1) &= cov(\frac{\sum y_i}{n}, \sum w_i y_i) = V(\sum \frac{w_i}{n} y_i) = (\frac{\sum w_i}{n})^2 V(y_i) = (0)(\sigma^2) = 0 \end{aligned}$$

Ahora es más sencillo calcular la varianza

$$\begin{aligned} V(\hat{\beta}_0) &= V(\bar{y} - \bar{x}\hat{\beta}_1) = V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\bar{x} cov(\bar{y}, \hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \\ &= \sigma^2 (\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}) = \sigma^2 \frac{\sum x_i^2}{\sum (x_i - \bar{x})^2} \end{aligned}$$

v.

$$cov(\hat{\beta}_0, \hat{\beta}_1) = cov(\bar{y} - \bar{x}\hat{\beta}_1, \hat{\beta}_1) = cov(\bar{y}, \hat{\beta}_1) - \bar{x}cov(\hat{\beta}_1, \hat{\beta}_1) = 0 - \bar{x}V(\hat{\beta}_1) = -\bar{x}\frac{\sigma^2}{S_{xx}}$$

Análogo:

$$\begin{aligned}
cov(\hat{\beta}_0, \hat{\beta}_1) &= E[\hat{\beta}_0\hat{\beta}_1] - E(\hat{\beta}_0)E(\hat{\beta}_0) \\
&= E[(\bar{y} - \bar{x}\hat{\beta}_1)\hat{\beta}_1] - \beta_0\beta_1 \\
&= \bar{y}\beta_1 - \bar{x}E[\hat{\beta}_1^{\,2}] - \beta_0\beta_1 \\
&= \bar{y}\beta_1 - \bar{x}(\frac{\sigma^2}{\sum(x_i - \bar{x})^2} + \beta_1^2) - \beta_0\beta_1 \\
&= \bar{y}\beta_1 - \bar{x}\frac{\sigma^2}{S_{xx}} - \bar{x}\beta_1^2 - \beta_0\beta_1 \\
&= (\bar{y} - \bar{x}\beta_1)\beta_1 - \beta_0\beta_1 - \bar{x}\frac{\sigma^2}{S_{xx}} \\
&= \beta_0\beta_1 - \beta_0\beta_1 - \frac{\bar{x}\sigma^2}{S_{xx}} \\
&= -\frac{\bar{x}\sigma^2}{S_{xx}}
\end{aligned}$$

□

Estos resultados son de utilidad para crear intervalos de confianza y pruebas de hipótesis que nos servirán para medir la adecuación del modelo estimado y hacer estimaciones sobre los datos. Si no conocemos la varianza (constante por suposición) de la v.a. ε, entonces se debe estimar. Idealmente esperamos que esta estimación no dependa de la adecuación del modelo y es posible si existen distintas observaciones y para al menos un valor de x. Cuando esto no es posible, podemos obtener una estimación insesgada a partir de los residuales.

Teorema 3. $\hat{\sigma^2}$ *es un estimador insesgado para la varianza poblacional* σ^2*, donde*

$$\hat{\sigma^2} = \frac{\sum e_i^2}{n-2}.$$

Demostración.

$$E[\hat{\sigma^2}] = E[\frac{\sum(y_i - \hat{y_i})^2}{n-2}] = \frac{1}{n-2}E[\sum(y_i - \hat{y_i})^2]$$

Donde

$$\begin{aligned}
E[\sum(y_i-\hat{y}_i)^2] &= E[\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2] \\
&= E[\sum(y_i-\bar{y}+\bar{x}\hat{\beta}_1-\hat{\beta}_1x_i)^2] \\
&= E[\sum[(y_i-\bar{y})-\hat{\beta}_1(x_i-\bar{x})]^2] \\
&= E[\sum(y_i-\bar{y})^2-2\hat{\beta}_1\sum(x_i-\bar{x})(y_i-\bar{y})+{\hat{\beta}_1}^2\sum(x_i-\bar{x})^2] \\
&= E[\sum y_i^2-n\bar{y}^2-\hat{\beta}_1S_{xx}] \qquad (*) \\
&= \sum E(y_i^2)-nE(\bar{y}^2)-S_{xx}E({\hat{\beta}_1}^2) \\
&= \sum[V(y_i)+E^2(y_i)]-n[V(\bar{y})+E^2(\bar{y})]-S_{xx}[V(\hat{\beta}_1)+E^2(\hat{\beta}_1)] \qquad (**) \\
&= n\sigma^2+\sum(\beta_0+\beta_1x_i)^2-n[\frac{\sigma^2}{n}+(\beta_0+\beta_1\bar{x})^2]-S_{xx}[\frac{\sigma^2}{S_{xx}}+\beta_1^2] \\
&= n\sigma^2+\sum\beta_0^2+2\beta_0\beta_1\sum x_i+\beta_1^2\sum x_i^2-\sigma^2-n\beta_0^2-2n\beta_0\beta_1\bar{x}-n\beta_1^2\bar{x}^2-\sigma^2-S_{xx}\beta_1^2 \\
&= (n-2)\sigma^2+n\beta_0^2-n\beta_0^2+2\beta_0\beta_1\sum x_i-2\beta_0\beta_1\sum x_i+\beta_1^2\sum x_i^2-\beta_1^2\frac{(\sum x_i)^2}{n}-S_{xx}\beta_1^2 \\
&= (n-2)\sigma^2+(\sum x_i^2-\frac{1}{n}(\sum x_i)^2)\beta_1^2-S_{xx}\beta_1^2 \\
&= (n-2)\sigma^2+S_{xx}\beta_1^2-S_{xx}\beta_1^2 \\
&= (n-2)\sigma^2
\end{aligned}$$

En $(*)$, como $\sum(x_i-\bar{x})(y_i-\bar{y})=\sum(x_i-\bar{x})^2\hat{\beta}_1$, entonces $-2\hat{\beta}_1\sum(x_i-\bar{x})(y_i-\bar{y})+{\hat{\beta}_1}^2\sum(x_i-\bar{x})^2=-\hat{\beta}_1S_{xx}$.
En $(**)$, para cualquier variable aleatoria U, se cumple $E(U^2)=V(U)+E^2(U)$.

Por lo tanto

$$\frac{1}{n-2}E[\sum(y_i-\hat{y}_i)^2]=\frac{1}{n-2}((n-2)\sigma^2)=\sigma^2$$

□

Una fórmula análoga útil para cálculos es

$$\hat{\sigma}^2=\frac{S_{yy}-\hat{\beta}_1S_{xy}}{n-2}$$

1.4. Pruebas de hipótesis

Hasta ahora hemos usado la esperanza y la varianza de los errores ε, pero no había sido necesario que tuvieran una distribución normal. En esta sección nos interesa crear pruebas de hipótesis sobre los parámetros del modelo y para ello es conveniente conocer la distribución de sus estimadores.

Suponiendo que deseamos probar la hipótesis de que algún beta es igual a una constante. Por ejemplo, para β_1

$$H_0:\ \beta_1=\beta_1^*\ vs\ H_a:\beta_1\neq\beta_1^*$$

Como los errores son normales, independientes e idénticamente distribuidos $\varepsilon_i\sim N(0,\sigma^2)$, las observaciones y_i también lo son $y_i\sim N(\beta_0+\beta_1x_i,\sigma^2)$. Dado que $\hat{\beta}_1$ es una combinación lineal de estas observaciones,

entonces está normalmente distribuido. Gracias al teorema 2 conocemos la media y la varianza de los estimadores, de esta manera $\beta_1 \sim N(\beta_1, \sigma^2/S_{xx})$. Con esto, el estadístico de prueba

$$Z_0 = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2/S_{xx}}}$$

tiene una distribución normal estándar si la hipótesis nula es cierta. Si conocemos la varianza σ^2 del error podemos utilizar Z_0, pero es poco común saber el valor real, sin embargo, es posible estimarla con $\hat{\sigma}^2$ dada en el teorema 3. Note que $(n-2)\hat{\sigma}^2/\sigma^2$ tiene una distribución χ^2_{n-2} y es independiente de $\hat{\beta}_1$. Por definición, el estadístico t dado por

$$t_0 = \frac{Z}{\sqrt{\chi^2/n}} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}^2/S_{xx}}} \tag{8}$$

sigue una distribución t_{n-2} si la hipótesis nula es verdadera. Definamos al nivel de confianza α como la probabilidad de rechazar H_0 cuando es cierta. Esto nos da la posibilidad de rechazar H_0 si

$$|t_0| > t_{n-2}^{\alpha/2}$$

De manera análoga podemos encontrar un estadístico para la prueba

$$H_0 : \ \beta_0 = \beta_0^* \ vs \ H_a : \beta_0 \neq \beta_0^*$$

utilizamos el estadístico

$$t_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}$$

Note que el denominador en ambos casos es la desviación estándar del parámetro. Tomando esto en cuenta se puede hacer una pequeña generalización que toma más importancia en el caso multiparamétrico. El estadístico de prueba t_0 para

$$H_0 : \ \beta_i = \beta_i^* \ vs \ H_a : \beta_i \neq \beta_i^*$$

Está dado por

$$t_0 = \frac{\hat{\beta}_i - \beta_i^*}{se(\hat{\beta}_i)} \tag{9}$$

Veremos el **enfoque de razón de verosimilitudes** para esta prueba y, posterior, concluir los mismos resultados. Sea $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ el modelo de regresión y

$$H_0 : \beta_1 = 0 \ vs \ H_a : \beta_1 \neq 0$$

En el espacio paramétrico $\underline{\theta} = (\beta_0, \beta_1, \sigma^2)$ y función de distribución $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$, tenemos la función de verosimilitud

$$L(y; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(\frac{-(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(\frac{-\sum(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

Bajo H_0

$$\omega = \{(\beta_0, \beta_1, \sigma^2) | \beta_0 \in \mathbb{R}, \beta_1 = 0, \sigma^2 > 0\} \ \Rightarrow \ L(y; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(\frac{-\sum(y_i - \beta_0)^2}{2\sigma^2}\right) \tag{10}$$

Bajo H_a

$$H = \{(\beta_0, \beta_1, \sigma^2) | \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R} - \{0\}, \sigma^2 > 0\} \Rightarrow L(y; \beta_0, \beta_1, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(\frac{-\sum(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right)$$

Usando (10)

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum(y_i - \beta_0)^2}{2\sigma^2}$$

Encontrando los estimadores de máx-vero.

$$\frac{\partial \ln L}{\partial \beta_0} = \frac{-1}{2\sigma^2} 2\sum(y_i - \beta_0)(-1) = \frac{\sum(y_i - \beta_0)}{2\sigma^2} = 0$$
$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{2\sum(y_i - \beta_0)^2}{4\sigma^2} = \frac{-n}{2\sigma^2} + \frac{\sum(y_i - \beta_0)^2}{2\sigma^2} = 0$$

Entonces

$$\sum y_i = n\beta_0 \Rightarrow \frac{\sum y_i}{n} = \beta_0 \Rightarrow \bar{y} = \hat{\beta}_0$$
$$\frac{n}{2\sigma^2} = \frac{\sum(y_i - \beta_0)^2}{2\sigma^4} \Rightarrow \hat{\sigma}^2 = \frac{\sum(y_i - \beta_0)^2}{n} = \frac{\sum(y_i - \bar{y})^2}{n}$$

Así, $\hat{\theta}_\omega = (\bar{y}, \hat{\sigma}^2)$. De donde

$$L(\hat{\omega}) = \left(\frac{n}{2\pi\sum(y_i - \bar{y})^2}\right)^{n/2} \exp\left(\frac{-\sum(y_i - \bar{y})^2 n}{2\sum(y_i - \bar{y})^2}\right) = \left(\frac{n}{2\pi\sum(y_i - \bar{y})^2}\right)^{n/2} \exp\left(-\frac{n}{2}\right)$$

Bajo H_a note que tenemos la función de verosimilitud completa y ya hemos el espacio paramétrico estimado por

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$
$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}$$
$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$$

Es decir, $\hat{\theta}_H = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$. De esta manera

$$L(\hat{H}) = \left(\frac{1}{2\pi\frac{\sum e_i^2}{n}}\right)^{n/2} \exp\left(-\frac{\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{2\frac{\sum e_i^2}{n}}\right) = \left(\frac{n}{2\pi\sum e_i^2}\right)^{n/2} \exp\left(-\frac{n}{2}\right)$$

Nuestro criterio de razón de verosimilitudes nos dice que si

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{H})} < \lambda_0$$

se rechaza H_0. Y si

$$\lambda = \frac{L(\hat{\omega})}{L(\hat{H})} \geq \lambda_0$$

no se rechaza H_0. Por lo cual

$$\lambda = \frac{\left(\frac{n}{2\pi\sum(y_i-\bar{y})^2}\right)^{n/2}\exp\left(-\frac{n}{2}\right)}{\left(\frac{n}{2\pi\sum e_i^2}\right)^{n/2}\exp\left(-\frac{n}{2}\right)} = \left(\frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}\right)^{n/2} < \lambda_0 \iff \lambda^* = \frac{\sum e_i^2}{\sum(y_i-\bar{y})^2} < \lambda_0^{2/n} = \lambda_0^*$$

Hagamos el desarrollo de los residuales

$$\begin{aligned}\sum e_i^2 &= \sum(y_i-\hat{y}_i)^2 = \sum(y_i-\bar{y}+\bar{y}-\hat{y}_i)^2 \sum[(y_i-\bar{y})+(\bar{y}-\hat{y}_i)]^2 \\ &= \sum(y_i-\bar{y})^2 - 2\sum(y_i-\bar{y})(\hat{y}_i-\bar{y}) + \sum(\hat{y}_i-\bar{y})^2\end{aligned} \tag{11}$$

Donde

$$\begin{aligned}2\sum(y_i-\bar{y})(\hat{y}_i-\bar{y}) &= 2\sum(y_i-\bar{y})(\hat{\beta}_0+\hat{\beta}_1 x_i-\bar{y}) \\ &= 2\sum(y_i-\bar{y})(\bar{y}-\hat{\beta}_1\bar{x}+\hat{\beta}_1 x_i-\bar{y}) \\ &= 2\hat{\beta}_1\sum(y_i-\bar{y})(x_i-\bar{x}) \\ &= 2\hat{\beta}_1\sum[(y_i-\bar{y})x_i-(y_i-\bar{y})\bar{x}] \\ &= 2\hat{\beta}_1[\sum(y_i-\bar{y})x_i-\bar{x}(\sum y_i-n\bar{y})] \\ &= 2\hat{\beta}_1[\sum(y_i-\bar{y})x_i-\bar{x}(\sum y_i-\sum y_i)] \\ &= 2\hat{\beta}_1\sum(y_i-\bar{y})x_i\left(\frac{\sum(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) \\ &= 2\hat{\beta}_1\sum(x_i-\bar{x})^2\hat{\beta}_1 \\ &= 2\hat{\beta}_1^2\sum(x_i-\bar{x})^2 \\ &= 2\sum(\hat{y}_i-\bar{y})^2 \qquad (*)\end{aligned}$$

En $(*)$; sabemos que $\hat{y}_i = \hat{\beta}_0+\hat{\beta}_1 x_i$, entonces

$$\begin{aligned}\hat{y}_i-\bar{y} = \hat{\beta}_0+\hat{\beta}_1 x_i-\bar{y} = \bar{y}-\hat{\beta}_1\bar{x}+\hat{\beta}_1 x_i-\bar{y} = \hat{\beta}_1(x_i-\bar{x}) \\ \therefore\ (\hat{y}_i-\bar{y})^2 = \hat{\beta}_1^2(x_i-\bar{x})^2\end{aligned} \tag{12}$$

Regresando a (11)

$$\sum(y_i-\hat{y}_i)^2 = \sum(y_i-\bar{y})^2 - 2\sum(\hat{y}_i-\bar{y})^2 + \sum(\hat{y}_i-\bar{y})^2 = \sum(y_i-\bar{y})^2 - \sum(\hat{y}_i-\bar{y})^2$$

Por lo tanto

$$\sum(y_i-\hat{y}_i)^2 = \sum(y_i-\hat{y}_i)^2 + \sum(\hat{y}_i-\bar{y})^2$$

$$\text{Realidad total} = \text{Error} + \text{Estimación de la regresión}$$

Regresando a la ecuación de λ^*

$$\lambda^* = \frac{\sum e_i^2}{\sum(y_i-\bar{y})^2} = \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2} = \frac{1}{1+\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}} \leq \lambda_0^*$$

Entonces

$$1+\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}>\frac{1}{\lambda_0^*} \iff \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}>\frac{1}{\lambda_0^*}-1$$

De esta manera, la zona de rechazo es

$$\mathfrak{C}=\left\{(x_i,y_i), i=\overline{1,n} \mid \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}>\frac{1}{\lambda_0^*}-1\right\}$$

Recordemos que

$$\frac{(\hat{\beta}_1-\beta_1)^2}{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}}\sim\chi^2_{(1)}$$

Si $H_0:\beta_1=0$ es verdadero, lo anterior es

$$\frac{\sum(x_i-\bar{x})\hat{\beta}_1^2}{\sigma^2}\sim\chi^2_{(1)}$$

Haciendo la estimación común S^2 para σ^2, tenemos

$$\frac{\frac{\sum(x_i-\bar{x})^2\hat{\beta}_1^2}{\sigma^2}}{\frac{\sum e_i^2}{\sigma^2(n-2)}}=\frac{\hat{\beta}_1^2\sum(x_i-\bar{x})^2(n-2)}{\sum e_i^2}\sim F_{(1,n-2)}$$

Así, para construir el criterio de decisión

$$\frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}>C_\alpha \Rightarrow \frac{(n-2)\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}>C_\alpha(n-2)=k_\alpha$$

Tal que

$$\alpha=P(\text{Rechazar } H_0|H_0 \text{ es verdadera})=P\left(\frac{(n-2)\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\hat{y}_i)^2}>k_\alpha \text{ con } F_{(1,n-2)}\right)$$

Así, por último, sabemos que se rechaza H_0 si

$$\frac{(n-2)}{\sum(y_i-\bar{y})^2}\sum(\hat{y}_i-\bar{y})^2=\frac{\sum(\hat{y}_i-\bar{y})^2}{\hat{\sigma}^2}>F^{1-\alpha}_{(1,n-2)}$$

1.5. Intervalos de confianza

Además de obtener la estimación de los parámetros β_0, β_1 y σ^2, puede resultar interesante obtener los intervalos de confianza de estos parámetros. La anchura de estos intervalos puede ser una medida de la calidad de una línea de regresión, donde un intervalo amplio implica menos certeza sobre el verdadero valor del parámetro. Con la información presentada en la sección anterior, crear un intervalo de confianza resulta casi natural. Sin embargo, se aprovecha este nuevo espacio para hacer algunas especificaciones.

Como sabemos que $\hat{\beta}_0\sim N(\beta_0,\frac{\sigma^2\sum x_i}{n\sum(x_i-\bar{x})^2})$ y que $\hat{\beta}_1\sim N(\beta_1,\frac{\sigma^2}{\sum(x_i-\bar{x})^2})$ podemos estandarizar estas variables, de tal forma que

$$z=\frac{\hat{\beta}_0-\beta_0}{\sqrt{\frac{\sigma^2\sum x_i}{n\sum(x_i-\bar{x})^2}}}=(\hat{\beta}_0-\beta_0)\frac{\sqrt{n\sum(x_i-\bar{x})^2}}{\sigma\sqrt{\sum x_i^2}}\sim N(0,1)$$

Como no conocemos a σ^2 hay que estimarla con su estimador insesgado $\hat{\sigma^2} = \frac{\sum e_i^2}{n-2} \sim \chi^2_{n-2}$. Entonces, análogo a la ecuación (8)

$$\frac{Z}{\sqrt{\chi^2_n/n}} = \frac{(\hat{\beta_0} - \beta_0)\frac{\sqrt{n\sum(x_i-\bar{x})^2}}{\sigma\sqrt{\sum x_i^2}}}{\sqrt{\frac{(n-2)\hat{\sigma^2}}{\sigma^2(n-2)}}} = \frac{(\hat{\beta_0} - \beta_0)\sqrt{n\sum(x_i-\bar{x})^2}}{\hat{\sigma}\sqrt{\sum x_i^2}} \sim t_{(n-2)}$$

Observaciones:

- Este error es muy pequeño y tiende a desaparecer de la anterior expresión.
- La muestra depende de las observaciones y del valor hipotético de β_0.
- Para saber en que región cae el valor de $\frac{Z}{\sqrt{\chi^2_n/n}}$ con la distribución $t_{(n-2)}$ obtenemos, por ejemplo, al nivel $1-\alpha$ el intervalo de confianza para el parámetro β_0 es

$$\begin{aligned}
1-\alpha &= P\left(-t_{n-2}^{\alpha/2} < \frac{Z}{\sqrt{\chi^2_n/n}} < t_{n-2}^{\alpha/2}\right) \\
&= P\left(-t_{n-2}^{\alpha/2} < \frac{(\hat{\beta_0} - \beta_0)\sqrt{n\sum(x_i-\bar{x})^2}}{\hat{\sigma}\sqrt{\sum x_i^2}} < t_{n-2}^{\alpha/2}\right) \\
&= P\left(\hat{\beta_0} - t_{n-2}^{\alpha/2}\frac{\hat{\sigma}\sqrt{\sum x_i^2}}{\sqrt{n\sum(x_i-\bar{x})^2}} < \beta_0 < \hat{\beta_0} + t_{n-2}^{\alpha/2}\frac{\hat{\sigma}\sqrt{\sum x_i^2}}{\sqrt{n\sum(x_i-\bar{x})^2}}\right)
\end{aligned}$$

Análogamente para β_1, tenemos la estandarización

$$z = \frac{\hat{\beta_1} - \beta_1}{\sqrt{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}}} = \frac{(\hat{\beta_1} - \beta_1)\sqrt{\sum(x_i-\bar{x})^2}}{\sigma} \sim N(0,1)$$

Nuevamente, estimamos la varianza poblacional a conciencia de la independencia que tiene con $\hat{\beta_1}$ y $\hat{\beta_0}$. Entonces

$$\frac{Z}{\sqrt{\chi^2_n/n}} = \frac{\frac{(\hat{\beta_1} - \beta_1)\sum(x_i-\bar{x})^2}{\sigma}}{\sqrt{\frac{\hat{\sigma^2}(n-2)}{\sigma^2(n-2)}}} = \frac{(\hat{\beta_1} - \beta_1)\sum(x_i-\bar{x})^2}{\hat{\sigma}} \sim t_{(n-2)}$$

Construyendo el intervalo de confianza

$$\begin{aligned}
1-\alpha &= P\left(-t_{n-2}^{\alpha/2} < \frac{(\hat{\beta_1} - \beta_1)\sum(x_i-\bar{x})^2}{\hat{\sigma}} < t_{n-2}^{\alpha/2}\right) \\
&= P\left(\hat{\beta_1} - t_{n-2}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}} < \beta_1 < \hat{\beta_1} + t_{n-2}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}}\right)
\end{aligned}$$

Si utilizamos la notación de la ecuación (9) llegamos a la forma compacta

$$P\left(\hat{\beta_i} - t_{n-2}^{\alpha/2}\sqrt{V(\beta_i)} < \beta_i < \hat{\beta_i} + t_{n-2}^{\alpha/2}\sqrt{V(\beta_i)}\right) = 1-\alpha \tag{13}$$

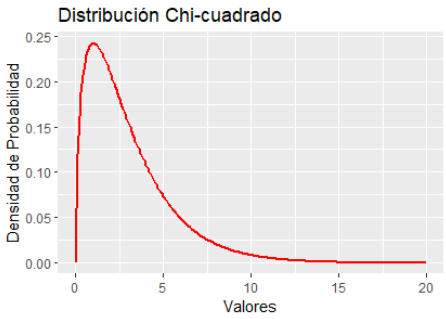


Figura 3: Distribución χ^2.

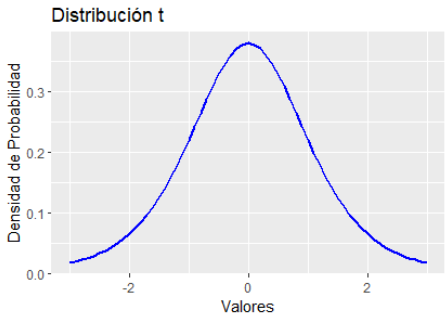


Figura 4: Distribución t.

La simetría de la distribución t fig. (4) nos da la ventaja de tener el mismo valor de $t_{(n-2)}$, sin embargo, para estimar la varianza poblacional usamos $\hat{\sigma^2}$ que sigue una distribución chi cuadrada χ^2 fig. (3) sin simetría. Es por ello que los valores críticos son distintos para cola izquierda y derecha de tal manera que $P(x \leq a) = \alpha/2$ y $P(x \geq b) = \alpha/2$. Entonces, el intervalo de confianza para σ^2 con el estimador $\hat{\sigma^2} \sim \chi^2_{(n-2)}$ y cola izquierda para la distribución chi es

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{\alpha/2,n-2} < \frac{\sum e_i^2}{\sigma^2} < \chi^2_{1-\alpha/2,n-2}\right) \\
&= P\left(\chi^2_{\alpha/2,n-2} < \frac{\hat{\sigma^2}(n-2)}{\sigma^2} < \chi^2_{1-\alpha/2,n-2}\right) \\
&= P\left(\frac{1}{\chi^2_{\alpha/2,n-2}} > \frac{\sigma^2}{\hat{\sigma^2}(n-2)} > \frac{1}{\chi^2_{1-\alpha/2,n-2}}\right) \\
&= P\left(\frac{\hat{\sigma^2}(n-2)}{\chi^2_{\alpha/2,n-2}} < \sigma^2 < \frac{\hat{\sigma^2}(n-2)}{\chi^2_{1-\alpha/2,n-2}}\right)
\end{aligned}$$

1.6. Predicción

Es común que la mayoría de los casos de regresión, sólo se tenga el objetivo de conocer los valores a futuro o pasado del fenómeno estudiado, por lo cual se construye una herramienta estadística que nos ayuda y nos brinda cierta confianza de los posibles resultados.

Supongamos que seguimos con el modelo lineal, donde el problema es predecir el valor medio de y que corresponde a un valor de x, sea x_0 $E(y|x_0)$, el cual puede o no puede pertenecer al rango de variación de los valores muestrales. Existen predicciones puntuales o de intervalos. Suponga que definimos a un estimador como una función lineal de los y_i, $i = \overline{1,n}$. Sea $\hat{y_0} = \sum c_i y_i$ donde puede suceder que la c_i es una constante que permite ser a $\hat{y_0}$ el mejor estimador lineal e insesgado. Nos gustaría que

$$E(y_0|x_0) = E(\beta_0 + \beta_1 x_0 + \varepsilon_0) = \beta_0 + \beta_1 x_1$$

En realidad tenemos

$$\hat{y_0} = \sum c_i(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 \sum c_i + \beta_1 \sum c_i x_i + \sum c_i \varepsilon_i \tag{14}$$

De esta manera, al obtener la esperanza condicionada a un valor particular x_0 tenemos que

$$\begin{aligned} E(\hat{y}_0|x_0) &= E(\beta_0 \sum c_i + \beta_1 \sum c_i x_i + \sum c_i \varepsilon_i) = \beta_0 E(\sum c_i) + \beta_1 E(\sum c_i x_i) + E(\sum c_i \varepsilon_i) \\ &= \beta_0 E(\sum c_i) + \beta_1 E(\sum c_i x_i) = \beta_0 + \beta_1 x_0 \iff \sum c_i = 1 \wedge \sum c_i x_i = x_0 \end{aligned} \tag{15}$$

Si lo anterior se cumple podemos afirmar que nuestro estimador es insesgado. ¿Será de varianza mínima?

$$\begin{aligned} V(\hat{y}_0) &= E[\hat{y}_0 - E(\hat{y}_0|x_0)]^2 = E[\beta_0 + \beta_1 x_0 - (\beta_0 + \beta_1 x_0 + \sum c_i \varepsilon_i)]^2 \\ &= E[-\sum c_i \varepsilon_i]^2 = E[\sum c_i \varepsilon_i]^2 = E[\sum c_i^2 \varepsilon_i^2 + \sum \sum_{i \neq j} c_i c_j \varepsilon_i \varepsilon_j] \\ &= \sum c_i^2 E(\varepsilon_i^2) + \sum \sum_{i \neq j} c_i c_j E(\varepsilon_i \varepsilon_j) = \sum c_i^2 \sigma^2 \end{aligned} \tag{16}$$

Como $\sigma^2 > 0$, utilizaremos el criterio de multiplicadores de Lagrange. De (15) vemos que $\sum c_i - 1 = 0$ y $\sum c_i x_i - x_0 = 0$. Sea $\varphi = \sum c_i^2 - 2\lambda(\sum c_i - 1) - 2\mu(\sum c_i x_i - x_0) = 0$ entonces

$$\frac{\partial \varphi}{\partial c_i} = 2c_i - 2\lambda - 2\mu x_i = 0 \qquad \Rightarrow \qquad c_i - \lambda - \mu x_i = 0 \tag{17}$$

$$\frac{\partial \varphi}{\partial \lambda} = -2(\sum c_i - 1) = 0 \qquad \Rightarrow \qquad \sum c_i = 1 \tag{18}$$

$$\frac{\partial \varphi}{\partial \mu} = -2(\sum c_i x_i - x_0) = 0 \qquad \Rightarrow \qquad \sum c_i x_i = x_0 \tag{19}$$

Sumando en (17)

$$\sum c_i - n\lambda - \mu \sum x_i = 0$$

Despejando λ y usando (18)

$$\lambda = \frac{\mu \sum x_i - \sum c_i}{-n} = -\mu \bar{x} + \frac{1}{n} \tag{20}$$

Sustituyendo (20) en (17)

$$c_i = \lambda + \mu x_i = -\mu \bar{x} + \frac{1}{n} + \mu x_i \tag{21}$$

Ahora, usando (21) en (19) llegamos a

$$\begin{aligned} \sum c_i x_i &= \sum (-\mu \bar{x} + \frac{1}{n} + \mu x_i) x_i = -\mu \bar{x} \sum x_i + \frac{\sum x_i}{n} + \mu \sum x_i^2 \\ &= \bar{x} + \mu \sum x_i (x_i - \bar{x}) = x_0 \end{aligned} \tag{22}$$

Si de la última igualdad despejamos a μ obtenemos

$$\mu = \frac{x_0 - \bar{x}}{\sum x_i (x_i - \bar{x})} \tag{23}$$

Con esto, ya podemos encontrar un valor explícito de c_i si sustituimos μ en (21)

$$c_i = -\mu \bar{x} + \frac{1}{n} + \mu x_i = \frac{1}{n} + \mu (x_i - \bar{x}) = \frac{1}{n} + \frac{x_0 - \bar{x}}{\sum x_i (x_i - \bar{x})} (x_i - \bar{x}) \tag{24}$$

Entonces, podemos expresar a $\hat{y}_0$ de (14) en términos de (24)

$$\begin{aligned}
\hat{y}_0 &= \sum c_i(\beta_0+\beta_1 x_i+\varepsilon_i) = \sum c_i y_i \\
&= \sum [\frac{1}{n}+\frac{x_0-\bar{x}}{\sum x_i(x_i-\bar{x})}(x_i-\bar{x})]y_i \\
&= \frac{\sum y_i}{n}+\sum \frac{(x_0-\bar{x})(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})x_i} \\
&= \bar{y}+(x_0-\bar{x})\hat{\beta}_1 \\
&= \bar{y}+x_0\hat{\beta}_1-\bar{x}\hat{\beta}_1 \\
&= \bar{y}-\bar{x}\hat{\beta}_1+x_0\hat{\beta}_1 \\
&= \hat{\beta}_0+x_0\hat{\beta}_1
\end{aligned}$$

Por lo tanto, el mejor estimador de $\beta_0+\beta_1 x_0$ es $\hat{\beta}_0+\hat{\beta}_1 x_0$. Desde (15) conocemos la esperanza de $\hat{y}_0$, ahora que conocemos c_i podemos sustituir en (16) para encontrar la varianza.

$$\begin{aligned}
V(\hat{y}_0) &= \sum c_i\sigma^2 = \sigma^2\sum[\frac{1}{n}+\frac{(x_i-x_0)(x_0-\bar{x})}{\sum(x_i-\bar{x})x_i}] \\
&= \sigma^2\sum[\frac{1}{n^2}+\frac{2(x_i-x_0)(x_0-\bar{x})}{n\sum(x_i-\bar{x})x_i}+(\frac{(x_i-x_0)(x_0-\bar{x})}{\sum(x_i-\bar{x})x_i})^2] \\
&= \sigma^2[\frac{1}{n}+\frac{(x_0-\bar{x})^2\sum(x_i-\bar{x})^2}{(\sum(x_i-\bar{x})x_i)^2}] \qquad (*) \\
&= \sigma^2[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}] \qquad (**)
\end{aligned}$$

En $(*)$, veamos que $\sum \frac{2(x_i-x_0)(x_0-\bar{x})}{n\sum(x_i-\bar{x})x_i}=0$

$$\begin{aligned}
\sum \frac{2(x_i-x_0)(x_0-\bar{x})}{n\sum(x_i-\bar{x})x_i} &= 2(x_0-\bar{x})\frac{\sum(x_i-\bar{x})}{n\sum(x_i-\bar{x})x_i} = 2(x_0-\bar{x})\frac{\sum x_i-\sum\bar{x}}{n\sum(x_i-\bar{x})x_i} \\
&= 2(x_0-\bar{x})[\frac{\bar{x}}{\sum(x_i-\bar{x})x_i}-\frac{\bar{x}}{\sum(x_i-\bar{x})x_i}] = 0
\end{aligned}$$

En $(**)$, demostremos que $\frac{\sum(x_i-\bar{x})^2}{(\sum(x_i-\bar{x})x_i)^2}=\frac{1}{\sum(x_i-\bar{x})^2}$. Sabemos que $\sum(x_i-\bar{x})^2=\sum(x_i-\bar{x})x_i$, entonces

$$1=\frac{\sum(x_i-\bar{x})^2}{\sum(x_i-\bar{x})x_i} \Rightarrow 1=\frac{(\sum(x_i-\bar{x})^2)^2}{(\sum(x_i-\bar{x})x_i)^2} \Rightarrow \frac{\sum(x_i-\bar{x})^2}{(\sum(x_i-\bar{x})x_i)^2}=\frac{1}{\sum(x_i-\bar{x})^2}$$

Se puede obtener el mismo resultado calculando la varianza por definición, de tal forma que $V(\hat{y}_0)=E(\hat{y}_0-E(\hat{y}_0|x_0))^2=E(\hat{\beta}_0+\hat{\beta}_1x_0-(\beta_0+\beta_1x_0))^2$. La demostración se deja como entretenida actividad al lector. En resumen, podemos decir que

$$\begin{aligned}
E(\hat{y}_0|x_0) &= \beta_0+\beta_1x_0 \\
V(\hat{y}_0) &= \sigma^2[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}]
\end{aligned} \tag{25}$$

Donde $\hat{y}_0$ es una transformación lineal de una v.a. normal, de tal manera que

$$\hat{y}_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}]\right)$$

Una vez que conocemos la distribución, podemos encontrar un **intervalo de confianza** para la respuesta media de y dado un x_0, es decir, $E(\hat{y}_0|x_0)$. Estandarizando $\hat{y}_0$ y dividiendo entre la raíz de una chi cuadrada llegamos a una distribución t.

$$\frac{\frac{\hat{y}_0 - E(\hat{y}_0|x_0)}{\sqrt{\sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}]}}}{\sqrt{\frac{(n-2)\hat{\sigma^2}}{\sigma^2(n-2)}}} = \frac{\hat{y}_0 - E(\hat{y}_0|x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}} \sim t_{(n-2)}$$

De esta manera, nuestro intervalo al $(100)(1-\alpha)\,\%$ para $E(\hat{y}_0|x_0)$ es de la forma

$$\begin{aligned} 1-\alpha &= P\left(-t^{\alpha/2}_{(n-2)} \leq \frac{\hat{y}_0 - E(\hat{y}_0|x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}} \leq t^{\alpha/2}_{(n-2)}\right) \\ &= P\left(\hat{y}_0 - t^{\alpha/2}_{(n-2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \leq E(\hat{y}_0|x_0) \leq \hat{y}_0 + t^{\alpha/2}_{(n-2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right) \end{aligned} \tag{26}$$

Mientras los valores a predecir se alejan del centro de los datos es razonable perder predictivilidad. Eso lo podemos ver en el intervalo de confianza, pues mientras $|x_0 - \bar{x}|$ toma valores mayores, el intervalo se amplía.

1.6.1. Predicción de nuevas observaciones (x_0, y_0)

Hasta ahora, conocemos un intervalo de confianza para el valor medio (esperanza) de y dado un valor fijo de x, dicho x_0. Pero, dada una pareja (x_0, y_0) observada ¿pertenece a la estructura lineal? En otras palabras, queremos saber si el modelo $y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ sigue siendo el apropiado. Definamos

$$\eta = y_0 - \hat{y}_0 = (\beta_0 + \hat{\beta}_0) + x_0(\beta_1 - \hat{\beta}_1) + \varepsilon_0 \tag{27}$$

Como y_0 y $\hat{y}_0$ son independientes y normalmente distribuidos, entonces η también lo es. Veamos sus propiedades.

$$E(\eta) = E(y_0 - \hat{y}_0) = E(y_0) - E(\hat{y}_0) = \hat{y}_0 - \hat{y}_0 = 0$$

$$
\begin{aligned}
V(\eta) &= E(\eta - E(\eta))^2 = E(\eta)^2 \\
&= E((\beta_0 + \hat{\beta}_0) + x_0(\beta_1 - \hat{\beta}_1) + \varepsilon_0)^2 \\
&= E[\varepsilon_0^2 + ((\beta_0 + \hat{\beta}_0) + x_0(\beta_1 - \hat{\beta}_1))^2 + 2\varepsilon_0((\beta_0 + \hat{\beta}_0) + x_0(\beta_1 - \hat{\beta}_1))] \\
&= E[\varepsilon_0^2] + E[((\beta_0 + \hat{\beta}_0) + x_0(\beta_1 - \hat{\beta}_1))^2] + 2E[\varepsilon_0((\beta_0 + \hat{\beta}_0) + x_0(\beta_1 - \hat{\beta}_1))] \\
&= \sigma^2 + E[\beta_0 - \hat{\beta}_0]^2 + E[x_0(\beta_1 - \hat{\beta}_1)]^2 + 2E[x_0(\beta_0 - \hat{\beta}_0)(\beta_1 - \hat{\beta}_1)] \\
&= \sigma^2 + V(\hat{\beta}_0) + x_0^2 V(\hat{\beta}_1) + 2x_0 E[(\beta_0 - \hat{\beta}_0)(\beta_1 - \hat{\beta}_1)] \\
&= \sigma^2 + \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} + \frac{\sigma^2 x_0^2}{\sum(x_i - \bar{x})^2} + 2x_0 Cov(\hat{\beta}_0, \hat{\beta}_1) \\
&= \sigma^2 + \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2} + \frac{\sigma^2 x_0^2}{\sum(x_i - \bar{x})^2} - \frac{2x_0 \bar{x}\sigma^2}{\sum(x_i - \bar{x})^2} \\
&= \sigma^2[1 + \frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2} + \frac{x_0^2}{\sum(x_i - \bar{x})^2} - \frac{2x_0\bar{x}}{\sum(x_i - \bar{x})^2}] \\
&= \sigma^2[1 + \frac{\sum(x_i - \bar{x})^2 + n\bar{x}^2}{n\sum(x_i - \bar{x})^2} + \frac{x_0^2 - 2x_0\bar{x}}{\sum(x_i - \bar{x})^2}] \\
&= \sigma^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}]
\end{aligned}
$$

De esta manera, podemos dar la distribución explícita de η

$$
\eta \sim N\left(0, \sigma^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}]\right)
$$

Ahora, con avidez, somos capaces de crear un intervalo de confianza para $y_0 - \hat{y}_0 = \eta$. Análogo al caso anterior, encontramos una distribución Student.

$$
\frac{\frac{\eta - 0}{\sigma\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}}{\sqrt{\frac{(n-2)\hat{\sigma^2}}{\sigma^2(n-2)}}} = \frac{\eta}{\sqrt{\hat{\sigma^2}\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}} = \frac{y - \hat{y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}} \sim t_{(n-2)}
$$

Así, nuestro intervalo es de la forma

$$
\begin{aligned}
1 - \alpha &= P\left(-t_{(n-2)}^{\alpha/2} \leq \frac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}} \leq t_{(n-2)}^{\alpha/2}\right) \\
&= P\left(\hat{y}_0 - t_{(n-2)}^{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \leq y_0 \leq \hat{y}_0 + t_{(n-2)}^{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}\right)
\end{aligned} \tag{28}
$$

Igual al caso anterior, mientras los valores se alejan del centro $|x_0 - \bar{x}|$, el intervalo es más grande y deja de explicar la realidad. A este hecho se le llama *PELIGRO MATEMÁTICO*, ya que es independiente la validez del modelo, es decir, aún siendo bueno el modelo se corre el riesgo de predecir en forma poco precisa. Otro peligro es el llamado *PRÁCTICO*; este consiste en que el modelo es sólo una aproximación de la realidad y nunca un modelo puede absolutamente correcto.

Aunque los intervalos (26) y (28) son muy similares, son para diferentes elementos. El primero es un intervalo de confianza para la respuesta media (esperanza) de los valores de y dado un valor de x fijo; mientras que el segundo intervalo es para un fijo de y dado un x igualmente fijo. Note que el primer intervalo siempre será más angosto, por tener menos variabilidad en las bandas Fig. (5).

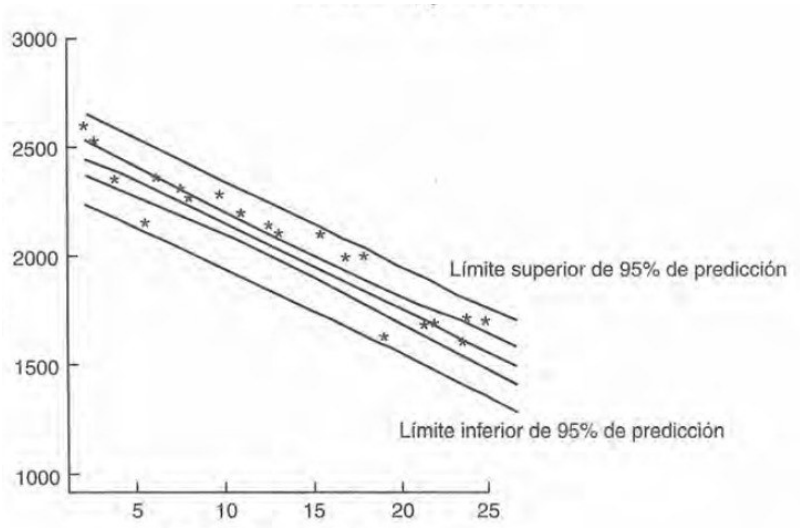


Figura 5: Intervalos de confianza para $E(y_0|x_0)$ & y_0.

1.7. Análisis de varianza

Ya hemos definido al error e, pero se puede definir de manera más general que tanto se aleja un valor estimado de la media de los datos y del valor estimado. Suponga, para un valor x_i, un valor observado y_i y un valor estimado $\hat{y}_i$, además del valor medio $\bar{y}$ para todos los valores observados de y. En la figura (6) se puede ver de manera explícita la siguiente relación

$$\begin{aligned}(y_i-\bar{y}) &= (\hat{y}_i-\bar{y})+(y_i-\hat{y}_i)\\ &= (\hat{y}_i-\bar{y})+e_i\end{aligned}$$

Si hacemos la suma para $i=1,2,\ldots,n$ de los cuadrados, podemos llegar a

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+2\sum(\hat{y}_i-\bar{y})(y_i-\hat{y})+\sum(y_i-\hat{y}_i)^2$$

Note que

$$\begin{aligned}2\sum(\hat{y}_i-\bar{y})(y_i-\hat{y}) &= 2\sum\hat{y}_i(y_i-\hat{y})+2\bar{y}\sum(y_i-\hat{y})\\ &= 2\sum\hat{y}_i e_i-2\hat{y}\sum e_i=0 \qquad (*)\end{aligned}$$

$(*)$ por la propiedad [7].

Entonces, esta identidad se reduce a

$$\sum(y_i-\bar{y})^2=\sum(\hat{y}_i-\bar{y})^2+\sum(y_i-\hat{y}_i)^2 \tag{29}$$

Note que

$$\sum(\hat{y}_i - \bar{y})^2 = \sum \hat{\beta}_1^2 S_{xx} = \hat{\beta}_1 S_{xy} = \frac{S_{xy}^2}{S_{xx}} \tag{30}$$

Estos términos son tan importantes para el estudio del modelo que reciben el nombre de la *igualdad fundamental del análisis de varianza*. $\sum(y_i - \bar{y})^2$ es la suma de cuadrados debido a la media o suma total SS_T (total sum of squares) que mide la variabilidad total de las observaciones, $\sum(\hat{y}_i - \bar{y})^2$ es la cantidad de variabilidad en las observaciones y_i explicada por la línea de regresión SS_R y $\sum(y_i - \hat{y}_i)^2$ la variación residual que queda sin explicar por la línea de regresión SS_{Res} o SS_e.

$$SS_T = SS_R + SS_e \tag{31}$$

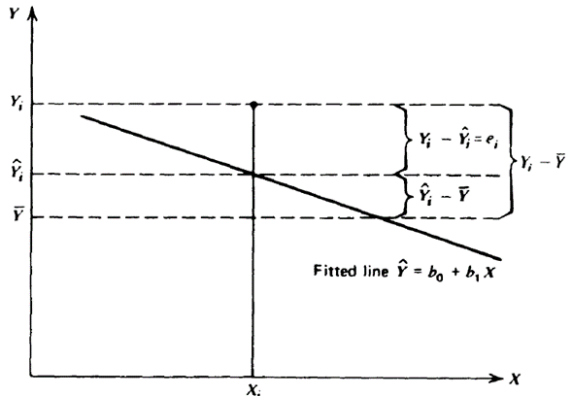


Figura 6: Suma de cuadrados.

Cada término tiene un número de grados de libertad asociado y se describe a partir de 'piezas' independientes de información involucrando los n números independientes $y_1, \ldots, y_n$ que se necesitan para completar la suma de cuadrados. SS_T tiene $(n-1)$ piezas independientes (de los números $y_1 - \bar{y}, \ldots, y_n - \bar{y}$, sólo $(n-2)$ son independientes, ya que todos los n números suman cero por definición de media). Podemos calcular SS_R de una función única de $y_1, \ldots, y_n$, llamada $\hat{\beta}_1$ (ya que $\sum(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum(x_i - \bar{x})^2$ como en la ec. (30)), y de esta manera sólo tiene 1 grados de libertad. Por último, SS_e tiene $(n-2)$ grados de libertad. Note como los grados de libertad son una extensión natural de la ec. (31):

$$n - 1 = 1 + (n-2)$$

Esta información es útil si nos interesa conocer la significancia del modelo. Dado que β_0 casi nunca es cero, una hipótesis interesante es $H_0: \beta_1 = 0$ contra $H_a: \beta_1 \neq 0$, donde evaluamos si la variable X realmente está explicando a Y. Para esta hipótesis conviene la prueba F del análisis de varianza Cuadro (1) que reúne la información necesaria. Sabemos que $SS_e = (n-2)MS_e$ tiene una distribución χ^2_{n-2}; si H_0 es cierta, entonces SS_R tiene también distribución chi-cuadrada e independiente a la anterior. Por definición de la distribución F:

$$F_0 = \frac{SS_R/1}{SS_e/(n-2)} = \frac{MS_R}{MS_e} \sim F_{1,n-2}$$

Bajo H_a, F_0 sigue una distribución F no centralizada con los mismos grados de libertad y el parámetro de no centralidad λ;

$$\lambda = \frac{\beta_1^2 S_{xx}}{\sigma^2}$$

Esto nos indica que para valores grandes de F_0 se rechaza H_0, es decir $\beta_1 \neq 0$. La regla de decisión es rechazar H_0 si $F_0 > F_{1,n-2}^{\alpha}$. Esta información se resume en el Cuadro (1) usualmente llamado ANOVA (analysis of variance).

Cuadro 1: Análisis de varianza de la regresión.

Fuente de Variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	F_0
Regresión	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	$\frac{MS_R}{MS_e}$
Residuales	$SS_e = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_e	
Total	$SS_T = \sum (y_i - \bar{y})^2$	$n-1$		

1.7.1. Estadístico R^2

Una medida de ajuste de nuestro modelo a los datos observados es el estadístico R^2, definido como

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SS_R}{SS_e} \tag{32}$$

Este valor está en el rango $[0,1]$, donde la cota inferior representa que la línea estimada no explica los datos y, por otro lado, el valor 1 representa que todas las observaciones están explicadas por la recta de regresión. Estos supuestos son teóricos, pues si $R^2 = 1$, entonces todos los puntos observados están sobre la misma recta y no existe un error aleatorio ε, de está manera tendríamos un modelo determinista. Este cociente es la proporción de la variación total debido a la media explicado por explicado por la regresión. Una propiedad útil para evaluar este cociente es

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

1.8. Correlación entre x & y

Sea u y w dos variables aleatorias, el coeficiente de correlación se define como

$$\rho_{uw} = \frac{cov(u,w)}{[V(u)V(w)]^{1/2}}$$

Donde

$$cov(u,w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[u-E(u)][w-E(w)]f(u,w)dudw$$
$$V(u) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[u-E(u)]^2 f(u,w)dudw$$
$$E(u) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}uf(u,w)dudw$$

Se puede demostrar que $-1 \leq \rho_{uw} \leq 1$. La cantidad de ρ_{uw} es una medida de la asociación lineal entre dos variables aleatorias, dichas u y w. Por ejemplo, si $\rho_{uw} = 1$, u y w están perfecta y positivamente correlacionadas, de tal manera que posibles valores de u y w ajustarían de manera perfecta sobre una línea con pendiente positiva en el plano (u,w). Por el contrario, si $\rho_{uw} = 0$ las variables no tienen correlación lineal (esto no implica la independencia estadística). Es importante notar que el valor de la correlación entre u y w sólo es una medida de asociación lineal y no implica por si misma una relación de causalidad. Si tenemos una muestra n de la distribución conjunta $f(u,w)$, dicho $(u_1,w_1),\ldots,(u_n,w_n)$, el coeficiente de correlación muestral se define como

$$r_{uw} = \frac{\sum(u_i-\bar{u})(w_i-\bar{w})}{\sum(u_i-\bar{u})^2]^{1/2}[\sum(w_i-\bar{w})^2]^{1/2}} \tag{33}$$

Suponiendo x e y dos variables aleatorias con el modelo $y = \beta_0 + \beta_1 x + \varepsilon$ y una muestra $(x_1,y_1),\ldots,(x_n,y_n)$, utilizando la ec. (33) es sencillo ver que existe una relación entre la estimación de β_1 y el factor de correlación entre x e y, de tal manera que $\hat{\beta}_1 = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} r_{xy}$. Si el lector no logra discernir esta identidad, a continuación se presenta el desarrollo.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sqrt{\sum(y_i-\bar{y})^2}}{\sqrt{\sum(y_i-\bar{y})^2}}\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \\
&= \frac{\sqrt{\sum(y_i-\bar{y})^2}}{\sqrt{\sum(y_i-\bar{y})^2}}\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(x_i-\bar{x})^2}} \\
&= \frac{\sqrt{\sum(y_i-\bar{y})^2}}{\sqrt{\sum(x_i-\bar{x})^2}}\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \\
&= \frac{\sqrt{\sum(y_i-\bar{y})^2}}{\sqrt{\sum(x_i-\bar{x})^2}}\cdot r_{xy} = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}\cdot r_{xy}
\end{aligned}$$

De esta manera, $\hat{\beta}_1$ y r_{xy} guardan una relación, pero proveen diferentes interpretaciones. r_{xy} mide la asociación lineal entre x e y, mientras que $\hat{\beta}_1$ mide el tamaño de cambio en y que puede ser predicho para una unidad de cambio en x. Un cambio de escala en los datos afecta a $\hat{\beta}_1$, pero no a r_{xy}.

El estadístico R^2 también tiene una relación directa con el coeficiente de correlación, tanto que se puede expresar como la correlación entre las observaciones de y_i y el correspondiente valor estimado $\hat{y}_i$.

$$R^2 = r^2_{y\hat{y}}$$

1.9. Transformación de variables

Un punto de partida conveniente en el análisis de regresión es que el modelo que describen los datos es lineal en las variables aleatorias. Para poder llevar a cabo esto el análisis se realiza muy frecuentemente sobre

variables transformadas. La necesidad de transformar los datos surge debido a que la variable original, o el modelo en términos de la variable original, violan una o más de las suposiciones estándares. Las suposiciones que se violan con más frecuencia son las que conciernen a la linealidad del modelo y a la constancia de la varianza de los errores. Un modelo de regresión es lineal cuando los parámetros presentados en el modelo ocurren linealmente. Por ejemplo, los siguiente modelos son lineales.

$$y = \beta_0 + \beta_1 x + \mu$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \mu$$
$$y = \beta_0 + \beta_1 \ln x + \mu$$
$$y = \beta_0 + \beta_1 \sqrt{x} + \mu$$

Debido a que los parámetros $\beta' s$ en el modelo entran linealmente. Pero, $y = \beta_0 + e^{\beta_1 x} + \mu$ no es un modelo lineal dado que el parámetro β_1 no es una entrada lineal en el modelo. Para satisfacer las suposiciones del modelo de regresión estándar, en lugar de trabajar con las variables originales, algunas veces trabajamos con variables transformadas. Las transformaciones pueden ser necesarias por varias razones que se puden resumir de la siguiente manera:

I. Las consideraciones teóricas pueden especificar que la relación entre dos variables e no lineal. Una transformación apropiada de las variables puede hacer que la relación entre las variables transformadas sea lineal. Consideremos un ejemplo en el área teórica de aprendizaje.

   Un modelo de aprendizaje que es ampliamente utilizado establece que el tiempo que toma llevar a cabo una tarea por i-ésima ocasión (T_i) es

   $$T_i = \alpha\beta^i \quad \alpha > 0,\ 0 < \beta < 1$$

   La razón entre T_i e i dada de esta manera no es lineal. No se pueden aplicar las técnicas de regresión lineal directamente. Entonces, hay que transformar. Aplicando el logaritmo en ambos lados, obtenemos

   $$\ln T_i = \ln \alpha + i \ln \beta$$

   La transformación nos permite usar los métodos que conocemos de regresión. Apenar de que la relación entre las variables originales era no lineal, la relación entre las variables transformadas es lineal.

II. La variable dependiente y analizada, puede tener una distribución de probabilidad cuya varianza está relacionada con la media. Si la media está relacionada con el valor de la variable independiente x, entonces la varianza de y cambiará con x y la varianza no será constante. Usualmente la distribución de las n observaciones de y será igualmente distinta de la normal bajo esta situación invalidando las pruebas de significancia.

   Por ejemplo: la varianza no constante de los errores producirá estimadores que son sesgados. En estas situaciones frecuentemente se transforman los datos para asegurar la normalidad y constancia de la varianza de los errores. En la práctica, se eligen las transformaciones para asegurar la constancia de la varianza. Por fortuna, es una coincidencia que las transformaciones que estabilizan la varianza también son buenas transformaciones normalizadoras.

1.9.1. Transformaciones para conseguir linealidad

Hay varios modelos no lineales que por transformaciones adecuadas pueden ser convertidas en lineales. A continuación se dan algunas curvas linealizables y las gráficas correspondientes.

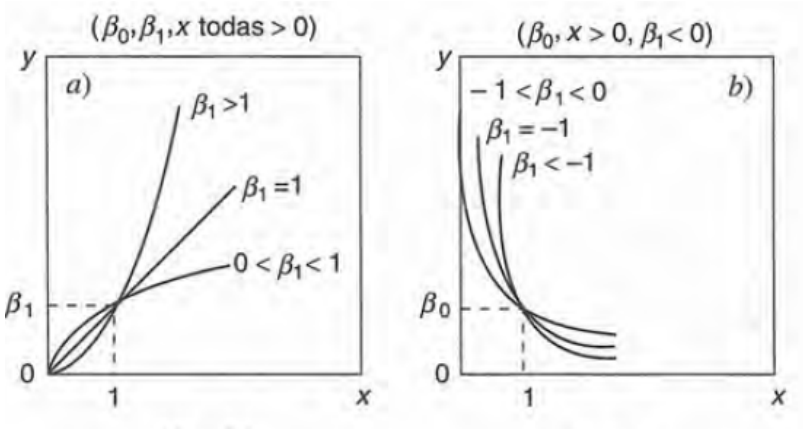


Figura 7:

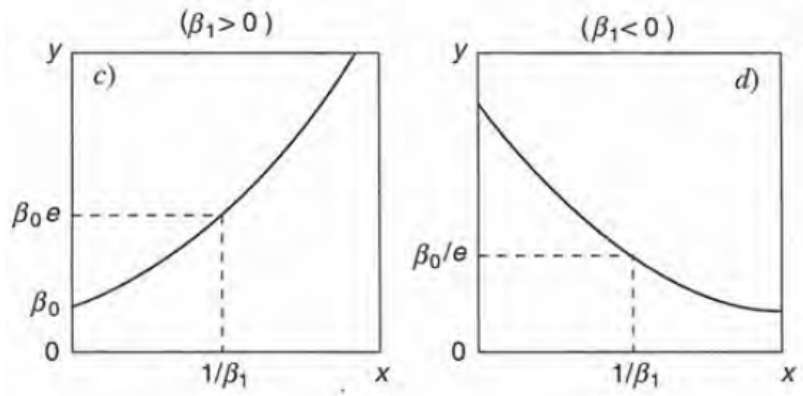


Figura 8:

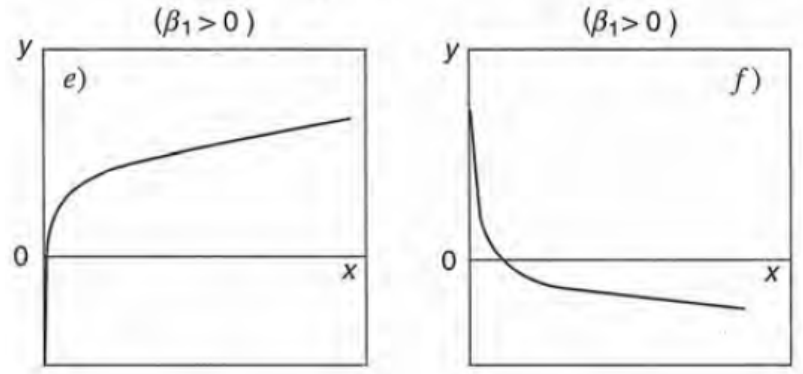


Figura 9:

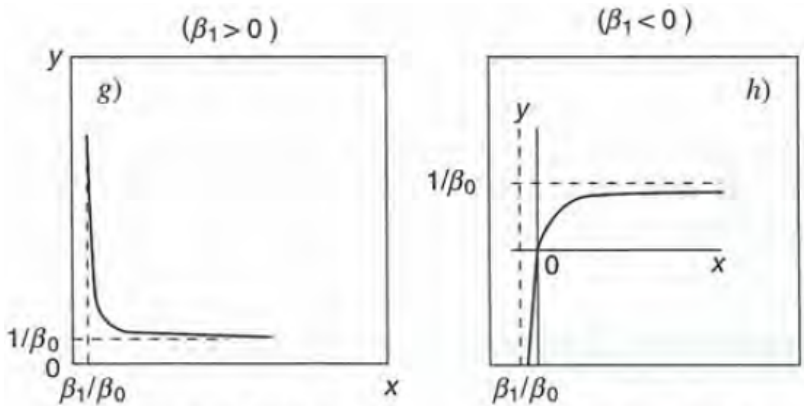


Figura 10:

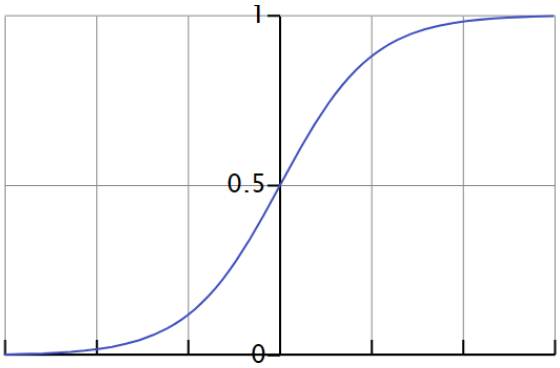


Figura 11:

Cuadro 2: Funciones linealizables.

Figura	Función inicial	Transformación	Función resultado
Fig. 7	$y = \beta_0 x^{\beta_1}$	$y^* = \log y, x^* = \log x$	$y^* = \log \beta_0 + \beta_1 x^*$
Fig. 8	$y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	$y^* = \ln \beta_0 + \beta_1 x$
Fig. 9	$y = \beta_0 + \beta_1 \log x$	$x^* = \log x$	$y^* = \beta_0 + \beta_1 x^*$
Fig. 10	$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y^* = \frac{1}{y}, x^* = \frac{1}{x}$	$y^* = \beta_0 - \beta_1 x^*$
Fig. 11	$y = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$	$y^* = \ln\left(\dfrac{y}{1-y}\right)$	$y^* = \beta_0 + \beta_1 x$

1.10. Ejercicios resueltos

Ejercicio 2.17 en [11] Para el modelo de regresión lineal simple $y = 50 + 10x + \varepsilon$, donde ε es i.i.d como $N(0,16)$, suponer que se usan n = 20 pares de observaciones para ajustar este modelo. Generar 500 muestras de 20 observaciones, tomando una observación para cada valor de x = 1, 1.5, 2, . . . , 10 para cada muestra.

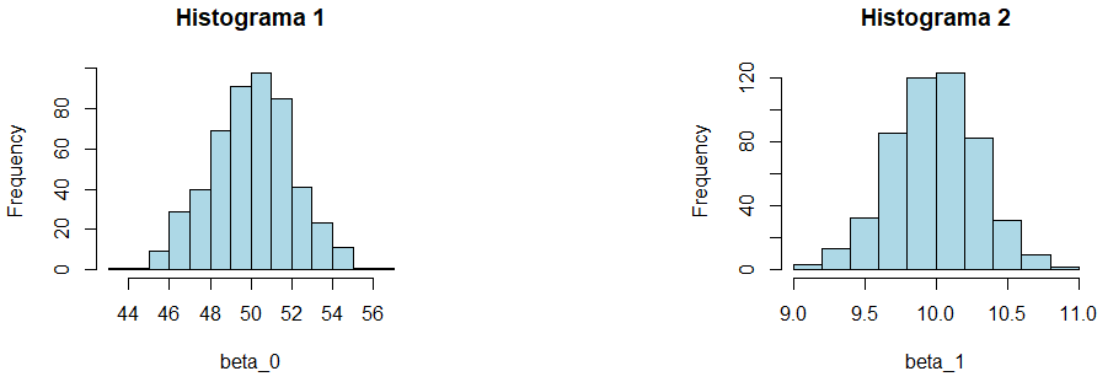
a. Para cada muestra, calcular los estimados de la pendiente y la ordenada al origen por mínimos cuadrados. Trazar histogramas de los valores muestrales de $\hat{\beta}_0$ y $\hat{\beta}_1$. Comentar la forma de esos histogramas.

   Se tienen que crear 500 muestras de 20 valores en x y su valor y correspondiente. Para ello, se crean 3 funciones, la primera añade la aleatoriedad a la ecuación $y = 50 + 10x + \varepsilon$, donde $\varepsilon \sim N(0,16)$ Posterior, una función que arroje una muestra aleatoria de tamaño 2n con los valores (x,y). La última función repite m cantidad de veces la muestra aleatoria de tamaño n.

   Con esta última función creamos una matriz que almacene todos los valores. Posterior, calculamos los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ para cada muestra. Por último, se presentan los histogramas. Note que presentan normalidad con centro en el valor verdadero de $\beta_0 = 50$ y $\beta_1 = 10$, tal y como se demostró en el teorema (2). Se sugiere al lector reescribir en código (en lenguaje R) y correrlo para distintos valores de n y m.

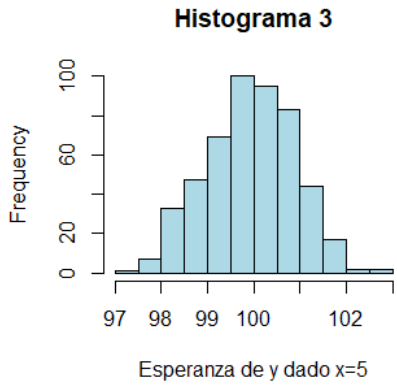
```r
#Creamos la funcion y=50+10x+e
y<-function(x) {
  valor<-50+10*x+rnorm(1,0,4)
  return(valor)
}
#Creamos una funcion que nos arroje una muestra de tamano n*2
#con valores (x,y) usando la funcion anterior
muestra<-function(n){
  df <- data.frame( x = rep(NA, 2*n), y = rep(NA, 2*n))
  for (i in seq(from = 1, to = n+0.5, by = 0.5)) {
   df[2*(i-0.5),2] <- round(y(i),4)
   df[2*(i-0.5),1]<- i
  }
  return(df)
}

#Asi, somos capaces de crear un dataframe con m cantidad de muestras
#de n observaciones.
muestraConjunta<-function(m,n){
  mc<-data.frame()
  mc<-muestra(n)
  for(i in 1:m-1){
    mc<-cbind(mc,muestra(n))
  }
  return(mc)
#Creamos 500 muestras de tamano 10
muestra500<-muestraConjunta(500,10)

# a)
#Obtenemos los coeficientes de las 500 muestras, es decir,
#500 valores de beta_0 y 500 de beta_1
mm1<-lm(muestra500[,2]~muestra500[,1])
coef_estimados<-coefficients(mm1)
for (k in 2:500){
  coef_estimados<-rbind(coef_estimados,
                        coefficients(lm(muestra500[,2*k]~muestra500[,1])))
```

```r
}
#Creamos el histograma de los coeficientes estimados
hist(coef_estimados[,1], xlab="beta_0", main = "Histograma 1",
     col = "lightblue")
hist(coef_estimados[,2], xlab="beta_1", main = "Histograma 2",
     col = "lightblue")
```



b. Para cada muestra, calcular un estimado de $E(y|x=5)$. Trazar un histograma de los estimados obtenidos. Comentar la forma del histograma.

El mejor estimador para $E(y|x=5)$ es, precisamente, $\hat{\beta}_0 + \hat{\beta}_1 \cdot 5$, con valor central en $\beta_0 + \beta_1 \cdot 5$. Conocemos los valores reales de los estimadores, entonces $y = 50 + 10 \cdot 5 = 100$. El histograma 3 presenta la frecuencia de los valores estimados para las 500 muestras con aparente normalidad y centro en el valor 99.93, muy cercano al valor real de 100. Este ejercicio es muy ilustrativo al presentar de manera empírica las propiedades de normalidad y valor esperado de los parámetros estimados y de la esperanza de la variable de respuesta dado un valor en x. Al hacer la comparación de los estimadores (inciso a) y del valor esperado (inciso b) con los valores reales de la ecuación $y = 50 + 10x$ podemos ver una aproximación asertiva, como supusimos de manera teórica.




```r
# b)
#Calculemos la esperanza de y dado x=5 para las 500 muestras
Esp_y_x5<-numeric(length = 500)
for (k in 1:500){
  Esp_y_x5[k]<-coef_estimados[k,1]+coef_estimados[k,2]*5
}
#creamos el histograma
hist(Esp_y_x5, xlab = "Esperanza de y dado x=5",
     main = "Histograma 3", col = "lightblue")
```

c. Determinar un intervalo de confianza de 95 % para la pendiente en cada muestra. ¿Cuántos de los intervalos contienen el valor verdadero $\beta_1 = 10$? ¿Es lo que se esperaba?

Cuando establecemos un intervalo de confianza con un $\alpha = 0.05$ estimamos que el 95 % de las veces nuestro intervalo realmente contiene al verdadero valor del parámetro. Si aquí creamos 500 intervalos de confianza para β_1 en cada muestra, al menos 95 % de ellos deberían contener al valor de 10. Nuevamente, se puede verificar de manera empírica que el 95.2 % de los intervalos creados tiene el valor de 10.

```r
# c)
#Creamos un intervalo de confianza para cada muestra
IC<-data.frame()
IC<-confint((lm(muestra500[,2]~muestra500[,1])),
                      muestra500[4,1], level = 0.95)
for (k in 2:500){
  IC<-rbind(IC,confint((lm(muestra500[,2*k]~muestra500[,1])),
            muestra500[4,1], level = 0.95))
}
IC

#contabilizamos cuantos intervalos contienen al valor 10
booleano_IC10<-numeric(length=500)
for (k in 1:500){
booleano_IC10[k]<-(10 >= IC[k,1] && 10 <= IC[k,2])
}
#calculamos el cociente de los intervalos que contienen al verdadero
#valor 10 de beta_1, entre el total de intervalos
sum(booleano_IC10)/500 #0.952
```

d. Para cada estimado de $E(y|x=5)$ en la parte b, calcular el intervalo de confianza de 95 %. ¿Cuántos de esos intervalos contienen el valor verdadero de $E(y|x=5)=100$? ¿Es lo que se esperaba?

A esta altura, el lector debería ser capaz de resolver este ejercicio sin mayor esfuerzo.

Ejercicio 2.12 en [11] Se cree que la cantidad de libras de vapor usadas en una planta por mes está relacionada con la temperatura ambiente promedio. A continuación se presentan los consumos y las temperaturas del último año.

Mes	Temperatura	Uso/1000
Ene.	21	185.79
Feb.	24	214.47
Mar.	32	288.03
Abr.	47	424.84
Mayo	50	454.68
Jun.	59	539.03
Jul.	68	621.55
Ago.	74	675.06
Sep.	62	562.03
Oct.	50	452.93
Nov.	41	369.95
Dic.	30	273.98

a) Ajustar el modelo de regresión lineal simple a los datos.

Supongamos que la variable y=Uso de vapor depende de x=Temperatura ambiente promedio. Es interesante ver que matemáticamente es igual si invertimos la asignación de las variables. Sin embargo, en el contexto del problema no tiene sentido pensar que la temperatura ambiente depende de la producción de una única planta.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 9.2084$$
$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = -6.3320$$
$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = 3.7854$$

b) Probar la significancia de la regresión.

Para probar la significancia nos ayudaremos de la tabla ANOVA. Note que el valor del estadístico F_0 es tan grande es casi imposible no rechazar la hipótesis nula de no significancia. Esto se verá más claro en el gráfico adjunto al final del ejercicio.

c) En la administración de la planta se cree que un aumento de 1 grado en la temperatura ambiente promedio hace aumentar 10 000 libras en el consumo mensual de vapor. ¿Estos datos respaldan la afirmación?

Recuerde que se hizo un escalamiento sobre mil. De esta manera, contestar a esta pregunta es análogo a la prueba de hipótesis $H_0: \beta_1 = 10 \quad vs \quad H_a: \beta_1 \neq 10$. Supongamos un nivel de

Fuente de Variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	F_0
Regresión	$SS_R = 280589.6$	1	280589.6	74124.16
Residuales	$SS_e = 37.8547$	10	3.7854	
Total	$SS_T = 280627.4$	11		

confianza $\alpha = 0.05$ Podemos utilizar el estadístico t_0

$$t_0 = \frac{\hat{\beta}_1 - 10}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{-0.7916}{\sqrt{3.7854/3309}} = -23.4044$$

mientras que nuestro valor crítico es $t_{n-2}^{\alpha/2} = t_{10}^{0.025} = 1.8124$. Como $|t_0| > t_{10}^{0.025}$, podemos rechazar la hipótesis nula, es decir, no existe suficiente evidencia estadística para afirmar que la taza de cambio $\beta_1 = 10$, tal que al aumentar una unidad el grado de la temperatura promedio no hace aumentar en diez mil libras el consumo mensual de vapor. Este modelo es muy estricto debido al alto nivel de ajuste. Si se calcula el valor $R^2 = 0.999$ se puede ver que existe un ajuste muy preciso. Tal parece que los datos siguen más una ley determinista que un modelo de probabilidad. Incluso si se intentara hacer la misma prueba para un valor $\beta_1 = 9$ también rechazaríamos la hipótesis nula.

d) Determina un intervalo de predicción de 99 % para el uso de vapor en un mes con temperatura ambiente promedio de 58 %.

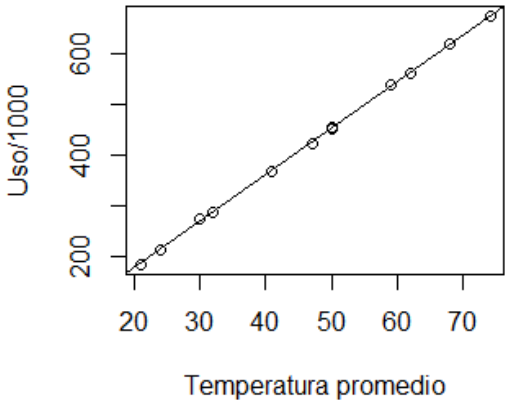
Recuerde que se puede hacer un intervalo de confianza para el valor medio de respuesta dado un valor fijo de x_0 y también un intervalo para un valor fijo de respuesta y_0 dado x_0 igual fijo. Considerando un nivel de significancia del 99 %, entonces $\alpha = 0.01$. Para el segundo caso, el intervalo de confianza es de la forma

$$\left(\hat{y}_0 - t_{(n-2)}^{\alpha/2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}} \leq y_0 \leq \hat{y}_0 + t_{(n-2)}^{\alpha/2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right)$$

$$\left(527.7552 - 3.1692\sqrt{3.7854\left(1+\frac{1}{12}+\frac{132.25}{3309}\right)} \leq y_0 \leq 527.7552 + 3.1692\sqrt{3.7854\left(1+\frac{1}{12}+\frac{132.25}{3309}\right)}\right)$$

$$(521.2200 \leq y_0 \leq 534.2903)$$

Un intervalo estrecho considerando la magnitud de los datos.



Ejemplo 11.9 en [14] En su tesis para obtener el doctorado, H. Behbahani estudió el efecto de la variación de la razón agua/cemento en la resistencia del concreto después de 28 días. Para el concreto que contiene 200 libras por yarda cúbica de cemento obtuvo los datos que se presentan en la siguiente tabla. Sea y la resistencia y x la razón de agua/cemento.

Agua/Cemento	Resistencia
1.21	1.302
1.29	1.231
1.37	1.061
1.46	1.040
1.62	0.803
1.79	0.711

a) Ajuste el modelo $E(y) = \beta_0 + \beta_1 x$.
Apliquemos las fórmulas para obtener los estimadores.

$$S_{xy} = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i = 8.709 - \frac{1}{6}(8.74)(6.148) = -0.247$$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 = 12.965 - \frac{1}{6}(8.74)^2 = 0.234$$

$$S_{yy} = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 = 6.569 - \frac{1}{6}(6.148)^2 = 0.269$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-0.247}{0.234} = -1.056$$

y

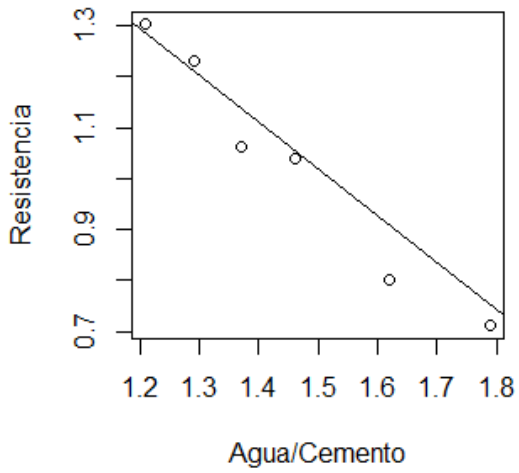
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{6.148}{6} - (-1.056)\left(\frac{8.74}{6}\right) = 2.563$$

(Para este ejemplo se llevaron a cabo los cálculos con tres decimales). Por lo tanto, el modelo de línea recta que mejor ajusta a los datos es

$$\hat{y} = 2.563 - 1.056x$$

Además, calculemos el nivel de ajuste del modelo R^2.

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(-0.247)^2}{(0.234)(0.269)} = 0.969$$



b) Pruebe $H_0: \beta_0 = 0$ frente a $H_a: \beta_1 < 0$ con $\alpha = 0.05$. (Advierta que si rechazamos H_0, concluimos que $\beta_1 < 0$, y que la resistencia tiene a disminuir con un incremento en la razón agua/-cemento). Obtenga el nivel de significancia correspondiente alcanzado.

Como deseamos probar si hay evidencia de que $\beta_1 < 0$ con $\alpha = 0.05$, el estadístico de prueba apropiado es

$$t_0 = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}\sqrt{1/S_{xx}}}$$

donde

$$\hat{\sigma} = \sqrt{\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}} = \sqrt{\frac{0.269 - (-1.056)(-0.247)}{4}} = \sqrt{\frac{0.008}{4}} = 0.045$$

Por lo tanto, el valor del estadístico de prueba apropiado para $H_0: \beta_1 = 0$ frente a $H_a: \beta_1 < 0$ es

$$t_0 = \frac{-1.056 - 0}{0.045\sqrt{1/0.234}} = -11.355$$

Como el estadístico anterior se basa en n-2=4 grados de libertad y la región de rechazo apropiada es $t < -t_{0.05} = -2.132$, rechazamos H_0 a favor de H_a a un nivel de significancia $\alpha = 0.05$. Como la prueba apropiada es de cola inferior, el valor p es $p = P(t < -11.355)$, donde t tiene una distribución con 4 grados de libertad. Por tanto, $p < 0.05$. De hecho, tablas más extensas de la distribución t indican que el valor p es considerablemente menor que 0.005, en realidad es de 0.00017. En consecuencia, para los valores de $\alpha > 0.00017$ que se usan por lo común común concluimos que hay evidencia suficiente para indicar que la resistencia disminuye con un incremento en la razón agua/cemento en la región donde se realizó el experimento. Desde un punto de vista práctico, la razón agua/cemento debe ser suficiente grande para humedecer el cemento, la arena y otros elementos que forman el concreto. Pero si la razón agua/cemento aumenta demasiado, el concreto no servirá.

c) Encuentre un intervalo de confianza al 90 % de la resistencia esperada del concreto cuando la razón agua/cemento es de 1.5. ¿Qué pasará con el intervalo de confianza si tratamos de estimar la resistencia media para razones de agua/cemento de 0.3 o 2.7?

El intervalo de confianza se puede obtener mediante la fórmula

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}}$$

Queremos un intervalo de confianza cuando $x_0 = 1.5$; por tanto, el intervalo es de la forma

$$\left(0.979 - (2.132)(0.045)\sqrt{\frac{1}{6} + \frac{1.5 - 1.457)^2}{0.234}}\ ,\ 0.979 + (2.132)(0.045)\sqrt{\frac{1}{6} + \frac{1.5 - 1.457)^2}{0.234}}\right)$$
$$= (0.938, 1.020)$$

Por tanto, estimaríamos la resistencia media del concreto con una razón agua/cemento de 1.5 entre 0.938 y 1.020. A partir de la expresión de la varianza, podemos ver que el intervalo de confianza se vuelve más grande conforme x_0 se aleja de $\bar{x} = 1.457$. Además, los valores $x_0 = 0.3$ y $x_0 = 2.7$ están lejos de los valores que se utilizaron en el experimento. Hay que ser muy cauteloso antes de construir un intervalo de confianza para $E(y)$ cuando los valores de x_0 se alejan de la región de experimentación. Razones de agua/cemento de 0.3 y 2.7 quizá producirían un concreto completamente inservible.

Ejemplo 11.10 en [14] En los datos de la tabla siguiente, W denota el peso (en libras) y l la longitud (en pulgadas) de la parte posterior de la cabeza a la punta de la nariz de 15 caimanes capturados en Florida. Como es más fácil observar l que W en caimanes en su hábitat natural, se desea construir un modelo que relacione el peso con la longitud. Se puede utilizar tal modelo para predecir el peso de los caimanes de longitudes específicas. Ajuste el modelo

$$\ln W = \ln \alpha_0 + \alpha_1 \ln l + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$$

x=lnL	y=lnW
3.87	4.87
3.61	3.93
4.33	6.46
3.43	3.33
3.81	4.38
3.83	4.7
3.46	3.5
3.76	4.5
3.5	3.58
3.58	3.64
4.19	5.9
3.78	4.43
3.71	4.38
3.73	4.42
3.78	4.25

Comencemos calculando las cantidades que se aplican en forma rutinaria en nuestra solución.

$$S_{xy} = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i = 251.9757 + \frac{1}{15}(56.37)(66.27) = 2.933$$

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 = 212.6933 - \frac{1}{15}(56.37)^2 = 0.8548$$

$$S_{yy} = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 = 303.0409 - \frac{1}{15}(66.27)^2 = 10.26$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{2.933}{0.8548} = 3.4312$$

y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{66.27}{15} - (3.4312)\left(\frac{56.37}{15}\right) = -8.476$$

Ahora podemos estimar α_0 mediante

$$\hat{\alpha}_0 = e^{\hat{\beta}_0} = e^{-8.476} = 0.0002$$

y α_1 con $\hat{\alpha}_1 = \hat{\beta}_1$ para obetener el modelo estimado

$$\hat{w} = \hat{\alpha}_0 l^{\hat{\alpha}_1} = (0.0002) l^{3.4312}$$

En muchos casos α_1 estará cerca de 3, ya que el peso o el volumen generalmente son proporcionales al cubo de una medición lineal.

Para estos datos SSE=0.1963, n=15, $\hat{\sigma} = \sqrt{SSE/(n-2)} = 0.123$. Los cálculos realizados para obtener estos valores numéricos son análogos a los cálculos del ejemplo anterior (11.9 en [14]).

Para encontrar un intervalo de predicción de W donde $x = \ln l = 4$, primero debemos construir un intervalo de predicción para $y = \ln W$. Como antes, el intervalo de predicción es

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{0.05}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

donde $t_{0.05}$ tiene $n-2=13$ grados de libertad. Por lo tanto, $t_{0.05}=1.771$ y el intervalo de predicción al 90 % para $y=\ln W$ es

$$-8.476+3.4312(4)\pm 1.771(0.123)\sqrt{1+\frac{1}{15}+\frac{(4-3.758)^2}{0.8548}}$$

o bien

$$(5.0167, 5.4809)$$

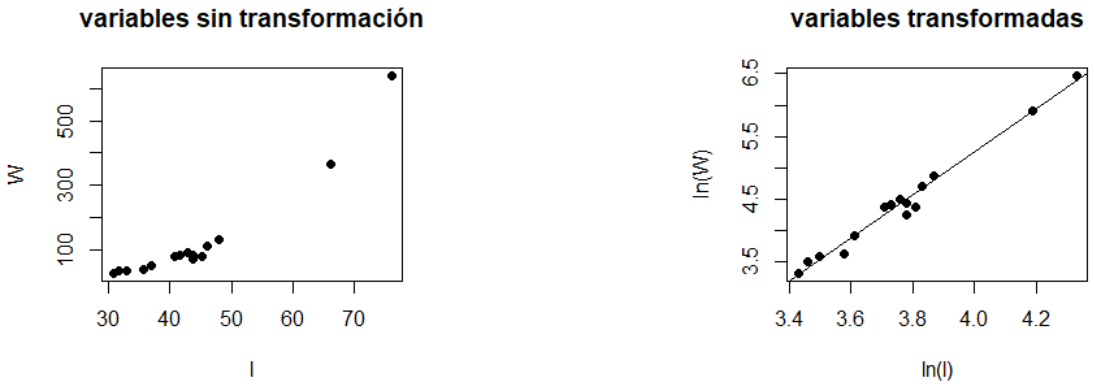
Como $\hat{y}=\ln\hat{W}$, podemos predecir W mediante $e^{\hat{y}}=e^{5.2488}=190.3377$. El intervalo de predicción de 90 % observado para W es

$$(e^{5.0167}, e^{5.4809})$$

o bien

$$(150.9125, 240.0627)$$

Cuando $x=\ln x=4$, entonces $l=e^4=54.598$. Por tanto, para un caimán con 54.598 pulgadas de longitud, predecimos que su peso está entre 150.91 y 240.06 libras. El intervalo hasta cierto punto estrecho en la escala logarítmica natural se vuelve muy grande cunado se transforma a la escala original.



Ejemplo 6.2 en [12] Los estudiantes de la clase de estadística sugieren que hacer tarea no les ha ayudado a prepararse para el examen de medio curso. Los puntajes del examen y y los puntajes de la tarea x para los 18 estudiantes en la clase son los siguientes:

y	x
95	96
80	77
0	0
0	0
79	78
77	64
72	89
66	47
98	90
90	93
0	18
95	86
35	0
50	30
72	59
55	77
75	74
66	67

Primero, obtenemos los coeficientes estimados

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{81195 - 18(58.056)(61.389)}{80199 - 18(58.056)^2} = 0.8726$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 61.389 - (0.8726)(58.056) = 10.73$$

Así, la ecuación está dada por

$$\hat{y} = 10.73 - 0.8726x$$

En el gráfico es sencillo ver que la pendiente $\hat{\beta}_1$ es la tasa de cambio en $\hat{y}$ a la variación en x y el intercepto $\hat{\beta}_0$ es el valor de $\hat{y}$ para $x = 0$.

La línea aparente de tendencia no establece causalidad entre la tarea y los resultados del examen. La asunción de varianza constante $V(\varepsilon_i) = \sigma^2$ para todo $i = 1, 2, \ldots, 18$ parece ser razonable.

¿Por qué calculamos los residuales al cuadrado? Como e_i es un estimador de ε_i y $E(\varepsilon_i) = 0$, se esperaría que el estimador insesgado de ε, llamado $\bar{\varepsilon}$, también se aproxime a cero. En este ejemplo podemos ver como

$$\bar{e} = \frac{\sum e_i}{n} = \frac{1.16 \cdot 10^{-14}}{18} = 6.44 \cdot 10^{-16} \approx 0$$

Por otro lado, calculemos el estimador $\hat{\sigma}^2$ para, posteriormente, obtener pruebas de hipótesis e intervalos de confianza.

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{3071.229}{16} = 191.9518$$

Podemos probar la hipótesis $H_0 : \beta_1 = \beta_1^*$ *vs* $H_a : \beta_1 \neq \beta_1^*$ para cualquier valor de $\beta_1^* \in R$. Sin embargo, suele resultar interesante para $\beta_1^* = 0$. ¿Qué pasaría si $\beta_1^* = \hat{\beta}_1$? Nunca rechazaríamos la hipótesis nula

$H_0 : \beta_1 = \hat{\beta}_1$, pues nuestro estadístico quedaría de la forma

$$t_0 = \frac{\hat{\beta}_1 - \hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = 0.$$

Como la distribución t es centrada en cero, t_0 nunca caerá dentro de la región de rechazo para un $\alpha \neq 0.5$ (que es análogo a una apuesta lanzando una moneda). Entonces, hagamos la prueba de hipótesis para $\beta_1^* = 0$

$$t_0 = \frac{0.8726}{\sqrt{191.9518/19530.94}} = 8.8019$$

Al cual le corresponde un p-value de $1.57 \cdot 10^{-7} \approx 0$. Es decir, rechazamos la hipótesis nula casi para cualquier valor de α. Por lo tanto, nuestro parámetro β_1 no toma el valor cero (la variable x si tiene influencia sobre y de tal manera que es significativa). Vimos, por el consciente de verosimilitudes, que también es posible probar la significancia mediante una prueba F. Esta prueba toma más relevancia en el caso multiparamétrico, donde la hipótesis es sobre la significancia del modelo completo. Por ahora, probaremos el mismo juego de hipótesis $H_0 : \beta_1 = 0$ *vs* $H_a : \beta_1 \neq 0$. El estadístico es

$$F_0 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\hat{\sigma}^2} = 77.4832$$

Con un p-value de $1.57 \cdot 10^{-7}$. Podemos concluir lo mismo que en la prueba t antes realizada.

Ahora, creemos intervalos de confianza para los parámetros β_0, β_1 y σ^2 con un $\alpha = 0.1$.Recordemos que el primer intervalo es de la forma

$$\left(\hat{\beta}_0 - t_{n-2}^{\alpha/2} \frac{\hat{\sigma}\sqrt{\sum x_i^2}}{\sqrt{nS_{xx}}} < \beta_0 < \hat{\beta}_0 + t_{n-2}^{\alpha/2} \frac{\hat{\sigma}\sqrt{\sum x_i^2}}{\sqrt{nS_{xx}}} \right)$$

Entonces

$$\left(10.73 - 1.7458 \frac{\sqrt{(191.9518)(80199)}}{\sqrt{18(19530.94)}} < \beta_0 < 10.73 + 1.7458 \frac{\sqrt{(191.9518)(80199)}}{\sqrt{18(19530.94)}} \right)$$
$$(-0.8261 < \beta_0 < 22.2799)$$

Para β_1

$$\left(\hat{\beta}_1 - t_{n-2}^{\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{n-2}^{\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \right)$$
$$\left(0.8726 - 1.7458 \sqrt{\frac{191.9518}{19530.94}} < \beta_1 < 0.8726 + 1.7458 \sqrt{\frac{191.9518}{19530.94}} \right)$$
$$(0.6995 < \beta_1 < 1.0456)$$

Para σ^2, utilizando una distribución de cola superior para chi-cuadrada

$$\left(\frac{\hat{\sigma}^2(n-2)}{\chi^2_{\alpha/2,n-2}} < \sigma^2 < \frac{\hat{\sigma}^2(n-2)}{\chi^2_{1-\alpha/2,n-2}}\right)$$
$$\left(\frac{191.9518(16)}{26.2962} < \sigma^2 < \frac{191.9518(16)}{7.9616}\right)$$
$$(116.7969 < \sigma^2 < 385.7552)$$

Recordemos, de igual manera, los intervalos para la respuesta esperada de y dado un valor fijo de x y, casi igual, un valor puntual de y dado x. Supongamos que nos interesa saber el valor esperado del puntaje en el examen cuando los alumnos obtienen 80 puntos en la tarea $x = x_0 = 80$ a un $\alpha = 0.05$. Primero, vea que $\hat{y}_0 = 10.7269 + 0.8726(80) = 80.5349$. Entonces

$$\left(\hat{y}_0 - t^{\alpha/2}_{(n-2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \leq E(\hat{y}_0|x_0) \leq \hat{y}_0 + t^{\alpha/2}_{(n-2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right)$$
$$[80.5349 - 2.1199\sqrt{(191.9518)\left(\frac{1}{16} + \frac{(80-58.0555)^2}{19530.94}\right)} \leq E(\hat{y}_0|x_0) \leq$$
$$80.5349 + 2.1199\sqrt{(191.9518)\left(\frac{1}{16} + \frac{(80-58.0555)^2}{19530.94}\right)}]$$

Esto es

$$(71.8640 \leq E(\hat{y}_0|x_0) \leq 89.5057)$$

Recuerde que con forme x_0 se aleja del centro de los datos $\bar{x}$, el intervalo se hace más grande y menos preciso.

De manera muy similar calculamos el intervalo para el valor puntual y_0

$$\left(\hat{y}_0 - t^{\alpha/2}_{(n-2)}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \leq y_0 \leq \hat{y}_0 + t^{\alpha/2}_{(n-2)}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}}\right)$$
$$(40.3769 \leq y_0 \leq 120.6928)$$

Note como la imprecisión aumenta aún más si queremos estimar un valor puntual en y.

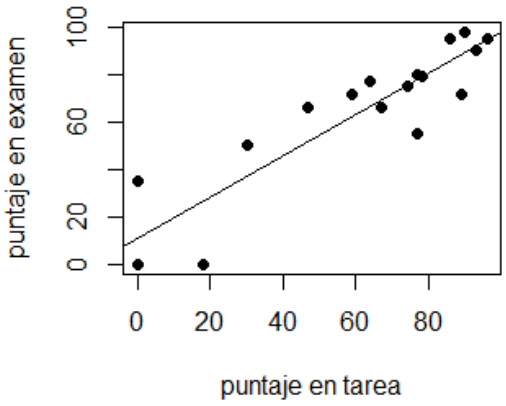
Casi por último, resulta útil medir el nivel de ajuste de nuestro modelo mediante R^2.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{14873.05}{17944.28} = 0.8288$$

Un nivel alto considerando que el tope es la unidad. Esto nos dice que gran parte de la variabilidad del modelo está siendo explicado por la línea estimada.

Por último, si x es una variable aleatoria, es posible calcular el coeficiente de correlación lineal entre x e y.

$$\hat{\beta}_1 = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}r_{xy} \Rightarrow r_{xy} = \hat{\beta}_1\frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} = (0.8726)\sqrt{\frac{19530.94}{17944.28}} = 0.9110$$



Ejercicio 6.14 en [12] La siguiente tabla (Weisberg 1985, p. 231) proporciona los datos de las erupciones diurnas del géiser Old Faithful en el parque nacional de Yellowstone entre el 1 y el 4 de agosto de 1978. Las variables x=duración de la erupción e y=intervalo hasta la siguiente erupción. ¿x puede ser usado para predecir satisfactoriamente y a través de un modelo de regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$?

x	y	x	y
4.4	78	4.5	76
3.9	74	3.9	82
4	68	4.3	84
4	76	2.3	53
3.5	80	3.8	86
4.1	84	1.9	51
2.3	50	4.6	85
4.7	93	1.8	45
1.7	55	4.7	88
4.9	76	1.8	51
1.7	58	4.6	80
4.6	74	1.9	49
3.4	75	3.5	82
4.3	80	4	75
1.7	56	3.7	73
3.9	80	3.7	67
3.7	69	4.3	68
3.1	57	3.6	86
4	90	3.8	72
1.8	42	3.8	75
4.1	91	3.8	75
1.8	51	2.5	66
3.2	79	4.5	84
1.9	53	4.1	70
4.6	82	3.7	79
2	51	3.8	60
		3.4	86

a) Encuentre $\hat{\beta}_1$ y $\hat{\beta}_0$.

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i \sum y_i\right)}{\sum x_i^2 - \left(\frac{1}{n}\sum x_i\right)^2} = \frac{601.1509}{52.8818} = 11.3678$$

Por consiguiente

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = 71.1320 - 3.4641(11.3678) = 31.7523$$

b) Haga la prueba de significancia $H_0: \ \beta_1 = 0$.
Estableciendo $H_a: \ \beta_1 \neq 0$. El estadístico de prueba es

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{11.3678}{\sqrt{55.3786/52.8818}} = 11.1085$$

Cuyo p-value es casi cero. Es decir, rechazamos la hipótesis nula para casi cualquier valor de α, lo que nos lleva a la afirmación de que existe suficiente evidencia estadística para suponer que el parámetro β_1 es significativo (distinto de cero). Esta muestra es lo suficiente grande para hacer una aproximación normal, de tal manera que llegaríamos a la misma conclusión.

c) Encuentre un intervalo de confianza para β_1.
Recuerde que el intervalo es de la forma

$$\left(\hat{\beta}_1 - t_{n-2}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} < \beta_1 < \hat{\beta}_1 + t_{n-2}^{\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right)$$

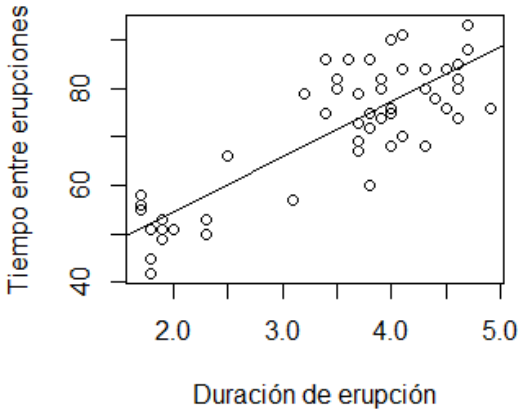
Utilizando un $\alpha = 0.05$, entonces $t_{n-2}^{\alpha/2} = 2.0075$. Además, ya sabemos que $\hat{\sigma}^2 = 55.3786$ y $S_{xx} = 52.8818$. De esta manera, nuestro intervalo es

$$\left(11.3678 - (2.0075)\sqrt{\frac{55.3786}{52.8818}} < \beta_1 < 11.3678 + (2.0075)\sqrt{\frac{55.3786}{52.8818}}\right)$$

$$(9.3134 < \beta_1 < 13.4221)$$

d) Encuentre R^2.
Por último

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{6833.766}{9658.075} = 0.7075$$



2. Regresión lineal múltiple

2.1. Modelo Lineal General

El tema de Modelo lineal General (LM) involucra los modelos de Regresión, Análisis de varianza (ANOVA) y los de Análisis de Covarianza (ANCOVA). Estos modelos tienen supuestos que deben de comprobarse una vez ajustado el modelo, los cuales son:

i) Independencia entre las y_i por lo que también lo serán los residuales del modelo.

ii) Linealidad entre la variable respuesta con respecto a la(s) variable(s) explicativa(s).

iii) Normalidad en las residuos, y que cumplan que su media sea cero y σ^2 constante para todas las observaciones.

iv) Homocedasticidad. Las varianzas tienen que ser homogéneas en los diferentes factores e iguales.

2.1.1. Modelo de Regresión Lineal

Sea Y la variable respuesta (variable dependiente) que está relacionada con las p variables explicativas (variables independientes) $X_1, X_2, \ldots, X_p$ por una función f. Tanto la variable respuesta como las explicativas son continuas y debido a que esta relación no es exacta, se escribe de la siguiente forma.

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon \tag{34}$$

donde ε es un error aleatorio.
Las Y y las X's se observan sobre n individuos. Cuando f es lineal, la ecuación (34) observada en un individuo con i=1,2,...,n se denota como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{35}$$

y se llama **modelo de regresión lineal**; los parámetros $\beta_j, j = 0, 1, \ldots, p$, se llaman **coeficientes de regresión**, los cuales se interesan estimar.
Escribiendo el modelo lineal (35) en forma de matricial, tenemos que con $i = 1, 2, ..., n.$ y $j = 0, 1, ..., p.$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{i1} & x_{i2} & x_{ij} & x_{ip} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Y una forma más compacta es

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{36}$$

donde $\mathbf{Y}$ es un vector aleatorio de dimensión $n\times 1$,
$\mathbf{X}$ es llamada la *matriz de diseño* de $n\times(p+1)$ no aleatoria,
β es un vector de parámetros desconocidos $(p+1)\times 1$,
ε es vector de errores aleatorio de dimensión $n\times 1$.
Los supuestos sobre el vector de errores se expresan como

$$E(\varepsilon)=0 \;\; y \;\; Var(\varepsilon)=\sigma^2 I,$$

donde I es la matriz identidad de dimensión (n x n), los errores tienen varianza constante y no están correlacionadas. La esperanza de $\mathbf{Y}$ es

$$E(\mathbf{Y})=E(\mathbf{X}\beta+\varepsilon)=E(\mathbf{X}\beta)+E(\varepsilon)=E(\mathbf{X}\beta)=\mathbf{X}\beta$$

y la varianza de $\mathbf{Y}$ es

$$Var(\mathbf{Y})=Var(\mathbf{X}\beta+\varepsilon)=Var(\mathbf{X}\beta)+Var(\varepsilon)=Var(\varepsilon)=\sigma^2 I.$$

La función para realizar una regresión lineal en el paquete R es lm ("modelo lineal"), en [3] se detalla el uso de la función.

2.1.2. Análisis de Varianza (*ANOVA*)

En la modelación estadística se desea conocer el efecto de una o más variables explicativas sobre una respuesta (continua). Las variables explicativas que pueden ser controladas en un experimento reciben el nombre de ***factores*** y el nivel de intensidad de un factor se le denomina ***nivel*** del factor. Si existe un solo factor, a los niveles de ese factor se le llaman tratamientos (T) [3], [8].
Cuando la(s) variable(s) explicativa(s) son categóricas en vez de continuas y se desea hacer comparaciones entre los niveles del factor entonces se tiene un problema de análisis de varianza (*ANOVA*).
Cuando se modela un solo factor, la expresión:

$$Y \quad = \quad \beta_0+\beta_1 X_1+\varepsilon, \tag{37}$$

se puede escribir como:

$$Y \quad = \quad \mu+\beta\, factor+\varepsilon. \tag{38}$$

Los valores tanto de Y como del factor se observan sobre los n individuos.

En el *ANOVA* implica el cálculo de la variación total en la variable respuesta y la partición de ella en componentes informativos. Para el caso univariado, dividimos la variación total (SSY) en sólo dos componentes: la variación explicada (SSA) y la variación no explicada (SSE), es decir SSY = SSA+ SSE. El Cuadro 3 expresa a detalle la información de las sumas de cuadrados.

Supongamos que hay m repeticiones de cada tratamiento y existen k niveles en el factor. Para identificar **los grados de libertad de SSE**, se tiene que hay m repeticiones de cada tratamiento por lo que se necesitan $(m-1)$ parámetros para estimarlos y como hay k niveles de cada factor, se concluye que hay k x $(m-1)$ grados de libertad para el error en el experimento.

La componente de variación entre los tratamientos se representa en SSA, la suma de cuadrados de los tratamientos, explica las diferencias entre las medias de los tratamientos en promedio. Cuando se tienen dos o más variables categóricas independientes, la SSB denotará la suma de cuadrados que se atribuye a las diferencias entre las medias del segundo factor, SSC se denotará la suma de cuadrados que se atribuye al

tercer factor, así sucesivamente hasta tener todos los factores incluídos.
Generalmente esta información se presenta en la tabla de Análisis de varianza llamada Anova Tabla [38], ésta tiene seis columnas, de izquierda a derecha se describen sus características, la primera columna es la fuente de variación, le sigue la suma de cuadrados que depende de la fuente, le sigue los grados de libertad para la fuente, cuadrados medios de la fuente (varianza) y la proporción F donde se esta probando una hipótesis, la cual es

H_0: la fuente de variación no es significativamente distinta de cero vs H_1: $\neg H_0$),
el valor de p asociado a F (si $p < 0.05$ se rechazo H_0). Los cuadrados medios se obtienen haciendo el cociente de la suma de cuadrados con sus respectivos grados de libertad. La varianza del error, s^2 también es llamada cuadrado medio residual o cuadrado medio de la variación no explicada o varianza del error no agrupado debido a se calcula por medio de todos los tratamientos.

Cuadro 3: Sumas de Cuadrados ANOVA unifactorial.

La suma total de cuadrados, SSY, es la suma de los cuadrados de las diferencias entre los puntos de datos, y_{ij}, y la media general, $\bar{\bar{y}}$.

$$SSY = \sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \bar{\bar{y}})^2$$

donde $\sum_{j=1}^{m} y_{ij}$ significa la suma sobre las m repeticiones dentro de cada uno de los k niveles de factor.
La suma de cuadrados del error, SSE, es la suma de los cuadrados de las diferencias entre los puntos de datos, y_{ij}, y sus medias de los tratamiento individuales, $\bar{y}_i$

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{m} (y_{ij} - \bar{y}_i)^2.$$

La suma de los cuadrados de tratamiento, SSR, es la suma de los cuadrados de las diferencias entre el tratamiento individual, $\bar{y}_i$ y la media general, $\bar{\bar{y}}$

$$SSR = \sum_{i=1}^{k} \sum_{j=1}^{m} (\bar{y}_i - \bar{\bar{y}})^2 = m \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}})^2.$$

Elevando el término al cuadrado en paréntesis y aplicando la suma nos da

$$m \sum_{i=1}^{k} (\bar{y}_i - \bar{\bar{y}}_i)^2 = \sum \bar{y}_i^2 - 2\bar{\bar{y}} \sum \bar{y}_i + k\bar{\bar{y}}^2.$$

Denotando el total de todos los valores de la variable respuesta $\sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij} = \sum y$;
Ahora reemplazamos $\bar{y}_i$ por T_i/m (donde T es el nombre convencional para los k tratamientos totales individuales) y reemplazando $\bar{\bar{y}}$ por $\sum y/km$ se obtiene

$$\frac{\sum_{i=1}^{k} T_i^2}{m^2} - 2\frac{\sum y \sum_{i=1}^{k} T_i}{mkm} + k\frac{\sum y \sum y}{kmkm}.$$

Note que $\sum_{i=1}^{k} T_i = \sum_{j=1}^{m} y_{ij}$, Por lo que los términos de la derecha positivos y negativos ambos tienen la forma $(\sum y)^2/km^2$. Finalmente, multiplicando por m se tiene

$$SSR = \frac{\sum_{i=1}^{k} T_i^2}{m} - \frac{(\sum y)^2}{km}.$$

Se puede probar que $SSY = SSR + SSE$.

Información obtenida de [11].

El cálculo de las sumas de cuadrados se presenta de manera tradicional como el Cuadro 4. Hay seis columnas que indican, de izquierda a derecha, la fuente de variación, la suma de cuadrados atribuibles a esa fuente, los grados de libertad para esa fuente, la varianza para esa fuente (tradicionalmente llamados el cuadrado medio en lugar de la varianza), el estadístico F y el valor p asociado con el valor F.

Cuadro 4: ANOVA.

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p valor $Pr(>F)$
Tratamiento	SSR	k	$\frac{SSR}{k}$	$\frac{SSR/k}{SSE/m-k-1}$	p
Error	SSE	$m-k-1$	$\frac{SSE}{m-k-1}$		
Total	SSY	$m-1$			

Los cuadrados medios se obtienen simplemente dividiendo cada suma de los cuadrados por sus respectivos grados de libertad (en la misma fila). La varianza del error, s^2, es el cuadrado medio residual (el cuadrado medio de la variación no explicada).

2.1.3. Análisis de Covarianza (*ANCOVA*)

El análisis de covarianza (*ANCOVA*) combina elementos de la regresión y el análisis de varianza. La variable de respuesta es continua, y hay al menos una variable explicativa continua (covariables) y con al menos una variable explicativa categórica. Según [3] el *ANCOVA* se resumen con las siguientes relaciones:

- Colocar dos o más regresiones lineales de Y contra X (uno para cada nivel del factor).
- Estimar diferentes pendientes e interceptos para cada nivel.
- Usar la simplificación del modelo (las pruebas de eliminación) para los parámetros innecesarios.

Los modelos antes descritos, se pueden resumir en el Cuadro 5.

Cuadro 5: Variables explicatorias de los modelos de Regresión, ANOVA y ANCOVA.

MODELO	VARIABLE RESPUESTA	VARIABLES EXPLICATIVAS
Regresión	Continua	Continua
ANOVA	Continua	Categórica
ANCOVA	Continua	Continuas y Categóricas

2.2. Métodos de estimación

A continuación se explican distintos métodos de estimación que ayudan a estimar los parámetros β y σ^2 de los modelos (36) y (37). En el contexto de los LM clásicos, el método de estimación más común es el método de Mínimos Cuadrados Ordinarios.

2.2.1. Mínimos Cuadrados Ordinarios (OLS)

Los valores de β son desconocidos, pero se pueden estimar, utilizando los datos de la muestra. Para estimarlos se usa el método **"Mínimos Cuadrados Ordinarios"** [11], que en inglés es Ordinary Least Squares (*OLS*).
Cuando la matriz $\mathbf{X}$ tiene rango completo p, el estimador *OLS* se obtiene minimizando la suma de cuadrados de los residuos, donde el i-ésimo residuo es la diferencia entre el valor observado y_i y el ajustado $\hat{y}_i$.
Los residuos o residuales se pueden escribir en forma matricial como sigue:

$$e_{(n\times 1)} = \mathbf{Y} - \hat{\mathbf{Y}}. \tag{39}$$

La suma de cuadrados de los residuos es:

$$\begin{aligned} S(\hat{\beta}) &= \sum_{i=1}^{n} e_i^2 \\ &= e'e \\ &= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}. \end{aligned} \tag{40}$$

Observemos que $\hat{\beta}'\mathbf{X}'\mathbf{Y}$ es una matriz de dimensión 1×1, es decir, un escalar, y que su transpuesta $(\hat{\beta}'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\hat{\beta}$, es el mismo escalar. Los estimadores de mínimos cuadrados deben satisfacer

$$\frac{\partial S}{\partial \hat{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0, \tag{41}$$

que se simplifica en

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y}. \tag{42}$$

El sistema lineal de (42) se denominan **ecuaciones normales de mínimos cuadrados** [6]. Para resolver las ecuaciones normales se multiplican ambos lados de (42) por la inversa de $\mathbf{X}'\mathbf{X}$. El estimador $\hat{\beta}$ por mínimos cuadrados es un vector de dimensión $p\times 1$ cuya expresión es

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{43}$$

Los estimadores mínimos cuadrados son insesgados y tienen matriz de varianza y covarianza $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, como se muestra a continuación.

$$\begin{aligned} E(\hat{\beta}_{OLS}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \beta. \end{aligned} \tag{44}$$

$$
\begin{aligned}
Cov(\hat{\beta}_{OLS}) &= Cov[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Cov[\mathbf{Y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2 I(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned} \tag{45}
$$

Se puede demostrar que el estimador *OLS* es el mejor estimador lineal insesgado, que en inglés se escribe Best Linear Unbiased Estimaton (*BLUE*) de los parámetros del modelo (36). Esto significa que, de entre los estimadores que son insesgados y lineales con respecto a las observaciones, el estimador *OLS* tiene la menor varianza (teorema Gauss-Markov, [6]). Sin embargo, esto es sólo cuando los supuestos (varianza constante y no correlación) en los residuos se mantengan.

2.2.2. Máxima Verosimilitud (ML)

El método de Máxima Verosimilitud, que en inglés se escribe Maximum Likelihood (*ML*) es un método alternativo para estimar los parámetros en (36), suponiendo que los errores son independientes e idénticamente distribuidos según la normal con varianza constante igual a σ^2, esto es $N(0,\sigma^2 I)$ [11] y [5].
El método *ML* inicia con la función de densidad de los errores

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2}, \quad i = 1,\ldots,n.$$

La función de verosimilitud es:

$$L(\varepsilon,\beta,\sigma^2) = \prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\varepsilon_i'\varepsilon}, \tag{46}$$

de (36) tenemos que $\varepsilon = \mathbf{Y} - \mathbf{X}\beta$. Así (46) se transforma en

$$L(\varepsilon,\beta,\sigma^2) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)}. \tag{47}$$

Ahora la función log-verosimilitud es:

$$
\begin{aligned}
l &= \log L(\varepsilon,\beta,\sigma^2). \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta).
\end{aligned} \tag{48}
$$

Para un valor fijo de σ, la función log$-$verosimilitud se maximiza cuando se minimiza el término $(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)$.
Por lo tanto, el estimador ML de β bajo los errores normales equivale al estimador de mínimos cuadrados $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ [6]. El estimador de máxima verosimilitud de σ^2 es

$$\hat{\sigma}^2_{ML} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\beta}_{OLS})'(\mathbf{Y}-\mathbf{X}\hat{\beta}_{OLS})}{n}. \tag{49}$$

donde $\hat{\sigma}^2_{ML}$ es un estimador sesgado, pero asintóticamente insesgado (consistencia, ver más en Apéndice 9.1).

2.2.3. Máxima Verosimilitud Restringida (*REML*)

El método de Máxima Verosimilitud Restringida, que en inglés se escribe Restricted maximum Likelihood (*REML*) soluciona el problema del sesgo en la estimación de σ^2.
El *REML*, consiste en aplicar el método de máxima verosimilitud a un vector $K'Y$ en vez de aplicarlo al vector de observaciones originales $\mathbf{Y}$. La matriz K se define de manera que se elimina del vector $\mathbf{Y}$ toda la variación que se explica por la matriz $\mathbf{X}$ del modelo. Una diferencia importante entre los vectores $\mathbf{Y}$ y $K'\mathbf{Y}$ es que la longitud de $K'\mathbf{Y}$ es $n-p$ [10]. Por lo tanto, un ajuste *ML* de un modelo lineal de n observaciones ofrece un estimador de la varianza residual con n en el denominador, mientras que el estimador correspondiente para el vector $K'Y$ ofrece un estimador con $(n-p)$ en el denominador.
Surge la cuestión de cómo encontrar una matriz K tal que elimine toda esa variación de $\mathbf{Y}$ que puede ser explicada por $\mathbf{X}$. La condición clave para la eliminación de toda la variación explicada por $\mathbf{X}$ es definir cada columna de la matriz K, denotada por $k_1,...,k_{n-p}$, tal que $k_i'\mathbf{X}=0$ para $i=1,2,3,\ldots,n-p$. En [17] hay un resultado matricial que establece que el número máximo de columnas linealmente independientes que cumplan la condición anterior es $n-p$, esto es la diferencia entre el número de filas y columnas de la matriz de modelo (la matriz $\mathbf{X}$ tiene rango completo). Por lo tanto, solo necesitamos encontrar un número máximo de tales vectores linealmente independientes y apilarlos en la matriz K. Una forma de encontrar al vector k es utilizar la propuesta en [9].

$$k' = c'[I - XX^-], \tag{50}$$

donde c' es un vector arbitrario de longitud n y X^- es una matriz inversa generalizada de $\mathbf{X}$.
Hay que tener en cuenta que puede haber varios valores posibles de X^-. Sin embargo, las estimaciones finales *REML* no se ven afectados por la elección de X^-.
Una vez que la matriz K se encuentra, vemos que multiplicando el modelo.

$$Y = X\beta + \varepsilon,$$

por K', por la izquierda, se obtiene

$$K'Y = K'X\beta + K'\varepsilon. \tag{51}$$

Por construcción de la matriz K, es decir, $K'\mathbf{X}=0$, obtenemos

$$K'\mathbf{Y} = K'\varepsilon.$$

Si $Var(\varepsilon)=V$, entonces $Var(K'\varepsilon)=K'VK$.
En forma más general, si $\mathbf{Y}\sim N(\mathbf{X}\beta, V)$,
entonces

$$K'\mathbf{Y} \sim N(0, K'VK). \tag{52}$$

Así

$$\hat{\beta} = (\mathbf{X}'\hat{V}\mathbf{X})^{-1}\mathbf{X}'\hat{V}\mathbf{Y}. \tag{53}$$

Para (52), la función de verosimilitud se transforma en

$$L(\sigma^2) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)}. \tag{54}$$

Ahora la función log-verosimilitud es:

$$l = \log L(\sigma^2) \tag{55}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log K'VK - \frac{1}{2K'VK}(K'Y-0)'(K'Y-0). \tag{56}$$

tomando $V = \sigma^2 I$, la función log-verosimilitud se transforma en :

$$l = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\sigma^{2(n-p)}(K'K) - \tag{57}$$

$$\frac{1}{2(n-p)}(K'K)Y'K(K'K)^{-1}K'Y. \tag{58}$$

Para estimar σ^2, diferenciamos el log- verosimilitud con respecto a σ^2

$$\frac{\partial l}{\partial \sigma^2} = \frac{n-p}{2\sigma^2} + \frac{1}{2\sigma^4}Y'K(K'K)^{-1}K'Y. \tag{59}$$

igualando a cero, y despejando σ^2. Así tenemos que el estimador *REML* de σ^2 es:

$$\hat{\sigma}^2_{REML} = \frac{1}{n-p}Y'K(K'K)^{-1}K'Y. \tag{60}$$

Por lo tanto, el modelo se puede ajustar utilizando una verosimilitud restringida basado en la normalidad de $K'Y$. Una diferencia esencial entre estas dos verosimilitudes, además de las longitudes de Y y $K'Y$, es que la verosimilitud *REML* no involucra a $X\beta$. Por lo tanto, REML se puede utilizar solo para la estimación de parámetros relacionados con V. Mientras que el vector de parámetros de β puede ser estimado usando el estimador GLS que se explica a continuación.

2.2.4. Mínimos Cuadrados Generalizados (*GLS*)

Si las suposiciones sobre la varianza constante y correlación cero entre los residuales no se cumplen, el estimador *OLS* del LM sigue siendo insesgado. Sin embargo, ya no es un estimador de mínima varianza. Ahora se puede utilizar el estimador por mínimos cuadrados generalizados, que en inglés se escribe Generalized Least Squares (*GLS*).
Como $\mathbf{V}$ es la matriz de varianza y covarianza del error, es decir, en forma más general $Var(\varepsilon) = \sigma^2\mathbf{V}$, el estimador *GLS* de β minimiza la suma de cuadrados $(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\beta)$. Así el estimador *GLS* es

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}. \tag{61}$$

Para el modelo lineal con $Var(\varepsilon) = \sigma^2\mathbf{V}$, este estimador es el *BLUE*, es decir, tiene la variación más pequeña de entre todos las posibles estimadores insesgados (generalización del teorema de Gauss-Markov).
El estimador *GLS* es insesgado y tiene matriz de varianza covarianza $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}^{-1}$

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\beta \\ &= \beta, \end{aligned}$$

$$
\begin{aligned}
Cov(\hat{\beta}) &= Cov((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Cov(\mathbf{Y})((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})' \\
&= \sigma^2((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}) \\
&= \sigma^2((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}) \\
&= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.
\end{aligned}
$$

El estimador *OLS* se puede derivar de los métodos de estimación anteriores, sustituyendo $\sigma^2\mathbf{V}$ por $\sigma^2 I$ [10].

2.3. Bondad de ajuste del modelo

En esta sección se presentan distintas maneras de evaluar la bondad de ajuste del LM, lo que se refiere al grado en que éste es conveniente como modelo que representa a las variables implicadas en él. Con el uso del software R, anteriormente se comentó que el ajuste de los Modelos Lineales se realizan usando la función lm(), mientras que para ver la bondad del ajuste se usa la función summary() [3]. Esta función da el resumen del modelo ajustado, permitiendo observar las distintas pruebas de la bondad de ajuste del modelo, estas pruebas se describen a continuación.

2.3.1. Prueba F para el ajuste del modelo

La prueba *F* del modelo nos permite determinar estadísticamente si las variables explicativas (en conjunto) tienen efecto o no sobre la variable respuesta. Este procedimiento suele considerarse como una prueba general o global del ajuste del modelo. Las prueba de hipótesis es:

$$
\begin{gathered}
H_0 : \beta_0 = \beta_1 = \ldots = \beta_p = 0 \\
\text{vs} \\
H_a : \beta_j \neq 0 \ \ al \ \ menos \ \ para \ \ un \ \ j \ \ (j = 1, \ldots, p).
\end{gathered}
$$

El rechazo de la hipótesis nula implica que al menos uno de los regresores contribuye al modelo en forma significativa.
Aunque en el Cuadro 4 desarrollamos como elaborar una ANOVA que muestra el cálculo del estadístico *F* para esta prueba, creemos conveniente repetirlo con la notación comúnmente utilizada en los modelos de regresión.
Este método consiste en una partición de la variabilidad total de la variable *y* de respuesta. Para obtener esta partición se comienza con la identidad

$$
y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i). \tag{62}
$$

Ahora procedemos a elevar al cuadrado en ambos lados de la ecuación (62), y se suman para todas las n observaciones. Así se obtiene la siguiente ecuación:

$$
\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{63}
$$

Por otro lado,

$$
\begin{aligned}
2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) &= 2\sum_{i=1}^{n}\hat{y}_i(y_i-\hat{y}_i)-2\sum_{i=1}^{n}\bar{y}(y_i-\hat{y}_i)\\
&= 2\sum_{i=1}^{n}\hat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n} e_i = 0.
\end{aligned}
\tag{64}
$$

Ya que la suma de los residuales siempre es igual a cero y la suma de los residuales ponderados por el valor ajustado $(\bar{y}-y_i)$ correspondiente también es igual a cero, la ecuación (63), se reduce a

$$
\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + \sum_{i=1}^{n}(y_i-\hat{y}_i)^2. \tag{65}
$$

Esta igualdad nos dice que **suma total de cuadrados** que se denota por SS_T se puede escribir como una **suma de cuadrados debido a la regresión**, SS_R, y una **suma de cuadrados de residuales**, SS_{Res}.

$$
SS_T = SS_R + SS_{Res}. \tag{66}
$$

En [11] se demuestra que $\frac{SS_R}{\sigma_2}$ tiene una distribución χ^2_p, con el mismo número de grados de libertad que la cantidad de variables regresoras, $\frac{SS_{Res}}{\sigma_2} \sim \chi^2_{n-p-1}$ y que SS_R y SS_{Res} son independientes.
Tomando a

$$
F_0 = \frac{SS_R/p}{SS_{Res}/(n-p-1)} = \frac{MS_R}{MS_{Res}}
$$

este tiene una distribución $F_{p,n-p-1}$ y rechazamos H_0 si

$$
F_0 > F_{\alpha,p,n-p-1}.
$$

El procedimiento se resume en el Cuadro 6 de análisis de varianza de la regresión.

Cuadro 6: Análisis de varianza de la regresión.

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_O	P valor $0<P<1$
Regresión	SS_R	p	MS_R	$\frac{MS_R}{MS_{Res}}$	p
Residuales	SS_{Res}	$n-p-1$	MS_{Res}		
Total	SS_T	$n-1$			

citado en [11].

2.3.2. Coeficiente de determinación (R^2) y ajustado (R_{Adj})

El coeficiente de determinación nos permite expresar la cantidad de la variabilidad presente en las observaciones de Y, que se explica mediante el modelo LM. La cantidad

$$
R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \tag{67}
$$

se llama **coeficiente de determinación**, R^2 indica la proporción de variabilidad explicada por la regresión. Ya que $0 \leq SS_{Res} \leq SS_T$, entonces $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte

de la variabilidad de Y está explicada por el modelo de regresión. A medida que el coeficiente se aproxime a cero el modelo deja de ser adecuado, ya que la cantidad de la variabilidad explicada mediante el modelo es pobre [11].
En general, R^2 aumenta siempre que se agrega un regresor al modelo, independientemente del valor de la contribución de esa variable. En consecuencia, es difícil juzgar si un aumento de R^2 dice en realidad algo importante.
Es posible usar el estadístico R^2_{Adj} , que se define como sigue:

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-(p+1))}{SS_T/(n-1)}. \tag{68}$$

En vista de que $\frac{SS_{Res}}{(n-(p+1))}$ es el cuadrado medio de los residuales, $\frac{SS_T}{(n-1)}$ es constante e independiente de cuántas variables hay en el modelo, R^2_{Adj} sólo aumentará al agregar una variable al modelo si esa variable reduce el cuadrado medio residual. La R^2_{Adj} penaliza el aumento de términos que no son útiles. Tanto R^2 como R^2_{Adj}, suelen utilizarse como procedimientos para evaluar y comparar los posibles modelos de regresión.

2.3.3. Prueba t de Student sobre coeficientes individuales

Una vez determinado que al menos una de las variables independientes es importante, la siguiente pregunta es: ¿Cuál(es) variable(s) (son) es importante(s)? Si agregamos una variable al modelo de regresión, la suma de cuadrados de la regresión aumenta y la suma de cuadrados residuales disminuye. Se debe decidir si el aumento de la suma de cuadrados de la regresión es suficiente para garantizar el uso del regresor adicional en el modelo.
En [11] se sugiere tener cuidado al agregar una variable explicativa, ya que también aumenta la varianza del valor ajustado $\hat{Y}$, por lo que se sugiere incluir sólo variables explicativas que tengan valor para explicar la respuesta. Además, si agregamos una variable explicativa que no es importante se puede aumentar el cuadrado medio de residuales y con eso disminuye la utilidad del modelo.

Las hipótesis para probar la significancia de cualquier coeficiente $\beta_i, i = 1,2,\ldots,p$, está dado por:

$$H_0 : \beta_i = 0 \quad \text{vs} \quad H_a : \beta_i \neq 0 \quad \text{para} \quad i = 1,2,\ldots,p.$$

El estadístico de prueba para esta hipótesis es,

$$t_0 = \frac{\hat{\beta}_i}{\sqrt{\sigma^2 C_{ii}}}, \tag{69}$$

donde C_{ii} es el elemento diagonal de $(X'X)^{-1}$ que corresponde a $\hat{\beta}_i$. Se rechaza la hipótesis nula $H_0 : \beta_i = 0$ si

$$|t_0| > t_{\alpha/2,n-p-1}.$$

Se observa que ésta es una prueba parcial, porque el coeficiente de regresión $\hat{\beta}_i$ depende de todas las demás variables explicativas x_j, que hay en el modelo. Esto es una prueba de la contribución de x_i dadas las demás variables del modelo.
En general, el cuadrado de una variable aleatoria t con f grados de libertad es una variable aleatoria F con 1 y f grados de libertad, respectivamente [11]. Aunque la prueba t para $H_0 : \beta_1 = 0$ equivale a la prueba F en la regresión lineal simple, la prueba t es algo más adaptable, porque se podría usar para probar hipótesis alternativas unilaterales (Sea $H_1 : \beta_1 < 0$ o $H_1 : \beta_1 > 0$), mientras que la prueba F sólo considera la alternativa bilateral [11].

2.3.4. Intervalos de confianza

Para construir intervalos de confianza par los coeficientes en β es importante reiterar que $\varepsilon \sim N(0, \sigma^2 I)$ y $y_i \sim N(0, \beta X)$ son independientes para cualesquiera $i \neq j$. Note que $\hat{\beta}$ es una combinación lineal de las observaciones, entonces se distribuye normal con media cero y matriz de covarianza $\sigma^2(X'X)^{-1}$. De esta manera, cada estadístico

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, \quad j = \overline{0,k} \tag{70}$$

se distribuye $t_{(n-p)}$. Con esta información es sencillo construir un intervalo de confianza para el j-ésimo coeficiente

$$\left(\hat{\beta}_j - t_{(n-p)}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{(n-p)}^{\alpha/2}\sqrt{\hat{\sigma}^2 C_{jj}}\right) \tag{71}$$

Donde C_{jj} es el j-ésimo elemento en la diagonal de $(X'X)^{-1}$.

2.3.5. Predicción de nuevas observaciones

Análogo al caso de regresión lineal simple, el modelo puede ser utilizado para hacer predicciones de observaciones futuras y, para valores particulares de X. Si $X_0' = [1, x_{01}, \ldots, x_{0k}]$, entonces el punto estimado de una observación futura y_0 es

$$\hat{y}_0 = X_0'\hat{\beta}$$

y el intervalo de confianza para y_0 es

$$\left(\hat{y}_0 - t_{(n-p)}^{\alpha/2}\sqrt{\hat{\sigma}^2(1 + X_0'(X'X)^{-1}X_0)} \leq y_0 \leq \hat{y}_0 + t_{(n-p)}^{\alpha/2}\sqrt{\hat{\sigma}^2(1 + X_0'(X'X)^{-1}X_0)}\right) \tag{72}$$

Que es una generalización de la Ec. (28).

3. Revisión de la adecuación del modelo

Hasta ahora se ha discutido sobre el ajuste de la recta (o plano en R^n) de regresión y sus propiedades asumiendo que los supuestos se cumplen. Sin embargo, es igual de importante verificar la validez de las estas premisas para que el modelo sea insesgado y de mínima varianza, además de la posibilidad de las pruebas estadísticas. Para recordar, los supuestos son:

1. La relación entre la respuesta y y los regresores es lineal, al menos en forma aproximada.
2. El error ε tiene media cero.
3. El error ε tiene varianza σ^2 constante.
4. Los errores no están correalcionados.
5. Los errores tienen distribución normal.

En este capitulo se introducen análisis del cumplimiento de estos puntos.

3.1. Residuales, escalamiento y error puro

Note que 4 de los 5 supuestos están relacionados con el error ε de nuestro modelo $Y = X\beta + \varepsilon$, de tal manera que la violación de alguno de estos puntos se ve reflejada en los errores. Aunque no es del todo preciso, se puede entender a los residuales como los errores observados ($e = Y - \hat{Y}$), siendo estos la desviación entre los datos y el ajuste. Sin embargo, los residuales no son independientes, ya que los n residuales sólo tienen $n - p$ grados de libertad asociados a ellos. Si $n \gg p$, la dependencia tiene poco efecto al utilizar los residuales para la comprobación del modelo. Como se ha visto, los residuales tienen media cero y matriz de covarianza es

$$cov(e) = \sigma^2[I - X(X'X)^{-1}X'] = \sigma^2[I - H]$$

donde $H = X(X'X)^{-1}X'$ es la matriz gorro. Algunas características interesantes de la matriz gorro son, por ejemplo, que transforma a Y en $\hat{Y}$, $\hat{Y} = HY$; además, $HX = X(X'X)^{-1}X'X = X$; es simétrica e idempotente; los residuales se pueden expresar como

$$e = (I-H)Y = (I-H)(X\beta+\varepsilon) = X\beta - HX\beta + (I-H)\varepsilon = X\beta - X(X'X)^{-1}X'X\beta + (I-H)\varepsilon = (I-H)\varepsilon$$

es decir, si h_{ij} es el elemento de la i-ésima fila y la j-ésima columna, entonces $e_i = \varepsilon - \sum_{j=1}^{n} h_{ij}\varepsilon_j \; i = \overline{1,n}$.

La varianza se estima por medio de

$$\hat{\sigma}^2 = \frac{1}{n-k-1}(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \frac{Y'Y - \hat{\beta}'X'Y}{n-k-1} = \frac{SSE}{n-k-1}$$

La matriz de covarianza de $\hat{\beta}$ es

$$cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

y se puede estimar sustituyendo σ^2 por $\hat{\sigma}^2$.

Como los residuales tienen distribución normal, es posible obtener **residuales estandarizados**

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}} \tag{73}$$

Dado que la varianza del i-ésimo residual se encuentra en la matriz $(I-H)$, entonces una aproximación más precisa se logra sabiendo que $V(e_i) = \sigma^2(1-h_{ii})$ y $cov(e_i, e_j) = -\sigma^2 h_{ij}$. Ahora bien, ya que $0 \leq h_{ii} \leq 1$, si se usa la estimación $\hat{\sigma}^2$, en realidad se sobre estima $V(e_i)$. Además, como h_{ii} es una medida de ubicación del i-ésimo punto en el espacio de X, la varianza de e_i depende de dónde esta el punto x_i. Bajo los supuestos, $V(r_i) = 1$, que es independiente de la posición de x_i, donde r_i es el i-ésimo **residual estudentizado** dado por

$$r_i = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}} \tag{74}$$

Un tercer método de escalar los residuales usa una estimación de σ^2 que ha excluido la i-ésima observación

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}_{(i)}(1-h_{ii})}} \tag{75}$$

donde $\hat{\sigma}_{(i)}$ es calculado con $n-1$ observaciones restantes después de omitir $(y_i, x_i') = (y_i, 1, x_{i1}, \ldots, x_{ik})$. Si la observación omitida es atípica, será más probable que aparezca con esta nueva estandarización llamada estudentización externa de los residuales o sólo R de Student.

Otra opción es examinar los residuales omitidos **PRESS** (prediction error sum of equares). El i-ésimo residual omitido $e_{(i)}$ es calculado con $\hat{\beta}_{(i)}$, que se basa en $n-1$ observaciones restantes después de omitir (y_i, x_i'), es decir

$$e_{(i)} = y_i - \hat{y}_{(i)} = y_i - x_i'\hat{\beta}_{(i)}$$

Por definición,

$$\hat{\beta}_{(i)} = (X_{(i)}'X_{(i)})^{-1}X_{(i)}'Y_{(i)} \tag{76}$$

Donde $X_{(i)}$ e $Y_{(i)}$ siguen el mismo sentido de omisión del i-ésimo elemento. Parecería que es necesario hacer un modelo de regresión para cada subconjunto de valores después de ir iterando el valor omitido, pero no es así, ya que

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{e_i}{1-h_{ii}}(X'X)^{-1}x_i \tag{77}$$

Siguiendo esto, se puede calcular el valor PRESS

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \tag{78}$$

Así, los n residuales PRESS pueden ser calculados sin haces n regresiones. Los residuales escalados t_i pueden expresarse en términos de PRESS como

$$t_i = \frac{e_{(i)}}{\sqrt{\hat{V}(e_{(i)})}} \tag{79}$$

Una manera de detectar los puntos atípicos es graficar los residuales ordinarios contra los residuales PRESS. Si no existe un cambio sustancial al calcular $\hat{\beta}$ sin alguna observación, el anterior gráfico debería seguir, aproximadamente, una línea recta de pendiente 1. Cualquier punto lejano a esta recta es potencialmente un valor atípico.

Si un punto atípico viene de una distribución con una media diferente, el modelo puede ser expresado como $E(y_i) = x_i'\beta + \theta$. La distribución de t_i en (75) o (79) es una t de Student con $(n-k-1)$ grados de libertad, de tal manera que t_i sirve para la prueba de hipótesis $H_0: \theta = 0$. Como se harán n pruebas, podemos

sólo enfocarnos en el tamaño de los valores de t_i. Con los residuales PRESS es posible definir el **estadístico PRESS**

$$PRESS = \sum e_{(i)}^2 = \sum [y_i - \hat{y}_{(i)}]^2 = \sum \left(\frac{e_i^2}{1 - h_{ii}} \right) \tag{80}$$

Los residuales con valores grandes en h_{ii} contribuyen más al estadístico PRESS. Para un conjunto dado de datos, PRESS puede ser un mejor evaluador de la calidad predictiva del modelo que SSE. Se prefiere valores pequeños de este estadístico para un modelo con mejor predictibilidad. Si un valor alto de algún residual PRESS se presenta, este dato es lejano al resto, tanto que podría ser un dato atípico dado por la naturaleza del experimento y no precisamente por un error de medición. En esta situación el experimento tiene comportamientos diferentes para valores cercanos a la observación correspondiente a dicho residual. Por ejemplo, si se estima la proliferación de una bacteria dada una temperatura, es posible que haya un punto de inflexión donde en vez de seguir aumentando el número de bacterias, la mayoría muera.

Si se tienen observaciones repetidas Y para un mismo punto en el espacio X, es posible hacer un nuevo análisis sobre el **error puro**. Estas repeticiones deben ser genuinas, es decir, replicar el experimento para el mismo punto en X y obtener las observaciones. Una repetición no genuina está dada por observar varias veces un experimento que no se ha repetido. Por ejemplo, si queremos estimar la relación entre la inteligencia IQ y la altura, una repetición genuina estaría dada si observamos el IQ de dos personas con la misma altura. Por el contrario, no sería genuina si observamos dos veces el IQ de la misma persona, en ese caso se recomienda mejor solo registrar una observación.

Spongamos que tenemos m diferentes valores de X y para el j-ésimo valor existen n_j observaciones. Es decir, existen $y_{j1}, y_{j2}, \ldots, y_{jn_j}$ observaciones para x_j. Todas las observaciones juntas son

$$n = \sum_{j=1}^{m} \sum_{u=1}^{n_j} 1 = \sum_{j=1}^{m} n_j$$

observaciones.La contribución de la suma de cuadrados del error puro de n_1 observaciones en x_1 es la suma interna de cuadrados de Y_{1u} sobre su promedio $\bar{y}_1$; que es

$$\sum_{u=1}^{n_1} (y_{1u} - \bar{y}_1)^2 = \sum_{u=1}^{n_1} y_{1u}^2 - n_1 \bar{y}_1^2 = \sum_{u=1}^{n_1} y_{1u}^2 - \frac{1}{n_1} \left(\sum_{u=1}^{n_1} y_{1u} \right)^2 \tag{81}$$

Siempre que estemos seguros que la variación del error puro es de la misma magnitud a través de los datos, es conveniente hacer la agrupacion de las sumas internas de cuadrados de todos los lugares con observaciones repetidas para obtener el error puro general SS como

$$\sum_{j=1}^{m} \sum_{u=1}^{n_1} (y_{ju} - \bar{y}_j)^2$$

con grados de libertad

$$n_e = \sum_{j=1}^{m} (n_j - 1) = \sum_{j=1}^{m} (n_j - m)$$

Por tanto, el cuadrado medio del error puro es

$$S_e^2 = \frac{\sum_{j=1}^{m} \sum_{u=1}^{n_1} (y_{ju} - \bar{y}_j)^2}{\sum_{j=1}^{m} (n_j - m)}$$

Se puede entender esta cantidad como la suma total de cuadrados dentro de repeticiones entre sus grados de libertad y es una estimación de σ^2.

La **prueba de Barlett** muestra si existe homogeneidad en el error puro. Esta prueba requiere de normalidad y es sensible a la violación de este supuesto, es decir, bajo no normalidad la validez de la prueba es afectada.

Sea $S_1^2, S_2^2, \ldots, S_m^2$ estimaciones de σ^2 provenientes de m de repeticiones con $v_1, \ldots, v_m$ grados de libertad, respectivamente, donde $v_j = n_j - 1$,

$$S_j^2 = \frac{\sum_{u=1}^{n_j}(Y_{uj} - \bar{Y}_j)^2}{n_j - 1},$$

$$S_e^2 = \frac{v_1 S_1^2, \ldots, v_m S_m^2}{v_1 + \cdots + v_m}$$

Denotemos $v = \sum_{u=1}^{m} v_u$ y la constante C como

$$C = 1 + \frac{v_1^{-1} + \cdots + v_m^{-1} - v^{-1}}{3(m-1)}$$

Entonces, el estadístico de prueba es

$$B = \frac{v \ln S_e^2 - \sum_{j=1}^{m} v_j \ln S_j^2}{C} \tag{82}$$

Cuando las varianzas de los m grupos son iguales, B se distribuye aproximadamente como una χ_{m-1}^2. Valores grandes de B podrían indicar heterogeneidad de las varianzas, así como falta de normalidad.

3.2. Gráficos de los residuales

Gráfica de probabilidad normal

Las pequeñas desviaciones respecto a la suposición de normalidad no afectan mucho al modelo, pero tener desviaciones grandes de no normalidad es potencialmente peligroso, porque los estadísticos t o F dependen de la suposición de normalidad. En [11] se comentó que si los errores provienen de una distribución con colas más gruesas (cuando la frecuencia de ocurrencia de eventos que están situados en los extremos de la distribución no es muy baja) que la normal, el ajuste por mínimos cuadrados será sensible a un subconjunto menor de datos.

Un método muy sencillo de comprobar la suposición de **normalidad** es trazar una gráfica de **probabilidad normal** de los residuales. Esta es una gráfica diseñada para que se dibuje una línea recta, que representa a una normal acumulada.

Sea $e_1 < e_2 < \ldots < e_n$ los residuales ordenados en orden creciente. Si se grafica e_i en función de la probabilidad acumulada $P_i = (i - \frac{1}{2})/n, i = 1, 2, \ldots, n$, los puntos que resulten deberían estar aproximadamente sobre una línea recta.

La recta se suele determinar en forma visual, con énfasis en los valores centrales (por ejemplo, los puntos de probabilidad acumulada 0.33 y 0.67) y no en los extremos. Las diferencias apreciables en distancia respecto a la recta indican que la distribución no es normal.

A veces, las gráficas de probabilidad normal se trazan graficando el residual clasificando e_i en función del valor normal esperado, $\phi^{-1}[(i - \frac{1}{2})/n]$, donde ϕ representa a la función de distribución acumulada de la distribución normal estándar. Esto es consecuencia de $E(e_i) \simeq \phi^{-1}[(i - \frac{1}{2})/n]$ [11].

El estudio de las gráficas ayuda a adquirir un grado de percepción de cuánta desviación de la recta es aceptable. Con frecuencia, los tamaños pequeños de muestra $(n \leq 16)$ producen gráficas de probabilidad normal que se desvían bastante de línea recta que representa la normal acumulad. Para muestras mayores $(n \geq 32)$, las gráficas se comportan mucho mejor. Por lo general, se requieren alrededor de 20 puntos para producir gráficas de probabilidad suficientemente estables como para poder interpretarse con facilidad.

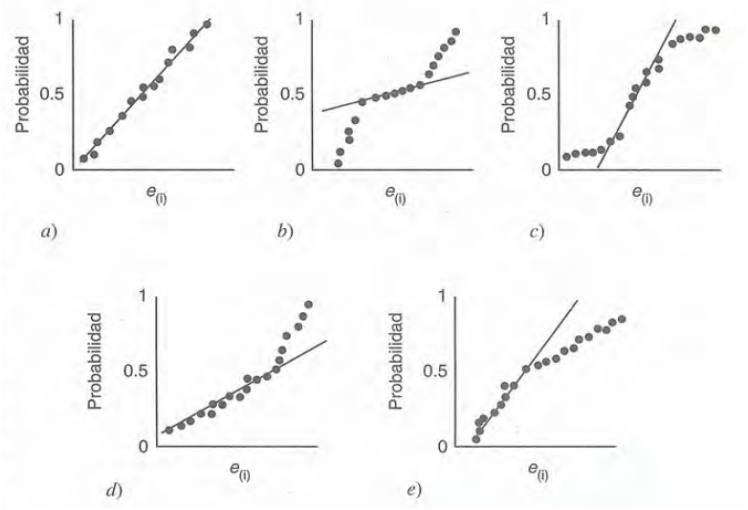


Figura 12: Gráficas de probabilidad normal: a) ideal; b) distribución con colas gruesas; e) distribución con colas delgadas; d) asimetría positiva; e) asimetría negativa. Fuente: [11].

3.2.1. Gráfica de residuales en función de los valores ajustados $\hat{y}_i$

Para poder detectar algunas inadecuaciones del modelo, es útil tener una gráfica de los residuales en función de los valores ajustados correspondientes $\hat{y}_i$. Está gráfica permite detectar diferentes problemas, tales como:

- **Heterocedasticidad**, la varianza no es constante y se deben de transformar los datos (la variable Y) o aplicar otros métodos de estimación.
- **Error en el análisis**, se ha realizado mal el ajuste y se verifica que los residuos negativos se corresponden con los valores pequeños $\hat{y}_i$ y los errores positivos se corresponden con los valores grandes de y_i, o al revés.
- El modelo es inadecuado por **falta de linealidad** (no lineal) y se deben transformar los datos o introducir nuevas variables que pueden ser cuadrados de las existentes o productos de las mismas, o bien se deben introducir nuevas variables explicativas.
- Existencia de **observaciones atípicas** o puntos extremos.
- **Falta de independencia**, los residuales se presentan formando grupos (clusters).

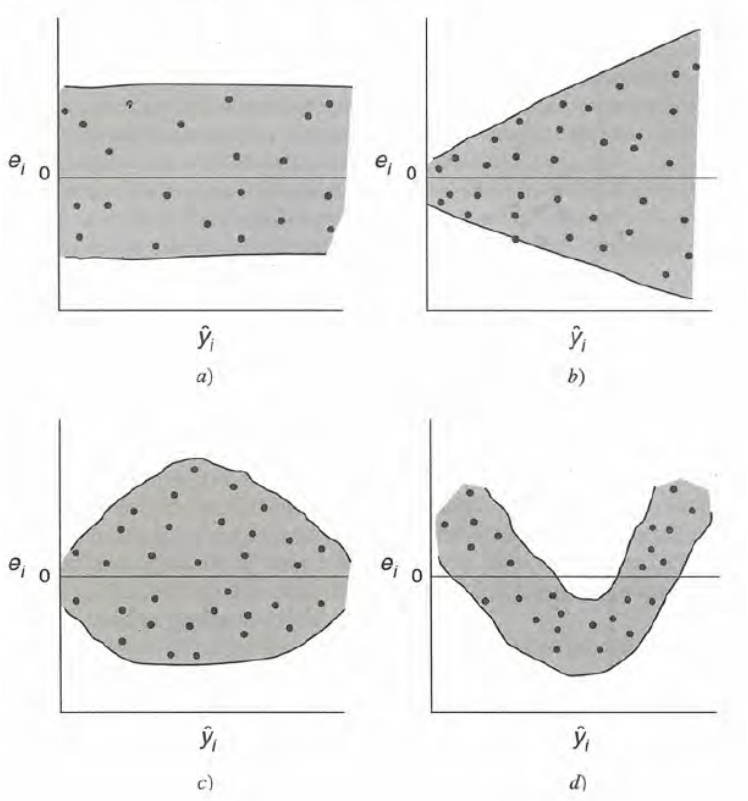


Figura 13: Patrones en las gráficas de residuales: a) satisfactorio; b) en embudo; e) en doble arco; d) no lineal
Fuente: [11].

En la Fig. 13, las distribuciones en las partes b y c indican que la varianza de los errores no es constante. La figura de embudo abierto hacia afuera en la parte b implica que la varianza es función creciente de y. También es posible un embudo abierto hacia dentro, que indica que $V(e)$ aumenta a medida que y disminuye. La distribución en doble arco en la parte c se presenta con frecuencia cuando y es una proporción entre O y 1. Se grafican los residuales contra $\hat{y}_i$ y no y_i porque usualmente existe una correlación entre e_i e y_i, formando patrones incluso si no existe ningún problema en el modelo. También es plausible tener a los residuales e_i en el en el eje de las ordenadas y algún regresor x_{ji} en las abscisas. Se espera, de manera ideal, tener aleatoriedad como en la Figura 13(a). De otra manera, como en (b) y (c), puede existir varianza no constante o una relación de orden superior entre la variable Y y X_j, como se muestra en 13(d).

Un supuesto del modelo de regresión lineal es la independencia entre las observaciones o en análogo la independencia entre los errores. Cuando se toma una muestra que tiene un orden temporal marcado es posible tener **autocorrelación temporal**. Si conocemos el orden en el tiempo de los datos, se puede graficar los residuales con la esperanza de que sigan mostrando aleatoriedad, similar a la Figura 13(a), de lo contrario es posible que se necesite una transformación, o incluso cambiar el enfoque de regresión lineal a series de tiempo. Incluso, en el caso multivariado, si deseamos conocer la correlación entre variables explicativas X se puede hacer una gráfica de X_j contra X_i para $i \neq j$. Como se espera que la correlación sea cercana a

cero, no debe existir ningún patrón en este gráfico. La correlación es una medida de asociación lineal entre dos variables, de tal manera que si la relación es no lineal, puede que no se vea reflejada en el factor de correlación, por lo cual un gráfico no da una visión más general de la influencia de una variable regresora con otra.

Si conocemos información extra sobre el experimento, como la fuente de las observaciones, algún orden temporal o espacial, además de alguna relación a priori de las variables, es mejor hacer cualquier gráfica que resulte de la imaginación del investigador para no perder generalidad de la información y establecer una correcta adecuación al modelo de regresión lineal.

3.3. homogeneidad en las varianzas

En la ecuación (82) se expuso una prueba estadística para la homogeneidad de las varianzas. Aquí se hace una continuación de estas pruebas. Existe una ventaja sustancial de una prueba estadística sobre un gráfico al no ser necesaria una interpretación que puede ser hasta arbitraria sujeta a la experiencia del investigador. Por otro lado, la prueba estadística resume tanto la información que para un conjunto de datos obtenemos un único resultado lógico de verdadero o falso y no es posible discernir el tipo de comportamiento que posiblemente tenga la varianza, como en un gráfico.

En la **prueba de Barlett modificada para curtosis** el estadístico B de la Ec. (82) es multiplicado por $d = 2/(\hat{\beta} - 1)$, donde

$$\hat{\vartheta} = \frac{N \sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)}{\left(\sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2 \right)^2}$$

estima la curtosis de los conjuntos de repeticiones. Para una distribución normal de los datos se espera que el verdadero valor de ϑ sea 3 y d tenga un valor cercano a 1. La misma prueba χ^2 es utilizada. N es el número total de observaciones (usualmente reducidas) del conjunto de datos usados para la prueba, que es, el número total de observaciones en todos los conjuntos de repeticiones, ignorando las observaciones sin repeticiones.

La **prueba de Levene con medias** considera, en el j-ésimo grupo de repeticiones, la desviación absoluta

$$z_{ju} = |Y_{ju} - \bar{Y}_j|, \quad u = 1, 2, \ldots, n_j$$

de los $Y's$ de las medias de sus grupos repetidos. Entiéndase esto como una manera de clasificación y comparación de los cuadrados medios .[en]tre gruposçon los cuadrados medios "dentro de grupos.[a] través del estadístico F. El estadístico F apropiado es

$$F_0 = \frac{\sum_{j=1}^{m} n_j (\bar{z}_j - \bar{z})^2 / (m-1)}{\sum_{j=1}^{m} \sum_{u=1}^{n_j} (z_{ju} - \bar{z}_j)^2 / \sum_{j=1}^{m} (n_j - 1)} \tag{83}$$

donde

$$\bar{z}_j = \sum_{u=1}^{n_j} \frac{z_{ju}}{n_j} \quad \& \quad \bar{z} = \sum_{j=1}^{m} \sum_{u=1}^{n_j} \frac{z_{ju}}{\sum_{j=1}^{m} n_j}$$

El valor F se refiere a $F_{(m-1),\sum_{j=1}^{m}(n_j-1)}$ de cola superior.

3.4. normalidad en los residuales

Usualmente asumimos que los residuales se distribuyen normal $e_i \sim N(0,\sigma^2)$, supuesto fuerte para los intervalos de confianza y pruebas de hipótesis, y todos los errores son independientes unos de otros. Pero sus estimados, los residuales, no pueden ser independientes unos de otros. La estimación de los parámetros nos dice que n residuales sólo tienen $(n-p)$ grados de libertad. Las p ecuaciones normales son restricciones sobre los e_i. A menos que p sea grande en comparación con n, esto tiene poco efecto sobre la revisión de la normalidad.

Si el modelo ajustado es $E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, la ecuación puede ser escrita como

$$-2\sum(y_i - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}) = 0$$

Esto se reduce a

$$\sum(y_i - \hat{y}_i) = 0$$

Por lo tanto

$$\sum e_i = 0 \tag{84}$$

De esta manera, no es necesario rectificar que el residual medio $\bar{e} = \sum e_i/n$ es cero, pues es consecuencia directa de la Ec. (84). Otra manera es hacer pequeños intervalos y después crear un histograma para la frecuencia de los intervalos. Si los residuales se distribuyen aproximadamente normal deberían presentar la forma de la campana normal en este histograma. Para pruebas estadísticas de normalidad podemos encontrar la *Shapiro-Wilk* para muestras menores a 50; la prueba *Kolmogorov-Smirnov* es más general y sirve para comprobar si los datos siguen una distribución normal, uniforme, Poisson o exponencial; la prueba *Anderson-Darling* es para casi cualquier distribución y entre menor sea el estadístico, mejor es el ajuste, de tal manera que es posible comparar el ajuste de los datos a distintas distribuciones Fig. (14), incluida la normal.

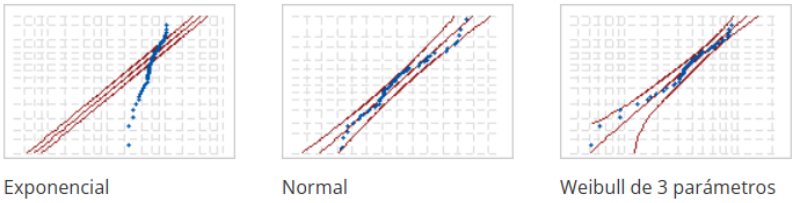


Figura 14: Prueba Anderson-Darling para los mismos datos y distintas distribuciones.

3.5. Prueba Durbin-Watson

Una prueba popular para detectar algún tipo de correlación serial es la prueba Durbin-Watson. Supongase que se quiere ajustar un modelo lineal

$$Y_u = \beta_0 + \sum_{i=1}^{k} \beta_k X_{iu} + \varepsilon_u$$

por mínimos cuadrados para las observaciones $(Y_u, X_{1u}, \ldots, X_{ku})$, para $u = 1, \ldots, n$. Usualmente asumimos que los errores son variables independientes $\varepsilon_i \sim N(0,\sigma^2)$, por lo cual todas las correlaciones seriales son

$\rho_s = 0$. Deseamos verificar este supuesto a través de los residuales. Se plantea el juego de hipótesis

$$H_0 : \rho_s = 0 \quad \forall s \quad \text{contra} \quad H_1 : \rho_s = \rho^s$$

($\rho \neq 0$, $|\rho| < 1$). La hipótesis alternativa surge de asumir que los errores ε_u son tales que

$$\varepsilon_u = \rho \varepsilon_{u-1} + z_u$$

donde $z_u \sim N(0, \sigma^2)$ y es independiente de $\varepsilon_{u-1}, \varepsilon_{u-2}, \ldots$, y $z_{u-1}, z_{u-2}, \ldots$. Se asume, también, que la media y la varianza de ε_u son constantes, independientes de u, de lo cual se sigue necesariamente que

$$\varepsilon_u \sim N\left(0, \sigma^2/(1-\rho^2)\right)$$

Note que bajo la hipótesis nula esto se reduce a $\varepsilon_u \sim N(0, \sigma^2)$. El estadístico es

$$d = \frac{\sum_{u=2}^{n} (e_u - e_{u-1})^2}{\sum_{u=1}^{n} e_u^2} \tag{85}$$

y determina si se rechaza o no la hipótesis nula basado en el valor d. La distribución de d depende de los datos en X y no es independiente de ellos. Esta distribución cae entre 0 y 4, siendo simétrica al rededor de 2. Los puntos porcentuales también dependen de los datos de X y tendrían que calcularse para cada aplicación para realizar la prueba correctamente. Debido a la dificultad de hacer esto de manera rutinaria, la prueba generalmente se realiza utilizando límites tabulados (d_L, d_U). Por lo tanto, en lugar de buscar un único valor crítico, tenemos que buscar dos valores críticos. Además, d se utiliza sólo para prueba de cola inferior contra la alternativa $\rho < 0$, teóricamente necesitamos una prueba de cola superior; afortunadamente, esto puede manejarse simplemente como una prueba de cola inferior utilizando el estadístico (4-d).

Note que los extremos 0 y 4 son alcanzables para muestras muy grandes. Los valores mínimos alcanzables dependen del tamaño de la muestra n de la siguiente manera

n	15	30	50	100	200	300	500
d mínimo	0.0437	0.0110	0.0039	0.0010	0.0002	0.0001	0.0000

El correspondiente valor d máximo es (4- d mínimo). Para tablas más detalladas con distintos n (observaciones), k(número de variables independientes) y α (nivel de significancia) revise [13], páginas 184-192, junto con las reglas de decisión que las acompañan.

3.6. Puntos atípicos

Un valor atípico es una observación extrema. Los residuales cuyo valor absoluto es bastante mayor que los demás, digamos de tres a cuatro desviaciones estándar respecto a la media, indican que hay valores atípicos potenciales en el espacio de Y. Los valores atípicos son puntos que no son representativos del resto de los datos. De acuerdo con su ubicación en el espacio de X, los valores atípicos pueden tener efectos de moderados a graves sobre el modelo de regresión. Las gráficas de residuales en función de y_i y la gráfica de probabilidad normal son útiles para identificar puntos atípicos. El examen de los residuales escalados, como por ejemplo los residuales estudentizados y los R de Student es una forma excelente de identificar puntos atípicos potenciales. Los valores atípicos se deben investigar con cuidado, para ver si se puede encontrar una razón de su comportamiento extraordinario. A veces, los valores atípicos son "malosz se deben a eventos desacostumbrados, pero explicables. Entre los ejemplos están la medición o el análisis incorrectos, el registro incorrecto de los datos y la falla de un instrumento de medición. Si éste es el caso, el valor atípico

se debería corregir (si es posible) o eliminar del conjunto de datos. Es claro que el eliminar valores malos es conveniente, porque los mínimos cuadrados jalan la ecuación ajustada hacia el valor atípico, ya que eso minimiza la suma de cuadrados de residuales, sin embargo, se hace notar que debe contarse con una fuerte evidencia no estadística de que el valor atípico es malo, para entonces descartarlo. A veces se encuentra que el valor atípico es una observación extraordinaria, pero perfectamente factible. Puede ser peligroso eliminar estos puntos para ”mejorar el ajuste de la ecuación”, porque puede dar al usuario una sensación falsa de precisión de la estimación o la predicción. A veces se ve que el valor atípico es más importante que el resto de los datos, porque puede controlar muchas propiedades clave del modelo. También, los valores atípicos pueden hacer resaltar inadecuaciones en el modelo, como la falla de tener un buen ajuste con los datos en cierta región del espacio de X. Si el valor atípico es un punto de respuesta especialmente deseable (por ejemplo, bajo costo o alto rendimiento), sería en extremo valioso conocer los valores de los regresores, cuando se observó esa respuesta. Los análisis de identificación y de seguimiento de los valores atípicos con frecuencia dar como resultado mejoras en el proceso, o nuevos conocimientos acerca de factores cuyo efecto sobre la respuesta se desconocía antes. Se han propuesto diversas pruebas estadísticas para detectar y rechazar los valores atípicos. Por ejemplo, basado en el residual máximo normado $|e_i|/\sqrt{\sum e_i^2}$ cuya aplicación es bastante fácil. El efecto de los valores atípicos sobre el modelo de regresión se puede comprobar con facilidad eliminándolos y volviendo a ajustar la ecuación de regresión, teniendo en cuenta que si son observaciones factibles, entonces el modelo no tiene buena predictibilidad para esa región. Se podrá encontrar que los valores de los coeficientes de regresión, o de los estadísticos de resumen como t, F o R^2, y que el cuadrado medio de residuales pueden ser muy sensibles a los valores atípicos. Los casos en los que un porcentaje relativamente pequeño de los datos tiene un gran impacto sobre el modelo podrán no ser aceptables para el usuario de la ecuación de regresión. En general, uno se siente más cómodo suponiendo que una ecuación de regresión es válida si no es muy sensible a unas pocas observaciones. Se preferiría que la relación de regresión estuviera embebida en todas las observaciones, y no sólo fuera un artificio de unos pocos puntos.

Una forma frecuente de modelar un valor atípico es con el **modelo del valor atípico con media desplazada**. Supóngase que se ajusta el modelo $Y = X\beta + \varepsilon$, cuando el modelo verdadero es

$$Y = X\beta + \delta + \varepsilon$$

donde δ es un factor de $n \times 1$, de ceros excepto por la u-ésima observación, cuyo valor es δ_u. Así,

$$\delta = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \delta \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Supóngase que tanto para el modelo que se ajusta y el modelo con el valor atípico con media desplazada, $E(\varepsilon) \sim N(0, \sigma^2 I)$. Se trata de determinar un estadístico de prueba adecuado para las hipótesis

$$H_0 : \delta_u = 0 \quad \text{contra} \quad H_1 : \delta_u \neq 0$$

En este procedimiento se supone que el interés específico es la u-ésima observación, es decir, que se cuenta con información a priori que la u-ésima observación puede ser un valor atípico. El primer paso es determinar

un estimador adecuado para δ_u. Un candidato es el u-ésimo residual. Sea $e = [I - H]Y$ el vector de los residuales de $n \times 1$. El valor esperado de e es

$$\begin{aligned} E(e) &= E\left([I-H]Y\right) \\ &= [I-H]E(Y) \\ &= [I-H][X\beta + \delta] \\ &= [I-H][X\beta] + [I-H][\delta] \\ &= [X-X]\beta + [I-H][\delta] \\ &= [I-H][\delta] \end{aligned}$$

Así,

$$E(e_u) = (1 - h_{uu})\delta_u$$

siendo h_{uu} el u-ésimo elemento de la diagonal en la matriz sombrero H. En consecuencia, un estimador insesgado de δ_u es

$$\hat{\delta}_u = \frac{e_u}{1 - h_{uu}}$$

Note que esta expresión es equivalente a la expresión en Ec. (78), de tal manera que $\hat{\delta}_u$ es un residual PRESS, cuya varianza es

$$\begin{aligned} V(e) &= V\left([I-H]Y\right) \\ &= [I-H]\sigma^2 I[I-H]' \\ &= \sigma^2[I-H][I-H]' \\ &= \sigma[I-H] \end{aligned}$$

Así, $V(e) = (1 - h_{uu})\sigma^2$. Entonces, la varianza $\hat{\delta}_u$ es

$$\begin{aligned} V\left(\frac{e_u}{1-h_{uu}}\right) &= \frac{1}{1-h_{uu}} V(e_u) \\ &= \frac{(1-huu)\sigma^2}{(1-h_{uu})^2} \\ &= \frac{\sigma^2}{1-h_{uu}} \end{aligned}$$

A continuación, se observará que e es una combinación lineal de Y. Así, e es una combinación lineal de variables aleatorias normalmente distribuidas. Por lo anterior, e sigue una distribución normal, al igual que $\hat{\delta}_u$. Entonces, bajo la hipótesis nula, $H_0 :\ \delta_u = 0$,

$$\frac{e_u/(1-h_{uu})}{\sigma/\left(\sqrt{1-h_{uu}}\right)} = \frac{e_u}{\sigma\sqrt{1-h_{uu}}}$$

sigue una distribución normal estándar. Se ve que esta cantidad no es más que un ejemplo de un residual estudentizado, como se vio anteriormente. En general, se desconoce σ^2. Se vio que $\hat{\sigma}^2$ es un estimador insesgado de σ^2. Además, se vio que $\hat{\sigma}^2/\sigma^2$ tiene una distribución χ^2, dividida entre sus grados de libertad. En consecuencia, un estadístico probable de prueba es

$$\frac{e_u}{S\sigma_{(u)}\sqrt{(1-h_{uu})}}$$

Que es el residual estudentizado externamente como se vio en la Ec. (75). Si H_0 es verdadera, el estadístico sigue una distribución central t_{n-p-1} y bajo H_1, sigue una distribución no central $t'_{n-p-1,\gamma}$, donde

$$\gamma = \frac{\delta_u}{\sigma/\sqrt{1-h_{uu}}} = \frac{\delta_u\sqrt{1-h_{uu}}}{\sigma}$$

Es importante notar que la potencia de esta prueba depende de h_{uu}. Recuérdese que si se ajusta una ordenada al origen al modelo, entonces $1/n \leq h_{uu} \leq 1$. La potencia máxima se tiene cuando $h_{uu} = 1/n$, que está en el centro de la nube de datos, en términos de las X. Cuando $h_{uu} \to 1$, la potencia baja a 0. En otras palabras, esta prueba tiene menos capacidad de detectar valores atípicos en los puntos de datos de alta influencia.

4. Transformaciones y ponderación para corregir inadecuaciones del modelo

En el Cuadro 2 se presenta un método gráfico para identificar una función y su posible transformación para obtener un modelo lineal que es parte esencial de los supuestos para el modelo de regresión lineal. Sin embargo, es posible que trabajar en el espacio Y transformado no sea cómodo para el investigador. Afortunadamente [13] menciona que regresar al espacio Y original es posible una vez que se hayan obtenidos los valores estimados. Por ejemplo, suponga que se han obtenido valores de ajuste para el modelo $\ln Y$ y una predicción para $\hat{\ln} Y$ de un cierto conjunto de X. Si se desea, es posible evaluar $\hat{Y} = \exp \hat{\ln} Y$ y hacer la predicción en el espacio original, con la salvedad de que al aplicar la transformación inversa en forma directa a los valores predichos se obtiene un estimado de la mediana de la distribución de la respuesta, y no de la media. Por otro lado, los intervalos de confianza se pueden convertir en forma directa de un espacio a otro, porque esos estimados son percentiles de una distribución, invariantes ante una transformación [11]. Entonces un intervalo de confianza para $E(\ln Y)$ con intervalo en (a,b) puede trasladarse al intervalo (e^a, e^b) en el espacio de Y, aunque evidentemente pierda simetría al rededor de $\hat{Y}$. De manera análoga se pueden obtener los residuales $Y_i - \hat{Y}_i$, pero estos residuales no son los que se deben usar para verificar los supuestos.

4.1. Método Box-Cox

Existe el **método Box-Cox**, útil para encontrar una transformación apropiada de manera analítica. Se parte de la transformación de potencia (TP) Y^λ, donde λ es un parámetro a determinar. Un detalle particular de esta transformación es $\lim_{\lambda \to 0} Y^\lambda = 1$, lo cual sería una transformación sin sentido. Para esta situación de discontinuidad en $\lambda = 0$ se usa $(Y^\lambda - 1)/\lambda$ como variable de respuesta. Así, cuando $\lim_{\lambda \to 0}(Y^\lambda - 1)/\lambda = \ln Y$. Para valores distintos de lambda, esta última función cambian significativamente sus resultados, lo cual conduce a un problema de comparación de adecuación para distintas λ. Tal parece un mejor resultado usar la transformación $W = (Y^\lambda - 1)/\lambda$ para $\lambda = 0$, ahora fuera de moda, Se sugiere, en consecuencia, usar la función $V = W/\dot{Y}^{\lambda - 1}$

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda \dot{Y}^{\lambda - 1}}, & \lambda \neq 0 \\ \dot{Y} \ln Y, & \lambda = 0 \end{cases} \tag{86}$$

Donde $\dot{Y}$ es el promedio geométrico de las observaciones y_i, de tal manera que

$$\dot{Y} = \ln^{-1}\left(\frac{\sum \ln y_i}{n}\right)$$

que está relacionado con el jacobiano de la transformación que convierte las variables de respuesta Y en Y^λ. Es, de hecho, un factor de escala que segura que la suma de los cuadrados de los residuales sean comparables por modelos con distintos valores de λ.

Note que para poder aplicar la transformación, todos los valores Y deben ser positivos. En general, cuando hacemos una transformación, es imposible relacionar los parámetros del modelo para datos transformados con los parámetros del modelo no transformado. Por lo general, no existe una equivalencia matemática a no ser de una aproximación por series de Taylor. Por ejemplo, si en vez de ajustar $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$ ajustamos $Y^\lambda = \alpha_0 + \alpha_1 X + \varepsilon$, la relación entre β_0, β_1, β_{11} y λ, α_0, α_1 no está clara.

Suponga que se tienen los datos $(y_1, \ldots, y_n)$, donde $y_j > 0$ para $j = \overline{1,n}$. Si el cociente entre el valor más grande y el más pequeño de Y es considerable (por decir 10 o más), es posible considerar la necesidad de

una transformación. A pesar de que existen varias transformaciones (por ejemplo, en el Cuadro 2) una útil para diversos casos es la descrita en la Ec. (86). Cuando esta transformación es aplicada a cada valor de Y_i, creamos un vector $V = (v_1, \ldots, v_n)'$ y lo usamos para ajustar el modelo lineal

$$V = X\beta + \varepsilon \tag{87}$$

por mínimos cuadrados para cualquier valor específico de λ. Más general, debemos estimar el valor λ y β. Esto se logra con el principio de máxima verosimilitud bajo el supuesto de $\varepsilon \sim N(0, \sigma^2 I)$. La idea básica es que si puede ser encontrado un valor apropiado de λ, un modelo aditivo con distribución normal, independiente y con estructura de error homogéneo puede ser ajustado por máxima verosimilitud. Es conveniente (pero no necesario) entender la máxima verosimilitud, la estadística bayesiana y el jacobiano. Los pasos a seguir son los siguientes:

1. Seleccione un valor para λ de un rango seleccionado. Usualmente se selecciona un rango $(-2, 2)$ con 11 a 21 particiones y se va ampliando de ser necesario.

2. Para cada valor de λ evalúe en la Ec. (86). Ahora, ajuste en la Ec. (87) y registre la suma de los cuadrados de los residuales $S(\lambda, V)$ (se puede valer de algún software).

3. Grafique $S(\lambda, V)$ contra λ. Dibuje una curva suavizada a través de los puntos graficados y encuentre para que valor de λ cae el punto más bajo de la curva. El valor, $\hat{\lambda}$, es la máxima estimación de λ. Es común que se use el valor múltiplo de 0.5 más cercano en vez del valor exacto de $\hat{\lambda}$. Por ejemplo, si $\hat{\lambda} = 1.445$ se usa el valor $\hat{\lambda} = 1.5$, pero esto queda a preferencia del investigador.

Una vez seleccionado el valor de λ adecuado, hacemos la transformación $Y^{(\lambda)}$ tal como en la Ec. (87). Alternativamente, si $\lambda \neq 0$, es posible hacer la TP Y^{λ} o si $\lambda = 0$ también disponemos de la transformación $\ln Y$. La razón de sugerir la transformación en la Ec. (86) es para encontrar el valor estimado de máxima verosimilitud de lambda y una vez conocido, ya es posible elegir la transformación más simple.

4.2. Aproximación del intervalo de confianza para λ

Anteriormente encontramos el valor de λ minimizando la función de la suma de los residuales $S(\lambda, V)$. Desde este punto, un paso atrás es el criterio equivalente de maximizar la función

$$L(\lambda) = -\frac{1}{2} \ln S(\lambda, V)/n \tag{88}$$

Un intervalo al $100(1-\alpha)\,\%$ de confianza para λ consiste en los valores de lambda que satisfagan la ecuación

$$L(\hat{\lambda}) - L(\lambda) \leq \frac{1}{2}\chi^2_{1,(1-\alpha)} \tag{89}$$

Donde $\chi^2_{1,(1-\alpha)}$ es el cuantil de la distribución chi cuadrada usando cola superior. Para ilustrar esta desigualdad podemos graficar la función $L(\lambda)$ contra λ y establecer una línea horizontal a la altura de $L(\lambda) = L(\hat{\lambda}) - \frac{1}{2}\chi^2_{1,(1-\alpha)}$. Esta línea cortará a la curva en dos puntos, es decir, para dos valores de la variable λ. Dichos puntos serán los valores aproximados del intervalo de confianza.

Análogo, si se está minimizando la suma de cuadrados $S(\lambda, V)$, entonces el gráfico debe ser $S(\lambda, V)$ contra λ y la línea horizontal debe estar a la altura

$$S(\lambda, V) = S(\hat{\lambda}, V) \cdot \exp\left(\chi^2_{1,(1-\alpha)}/n\right) \quad \text{para} \quad S(\lambda, V) \tag{90}$$

ó

$$\ln S(\lambda, V) = \ln S(\hat{\lambda}, V) + \chi^2_{1,(1-\alpha)}/n \quad \text{para} \quad \ln S(\lambda, V) \tag{91}$$

Es posible (sin cambios significativos en la mayoría de los casos), que se usen los factores

- $1 + t_v^2/v$
- $1 + z^2/v$
- $1 + \chi^2_{1,(1-\alpha)/v}$
- $1 + \chi^2_{1,(1-\alpha)/n}$
- $1 + z^2/n$

donde t_v y z son los cuantiles de dos colas de la distribución t con v grados de libertad y de la distribución normal estándar, respectivamente. Estas alternativas tienen soporte en que

- $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$ que, reducido a $1 + x$ como primera aproximación, representa la porción "1+".
- $\chi_1^2 = z^2 \approx t_v^2$ a menos que v sea pequeño (quizá menor a 30).
- Se puede discutir si usar n o v.

4.3. Transformación de las variables regresoras

En la transformación por el método de Box-Cox, se hace una transformación sobre la variable de respuesta Y que resulta equivalente a realizar la transformación inversa al lado derecho de la ecuación $Y = X\beta + \varepsilon$ y puede resultar conveniente para mejorar la adecuación del modelo o recuperar supuestos que se estaban violando. Ahora, si existe una (o más) variable regresora que se relaciona en forma de potencia con la variable de respuesta, por decir $\xi = x^\alpha$, y los demás supuestos (normalidad, independencia, homocedasticidad) se cumplen, por lo menos de manera aproximada, entonces es posible encontrar de forma analítica la transformación pertinente mediante el **método Box-Tidwell**. Para el caso de regresión lineal simple, suponga la relación

$$E(y) = \beta_0 + \beta_1 x^\alpha$$

Si definimos una nueva variable *xi* tal que

$$\xi = \begin{cases} x^\alpha, & \alpha \neq 0 \\ \ln x, & \alpha = 0 \end{cases}$$

Podemos expresar nuestro modelo como una función de los parámetros desconocidos y la variable *xi*

$$E(y) = \beta_0 + \beta_1 \xi = f(\xi, \beta_0, \beta_1) \tag{92}$$

Donde β_0, β_1 y α son parámetros desconocidos. Si α_0 es un tanteo inicial de α es común usar $\alpha_0 = 1$. Por lo que $\xi_0 = x^1 = x$, es decir, no se hizo ninguna transformación. Al desarrollar en una serie de Taylor respecto al tanteo inicial, e ignorar los términos de orden mayor que uno, se obtiene

$$E(y) = f(\xi_0, \beta_0, \beta_1) + (\alpha - \alpha_0) \left[\frac{df(\xi, \beta_0, \beta_1)}{d\alpha} \right]_{\alpha=\alpha_0, \xi=\xi_0} \tag{93}$$

$$= \beta_0 + \beta_1 + (\alpha - 1) \left[\frac{df(\xi, \beta_0, \beta_1)}{d\alpha} \right]_{\alpha=\alpha_0, \xi=\xi_0} \tag{94}$$

Ahora bien, si se conociera el término entre llaves de esta ecuación, se podría manejar como una variable regresora adicional y sería posible estimar los parámetros β_0, β_1, α mediante mínimos cuadrados. El estimado de α se podría tomar entonces como un estimado mejorado del parámetro de transformación. El término entre llaves de la ecuación (93) se puede escribir en la siguiente forma:

$$\left[\frac{df(\xi,\beta_0,\beta_1)}{d\alpha}\right]_{\alpha=\alpha_0,\xi=\xi_0} = \left[\frac{df(\xi,\beta_0,\beta_1)}{d\xi}\right]_{\xi=\xi_0} \cdot \left[\frac{d\xi}{d\alpha}\right]_{\alpha=\alpha_0}$$

Como $\xi = x^\alpha$ para $\alpha \neq 0$, entonces

$$\frac{d\xi}{d\alpha} = x^\alpha \ln x$$

y

$$\left[\frac{df(\xi,\beta_0,\beta_1)}{d\xi}\right]_{\xi=\xi_0} = \frac{d(\beta_0+\beta_1 x)}{dx} = \beta_1$$

Estos parámetros se pueden ajustar de forma conveniente ajustando por mínimos cuadrados el modelo

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Se puede calcular un ajuste del tanteo inicial $\alpha_0 = 1$ definiendo una segunda variable regresora como $w = x^\alpha \ln x$ y estimando por mínimos cuadrados los parámetros en

$$\begin{aligned} E(y) &= \beta_0^* + \beta_1^* x + (\alpha-1)\beta_1 w \\ &= \beta_0^* + \beta_1^* x + \gamma w \end{aligned}$$

para obtener

$$\hat{y} = \hat{\beta}_0^* + \hat{\beta}_1^* + \hat{\gamma} w$$

y definiendo como estimador de α

$$\alpha_1 = \frac{\hat{\gamma}}{\hat{\beta}_1} + 1 \tag{95}$$

En general, $\hat{\beta}_1$ y $\hat{\beta}_1^*$ serán distintas. Ahora se puede repetir este procedimiento usando un nuevo regresor $x' = x^{\alpha_1}$ en los cálculos. Box y Tidwell [1962] hacen notar que en general este procedimiento converge con mucha rapidez, y que con frecuencia la α_1 que resulta en la primera etapa es un estimado satisfactorio de α. También previenen que el error de redondeo es un problema potencial, y que los valores sucesivos de α pueden oscilar locamente a menos que se conserven los suficientes lugares decimales. Se pueden encontrar problemas de convergencia en casos en los que la desviación estándar σ del error es grande, o cuando el rango del regresor es muy pequeño en comparación con su media. Esta situación implica que los datos no respaldan la necesidad de transformación alguna.

4.4. Mínimos cuadrados generalizados

En la Ec. (61) se vio por primera vez los mínimos cuadrados generalizados. En esta sección se explica un poco más acerca del tema. Las transformaciones vistas anteriormente sirven para retomar adecuaciones del modelo, como linealidad u homogeneidad de las varianzas. Si se conoce la matriz V tal que $V(\varepsilon) = \sigma^2 V$ en lugar de $V(\varepsilon) = \sigma^2 I$, donde V es una matriz no singular, positiva definida y de tamaño $n \times n$, entonces es posible encontrar los estimadores insesgados y de mínima varianza sin hacer ninguna transformación a las

variables. Estos estimadores son distintos a los estimadores por mínimos cuadrados ordinarios. Supongamos, por ejemplificar, que la matriz V es de la forma

$$V = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$$

Cualquier valor en la matriz V fuera de la diagonal principal representa la covarianza de los errores, mientras que un elemento en la diagonal principal es la varianza de dicho error. Para este ejemplo, las varianzas son no constantes, pues $V \neq I$, y los errores son no correlacionados, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$.

La generalización del modelo es, para cualquier V,

$$\begin{aligned} Y &= X\beta + \varepsilon \\ E(\varepsilon) &= 0 \\ V(\varepsilon) &= \sigma^2 V \end{aligned}$$

Dada la definición de V, también existe una matriz K no singular, simétrica y de tamaño $n \times n$ de modo que $K'K = KK = V$. Con esta matriz definimos las variables

$$z = K^{-1}Y, \qquad B = K^{-1}X, \qquad g = K^{-1}\varepsilon$$

Así, nuestro modelo lineal se transforma en

$$K^{-1}Y = K^{-1}X + K^{-1}\varepsilon$$

que es

$$z = B\beta + g$$

En este modelo los errores tienen valor esperado cero y varianza constante, lo cual nos lleva nuevamente al método de mínimos cuadrados, pero ahora para la variable g.

$$E(g) = E(K^{-1}\varepsilon) = K^{-1}E(\varepsilon) = 0$$

y

$$\begin{aligned} V(g) &= [g - E(g)][g - E(g)]' \\ &= E(gg') \\ &= E(K^{-1}\varepsilon\varepsilon' K^{-1}) \\ &= K^{-1}E(\varepsilon\varepsilon')K^{-1} \\ &= \sigma^2 K^{-1} V K^{-1} \\ &= \sigma^2 K^{-1} K K K^{-1} \\ &= \sigma^2 I \end{aligned}$$

Recuerde que la función a minimizar es la suma de cuadrados $S(\beta)$ para los parámetros dados, donde

$$S(g) = g'g = \varepsilon' V^{-1}\varepsilon = (Y - X\beta)' V^{-1}(Y - X\beta)$$

Entonces, la ecuación resulta de la forma

$$(X'V^{-1}X)\beta = X'V^{-1}Y$$

Despejando β obtenemos la ecuación para los valores estimados de los parámetros $\hat{\beta}$, ahora llamados mínimos cuadrados generalizados (*mcg*).

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$$

Teorema 4. *El estimador de mínimos cuadrados generalizados es el mejor estimador lineal insesgado para cualquier combinación lineal de los coeficientes estimados, $\ell'\hat{\beta}$. BLUE por sus siglas en inglés (Best Linear Unbiased Estimator).*

Demostración. Veamos que $\hat{\beta}$ dado por *mcg* es insesgado.

$$\begin{aligned} E(\hat{\beta}) &= E((X'V^{-1}X)^{-1}X'V^{-1}Y) \\ &= (X'V^{-1}X)^{-1}X'V^{-1}E(Y) \\ &= (X'V^{-1}X)^{-1}X'V^{-1}X\beta \\ &= I\beta \\ &= \beta \end{aligned}$$

Ahora, veamos que es el estimador de mínima varianza. La varianza de $\hat{\beta}$ es

$$\begin{aligned} V(\hat{\beta}) &= V((X'V^{-1}X)^{-1}X'V^{-1}Y) \\ &= [(X'V^{-1}X)^{-1}X'V^{-1}]V(Y)[(X'V^{-1}X)^{-1}X'V^{-1}]' \\ &= [(X'V^{-1}X)^{-1}X'V^{-1}]V[(X'V^{-1}X)^{-1}X'V^{-1}]' \\ &= [(X'V^{-1}X)^{-1}X'V^{-1}]V[V^{-1}X(XV^{-1}X)^{-1}] \\ &= (X'V^{-1}X)^{-1} \end{aligned}$$

En consecuencia,

$$V(\ell'\hat{\beta}) = \ell'V(\hat{\beta})\ell = \ell'[(X'V^{-1}X)^{-1}]\ell$$

Supongamos a $\tilde{\beta}$ como otro estimador insesgado de β que sea combinación lineal de los datos. Se quiere demostrar que $V(\ell'\tilde{\beta} \geq \ell'[(X'V^{-1}X)^{-1}]\ell)$ para algún ℓ. Note que

$$\tilde{\beta} = \left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]Y + b_0$$

donde B es una matriz de $p \times n$ y b_0 es un vector de constantes de tamaño $p \times 1$, que se ajusta en forma adecuada al estimador *mcg*, para formar el estimador alternativo. Si el modelo es correcto, entonces

$$\begin{aligned} E(\tilde{\beta}) &= E\left(\left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]Y + b_0\right) \\ &= \left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]E(Y) + b_0 \\ &= \left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]X\beta + b_0 \\ &= (X'V^{-1}X)^{-1}X'V^{-1}X\beta + BX\beta + b_0 \\ &= \beta + BX\beta + b_0 \end{aligned}$$

Por lo tanto, $\tilde{\beta}$ es insesgado si y sólo si $\beta_0 = 0$ y $BX = 0$. Por otro lado, la varianza de $\tilde{\beta}$ es

$$\begin{aligned}
V(\tilde{\beta}) &= V\left(\left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]Y + b_0\right) \\
&= V\left(\left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]Y\right) \\
&= \left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]V(Y)\left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]' \\
&= \left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]V\left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]' \\
&= \left[(X'V^{-1}X)^{-1}X'V^{-1} + B\right]V(Y)\left[V^{-1}X(X'V^{-1}X)^{-1} + B'\right] \\
&= \left[(X'V^{-1}X)^{-1} + BVB'\right] \\
&=
\end{aligned}$$

De esta manera,

$$\begin{aligned}
V(\ell'\tilde{\beta}) &= \ell' V(\tilde{\beta})\ell \\
&= \ell'\left[(X'V^{-1}X)^{-1} + BVB'\right]\ell \\
&= \ell'(X'V^{-1}X)^{-1}\ell + \ell'BVB'\ell \\
&= V(\ell'\tilde{\beta}) + \ell'BVB'\ell
\end{aligned}$$

Por hipótesis, V es una matriz positiva definida. Por consiguiente, existe una matriz no singular, T, tal que $V = T'T$. El resultado $BVB' = BT'TB'$ es, a lo menos una matriz positiva semidefinida, lo cual implica que $\ell'BVB'\ell \geq 0$. Definamos a $\ell^* = TB'\ell$, con virtud de que

$$\ell'BVB'\ell = \ell^{*'}\ell^* = \sum_{i=1}^{p} \ell_i^{*2}$$

Que es estrictamente mayor a cero para $\ell \neq 0$, a menos que $B = 0$, en cuyo caso $V(\hat{\beta}) = V(\tilde{\beta})$. □

Cuadro 7: Análisis de varianza para mínimos cuadrados generalizados.

Fuente de Variación	**Suma de cuadrados**	**Grados de libertad**	**Cuadrado medio**	F_0
Regresión	$SS_R = \hat{\beta}'B'z$	p	SS_R/p	$\dfrac{MS_R}{MS_e}$
Residuales	$SS_e = z'z - \hat{\beta}'B'z$	$n-p$	$SS_e/(n-p)$	
Total	$SS_T = z'z = Y'V^{-1}Y$	n		

4.5. Mínimos cuadrados ponderados

Los mínimos cuadrados ponderados son un caso particular de mínimos cuadrados generalizados, donde la matriz V es diagonal, definida positiva y con valores $1/w_i$.

$$V = \begin{pmatrix} \frac{1}{w_1} & 0 & \dots & 0 \\ 0 & \frac{1}{w_2} & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \frac{1}{w_n} \end{pmatrix}$$

En este método de estimación se multiplica la diferencia entre los valores observados y esperados de y_i por un peso w_i, o factor de ponderación, que se escoge como inversamente proporcional a la varianza de y_i. Para el caso de la regresión lineal simple, la función de mínimos cuadrados ponderados es

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Con las ecuaciones normales como

$$\hat{\beta}_0 \sum_{i=1}^{n} w_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} w_i x_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i^2 = \sum_{i=1}^{n} w_i x_i y_i$$

Para el caso multiparamétrico, la solución viene dada por la ecuación

$$(X' V^{-1} X)\beta = X' W Y$$

Cuyo despeje para estimar β es

$$\hat{\beta} = (X' V^{-1} X)^{-1} X' W Y$$

Aquí, la matriz V^{-1} también es diagonal con elementos $w_1, \dots, w_n$, que son usualmente llamados pesos o factores de ponderación. Para usar mínimos cuadrados ponderados se deben conocer la matriz V, al igual que en el caso anterior. Sin embargo, A veces se puede recurrir a la experiencia o conocimientos anteriores, o a la información de un modelo teórico, para determinar los pesos w_i, únicos componentes de V. También, el análisis de residuales puede indicar que la varianza de los errores puede ser una función de uno de los regresores, por ejemplo, $V(\varepsilon) = \sigma^2 x_{ij}$, de modo que $w_i = 1/x_{ij}$. En algunos casos, en realidad y_i es un promedio de n_i observaciones en x_i, y si todas las observaciones originales tienen varianza constante σ^2, entonces la varianza de y_i es $V(y_i) = V(\varepsilon_i) = \sigma^2/n_i$, y los pesos se escogerían como $w_i = n_i$. A veces, la fuente principal de error es la del error de medición, y distintas observaciones se miden con distintos instrumentos de precisión desigual (pero bien estimada). En ese caso los pesos se podrían elegir inversamente proporcionales a las varianzas del error de medición. En muchos casos prácticos se podrán adivinar los pesos, hacer el análisis para después volver a estimar los pesos con base en los resultados. Pueden ser necesarias varias iteraciones.

En el caso ideal, tenemos varias observaciones de respuesta y_i para los mismos valores en las variables regresoras X. Si la varianza está en función de las variables regresoras, entonces se puede estimar esta función calculando la varianza para cada subconjunto de observaciones repetidas, j, es decir $V(y_{ji})$, y posterior graficar la varianza contra X para estimar la función. Si X es multiparamétrico, se requiere de otros métodos de estimación distintos a la gráfica. Los pesos w_i serán el inverso de la varianza estimada $\hat{V}(y_j)$.

5. Diagnóstico para balanceo e influencia

Al estimar una recta de regresión es posible que todos los supuestos necesarios se cumplan y aún así tener puntos que tienen una fuerte influencia en los valores de beta estimada, $\hat{\beta}$. Estos puntos no necesariamente son observaciones atípicas y para su identificación se requieren nuevas técnicas. Responder a la cuestión de si se deben eliminar o no es análogo a la misma pregunta para datos atípicos. Si existe evidencia de que este punto es un error de medición, se puede eliminar sin mayor remordimiento; si es un valor válido, entonces su eliminación crearía valores estimados no representativos de la realidad. Un **punto de influencia** Fig. 15 es aquel que modifica de manera significativa los valores de los parámetros estimados $\hat{\beta}$; Un **punto de balanceo** Fig. 16, por el contrario, tiene poco efecto en $\hat{\beta}$, pero sí modifica los estadísticos de resumen, como R^2.

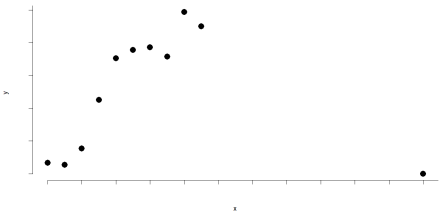


Figura 15: Punto de influencia

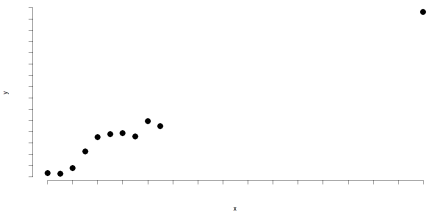


Figura 16: Punto de balanceo

5.1. Métodos de diagnóstico

Como $V(\hat{Y}) = \sigma^2 H$ y $V(e) = \sigma(I-H)$, donde $H = X(X'X)^{-1}X'$ es la matriz sombrero, podemos ver que H determina las varianzas y covarianzas de $\hat{Y}$ y de e, de tal manera que resulta razonable considerar los elementos h_{ij} de la matriz H como los elementos de balanceo que tiene la i-ésima observación, y_i, sobre el i-ésimo valor ajustado, $\hat{y}_i$. La diagonal principal de H es una medida estandarizada de la distancia de la i-ésima observación al centro del espacio X de tal manera que si $h_{ii} = x_i'(X'X)^{-1}x_i$ es grande, entonces estamos ante un posible punto influyente por la lejanía al centro.

El **método de la matriz sombrero** nos ayuda a identificar posibles puntos de balanceo. Dado que el valor promedio de un elemento diagonal es $\bar{h} = p/n$ (porque $\sum_{i=1}^{n} h_{ii} = ran(H) = ran(X) = p$), se puede considerar una distancia, digamos $2p/n$, desde la cual podemos acusar a una observación de estar modificando las estadísticas de resumen. Para ver la influencia de este punto (o puntos), podemos calcular $\hat{\beta}$, $\hat{\sigma}^2$, R^2 y otros, con y sin dichas observaciones para ver que tanto varían sus valores. Este método es recomendado para muestras grandes donde, al menos, $2p/n < 1$.

La D **de Cook** es una medida de influencia propuesta por Cook [1977,1979] y es, precisamente, una formalización de la idea de distancia entre el los parámetros estimados con y sin la observación potencialmente influyente. En forma general, se esxpresa como

$$D_i(M,c) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' M (\hat{\beta}_{(i)} - \hat{\beta})}{c}, \quad \text{para} \quad i = \overline{1,n} \tag{96}$$

Es usual la utilización de $M = X'X$ y $c = p\hat{\sigma}^2$, de tal manera que

$$D_i(M,c) = D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X'X (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2}, \quad \text{para} \quad i = \overline{1,n} \tag{97}$$

Si el investigador tiene conocimiento de un conjunto de datos que son posiblemente influyentes, esta medida sigue siendo válida con la salvedad de que $\hat{\beta}_{(i)}$ ahora será el vector beta estimado sin considerar las m observaciones influyentes. Para valores grandes de D_i se tiene mayor influencia del i-ésimo elemento (o conjunto de elementos). Si $D_i = F^{0.5}_{(p,n-p)}$, entonces al eliminar el punto i se movería $\hat{\beta}_{(i)}$ hacia la frontera de una región de confianza aproximada del 50 % para β, basándose en el conjunto completo de datos. Sabiendo que $F^{0.5}_{(p,n-p)} \approx 1$, podemos considerar a un punto i como influyente si $D_i > 1$. Aunque se ha asumido que D_i es una estadística F y que $F^{0.5}_{(p,n-p)} \approx 1$, esto es matemáticamente incorrecto. Sin embargo, en la práctica resulta ser conveniente en la mayoría de los casos (pero no en todos).

La estadística D_i se puede expresar como

$$D_i = \frac{r_i^2}{p}\frac{V(\hat{y}_i)}{V(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$$

Donde r_i es el i-ésimo residual estudentizado. De esta manera se puede observar como D_i esta formada por un componente que refleja lo bien que se ajusta el modelo a la i-ésima observación y un componente que mide lo alejado está el punto dentro del resto de datos. Recuerde que h_{ii} representa la ubicación de ese punto en el espacio de X y r_i un residual, ambos componentes de D_i, que evalúan la influencia de y_i.

Una tercera forma de expresar a D_i es con el cuadrado de la distancia euclidiana del vector de los valores ajustados con y sin y_i. Es decir

$$D_i = \frac{(X\hat{\beta}_{(i)} - X\hat{\beta})'(X\hat{\beta}_{(i)} - X\hat{\beta})}{p\hat{\sigma}^2} = \frac{(\hat{y}_{(i)} - \hat{y}_i)'(\hat{y}_{(i)} - \hat{y}_i)}{p\hat{\sigma}^2}$$

Aunque pareciera que se necesitan calcular n regresiones, cada una para una observación omitida, esto no es necesario utilizando la Ec. (77).

De manera análoga, **DFBETAS** es una estadística que indica cuánto cambia el coeficiente de regresión $\hat{\beta}_j$, en unidades de desviación estándar, si se omitiera la i-ésima observación. Esta estadística es

$$DES_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)} C_{jj}}} \tag{98}$$

siendo $\hat{\sigma}^2_{(i)}$ el cuadrado medio de los residuales con omisión de la observación y_i, C_{jj} el j-ésimo elemento diagonal de $(X'X)^{-1}$ y $\hat{\beta}_{j(i)}$ el j-ésimo coeficiente estimado sin y_i. $DES_{j,i}$ es una matriz de $n \times p$ cuyos elementos indican la influencia de cada observación en cada coeficiente estimado de la regresión. A valores grandes, mayor influencia.

Supongamos la relación

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(X'X)^{-1}x_i e_i}{1-h_{ii}} \tag{99}$$

$DES_{j,i}$ es el j-ésimo elemento de $\hat{\beta} - \hat{\beta}_{(i)}$ dividido entre un factor de estudentización. Por otro lado

$$\hat{\beta}_j - \hat{\beta}_{j(i)} = \frac{r_{j,i} e_i}{1-h_{ii}}$$

Definamos a $R = (X'X)^{-1}X$, por lo que

$$(RR')' = \left[(X'X)^{-1}X'X(X'X)^{-1}\right]' = (X'X)^{-1} = C = R'R$$

Por lo anterior, $C_{jj} = r_j' r_j$ y entonces se puede escribir el factor de estandarización como

$$\sqrt{\hat{\sigma}^2_{(i)} C_{jj}} = \sqrt{\hat{\sigma}^2_{(i)} r_j' r_j}$$

De tal manera que una forma computacional de escribir a DES es

$$\begin{aligned} DES_{j,i} &= \left[\frac{r_{j,i} e_i}{1 - h_{ii}} \frac{1}{\sqrt{\hat{\sigma}^2_{(i)} r_j' r_j}} \right] \\ &= \frac{r_{j,i}}{\sqrt{r_j' r_j}} \frac{t_i}{\sqrt{1 - h_{ii}}} \end{aligned}$$

En donde t_i es el residual estudentizado. Se sugiere un punto de corte $|DES| > 2/\sqrt{n}$ para considerar a un punto como influyente.

También se puede investigar la influencia de la eliminación de la i-ésima observación sobre el valor predicho o ajustado a través del método $DFFITS$ y basta con sustituir el numerador en la Ec. (98) por $\hat{y}_i - \hat{y}_{(i)}$, resultando en

$$DIS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} C_{jj}}} \tag{100}$$

DIS_i es la cantidad de desviaciones estándar que cambia el valor ajustado $\hat{y}_i$ si se elimina la observación i. Ahora, si en la Ec. (99) multiplicamos ambos lados por x_i', se obitiene

$$\hat{y}_i - \hat{y}_{(i)} = \frac{h_{ii} e_i}{1 - h_{ii}}$$

Ambos lados de esta ecuación se multiplican por el factor $\sqrt{\hat{\sigma}^2_{(i)} C_{jj}}$ para obtener DIS_i

$$\begin{aligned} DIS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)} C_{jj}}} &= \frac{h_{ii} e_i}{1 - h_{ii}} \frac{1}{\sqrt{\hat{\sigma}^2_{(i)} C_{jj}}} \\ &= \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)} (1 - h_{ii})}} \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} \\ &= t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} \end{aligned}$$

Esta forma es más sencilla al momento de hacer cálculos. Si el dato es atípico, el residual R de Student tendrá magnitud grande, mientras que si el dato tiene gran balanceo, h_{ii} se aproximará a la unidad. En cualquiera de esos casos, DIS_i puede ser grande. Se sugiere el punto de corte $|DIS_i| > 2\sqrt{p/n}$.

Una **medida del desempeño del modelo** se puede obtener usando los métodos anteriores, DFBETAS y DFFITS. Primero, definamos la varianza generalizada de $\hat{\beta}$ como

$$GV(\hat{\beta}) = |V(\hat{\beta})| = |\sigma^2 (X'X)^{-1}|$$

y la razón de covarianzas como

$$COVRATIO_i = \frac{|(X'_{(i)}X_{(i)})^{-1}\hat{\sigma}^2_{(i)}|}{|(X'X)^{-1}\hat{\sigma}^2|} \quad i = \overline{1,n}$$

Que representa la precisión de la estimación en la i-ésima observación. Note que para un $COVRATIO_i > 1$, la i-ésima observación mejora la precisión de la estimación, mientras que para valores menores a la unidad ocurre lo contrario. Una manera del $COVRATIO_i$ análoga útil para los cálculos es

$$COVRATIO_i = \frac{(\hat{\sigma}^2_{(i)})^p}{(\hat{\sigma}^2)^p}\left(\frac{1}{1-h_{ii}}\right)$$

este nuevo estadístico también es útil para detectar puntos influyentes para muestras los suficiente grandes (al menos $n > 3p$). Si $|COVRATIO_i - 1| > 3p/n$, entonces es un buen candidato a punto influyente.

6. Modelos polinomiales de regresión

Suponga que existe una función determinista f que tiene derivadas de cualquier orden en el punto $x = a$, entonces la **serie de Taylor** para la función f en el punto a es

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots$$

Si limitamos este resultado a n+1 términos obtenemos el n-ésimo polinomio de Taylor para f en el punto a, que es

$$p_n(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!} + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{101}$$

Dicho polinomio nos proporciona una aproximación a la función f, siempre que conozcamos su valor en $x = a$ para cualquier derivada de orden n o menor. Por ejemplo, si deseamos conocer p_3 para $f(x) = \ln x$ en $x = 1$ primero debemos derivar y evaluar:

$$\begin{aligned} f(x) &= \ln x & f(1) &= 0 \\ f'(x) &= \frac{1}{x} & f'(1) &= 1 \\ f''(x) &= -\frac{1}{x^2} & f''(1) &= -1 \\ f'''(x) &= \frac{2}{x^3} & f'''(1) &= 2 \end{aligned}$$

usando lo anterior y sustituyendo en (101)

$$p_3(x) = 0 + 1 \cdot (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 \tag{102}$$

Podemos ver esta aproximación en la Fig. (17 que el polinomio toma valores cercanos a la función en el punto $x = 1$. Para un polinomio de mayor grado la aproximación es válida para puntos más lejanos. Es posible que la utilidad de esta aproximación no parezca útil a primera instancia, pero sucede lo contrario. De manera computacional es más eficiente evaluar funciones polinomiales que de senoidales o logarítmicas; si no se conoce la función $f(x)$, es posible aproximarla por este polinomio; es posible definir funciones por series infinitas de Taylor.

Si se agrega la aleatoriedad a esta función basta con agregar el error de tal manera que

$$E(y) = p(x) + \varepsilon$$

donde $p(x)$ es un polinomio de grado n de la forma

$$p(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

Que resulta análogo al polinomio de Taylor Maclaurin (en el punto $x = 0$). Con esta definición es posible establecer una aproximación a cualquier función (de existir) sin conocerla estimando los parámetros por mínimos cuadrados como se ha hecho hasta ahora salvo las siguientes advertencias.

- Orden del modelo. Como es usual, información previa sobre el fenómeno a presentar un mejor modelo. De tal manera que si se conoce, por ejemplo, que la relación entre X y Y es cúbica, se tiene una razón justificada para tener un polinomio de tercer orden. De cualquier otra forma se sugiere no tener polinomios del menor grado posible (sentido de parsimonia) para evitar un sobre ajuste. Recuerde que siempre existe un polinomio de grado $n-1$ que pasa exactamente por n puntos.
- Extrapolación. Otro aspecto importante es entender las limitaciones de la estimación. Cuando no es clara la relación entre las variables regresoras y la dependiente es tener un buen ajuste para las observaciones y aún así tener muy mala capacidad predictiva. Por ejemplo, supongamos que la relación funcional real es $f(x) = \ln x$ más un error de medición ε y que se estima con un polinomio de tercer grado cuyos parámetros estimados $\hat{\beta}$ coinciden con (102). Se puede observar en la Fig. (18) que el ajuste en el intervalo [0,2] es suficiente bueno, pero la realidad es que para valores fuera de este intervalo el polinomio presenta un comportamiento cada vez más alejado a $f(x)$.
- Inexactitud y multicolinealidad. Cuando se introducen variables regresoras que son función de otras variables regresoras existentes es posible caer un mal acondicionamiento en el sentido de $(X'X)^{-1}$ tendrá un cálculo inexacto. Una posible (pero no infalible) solución es centrar las variables a su promedio. Nuevamente se puede ver esta analogía con la Ec. (101) para $a = \bar{x}$. Otro mal acondicionamiento es el de multicolinealidad, presente cuando el intervalo de valores de X es estrecho.
- Jerarquía y construcción del modelo. Un modelo jerárquico es aquel que contiene

$$\sum_{j=1}^{m}\sum_{k=1}^{n}\beta_k x_j^k$$

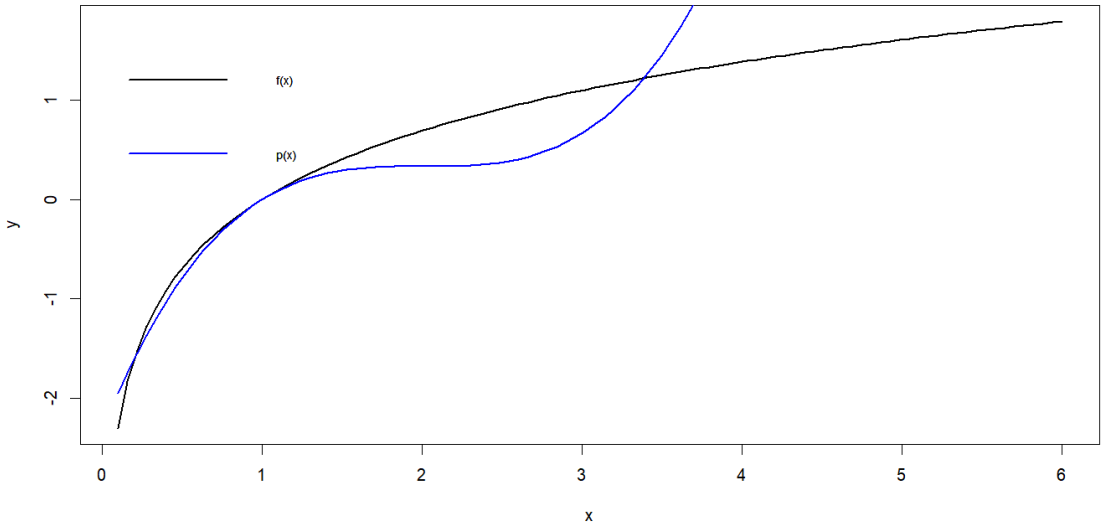


Figura 17: Aproximación por polinomio de Taylor.

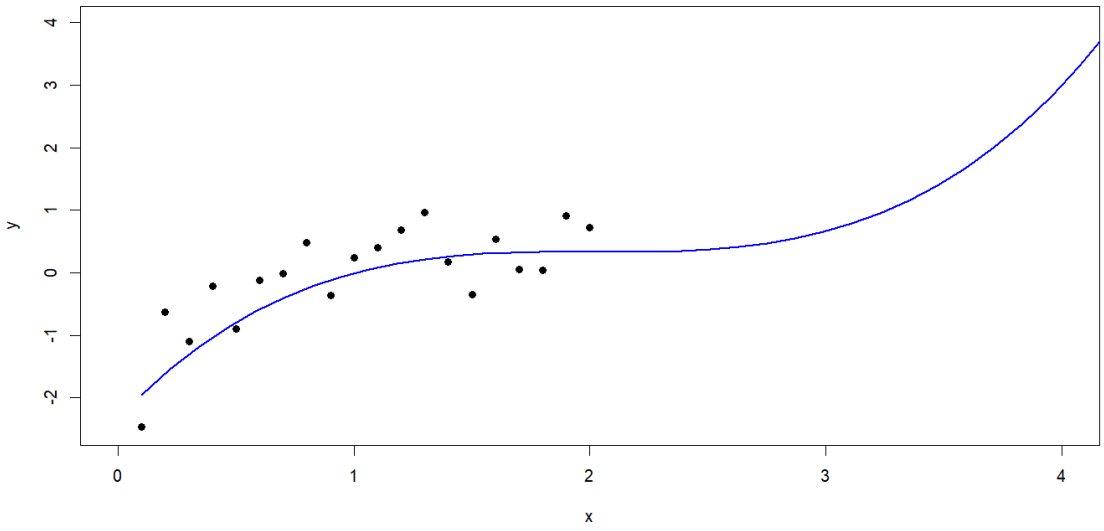


Figura 18: Estimación polinomial de $Y = \ln X + \varepsilon$

en el polinomio de grado n con m variables regresoras, además del producto de todas las posibles combinaciones de las m variables. Por ejemplo, Supongamos que existen dos variables x_1 y x_2, entonces nuestro modelo sería de la forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 x_2 + \beta_4 x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

Una estrategia de construcción del modelo que se hace en [11] es ajustar en forma sucesiva modelos de orden creciente hasta que la prueba t para el término de orden máximo sea no significativa o ajustar el modelo de orden máximo adecuado, y a continuación eliminar términos, uno por uno, comenzando con el de orden máximo hasta que el término que quede de orden máximo tenga una estadística t significativa. Esos dos procedimientos se llaman selección en avance y eliminación en reversa, respectivamente, no necesariamente conducen al mismo modelo.

Ejemplo en [11]

El cuadro (6) presenta datos acerca de la resistencia del papel kraft y el porcentaje de madera dura en el lote de pulpa con el que se fabricó. En la Fig. (19) se ve el diagrama de dispersión para esos datos que muestra a simple vista una relación cuadrática. Siguiendo el consejo que se sugiere para la inexactitud de $(X'X)^{-1}$, centramos los datos con la variable $(x - \bar{x})$ (Fig. 20) de tal manera que el modelo a priori es

$$y = \beta_0 + \beta_1 (x - \bar{x}) + \beta_2 (x - \bar{x})^2 + \varepsilon$$

Nuestro modelo estimado resulta en

$$\hat{y} = 45.2949731 + 2.5463440 \cdot (x - 7.2632) - 0.6345492 \cdot (x - 7.2632)^2$$

	X	**Y**	$X - \bar{X}$
1	1.0	6.3	-6.2631579
2	1.5	11.1	-5.7631579
3	2.0	20.0	-5.2631579
4	3.0	24.0	-4.2631579
5	4.0	26.1	-3.2631579
6	4.5	30.0	-2.7631579
7	5.0	33.8	-2.2631579
8	5.5	34.0	-1.7631579
9	6.0	38.1	-1.2631579
10	6.5	39.9	-0.7631579
11	7.0	42.0	-0.2631579
12	8.0	46.1	0.7368421
13	9.0	53.1	1.7368421
14	10.0	52.0	2.7368421
15	11.0	52.5	3.7368421
16	12.0	48.0	4.7368421
17	13.0	42.8	5.7368421
18	14.0	27.8	6.7368421
19	15.0	21.9	7.7368421

Cuadro 8: Datos de resistencia del papel Kraft y el porcentaje de madera dura.

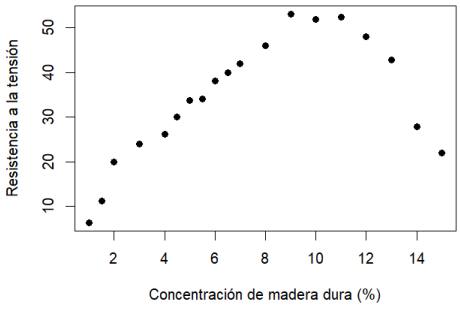


Figura 19: Diagrama de dispersión. $Y \sim X$

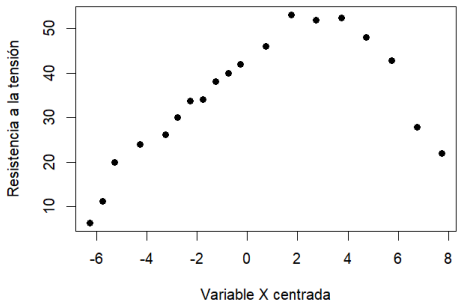


Figura 20: Diagrama de dispersión. $Y \sim (X - \bar{X})$

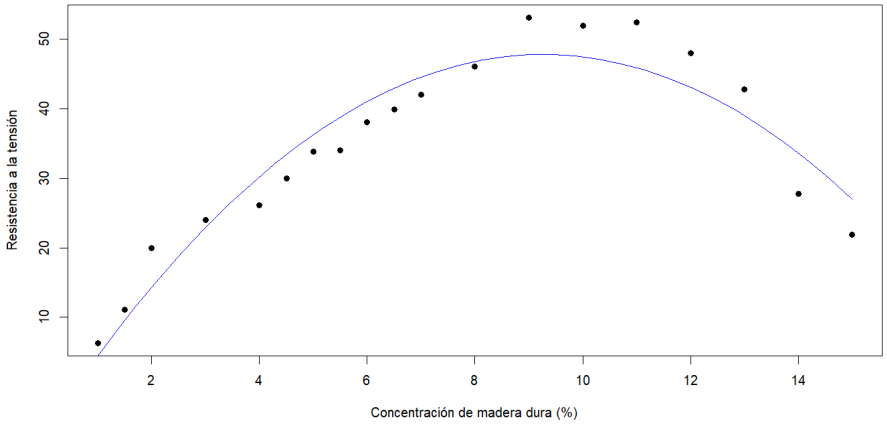


Figura 21: Estimación polinomial por mínimos cuadrados.

Note el lector que de haber estimado el modelo para X no centrada el resultado sería el mismo (después de reducir la expresión anterior)

$$\hat{y} = -6.6741916 + 11.7640057x - 0.6345492x^2$$

Entonces, centrar o no la variable es sólo para omitir la inexactitud de la inversión de $X'X$, pero resultan en el mismo $\hat{\beta}$. podemos ver el buen ajuste de nuestro modelo en la Fig. (21), con un valor $F = 79.43$ suficiente alto para que el modelo sea significativo, al igual que las pruebas t para cada variable y un R^2 ajustada de 0.8971. Si se analizan los residuales se puede ver que no existe un problema serio de los supuestos del modelo para aplicar mínimos cuadrados.

6.1. Ajuste polinomial por segmentos

Suponga que el aumento del grado polinomial no refleja un ajuste significativo en las variables. Para mejorar el modelo se puede hacer una transformación de las variables o el método de **funciones SPLINE**, que se aborda en este capítulo.

Ahora, se busca hacer un modelo polinomial para segmentos de datos que parezcan tener un comportamiento similar. Los puntos de unión entre segmento y segmento serán llamados nudos. Para darle continuidad a la función que se arma por trozos es conveniente que las primeras $k-1$ derivadas concuerden en los nudos, como amarrar de los extremos tramos de listón, de ahí el sentido de nudo. En [11] se sugiere que la spline cúbica ($k = 3$) es adecuada para la mayor parte de los problemas prácticos.

En general, una función spline cúbica con h nudos, ubicados en $t_1, t_2, \ldots, t_h$ de tal manera que sean ordenados, $t_1 < t_2 < \ldots < t_h$, con primera y segunda derivada continuas, se puede escribir como

$$E(y) = S(x) = \sum_{j=0}^{3} \beta_{0,j} x^j + \sum_{i=1}^{h} \beta_i (x - t_i)_+^3 \tag{103}$$

Donde

$$(x-t_i)_+ = \begin{cases} (x-t_i), & x-t_i > 0 \\ 0, & x-t_i \geq 0 \end{cases}$$

Decidir la posición y el número de nudos no resulta trivial, sin embargo, Wold[1974] sugiere que debería haber el menor número de nudos posibles con al menos 5 puntos en cada segmento, teniendo cuidado con el sobre ajuste que puede ocasionar el método SPLINE. También se sugiere que cada nudo debe tener a lo más un máximo (o mínimo) centrado y un punto de inflexión cerca del nudo. A continuación se muestra un código en lenguaje R para la Ec. (103) con su respectiva gráfica. los parámetros de beta y los nudos t se sugieren, pero no son correspondientes a un cálculo por mínimos cuadrados.

```
#library(dplyr)
b<-seq(-10,10, length=7) #valores beta
t<-c(1,2,3) #valores de t_i

x<-seq(1,15,by=0.1) #valores sobre los que corre x

y<-function(x){ #funcion spline cubica
  case_when(
    x<t[1]~sum(b*c(1,x,x^2,x^3,0,0,0)),
    x<t[2]~sum(b*c(1,x,x^2,x^3,(x-t[1])^3,0,0)),
    x<t[3]~sum(b*c(1,x,x^2,x^3,(x-t[1])^3,(x-t[2])^3,0)),
    TRUE~sum(b*c(1,x,x^2,x^3,(x-t[1])^3,(x-t[2])^3,(x-t[3])^3))
  )
}

w<-numeric(length = (length(x))) #vector vacio

for (i in seq(1,length(x))){ #llenado del vector w=f(x)=y
  w[i]<-y(x[i])
}

plot(x,w, type="l", ylab = "E(Y)", xlab = "X") #grafica
```

Note que los términos de la función spline en la Ec. (103) sólo tienen un término cúbico en cada nudo, sin embargo esto se puede ajustar fácil para tener un polinomio en cada nudo

$$E(y) = S(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \sum_{j=0}^{3} \beta_{ij} (x-t_i)_+^j$$

La nueva cuestión es la discontinuidad que causa en la función estos nuevos términos. Por ejemplo, el término $\beta_{ij}(x-t_i)_+^j$ causa discontinuidad de la j-ésima derivada de $S(x)$ en t_i y eliminar este término se considera como restricción de continuidad. Mientras menores restricciones de continuidad se requieran, el ajuste es mejor, porque en el modelo habrá más parámetros; mientras que cuanto más restricciones de continuidad se requieran el ajuste será peor, pero la curva final será más uniforme. Se puede determinar. Para conocer si es conveniente o no incluir estos términos se puede hacer una prueba de significancia $\beta_{ij} = 0$ para los subíndices de interés.

Para solucionar el problema del mal acondicionamiento de la matriz $X'X$ se crea una representación distinta de la función spline, llamada **spline B cúbica** que se define en función de diferencias divididas

$$\beta_i(x) = \sum_{j=i-4}^{i} \left[\frac{(x-t_j)_+^3}{\prod_{m=i-4}^{i}(t_i - t_m)} \right] \quad \text{para} \quad i = \overline{1, h+4} \text{ y } m \neq j$$

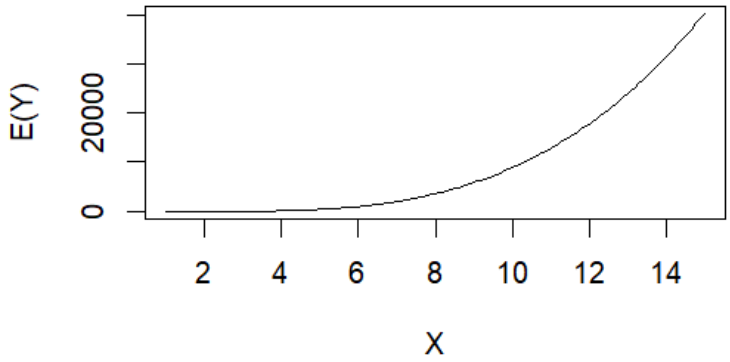


Figura 22: gráfica de la Ec. (103) para valores arbitrarios

siendo

$$E(y) = S(x) = \sum_{i=1}^{h+4} \gamma_i \beta_i(x)$$

en donde $\gamma_i = \overline{1, h+4}$ son parámetros a estimar.

Ejemplo en [11]

La caída de voltaje en la batería del motor de un misil guiado, que se observa durante el tiempo de vuelo del misil, se muestra en el cuadro (6.1). El diagrama de dispersión de la Fig. (23) parece indicar que la caída de voltaje se comporta en forma distinta en diferentes intervalos de tiempo, por lo que se modelarán los datos con una spline cúbica usando dos nudos en $t_1 = 6.5$ y $t_2 = 13$ segundos después del lanzamiento, respectivamente. La colocación de los nudos concuerda en forma aproximada con los cambios de curso del proyectil (con los cambios asociados en necesidades de energía), que se conocen por los datos de la trayectoria. De tal manera que el modelo es de la forma

$$y = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \beta_1(x-6.5)_+^3 + \beta_2(x-13)_+^3 + \varepsilon$$

Note que este modelo tiene restricciones de continuidad y aún así se verá que es eficiente. Sería interesante que el lector pruebe la eficiencia del modelo sin dichas restricciones. El código en lenguaje R para este problema se muestra a continuación. Se puede apreciar el buen ajuste del modelo para el vector β en las pruebas t de significancia individual, además de un valor p para la prueba F de casi cero, lo que indica la significancia del modelo completo. El nivel de explicativo del modelo R ajustada toma un valor de 0.98 igualmente satisfactorio. Además, si se aplican pruebas de normalidad en los errores o la homogeneidad de las varianzas tampoco se detecta un problema grave.

```
library(readxl)
library(dplyr)
volt<-read_xlsx("Ejemplo 7.2 Montgomery.xlsx")
volt<-volt %>% mutate(x2=Segundos^2,x3=Segundos^3,
```

Observación	Segundos x_i	Caída del voltaje y_i
1	0	8.33
2	0.5	8.23
3	1	7.17
4	1.5	7.14
5	2	7.31
6	2.5	7.6
7	3	7.94
8	3.5	8.3
9	4	8.76
10	4.5	8.71
11	5	9.71
12	5.5	10.26
13	6	10.91
14	6.5	11.67
15	7	11.76
16	7.5	12.81
17	8	13.3
18	8.5	13.88
19	9	14.59
20	9.5	14.05
21	10	14.48
22	10.5	14.92
23	11	14.37
24	11.5	14.63
25	12	15.18
26	12.5	14.51
27	13	14.34
28	13.5	13.81
29	14	13.79
30	14.5	13.05
31	15	13.04
32	15.5	12.6
33	16	12.05
34	16.5	11.15
35	17	11.15
36	17.5	10.14
37	18	10.08
38	18.5	9.78
39	19	9.8
40	19.5	9.95
41	20	9.51

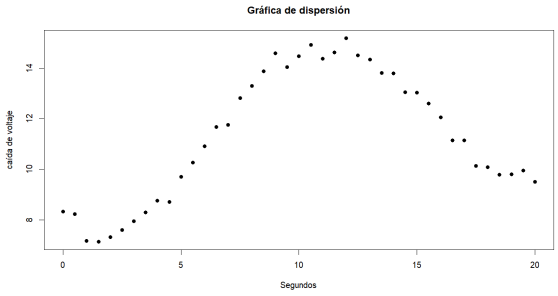


Figura 23: Gráfica de dispersión para la caída de voltaje

```
                        t1=case_when(
                          Segundos<6.5~0,
                          Segundos>=6.5~(Segundos-6.5)^3,
                        ),
                        t2=case_when(
                          Segundos<13~0,
                          Segundos>=13~(Segundos-13)^3
                        ))


modelo<-lm(Voltaje~Segundos+x2+x3+t1+t2,data=volt)
summary(modelo)

#Consola:

Residuals:
     Min       1Q   Median       3Q      Max
-0.45168 -0.18499 -0.03547  0.20577  0.61694

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.465678   0.200520  42.219  < 2e-16 ***
Segundos    -1.453124   0.181586  -8.002 2.04e-09 ***
x2           0.489889   0.043018  11.388 2.54e-13 ***
x3          -0.029467   0.002848 -10.347 3.44e-12 ***
t1           0.024706   0.004039   6.116 5.43e-07 ***
t2           0.027112   0.003578   7.577 6.98e-09 ***
---

Residual standard error: 0.2678 on 35 degrees of freedom
Multiple R-squared:  0.9904,  Adjusted R-squared:  0.9891
F-statistic: 725.5 on 5 and 35 DF,  p-value: < 2.2e-16
```

6.2. Regresión no paramétrica

Una estadística se dice paramétrica cuando se conoce la distribución de los datos sus parámetros, salvo un número finito de ellos que conciernen a la prueba de hipótesis. Por otro lado, la estadística no paramétrica hacer referencia a un conjunto de datos de los cuales desconocemos la distribución. En [14] se menciona que las pruebas estadísticas no paramétricas son casi tan eficaces como los métodos paramétricos. Los siguientes

métodos apelan más al empirismo y en [11] se sugiere preferir una modelo paramétrico si éste presenta un buen ajuste. La gama de análisis disponible es mayor para la regresión lineal simple al tener la posibilidad de presentar pruebas tipo F o t para un sin fin de hipótesis, además de intervalos de confianza para los parámetros y de predicción para nuevas observaciones.

6.2.1. Regresión Kernel

El **alisador de Kernel** crea una banda de ancho b (arbitrario a elección del investigador) mediante pesos w_i (donde $\sum_{j=1}^{n} w_{ij} = 1$), de tal manera que para valores observados de respuesta fuera de la banda el peso es casi cero.

$$\tilde{y}_i = \sum_{j=1}^{n} w_{ij} y_i \tag{104}$$

Para especificar los pesos w_{ij} se utiliza la función kernel que puede ser cualquiera con las propiedades de una función de densidad simétrica, es decir

- $K(t) \geq 0$ para todo t
- $\int_{-\infty}^{\infty} K(t)dt = 1$
- $K(-t) = K(t)$

con lo anterior podemos definir

$$w_{ij} = \frac{K\left(\frac{x_i - x_j}{b}\right)}{\sum_{k=1}^{n} K\left(\frac{x_i - x_k}{b}\right)}$$

El concepto de kernel es importante para un vector supervisado de regresión. Aquí de de una breve explicación. El vector supervisado de regresión es un caso particular de un support vector machine (SVM), es por ello que resulta conveniente, primero, entender este tema. SVM es un algoritmo supervisado de aprendizaje que intenta predecir valores basado en la clasificación o la regresión analizando los datos y reconociendo patrones. El algoritmo que se usa para clasificaciones se llama SVC (support vector classifier) y para regresión es SVR (support vector regression). Las principales ventajas del SVR es que la complejidad computacional no depende de la dimensionalidad espacial de los datos de entrada y tiene una excelente capacidad de generación, con alta precisión de predicción.

SVR formula un problema de aproximación de funciones como un problema de optimización que intenta encontrar el tubo más estrecho centrado alrededor de la superficie, minimizando al mismo tiempo el error de predicción, es decir, la distancia entre los resultados previstos y los deseados. La primera condición produce la función objetivo en Ecuación 105, donde $\| w \|$ es la magnitud del vector normal (perpendicular) a la superficie que se está aproximando.

$$min_w \quad \frac{1}{2} \| w \|^2 \tag{105}$$

Supóngase que se está dando el conjunto de datos de entrenamiento $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset \chi \times R$, donde χ denota el espacio de las entradas ($\chi = R^d$). El objetivo es encontrar la función $f(x)$ que tiene a lo mucho ε desviación estándar de los objetivos realmente obtenidos y_i para todos los datos de entrenamiento.

Este concepto es más claro si se trabaja con un vector unidimensional Fig. 24. La función de valor continuo que se aproxima se puede escribir como en la ecuación 106, donde $< \cdot, \cdot >$ representa el producto

punto dentro de χ. En el caso de los datos multidimensionales, se aumenta x en uno y se incluye b en el vector w simplemente a la notación matemática, y se obtiene la regresión multivariante en la ecuación 107.

$$y = f(x) = < w,x > +b = \sum_{j=1}^{M} w_j x_j + b \quad y,b \in R, x,w \in R^m \tag{106}$$

$$f(x) = \begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + b \quad x,w \in R^{M+1} \tag{107}$$

Se expresa la función a minimizar en la Ec. (105) $||w||^2 = < w,w >$ como un problema de optimización convexa

$$min_w \frac{1}{2}||w||^2$$

Sujeto a

$$\begin{cases} y_i - < w,x_i > -b \leq \varepsilon \\ < w,x_i > +b - y_i \leq \varepsilon \end{cases} \tag{108}$$

La suposición en la Ec. (108) fue que realmente existe una función f que aproxima todos los pares (x_i, y_i) con ε de precisión. Sin embargo, puede que queramos algunos errores fuera del tubo delimitado por $\pm\varepsilon$. En este caso, introducimos las variables de holgura ξ_i, ξ_i^* llegando a la fórmula

$$min_w \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^2) \tag{109}$$

Sujeto a

$$\begin{cases} y_i - < w,x_i > -b \leq \varepsilon + \xi_i \\ < w,x_i > +b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{110}$$

Donde $C > 0$ determina la compensación entre la plenitud de f y la cantidad hasta la cual las desviaciones ε son toleradas. Esto corresponde a la función de pérdida de ε-insensibilidad $|\xi|_\varepsilon$ descrita por

$$|\xi|_\varepsilon = \begin{cases} 0 \quad \text{si} \quad |\xi| \leq \varepsilon \\ |\xi| - \varepsilon \quad \text{en otro caso} \end{cases} \tag{111}$$

Para entender la magnitud de los pesos como una medida de plenitud, veamos el siguiente ejemplo para una función en R^2:

$$f(x,w) = \sum_{i=1}^{M} w_i x^i, \quad x \in R,\ w \in R^M$$

Note que M es el orden polinomial de la función que usamos para aproximar los datos. Si $M = 0$, entonces $f(x,w) = w_1$, es decir, alguna constante que minimice la distancia de los puntos a la recta $y = w_1$. Para $M = 3$ de mayor orden obtendríamos $f(x,w) = w_1 x + w_2 x^2 + w_3 x^3$. Con forme la magnitud del vector w incrementa, un mayor número de w_i se vuelven diferentes de cero, resultando en soluciones de mayor orden. En la Figura 25 se puede apreciar la solución estimada para distintos órdenes. Note que el polinomio de grado 0 (u orden cero) es una línea horizontal que intenta cortar los datos por la mitad, sin embargo, genera un error (distancia de los puntos a la recta) grande. Para el primer orden se tiene una línea diagonal y para

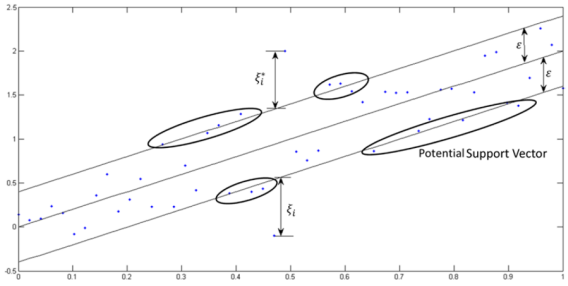


Figura 24: Caso del vector unidimensional

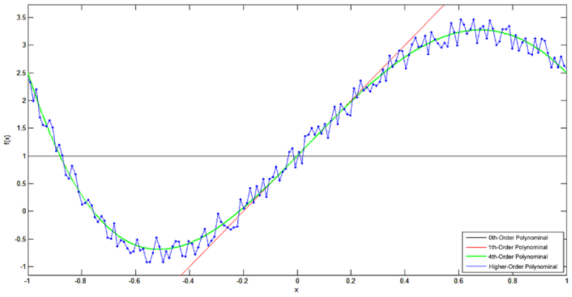


Figura 25: Soluciones para distintos órdenes polinomiales.

los demás órdenes se sigue el mismo análisis. la magnitud de w actúa como un regularizador de términos y provee un control en la optimización del problema sobre la plenitud del problema. La restricción objetivo es minimizar el error entre el valor predicho de la función para una entrada determinada y la salida real. SVR adopta una función de pérdida insensible al error ε, penalizando las predicciones que están más lejos que ε del resultado deseado, es decir ξ_i. En la Figura 24 se puede apreciar que el error ε crea dos rectas por arriba y debajo. El valor de ε determina el ancho del tubo; un valor menor indica una menor tolerancia al error y también afecta el número de vectores de soporte y, en consecuencia, la escasez de la solución.

Debido a que es menos sensible a entradas ruidosas, la región insensible a ε hace que el modelo sea más robusto. Se pueden adoptar varias funciones de pérdida, incluidas la lineal, la cuadrática y la de Huber. Como se demuestra en la Figura 26, la función de pérdida de Huber es más suave que las funciones lineal y cuadrática, pero penaliza todas las desviaciones de la salida deseada, con una penalización mayor a medida que aumenta el error. Las funciones de pérdida presentadas aquí son simétricas y convexas. Aunque se pueden adoptar funciones de pérdida asimétricas para limitar la subestimación o la sobreestimación, las funciones de pérdida deben ser convexas para garantizar que el problema de optimización tenga una solución única que pueda encontrarse en un número finito de pasos.

$$L_\varepsilon(y, f(x,w)) = \begin{cases} 0 \quad \text{para} \quad |y - f(x,w)| \leq \varepsilon \\ |y - f(x,w)| - \varepsilon \quad \text{en otro caso} \end{cases}$$

$$L_\varepsilon(y, f(x,w)) = \begin{cases} 0 & \text{para} \quad |y - f(x,w)| \leq \varepsilon \\ (|y - f(x,w)| - \varepsilon)^2 & \text{en otro caso} \end{cases}$$

$$L(y, f(x,w)) = \begin{cases} c|y - f(x,w)| - \frac{c^2}{2} & \text{para} \quad |y - f(x,w)| > c \\ \frac{1}{2}|y - f(x,w)|^2 & \text{para} \quad |y - f(x,w)| \leq c \end{cases}$$

Que son, respectivamente,

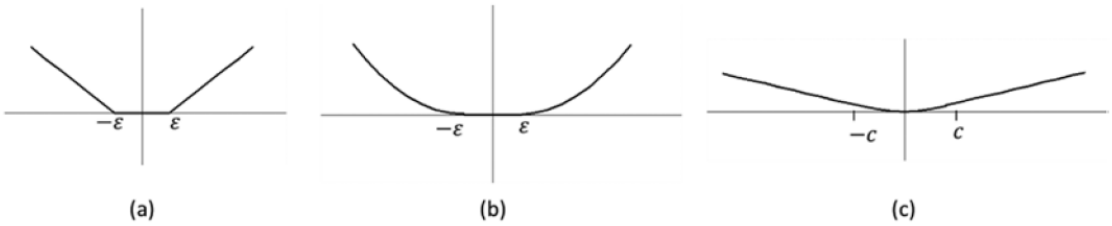


Figura 26: tipos de función de pérdida: (a) lineal, (b) cuadrática, (c) Huber

Si los datos presentan una función que no es lineal, se puede hacer un mapeo a un espacio dimensional mayor Fig 27, llamado **espacio kernel**, para alcanzar un mayor ajuste. Por ejemplo, considere el mapeo $\Phi : R^2 \to R^3$ con $\Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$. Posterior a estimar los pesos en una dimensión superior, se regresa a la original. Tres tipos de kernel de uso común son:

- Polinomial:
$$K(x,w) = (x^T w + b) \tag{112}$$
- Radial Basis (Gaussiano):
$$K(x,w) = exp\left(\frac{|x-w|^2}{2\sigma^2}\right)$$
- Exponencial
$$K(x,w) = exp\left(-\frac{|x-w|}{\sigma}\right)$$

6.2.2. Regresión ponderada localmente (Loess)

La idea de datos "próximos.[en] la regresión kernel se retoma para Loess. Esta proximidad se define el tramo, que es la fracción de los puntos totales que se utilizan para formar las proximidades. El procedimiento loess usa entonces los puntos en la proximidad para generar un estimado por mínimos cuadrados ponderados, de la respuesta específica. El procedimiento de mínimos cuadrados ponderados usa un polinomio de bajo orden, que suele ser la regresión lineal simple, o un modelo cuadrático de regresión [11]. Recordemos que para mínimos cuadrados ponderados es necesario establecer pesos de ponderación w_i. El software **R** utiliza la función tricubo para establecer estos pesos. Sea x_0 el punto de interés y $\Delta(x_0)$ la distancia máxima entre x_0 y los puntos en su proximidad, entonces la función tricubo se define como

$$W\left[\frac{|x_0 - x_j}{\Delta(x_0)}\right]$$

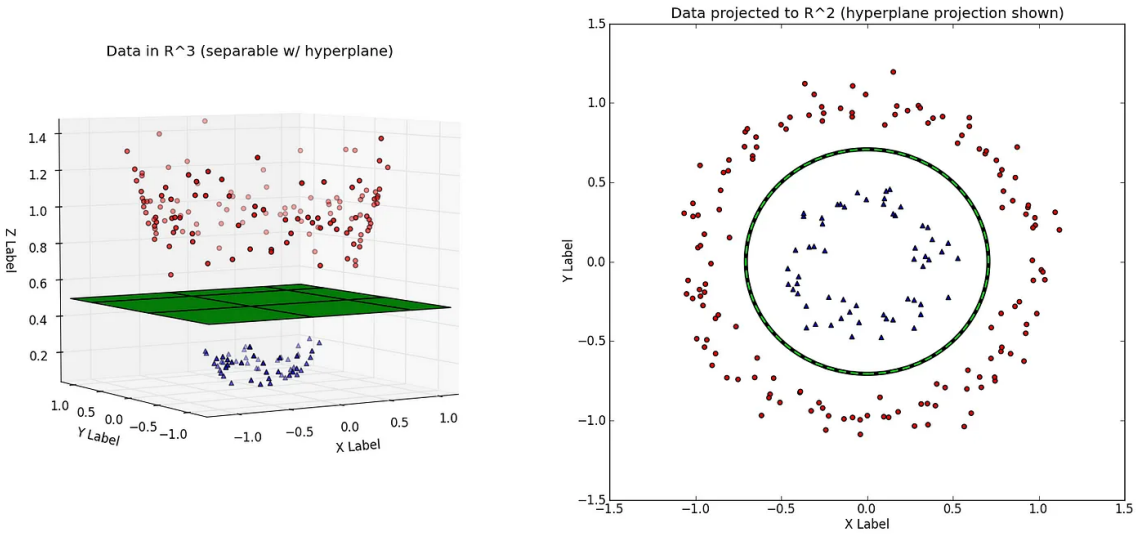


Figura 27: Aumento de dimensión

donde

$$W(u) = \begin{cases} (1-|u|^3)^3, & 0 \leq u < 1 \\ 0, & \text{en otro caso} \end{cases}$$

Donde u es la distancia normada entre x_0 y cualquier punto en su proximidad. Note que esta función asigna mayores pesos a puntos cercanos y penaliza la lejanía. Para cualquier punto fuera de la proximidad, su peso es nulo.

También podemos definir una estimación a σ^2 y un análogo a R^2. Antes, es necesario ver que se puede resumir la función loess como

$$\tilde{y} = Sy$$

donde S es la matriz de suavizado. La suma de cuadrados de los residuales sería, entonces

$$\begin{aligned} SS_{Res} &= \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \\ &= (y - Sy)'(y - Sy) \\ &= y'[I - S'][I - S]y \\ &= y'[I - S' - S + S'S]y \end{aligned}$$

Como este procedimiento es asintóticamente insesgado, se tiene que

$$E(SS_{Res}) = \sigma^2[n - 2tr(S) + tr(S'S)]$$

pues

$$\begin{aligned} tr[(I - S' - S + S'S)\sigma^2 I] &= \sigma^2 tr[I - S' - S + S'S] \\ &= \sigma^2 / [tr(I) - tr(S') - tr(S) + tr(S'S)] \end{aligned}$$

y, como S es cuadrada

$$tr(S') = tr(S)$$

En cierto sentido, $[n - 2tr(S) + tr(S'S)]$ son los grados de libertad asociados a todo el modelo. Es natural pensar en un estimador para σ^2 como

$$\hat{\sigma}^2 = \frac{SS_{Res}}{df} = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{[n - 2tr(S) + tr(S'S)]}$$

y a R^2 como

$$R^2 = \frac{SS_T - SS_{Res}}{SS_T}$$

Para el ejemplo de la caída de voltaje mostrado en la sección de ajuste polinomial por segmentos se le aplica ahora el método Loess para ejemplificar su desempeño. Además, se adjuntan las lineas de código en lenguaje **R** que son muy simples. Note que se establece un span (tramo) de 0.75, que indica que la el 75 % más próximo del total de puntos de datos se usa como proximidad. La Fig. (28) muestra la misma nube de puntos observados además de la curva azul ajustada por el método Loess.

```
modelo_loess<-loess(Voltaje~Segundos,data = volt)
lines(volt$Segundos, predict(modelo_loess), col = "blue", lwd = 2)
summary(modelo_loess)

#Consola:

Number of Observations: 41
Equivalent Number of Parameters: 4.43
Residual Standard Error: 0.542
Trace of smoother matrix: 4.84  (exact)

Control settings:
  span     :  0.75
  degree   :  2
  family   :  gaussian
  surface  :  interpolate   cell = 0.2
  normalize:  TRUE
 parametric:  FALSE
drop.square:  FALSE
```

6.3. Polinomios ortogonales

dos vectores de tamaño $n \times 1$ se dicen ortogonales si su producto resulta en el escalar nulo.

$$a'b = a_1b_1 + a_2b_2 + \cdots + a_nb_n = 0$$

Cuando dos vectores son ortogonales, $x_1'x_2 = 0$ sus parámetros estimados β_1 y β_2 no cambian si se calculan con un modelo completo $y = \beta_1x_1 + \beta_2x_2 + \varepsilon$ o con un modelo reducido $y = \beta_1x_1 + \varepsilon$ para β_1 o $y = \beta_2x_2 + \varepsilon$ para β_2. Para un modelo polinomial

$$y_i = \beta_0 + \beta_1x_i + \beta_2x_i^2 + \cdots + \beta_kx_i^k + \varepsilon_i$$

los vectores columna de X no son ortogonales. Para crear un modelo ortogonal con polinomios resulta útil primero definir polinomios que cumplan con

$$\sum_{i=1}^{n} P_r(x_i)P_s(x_i) = 0, \quad r \neq s, r, s = \overline{1,k}$$

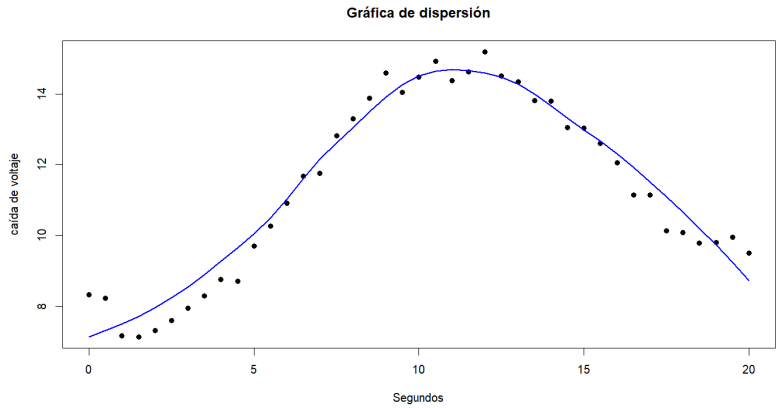


Figura 28: Ajuste Loess para la caída de voltaje.

y, por convención

$$P_0(x_i) = 1, \quad \forall i = \overline{1,n}$$

de tal manera que el modelo será

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \cdots + \alpha_k P_k(x_i) + \varepsilon_i \tag{113}$$

Con esta nueva estructura podemos expresar la matriz X que antes era de valores x_i como una matriz de polinomios de u-ésimo orden para cada x_i observación en la Ec. (113)

$$X = \begin{bmatrix} P_0(x_1) & P_1(x_1) & \cdots & P_k(x_1) \\ P_0(x_2) & P_1(x_2) & \cdots & P_k(x_2) \\ & & \ddots & \\ P_0(x_n) & P_1(x_n) & \cdots & P_k(x_n) \end{bmatrix}$$

y el modelo se puede escribir como

$$Y = X\alpha + \varepsilon$$

Como la matriz X tiene columnas ortogonales, $X'X$ es una matriz diagonal con elementos

$$\sum_{i=1}^{n} P_j^2(x_i)$$

para el elemento $X'X_{jj}$ con $j = \overline{0,k}$.

Por su parte, los parámetros $\hat{\alpha}$ se obtienen de la misma manera que en mínimos cuadrados con la matriz $(X'X)^{-1}X'Y$. Particularmente

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{n} P_j(x_i) y_i}{\sum_{i=1}^{n} P_j^2(x_i)}, \quad j = \overline{0,k}$$

La suma de cuadrados de los residuales y de la regresión se pueden calcular como

$$SS_{Res}(k) = SS_T - \sum_{i=1}^{k} \hat{\alpha}_j \left[\sum_{i=1}^{n} P_j(x_i) y_i \right]$$

y

$$SS_R(\alpha_j) = \hat{\alpha}_j \sum_{i=1}^{n} P_j(x_i) y_i$$

Con lo anterior es posible calcular un estadístico F útil para la prueba de significancia particular $H_0 : \beta_k = 0$, donde

$$F_0 = \frac{SS_R(\alpha_k)}{SS_{Res}(k)/(n-k-1)} = \frac{\hat{\alpha}_k \sum_{i=1}^{n} P_k(x_i) y_i}{SS_{Res}(k)/(n-k-1)}$$

Para las demostraciones revisar en [12].

7. Variables indicadoras

Las variables cuantitativas tienen una escala de medición definida, mientras que una variable cualitativa es un atributo no numeral. Por ejemplo, la distancia medida en centímetros, el tiempo medido en minutos, la riqueza medida en pesos, etc, son variables cuantitativas, mientras que el sexo, el color, marca, religión, forma, etc. son cualidades, es decir, variables cuantitativas. Un modelo de regresión lineal simple se calcula sobre una variable independiente cuantificable y su variable de respuesta igualmente cuantitativa. Sin embargo, es posible determinar un modelo de regresión introduciendo variables cualitativas dándole una categoría. A estas variables se les suele llamar indicadoras, categóricas o dummy (ficticias).

Imagine que un economista quiere conocer el ingreso de las personas dependiendo de las horas que han trabajado. Para ello, nota que las observaciones se agrupan en dos líneas aproximadas paralelas pero con distinta ordenada al origen Fig. (29). Naturalmente se da cuenta que esta diferencia se debe a la región de residencia. Decide hacer un modelo de regresión para cada región obteniendo

$$y^* = \beta_0^* + \beta_1^* x^* + \varepsilon^*$$

y

$$y^{**} = \beta_0^{**} + \beta_1^{**} x^{**} + \varepsilon^{**}$$

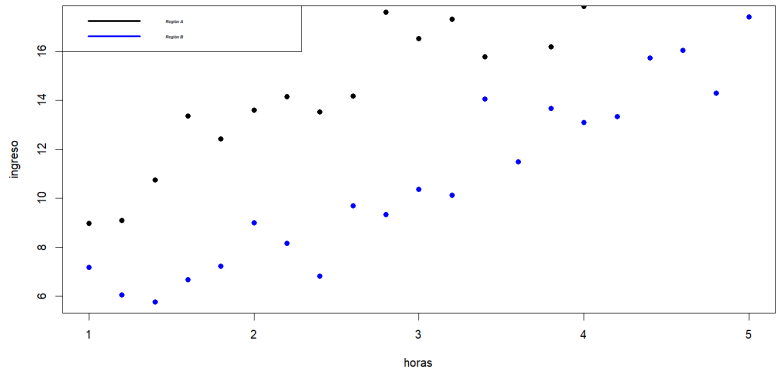


Figura 29: Ingreso contra horas trabajadas (Ilustrativo, datos no reales).

Si la varianza en la misma en ambos modelos, se pueden representar con un modelo más general introduciendo una variable indicadora x_2 de la forma

$$x_2 = \begin{cases} 0, & \text{si la obs. pertenece a la región A} \\ 1, & \text{si la obs. pertenece a la región B} \end{cases}$$

Entonces, el modelo conjunto es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

note que

$$E(y|x_2 = 0) = \beta_0 + \beta_1 x_1$$

representa el modelo para la región A, con ordenada al origen en β_0. Análogo

$$\begin{aligned} E(y|x_1 = 1) &= \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned}$$

representa el modelo para la región B, con ordenada al origen en $(\beta_0 + \beta_2)$.
Es posible trabajar con una variable indicadora con más de dos categorías o con múltiples variables indicadoras, salvo algunas advertencias. Si para el ejemplo anterior se tuviera más de dos regiones, entonces la variable indicadora podría ser de la forma

$$x_2 = \begin{cases} 0, & \text{si la obs. pertenece a la región A} \\ 1, & \text{si la obs. pertenece a la región B} \\ 2, & \text{si la obs. pertenece a la región C} \\ 3, & \text{si la obs. pertenece a la región D} \end{cases}$$

creando el mismo modelo general

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Esta variable sólo es posible si las categorías son mutuamente excluyentes, es decir que solo puedan tomar una categoría a la vez. Aunque sigue el principio de parsimonia, se pierde el sentido de ausencia o presencia de una categoría por la forma en la que se asignaron los códigos numerales a cada región llegado a conclusiones poco intuitivas como que

$$\begin{aligned} \beta_2 &= E(y|x_2 = 3) - E(y|x_2 = 2) \\ &= E(y|x_2 = 2) - E(y|x_2 = 1) \\ &= E(y|x_2 = 1) - E(y|x_2 = 0) \end{aligned}$$

Alternativo a esto, se crea una variable artificial para cada región siendo uno si pertenece a la región o cero si no. Con esta definición se necesitan 4 variables indicadoras para el mismo modelo

$$x_2 = \begin{cases} 1, & \text{si la obs. pertenece a la región A} \\ 0, & \text{otro caso} \end{cases}$$

$$x_3 = \begin{cases} 1, & \text{si la obs. pertenece a la región B} \\ 0, & \text{otro caso} \end{cases}$$

$$x_4 = \begin{cases} 1, & \text{si la obs. pertenece a la región C} \\ 0, & \text{otro caso} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{si la obs. pertenece a la región D} \\ 0, & \text{otro caso} \end{cases}$$

Resultando en el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \tag{114}$$

Con este modelo extendido no es necesario que las categorías sean excluyentes. Por ejemplo, si además de la región se añade el estado civil, ambas variables indicadoras pueden tomar el valor 1. La advertencia es

Observación	Horas y_i	rpm x_i	tipo
1	18.73	610	A
2	14.52	950	A
3	17.43	720	A
4	14.54	840	A
5	13.44	980	A
6	24.39	530	A
7	13.34	680	A
8	22.71	540	A
9	12.68	890	A
10	19.32	730	A
11	30.16	670	B
12	27.09	770	B
13	25.4	880	B
14	26.05	1000	B
15	33.49	760	B
16	35.62	590	B
17	26.07	910	B
18	36.78	650	B
19	34.95	810	B
20	43.67	500	B

que se pierde el sentido cuando dos variables excluyentes toman el valor unitario, por lo que el investigador debe poner especial atención en cuales variables indicadoras son excluyentes y cuales no.
Hasta ahora se trabajó con el supuesto de que cada categoría conservaba la misma pendiente β_1, pero esto tampoco es necesariamente cierto. Para el caso donde no se cumple, se añade una variable de interacción, por ejemplo $(\beta_{12}x_1x_2)$ en la Ec. (114), resultando en el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \varepsilon$$

Donde

$$\begin{aligned} E(y|x_2 = 1) &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_{12} x_1 \\ &= (\beta_0 + \beta_1) + (\beta_1 + \beta_{12}) x_1 \end{aligned}$$

modificando la ordenada al origen y la pendiente de x_1. Esta idea se puede extender para cualesquiera variables indicadoras.

Ejemplo en [11]

La siguiente tabla tiene el registro de las observaciones del tiempo de vida de una herramienta dependiendo de la velocidad del torno y del tipo de herramienta (A o B). Cómo se explicó anteriormente, la gráfica de dispersión Fig. (30) muestra dos tendencias distintas para las herramientas de tipo A y tipo B. Gracias a las variables dummy es posible simplificar estas tendencias en un único modelo de regresión lineal. Aunque uno podría advertir que las pendientes son similares, es más preciso hacer un modelo suponiendo que no lo son y luego corroborar con una prueba de hipótesis.

Dicho modelo es de la forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

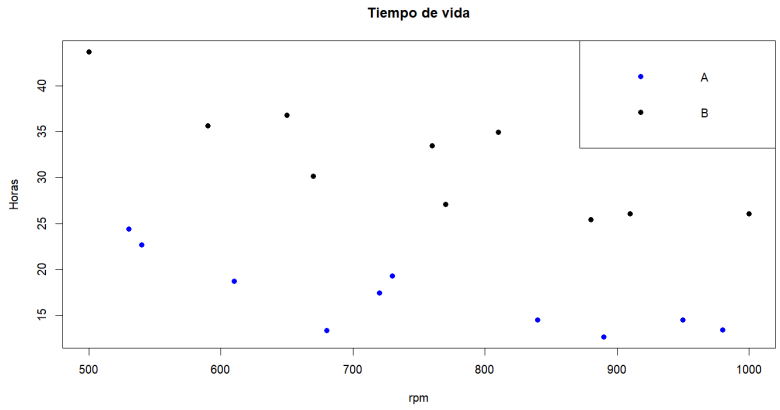


Figura 30: Diagrama de dispersión para el tiempo de vida.

Donde x_2 es una variable indicadora que toma el valor cero si la herramienta es de tipo A y uno si es tipo B. A continuación se adjunta el código en lenguaje **R**. Sólo con ver los cuantiles se puede intuir que no hay un problema de sesgo en los residuales, al igual que en el gráfico de dispersión Fig. (31) para el modelo simplificado.. para la prueba t de significancia se nota que la variable $(rpm * dummy)$ correspondiente al parámetro β_{12} tiene un valor p suficiente alto para no rechazar la hipótesis nula $H_0: \beta_{12} = 0$, lo que se interpreta como que no existe el cambio en la pendiente para los dos tipos de herramienta que ya se había intuido desde la gráfica. los valores β_1 y β_2 si son significativos, así que se reestructura el modelo por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Al disminuir el número de variables se espera una disminución en el nivel de ajuste, pero no es exorbitante, pasando de un R^2 ajustada de 0.8937 a 0.8886. Ambas pruebas F de significancia del modelo se concluyen en rechazar H_0 casi para cualquier valor de α. El valor de $\beta_1 = -0.0266$ nos dice que existe una relación negativa entre el tiempo de vida medida en horas y las revoluciones por minuto del torno, mientras que la variable artificial (o dummy) es el cambio en la duración promedio de la herramienta debido a un cambio del tipo A al tipo B, es decir de β_0 a $\beta_0 + \beta_2$. Para este ejemplo, cambiar al tipo B aumenta la vida media en 15.004 horas. Es preferible trabajar con el modelo simple por el principio de parsimonia, además de que presenta una significancia F mayor y la interpretación hace más sentido en el contexto del problema.

```
#Ejemplo 8.1 Montgomery
#datos
datos<-read_xlsx("Ejemplo 8.1 Montgomery.xlsx")
View(datos)
attach(datos)
#categorizacion
indicadora<-c('A'=0,'B'=1)
colores<-c('A'='blue','B'='black')
datos<-datos %>% mutate(dummy=indicadora[tipo])
#grafica
plot(y=Horas,x=rpm, pch=16, col=colores[tipo],main='Tiempo de vida')
legend("topright", legend = unique(tipo), col = unique(colores), pch = 16)
#modelo
modelo<-lm(Horas~rpm+dummy+I(rpm*dummy),data=datos)
```

```
summary(modelo)
modelo_simple<-lm(Horas~rpm+dummy,data=datos)
summary(modelo_simple)
#residuales
plot(fitted.values(modelo_simple),residuals(modelo_simple),ylab='residuales',xlab='y
     gorro',pch=16)


#Consola:
#modelo
Residuals:
    Min      1Q  Median      3Q     Max
-5.1750 -1.4999  0.4849  1.7830  4.8652

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    32.774760   4.633472   7.073 2.63e-06 ***
rpm            -0.020970   0.006074  -3.452  0.00328 **
dummy          23.970593   6.768973   3.541  0.00272 **
I(rpm * dummy) -0.011944   0.008842  -1.351  0.19553
---

Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared:  0.9105,  Adjusted R-squared:  0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08

#modelo simple
Residuals:
    Min      1Q  Median      3Q     Max
-5.5527 -1.7868 -0.0016  1.8395  4.9838

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.98560    3.51038  10.536 7.16e-09 ***
rpm         -0.02661    0.00452  -5.887 1.79e-05 ***
dummy       15.00425    1.35967  11.035 3.59e-09 ***
---

Residual standard error: 3.039 on 17 degrees of freedom
Multiple R-squared:  0.9003,  Adjusted R-squared:  0.8886
F-statistic: 76.75 on 2 and 17 DF,  p-value: 3.086e-09
```

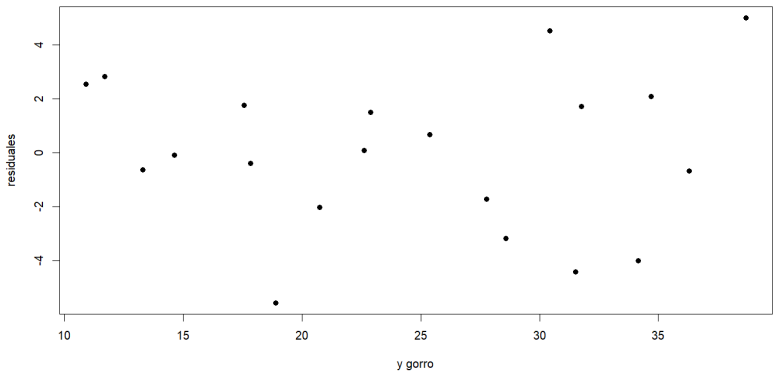


Figura 31: Gráfico de dispersión residuales contra $\hat{Y}$

Apéndices

8. Matrices

Para los cálculos de la regresión lineal multiparamétrica es importante recordar algunas cuestiones importantes de las matrices.

- Sea $A = (a_{ij})$ una matriz de $n \times m$ (n renglones y m columnas), con $i \leq n$ y $j \leq m$, la matriz transpuesta se define como $A' = (a_{ji})$. Es decir, intercambiamos rilas (renglones) por columnas. Note que

$$(A')' = (a_{ji})' = (a_{ij})$$

$$\begin{pmatrix} 6 & -2 \\ 4 & 7 \\ 1 & 3 \end{pmatrix}' = \begin{pmatrix} 6 & 4 & 1 \\ -2 & 7 & 3 \end{pmatrix}$$

- Si $A = A'$, entonces la matriz A se dice simétrica.
- Si $n = m$ (matriz cuadrada), la diagonal de la matriz son los elementos $a_{11}, a_{22}, \ldots, a_{nn}$.
- si la matriz contiene ceros en todos sus elementos excepto en la diagonal, es llamada matriz diagonal.
- Sea A una matriz cuadrada, $B = diag(A)$ es la matriz diagonal tal que $a_{ii} = b_{ii} \; \forall \; i \leq n$.
- Si $(a_{ii} = 1)$ en una matriz diagonal, entonces es llamada matriz identidad.

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

- Una matriz triangular superior tiene ceros en las entradas por debajo de la diagonal.

$$T = \begin{pmatrix} 7 & 2 & 4 & 5 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

- Una matriz (o vector) unitario es tal que $(a_{ij} = 1) \; \forall \; i \leq n, j \leq m$. Análogo para una matriz nula.

$$J = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad O = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

- Sean $A_{n \times p}$ y $B_{n \times p}$ dos matrices, entonces la suma se define como $C_{n \times p} = A + B = (c_{ij}) = (a_{ij} + b_{ij})$. Análogo para la resta.
- Note que $A + B = B + A$ y $(A + B)' = A' + B'$.
- Sean $A_{n \times p}$ y $B_{p \times m}$ dos matrices, el producto $C = AB$ se define como $c_{ij} = \sum_k a_{ik} b_{kj}$, donde C es de tamaño $n \times m$. Note que no es conmutativa.
- sea c un escalar y A una matriz, $cA = (ca_{ij}) = Ac$.

- Sean $A_{n\times p}$ y $B_{p\times m}$ dos matrices,
$$(AB)' = B'A'.$$

- Sea $A_{n\times p}$ una matriz, entonces $A'A$ y AA' tienen las siguientes propiedades

  i) $A'A$ es $p \times p$ y se obtiene como producto de columnas de A.

  ii) AA' es $n \times n$ y se obtiene como producto de filas de A.

  iii) AA' y $A'A$ son simétricas.

  iv) Si $A'A = O$, entonces $A = O$, donde O es una matriz nula.

- Sea I la matriz identidad, entonces
$$IA = AI = A.$$

- El rango de una matriz, llamado $ran(A)$, es el número de columnas (o filas) linealmente independientes.

- Sea $A_{n\times n}$ una matriz no singular y $ran(A) = n$. Entonces A tiene una única matriz inversa, denotada A^{-1} tal que
$$AA^{-1} = A^{-1}A = I$$
Note que $(A^{-1})^{-1} = A$. Además, se cumple que

  i) $(A')^{-1} = (A^{-1})'$

  ii) $(AB)^{-1} = B^{-1}A{-}1$

- Una inversa generalizada de la matriz $A_{n\times p}$ es cualquier matriz A^{-} que satisfaga
$$AA^{-}A = A$$

- El determinante de una matriz $A_{n\times n}$, denotado por $|A|$ o $det(A)$, es una función escalar de A definida como la suma de todos los $n!$ posibles productos de n elementos tal que

  - Cada producto contenga un elemento de todas las columnas y filas de A.
  - Los factores en cada producto son escritos tal que la columna subscrita aparezca en orden de magnitud y cada producto es precedido, entonces, por un signo positivo o negativo de acuerdo al número de inversiones en la columna es impar o par.

- La traza de una matriz es la suma de todos los elementos de la diagonal. $tr(A) = \sum_{i=1}^{n} a_{ii}$.

9. Propiedades

9.1. Consistencia

La consistencia es una propiedad de ciertos estimadores. Se dice que un estimador es consistente cuando éste converge a su valor verdadero cuando el número de datos de la muestra tiende a infinito.

9.2. Propiedades de Varianza

Para el cálculo de la varianza de y, se usa con frecuencia la ecuación de la varianza y el valor esperado condicional, se usa la siguiente propiedad para particionar la variabilidad:

$$var(y) = var(E[y|u]) + E[var(y|u)] \tag{115}$$

9.3. Función generadora de momentos

Sea X una variable aleatoria. El valor esperado

$$m_X(t) = E[\exp(tX)] \quad -c \leq t \leq c \tag{116}$$

recibe el nombre de función generadora de momentos.
Si X es una v.a. discreta

$$m_X(t) = E[\exp(tX)] = \Sigma_x exp(tx) * p(x) \tag{117}$$

Si X es una v.a. continua

$$m_X(t) = E[\exp(tX)] = \int_{-\infty}^{\infty} exp(tx) * p(x) \tag{118}$$

10. Distribuciones relacionadas con la distribución normal

10.1. Distribución normal

1. Si la variable aleatoria Y tiene la distribución normal con media μ y varianza σ^2, su función de densidad de probabilidad es que se denota por $Y \sim N(\mu, \sigma^2)$

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-[y-\mu]^2/2\sigma^2\}$$

2. La distribución normal con $\mu = 0$ y $\sigma^2 = 1, Y \sim N(0,1)$, es llamada la **distribución normal estándar**.

3. Sea $Y_1, Y_2, \ldots, Y_n$ denota las variables aleatorias distribuidas normalmente con $Y \sim N(\mu_i, \sigma 2)$ para $i = 1, \ldots, n$ y sea las covariables de Y_i y Y_j que se denota por

$$Cov(Y_i, Y_j) = \rho_{ij}\sigma_i\sigma_j,$$

donde ρ_{ij} es el coeficiente de correlación para Y_i y Y_j

4. Suponga que las variables aleatorias $Y_1, Y_2, \ldots, Y_n$ son independientes y distribuidos normalmente con las distribuciones $Y \sim N(\mu_i, \sigma_i^2)$ para $i = 1, \ldots, n$. Si

$$W = a_1Y_1 + a_2Y_2 + \ldots + a_nY_n$$

donde los $a_i's$ son constantes. Entonces W es también distribuida normalmente, así que

$$W = \Sigma_{i=1}^n a_iY_i \sim N(\Sigma_{i=1}^n a_i\mu_i, \Sigma_{i=1}^n a_i^2\sigma_i^2)$$

10.2. Distribución Chi cuadrada

1. La distribución chi cuadrada centra con n grados de libertad está definida como la suma de cuadrados de n variables aleatorias independientes $Z_1, \ldots, Z_n$ cada una con distribución normal estándar. Se denota por
$$X^2 = \Sigma_{i=1}^{n} Z_i^2 \sim \chi^2(n).$$

2. Si X^2 tiene la distribución $\chi^2(n)$, entonces el valor esperado es $E(X^2) = n$ y su varianza es $var(X^2) = 2n$.

3. Si $Y_1, Y_2, \ldots, Y_n$ son variables aleatorias independientes distribuidas normalmente cada una con la distribución $Y_i \sim N(\mu_i, \sigma_i^2)$ entonces
$$\Sigma_{i=1}^{n} X^2 = \Sigma_{i=1}^{n} (\frac{Y_i - \mu_i}{\sigma_i})^2 \sim \chi^2(n)$$
porque cada variable $Z_i = \frac{Y_i - \mu_i}{\sigma_i}$ tiene una distribución normal $N(0,1)$.

10.3. Distribución t

La distribución t con n grados de libertad está definida como el radio de dos variables aleatorias independientes. El numerador tiene distribución estándar y el denominador es la raíz de una v. a. chi- cuadrada central dividida por sus grados de libertad: esto es,
$$T = \frac{Z}{(X^2/n)^{1/2}} \quad denotado\ por\ T \sim t(n).$$
donde $Z \sim N(0,1)$, $X^2 \sim \chi^2(n)$, tanto Z y X^2 son independientes.

10.4. Distribución F

La distribución central con n y m grados de libertad está definida como el radio de dos variables aleatorias independientes chi cuadradas centrales cada una dividida por sus grados de libertad
$$F = \frac{X_1^2/n}{X_2^2/m} \quad denotado\ por\ F \sim F(n,m) \tag{119}$$
donde $X_1^2 \sim \chi^2(n)$, $X_2^2 \sim \chi^2(m)$ y X_1^2 y X_2^2 son independientes.

Referencias

[1] Calderon, JP. de los Godos, LA., *Regresión Logística Aplicada a la Epidemiología*, Rev Salud, Sexualidad y Sociedad, 1(4):78-84, 2009.

[2] Casella, G. & Berger, R., *Statistical Inference*, Duxbury Advanced Series, 2002.

[3] Crawley, M.J., *The R book*, Jhon Wiley & Sons, 2012.

[4] Lindsey, J., *Applying Generalized Linear Models*, Springer-Verlag, 1997.

[5] Gałecki A. and Burzykowski T., *Linear mixed-effects models using R: A step-by-step approach*, Springer Science & Business Media, 2013.

[6] Heumann, C., Rao, R.C., and Shalab, Toutenburg, H., *Linear Models and Generalizations: least Squares and Alternatives*, Springer, 2008.

[7] Hosmer, D. & Lemeshow, S., *Applied Logistic Regression*, Jhon Wiley & Sons, 2000.

[8] Infante, G.S. y P-Zarate de Lara, G., *Métodos Estadísticos*, Trillas, 2005.

[9] McCulloch E. C., Searle R. S. and Neuhaus M. J., *Generalized, Linear and Mixed Models*, Jhon Wiley & Sons, 2011.

[10] Mehtätalo L., *Linear mixed-effects models with examples in R*, University of Eastern Finland, Addison-Wesley, 2013.

[11] Montgomery, D. C., Peck E.A. and Vining G.G., *Introduction to linear regression Analysis*, Jhon Wiley & Sons, 2011.

[12] Alvin C. Rencher and G. Bruce Schaalje., *Linear Models In Statistics*, Jhon Wiley & Sons, 2008.

[13] Norman R. Draper, Harry Smith, *Applied Regression Analysis*, Jhon Wiley & Sons, 1998.

[14] Dennis D. Wacherly, William Mendenhall III, Richard L. Scheaffer, *Estadística matemática con aplicaciones*, International Thomson Editores, 2002.

[15] Navarro, E., Verbel A., Robles, D., Hurtado, K. R., *Regresión Logística Ordinal Aplicada a la identificación de factores de riesgo para cáncer de cuello uterino*, Ingeniare, 9(17):87-105, 2014.

[16] Oroza, H. A., *Modelos mixtos en la determinación del carbono orgánico en la hojarasca en una zona de Teziutlán*, Puebla, tesis de Doctorado-BUAP, 2015.

[17] Searle S. R. and Khuri A. I., *Matrix algebra useful for statistics*, John Wiley & Sons, 2017.

[18] Cárdenas Julián, Investigación cuantitativa. Berlin: TrAndes,Programa de Posgrado en Desarrollo Sostenible y Desigualdades Sociales en la Región Andina, 2018. https://networkianos.com/regresion-logistica-binaria/regresion-logistica-4/ (9 de agosto 2023)

[19] Team, R. C., *R: A language and environment for statistical computing*, 2013. Cárdenas, Julián (2018) Investigación cuantitativa. Berlin: TrAndes,Programa de Posgrado en Desarrollo Sostenible y Desigualdades Sociales en la Región Andina

[20] wad, M., Khanna, R. (2015). Support Vector Regression. In: Efficient Learning Machines. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_4

[21] mola, A.J., Schölkopf, B. A tutorial on support vector regression. Statistics and Computing 14, 199–222 (2004). https://doi.org/10.1023/B:STCO.0000035301.49549.88