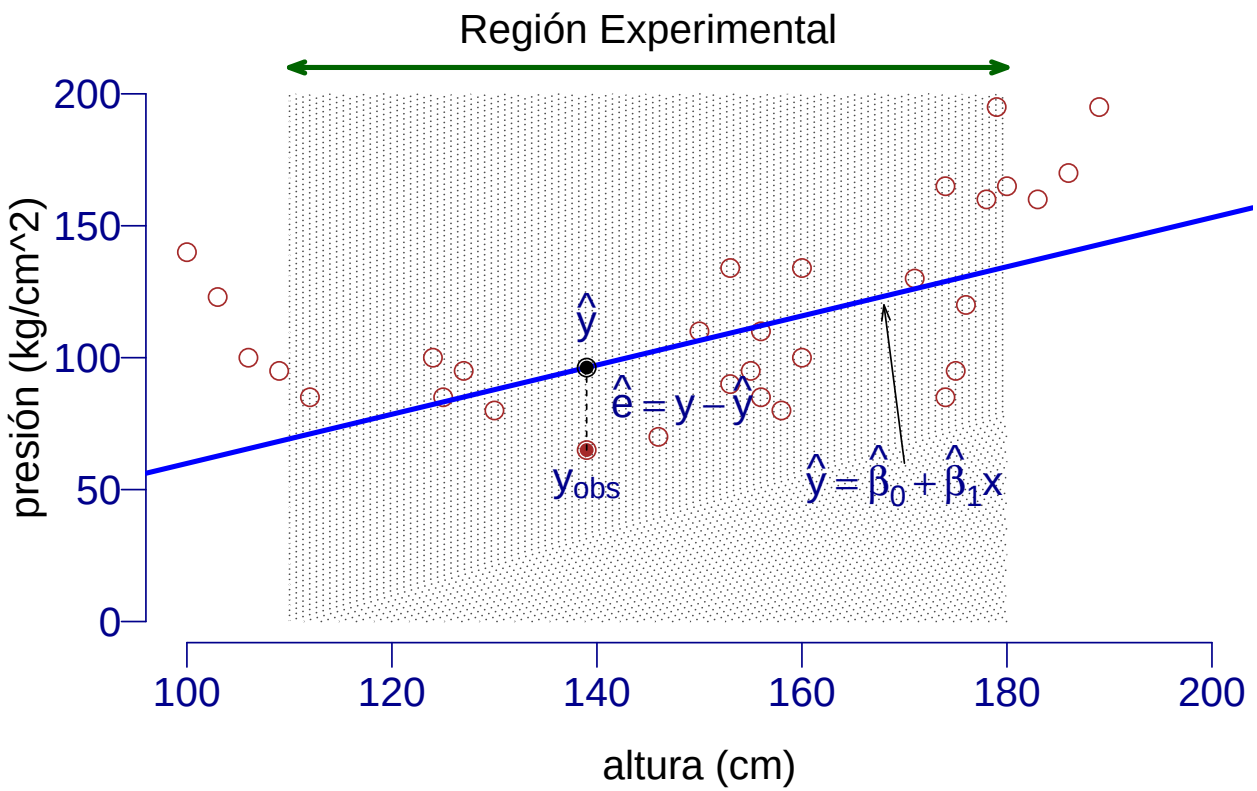
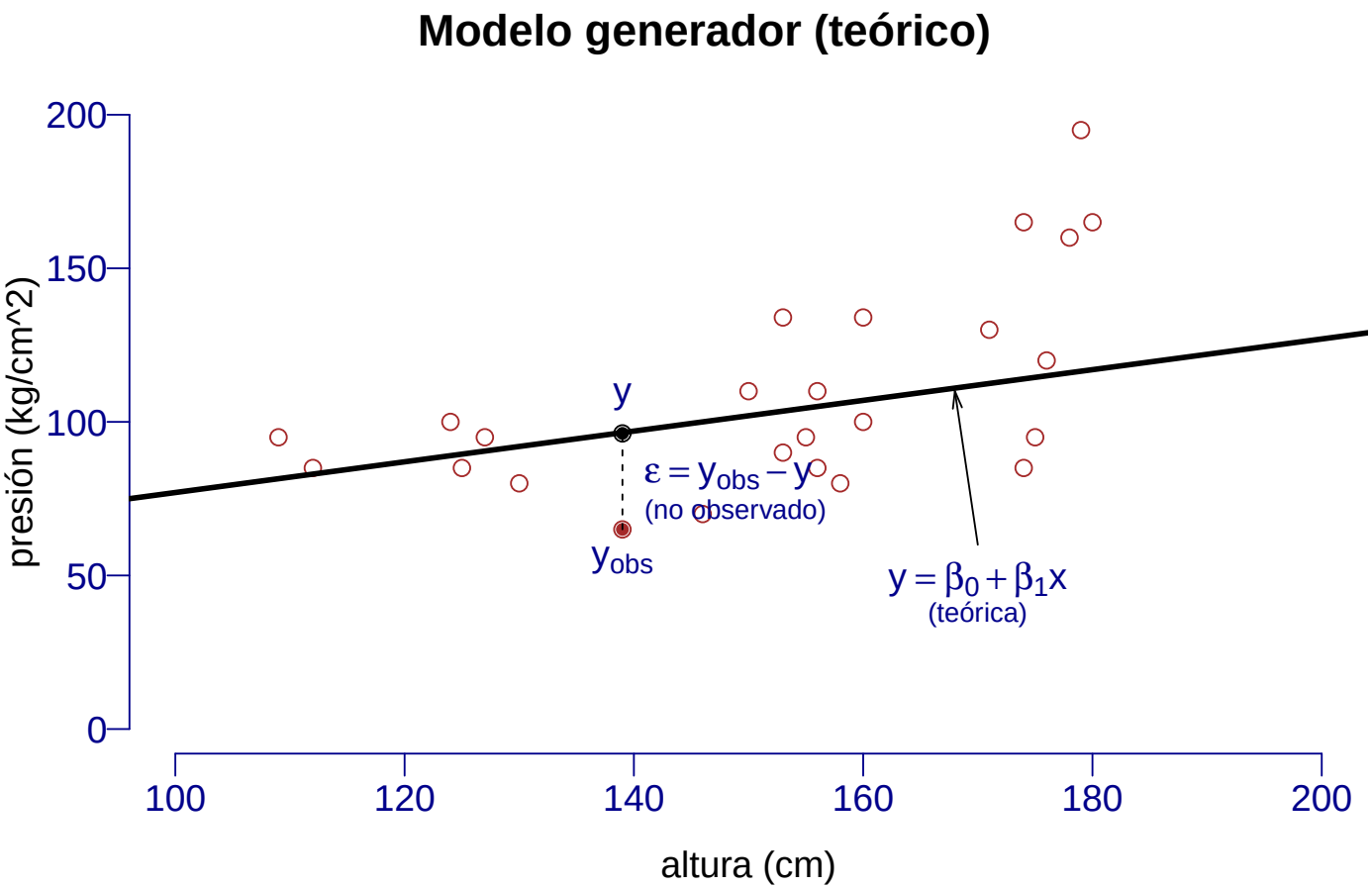
2 Regresión Lineal Simple



Contenido

Modelo generador teórico



Modelo de regresión lineal simple

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

donde,

$i = 1, \ldots, n$, observaciones.

y_i: variable de respuesta, variable dependiente.

x_i: variable de control, regresor, variable independiente.

β_0, β_1: coeficientes del modelo.

β_0: es el nivel de la respuesta cuando $x = 0$.

β_1: es el incremento en la respuesta cuando aumento al regresor x en una unidad (tasa de cambio).

ϵ_i: error aleatorio (no observado), diferencia entre el modelo generador y el valor observado.

Supuestos

Si se supone que el error ϵ es aleatorio y tal que

$$\mathbb{E}[\epsilon] = 0, \quad \text{var}(\epsilon) = \sigma^2, \quad \text{cov}(\epsilon, \epsilon') = 0$$

la respuesta y es aleatoria, pero para el nivel $x = x_0$,

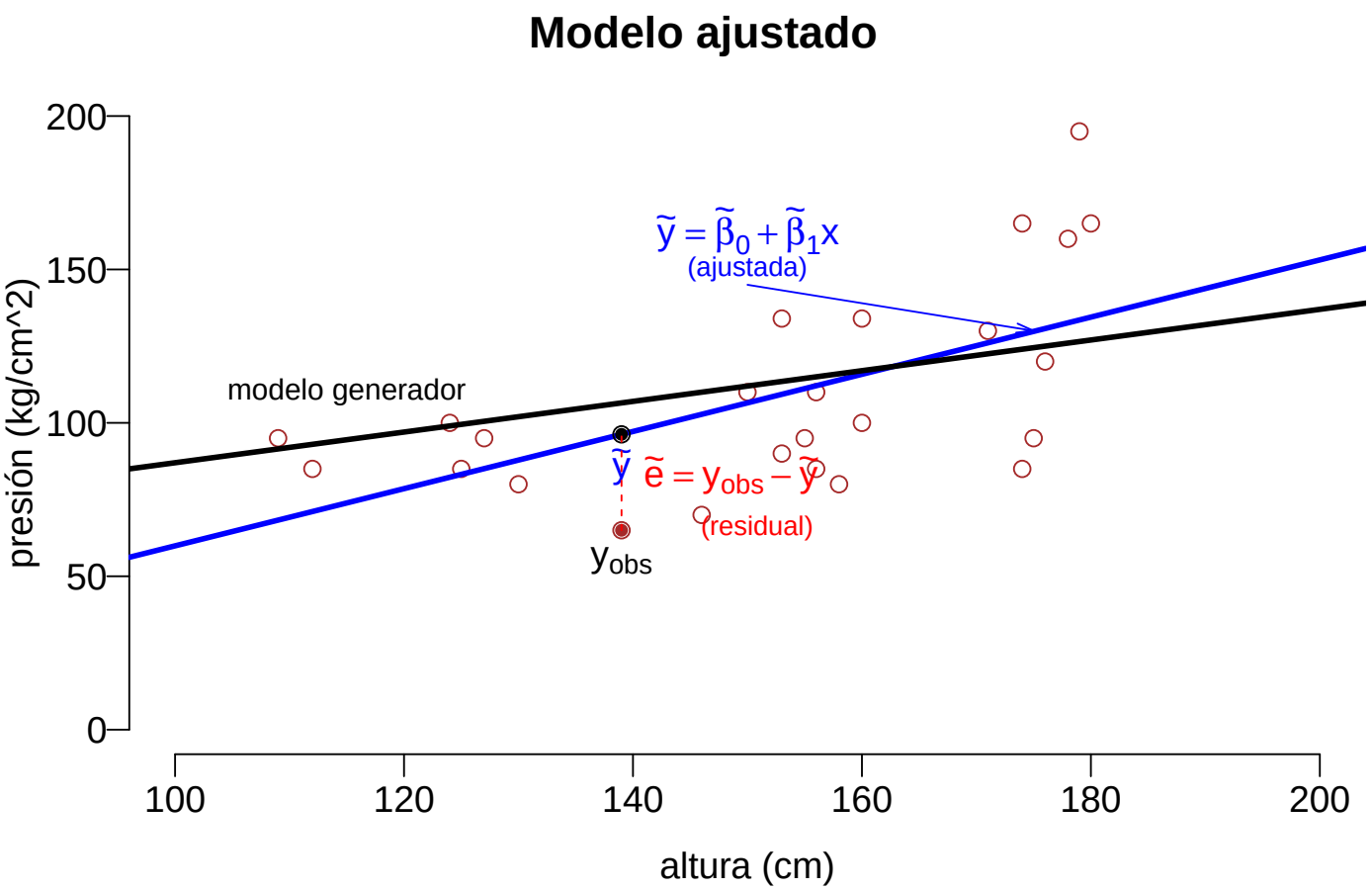
$$\begin{aligned} \mathbb{E}[y|x = x_0] &= \beta_0 + \beta_1 x_0 \\ \text{var}(y|x = x_0) &= \text{var}(\epsilon) = \sigma^2 \\ \text{cov}(y, y') &= 0 \end{aligned}$$

Esto es,

- El nivel medio de y depende del nivel de x.
- La variabilidad de la respuesta y *no* depende de x.
- $\text{cov}(y, y') = \text{cov}(\epsilon, \epsilon') = 0$.
- β_0 es la respuesta media para $x = 0$.
- β_1 es el incremento de la respuesta debido a un cambio unitario del regresor x. Es decir, para $\Delta_x = 1$, se tiene que $\Delta_y = \beta_1$.

–ebz

Modelo ajustado



Criterios para determinar "la mejor" línea recta

Problema

Encontrar la "mejor" recta $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$

1. Criterio L_0: *Suma de errores cero*

$$\sum_{i=1}^{n} \tilde{\epsilon}_i = \sum_i (y_i - \tilde{y}_i) = \sum_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0$$

¡¡¡Impráctico!!!

2. Criterio L_1: *Mínima Desviación Absoluta*

$$\min_{\beta} \sum_i |y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i|$$

3. Criterio L_2: *Mínimos Cuadrados*

$$\min_{\beta} \sum_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

4. Criterio L_∞: *Mínima Desviación Máxima*

$$\min_{\beta} \left\{ \max_i \left| y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \right| \right\}$$

–ebz

El problema de mínimos cuadrados

Considere la *suma de cuadrados*: $S(\beta) \doteq \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$

Criterio

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

Estimadores Mínimos Cuadrados (EMC)

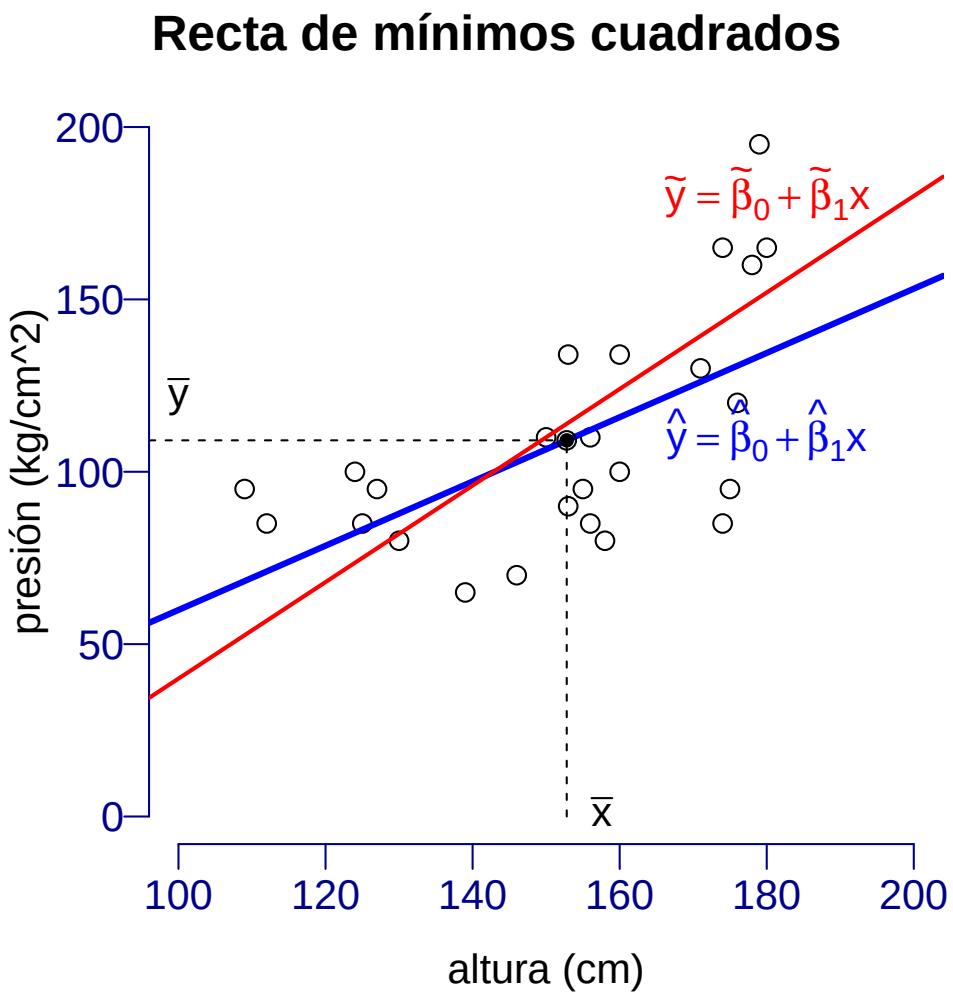
$$\frac{dS}{d\beta} \equiv 0 \Longrightarrow \begin{cases} \text{*Ecuaciones normales* (ortogonales) :} \\ n\beta_0 + \beta_1 \sum x_i = \sum y_i \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

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

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

–ebz

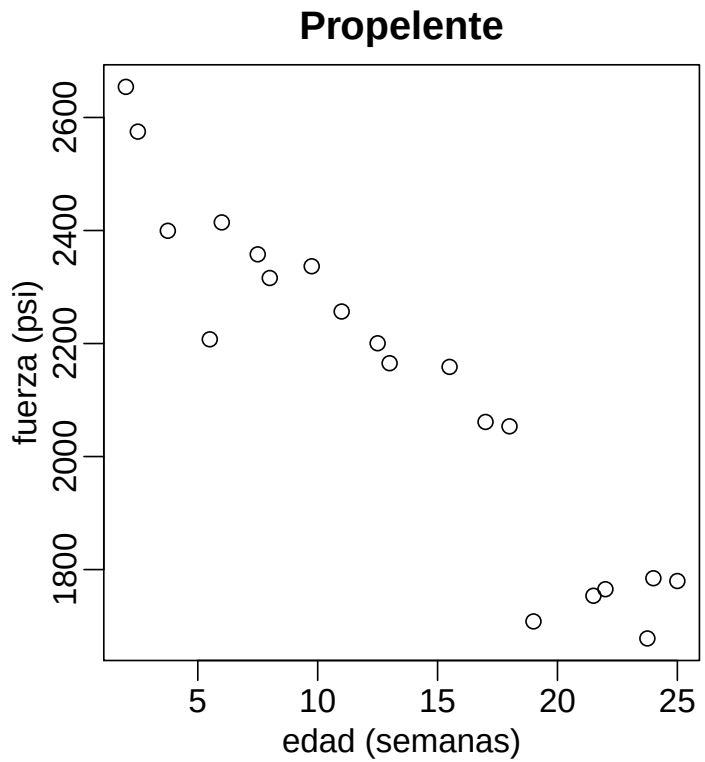
Recta ajustada por mínimos cuadrados



Ejemplo propelente[1]

Datos Se considera un motor de cohete estudiando el propelente de encendido dentro de un depósito de metal. La fuerza para separar la unión entre las componentes del combustible es la respuesta, que depende de la edad de propelente.

obs	fuerza	edad
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50



[1] Montgomery et al. (2001)

−ebz

Ejemplo propelente (cont.)

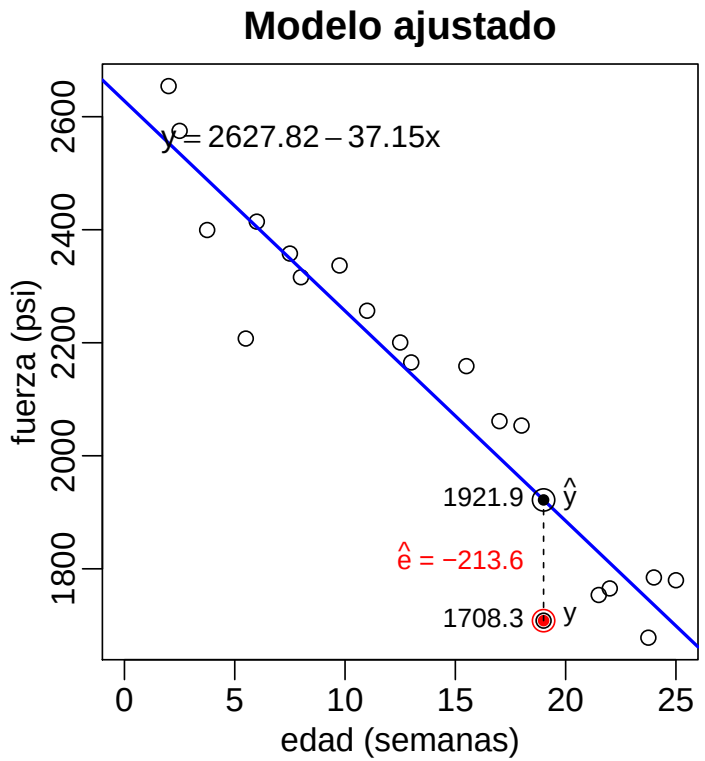
Ajuste

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16
edad         -37.154      2.889  -12.86 1.64e-10

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared: 0.9018,     Adjusted R-squared: 0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10

Analysis of Variance Table
Response: fuerza
          Df  Sum Sq Mean Sq F value    Pr(>F)
edad       1 1527483 1527483  165.38 1.643e-10
Residuals 18  166255    9236
```

Ejemplo propelente (cont.)

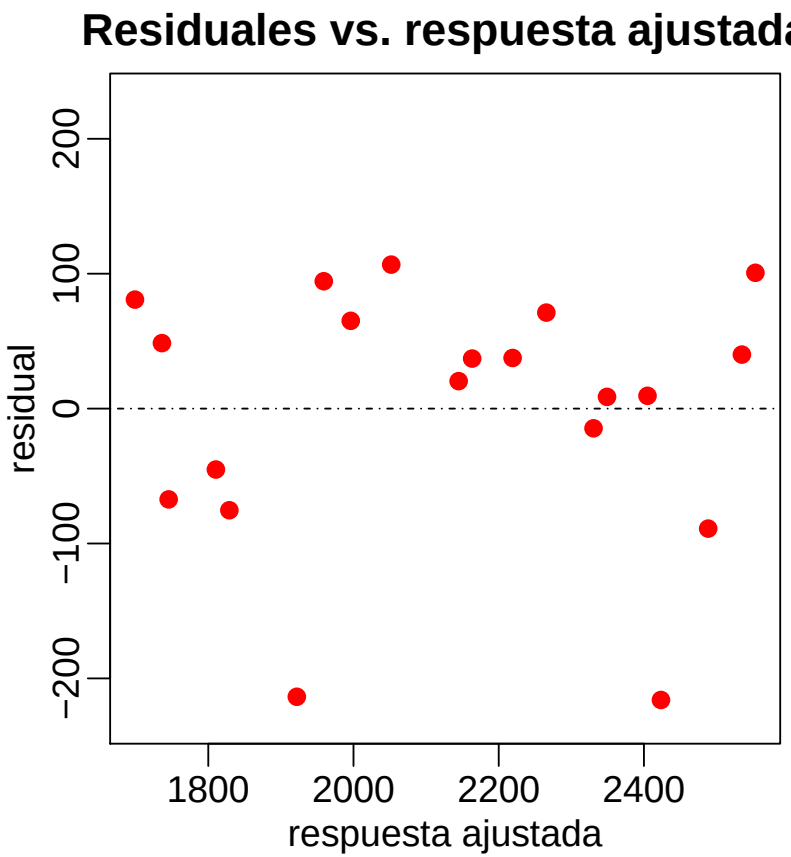
Datos observados, ajustados y residuales

obs	yobs	yhat	res
1	2158.70	2051.9	106.8
2	1678.15	1745.4	-67.3
3	2316.00	2330.6	-14.6
4	2061.30	1996.2	65.1
5	2207.50	2423.5	216.0
6	1708.30	1921.9	213.6
7	1784.70	1736.1	48.6
8	2575.00	2534.9	40.1
9	2357.90	2349.2	8.7
10	2256.70	2219.1	37.6
11	2165.20	2144.8	20.4
12	2399.55	2488.5	-88.9
13	1779.80	1699.0	80.8
14	2336.75	2265.6	71.2
15	1765.30	1810.4	-45.1
16	2053.50	1959.1	94.4
17	2414.40	2404.9	9.5
18	2200.50	2163.4	37.1
19	2654.20	2553.5	100.7
20	1753.70	1829.0	-75.3



Ejemplo propelente (cont.)

Gráfica de residuales



−ebz

Propiedades de los estimadores de mínimos cuadrados

$\sum(x_i - \bar{x}) = 0$ y por lo mismo $S_{xy} = \sum(x_i - \bar{x})y_i$. Luego $\hat{\beta}_1 = \sum c_i y_i$, donde $c_i = (x_i - \bar{x})/S_{xx}$.

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ \mathbb{E}[\hat{\beta}_1] &= \beta_1 \\ \text{var}(\hat{\beta}_1) &= \sigma^2 \frac{1}{S_{xx}} \end{aligned} \qquad \begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \mathbb{E}[\hat{\beta}_0] &= \beta_0 \\ \text{var}(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \end{aligned}$$

Teorema Gauss-Markov

Bajo los supuestos

$$\mathbb{E}[\epsilon_i] = 0, \quad \text{var}(\epsilon_i) = \sigma^2, \quad \text{cov}(\epsilon_i, \epsilon_j) = 0$$

los *estimadores de mínimos cuadrados* son los *mejores* estimadores lineales insesgados, en el sentido de que tienen varianza mínima.

–ebz

Propiedades de los estimadores de mínimos cuadrados

Se definen los *residuales de mínimos cuadrados* por

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores de mínimos cuadrados de los coeficientes del modelo. Se sigue entonces que

- $\sum \hat{e}_i = \sum (y_i - \hat{y}_i) = 0$
- $\sum \hat{y}_i = \sum y_i$
- $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
- $\sum x_i \hat{e}_i = 0$
- $\sum \hat{y}_i \hat{e}_i = 0$

Estimación de σ^2

Suma de Cuadrados: $\text{SC}_{\text{Res}} = S(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}$

Si se supone además que $\epsilon \sim \text{N}(0, \sigma^2)$, se puede mostrar que $\frac{\text{SC}_{\text{Res}}}{\sigma^2} \sim \chi^2_{n-2}$ independiente de $\hat{\beta}$. Luego,

$$\mathbb{E}\left[\frac{\text{SC}_{\text{Res}}}{\sigma^2}\right] = n - 2, \qquad \text{var}\left[\frac{\text{SC}_{\text{Res}}}{\sigma^2}\right] = 2(n-2)$$

Cuadrados Medios: $s^2 = \text{CM}_{\text{Res}} = \frac{1}{n-2}\text{SC}_{\text{Res}} = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$

Entonces,

$$\mathbb{E}\left[s^2\right] = \mathbb{E}\left[\frac{\text{SC}_{\text{Res}}}{n-2}\right] = \sigma^2, \qquad \text{var}\left[s^2\right] = \text{var}\left[\frac{\text{SC}_{\text{Res}}}{n-2}\right] = \frac{2\sigma^4}{n-2}$$

Error Estándar de la Regresión: $s = \sqrt{\text{CM}_{\text{Res}}} = \sqrt{\frac{\text{SC}_{\text{Res}}}{n-2}} = \sqrt{\frac{1}{n-2}\sum(y_i - \hat{y}_i)^2}$

Estimación por Máxima Verosimilitud

$$y_i - \beta_0 - \beta_1 x_i \sim \mathrm{N}(0, \sigma^2) \quad \textit{i.i.d.}$$

Función de verosimilitud del modelo de regresión:

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2; x, y) &= \prod_{i=1}^{n} f_\epsilon(x_i, y_i; \beta_0, \beta_1, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2 \right\} \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right\} \end{aligned}$$

$$\ell = \log L = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

−ebz

Estimación por Máxima Verosimilitud

Maximizar la función de verosimilitud equivale a minimizar la suma de cuadrados:

$$\left.\begin{array}{rcl} \frac{\partial \log L}{\partial \beta_0} & = & 0 \\ \frac{\partial \log L}{\partial \beta_1} & = & 0 \end{array}\right\}$$

La solución del sistema da lugar a los *Estimadores de Máxima Verosimilitud* (EMV).

$$\left\{\begin{array}{rcl} \hat{\beta}_1 & = & \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 & = & \bar{y} - \hat{\beta}_1 \bar{x} \end{array}\right.$$

Además,

$$\frac{\partial \log L}{\partial \sigma} = 0 \quad \Longrightarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Por lo tanto, los EMV son los mismos que los EMC, y

$$s^2 = \frac{n}{n-2}\,\hat{\sigma}^2, \qquad \hat{\sigma}^2 = \frac{n-2}{n}\,s^2$$

Nota: El hecho que los EMC coincidan con los EMV, añade las propiedades de estos últimos (consistencia, suficiencia) a la mínima varianza (eficiencia) de los EMC (Teorema Gauss-Markov).

–ebz

Observaciones

- El supuesto de normalidad de los errores es necesario para hacer inferencia pero no para que se cumplan las condiciones de *Gauss-Markov*.
- El supuesto $\epsilon \sim \mathrm{N}(0, \sigma^2)$ hace que los EMV y los EMC coincidan. (A excepción del estimador de la varianza.)
- **Teorema de Gauss-Markov.** Dentro de la clase de estimadores lineales insesgados para β_0 y β_1, los estimadores de mínimos cuadrados tienen varianza mínima.

  Se tiene que si $\tilde{\beta}_0$ es tal que $\mathbb{E}[\tilde{\beta}_0] = \beta_0 : \Rightarrow \mathrm{Var}(\tilde{\beta}_0) \geq \mathrm{Var}(\hat{\beta}_0)$

  Análogamente, si $\tilde{\beta}_1$ es tal que $\mathbb{E}[\tilde{\beta}_1] = \beta_1 : \Rightarrow \mathrm{Var}(\tilde{\beta}_1) \geq \mathrm{Var}(\hat{\beta}_1)$

  donde $\hat{\beta}_0$, $\hat{\beta}_1$ son los estimadores de mínimos cuadrados (EMC).

Distribuciones

Podemos escribir los EMC como funciones lineales de $y \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum c_i y_i, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \sum d_i y_i \qquad c_i, d_i \in \mathbb{R}$$

Luego,

$$\hat{\beta}_1 \sim \mathrm{N}(\beta_1, \sigma^2 \frac{1}{S_{xx}}), \qquad \hat{\beta}_0 \sim \mathrm{N}\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$$

e independientemente,

$$s^2 \sim \frac{\sigma^2}{n-2} \chi^2_{n-2}$$

Respuesta media ajustada $\hat{y}(x)$:

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \sim \mathrm{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\right)$$

Nueva observación $\dot{y}(x)$:

$$\dot{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon \sim \mathrm{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\right)$$

−ebz

Pruebas de Hipótesis sobre los coeficientes

Suponiendo σ^2 conocida

- Recta que pasa por β_0^0 al origen:

$$H_0 : \beta_0 = \beta_0^0 \quad vs. \quad H_1 : \beta_0 \neq \beta_0^0$$

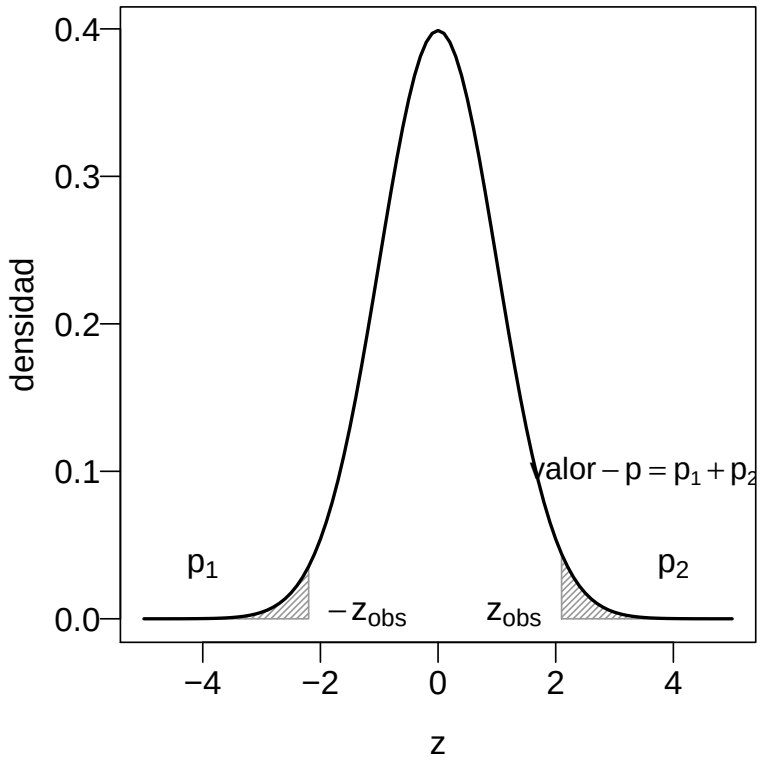
$$z_0 = \frac{\hat{\beta}_0 - \beta_0^0}{\text{de}(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0^0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim \text{N}(0,1)$$

- Recta de pendiente β_1^0:

$$H_0 : \beta_1 = \beta_1^0 \quad vs. \quad H_1 : \beta_1 \neq \beta_1^0$$

$$z_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{de}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1^0}{\sigma\sqrt{\frac{1}{S_{xx}}}} \sim \text{N}(0,1)$$

Distribución nula



Pruebas de Hipótesis sobre los coeficientes

Suponiendo σ^2 desconocida

- Recta que pasa por β_0^0 al origen:

$$H_0 : \beta_0 = \beta_0^0 \quad vs. \quad H_1 : \beta_0 \neq \beta_0^0$$

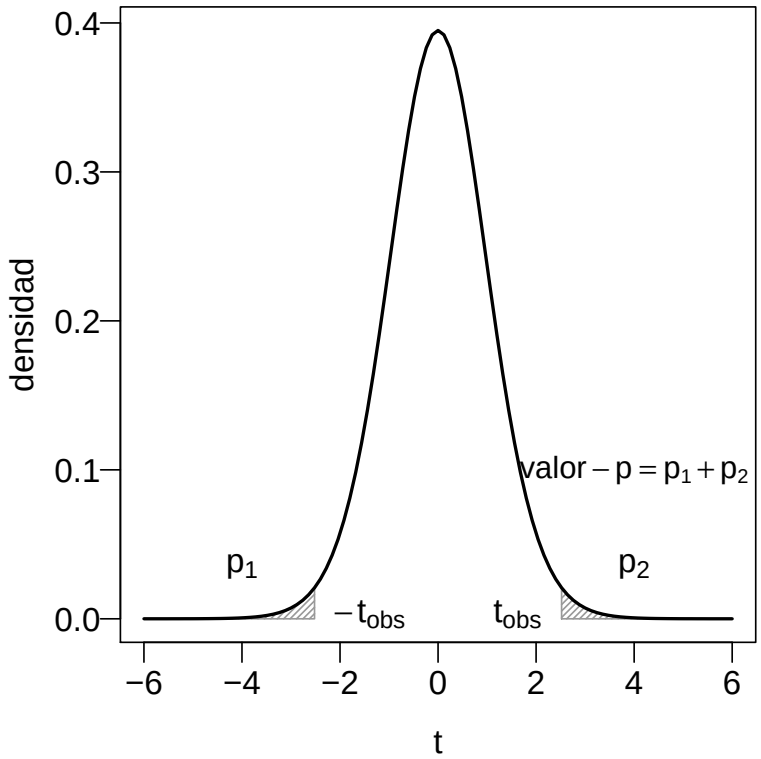
$$t_0 = \frac{\hat{\beta}_0 - \beta_0^0}{\text{ee}(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - \beta_0^0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2}$$

- Recta de pendiente β_1^0:

$$H_0 : \beta_1 = \beta_1^0 \quad vs. \quad H_1 : \beta_1 \neq \beta_1^0$$

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{ee}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1^0}{s\sqrt{\frac{1}{S_{xx}}}} \sim t_{n-2}$$

Distribución nula

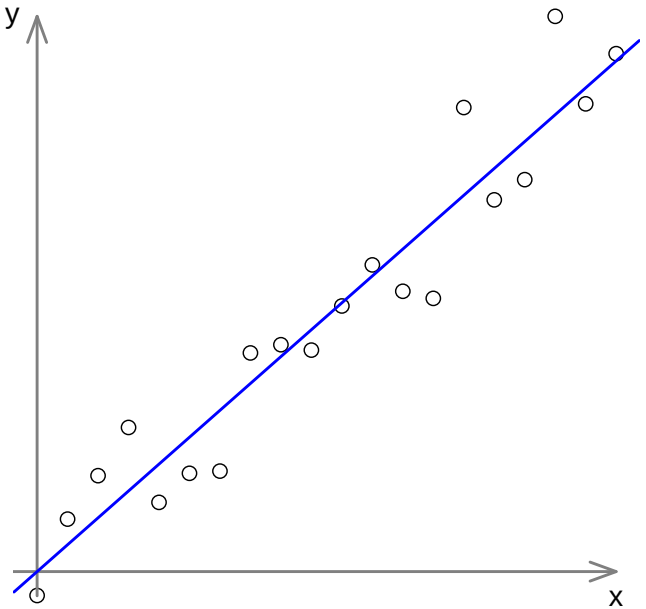


Pruebas de Hipótesis sobre los coeficientes

Casos particulares

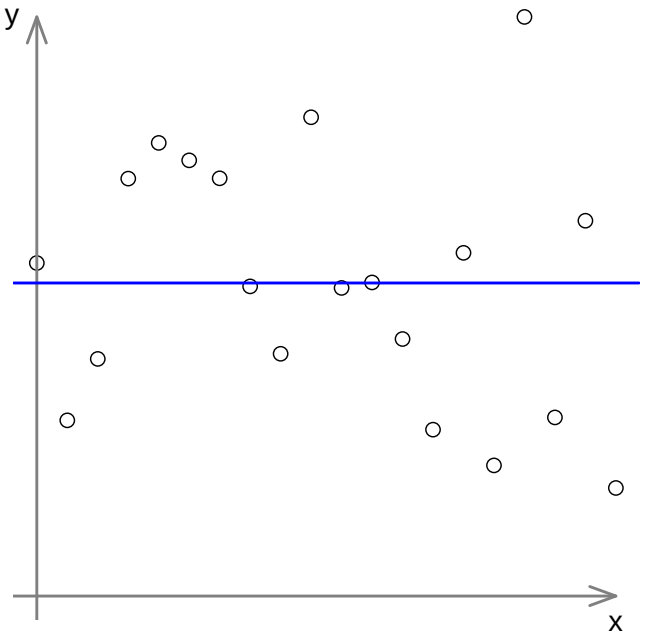
Recta que pasa por el origen:

$H_0 : \beta_0 = 0 \quad vs. \quad H_1 : \beta_0 \neq 0$



Significancia de la regresión:

$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$



Intervalos de Confianza

- Ordenada al origen β_0:

$$\hat{\beta}_0 \pm t_{(1-\alpha/2;n-2)} \cdot s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

- Pendiente β_1:

$$\hat{\beta}_1 \pm t_{(1-\alpha/2;n-2)} \cdot s\sqrt{\frac{1}{S_{xx}}}$$

- Varianza σ^2:

$$\left(\frac{(n-2)s^2}{\chi^2_{(1-\alpha/2,n-2)}}, \frac{(n-2)s^2}{\chi^2_{(\alpha/2,n-2)}} \right)$$

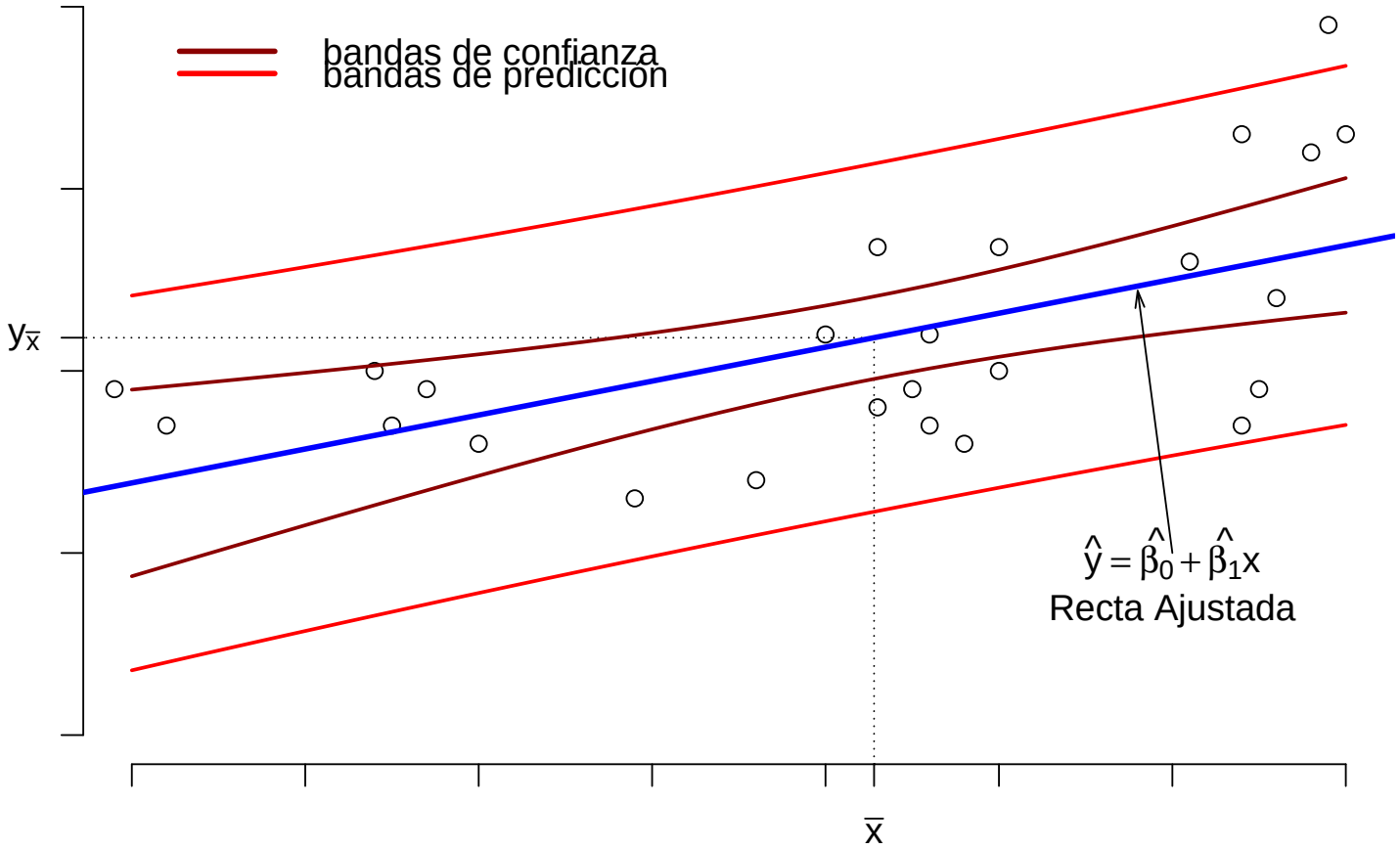
- Respuesta media $\hat{y}_x$, al nivel x:

$$\hat{y}_x \pm t_{(1-\alpha/2;n-2)} \cdot s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

- *Intervalos de predicción* – nueva observación $\dot{y}_x$:

$$\dot{y}_x \pm t_{(1-\alpha/2;n-2)} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

Bandas de confianza y bandas de predicción



Regiones de Confianza para $\beta = (\beta_0, \beta_1)$

Considere la reparametrización del modelo de regresión

$$y = \beta_0 + \beta_1 x + \epsilon = \beta_0' + \beta_1(x - \bar{x}) + \epsilon$$

Entonces

$$\hat{\beta}_0' = \bar{y} \quad \text{y} \quad \text{var}(\hat{\beta}_0') = \sigma^2/n$$

$$\left.\begin{array}{rcl} \left(\frac{\hat{\beta}_0' - \beta_0'}{\sigma/\sqrt{n}}\right)^2 & = & \frac{n(\hat{\beta}_0' - \beta_0')^2}{\sigma^2} \sim \chi_1^2 \\ \left(\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2 & = & \frac{S_{xx}(\hat{\beta}_1 - \beta_1)^2}{\sigma^2} \sim \chi_1^2 \end{array}\right\} \text{ Independientes}$$

Luego,

$$\frac{n(\hat{\beta}_0' - \beta_0')^2}{\sigma^2} + \frac{S_{xx}(\hat{\beta}_1 - \beta_1)^2}{\sigma^2} \sim \chi_2^2$$

Y por otro lado,

$$(n-2)s^2/\sigma^2 \sim \chi_{n-2}^2$$

independientemente de $\hat{\beta}_0'$ y $\hat{\beta}_1$. Entonces,

$$\frac{n(\hat{\beta}_0' - \beta_0')^2 + S_{xx}(\hat{\beta}_1 - \beta_1)^2}{2s^2} \sim F_{2,n-2}$$

Por lo que una región del $(1-\alpha)$ nivel de confianza para $\beta = (\beta_0, \beta_1)$ es:

$$\mathcal{RC}_{(1-\alpha)} = \left\{\beta \in \mathbb{R}^2 : n(\hat{\beta}_0 - \beta_0)^2 + 2(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)\sum x_i + (\hat{\beta}_1 - \beta_1)^2 \sum x_i^2 \leq 2s^2 F_{(1-\alpha;2,n-2)}\right\}$$

–ebz

Intervalos de Confianza y Método de Bonferroni

Sea IC_j el intervalo de $1-\alpha$ nivel de confianza para el parámetro β_j ($j=0,1$). Sea I_j el evento *"el intervalo IC_j efectivamente contiene el parámetro de interés β_j''. Luego,* $\mathbb{P}(I_j) = 1-\alpha$

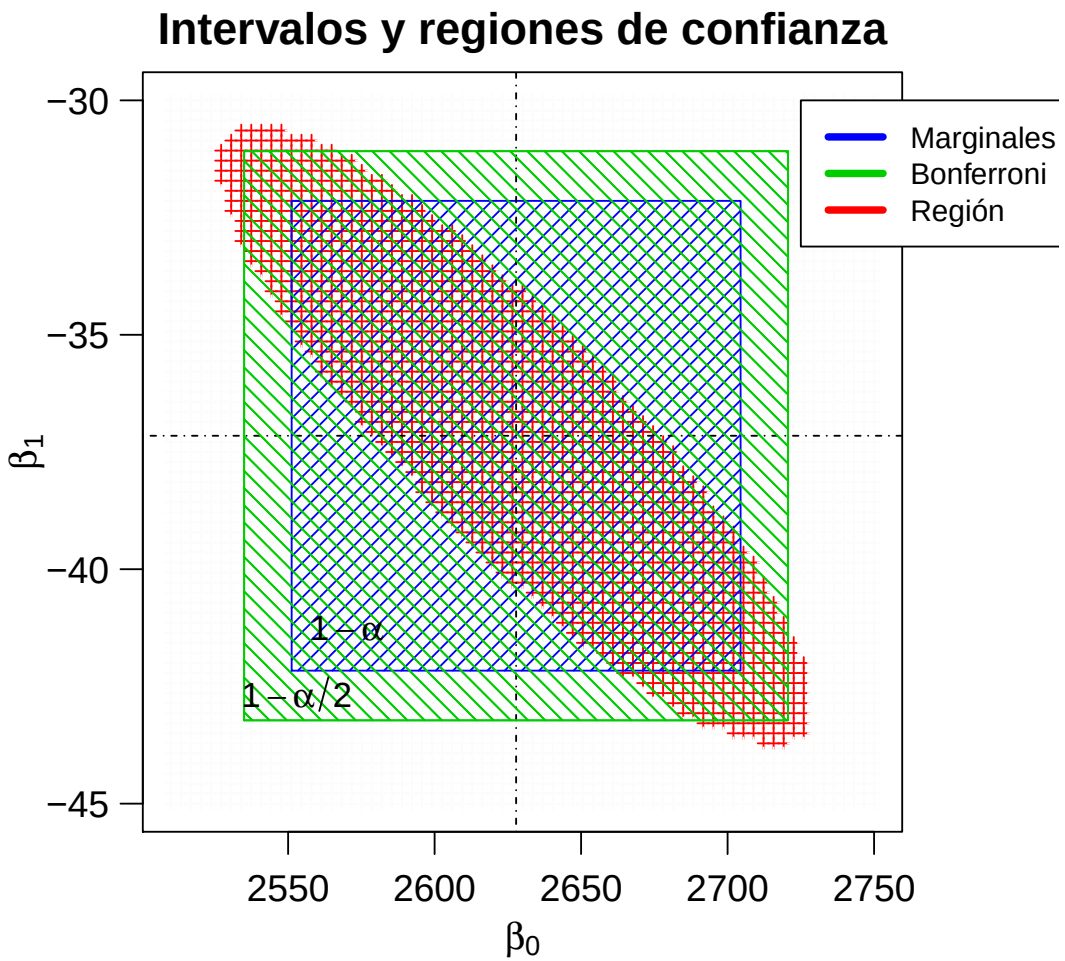
$$\begin{aligned}
\mathbb{P}(I_0 \cap I_1) &= 1 - \mathbb{P}((I_0 \cap I_1)^c) \\
&= 1 - \mathbb{P}(I_0^c \cup I_1^c) \\
&= 1 - \left[\mathbb{P}(I_0^c) + \mathbb{P}(I_1^c) - \mathbb{P}(I_0^c \cap I_1^c)\right] \\
&= 1 - 2\alpha + \mathbb{P}(I_0^c \cap I_1^c) \\
&\geq 1 - 2\alpha
\end{aligned}$$

Entonces, para garantizar un nivel de $(1-\alpha)$ confianza *conjunto*, construya marginalmente los intervalos IC_j con un nivel $(1-\alpha/2)$ de confianza.

Intervalos de Bonferroni

Para garantizar una *confianza conjunta* de $(1-\alpha)$) de k intervalos construya intervalos con niveles de confianza marginal de $(1-\alpha/k)$.

Regiones de Confianza para $\beta = (\beta_0, \beta_1)$



Ejemplo propelente (Montgomery et al)

−ebz

Análisis de la Varianza

Tabla de Análisis de Varianza

Fuente	gl	Suma de Cuadrados (SC)	Cuadrados Medios (CM)	F
Debido regresión	1	$\text{SC}_{\text{Reg}} = \sum(\hat{y}_i - \bar{y})^2$	$\text{CM}_{\text{Reg}} = \frac{\text{SC}_{\text{Reg}}}{1}$	$\frac{\text{CM}_{\text{Reg}}}{\text{CM}_{\text{Res}}}$
Residuales	$n-2$	$\text{SC}_{\text{Res}} = \sum(y_i - \hat{y}_i)^2$	$\text{CM}_{\text{Res}} = \frac{\text{SC}_{\text{Res}}}{n-2}$	
Total (Corregido)	$n-1$	$\text{SC}_{\text{Total}} = \sum(y_i - \bar{y})^2$		

Análisis de Varianza y Suma Extra de Cuadrados

Fuente	GL	Suma de Cuadrados	Cuadrados Medios	F
β_0	1	$n\bar{y}^2$		
$\beta_1 \mid \beta_0$	1	S_{xy}^2 / S_{xx}	CM_{Reg}	$\text{CM}_{\text{Reg}} / s^2$
Residuales	$n-2$	Por diferencia	s^2	
Total	n	$\sum y_i^2$		

Coeficiente de Determinación R^2

El coeficiente de determinación es el porcentaje de variabilidad debido a la regresión; es decir, el porcentaje de la variabilidad de los datos explicado por el modelo.

$$R^2 = \frac{\text{SC Debido a la regresión}}{\text{SC Total corregidos}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{SC}_{\text{Reg}}}{\text{SCT}_{\text{corr}}} = 1 - \frac{\text{SC}_{\text{Res}}}{\text{SCT}_{\text{corr}}}$$

Notas:

1. $R^2 = [\text{corr}(y, \hat{y})]^2$, de ahí que también se le conozca como *Coeficiente De Correlación Cuadrada*.
2. $\text{corr}(X, Y) = \text{signo}(\hat{\beta}_1) \cdot R$
3. En la definición de R^2, se considera siempre un *modelo base*, que en el caso de la RLM es el modelo *trivial*: $y = \bar{y} + \epsilon$. Evite el uso de R^2 en modelos sin ordenada al origen.

Coeficiente de Determinación R^2

Notas:

4. $0 \leq R^2 \leq 1$. Sin embargo, si se tienen *réplicas puras*, (varios $y(x)$'s para el mismo $X = x$), se tiene *error puro* y *no hay modelo* que capture la variación debida a este error. En este caso: $R^2 < 1$.
5. R^2 cercanas a 1 indican *"buen ajuste"*. Recuerde sin embargo que *"buen ajuste"* y *"mal ajuste"* depende del contexto.
6. En el caso de RLS, se puede mostrar que

$$\mathbb{E}[R^2] = \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2 + \hat{\beta}_1^2 S_{xx}}$$

   por lo que es posible incrementar R^2 aumentando el rango del regresor X.
7. Una R^2 alta no implica un buen modelo de predicción.
8. En RLM, siempre se puede incrementar R^2 aumentando el número de regresores.

−ebz

Coeficiente de Correlación (Muestral)

Considere la muestra $(x_1, y_1), \ldots, (x_n, y_n)$. La correlación muestral entre X y Y está dada por

$$r = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\left[\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2\right]^{1/2}} = \frac{S_{XY}}{[S_{XX} S_{YY}]^{1/2}}$$

Se puede mostrar que

$$\hat{\beta}_1 = r\left(\frac{S_{YY}}{S_{XX}}\right)^{1/2}$$

y que r^2 es el *Coeficiente de Determinación*. Es decir, $r^2 = R^2$.

Regresión Simple por el Origen

El modelo de una línea recta que pasa por el origen es:

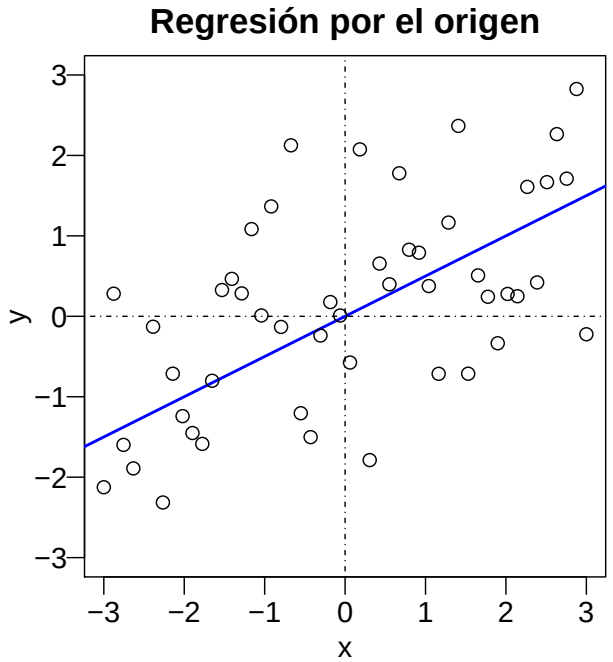
$$y_i = \beta x_i + \epsilon_i,$$

tiene los siguientes estimadores (insesgados) de mínimos cuadrados:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{S_{xy}}{S_{xx}}$$

y

$$s^2 = \frac{1}{n-1} \sum (y_i - \hat{y}_i)^2$$



Los estimadores $\hat{\beta}$ y s^2 tiene propiedades similares a las correspondientes en los modelos con ordenada al origen. A saber,

$$\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2 \frac{1}{S_{xx}}) \qquad \text{y} \qquad s^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}$$

Recuerde que en este caso la R^2 no es fácilmente interpretable.

–ebz