IECD-325: Regresión lineal simple

Felipe Osorio

felipe.osorio@uv.cl

Datos de Forbes

Suponga el modelo de regresión

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

donde y es la variable dependiente (respuesta) y x es la variable regresora o explicativa.

Los parámetros β_0 y β_1 corresponden al intercepto y la pendiente, respectivamente, mientras que $\{\epsilon_i\}$ representa disturbios aleatorios.

Para propósitos de inferencia estadística, supondremos

$$\mathsf{E}(\epsilon_i) = 0, \qquad \mathsf{var}(\epsilon_i) = \sigma^2, \qquad \mathsf{cov}(\epsilon_i, \epsilon_j) = 0,$$

para $i, j = 1, \ldots, n$ $(i \neq j)$.

De este modo, dispondremos de $(x_1, y_1), \ldots, (x_n, y_n)$ observaciones desde (x, y).

Regresión lineal (simple)

Deseamos hallar β_0 y β_1 tal que produzcan el mejor ajuste a los datos[1]. En este curso usaremos el método de mínimos cuadrados ordinarios, dado por

$$\min_{\theta} Q(\boldsymbol{\beta}),$$

con $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ y

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

Observación:

La función $Q(\boldsymbol{\beta})$ es conocida como suma de cuadrados de los errores.

[1] Es decir, deseamos obtener estimadores para α y β

Regresión lineal (simple)

Usando el método de minimos cuadrados obtenemos:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x},$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

La recta de regresión es dada por:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \qquad i = 1, \ldots, n,$$

y llamamos a $\widehat{y}_i$ es valor predicho (o valor ajustado). Además,

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \qquad i = 1, \ldots, n,$$

es conocido como el i-ésimo residuo.

Una medida de variabilidad es dada por:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

Mientras que

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2},$$

se denomina coeficiente de determinación.[2]

Interpretación:

R^2 es la varianza de los datos que puede ser explicada por el modelo.

[2] permite medir la calidad (bondad) del ajuste

Es posible notar que (cuando el modelo tiene intercepto):

$$R^2 = 1 - \frac{RSS}{s^2_{\text{DATOS}}},$$

con

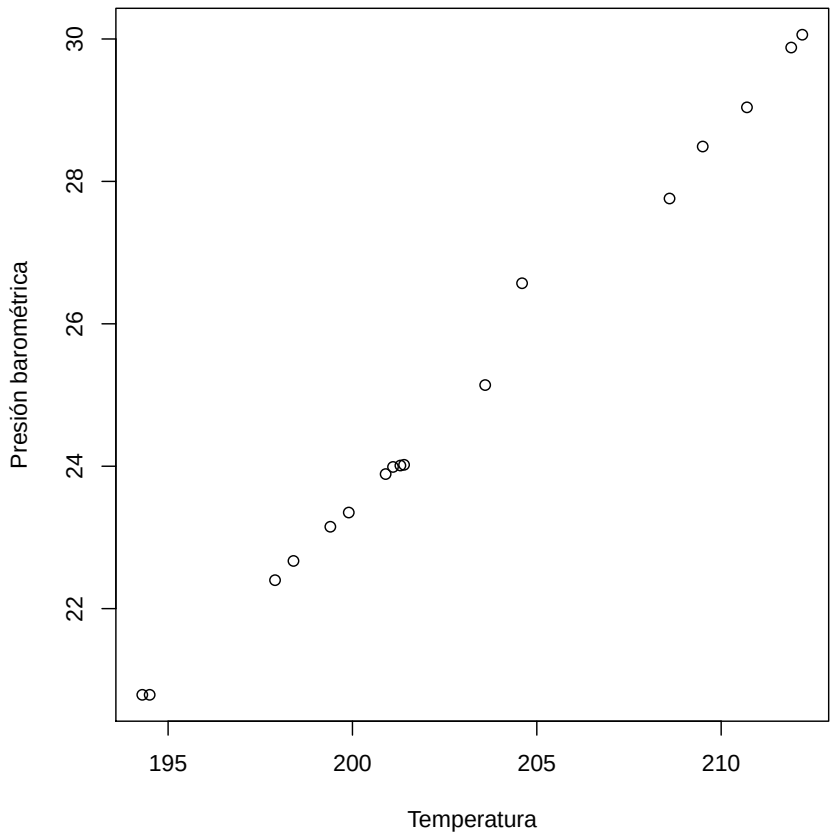
$$RSS = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2, \qquad s^2_{\text{DATOS}} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Interpretación:

En efecto, $0 \leq R^2 \leq 1$ permite medir la calidad (o bondad) del ajuste.

Datos de Forbes

Presión barométrica en pulgadas de mercurio y temperatura de ebullición del agua en grados Fahrenheit para 17 diferentes altitudes.



Datos de Forbes

```
> library(MASS)
> data(forbes) # disponibiliza los datos en la sesión

> forbes
      bp  pres
1  194.5 20.79
2  194.3 20.79
3  197.9 22.40
4  198.4 22.67
5  199.4 23.15
6  199.9 23.35
7  200.9 23.89
8  201.1 23.99
9  201.4 24.02
10 201.3 24.01
11 203.6 25.14
12 204.6 26.57
13 209.5 28.49
14 208.6 27.76
15 210.7 29.04
16 211.9 29.88
17 212.2 30.06

```

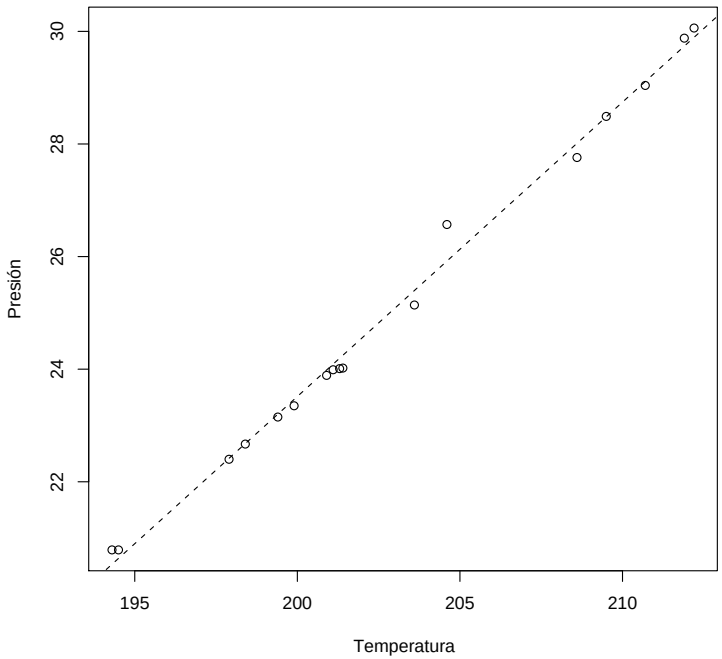
Datos de Forbes

```
# ajuste de un modelo de regresión
> fm <- lm(pres ~ bp, data = forbes)

# salida:
> fm

Call:
lm(formula = pres ~ bp, data = forbes)

Coefficients:
(Intercept)           bp
   -81.0637       0.5229

# residuos y valores ajustados
> res <- residuals(fm)
> fit <- fitted(fm)

# otra forma de calcular R^2
> cor(fit, forbes$pres)^2
[1] 0.9944282

```

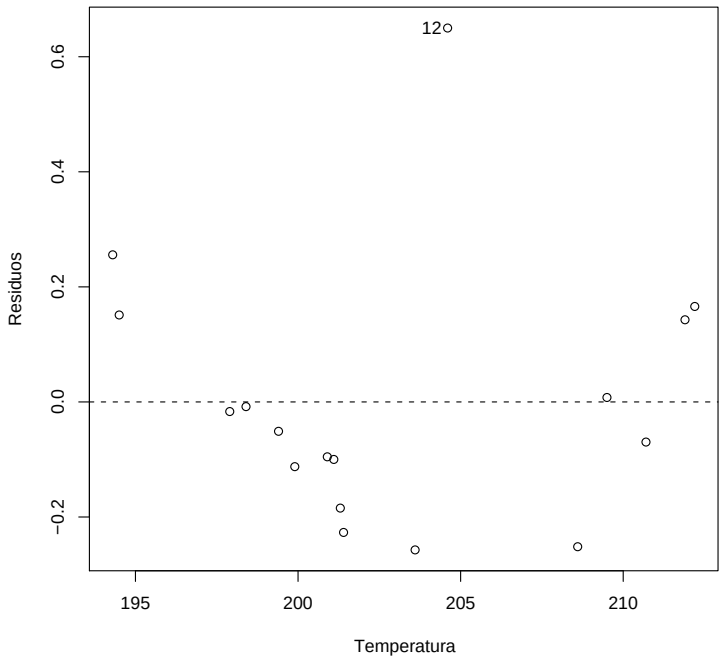
Datos de Forbes

```
# salida un poco más extensa
> summary(fm)

Call:
lm(formula = pres ~ bp, data = forbes)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25717 -0.11246 -0.05102  0.14283  0.64994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.06373    2.05182  -39.51   <2e-16 ***
bp            0.52289    0.01011   51.74   <2e-16 ***

Residual standard error: 0.2328 on 15 degrees of freedom
Multiple R-squared:  0.9944,  Adjusted R-squared:  0.9941
F-statistic:  2677 on 1 and 15 DF,  p-value: < 2.2e-16

```

Datos de Forbes



(a) recta ajustada



(b) residuos vs. ajuste

Datos de Forbes

Ahora consideramos el modelo

$$100 \times \log_{10}(\text{Presión}_i) = \beta_0 + \beta_1 \text{Temperatura}_i + \epsilon_i,$$

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

Se obtuvo (usando función `lm` de **R**)

$$\widehat{\boldsymbol{\beta}} = (-42.1378, 0.8955)^\top \quad \text{y} \quad s^2 = 0.1438$$

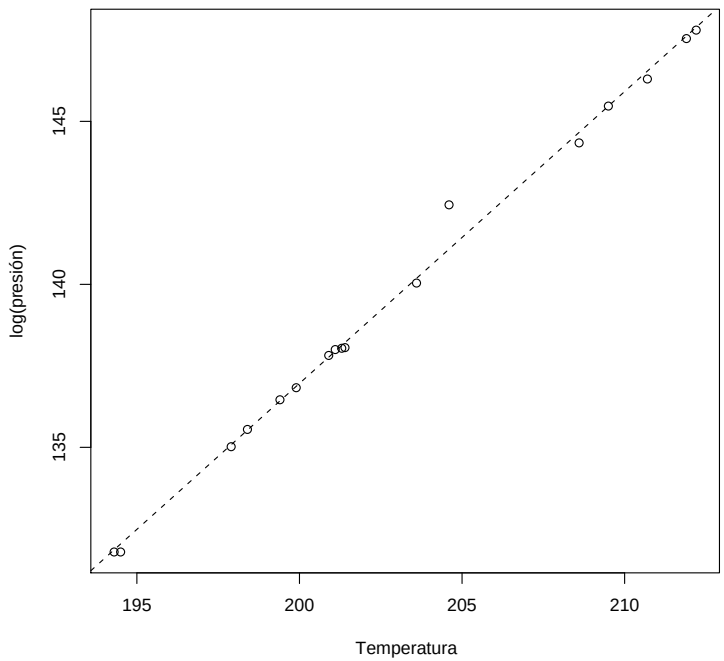
Además, $R^2 = 0.9950$.

Datos de Forbes

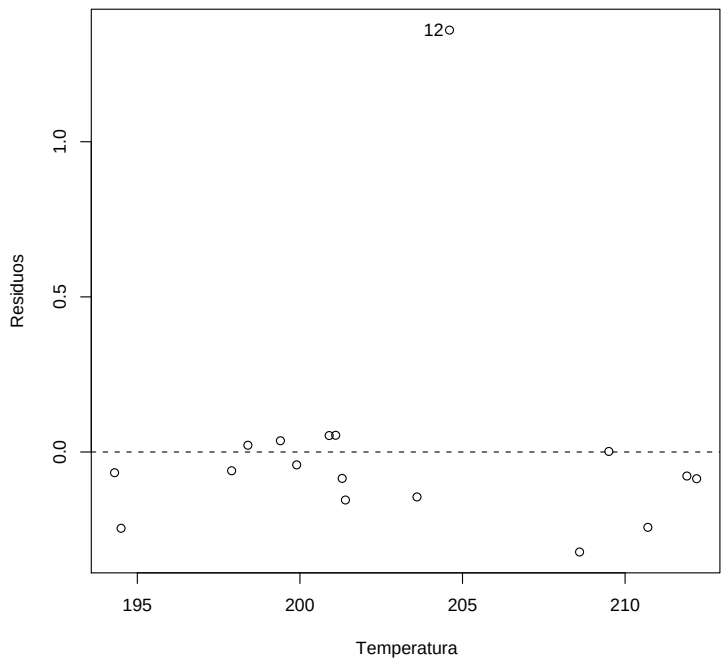
```
# modelo con datos transformados
> f1 <- lm(100 * log10(pres) ~ bp, data = forbes)

> summary(f1)

Call:
lm(formula = 100 * log10(pres) ~ bp, data = forbes)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31974 -0.14707 -0.06890  0.01877  1.35994

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.16418    3.34136  -12.62 2.17e-09 ***
bp            0.89562    0.01646   54.42  < 2e-16 ***

Residual standard error: 0.3792 on 15 degrees of freedom
Multiple R-squared:  0.995, Adjusted R-squared:  0.9946
F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16

```

Datos de Forbes

Recta de regresión y gráfico de residuos para los datos de Forbes[3].



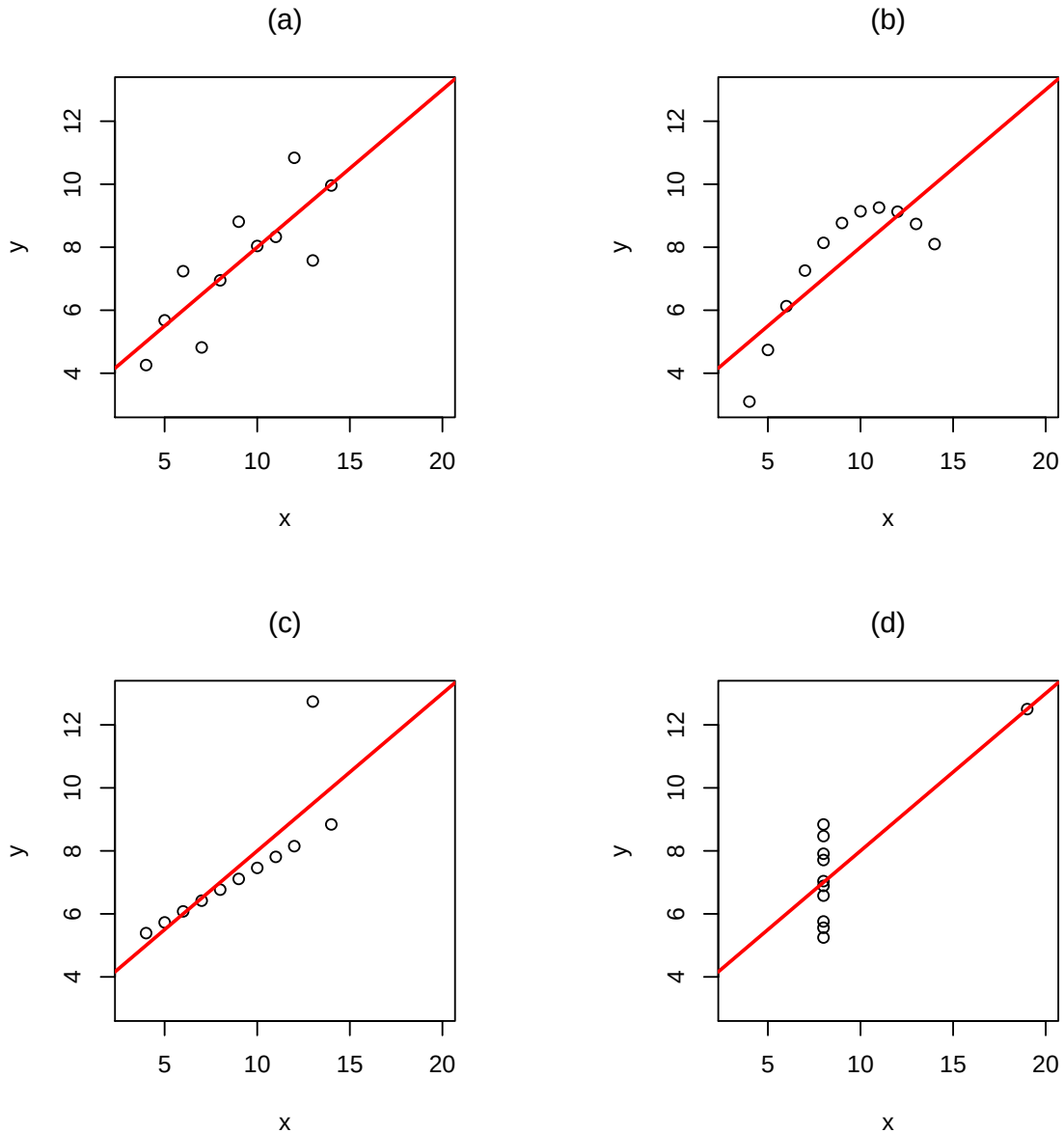
(a) recta ajustada



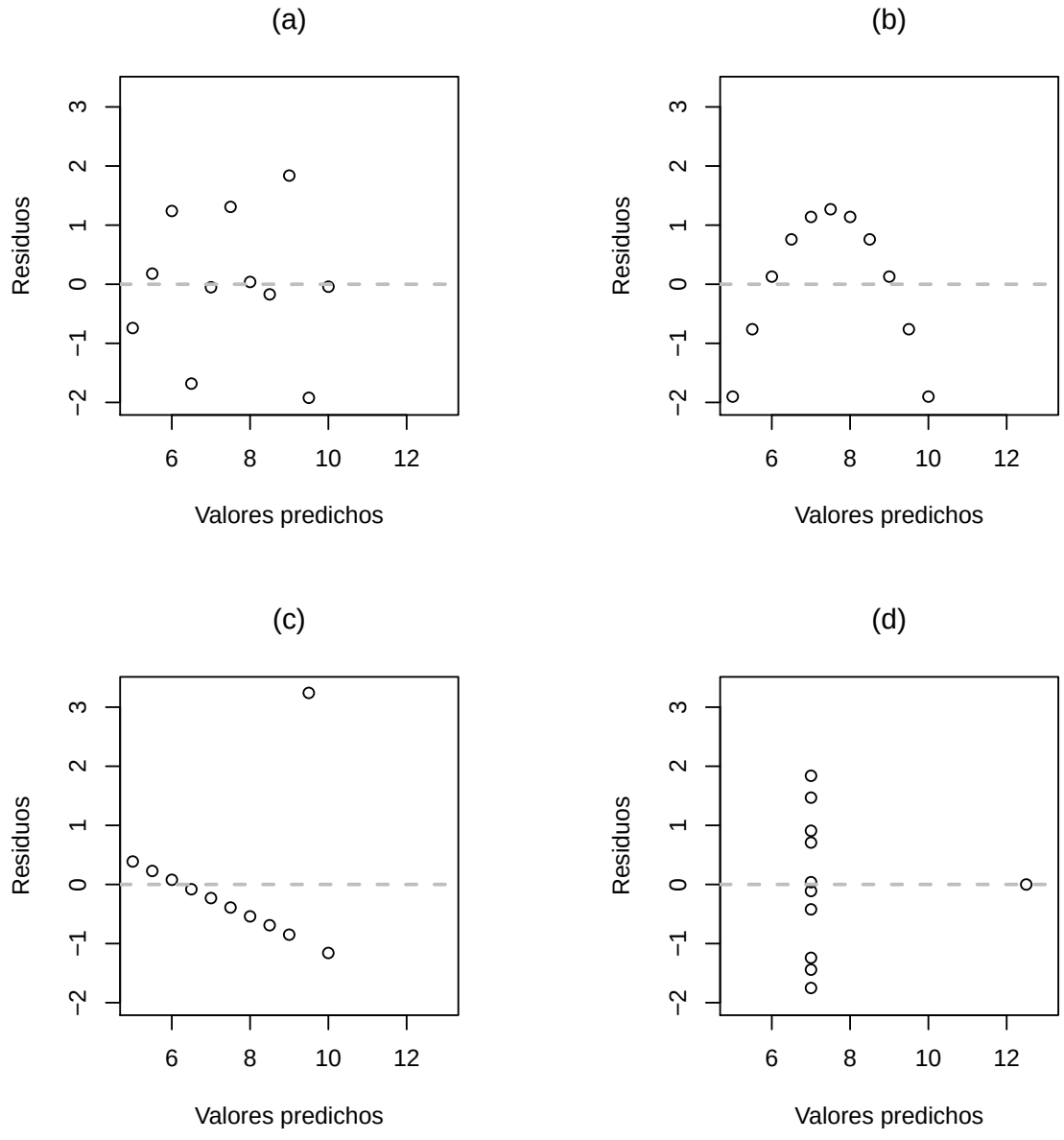
(b) residuos vs. ajuste

[3] datos transformados

Cuarteto de regresiones “idénticas” de Anscombe (1973)



Cuarteto de regresiones “idénticas” de Anscombe (1973)



Cuarteto de regresiones "idénticas" de Anscombe (1973)

Observaciones:

- Para el cuarteto de regresiones de Anscombe se obtiene (para todos los modelos):

  $$\widehat{\beta}_0 = 3.001, \quad \widehat{\beta}_1 = 0.500, \quad s^2 = 1.528, \quad R^2 = 0.666, \quad F = 17.97, \quad p = 0.002$$

  confiar solamente en medidas globales puede ser engañoso.

- En efecto, note que

  $$\sum_{i=1}^{n} e_i = 0, \qquad \sum_{i=1}^{n} e_i \widehat{y}_i = 0,$$

  de modo que el gráfico de dispersión de residuos vs. valores predichos no debería presentar algún comportamiento sistemático.

- Es recomendable realizar un análisis de residuos o de diagnóstico.