

# MAT-042: Correlación y Regresión lineal

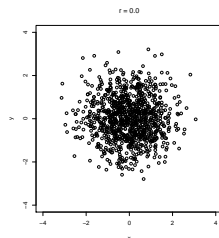
**Felipe Osorio**

[fosorios.mat.utfsm.cl](mailto:fosorios.mat.utfsm.cl)

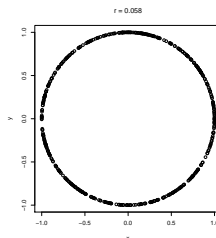
Departamento de Matemática, UTFSM



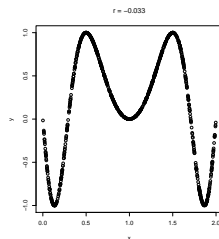
# Correlación: Midiendo asociación lineal



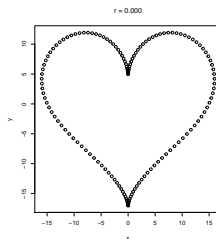
(a)  $r = 0.000$



(b)  $r = 0.058$

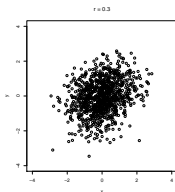


(c)  $r = -0.033$

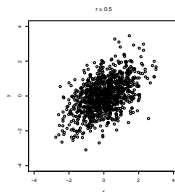


(d)  $r = 0.000$

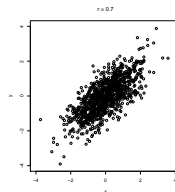
# Correlación: Midiendo asociación lineal



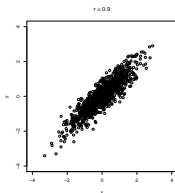
(a)  $r = 0.30$



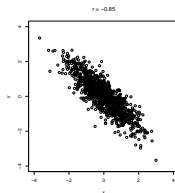
(b)  $r = 0.50$



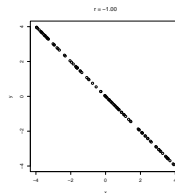
(c)  $r = 0.70$



(d)  $r = 0.90$



(e)  $r = -0.85$



(f)  $r = -1.00$

## Definición 1 (Covarianza):

Para el conjunto  $(x_1, y_1), \dots, (x_n, y_n)$ , se define la **covarianza** como una medida de **variabilidad conjunta** de dos variables cuantitativas, como:

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

### *Observación:*

Evidentemente,  $\text{cov}(\mathbf{x}, \mathbf{x}) = \text{var}(\mathbf{x}) = s_x^2$ .



## Propiedades:

(a)

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}.$$

(b)

$$\text{cov}(ax + b, cy + d) = ac \text{cov}(x, y).$$



## Definición 2 (Correlación):

La **correlación** entre  $\mathbf{x} = (x_1, \dots, x_n)^\top$  e  $\mathbf{y} = (y_1, \dots, y_n)^\top$  es la covarianza de sus versiones estandarizadas. Es decir,

$$\begin{aligned}\text{corr}(\mathbf{x}, \mathbf{y}) &= \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\sqrt{\text{var}(\mathbf{x}) \text{var}(\mathbf{y})}}.\end{aligned}$$

## Observación:

$\text{corr}(\mathbf{x}, \mathbf{y})$  es una medida adimensional.



## Propiedades:

(a)

$$\text{corr}(ax + b, cy + d) = \pm \text{corr}(x, y).$$

(b)

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2 \text{cov}(x, y).$$

(c)

$$\{\text{corr}(x, y)\}^2 \leq 1.$$

## Observación:

- ▶ Evidentemente,  $-1 \leq \text{corr}(x, y) \leq 1$ .
- ▶ Cuando  $\text{corr}(x, y) = 0$ , diremos que  $x$  e  $y$  son **no correlacionados**.



## Definición 3 (Coeficiente de correlación de Spearman):

Suponga los datos pareados  $(x_1, y_1), \dots, (x_n, y_n)$ . Sea  $R_i, S_i$  los rangos de  $x_i$  e  $y_i$ , respectivamente ( $i = 1, \dots, n$ ). Entonces el **coeficiente de correlación de Spearman** es dado por

$$r_S = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}}.$$

### Observación:

Sea

$$D = \sum_{i=1}^n (R_i - S_i)^2,$$

y suponga que no existen empates entre los  $x$ 's e  $y$ 's, entonces podemos escribir

$$r_S = 1 - \frac{6D}{n^2 - 1}.$$





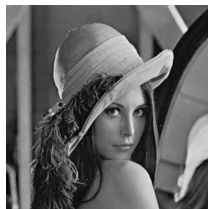
## Ejemplo: Distorsiones de Lenna<sup>1</sup>



(a)  $r = 1.000$



(b)  $r = 0.903$



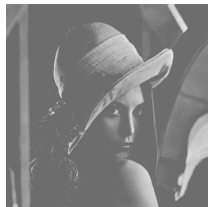
(c)  $r = 0.991$



(d)  $r = 0.915$



(e)  $r = 0.983$



(f)  $r = 0.799$

<sup>1</sup>(a) Original, (b) sal y pimienta, (c) filtro mediana, (d) ruido speckle, (e) filtro Lee, (f) imagen saturada.

## Ejemplo: Distorsiones de Lenna

```
> library(SpatialPack)
# https://github.com/faosorios/SSIM/blob/master/data/lena.rda
> load("lena.rda") # carga datos de Lenna
# aplica distorsiones y filtros
> lena.speckle <- imnoise(lena, type = "speckle")
> lena.sp <- imnoise(lena, type = "saltndpepper")
> lena.med <- denoise(lena.sp, type = "median")
> lena.lee <- denoise(lena.speckle, type = "Lee")
> lena.05 <- clipping(lena, low = 0.5) # saturación
# calculando correlaciones
> x <- as.vector(lena) # 262144 observaciones
> cor(x, x)
[1] 1
> cor(x, as.vector(lena.speckle))
[1] 0.9154696
> cor(x, as.vector(lena.sp))
[1] 0.9028631
> cor(x, as.vector(lena.med))
[1] 0.9907281
> cor(x, as.vector(lena.lee))
[1] 0.9829129
> cor(x, as.vector(lena.05))
[1] 0.7997093
```



## Datos de Flores Iris (Anderson, 1935; Fisher, 1936)



(a) *setosa*

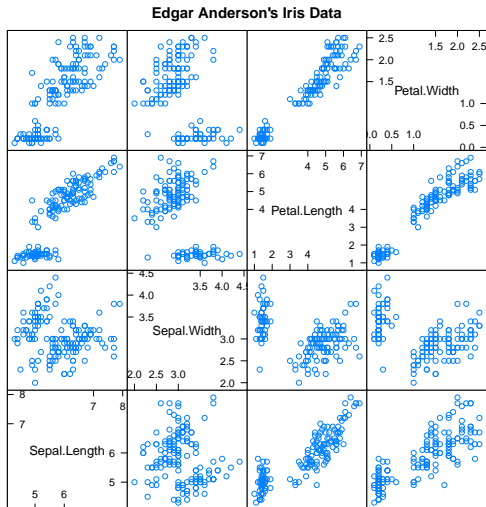


(b) *versicolor*



(c) *virginica*

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)



Deseamos estudiar  $p$  variables (características) de interés asociadas a una muestra aleatoria  $\mathbf{x}_1, \dots, \mathbf{x}_n$  donde cada  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$  es un vector  $p$ -dimensional.

Podemos disponer la información en una matriz

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}.$$



Análogamente a la media y varianza muestrales  $\bar{x}$  y  $s^2$ , podemos definir sus contrapartes multivariadas como:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i,$$

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

que representan el **vector de medias** y la **matriz de covarianza**, respectivamente.

## *Observación:*

En este caso, tenemos  $\mathbf{S} = (s_{ij})$ , donde

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j),$$

con  $\bar{x}_i = (\sum_{k=1}^n x_{ki})/n$ .



Los elementos anteriores permiten definir la **matriz de correlación** entre las  $p$  variables, como:

$$\mathbf{R} = (r_{ij})$$

donde

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

*Observación:*

Defina  $\mathbf{D} = \text{diag}(s_{11}, s_{22}, \dots, s_{pp})$ , de este modo, podemos definir

$$\mathbf{R} = \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2}.$$



# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

## Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

## Base de datos:

```
# Datos de flores Iris
```

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
...					
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica





# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

## Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

## Base de datos:

```
# Datos de flores Iris
```

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
...					
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica



## Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Matriz de Correlación ( $R$ ):

	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
Largo Sépalo	1.000	-0.118	0.872	0.818
Ancho Sépalo	-0.118	1.000	-0.428	-0.366
Largo Pétalo	0.872	-0.428	1.000	0.963
Ancho Pétalo	0.818	-0.366	0.963	1.000

Cálculo en R:

```
> cor(iris[,1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000	-0.1176	0.8718	0.8179
Sepal.Width	-0.1176	1.0000	-0.4284	-0.3661
Petal.Length	0.8718	-0.4284	1.0000	0.9629
Petal.Width	0.8179	-0.3661	0.9629	1.0000



## Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Matriz de Correlación ( $R$ ):

	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
Largo Sépalo	1.000	-0.118	0.872	0.818
Ancho Sépalo	-0.118	1.000	-0.428	-0.366
Largo Pétalo	0.872	-0.428	1.000	0.963
Ancho Pétalo	0.818	-0.366	0.963	1.000

Cálculo en R:

```
> cor(iris[,1:4])
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000	-0.1176	0.8718	0.8179
Sepal.Width	-0.1176	1.0000	-0.4284	-0.3661
Petal.Length	0.8718	-0.4284	1.0000	0.9629
Petal.Width	0.8179	-0.3661	0.9629	1.0000



## Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Se obtuvo además el **vector de medias** ( $\bar{x}$ ) y la **matriz de Covarianza** ( $S$ ):<sup>2</sup>

```
> z <- cov.wt(iris[,1:4])
> z
$cov
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	0.6856935	-0.0424340	1.2743154	0.5162707
Sepal.Width	-0.0424340	0.1899794	-0.3296564	-0.1216394
Petal.Length	1.2743154	-0.3296564	3.1162779	1.2956094
Petal.Width	0.5162707	-0.1216394	1.2956094	0.5810063

```
$center
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.843333	3.057333	3.758000	1.199333

```
$n.obs
[1] 150
```

---

<sup>2</sup>Análogamente podemos usar `cov(iris[,1:4])`.



# Objetivo del análisis de regresión

Estudiar una variable de **respuesta**,  $y$  [asumienda continua] como función de una variable explicativa o **regresor**,  $x$  [puede ser discreta y/o continua].

Entrada  $\rightarrow$  **Proceso**  $\rightarrow$  Salida

En ocasiones la relación funcional es **conocida** salvo algunos coeficientes (**parámetros**).

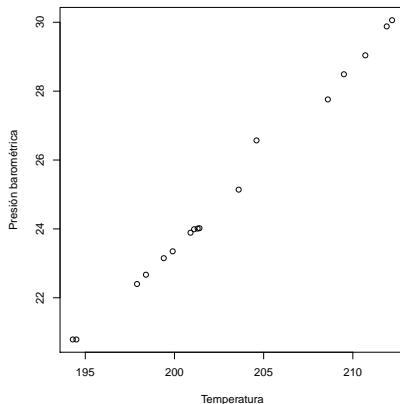
Es decir, la relación es gobernada por un **proceso físico** o por leyes bien aceptadas

$$y \approx f(x; \theta),$$

en cuyo caso, el interés recae en **estimar el vector de parámetros**  $\theta = (\alpha, \beta)^\top$ .



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



Para describir la relación entre la temperatura y la media de la presión barométrica, podemos considerar

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

El conjunto de datos consiste del **vector de respuestas**.

$$\mathbf{y} = (y_1, \dots, y_n)^\top,$$

y una **matriz de diseño**  $n \times 2$

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$



## Regresión lineal (simple)

Deseamos hallar  $\alpha$  y  $\beta$  tal que produzcan el **mejor ajuste** a los datos<sup>3</sup>. En este curso usaremos el **método de mínimos cuadrados** ordinarios, dado por

$$\min_{\theta} S(\theta),$$

con  $\theta = (\alpha, \beta)^\top$  y

$$S(\theta) = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^n \epsilon_i^2$$

### *Observación:*

La función  $S(\theta)$  es conocida como **suma de cuadrados de los errores**.

---

<sup>3</sup>Es decir, deseamos obtener estimadores para  $\alpha$  y  $\beta$





## Regresión lineal (simple)

Usando el **método de mínimos cuadrados** obtenemos:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\text{var}(\mathbf{x})}.$$

La **recta de regresión** es dada por:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i, \quad i = 1, \dots, n,$$

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

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} + \hat{\beta}x_i, \quad i = 1, \dots, n,$$

es conocido como el  $i$ -ésimo **residuo**.



## Regresión lineal (simple)

Una **medida de variabilidad** es dada por:

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

Mientras que

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

se denomina **coeficiente de determinación**.<sup>4</sup>

*Interpretación:*

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

---

<sup>4</sup>permite medir la calidad (bondad) del ajuste



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

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

con

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad s_{\text{DATOS}}^2 = \sum_{i=1}^n (y_i - \bar{y})^2.$$

*Interpretación:*

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



```
> 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
```



```
# 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
```



```
# 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:
```

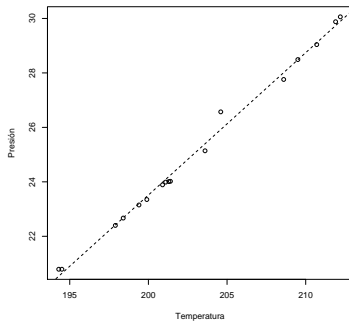
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-81.06373	2.05182	-39.51	<2e-16 ***
bp	0.52289	0.01011	51.74	<2e-16 ***

```
Residual standard error: 0.2328 on 15 degrees of freedom
```

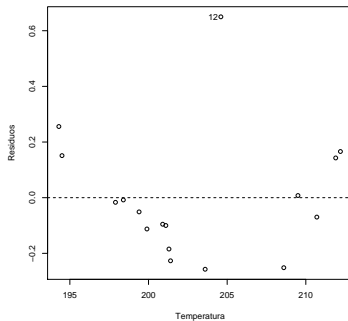
```
Multiple R-squared: 0.9944, Adjusted R-squared: 0.9941
```

```
F-statistic: 2677 on 1 and 15 DF, p-value: < 2.2e-16
```





(a) recta ajustada



(b) residuos vs. ajuste

Ahora consideramos el modelo

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

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

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

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

Además,  $R^2 = 0.9950$ .





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

> summary(f1)

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

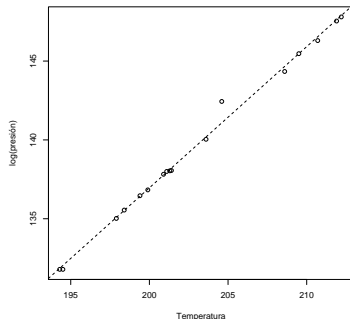
Residuals:
    Min       1Q   Median       3Q      Max
-0.31974 -0.14707 -0.06890  0.01877  1.35994

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 ***

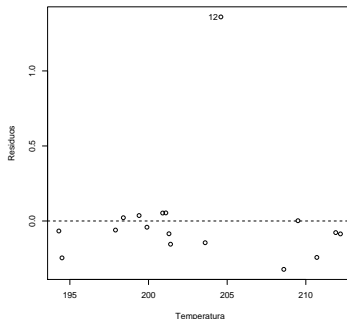
Residual standard error: 0.3792 on 15 degrees of freedom
Multiple R-squared:  0.995,    Adjusted R-squared:  0.9946
F-statistic: 2962 on 1 and 15 DF, p-value: < 2.2e-16
```



Recta de regresión y gráfico de residuos para los datos de Forbes<sup>5</sup>.



(a) recta ajustada

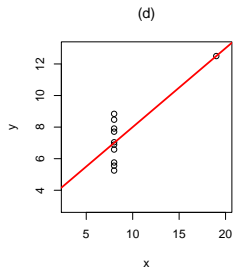
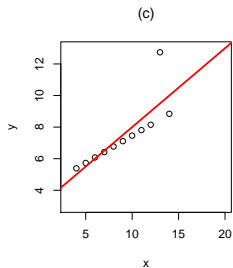
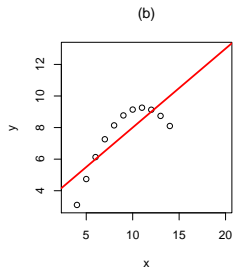
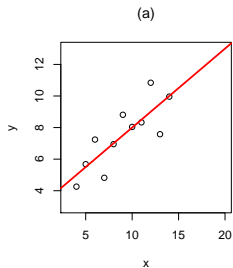


(b) residuos vs. ajuste

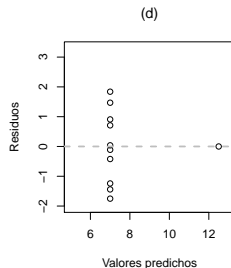
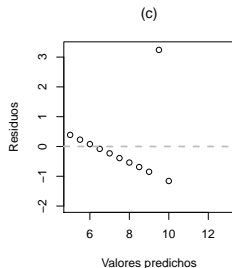
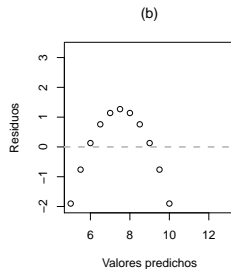
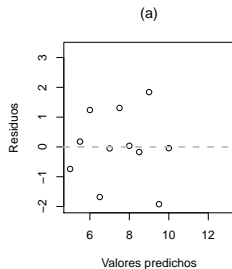
---

<sup>5</sup>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):

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

confiar **solamente** en medidas globales puede ser **engañoso**.

- ▶ En efecto, note que

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

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**.<sup>6</sup>

---

<sup>6</sup>Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, New York.

