Regresión Lineal
Modelo Lineal
Tópicos de Modelo Lineal

María Eugenia Szretter Noste

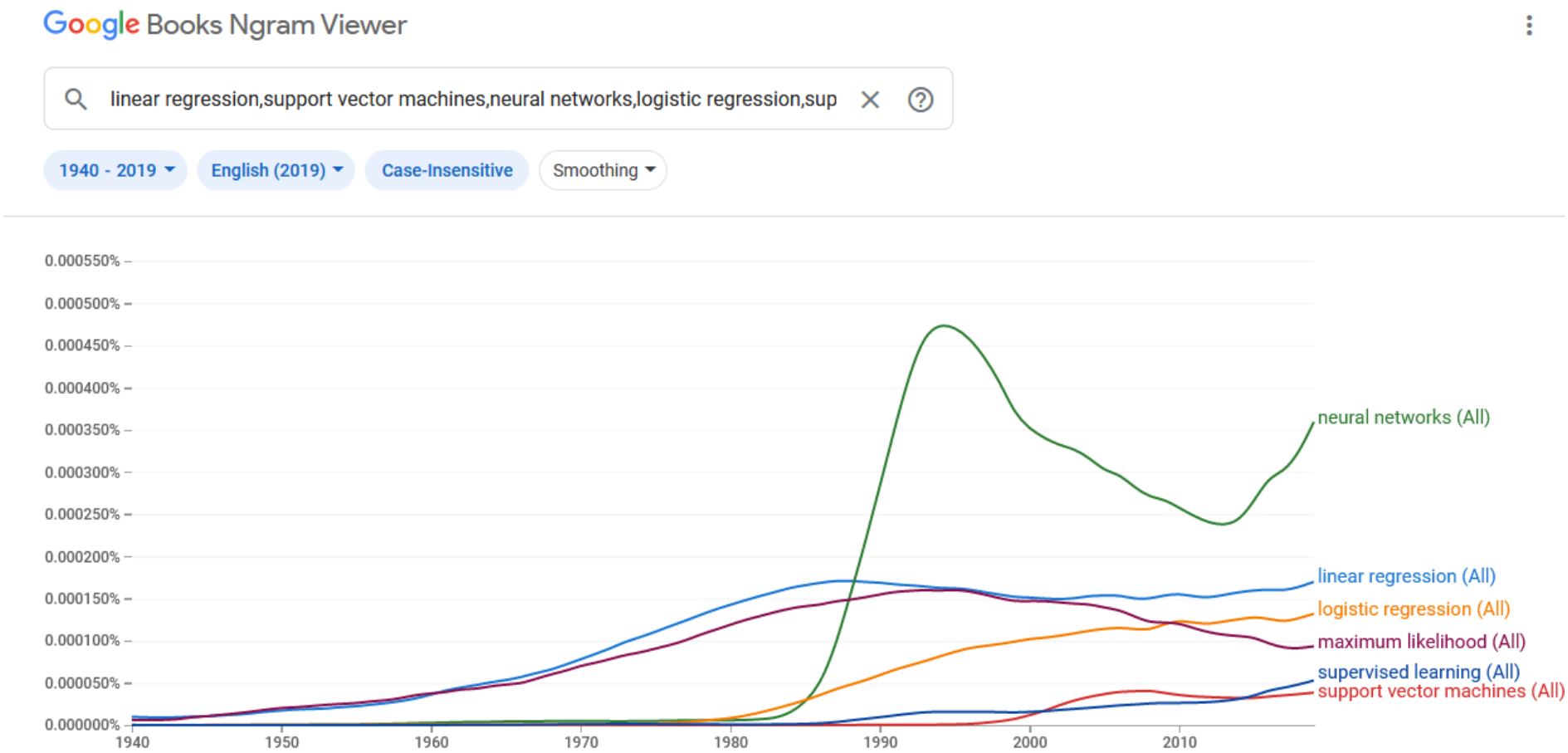
Instituto de Cálculo
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Segundo cuatrimestre 2023

En muchas disciplinas científicas interesa saber cómo se relacionan distintas variables entre sí. Una de las herramientas principales que tiene la estadística para hacer eso es la regresión

El modelo de regresión lineal es un método conceptualmente simple para investigar la relación entre dos o más variables. Esta relación se expresa en la forma de una ecuación o un modelo que conecta una variable respuesta o variable dependiente (continua) y una o muchas variables explicativas o covariables. Es una técnica clásica y muy utilizada.

Evolución de las metodologías en publicaciones



¿Por qué entonces estudiar modelos lineales?

1. La teoría de modelos lineales es un caso especial de la teoría más general que cubre modelos más flexibles y realistas. Precisamente porque es un caso tan especial, permite muchos atajos simplificadores, que pueden facilitan el aprendizaje, especialmente sin matemáticas avanzadas.
2. Debido a que los modelos lineales son tan simples, han sido y son tremendamente utilizados. Esto significa que muchas aplicaciones de la estadística se ha realizado sobre modelos lineales. También significa que muchos de los consumidores de estadística esperan modelos lineales o compararán los modelos obtenidos con modelos lineales. Por tanto, es importante entender a fondo tanto cómo funcionan como cuáles son sus limitaciones.

Arranquemos con dos variables

La regresión lineal se ocupa de investigar la relación entre dos o más variables continuas.

Comenzaremos tratando de describir el vínculo entre dos variables aleatorias continuas. Medimos ambas variables en la misma unidad: puede tratarse de un individuo, un país, un animal, una escuela, etc.

Ejemplo Se miden en el año 2015, para 187 países:

- Y : Expectativa de vida: El número promedio de años que un niño recién nacido espera vivir, si los patrones de mortalidad no cambiaran (`life`)
- X : Mortalidad 0 a 5: El número de niños de 0 a 5 años que mueren en un año, por cada 1000 niños vivos (`child`)

Ejemplo

Archivo: `esperanza2015.txt`

Los siguientes datos fueron curados por Gapminder, `https://www.gapminder.org` Primeros siete datos del archivo

	country	child	life
1	Afghanistan	73.20	57.90
2	Albania	14.00	77.60
3	Algeria	25.50	77.30
4	Andorra	2.80	82.50
5	Angola	86.50	64.00
6	Antigua and Barbuda	8.70	77.20
7	Argentina	11.60	76.50

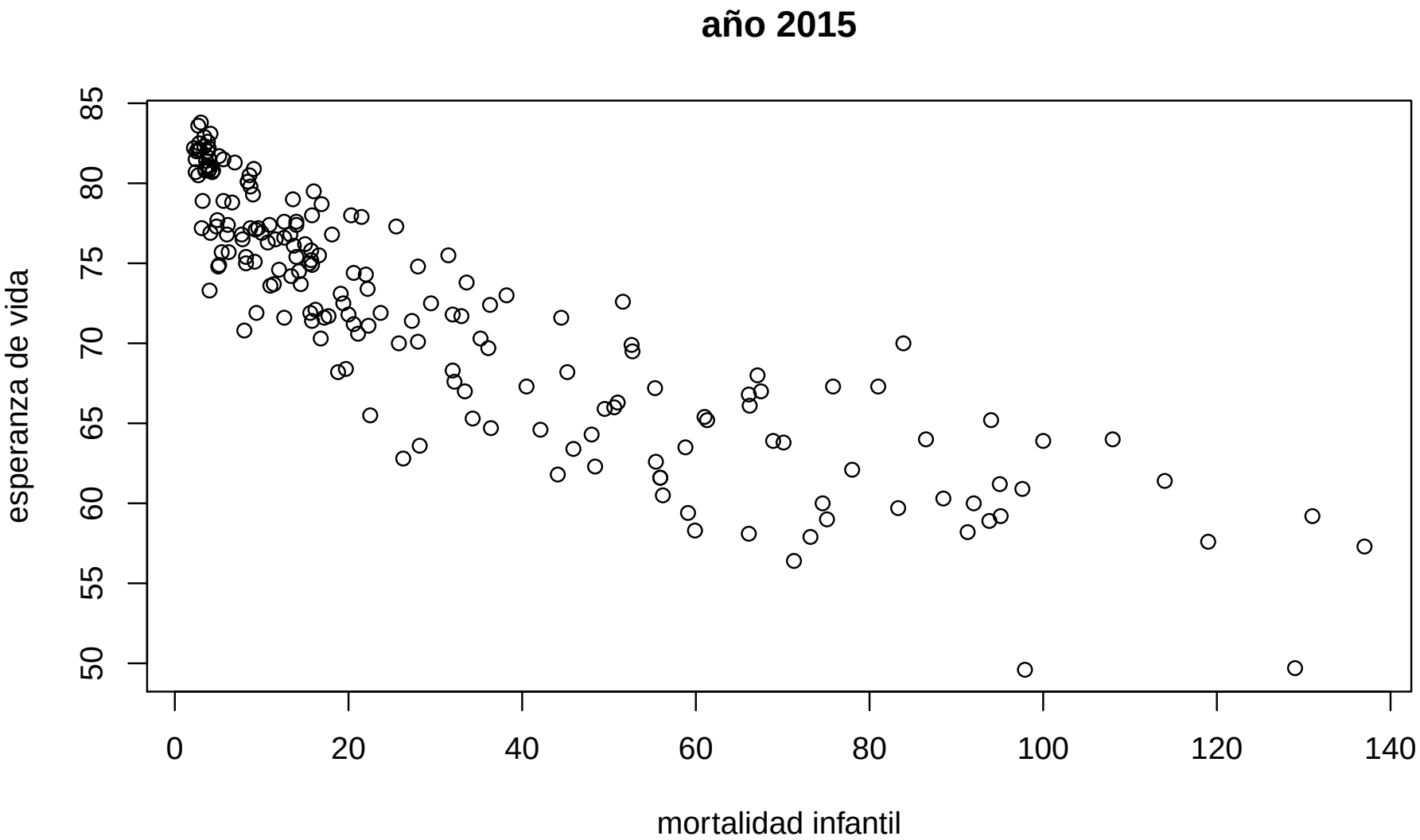
La notación matemática para las observaciones será (X_i, Y_i), donde

$X_i =$ `child` del país i-ésimo

$Y_i =$ `life` del país i-ésimo, $1 \leq i \leq n = 187$

Gráficos de dispersión (o scatter plots)

¿Cómo los visualizamos?



Gráficos de dispersión (o scatter plots)

En un scatter plot se ubican los resultados de una variable (X) en el eje horizontal y los de la otra variable (Y) en el eje vertical. Cada punto en el gráfico representa una observación (X_i, Y_i).

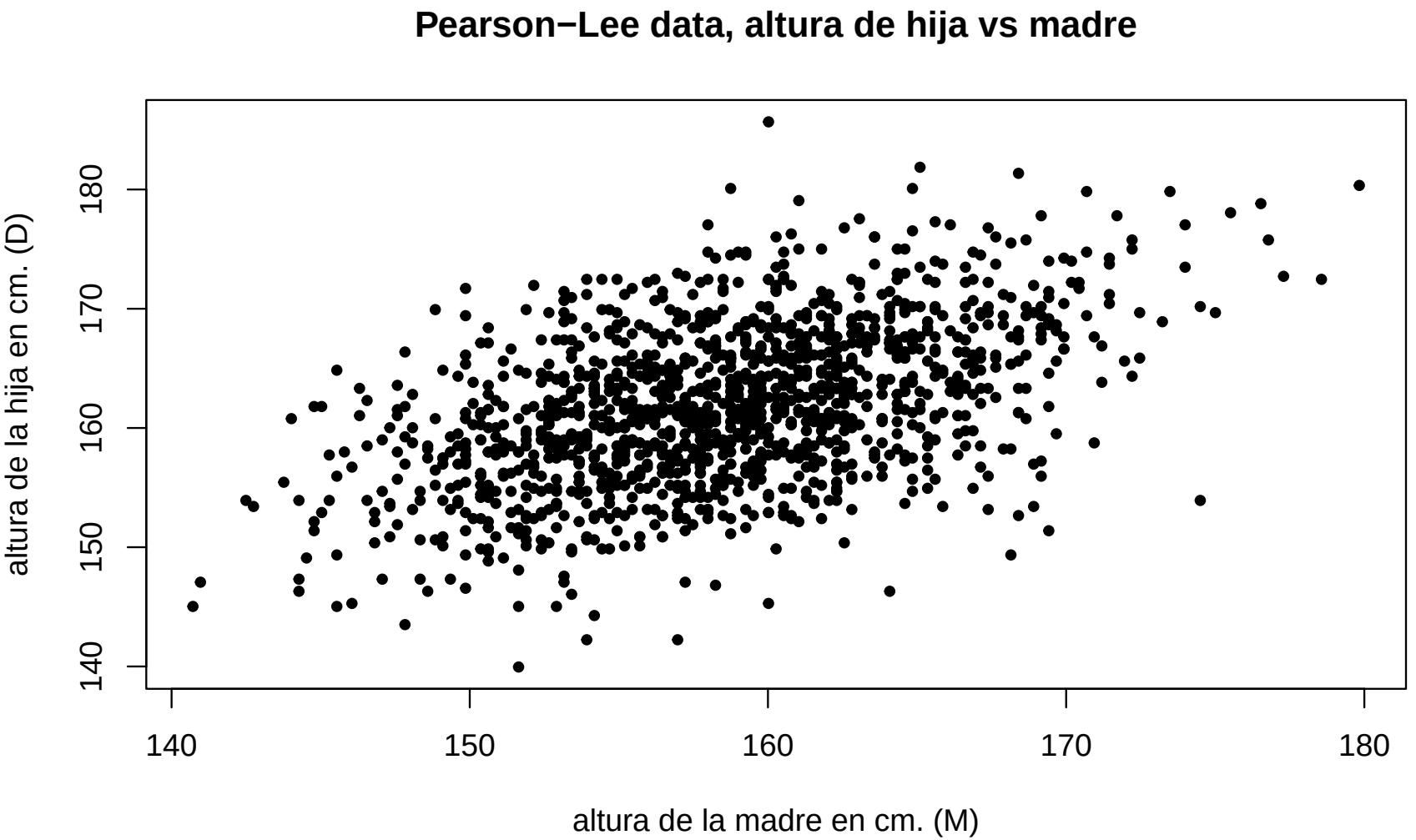
Se pierde la información del individuo (país) Con este gráfico podemos determinar si existe algún tipo de relación entre X e Y.

En este caso vemos que a medida que aumenta la mortalidad infantil, decrece la esperanza de vida. Queremos modelar la relación entre ambas variables. El objetivo es tratar de explicar la esperanza de vida a partir de la mortalidad infantil.

Otro ejemplo: Pearson-Lee data

Karl Pearson organizó la recolección de datos de 1100 familias en Inglaterra en el período 1893 a 1898. Este conjunto de datos en particular: `Heights` en el paquete `alr4` de R da la altura de madres e hijas (en pulgadas), con hasta dos hijas por madre. Todas las hijas tienen 18 años o más, y todas las madres son menores de 65 años. En la fuente los datos aparecen redondeados a la pulgada más cercana. En la librería se les agrega un error de redondeo para que el gráfico no sea discreto. Mostramos datos en cm.

	$X =$ altura madre	$Y =$ altura hija
1	151.64	139.95
2	147.83	143.51
3	153.92	142.24
4	154.18	144.27
5	156.97	142.24
6	140.97	147.07



¿Cuál es la unidad experimental acá? Queremos predecir la altura de la hija a partir de la altura de la madre. ¿Podremos? Vemos que a medida que aumenta X, Y también aumenta. Las madres más altas suelen tener hijas más altas. ¿Siempre?

Ejemplo con más variables

Para $n = 193$ países, medimos en 2015 las siguientes variables:

Y = life es la esperanza de vida al nacer (en años)

X_1 = income: Producto Bruto Interno, per cápita (en USD)

X_2 = child Tasa de Mortalidad de 0 a 5 años, por cada mil niños nacidos vivos en el año.

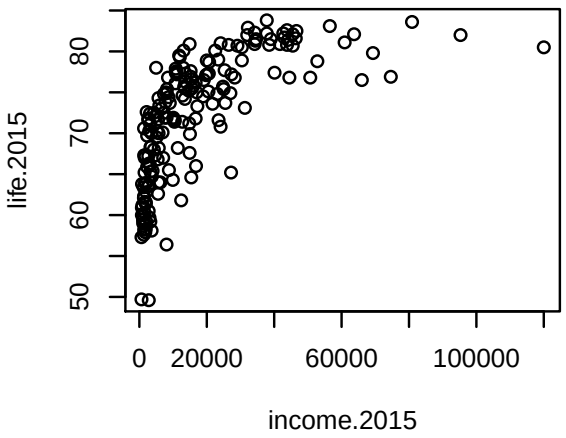
X_3 = dtp3: porcentaje de niños de un año inmunizados con tres dosis de vacuna contra la difteria, tétanos y pertussi (DTP3)

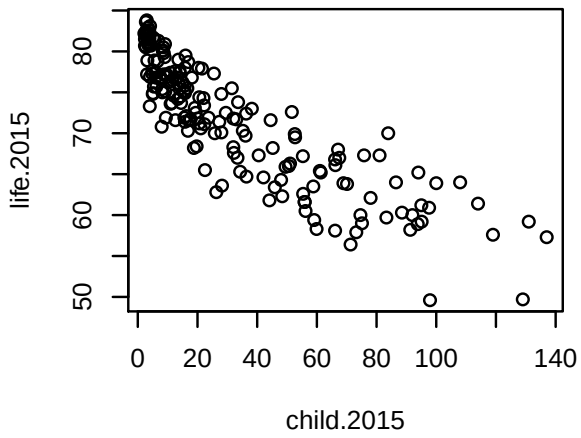
X_4 = school: número de años de escolaridad promedio en hombres de 25–34 años.

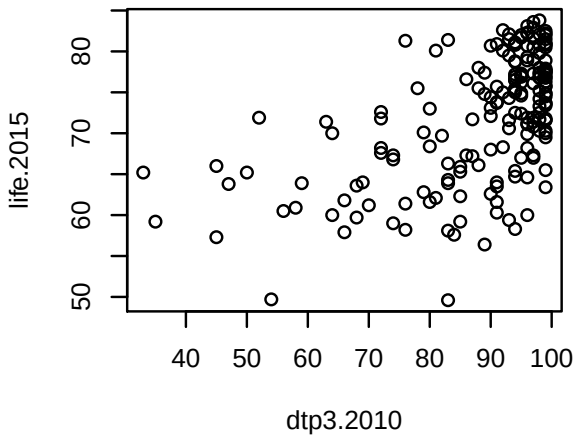
X_5 = status: grado de desarrollo del país ("developed" o "not.developed")

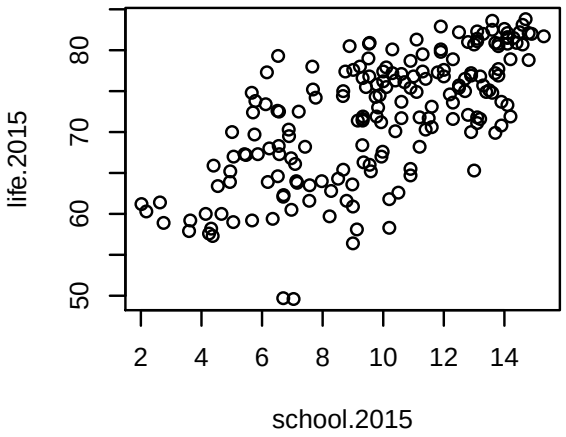
El objetivo es explicar a Y ¿Cómo lo visualizamos?

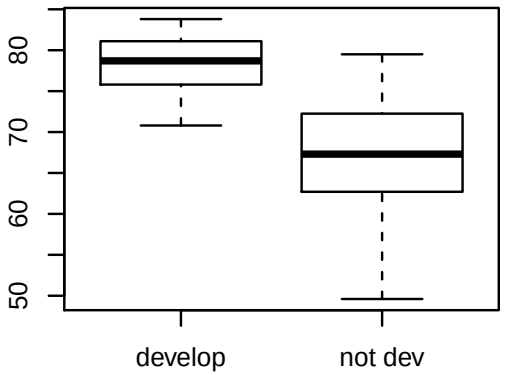
Scatterplot de Y versus cada explicativa











Correlación de Pearson

- La correlación (poblacional) de un vector aleatorio (X, Y):

$$\rho(X, Y) = \frac{E((X - E(X)(Y - E(Y)))}{\sqrt{V(X)V(Y)}}$$

- La correlación muestral:

$$\hat{\rho}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})/(n-1)}{\sqrt{\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

$$\hat{\rho}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

(copiamos) El coeficiente de correlación muestral, $\hat{\rho}(X, Y)$ ó r

$$\hat{\rho}(X, Y) = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}) (Y_i - \overline{Y})}{S_X \cdot S_Y}.$$

Al numerador, se lo denomina covarianza muestral entre X e Y,

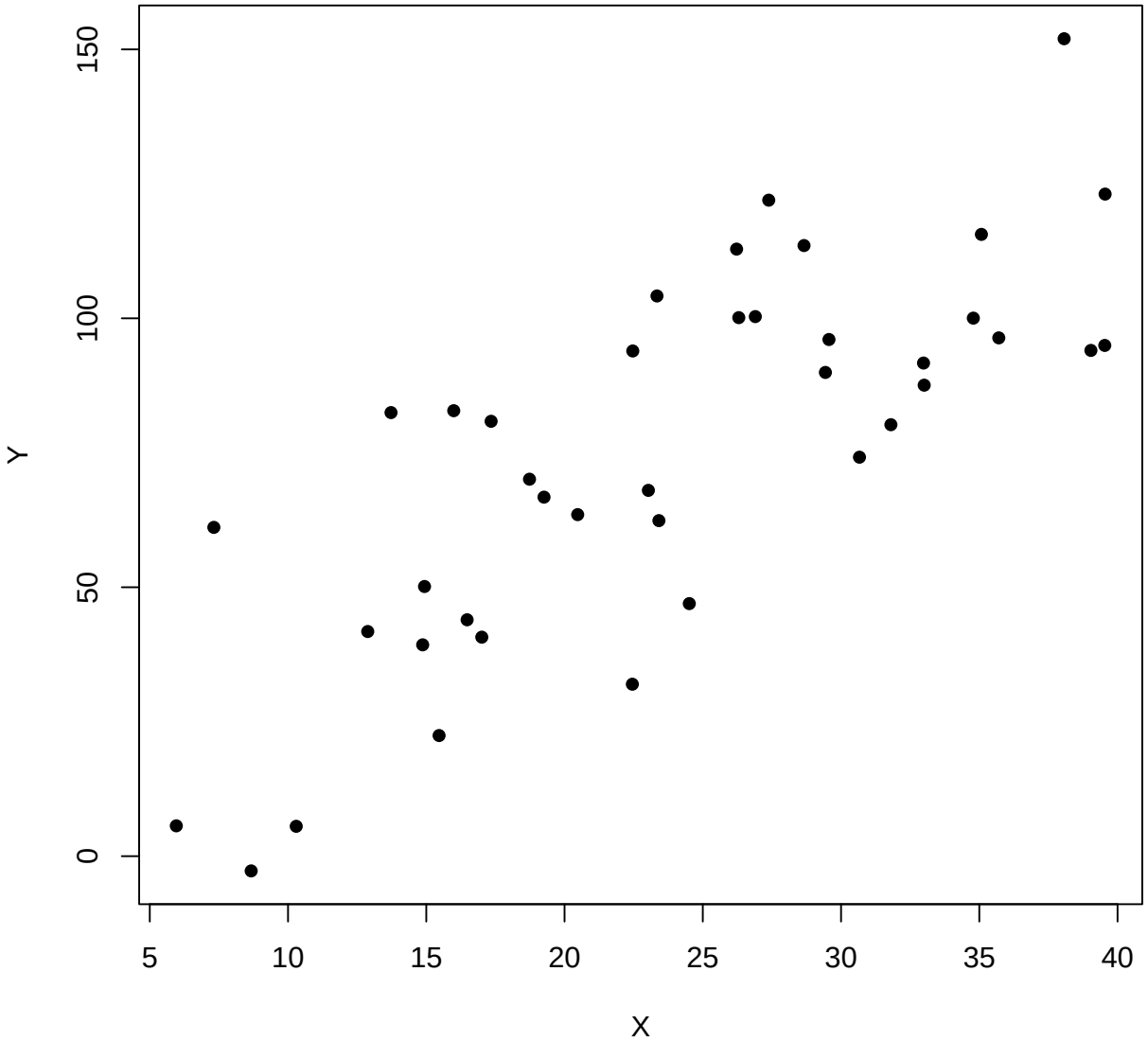
$$\text{covarianza muestral } = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}) (Y_i - \overline{Y})$$

y el denominador es el producto de los desvíos muestrales de cada muestra por separado

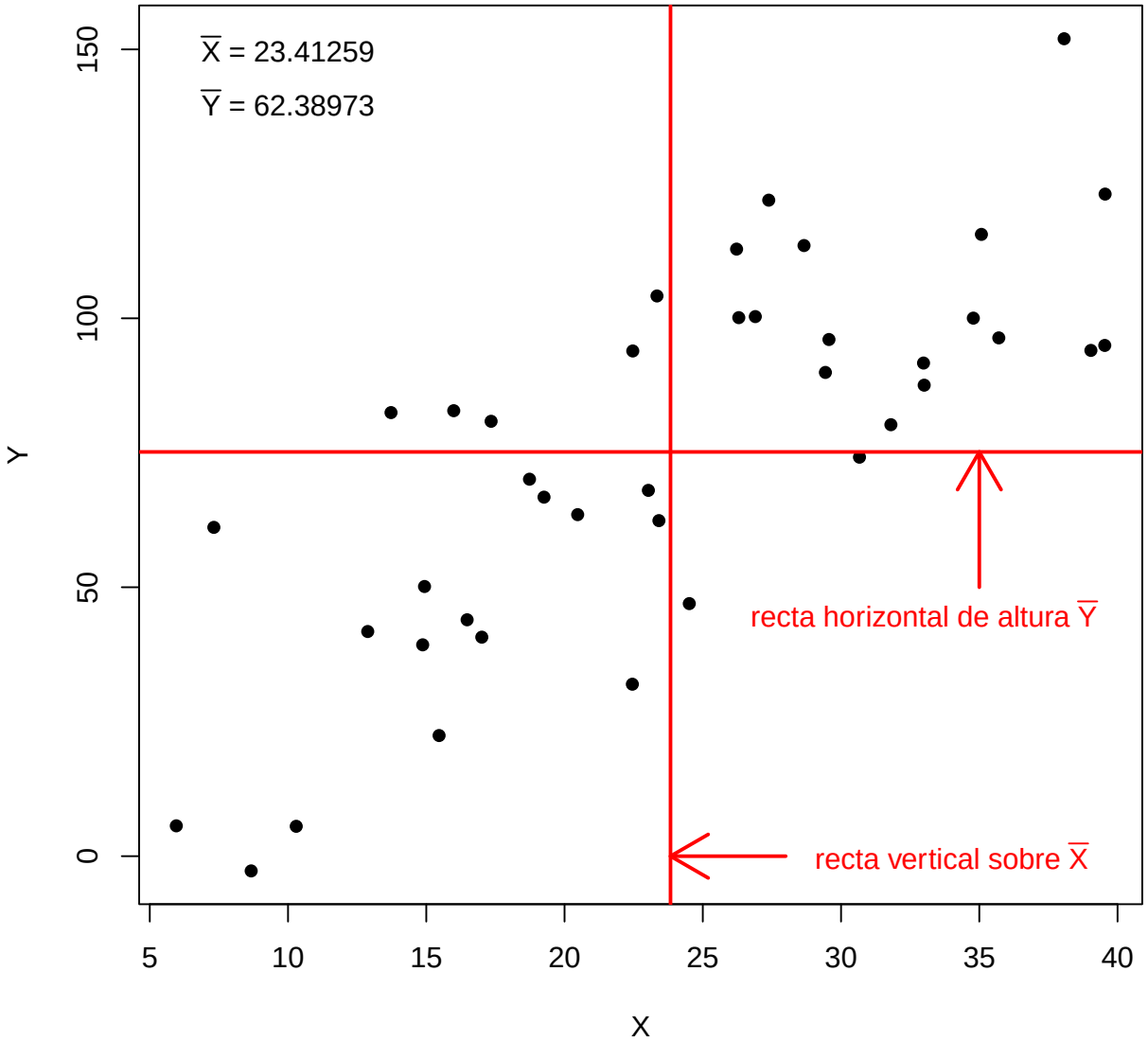
$$S_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (X_i - \overline{X})^2}, \; S_Y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

El numerador $\sum_{i=1}^{n} (X_i - \overline{X}) (Y_i - \overline{Y})$ puede ser positivo o negativo, pero el denominador $\sqrt{\left[\sum_{i=1}^{n} (X_i - \overline{X})^2\right] \left[\sum_{i=1}^{n} (Y_i - \overline{Y})^2\right]}$ siempre es positivo. Luego el signo de $\hat{\rho}(X, Y)$ está determinado por el del numerador. Veamos de qué depende.

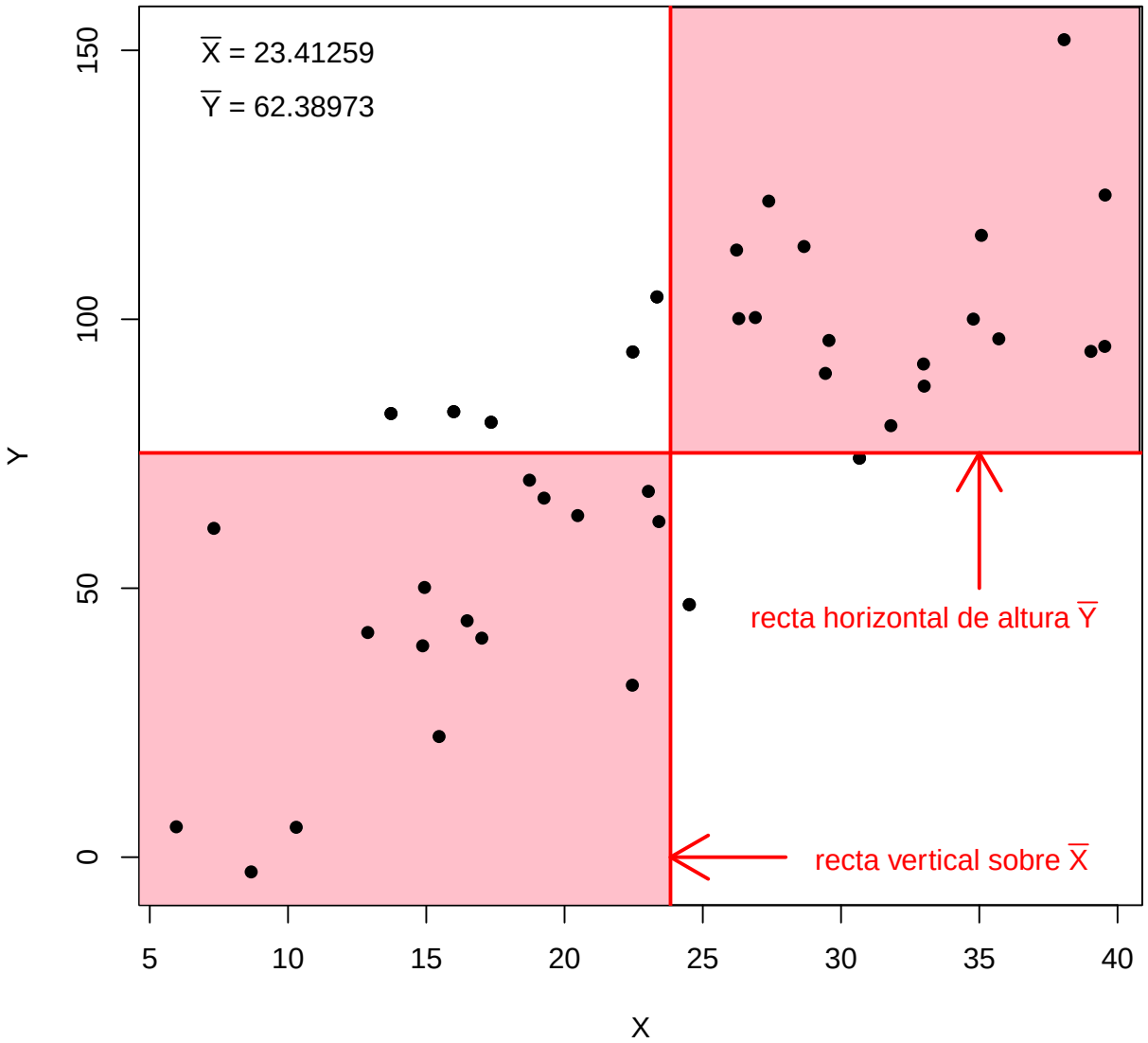
Interpretación de la Correlación



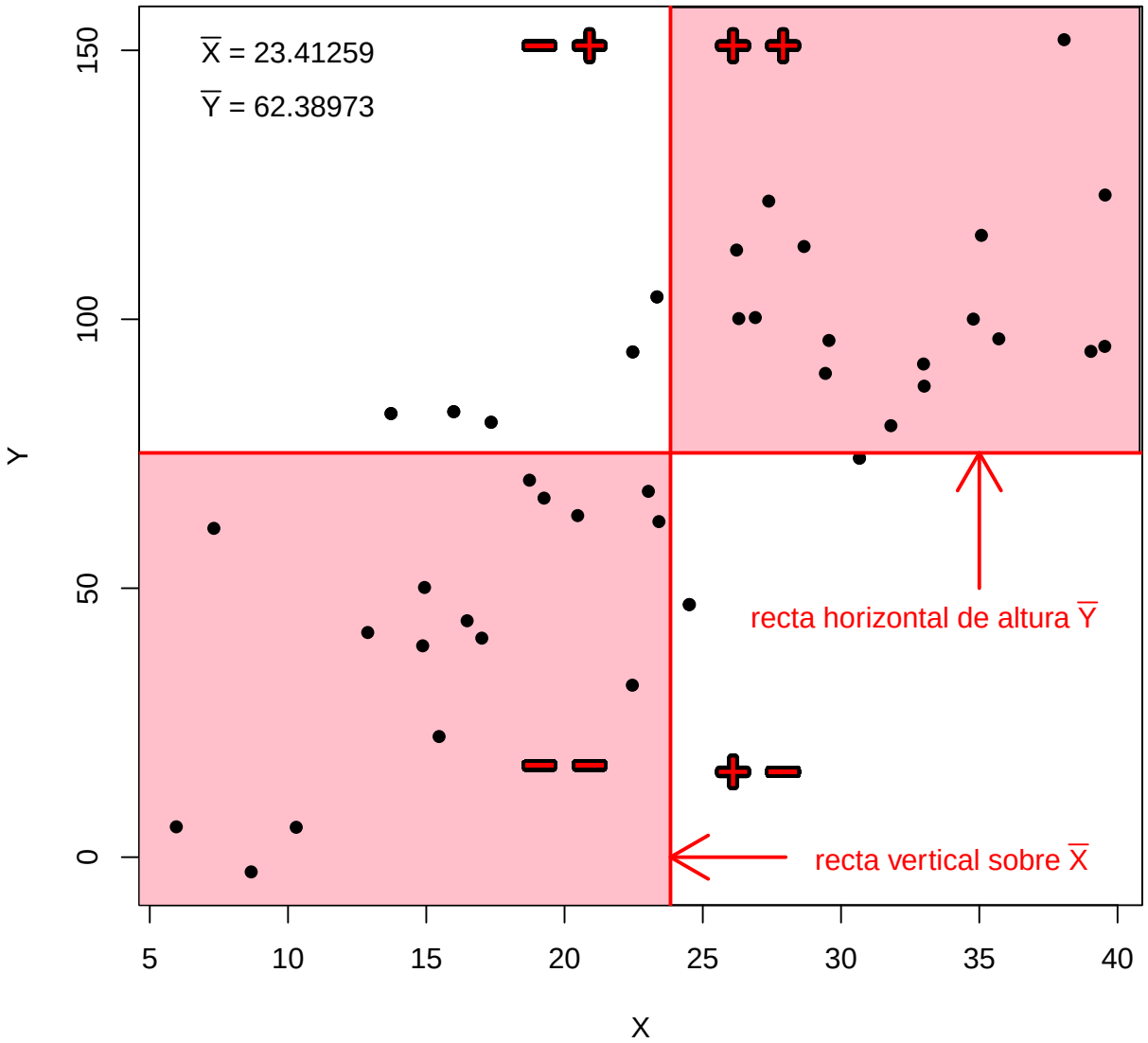
Interpretación de la Correlación



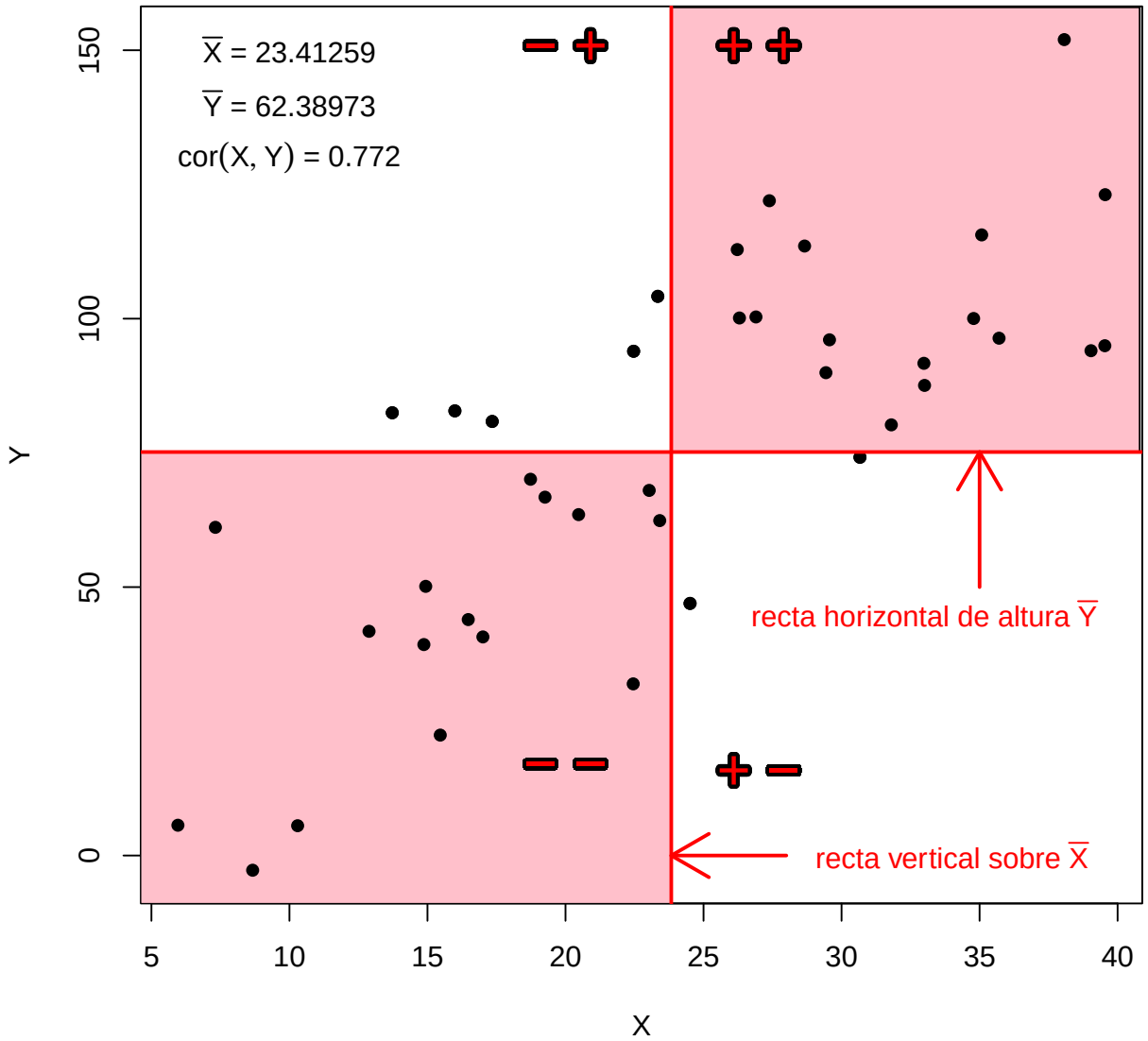
Interpretación de la Correlación



Interpretación de la Correlación



Interpretación de la Correlación



Propiedades del coeficiente de correlación muestral, $\hat{\rho}$ ó r

y también de ρ

1. $-1 \leq r \leq 1$.
2. El valor absoluto de r, $|r|$ mide la fuerza de la asociación lineal entre X e Y, a mayor valor absoluto, hay una asociación lineal más fuerte entre X e Y.
3. El caso particular $r = 0$ indica que no hay asociación lineal entre X e Y.
4. El caso $r = 1$ indica asociación lineal perfecta. O sea que los puntos están ubicados sobre una recta de pendiente (o inclinación) positiva.
5. En el caso $r = -1$ tenemos a los puntos ubicados sobre una recta de pendiente negativa (o sea, decreciente).

6. El signo de r indica que hay asociación positiva entre las variables (si $r > 0$); o asociación negativa entre ellas (si $r < 0$).

7. $r = 0{,}90$ indica que los puntos están ubicados muy cerca de una recta creciente, $r = 0{,}80$ indica que los puntos están cerca, pero no tanto, de una recta creciente.

8. r no depende de las unidades en que son medidas las variables (milímetros, centímetros, metros o kilómetros, por ejemplo) .

9. Los roles de X e Y son simétricos para el cálculo de r.

10. **Cuidado**: el coeficiente de correlación de Pearson es muy sensible a observaciones atípicas. Hay que hacer **siempre** un scatter plot de los datos antes de resumirlos con r.

Ejemplo de correlación: temperatura de ardillas

La temperatura corporal de mamíferos y pájaros tiende a fluctuar durante el día según un ritmo circadiano regular. En un estudio [1] se registra la temperatura corporal de 10 ardillas antílopes cada 6 minutos a lo largo de 10 días consecutivos en condiciones de laboratorio. Elegimos una ardilla y promediamos las temperaturas de los 10 días para obtener un conjunto de datos de 24×10 observaciones. Los autores trataban de contestar a la pregunta:

¿Hay una asociación entre la hora del día y la temperatura corporal?

Para contestarla, tenemos dos estrategias.

- Calcular la correlación entre la hora del día y la temperatura corporal de la ardilla
- Graficar ambas variables: horario y temperatura en un scatter plot

Los primeros 5 datos de la ardilla 6: $(X_i, Y_i)_{i=1,\ldots,240}$

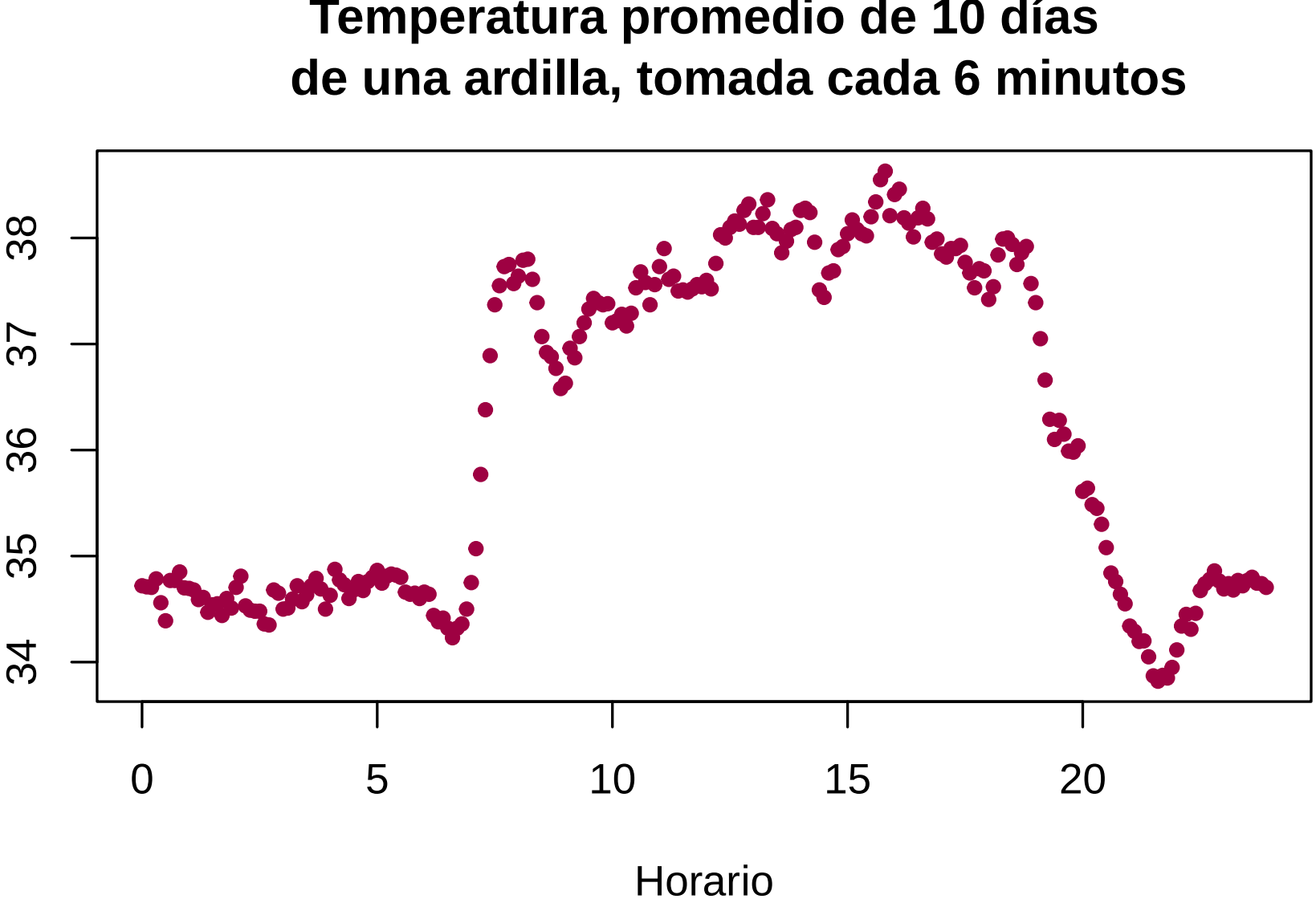
```
> head(ardillas)
horario   temperatura6
1      0.0 34.39
2      0.1 34.42
3      0.2 34.45
4      0.3 34.45
5      0.4 34.43
```

X_i = horario de la iésima medición
Y_i = temperatura promedio a lo largo de 10 días de las 10 mediciones realizadas en el horario X_i

Calculamos la correlación muestral:

```
> cor(horario, temperatura6)
[1] -0.05863851
```

Parece no haber relación entre ambas. ¿Eso mide la correlación? Casi no hay relación lineal entre ambas variables.



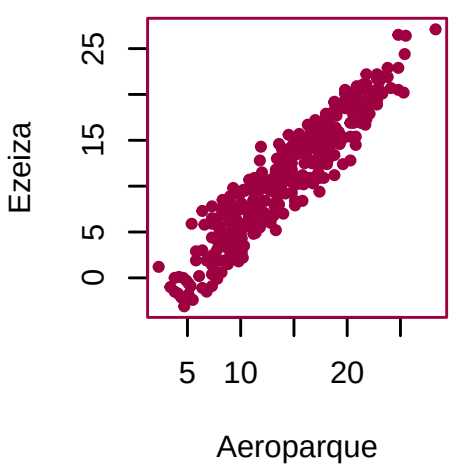
Una correlación cercana a cero no significa (necesariamente) que las dos variables no están asociadas: la correlación mide sólo la **fuerza de una relación lineal**

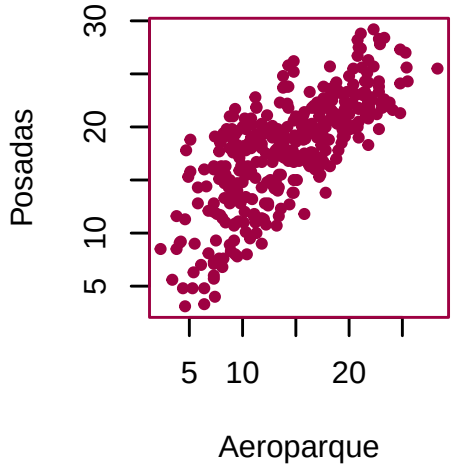
Más ejemplos de correlaciones

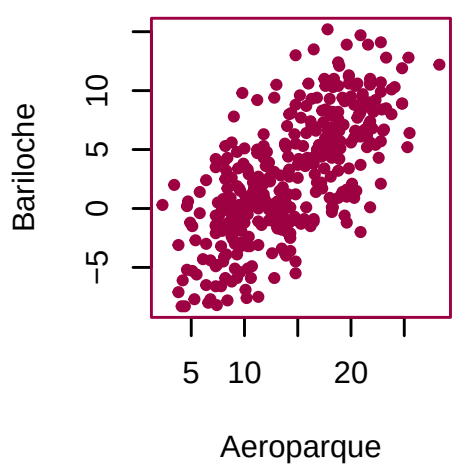
En el sitio del Servicio Meteorológico Nacional pueden bajarse los datos de temperaturas máximas y mínimas diarias de los distintos observatorios ubicados en el país. [2] Elegimos 5 localidades, queremos ver cómo se relacionan entre sí las temperaturas mínimas del mismo día. Así tenemos un vector aleatorio $(A_i, E_i, B_i, P_i, U_i)$, con $1 \leq i \leq n = 365$

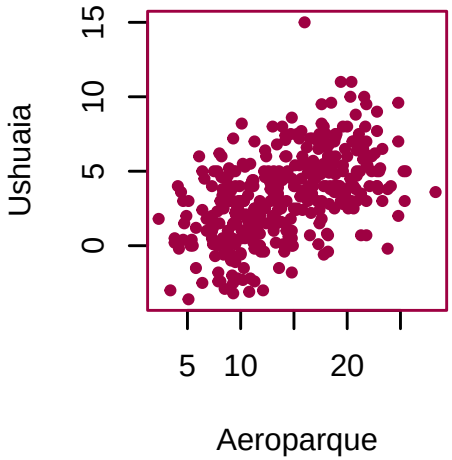
- A_i = temperatura mínima del día i en Aeroparque
- E_i = temperatura mínima del día i en Ezeiza
- B_i = temperatura mínima del día i en Bariloche
- P_i = temperatura mínima del día i en Posadas
- U_i = temperatura mínima del día i en Ushuaia

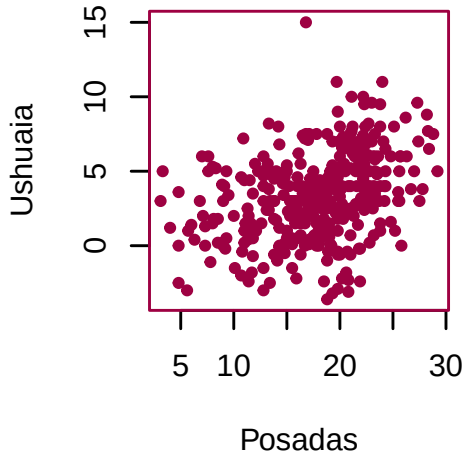
Gráficos de temperaturas mínimas



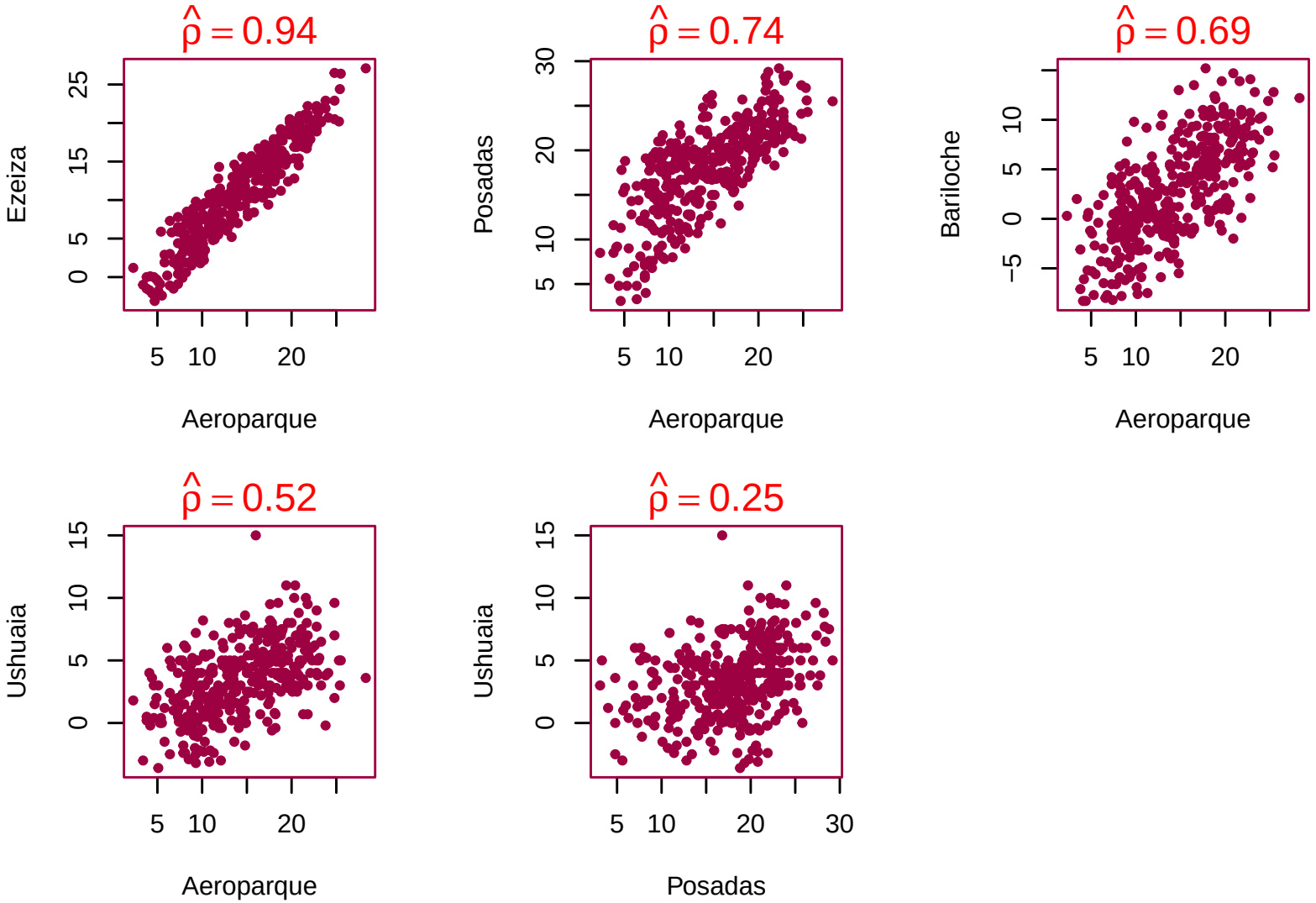




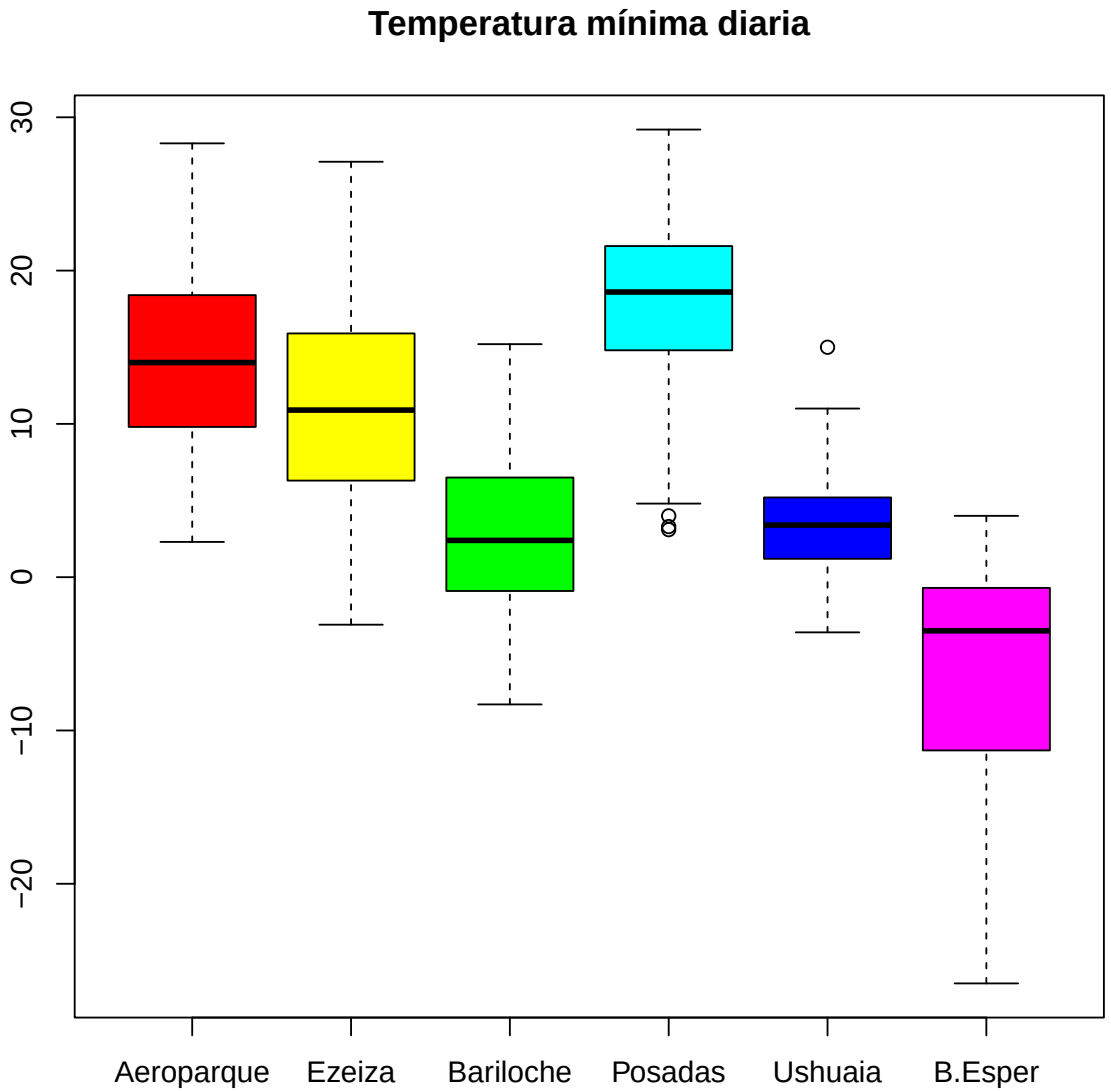




Gráficos de temperaturas mínimas, con correlaciones



Boxplot de temperaturas mínimas, por localidad



¿Por qué estudiar modelos lineales?

1. La teoría de modelos lineales es un caso especial de la teoría más general que cubre modelos más flexibles y realistas. Precisamente porque es un caso tan especial, permite muchos atajos simplificadores, que pueden facilitan el aprendizaje, especialmente sin matemáticas avanzadas.
2. Debido a que los modelos lineales son tan simples, han sido y son tremendamente utilizados. Esto significa que muchas aplicaciones de la estadística se ha realizado sobre modelos lineales. También significa que muchos de los consumidores de estadística esperan modelos lineales o compararán los modelos obtenidos con modelos lineales. Por tanto, es importante entender a fondo tanto cómo funcionan como cuáles son sus limitaciones.

Modelo lineal simple: caso dos variables

El modelo de regresión lineal es un modelo para el vínculo de dos variables aleatorias que denominaremos $X =$ *variable predictora o covariable* e $Y =$ *variable dependiente o de respuesta*. El modelo lineal (simple pues sólo vincula una variable predictora con Y) asume que

1. La distribución de X no está especificada, incluso puede ser determinística.
2. Proponemos el siguiente modelo para las variables:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1}$$

   donde ε es el término del error.
3. Asumimos que la variable aleatoria error ε tiene esperanza 0, varianza constante desconocida que llamaremos σ^2, no está correlacionado con X y no está correlacionado con los errores de otras observaciones.

En el modelo (2) los números β_0 y β_1 son constantes desconocidas que se denominan *parámetros* del modelo, o *coeficientes* de la ecuación.
Los parámetros se denominan

$$\beta_0 = \text{ordenada al origen}$$
$$\beta_1 = \text{pendiente.}$$

El supuesto de la relación funcional entre X e Y sea lineal es no trivial, ya dijimos que muchas variables no lo cumplen. El requisito de que el error tenga varianza constante, se lo suele llamar *homoscedasticidad*, y tampoco es no trivial. Lo mismo pasa con las no correlaciones. Pero el supuesto de que los errores tengan esperanza cero sí es trivial.

Verificar que los supuestos se cumplan para un conjunto de datos será uno de los objetivos que atacaremos más adelante en la materia.

Modelo lineal simple: caso dos variables

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{2}$$

En el caso que mejor se entiende, cuando pedimos ε independiente de X y con $E(\varepsilon) = 0$, el modelo de regresión lineal se puede escribir en términos de la esperanza condicional.

$$E[Y \mid X] = \beta_0 + \beta_1 X \tag{3}$$

Si agregamos el supuesto de que $\text{Var}(\varepsilon) = \sigma^2$, este se traduce en

$$\text{Var}[Y \mid X] = \sigma^2$$

Modelo lineal simple: estimación por mínimos cuadrados

El *error cuadrático medio muestral*, o basado en la muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$, o ECM de entrenamiento, está dado por

$$\widehat{ECM}(b_0, b_1) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Entonces, parece razonable tratar de minimizar el ECM muestral¿Qué obtenemos? Obtendremos los Estimadores de Mínimos Cuadrados (*ordinary least squares, OLS*)

Modelo lineal simple: estimación por mínimos cuadrados

Comencemos derivando con respecto a b_0 y b_1.

$$\frac{\partial \widehat{ECM}}{\partial b_0} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)(-2)$$

$$\frac{\partial \widehat{ECM}}{\partial b_1} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)(-2X_i)$$

Igualemos a cero en el óptimo $\left(\hat{\beta}_0, \hat{\beta}_1\right)$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right) = 0 \tag{4}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)(X_i) = 0 \tag{5}$$