Modelos Estadísticos Interpretables
Tópicos de Modelos Interpretables
1. Modelo Lineal Simple. Correlación

María Eugenia Szretter Noste

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

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.

Mejor predictor constante

Nos interesa estudiar una variable aleatoria Y. Pensemos en la altura que tendrá a los 18 años una niña recién nacida. Supongamos que estamos interesados en *predecir (o adivinar)* dicha altura, a una madre de una beba que la pregunta. ¿Cuál es la mejor predicción que podemos hacer? Necesitamos una medida de cuán buena es nuestra "adivinación". Digamos que nuestra adivinación es "c". La diferencia entre Y y c debe ser chica. Como no queremos que las diferencias positivas se compensen con las negativas, suele tomarse la diferencia al cuadrado: $(Y-c)^2$. Pero esta cantidad es aleatoria. ¿Cómo la resumimos? Con

$$E\left[(Y-c)^2\right]$$

Mejor predictor constante

De hecho, llamamos

Definición 1.1

El **error cuadrático medio de** c *para predecir a* Y *está dado por*

$$ECM(Y, c) = E\left[(Y - c)^2\right]$$

En inglés, se llama mean squared error of c *to predict* Y, $MSE(Y, c)$.

¿Cuál es el mejor predictor constante de una variable aleatoria Y?

Proposición 1.1

Sea Y una variable aleatoria con $E[Y^2]$ finita, entonces $\mu = E[Y]$ es la constante que mejor aproxima a Y en el sentido del ECM, es decir que

$$ECM(Y,\mu) = E\left[(Y-\mu)^2\right] \leq E\left[(Y-c)^2\right] = ECM(Y,c), \quad \forall c \in \mathbb{R}$$

Demostración.

$$\begin{aligned}
E\left[(Y-c)^2\right] &= E\left[(Y-\mu+\mu-c)^2\right] \\
&= E\left[(Y-\mu)^2 + (\mu-c)^2 + 2(Y-\mu)(\mu-c)\right] \\
&= E\left[(Y-\mu)^2\right] + E\left[(\mu-c)^2\right] + E\left[2(Y-\mu)(\mu-c)\right] \\
&= E\left[(Y-\mu)^2\right] + (\mu-c)^2 + 2(\mu-c)\,E\left[Y-\mu\right] \\
&= E\left[(Y-\mu)^2\right] + (\mu-c)^2 \geq E\left[(Y-\mu)^2\right].
\end{aligned}$$

□

Del último renglón de la demostración vemos que

$$E\left[(Y-c)^2\right] = E\left[(Y-\mu)^2\right] + (\mu - c)^2$$
$$= Var(Y) + (E[Y]-c)^2 = h(c)$$

Si queremos hallar el mínimo valor de esta función otro camino es

Ejercicio 1.1

Probar, usando derivadas, que la función h alcanza su mínimo en $c = E[Y]$.

De paso, ¿cuál es el $ECM(Y, E[Y])$? Este valor puede verse como el precio que pagamos por reemplazar a toda la variable aleatoria por una constante, aumenta cuánto más dispersa es dicha variable aleatoria, indicando que cuesta más ("perdemos más" de Y cuando la reemplazamos/resumimos por $E[Y]$ cuánto más dispersa sea Y)

Mejor predictor basado en otra variable aleatoria X

Supongamos que ahora tenemos dos variables aleatorias, digamos X e Y. Conocemos X y queremos utilizar ese conocimiento para mejorar nuestra conjetura sobre Y. Nuestra predicción será, por lo tanto, una función de X, digamos $g(X)$. Nuevamente queremos que

$$E\left[(Y - m(X))^2\right] = ECM[Y, m(X)]$$

sea lo más pequeño posible. ¿Cómo? ¿Cuál función debemos elegir si utilizamos el ECM como criterio a minimizar? Usando lo que ya sabemos de Esperanza condicional.

Mejor predictor basado en otra variable aleatoria X

Proposición 1.2

Sea Y una variable aleatoria con $E[Y^2]$ finita, entonces $m^*(X) = E[Y \mid X]$ es la función de X que mejor aproxima a Y en el sentido del ECM, es decir que $ECM(Y, m^*(X)) \leq ECM(Y, m(X))$

$$E\left[(Y - m^*(X))^2\right] \leq E\left[(Y - m(X))^2\right]$$

para toda variable m(X) con esperanza cuadrada finita.

Demostración.

$$E\left[(Y-m(X))^2\right] = E\left[(Y-m^*(X)+m^*(X)-m(X))^2\right]$$
$$= E\Big[(Y-m^*(X))^2+(m^*(X)-m(X))^2$$
$$+2\,(Y-m^*(X))\,(m^*(X)-m(X))\Big]$$

El primer sumando no depende de $m(X)$. El tercer sumando

$$E\left[\underbrace{(Y-m^*(X))}_{a(X,Y)}\underbrace{(m^*(X)-m(X))}_{b(X)}\right] = E\left[a(X,Y)b(X)\right]$$
$$= E\left[E\left[a(X,Y)b(X)|\,X\right]\right]$$
$$= E\left[b(X)E\left[a(X,Y)|\,X\right]\right]$$

Pero... $E\left[a(X,Y)|\,X\right] = E\left[Y-m^*(X)|\,X\right] = E\left[Y|\,X\right] - E\left[m^*(X)|\,X\right]$
$= m^*(X) - m^*(X) = 0$

Demostración (cont.)

Luego, minimizar el ECM es lo mismo que minimizar el segundo sumando. O sea, buscamos la función $m(X)$ que haga mínimo este término.

$$E\left[(m^*(X) - m(X))^2\right]$$

Respuesta: el mínimo se alcanza en $m^*(X) = E[Y \mid X]$

□

Esta función $m^*(X) = E[Y \mid X]$ se llama la función de regresión de Y en X verdadera o poblacional, y es uno de los objetivos de nuestro estudio. El problema es que necesitamos conocer la distribución conjunta teórica del vector (X, Y) para poder calcularla. Usualmente esta información no está disponible en las aplicaciones.

Entonces, podemos bajar la expectativa y hacer una pregunta menos ambiciosa.

Mejor predictor lineal de Y basado en X

¿Cuál es el mejor predictor de Y que es **función lineal** de X?

buscamos β_0, β_1 que minimicen la función

$$H(\beta_0, \beta_1) = E\Big[\Big(Y - (\beta_0 + \beta_1 X)\Big)^2\Big]$$

Ejercicio 1.2

Hallar los valores de las constantes β_0^* y β_1^* que minimizan el ECM, derivando la función H respecto de cada una de sus variables e igualando la derivada a cero y despejando los valores óptimos.

Observemos que $H(\beta_0, \beta_1) = ECM(Y, (\beta_0 + \beta_1 X))$

$$
\begin{aligned}
H(\beta_0, \beta_1) &= E\Big[\Big(Y - (\beta_0 + \beta_1 X)\Big)^2\Big] \\
&= E\Big[Y^2 + (\beta_0 + \beta_1 X)^2 - 2Y(\beta_0 + \beta_1 X)\Big] \\
&= E\Big[Y^2 + \Big(\beta_0^2 + 2\beta_0\beta_1 X + \beta_1^2 X^2) - 2\beta_0 Y - 2\beta_1 YX\Big] \\
&= E\big[Y^2\big] + \beta_0^2 + 2\beta_0\beta_1 E\big[X\big] + \beta_1^2 E\Big[X^2\Big] - 2\beta_0 E\left[Y\right] - 2\beta_1 E\left[YX\right]
\end{aligned}
$$

Derivamos

$$
\frac{\partial}{\partial \beta_0} H(\beta_0, \beta_1) = 2\beta_0 + 2\beta_1 E\big[X\big] - 2E\left[Y\right] \tag{1}
$$

$$
\frac{\partial}{\partial \beta_1} H(\beta_0, \beta_1) = 2\beta_0 E\big[X\big] + 2\beta_1 E\Big[X^2\Big] - 2E\left[XY\right] \tag{2}
$$

$$\frac{\partial}{\partial \beta_0} H(\beta_0, \beta_1) = 2\beta_0 + 2\beta_1 E[X] - 2E[Y] \quad (1)$$

$$\frac{\partial}{\partial \beta_1} H(\beta_0, \beta_1) = 2\beta_0 E[X] + 2\beta_1 E[X^2] - 2E[XY] \quad (2)$$

Las igualamos a cero: $\begin{cases} (1) \quad \beta_0^* + E(X)\beta_1^* - E(Y) = 0 \\ (2) \quad E(X)\beta_0^* + E(X^2)\beta_1^* - E(XY) = 0 \end{cases}$

Nos queda un sistema de ecuaciones lineales con 2 incógnitas (β_0^* y β_1^*) que podemos resolver.

de (1) $\beta_0^* = E(Y) - \beta_1^* E(X)$ Lo reemplazamos en (2):

$$E(X)\left[E(Y) - E(X)\beta_1^*\right] + E(X^2)\beta_1^* - E(XY) = 0$$

$$\beta_1^* \underbrace{\left[-E(X)^2 + E(X^2)\right]}_{Var(X)} = \underbrace{E(XY) - E(X)E(Y)}_{Cov(X,Y)}$$

Osea $\beta_1^* = \frac{Cov(X,Y)}{Var(X)}$ $\quad \beta_0^* = E(Y) - \frac{Cov(X,Y)}{Var(X)} E(X)$

Óptimo lineal

Luego, el predictor óptimo lineal de Y basado en X es

$$m_{OL}(X) = \frac{cov(X,Y)}{V(X)} X + E(Y) - \frac{cov(X,Y)}{V(X)} E(X)$$

1. El predictor óptimo lineal pasa por el punto $(E(X), E(Y))$.
2. La pendiente óptima es el cociente entre la covarianza de X e Y, y la varianza de X. La pendiente aumenta cuanto más tienden X e Y a fluctuar juntas y se acerca a cero cuanto más varíe X.
3. En ningún momento tuvimos que asumir que la relación entre X e Y realmente es lineal. Obtuvimos la aproximación lineal óptima a la verdadera relación entre ellas, fuera la que fuera.
4. La aproximación dada por el mejor predictor lineal a la verdadera relación entre X e Y puede ser horrible. (Imaginemos que $E[Y \mid X = x] = e^x$, ó $= \sin x$.)

¿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.

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 la distribución de Y condicional a X:
$$Y \mid X = \beta_0 + \beta_1 X + \varepsilon, \tag{3}$$
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 (3) 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.

Estimación en el Modelo lineal simple: enfoque *naive*

¿Cómo estimamos los parámetros a partir de una muestra? Es decir, a partir de conocer la muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$ de observaciones de alturas de madres e hijas, cómo conseguimos estimar β_0, β_1 y σ^2? Una forma es usar resultados teórico y estimadores *plug-in*.

$$\beta_1^* = \frac{cov(X, Y)}{V(X)}$$

$$\beta_0^* = E(Y) - \frac{cov(X, Y)}{V(X)} E(X)$$

Los podríamos estimar por sus versiones muestrales o estimadores puntuales:

$$\widehat{E(X)} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X} \qquad \widehat{V(X)} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

$$\widehat{cov(X, Y)} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) \left(Y_i - \overline{Y}\right)$$

Repasemos estimadores plug-in

Sean $X_1, \ldots, X_n$ variables aleatorias iid con distribución F. Estimamos a F con la distribución empírica.

Definición 1.2

La **función de distribución empírica** $\widehat{F}_n$ *es la función de distribución acumulada que asigna masa o probabilidad* $1/n$ *a cada observación* X_i *de la muestra,*

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, X_i]}(t),$$

donde $I_{(-\infty, X_i]}(t) = \begin{cases} 1 & si \quad X_i \leq t \\ 0 & si \quad X_i > t \end{cases}$ *, es la función indicadora.*

Luego, la distribución que induce $\widehat{F}_n$ es una distribución discreta que asigna la misma probabilidad a cada elemento de la muestra. ¿Cómo se calculan esperanzas con $\widehat{F}_n$?

Estimador plug-in de la esperanza

$$\widehat{E(X)} = E_{\widehat{F}_n}(X) = \sum_{a \in \text{Rango}\left(\widehat{F}_n\right)} a p_{\widehat{F}_n}(a) = \sum_{i=1}^{n} X_i \frac{1}{n} = \overline{X}_n$$

Más generalmente

$$E_{\widehat{F}_n}\left(g(X)\right) = \sum_{a \in \text{Rango}\left(\widehat{F}_n\right)} g(a) p_{\widehat{F}_n}(a) = \sum_{i=1}^{n} g\left(X_i\right) \frac{1}{n}$$

Y por lo tanto

$$V_{\widehat{F}_n}(X) = E_{\widehat{F}_n}\left(\left(X - E_{\widehat{F}_n}(X)\right)^2\right) = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

resulta ser el estimador plug-in de la varianza, es decir, la varianza muestral (dividida por n en vez de $n-1$).

Estimador plug-in para vectores aleatorios

¿Cómo opera con vectores aleatorios? La distribución empírica, en este caso, asigna peso $\frac{1}{n}$ a cada observación $(X_1, Y_1), \ldots, (X_n, Y_n)$. Luego

$$E_{\widehat{F}_n}(g(X,Y)) = \sum_{i=1}^{n} g(X_i, Y_i)\frac{1}{n}$$

En particular, el estimador plug-in $\widehat{cov(X,Y)}$ es

$$\widehat{cov(X,Y)} = E_{\widehat{F}_n}\left(\left[X - E_{\widehat{F}_n}(X)\right]\left[Y - E_{\widehat{F}_n}(Y)\right]\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

la covarianza muestral.

Estimación en el Modelo lineal simple: enfoque *naive*

Entonces:

$$\widehat{\beta_1} = \frac{\widehat{cov(X,Y)}}{\widehat{V(X)}}$$

$$\widehat{\beta_0} = \widehat{E(Y)} - \frac{\widehat{cov(X,Y)}}{\widehat{V(X)}}\widehat{E(X)}$$

O sea,

$$\widehat{\beta_1} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

$$\widehat{\beta_0} = \overline{Y} - \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\overline{X}$$

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

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

Asumimos que las observaciones de nuestra muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$ son independientes, entonces, para cada (b_0, b_1) fijo, la Ley de los Grandes Números nos garantiza que

$$\widehat{ECM}(b_0, b_1) \to E\left[Y - b_0 - b_1 X\right]^2 = ECM(b_0, b_1), \text{ cuando } n \to \infty$$

Entonces, parece razonable tratar de minimizar el ECM muestral, que podemos calcular, como una buena aproximación al ECM verdadero o poblaacional que no podemos calcular. ¿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 $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$: ecuaciones normales p=2 vamos a S^2 insesgado

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

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

Las ecuaciones (4) y (5) se denominan **ecuaciones normales** para mínimos cuadrados.

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1X_i\right)\right)=0$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1X_i\right)\right)(X_i)=0 \quad \Longleftrightarrow$$

$$\frac{1}{n}\sum_{i=1}^{n}Y_i-\frac{1}{n}\sum_{i=1}^{n}\widehat{\beta}_0-\frac{1}{n}\widehat{\beta}_1\sum_{i=1}^{n}X_i=0$$

$$\frac{1}{n}\sum_{i=1}^{n}Y_iX_i-\frac{1}{n}\widehat{\beta}_0\sum_{i=1}^{n}X_i-\widehat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}X_i^2=0 \quad \Longleftrightarrow$$

$$\overline{Y}-\widehat{\beta}_0-\hat{\beta}_1\overline{X}=0$$

$$\overline{XY}-\hat{\beta}_0\overline{X}-\hat{\beta}_1\overline{X^2}=0$$

(copiamos) $\overline{Y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{X} = 0$

$\overline{XY} - \hat{\beta}_0 \overline{X} - \hat{\beta}_1 \overline{X^2} = 0$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\overline{XY} - \left(\overline{Y} + \hat{\beta}_1 \overline{X}\right) \overline{X} - \hat{\beta}_1 \overline{X^2} = 0 \quad \Leftrightarrow \quad -\hat{\beta}_1 \left(\overline{X^2} - \overline{X}^2\right) + \overline{XY} - \overline{X}\,\overline{Y} = 0$$

Observemos que

$$\overline{X^2} - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \widehat{V(X)} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Ejercicio 1.3

Comprobar que $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$

Observemos que

$$\overline{XY} - \overline{X}\,\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$$

$$= \widehat{cov(X,Y)} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

Ejercicio 1.4

Comprobar que

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)$$

Notaremos $c_{XY} = \sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}$

(copiamos) $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

$$-\hat{\beta}_1 \left(\overline{X^2} - \overline{X}^2 \right) + \overline{XY} - \overline{X}\,\overline{Y} = 0 \iff$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\widehat{cov(X,Y)}}{\widehat{var(X)}} \iff$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

¡Son los mismos estimadores que obtuvimos antes!

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:

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

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

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

$$\widehat{\rho}(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{s_X \cdot s_Y}.$$

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

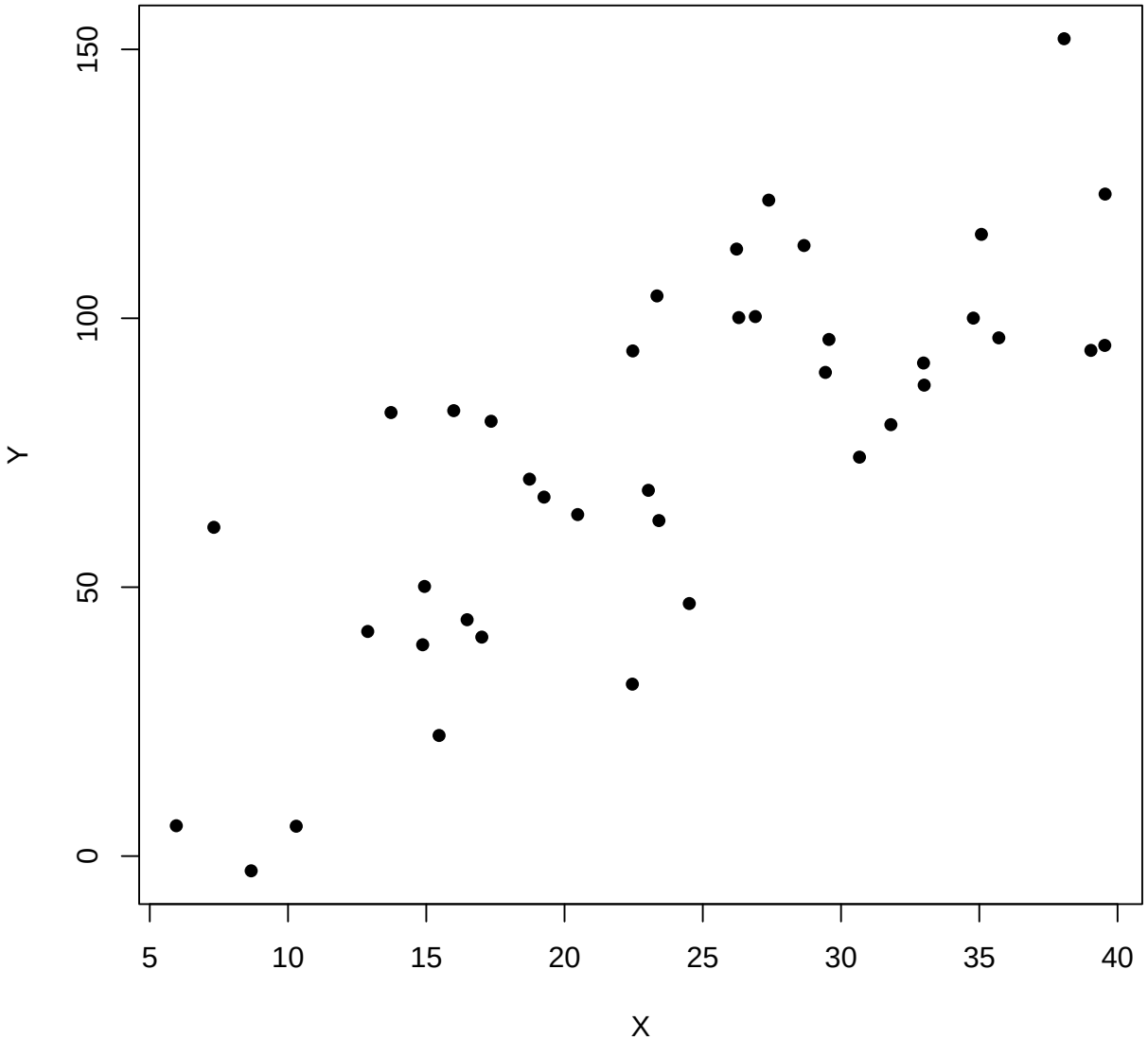
$$\text{covarianza muestral } = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

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

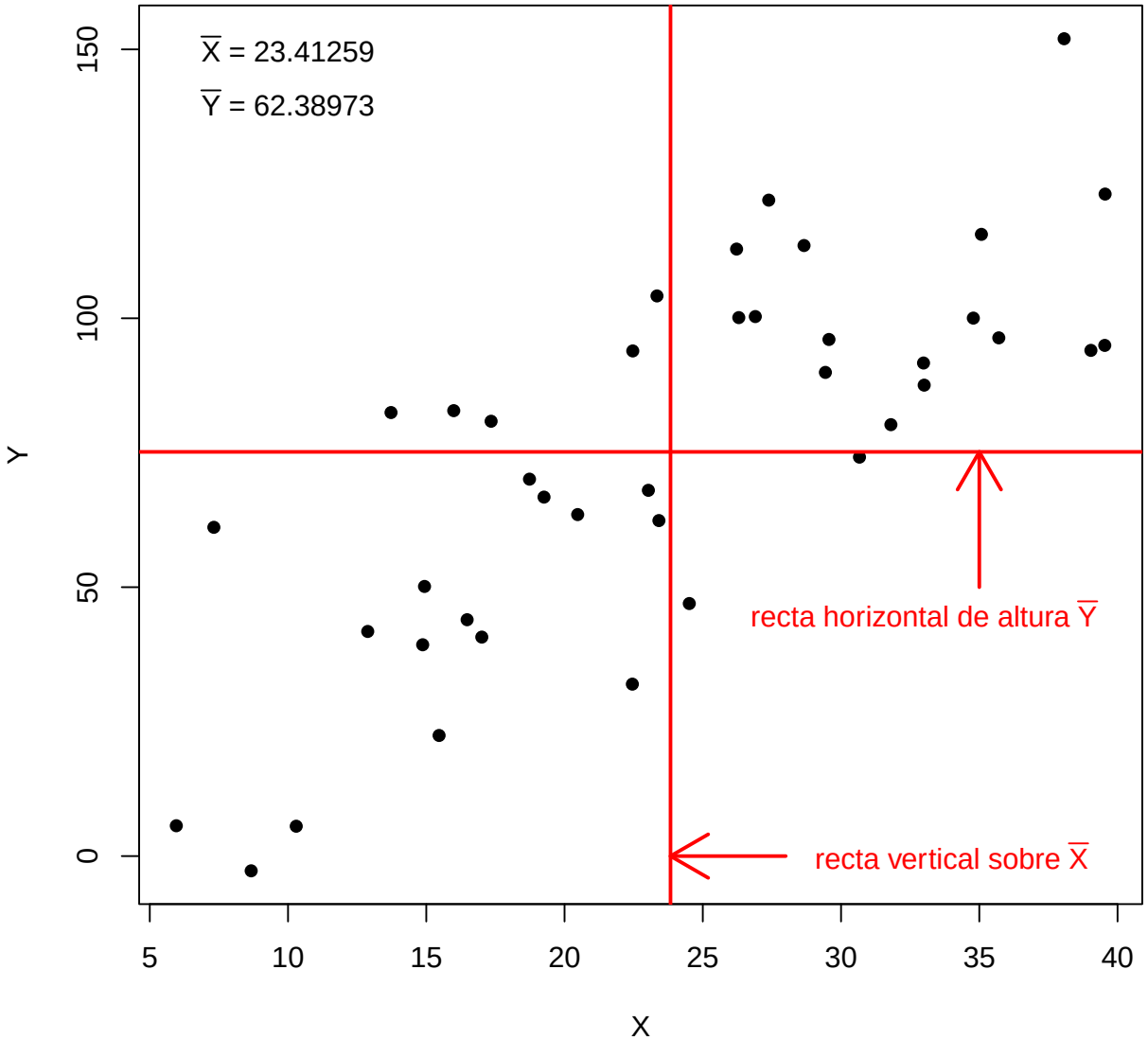
$$s_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2},\ s_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

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

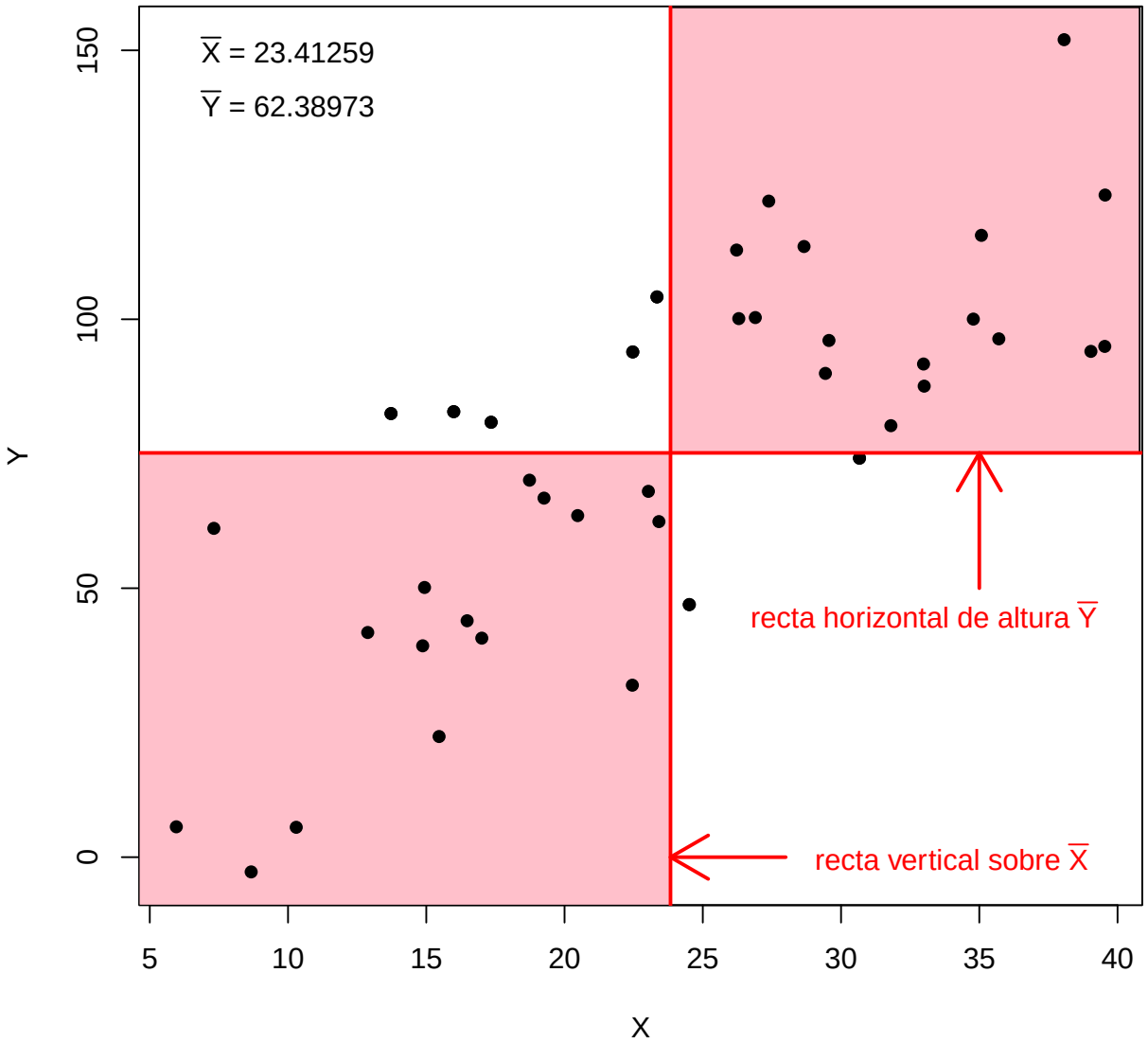
Interpretación de la Correlación



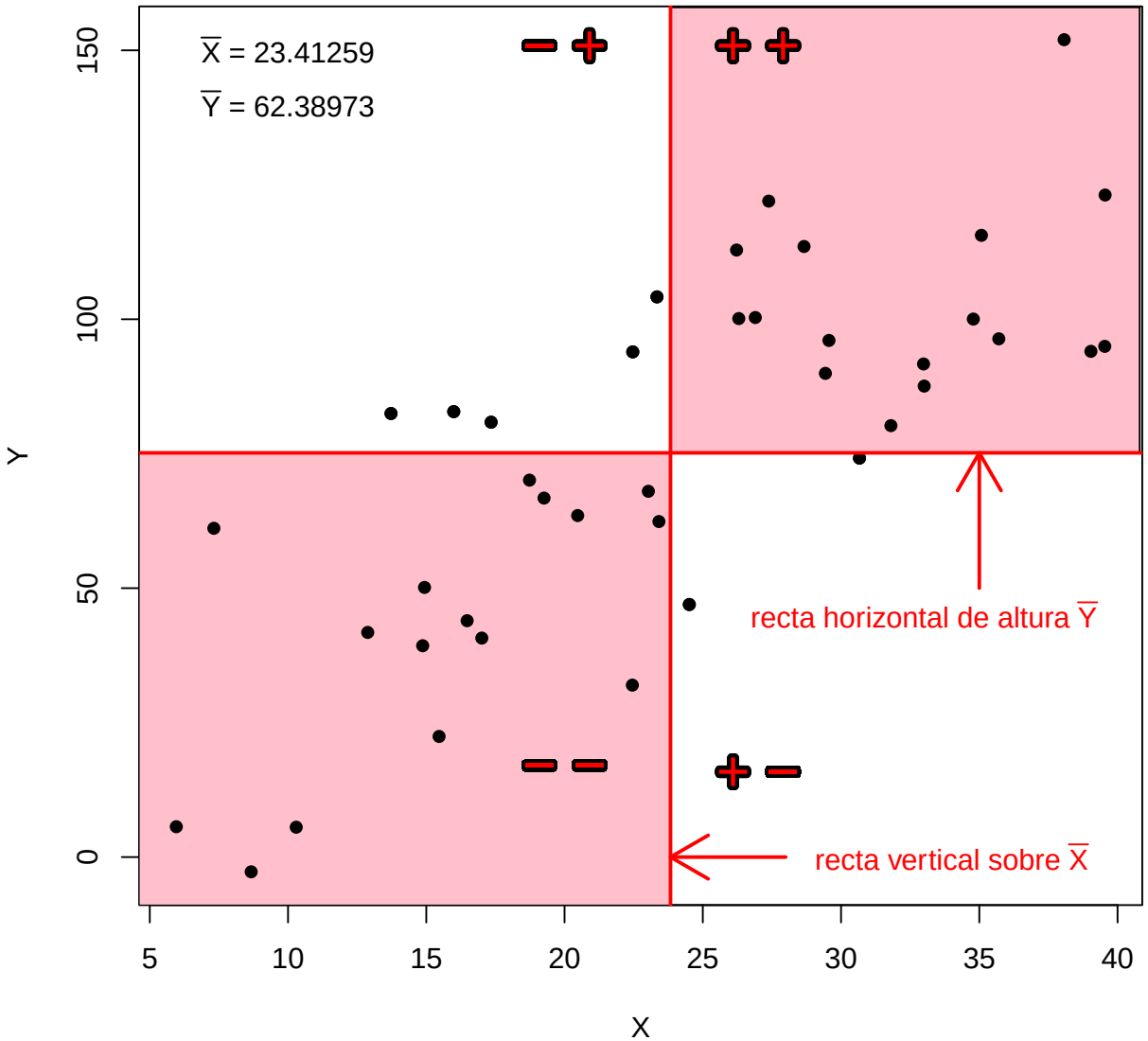
Interpretación de la Correlación



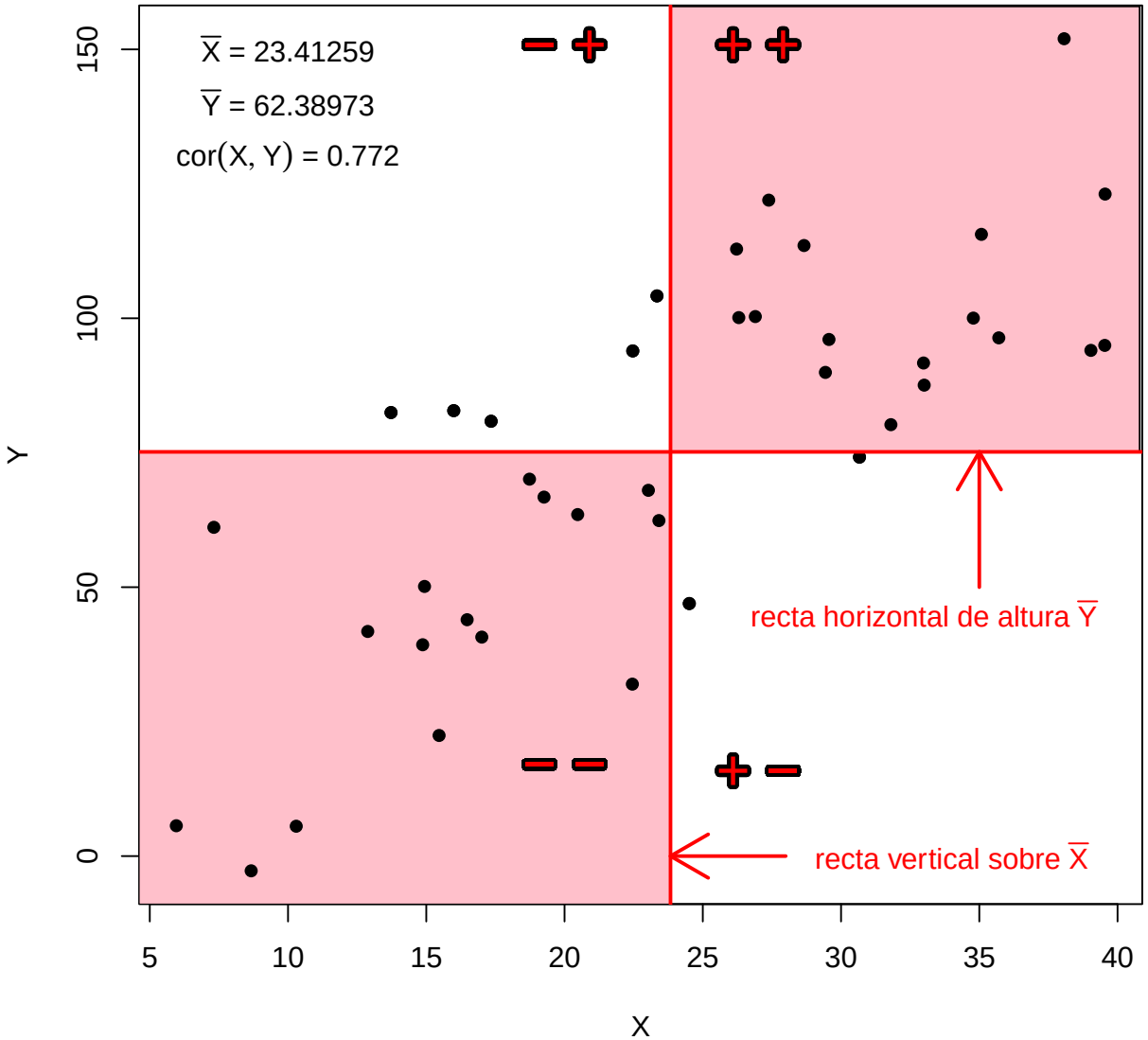
Interpretación de la Correlación



Interpretación de la Correlación



Interpretación de la Correlación



Propiedades del coeficiente de correlación muestral, $\widehat{\rho}$ 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: $(X_i, Y_i)_{i=1,\ldots,240}$

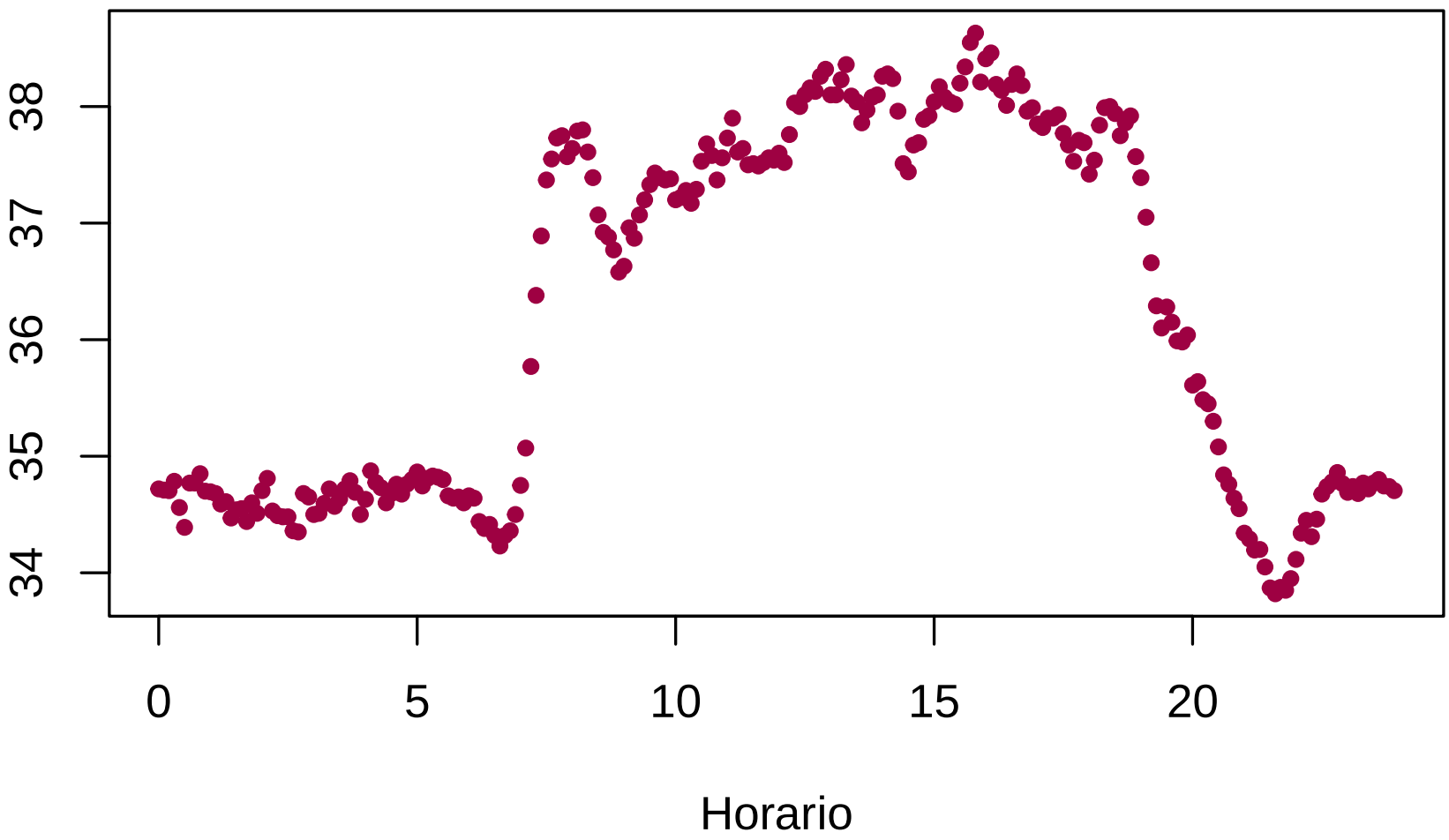
```
> head(ardillas)
horario   temperatura6
1      0.0 34.39
2      0.1 34.42
3      0.2 34.45
4      0.3 34.45
5      0.4 34.43
```

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.



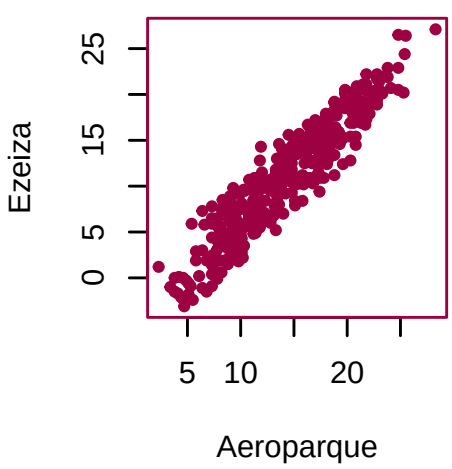
Correlación cercana a cero no significa (necesariamente) que las dos variables no están asociadas: la correlación mide sólo la **fuerza de una relación lineal**

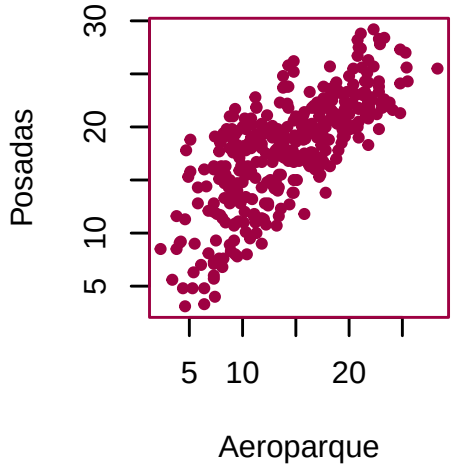
Más ejemplos de correlaciones

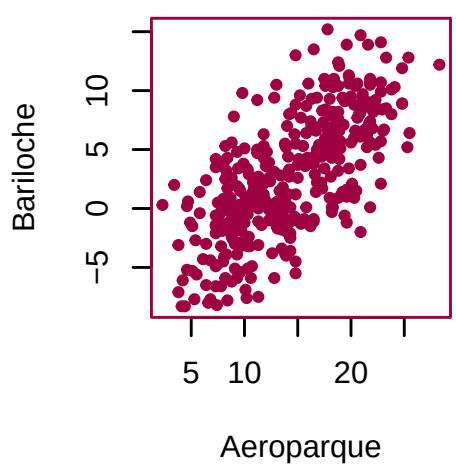
En el sitio del Servicio Meteorológico Nacional pueden bajarse los datos de temperaturas máximas y mínimas diarias de los distintos observatorios ubicados en el país. [2] Elegimos 5 localidades, queremos ver cómo se relacionan entre sí las temperaturas mínimas del mismo día. Así tenemos un vector aleatorio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

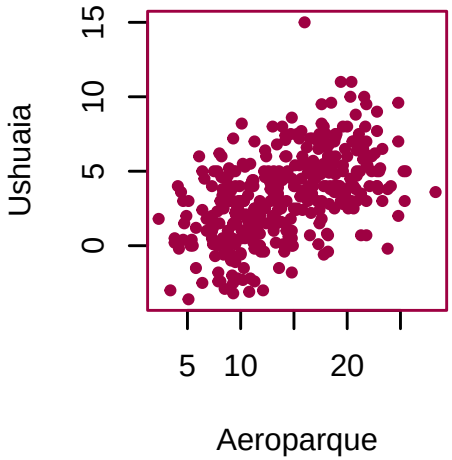
- A_i = temperatura mínima del día i en Aeroparque
- E_i = temperatura mínima del día i en Ezeiza
- B_i = temperatura mínima del día i en Bariloche
- P_i = temperatura mínima del día i en Posadas
- U_i = temperatura mínima del día i en Ushuaia

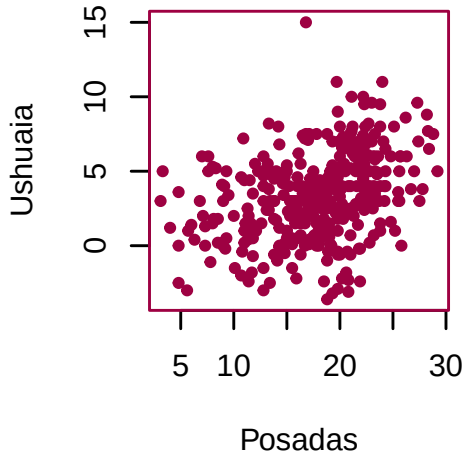
Gráficos de temperaturas mínimas



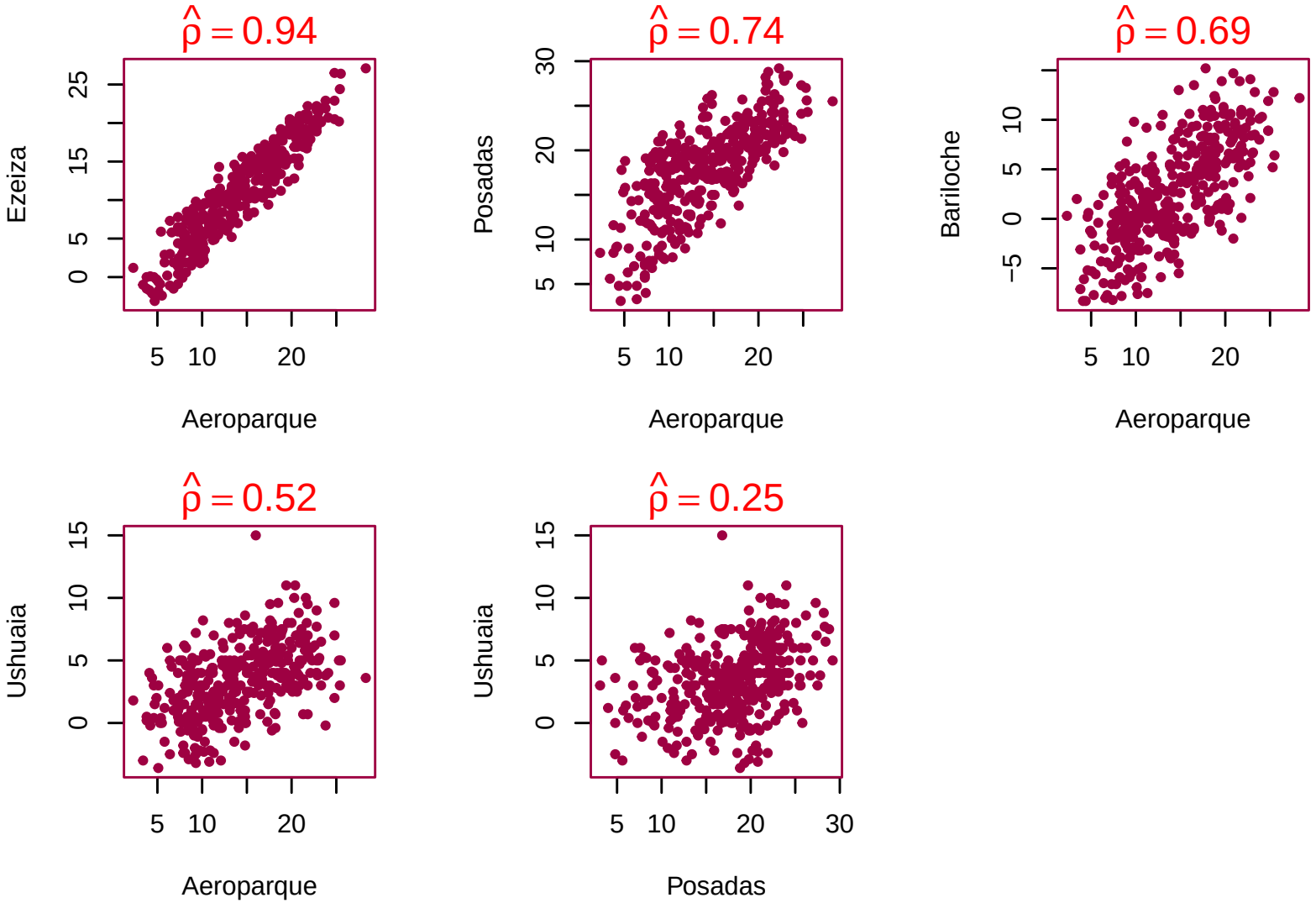




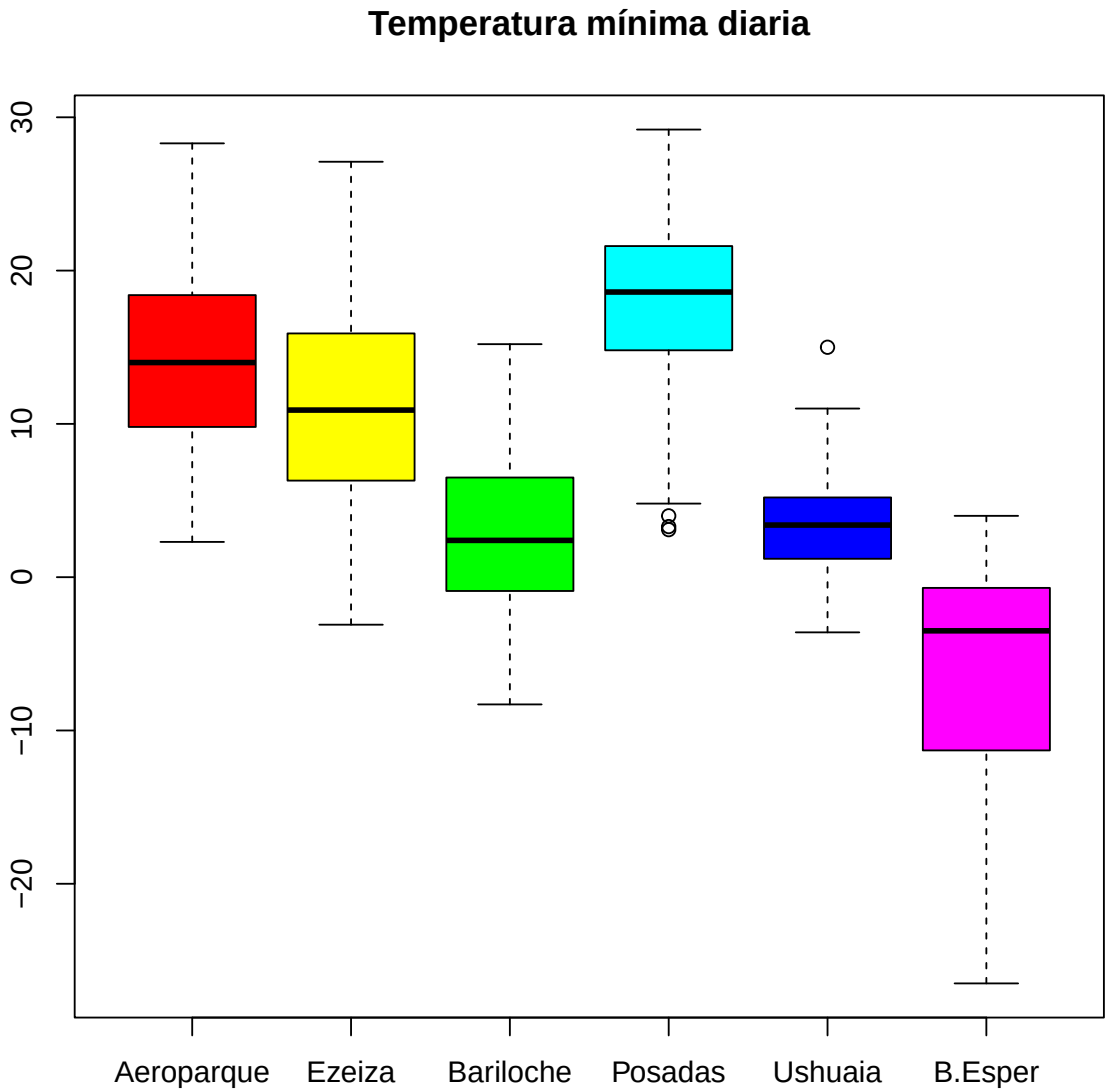




Gráficos de temperaturas mínimas, con correlaciones



Boxplot de temperaturas mínimas, por localidad



¿Cuánto vale el ECM óptimo?

$$H(\beta_0, \beta_1) = ECM(Y, (\beta_0 + \beta_1 X)) = E\left[\left(Y - (\beta_0 + \beta_1 X)\right)^2\right]$$

El predictor óptimo lineal de Y basado en X resulta ser

$$\beta_1^* = \frac{cov(X,Y)}{V(X)} = \frac{cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}\frac{\sqrt{V(Y)}}{\sqrt{V(X)}} = \rho(X,Y)\frac{\sqrt{V(Y)}}{\sqrt{V(X)}} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$$

$\beta_0^* = E(Y) - \rho_{XY}\frac{\sigma_Y}{\sigma_X}E(X)$ entonces

$$H(\beta_0^*, \beta_1^*) = E\left[\left(Y - E(Y) - \rho\frac{\sigma_Y}{\sigma_X}(X - E(X))\right)^2\right]$$

$$= E\left[(Y - E(Y))^2\right] + \rho^2\frac{\sigma_Y^2}{\sigma_X^2}E\left[(X - E(X))^2\right]$$

$$- 2\rho\frac{\sigma_Y}{\sigma_X}E\left[(Y - E(Y))(X - E(X))\right]$$

(copiamos)

$$H(\beta_0^*, \beta_1^*) = E\left[\left(Y - E(Y) - \rho\frac{\sigma_Y}{\sigma_X}(X - E(X))\right)^2\right]$$

$$= E\left[(Y - E(Y))^2\right] + \rho^2\frac{\sigma_Y^2}{\sigma_X^2}E\left[(X - E(X))^2\right]$$

$$- 2\rho\frac{\sigma_Y}{\sigma_X}E\left[(Y - E(Y))(X - E(X))\right]$$

$$= \sigma_Y^2 + \rho^2\frac{\sigma_Y^2}{\sigma_X^2}\sigma_X^2 - 2\rho\frac{\sigma_Y}{\sigma_X}Cov(X, Y)$$

$$= \sigma_Y^2 + \rho^2\sigma_Y^2 - 2\rho\sigma_Y\frac{Cov(X, Y)}{\sigma_X\sigma_Y}\sigma_Y$$

$$= \sigma_Y^2 + \rho^2\sigma_Y^2 - 2\rho\sigma_Y\rho\sigma_Y$$

$$= \sigma_Y^2 + \rho^2\sigma_Y^2 - 2\rho^2\sigma_Y^2 = \sigma_Y^2 - \rho^2\sigma_Y^2 = \sigma_Y^2(1 - \rho^2)$$

¿Cuánto ganamos con el óptimo lineal?

El óptimo lineal

$$H(\beta_0^*, \beta_1^*) = ECM(Y, (\beta_0^* + \beta_1^* X)) = E\left[\left(Y - (\beta_0^* + \beta_1^* X)\right)^2\right]$$
$$= \sigma_Y^2 (1 - \rho^2)$$

La mejor constante

$$ECM(Y, c_{opt}) = E\left[\left(Y - c_{opt}\right)^2\right] = E\left[\left(Y - E(Y)\right)^2\right] = \sigma_Y^2$$

¿Cuál es mayor? Claramente, $ECM(Y, c_{opt}) \geq ECM(Y, (\beta_0^* + \beta_1^* X))$

¿Cuánto ganamos con el óptimo lineal con respecto a la constante óptima?

$$\Delta ECM = ECM(Y, c_{opt}) - ECM(Y, (\beta_0^* + \beta_1^* X)) = \sigma_Y^2 \rho^2$$

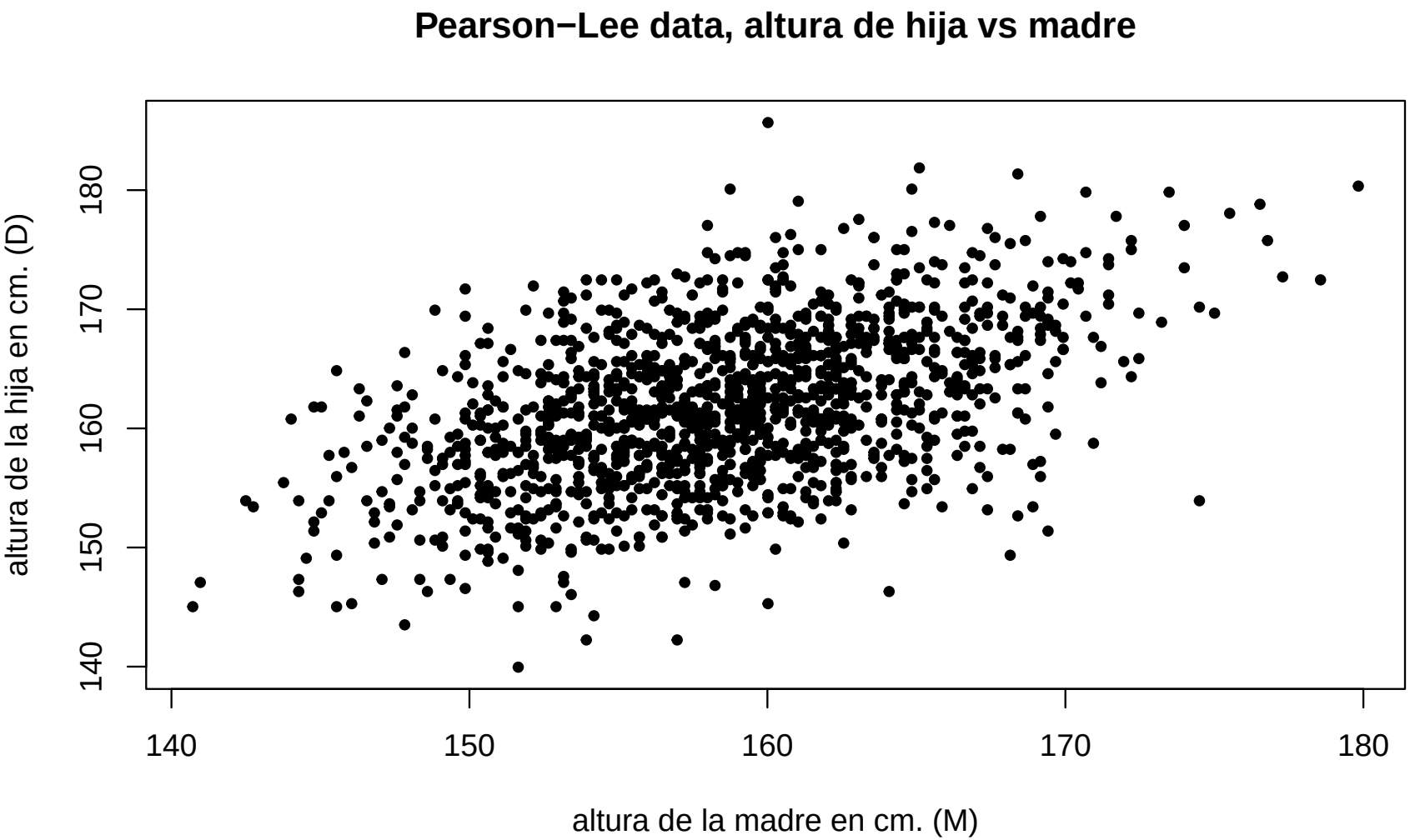
En términos relativos

$$\frac{\Delta ECM}{ECM(Y, c_{opt})} = \rho^2$$

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 `alr3` 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?

Modelo lineal simple

Hacemos los siguientes supuestos

A1. Observaciones independientes

Suponemos que $(X_1, Y_1), \ldots, (X_n, Y_n)$ son independientes.

A2. Esperanza condicional lineal

Suponemos que $E(Y \mid X = x) = \beta_0 + \beta_1 x$ para todo x.

A3. Varianza constante (homoscedasticidad)

Suponemos que $Var(Y \mid X = x) = \sigma^2$ para todo x.

Si las X_i son fijas, el supuesto $A1$ pueden pensarse como que solamente las $Y_1, \ldots, Y_n$ son independientes.

Formulación equivalente del modelo lineal simple

(escritos en términos del error)

Sea $\varepsilon = Y - E(Y \mid X)$ el error. Entonces

B1. Errores independientes

El supuesto A1 implica que los errores $\varepsilon_1, \ldots, \varepsilon_n$ son independientes. A1 es equivalente a que $(X_1, \varepsilon_1), \ldots, (X_n, \varepsilon_n)$ son independientes.

B2. Esperanza condicional cero del error

El supuesto A2: $E(Y \mid X) = \beta_0 + \beta_1 X$ es equivalente a pedir que $E(\varepsilon \mid X) = 0$

B3. Homoscedasticidad del error

El supuesto A3: $Var(Y \mid X = x) = \sigma^2$ para todo x. es equivalente a pedir que $Var(\varepsilon \mid X = x) = \sigma^2$ para todo x.

Del Supuesto $B2$, $E(\varepsilon \mid X) = 0$, se deduce que:

- la $E(\varepsilon) = 0$. Pues tomando esperanza de la esperanza condicional tenemos: $E(\varepsilon) = E\left(E\left[\varepsilon \mid X\right]\right) = E\left(0\right) = 0$.
- $Cov(\varepsilon, X) = 0$. Pues $Cov(\varepsilon, X) = E(\varepsilon X) - E(\varepsilon)E(X) = E(\varepsilon X) = E\left[E\left[\varepsilon X \mid X\right]\right] = E\left[XE\left[\varepsilon \mid X\right]\right] = 0$

Del Supuesto $B3$, $Var\left(\varepsilon \mid X = x\right) = \sigma^2$, se deduce que:

- $Var\left(\varepsilon\right) = \sigma^2$. Pues $Var(\varepsilon) = E(\varepsilon^2) - E(\varepsilon)^2 = E(\varepsilon^2) = E\left[E\left[\varepsilon^2 \mid X\right]\right] = E\left[Var\left[\varepsilon \mid X\right]\right] = E\left[\sigma^2\right] = \sigma^2$

Modelo lineal para las observaciones

Para las observaciones $\{(X_i, Y_i)\}_{i=1,\ldots,n}$, el modelo lineal simple asume que

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{6}$$

para cada $i = 1, \ldots, n$, donde ε_i es el término del error para el individuo i-ésimo. Sobre ellos asumimos:

1. las observaciones $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ son independientes entre sí. O, lo que es lo mismo, que los errores son independientes entre sí.
2. $E(\varepsilon_i \mid X_i) = 0$, para $i = 1, \ldots, n$. Como los errores de distintas observaciones son independientes entre sí, además, tenemos $E(\varepsilon_i \mid X_1, \ldots, X_n) = 0$.
3. $Var(\varepsilon_i \mid X_i) = \sigma^2$. Más aún, por la independencia, vale que $Var(\varepsilon_i \mid X_1, \ldots, X_n) = \sigma^2$. Y también se deduce que $Var(\varepsilon_i) = \sigma^2$, para $i = 1, \ldots, n$.

Estimadores de mínimos cuadrados

Proposición 1.3

Los estimadores de mínimos cuadrados del modelo lineal simple basado en las observaciones $\{(X_i, Y_i)\}_{i=1,\ldots,n}$, *resultan ser*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \overline{X}\,\overline{Y}}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

Son también los estimadores *plug-in*.

Modelo ajustado, valores predichos, residuos

Una vez que se ajustan los parámetros, tenemos lo que se conoce como **modelo ajustado**:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

Y a partir de él, el **valor predicho o ajustado** iésimo (o correspondiente a la iésima observación)

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

También tenemos el **residuo** iésimo

$$r_i = Y_i - \widehat{Y}_i$$

Ejemplo: *low birth weight data*

Datos publicados en

Leviton, A., Fenton, T., Kuban, K. C., y Pagano, M. (1991). Labor and deliver characteristics and the risk of germinal matrix hemorrhage in low birth weight infants. *Journal of child neurology*, 6 (1), 35-40.

Tratados en el libro de

Pagano, M., Gauvreau, K. (2018). *Principles of biostatistics*. Chapman and Hall/CRC.

(o en su versión anterior del año 2000).

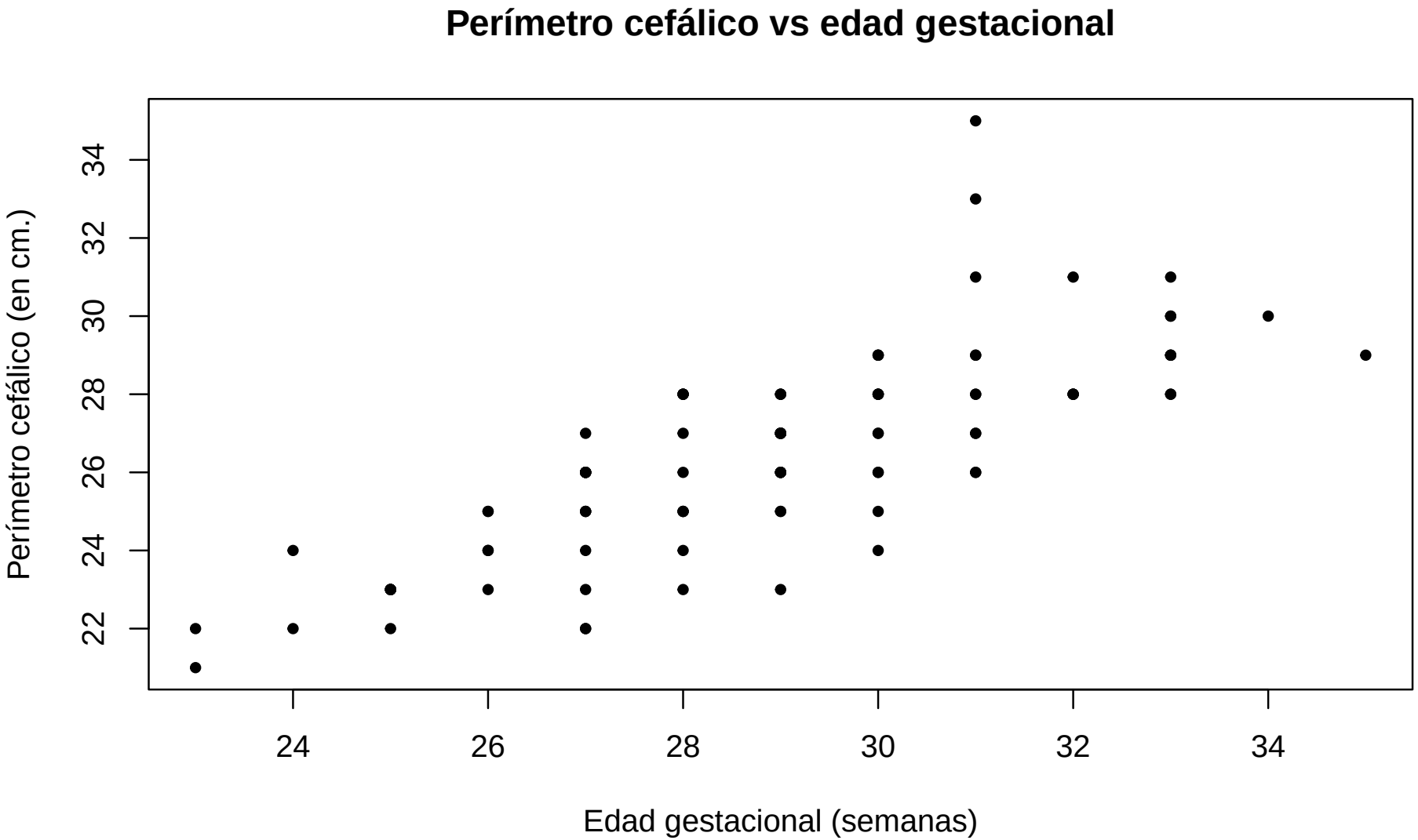
Ejemplo: *low birth weight data*

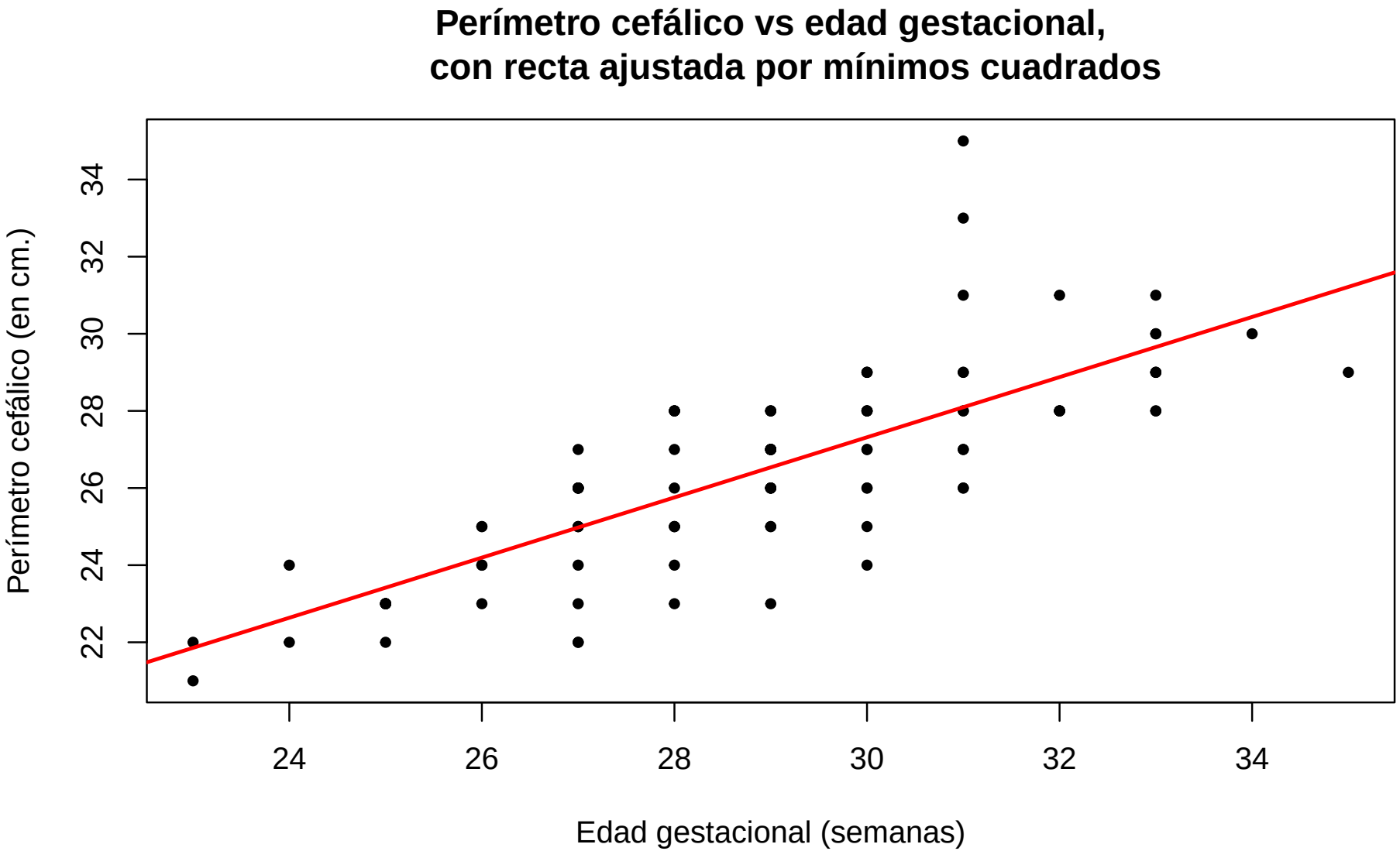
Los datos corresponden a mediciones de 100 niños nacidos con bajo peso (es decir, con menos de 1500g.) en Boston, Massachusetts. Para dichos bebés se miden varias variables. La variable que nos interesa es

- $Y =$ `headcirc`: el perímetro cefálico al nacer (medido en cm.)

La base tiene varias covariables:

- $X_1 =$ `length`: longitud del bebé al nacer, en cm.
- $X_2 =$ `gestage`: edad gestacional o duración del embarazo
- $X_3 =$ `birthwt`: peso del bebé al nacer, en gramos
- $X_4 =$ `momage`: edad de la madre al nacimiento, en años
- $X_5 =$ `toxemia`: indicadora de que la madre padeció una patología durante el embarazo


Perímetro cefálico vs edad gestacional
Perímetro cefálico (en cm.)
22
24
26
28
30
32
34
24
26
28
30
32
34
Edad gestacional (semanas)


Perímetro cefálico vs edad gestacional,
con recta ajustada por mínimos cuadrados
Perímetro cefálico (en cm.)
22
24
26
28
30
32
34
24
26
28
30
32
34
Edad gestacional (semanas)

Modelo Ajustado

```
> ajuste <- lm(headcirc ~ gestage, data = low)
> summary(ajuste)
Residuals:
Min      1Q  Median      3Q     Max
-3.5358 -0.8760 -0.1458  0.9041  6.9041

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.91426    1.82915    2.14   0.0348 *
gestage      0.78005    0.06307   12.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-squared:  0.6095,Adjusted R-squared:  0.6055
F-statistic: 152.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

Modelo ajustado

Cuadro 1: Coeficientes estimados para el modelo de regresión lineal aplicado a los datos de bebés recién nacidos.

```
> ajuste<-lm(headcirc ~ gestage)
> ajuste

Call:
lm(formula = headcirc ~ gestage)

Coefficients:
(Intercept)      gestage
     3.9143       0.7801
```

La recta ajustada resulta ser

$$\widehat{\texttt{headcirc}} = 3.91 + 0.78 \cdot \texttt{gestage}$$

$$\widehat{Y} = 3.91 + 0.78 \cdot X$$

Significado de los coeficientes estimados

Teóricamente, el valor de la ordenada al origen, es decir, 3.91 es el valor de perímetro cefálico esperado para una edad gestacional de 0 semanas. En este ejemplo, sin embargo, la edad 0 semanas no tiene sentido. La pendiente de la recta es 0.78, lo cual implica que para cada incremento de una semana en la edad gestacional, el perímetro cefálico del bebé aumenta 0.78 centímetros en promedio.

O bien, la diferencia esperada en el perímetro cefálico medio de dos grupos de bebés que difieren en una semana de edad getacional es 0.78, es decir,

$$E\left[\texttt{headcirc} \mid \widehat{\texttt{gestage} = x + 1}\right] - E\left[\texttt{headcirc} \mid \widehat{\texttt{gestage} = x}\right] = 0.78$$

para cualquier valor de x (edad gestacional).

A veces (no en este caso), tiene más sentido emplear un aumento de la variable explicativa mayor a una unidad, para expresar el significado de la pendiente, esto sucede cuando las unidades de medida de la covariable son muy pequeñas, por ejemplo.

Valores predichos y residuos, en el ejemplo

Cuadro 2: Tres datos de los bebés de bajo peso de la base de datos

Caso (i)	Y_i	X_i	$\widehat{Y}_i = 3.91 + 0.78 \cdot X_i$ (predicho)	$r_i = Y_i - \widehat{Y}_i$ (residuo)
1	27	29	$3.91 + 0.78 \cdot 29 = 26.537$	$27 - 26.537 = 0.463$
3	30	33	29.658	0.342
6	23	25	23.417	-0.417

Proposición 1.4

Bajo el modelo de regresión lineal simple, los estimadores de mínimos cuadrados satisfacen

$$E\left(\widehat{\beta}_1 \mid X_1, \ldots, X_n\right) = \beta_1 \quad y \quad E\left(\widehat{\beta}_1\right) = \beta_1$$
$$E\left(\widehat{\beta}_0 \mid X_1, \ldots, X_n\right) = \beta_0 \quad y \quad E\left(\widehat{\beta}_0\right) = \beta_0$$

Es decir, son insesgados y condicionalmente insesgados. Más aún podemos calcular su varianza condicional

$$Var\left(\widehat{\beta}_1 \mid X_1, \ldots, X_n\right) = \frac{\sigma^2}{n\, s_X^2} \quad donde\ s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

$$Var\left(\widehat{\beta}_0 \mid X_1, \ldots, X_n\right) = \frac{\sigma^2}{n}\left(1 + \frac{\overline{X}^2}{s_X^2}\right)$$

$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{n} E\left(\frac{1}{s_X^2}\right) \quad y \quad Var\left(\widehat{\beta}_0\right) = \frac{\sigma^2}{n} E\left(\frac{\overline{X}^2}{s_X^2}\right)$$

Demostración.

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \overline{X}\,\overline{Y}}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}}{s_X^2}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^n X_i \left(\beta_0 + \beta_1 X_i + \varepsilon_i\right) - \bar{X}\left(\beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}\right)}{s_X^2}$$

$$= \frac{\beta_0 \bar{X} + \beta_1 \overline{X^2} + \frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i - \bar{X}\beta_0 - \beta_1 \bar{X}^2 - \bar{X}\bar{\varepsilon}}{s_X^2}$$

$$= \frac{\beta_1 s_X^2 + \frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i - \bar{X}\bar{\varepsilon}}{s_X^2}$$

$$= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n X_i \varepsilon_i - \bar{X}\bar{\varepsilon}}{s_X^2}$$

Como $\bar{X}\bar{\varepsilon} = \frac{1}{n}\sum_i \bar{X}\varepsilon_i$ podemos escribir

$$\widehat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)\varepsilon_i}{s_X^2} = \beta_1 + \sum_{i=1}^n \frac{\frac{1}{n}\left(X_i - \bar{X}\right)}{s_X^2}\varepsilon_i \qquad (7)$$

Demostración (cont.)

Esta manera de escribir al estimador de la pendiente muestra que es igual a la verdadera pendiente más una expresión que depende de los errores. La vamos a usar para calcular esperanza y varianza de $\widehat{\beta}_1$.

Queremos $E\left[\widehat{\beta}_1\right] = E\left[E\left(\widehat{\beta}_1 \mid X_1, \ldots, X_n\right)\right]$ Calcular la esperanza condicional (roja) equivale a pensar los $X_1, \ldots, X_n$ como constantes.

$$E\left(\widehat{\beta}_1 \mid X_1, \ldots, X_n\right) = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)}{s_X^2} \underbrace{E\left(\varepsilon_i \mid X_1, \ldots, X_n\right)}_{=0} = \beta_1$$

Luego $E\left[\widehat{\beta}_1\right] = E\left[E\left(\widehat{\beta}_1 \mid X_1, \ldots, X_n\right)\right] = E\left[\beta_1\right] = \beta_1$. Luego, $\widehat{\beta}_1$ es un estimador insesgado de β_1.

Para el estimador de β_0, puede obtenerse la siguiente expresión análoga a (7) a partir de (7) y de

$$\widehat{\beta}_0 = \beta_0 + \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{X}\frac{\left(X_i - \bar{X}\right)}{n s_X^2}\right)\varepsilon_i \tag{8}$$

Demostración (cont.)

Hallemos la varianza. Comenzamos calculando la varianza condicional, que equivale a calcular la varianza de $\widehat{\beta}_1$ asumiendo que los valores de $X_1, \ldots, X_n$ son fijos.

$$
\begin{aligned}
Var\left(\widehat{\beta}_1 \mid X_1, \ldots, X_n\right) &= Var\left(\sum_{i=1}^{n} a_i \varepsilon_i \mid X_1, \ldots, X_n\right) \\
&= \sum_{i=1}^{n} a_i^2 \, Var\left(\varepsilon_i \mid X_1, \ldots, X_n\right) = \sum_{i=1}^{n} a_i^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} \frac{\frac{1}{n^2}\left(X_i - \bar{X}\right)^2}{s_X^4} = \frac{\sigma^2}{n\, s_X^4} s_X^2 = \frac{\sigma^2}{n\, s_X^2}
\end{aligned}
$$

Como las esperanzas y varianzas condicionales se relacionan, tenemos el siguiente resultado:

Lema 1.1

$$Var(Y) = E\Big[Var(Y \mid X)\Big] + Var\Big(E[Y \mid X]\Big) .$$

Podemos deducir la varianza de $\widehat{\beta}_1$.

Demostración (cuenta final).

$$\begin{aligned} V(\widehat{\beta}_1) &= E\Big[V(\widehat{\beta}_1 \mid X_1, \ldots, X_n)\Big] + V\Big(E[\widehat{\beta}_1 \mid X_1, \ldots, X_n]\Big) \\ &= E\left[\frac{\sigma^2}{n\, s_X^2}\right] + Var\Big(\beta_1\Big) \\ &= \frac{\sigma^2}{n} E\left[\frac{1}{s_X^2}\right] \end{aligned}$$

Estimación de σ^2

Ejercicio 1.5

Probar que $E\left[(Y-(\beta_0+\beta_1 X))^2\right]=\sigma^2$ *asumiendo que vale el modelo lineal simple.*

Esto nos sugiere que un estimador *natural* de σ^2 es

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1 X_i\right)\right)^2$$

Puede probarse (¡otro ejercicio!) que este resulta ser un estimador ligeramente sesgado de σ^2, un estimador insesgado es:

$$s^2=\frac{n}{n-2}\widehat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1 X_i\right)\right)^2=\frac{SSRes}{n-2}$$

Vamos a las ecuaciones normales p=2 | Volvemos a S^2 es insesgado

Estimación de σ^2

Otra maneras de nombrar al estimador de

- σ^2 : *mean squared residuals*, MSRes
- σ : *residual standard error*, como en `R`, que forma parte de la salida del summary

Miremos la salida del ejemplo de niños de bajo peso

```
summary(ajuste)
....
Residual standard error: 1.59 on 98 degrees of freedom
```

Luego, la estimación de σ en base a los datos ‘‘`low`’’ de bajo peso es $S = 1.59$

¿Errores o residuos? Encuentre las diferencias...

Los errores no son los residuos.

- Tanto los errores (ε_i) como los residuos (r_i) son variables aleatorias.
- Podemos ver que "juegan papeles similares". Cuando $X = x_i$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{(modelo propuesto)} \tag{9}$$

$$r_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \quad \text{(por definición)} \tag{10}$$

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + r_i \quad \text{(modelo ajustado)} \tag{11}$$

El lado derecho de (9) involucra parámetros verdaderos. En el lado derecho de (11) hay estimadores de los parámetros. En particular, los parámetros estimados son funciones de todas las observaciones, y por lo tanto, los residuos dependen de todas ellas. Los errores son independientes entre sí, e independientes de todo el resto.

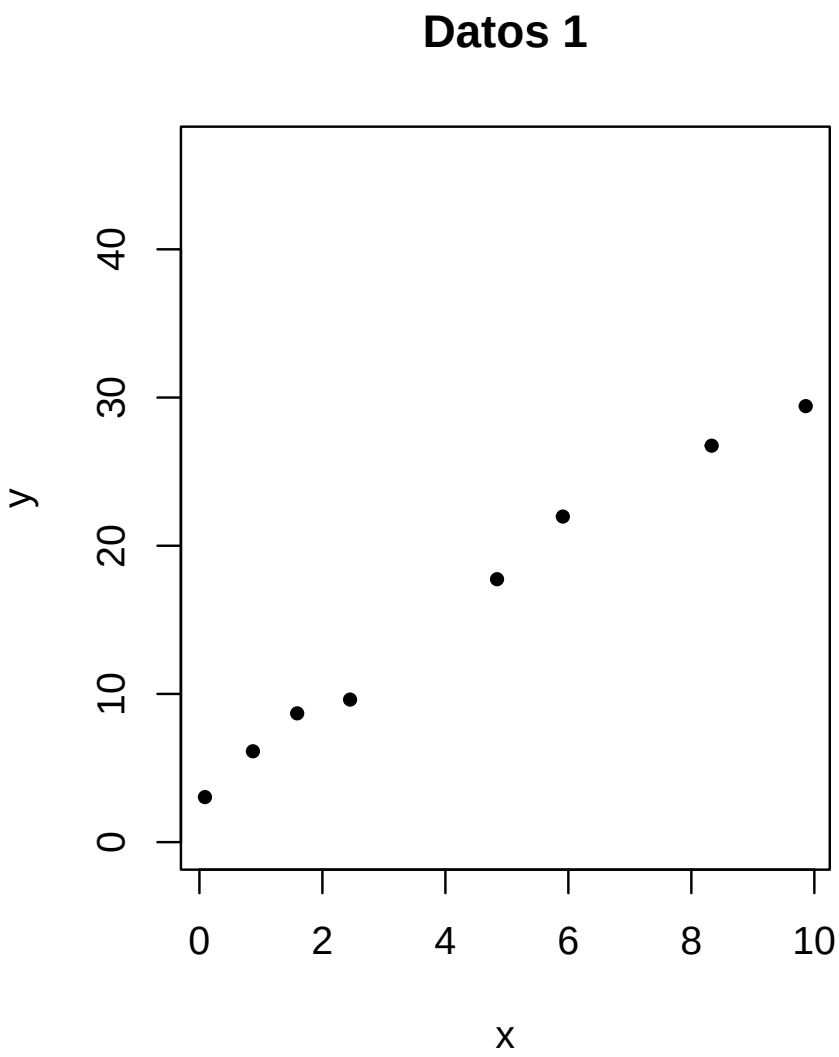
- Los errores no son observables.

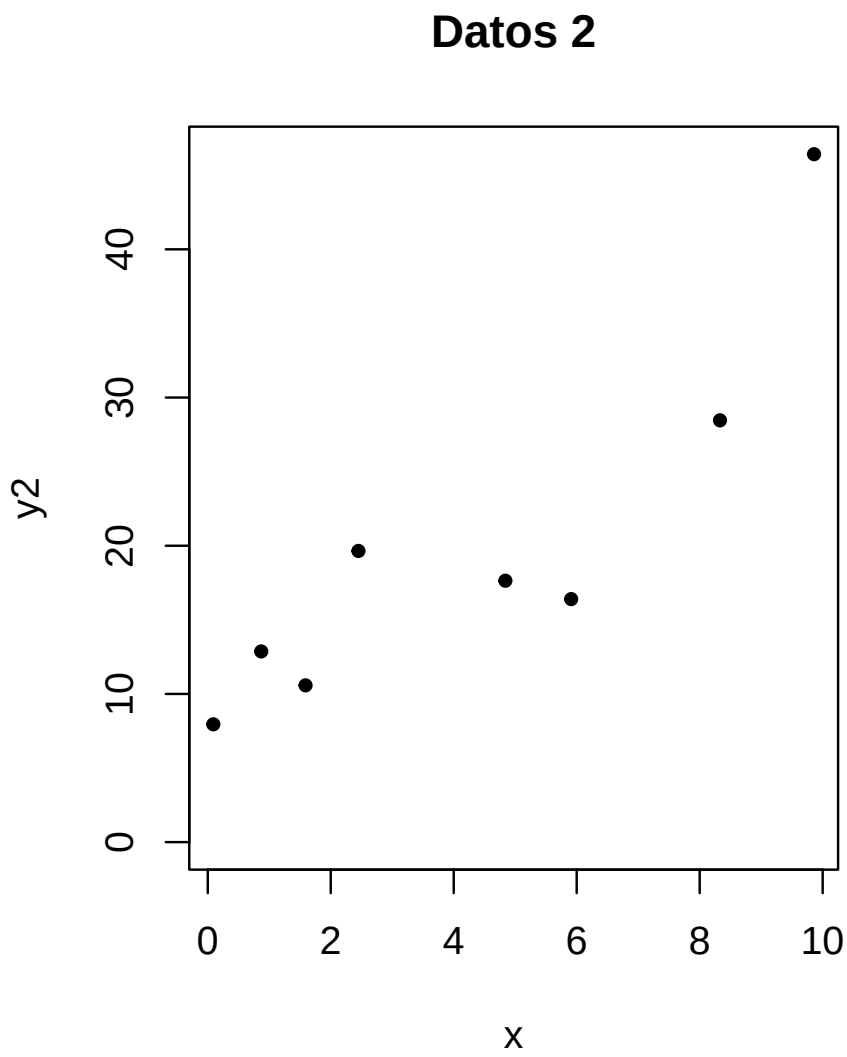
¿Errores o residuos? Encuentre las diferencias...

- La media muestral de los residuos es cero es decir, $\frac{1}{n}\sum_{i=1}^{n} r_i = 0$ (ejercicio de la Práctica 2)
- La covarianza muestral de la covariable y los residuos, (X_i, r_i) es cero: es decir, $\frac{1}{n}\sum_{i=1}^{n} r_i \left(X_i - \bar{X}\right) = 0$ (ejercicio de la Práctica 2)

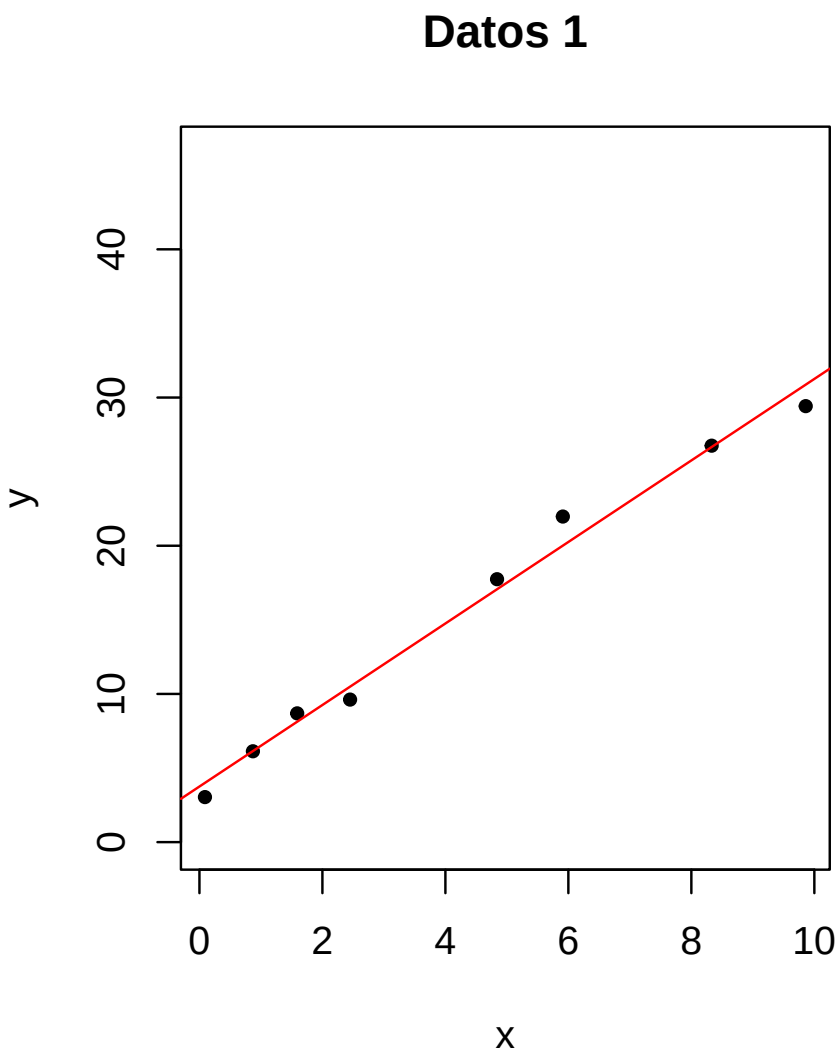
Ojo : estas dos afirmaciones son consecuencia de las ecuaciones normales, y serán ciertas aun si el modelo de regresión lineal simple **no** es verdadero para los datos.
Más allá de estas diferencias, hay mucho vínculo entre ambos. Los residuos son el "correlato empírico" de los errores. De hecho, la mayor parte del diagnóstico del modelo que haremos (es decir, técnicas y gráficos para chequear que los datos cumplan los supuestos que hacemos sobre el modelo) los haremos a través de los residuos. No sólo en modelo lineal.

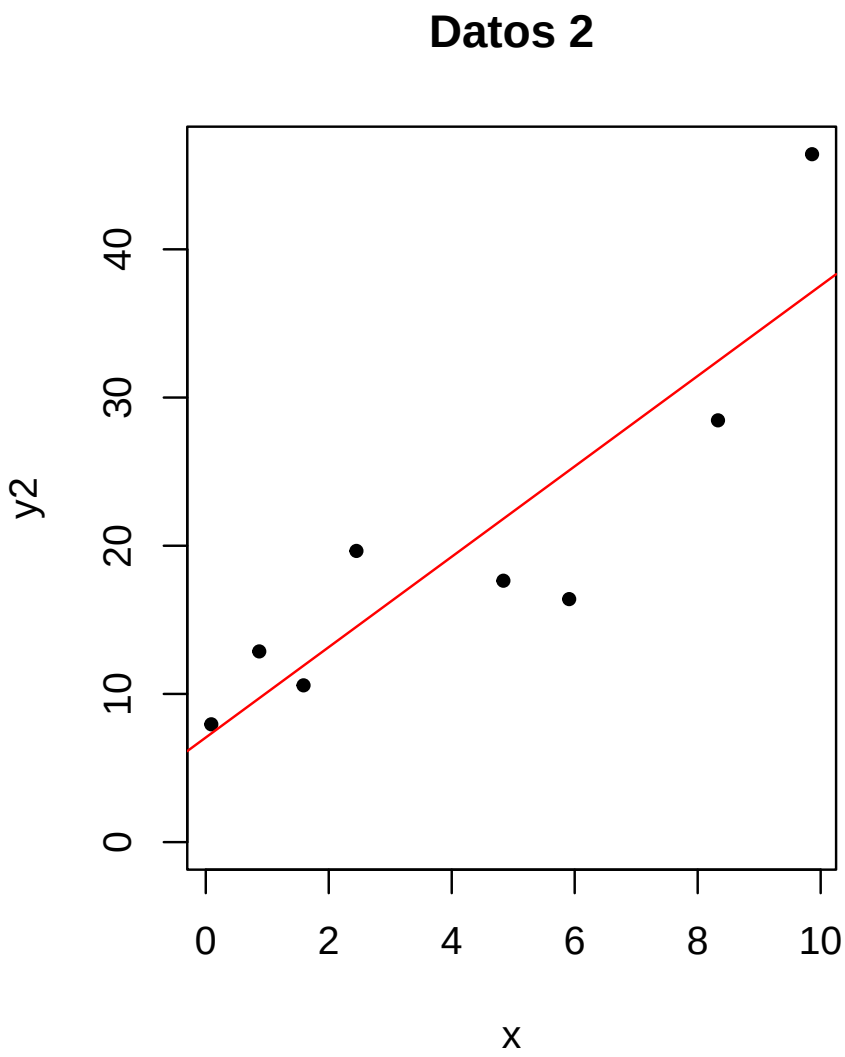
El efecto de σ^2: dos conjuntos de datos



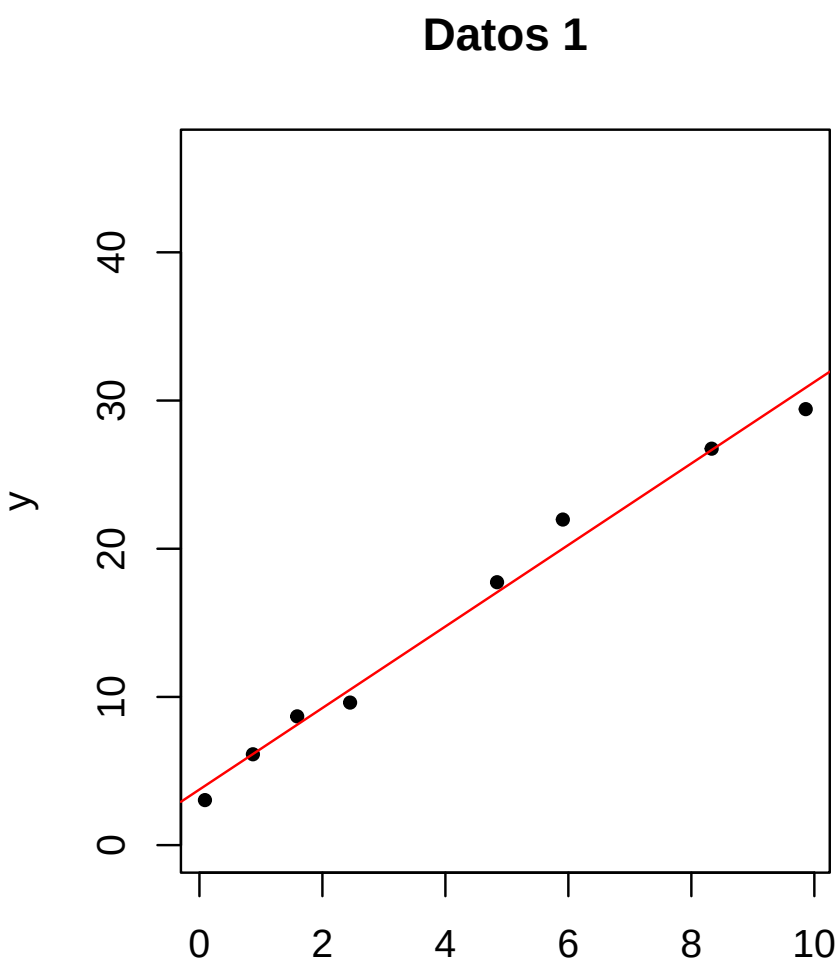


El efecto de σ^2: dos conjuntos de datos

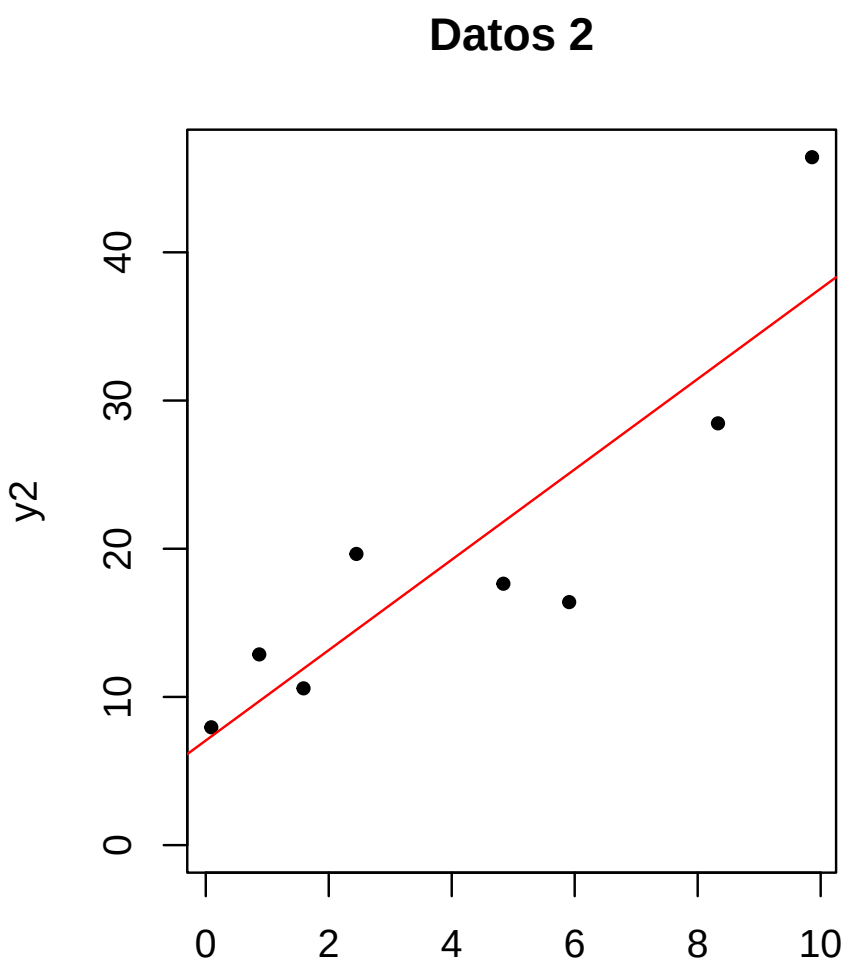




El efecto de σ^2: dos conjuntos de datos

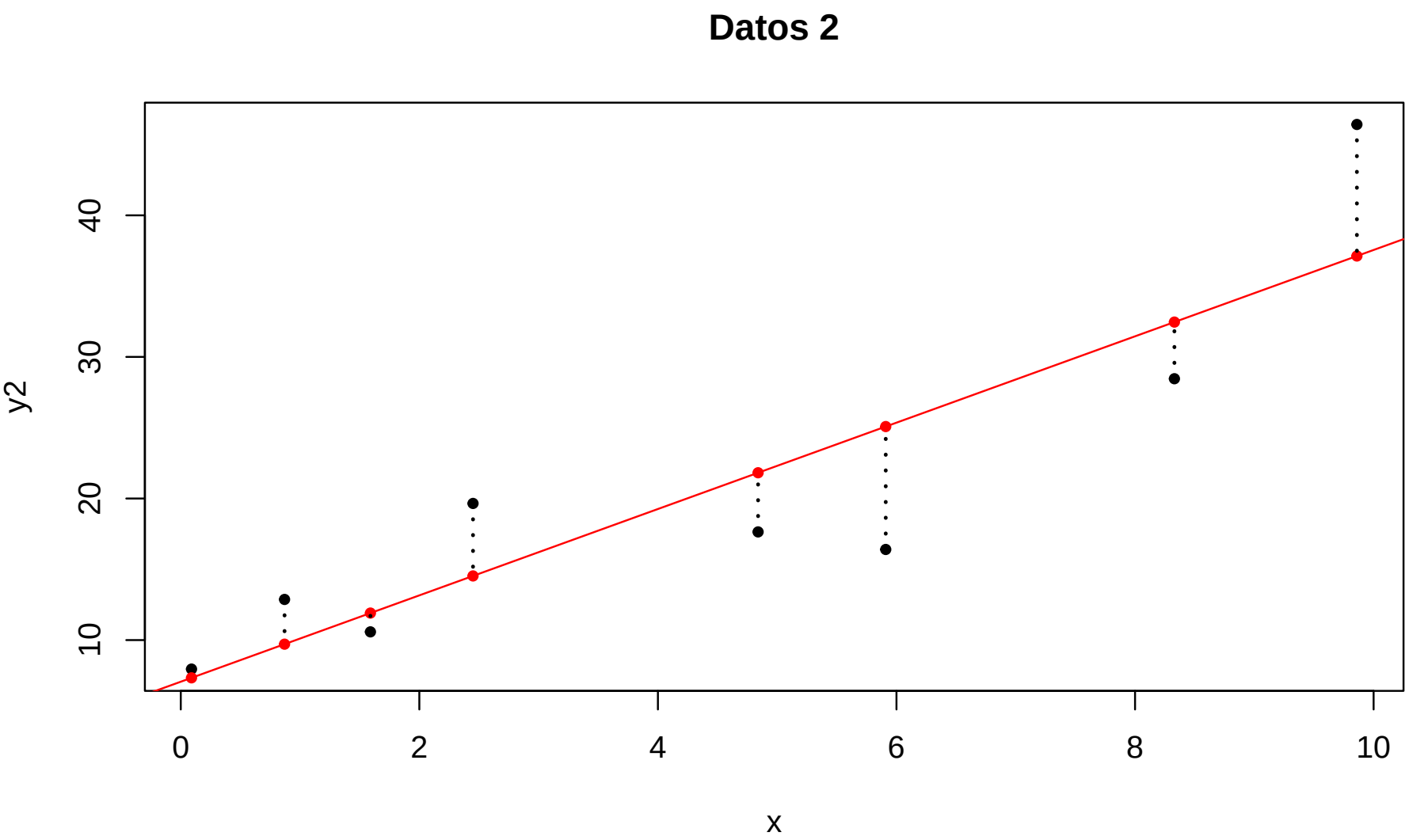


s = 1.18

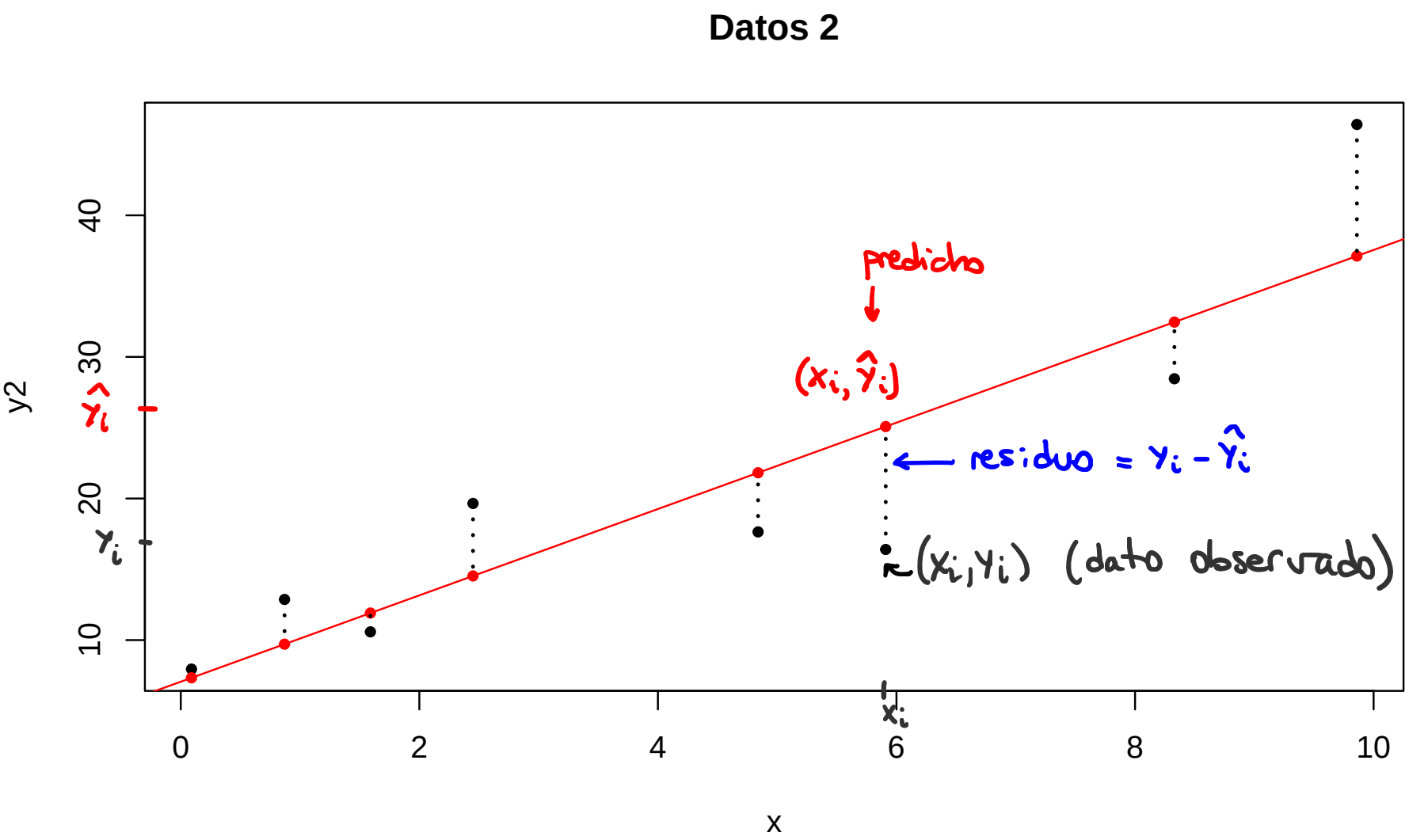


s = 6.24

Datos 2, con recta ajustada, valores predichos y residuos



Datos 2, con recta ajustada, valores predichos y residuos



¿Qué ganamos con el modelo lineal simple, para los datos de bajo peso?

sin covariable

- Resumimos la variable $Y =$ perímetro cefálico con una constante: $\overline{Y} = 26.45$ cm.
- Predecimos la variable $Y =$ perímetro cefálico de un bebito de la población de bajo peso por 26.45.
- Cuantificamos la incerteza total involucrada en ese resumen por $\widehat{ECM}(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = 6.3475$

con covariable

- Resumimos la variable $Y =$ perímetro cefálico con una recta, para valor de $X =$ edad gestacional tenemos $\widehat{m}(x) = 3.91 + 0.78x$
- Predecimos la variable $Y =$ perímetro cefálico para tres bebitos de la población de bajo peso según su edad gestacional:
  - Si $X = 23$, predecimos el perímetro cefálico por 23.417 cm.
  - Si $X = 29$, predecimos el perímetro cefálico por 26.537 cm.
  - Si $X = 33$, predecimos el perímetro cefálico por 29.658 cm.
- Cuantificamos la incerteza total involucrada en ese resumen $\widehat{ECM}(\widehat{\beta}_0, \widehat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \frac{1}{n}SSRes = 2.4788$

Agregamos la normalidad de los errores

Definición 1.3

Decimos que las observaciones siguen el modelo de regresión lineal con distribución condicional normal, es decir,

$$Y_i|X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

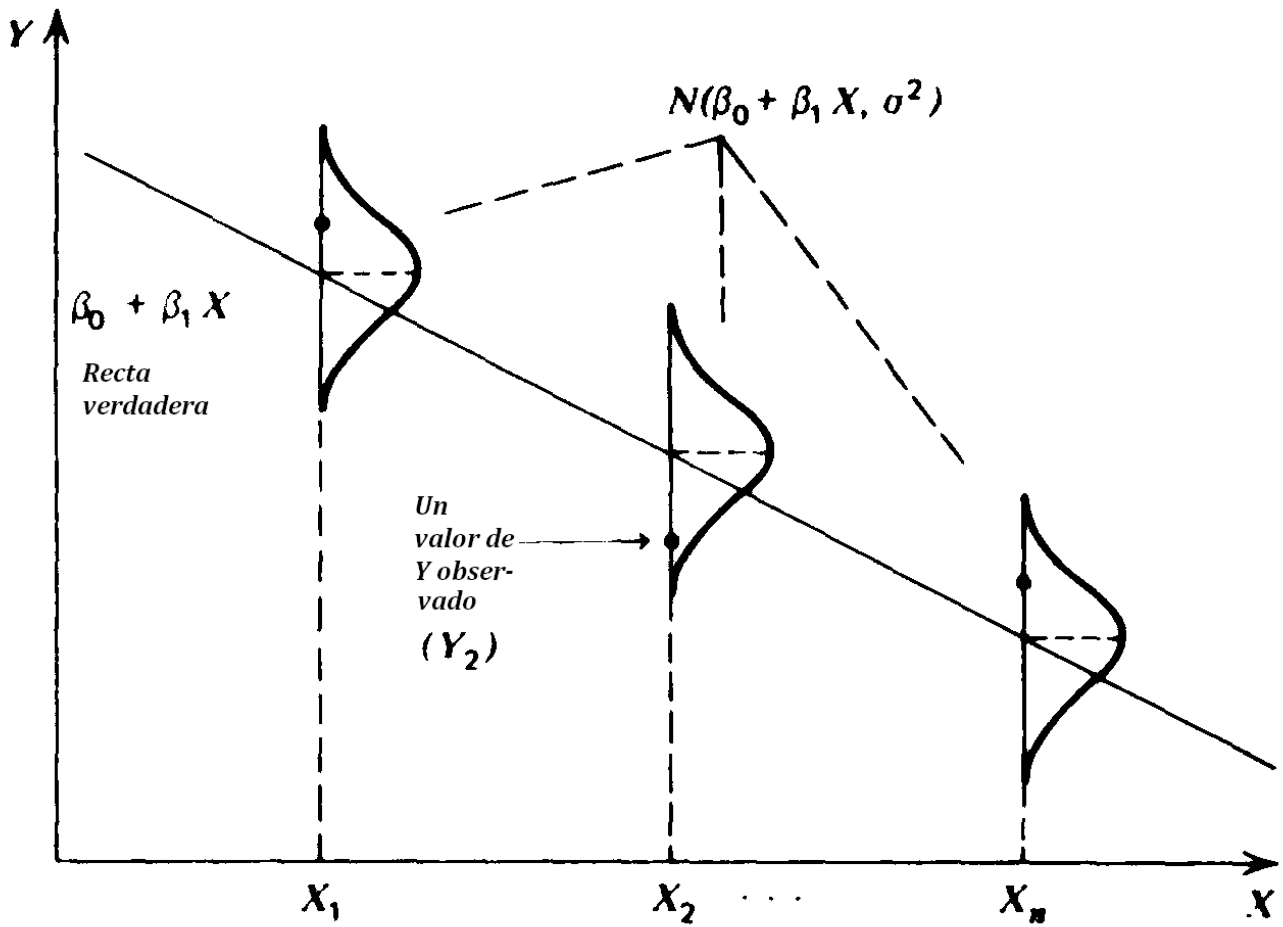
o equivalentemente

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \textit{con } \varepsilon_i \sim N(0, \sigma^2)$$

con observaciones independientes entre sí, y X_i y ε_i independientes.

Modelo y Supuestos vía la distribución condicional

Figura 1: Suponemos que cada observación de la variable respuesta proviene de una distribución normal centrada verticalmente en el nivel que da la recta. Asumimos que la varianza de cada distribución normal es la misma, σ^2.



Repaso de variables aleatorias I

Definición 1.4 (Distribución χ^2_1)

Si $Z \sim \mathrm{N}(0,1)$ *entonces la distribución de* $Y = Z^2$ *se denomina* **chi-cuadrado con un grado de libertad**. *La notaremos* $Y \sim \chi^2_1 = \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$.

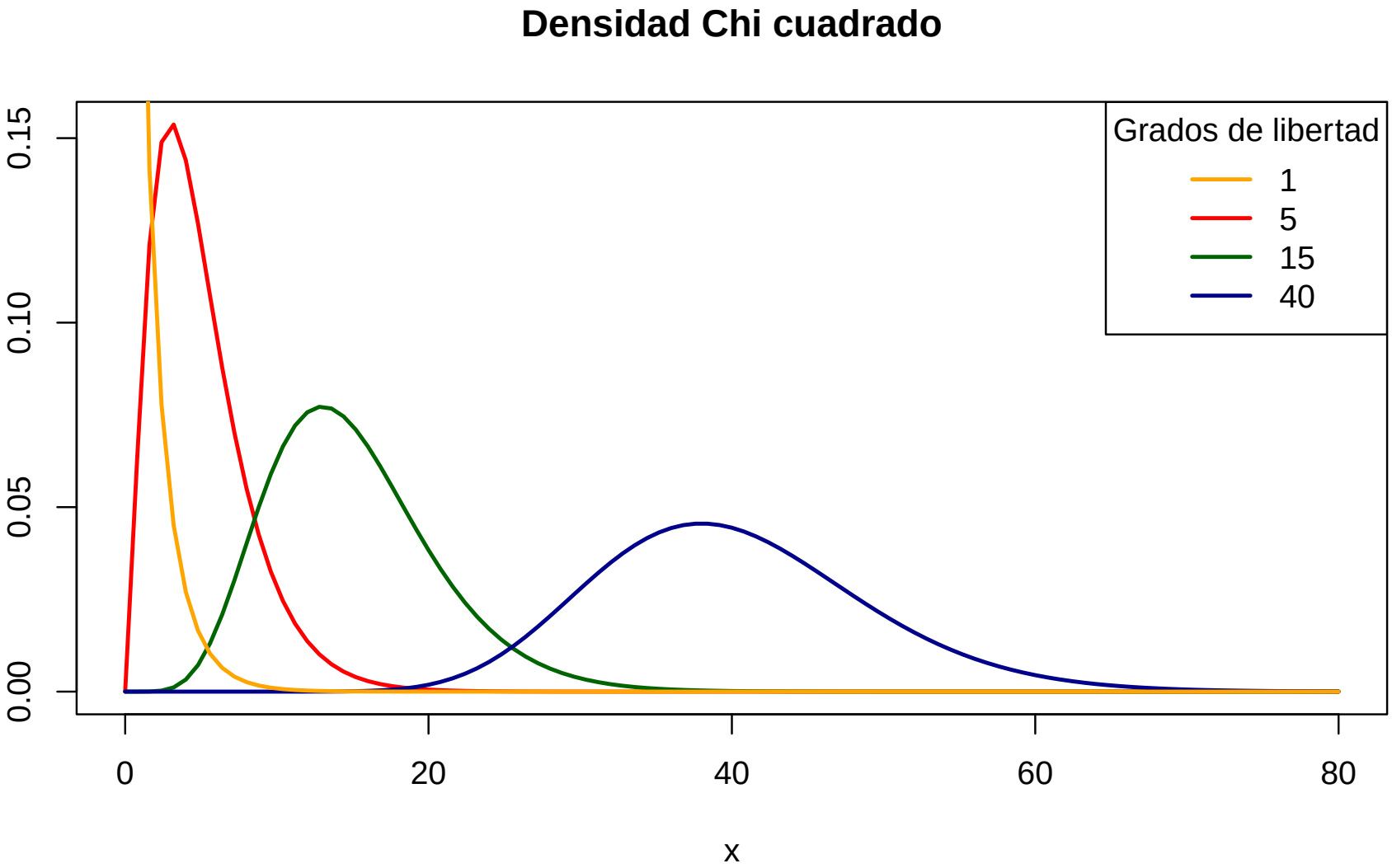
Proposición 1.5

Sean $Y_1 \sim \Gamma(\alpha_1, \lambda)$ *e* $Y_2 \sim \Gamma(\alpha_2, \lambda)$ *variables aleatorias independientes. Entonces* $Y_1 + Y_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.

Definición 1.5 (Distribución χ^2_n)

Sean $Z_i,\ i = 1, 2, \ldots, n$ *variables independientes con distribución* $\mathrm{N}(0,1)$. *La distribución de la variable aleatoria* $Y = \sum_{i=1}^{n} Z_i^2$ *la denominaremos* **distribución chi-cuadrada** *con n grados de libertad, que simbolizaremos por* χ^2_n. *Vale que* $\chi^2_n = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right),\ n \in \mathbb{N}$.

$$E(Y) = n \quad var(Y) = 2n$$


Densidad Chi cuadrado
Grados de libertad
1
5
15
40
0.00
0.05
0.10
0.15
0
20
40
60
80
x

Repaso de variables aleatorias II

Definición 1.6 (Distribución t de Student)

Sean $U \sim \mathrm{N}(0,1)$ y V con distribución χ^2_n con U y V independientes. Luego se define la **distribución t de Student** *con n grados de libertad, que simbolizaremos con t_n, como la distribución de*

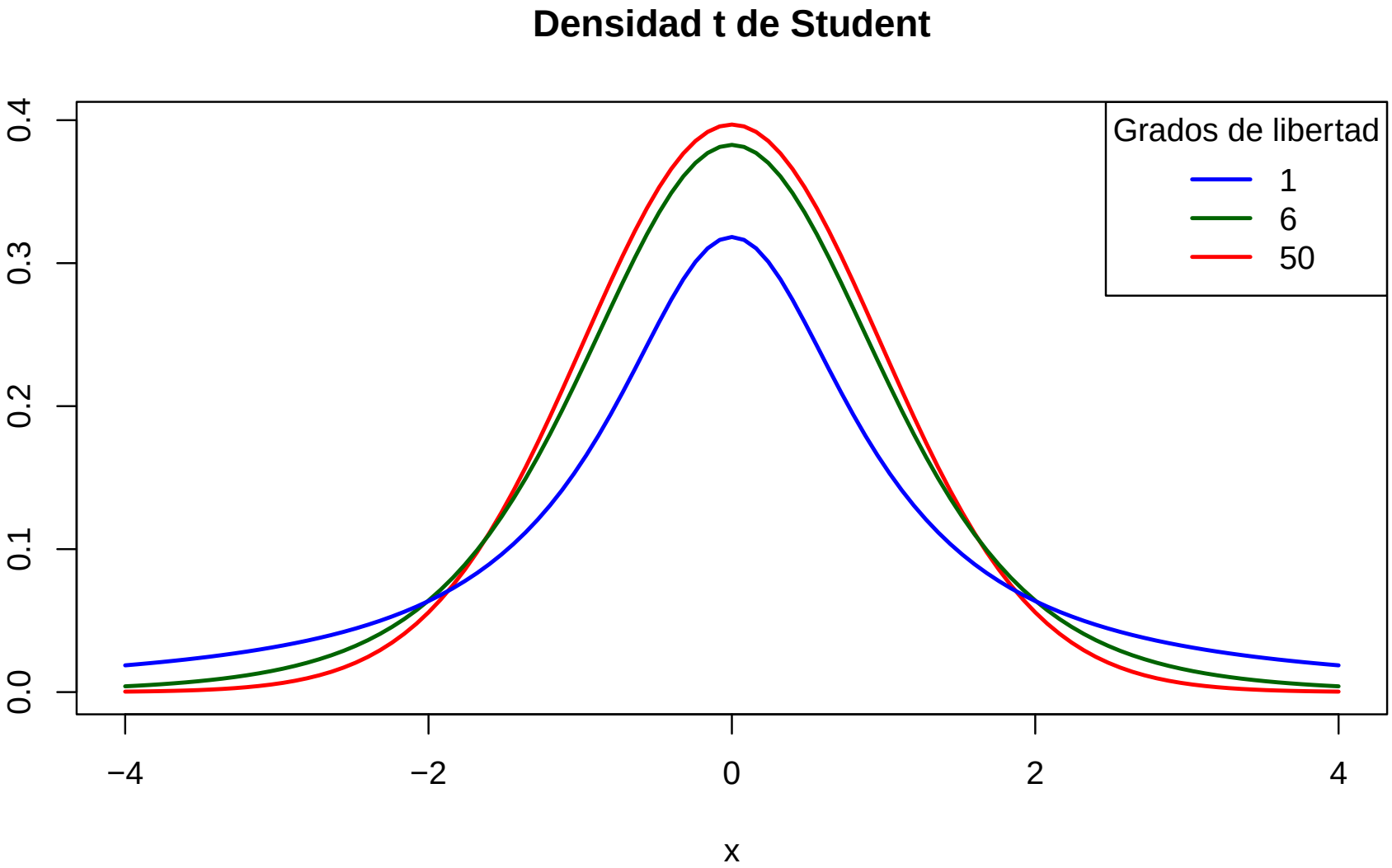
$$T = \frac{U}{\sqrt{V/n}}.$$

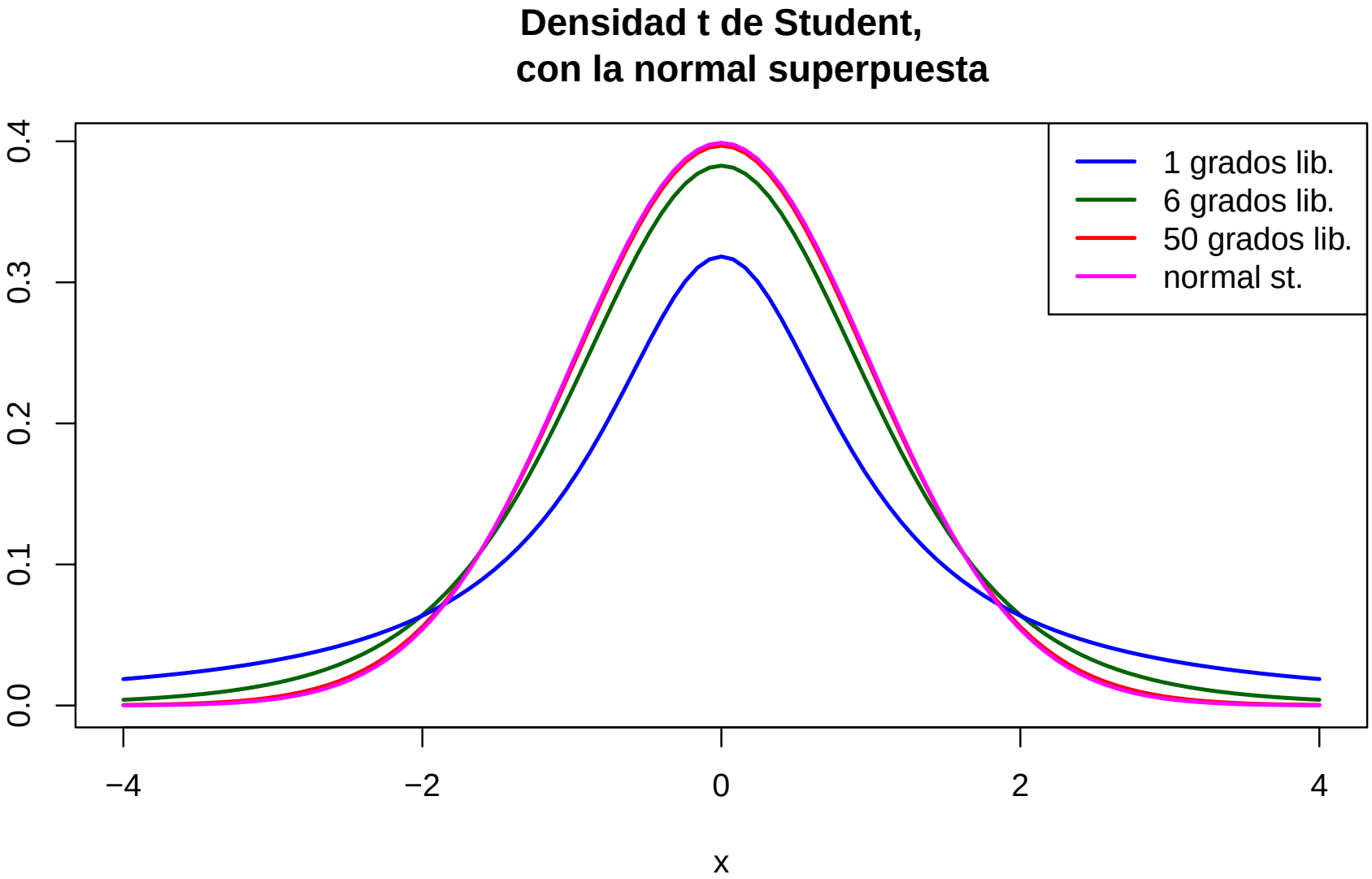
La densidad de T es

$$f_T(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

$$E(T) = 0, \quad n \geq 2$$

$$var(T) = \frac{n}{n-2}, \quad n > 2$$


Densidad t de Student
Grados de libertad
1
6
50
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x


Densidad t de Student,
con la normal superpuesta
1 grados lib.
6 grados lib.
50 grados lib.
normal st.
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

Repaso de variables aleatorias III

Definición 1.7 (Distribución F de Fisher)

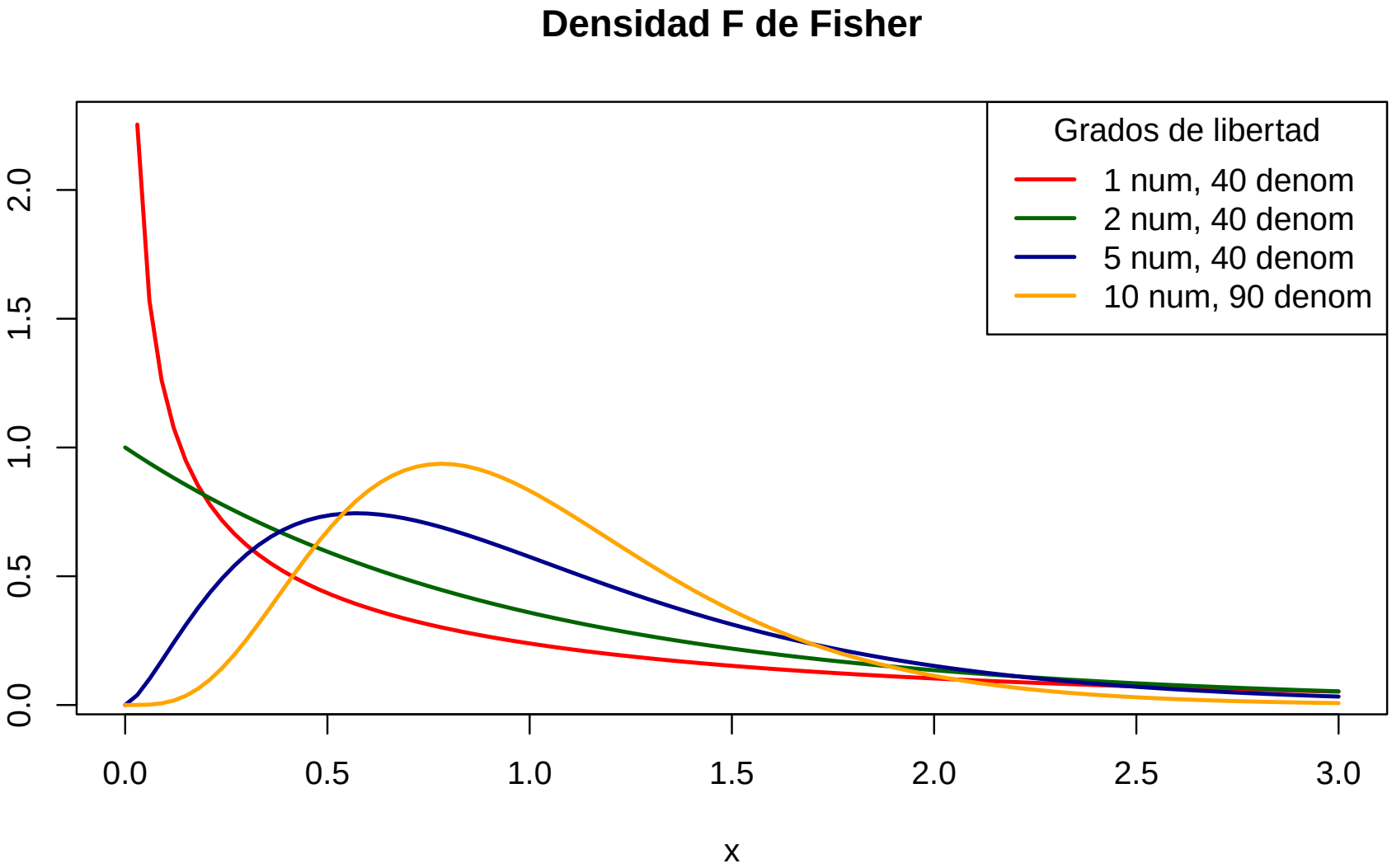
Sean $U \sim \chi^2_n$ y $V \sim \chi^2_m$ con U y V independientes. Luego se define la **distribución de F de Fisher o de Snedecor** *con n grados de libertad en el numerador y m grados de libertad en el denominador que notaremos $F_{n,m}$, como la distribución de*

$$F = \frac{U/n}{V/m}$$

Su función de densidad está dada por

$$f_F(x) = \frac{\Gamma((m+n)/2)}{\Gamma(n/2)\Gamma(m/2)} \left(\frac{n}{m}\right)^{n/2} x^{(n/2)-1} \left(1 + \frac{n}{m}x\right)^{-(n+m)/2} I_{(0,\infty)}(x).$$

$$E(X) = \frac{n}{n-2}, \text{ si } n \geq 2, \quad \text{Var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \text{ para } n > 4.$$


Densidad F de Fisher
Grados de libertad
1 num, 40 denom
2 num, 40 denom
5 num, 40 denom
10 num, 90 denom
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x

Repaso de variables aleatorias IV

OBSERVACIONES:

1. Sea $T \sim t_n$, y definimos $W = T^2 = \left(\frac{U}{\sqrt{V/n}}\right)^2$. Resulta que $W \sim F_{1,n}$.
2. La distribución $t_1 = Cauchy(0, 1)$ y no tiene esperanza ni varianza finitas.
3. La distribución t_2 tiene esperanza finita (0), pero no tiene varianza finita.

Teorema 1.2

Bajo el modelo de regresión lineal con distribución condicional normal,

1. *los estimadores de MV de* β_0 *y* β_1 *son iguales a los de mínimos cuadrados y el estimador de* σ^2 *resulta ser*
   $\widehat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right)^2 = \frac{SSRes}{n}$
2. $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ *y* $S^2 = \frac{n}{n-2}\widehat{\sigma}^2_{MV}$ *son independientes.*
3. $\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{ns_X^2}\right)$ *y* $\widehat{\beta}_0 \sim N(\beta_0, \underbrace{\sigma^2\left(\frac{1}{n} + \frac{\overline{X}^2}{s_X^2 n}\right)}_{\sigma^2 \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{ns_X^2}})$
4. $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$
5. $T = \frac{\widehat{\beta}_j - \beta_j}{\widehat{sd}\left(\widehat{\beta}_j\right)} \sim t_{n-2}$, *para* $j = 0, 1$ *donde en* $\widehat{sd}\left(\widehat{\beta}_j\right)$ *reemplazamos a* σ *por* S.

Recordar que $s_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

Este Teorema nos permite hallar intervalos de confianza, intervalos de predicción y hacer tests. El `R` hace algunos de forma automática con el comando `lm`. También `predict`, `confint`.

Modelo Ajustado

```
> ajuste <- lm(headcirc ~ gestage, data = low)
> summary(ajuste)
Residuals:
Min      1Q  Median      3Q     Max
-3.5358 -0.8760 -0.1458  0.9041  6.9041

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.91426    1.82915    2.14   0.0348 *
gestage      0.78005    0.06307   12.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-squared:  0.6095,Adjusted R-squared:  0.6055
F-statistic: 152.9 on 1 and 98 DF,  p-value: < 2.2e-16
```