Apunte de Regresión Lineal

María Eugenia Szretter Noste
Carrera de Especialización en Estadística
para Ciencias de la Salud
Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires

Agosto - Octubre de 2017

Tabla 1: Observaciones a nuestra disposición. Aquí X_1 quiere decir, la variable X medida en el individuo 1, etc.

Individuo	Variable X	Variable Y
1	X_1	Y_1
2	X_2	Y_2
$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n

En estas notas, estamos pensando en que medimos ambas variables en la misma unidad: puede tratarse de un individuo, un país, un animal, una escuela, etc. Comencemos con un ejemplo.

duo. El más utilizado de todos es el que se conoce como *coeficiente de correlación*, que se simboliza con una letra griega *rho*: ρ ó ρ_{XY} y se define por

$$\begin{aligned}\rho_{XY} &= E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right] \\ &= \frac{cov\,(X,Y)}{\sigma_X\sigma_Y},\end{aligned}$$

$$\widehat{\mu}_X = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

o sea, el número promedio a nivel población del pro

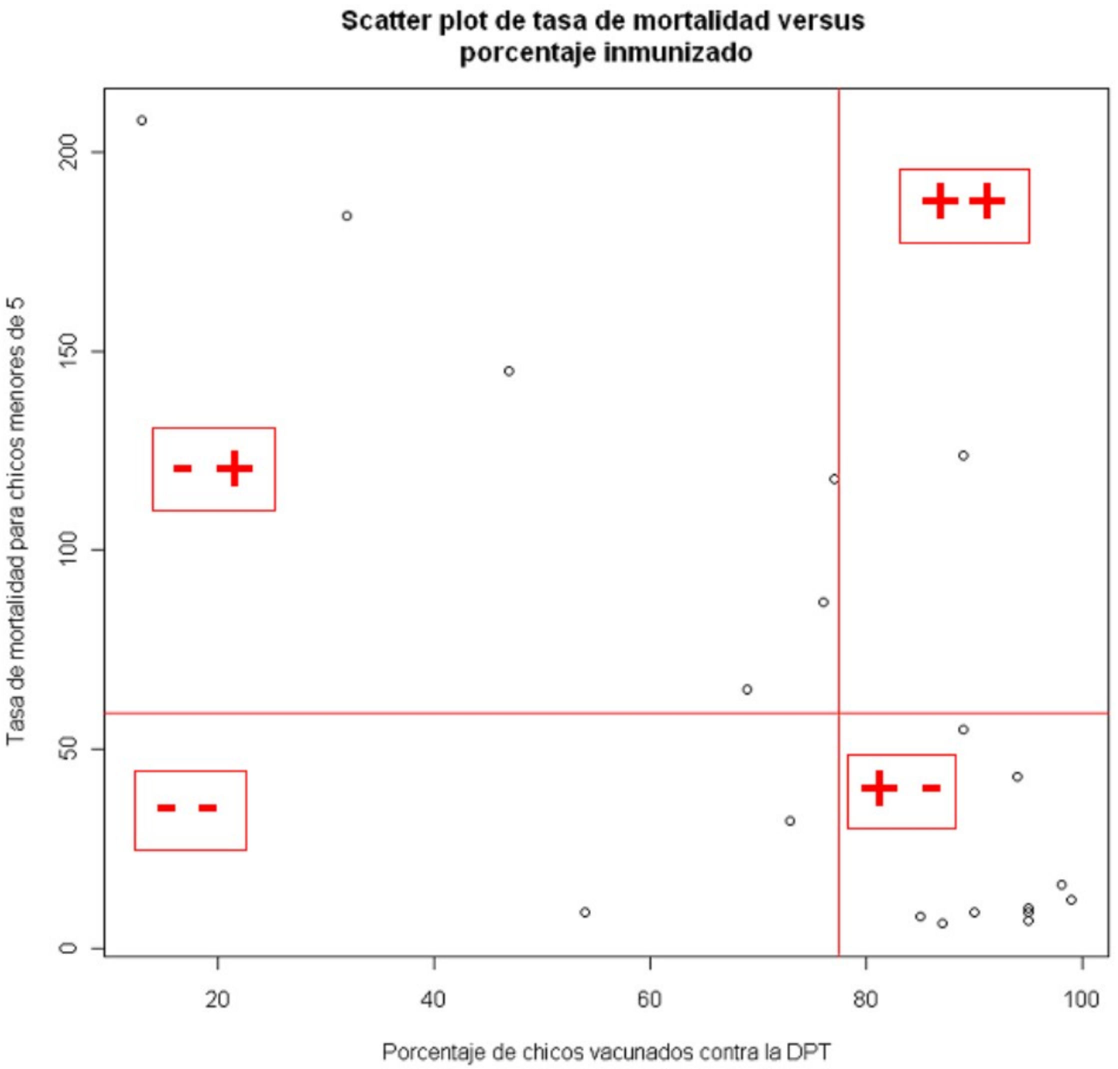
por ducto de los desvíos estandares. ¿Cómo

esti *de correlación de Pearson, o coeficiente*

de

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

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



Scatter plot de tasa de mortalidad versus
porcentaje inmunizado
Tasa de mortalidad para chicos menores de 5
Porcentaje de chicos vacunados contra la DPT
0
50
100
150
200
20
40
60
80
100
+ +
- +
- -
+ -

Test para $\rho = 0$ Los supuestos para llevar a cabo el test son que los pares de observaciones $(X_1, Y_1), \ldots, (X_n, Y_n)$ sean independientes entre sí, idénticamente distribuidos, y tengan distribución (conjunta) normal bivariada (ver la definición de esto en la Observación 1.1). En particular, esto implica que cada una de las muestras $X_1 \ldots, X_n$ e $Y_1 \ldots, Y_n$ tengan distribución normal. Si la hipótesis nula es verdadera, entonces el estadístico

$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

que no es más que $\widehat{\rho}$ dividido por un estimador de su desvío estándar, tiene distribución t de Student con $n-2$ grados de libertad, lo cual notaremos

$$T \sim t_{n-2} \text{ bajo } H_0.$$

1.3. Coeficiente de correlación de Spearman

1. Se ordena cada muestra por separado, de menor a mayor. A cada observación se le calcula el ranking que tiene (o rango, o número de observación en la muestra ordenada). De este modo, la observación más pequeña de las $X's$ recibe el número 1 como rango, la segunda recibe el número 2, etcétera, la más grande de todas las $X's$ recibirá el rango n. Si hubiera dos o más observaciones empatadas en algún puesto (por ejemplo, si las dos observaciones más pequeñas tomaran el mismo valor de X, entonces se promedian los rangos que les tocarían: cada una tendrá rango 1,5, en este ejemplo, ya que $\frac{1+2}{2} = 1{,}5$. En el caso en el que las tres primeras observaciones fueran empatadas, a las tres les tocaría el promedio entre 1, 2 y 3, que resultará ser $\frac{1+2+3}{3} = 2$). A este proceso se lo denomina *ranquear las observaciones* X. Llamemos $R(X_i)$ al rango obtenido por la i-ésima observación X.

2. Se reemplaza a cada observación X_i por su rango $R(X_i)$.

3. Se ranquean las observaciones Y, obteniéndose $R(Y_i)$ de la misma forma en que se hizo en el ítem 1 para las $X's$.

4. Se reemplaza a cada observación Y_i por su rango $R(Y_i)$. Observemos que conocemos la suma de todos los rangos de ambas muestras (es la suma de $1 + 2 + 3 + \cdots + n = \frac{n(n+1)}{2}$).

5. Se calcula el coeficiente de correlación de Pearson entre los pares $(R(X_i), R(Y_i))$. El valor obtenido es el coeficiente de correlación de Spearman, que denotaremos r_S.

2.2. Modelo lineal simple

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) propone que

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad (2)$$

donde ε es el término del error. Esto es que para cada valor de X, la correspondiente observación Y consiste en el valor $\beta_0 + \beta_1 X$ más una cantidad ε, que puede ser positiva o negativa, y que da cuenta de que la relación entre X e Y no es exactamente lineal, sino que está expuesta a variaciones individuales que hacen que el

1. La esperanza condicional de Y depende de X de manera lineal, es decir

$$E(Y \mid X) = \beta_0 + \beta_1 X \tag{6}$$

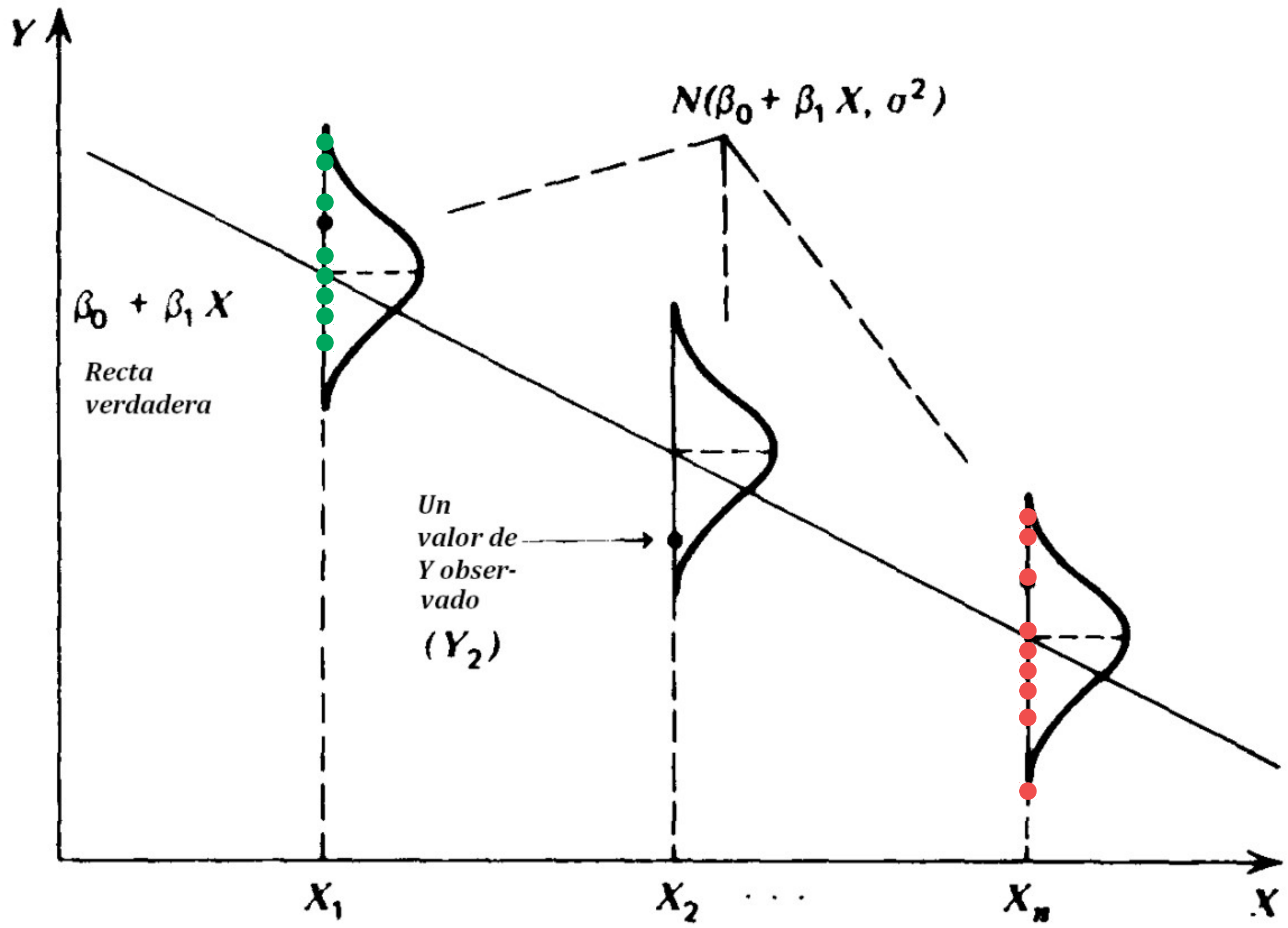
o, escrito de otro modo

$$E(Y \mid X = x_i) = \beta_0 + \beta_1 x_i \tag{7}$$

donde β_0, β_1 son los parámetros del modelo, o coeficientes de la ecuación. A la ecuación (6) se la suele llamar **función de respuesta**, es una recta.

2. La varianza de la variable respuesta Y dado que la predictora está fijada en $X = x$ la denotaremos por $Var(Y \mid X = x)$. Asumimos que satisface

$$Var(Y \mid X = x_i) = \sigma^2,$$


Y
$N(\beta_0 + \beta_1 X, \sigma^2)$
$\beta_0 + \beta_1 X$
Recta
verdadera
Un
valor de
Y obser-
vado
(Y_2)
X_1
X_2 . . .
X_n
X

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

2.5. Estimación de los parámetros β_0 y β_1

Los coeficientes del modelo se estiman a partir de la muestra aleatoria de n observaciones (X_i, Y_i) con $1 \leq i \leq n$. Llamaremos $\widehat{\beta}_0$ y $\widehat{\beta}_1$ a los estimadores de β_0 y β_1. Los valores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ corresponderán a la recta de ordenada al origen $\widehat{\beta}_0$ y pendiente $\widehat{\beta}_1$ que "mejor ajuste" a los datos $(X_1, Y_1), \ldots, (X_n, Y_n)$ observados. Para encontrarlos, debemos dar una noción de bondad de ajuste de una recta cualquiera con ordenada al origen a y pendiente b a nuestros datos. Tomemos las distancias verticales entre los puntos observados (X_i, Y_i) y los puntos que están sobre la recta $y = a + bx$, que están dados por los pares $(X_i, a + bX_i)$. La distancia entre ambos es $Y_i - (a + bX_i)$. Tomamos como función que mide el desajuste de la recta a los datos a

$$g(a, b) = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2, \tag{8}$$

$$\frac{\partial g(a,b)}{\partial a} = \sum_{i=1}^{n} 2\left(Y_i - (a + bX_i)\right)(-1)$$
$$\frac{\partial g(a,b)}{\partial b} = \sum_{i=1}^{n} 2\left(Y_i - (a + bX_i)\right)(-X_i)$$

Las igualamos a cero para encontrar $\widehat{\beta}_0$ y $\widehat{\beta}_1$, sus puntos críticos. Obtenemos

$$\sum_{i=1}^{n} \left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right) = 0 \tag{9}$$
$$\sum_{i=1}^{n} \left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i\right)\right) X_i = 0. \tag{10}$$

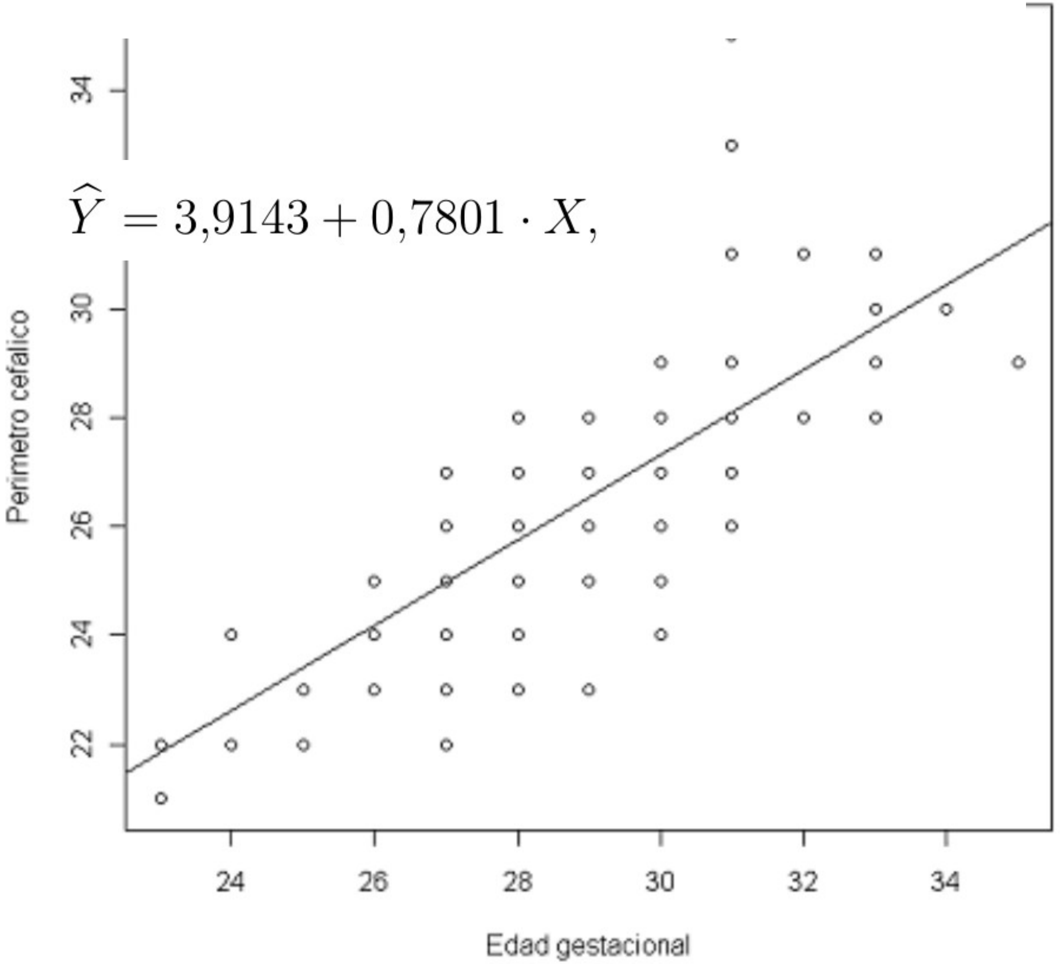
Las dos ecuaciones anteriores se denominan las *ecuaciones normales* para regresión lineal. Despejamos de ellas las estimaciones de los parámetros que resultan ser

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}, \tag{11}$$
$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}. \tag{12}$$

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i, \qquad \text{Modelo ajustado.}$$

```
> ajuste<-lm(headcirc~gestage)
> summary(ajuste)
  Coefficients:
              Estimate Std. Error
  (Intercept)  3.91426    1.82915
  gestage      0.78005    0.06307
  ---
  Signif. codes:  0 ‘***’ 0.001 ’*

  Residual standard error: 1.59 on
  Multiple R-squared:  0.6095,
  F-statistic: 152.9 on 1 and 98 D
```

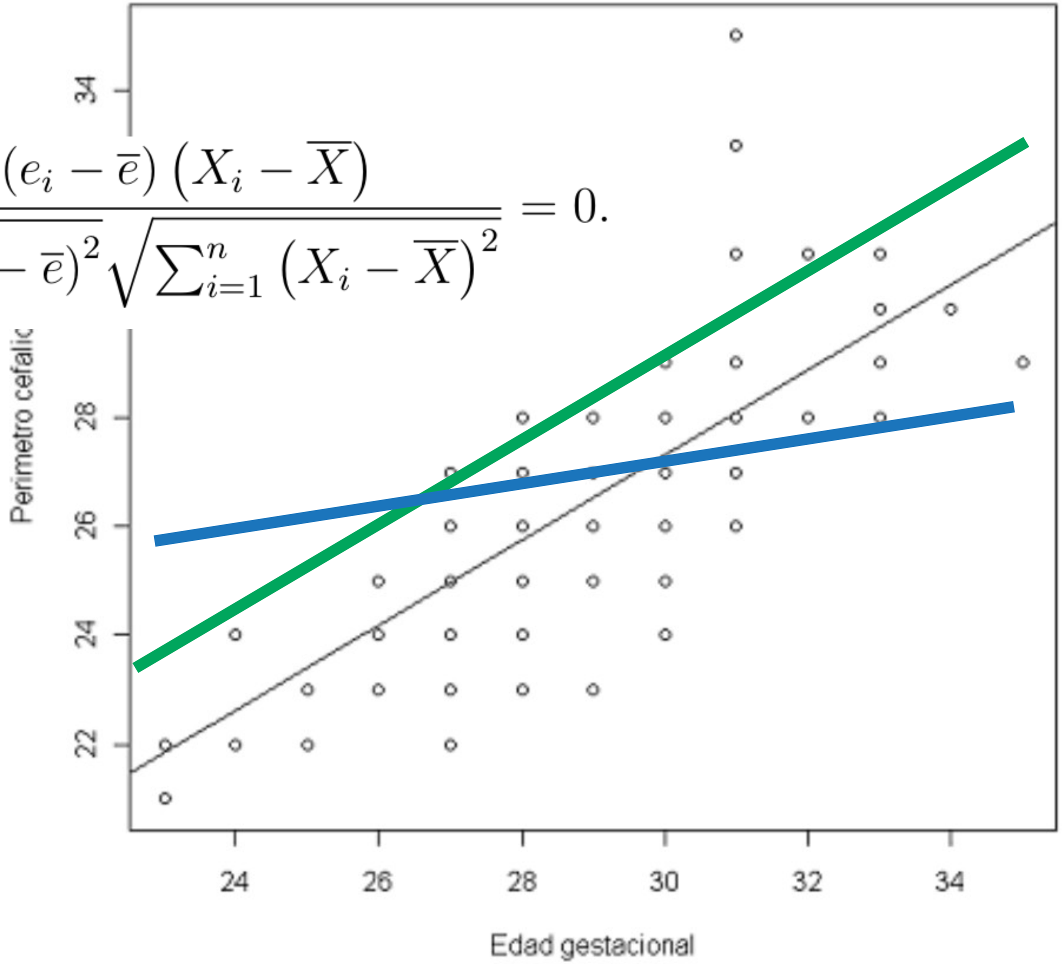
$$\widehat{Y} = 3{,}9143 + 0{,}7801 \cdot X,$$



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

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

$$r = r\left((X_1, e_i), \ldots, (X_n, e_n) \right) = \frac{\sum_{i=1}^{n} (e_i - \overline{e}) \left(X_i - \overline{X} \right)}{\sqrt{\sum_{i=1}^{n} (e_i - \overline{e})^2} \sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}} = 0.$$



2.4. Supuestos del modelo lineal

Los supuestos bajo los cuales serán válidas las inferencias que haremos más adelante sobre el modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{4}$$

son los siguientes:

$$\boxed{\varepsilon_i \sim N\left(0, \sigma^2\right), \quad 1 \leq i \leq n, \quad \text{independientes entre sí.}}$$

1. los ε_i tiene media cero, $E\left(\varepsilon_i\right) = 0$.

2. los ε_i tienen todos la misma varianza desconocida que llamaremos σ^2 y que es el otro parámetro del modelo, $Var\left(\varepsilon_i\right) = \sigma^2$. A este requisito se lo suele llamar *homoscedasticidad*.

3. los ε_i tienen distribución normal

4. los ε_i son independientes entre sí, y son no correlacionados con las X_i.

$$, Y \mid X = x_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} = \frac{1}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} \sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i \\
&= \sum_{i=1}^n \frac{\left(X_i - \overline{X}\right)}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} Y_i = \sum_{i=1}^n c_i Y_i \qquad (20)
\end{aligned}$$

donde

$$\begin{aligned}
c_i &= \frac{\left(X_i - \overline{X}\right)}{\sum_{j=1}^n \left(X_j - \overline{X}\right)^2} = \frac{\left(X_i - \overline{X}\right)}{S_{XX}}, \qquad (21) \\
S_{XX} &= \sum_{j=1}^n \left(X_j - \overline{X}\right)^2.
\end{aligned}$$

$$\begin{aligned}
E\left(\widehat{\beta}_1\right) &= \beta_1 \\
Var\left(\widehat{\beta}_1\right) &= \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}.
\end{aligned}$$

2.9. Inferencia sobre β_1

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right).$$

2.10. Inferencia sobre β_0

Esta inferencia despierta menos interés que la de β_1. Aunque los paquetes estadísticos la calculan es infrecuente encontrarla en aplicaciones. Bajo los supuestos del modelo lineal, puede calcularse la esperanza y varianza del estimador de β_0, que resultan ser

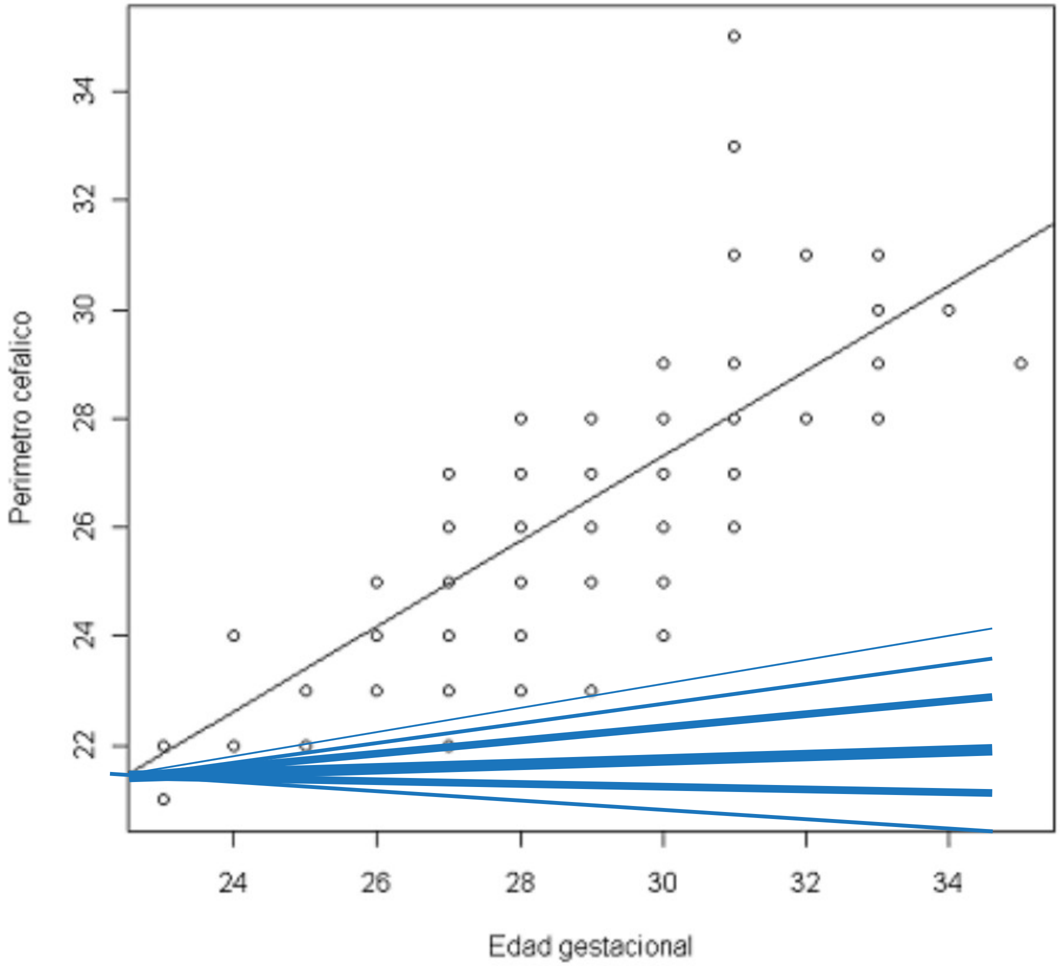
$$\begin{aligned} E\left(\widehat{\beta}_0\right) &= \beta_0 \\ Var\left(\widehat{\beta}_0\right) &= \sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{j=1}^n\left(X_j-\overline{X}\right)^2}\right). \end{aligned}$$

Nuevamente, las conclusiones son condicionales a los valores de los X‘s observados. La varianza puede estimarse por

$$\widehat{Var}\left(\widehat{\beta}_0\right)=\widehat{\sigma}^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{j=1}^n\left(X_j-\overline{X}\right)^2}\right)$$

Nuevamente, el estadístico $\widehat{\beta}_0$ tiene distribución normal, su distribución es $N\left(\beta_0,\sigma^2\left(\frac{1}{n}+\frac{\overline{X}^2}{\sum_{j=1}^n\left(X_j-\overline{X}\right)^2}\right)\right)$, luego

¿ Qué hago con la distribución del estadístico ?



$$E\left(\widehat{\beta}_1\right) = \beta_1$$

$$Var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right).$$

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right).$$

Un estimador de la varianza es

$$\widehat{Var}\left(\widehat{\beta}_1\right) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} = \frac{\text{SSRes}/\left(n-2\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}.$$

Finalmente, bajo los supuestos del modelo, puede probarse que

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - \beta_1}{se_{\widehat{\beta}_1}}$$

tiene distribución *t de Student con $n-2$ grados de libertad* si los datos siguen el modelo lineal, donde

$$se_{\widehat{\beta}_1} = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}} = \sqrt{\frac{\text{SSRes}/\left(n-2\right)}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}}.$$

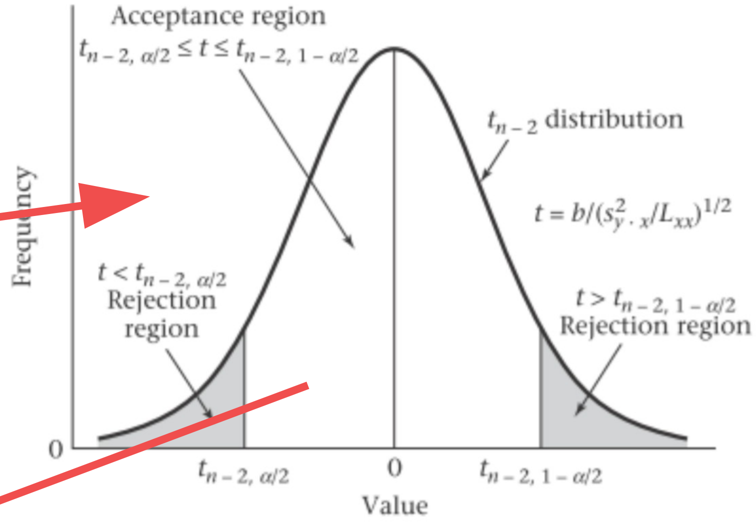
El estimador de σ^2 que usaremos será

$$\widehat{\sigma}^2 = \frac{1}{n-2}\text{SSRes} = \text{MSRes}. \qquad \widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \left(e_i - \overline{e}\right)^2 = \frac{1}{n-2}\sum_{i=1}^n e_i^2 = \frac{1}{n-2}\sum_{i=1}^n \left(Y_i - \widehat{Y}_i\right)^2. \tag{17}$$

Intervalos de Confianza

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - \beta_1}{se_{\widehat{\beta}_1}}$$

. *t de Student con $n-2$ grados de libertad*



Con esta distribución podemos construir un intervalo de confianza de nivel $1-\alpha$ para β_1 que resultará

$$\widehat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}, \quad \text{o bien} \tag{18}$$

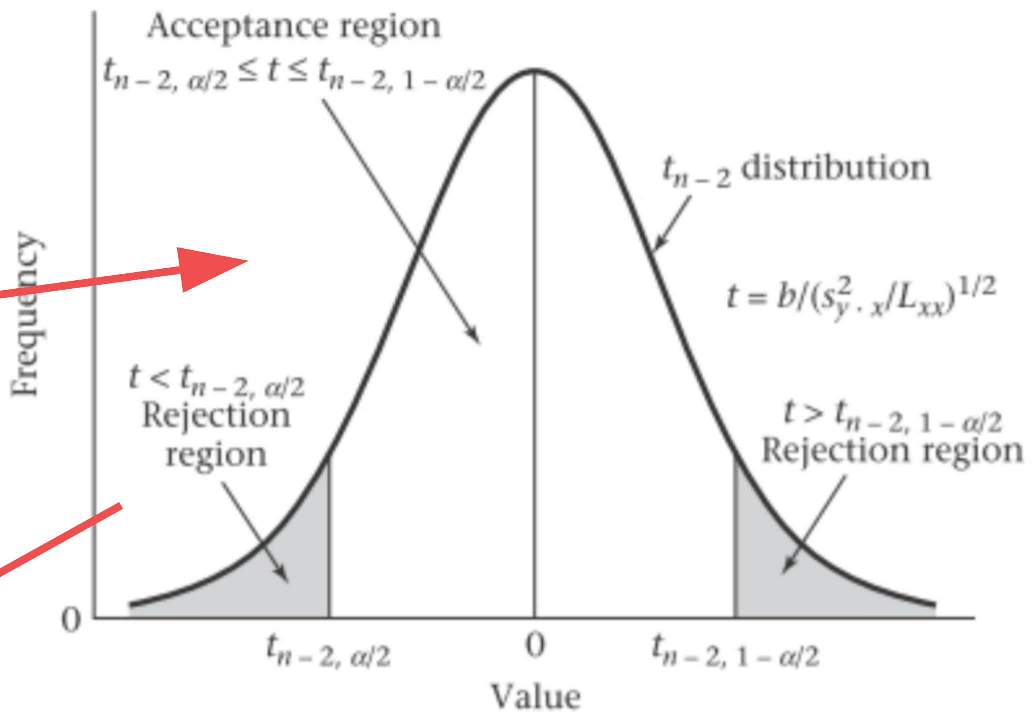
$$\widehat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \cdot se_{\widehat{\beta}_1}$$

Test de Hipótesis

$$H_0 \quad : \quad \beta_1 = b$$
$$H_1 \quad : \quad \beta_1 \neq b.$$

$$T_{obs} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - \beta_1}{se_{\widehat{\beta}_1}}$$

t de Student con $n-2$ grados de libertad



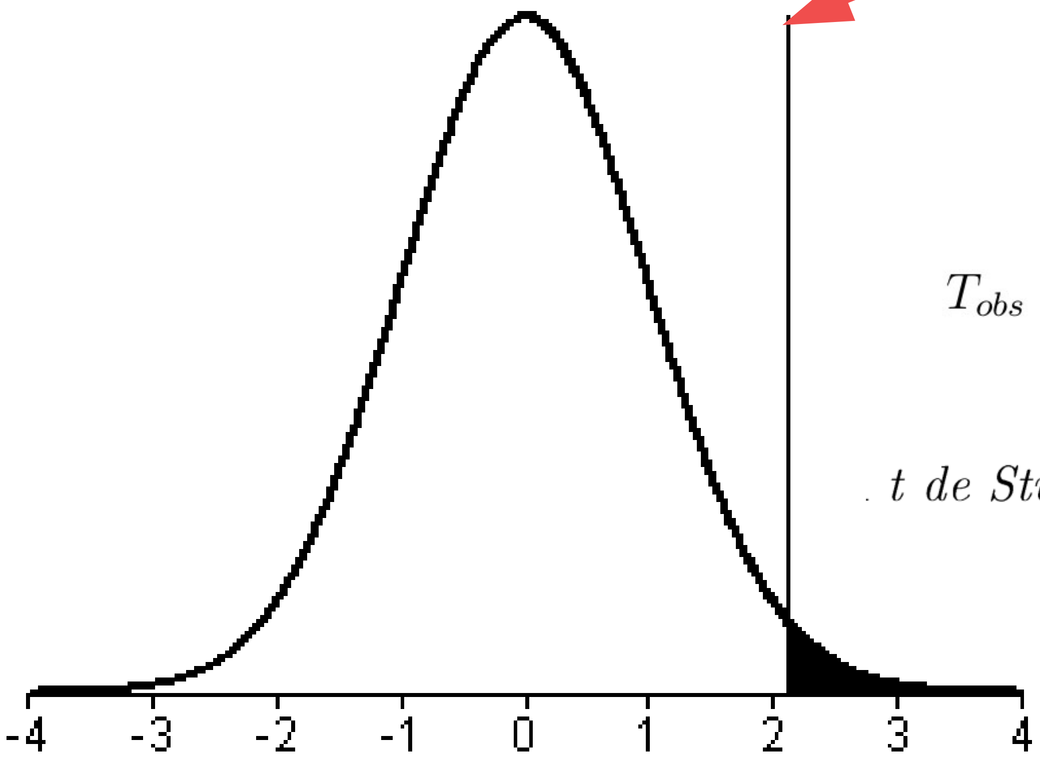
el test rechaza H_0 con nivel α si

$$T_{obs} \leq t_{n-2,\frac{\alpha}{2}} \text{ ó } t_{n-2,1-\frac{\alpha}{2}} \leq T_{obs},$$

P-valor

$$p - valor = 2P\left(T \geq |T_{obs}|\right),$$

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_1 &: \beta_1 \neq b. \end{aligned}$$



$$T_{obs} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - \beta_1}{se_{\widehat{\beta}_1}}$$

t de Student con n − 2 grados de libertad

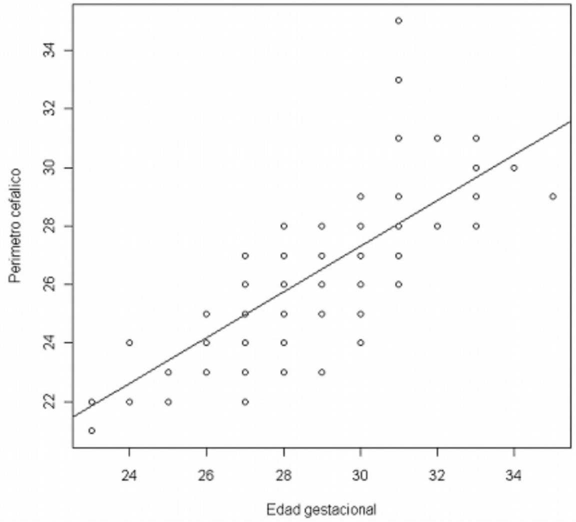
Test, Intervalo y p-valor en el Ejemplo

```
Call:
lm(formula = headcirc ~ gestage)

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
```

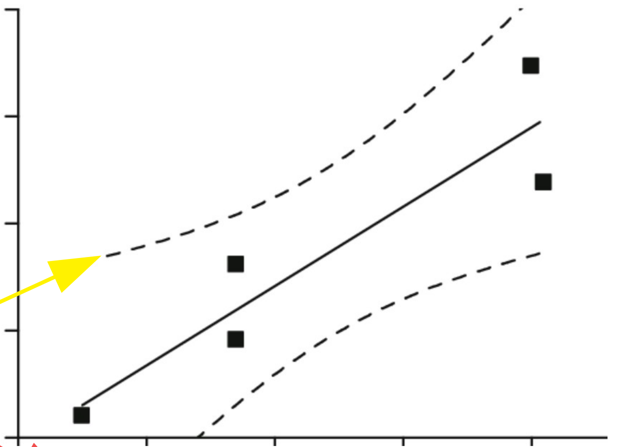
$$\widehat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \cdot se_{\widehat{\beta}_1}$$

$$0{,}7801 \pm 1{,}984467 \cdot 0{,}06307941$$

$$[0{,}654921, 0{,}905279]$$



Inferencia sobre E(Y|X=x)



$$E\left(Y_h \mid X = x_h\right) = \beta_0 + \beta_1 x_h.$$

$$\widehat{Y}_h = \widehat{\beta}_0 + \widehat{\beta}_1 x_h.$$

$$\begin{aligned} E\left(\widehat{Y}_h\right) &= E\left(Y_h\right) \\ Var\left(\widehat{Y}_h\right) &= \sigma^2 \cdot \left[\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]. \end{aligned}$$

$$\widehat{Y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}. \qquad (24)$$

Inferencia sobre Y

$$Y_h = \beta_0 + \beta_1 X_h + \varepsilon_h$$

$$\widehat{Y}_h = \widehat{\beta}_0 + \widehat{\beta}_1 x_h.$$

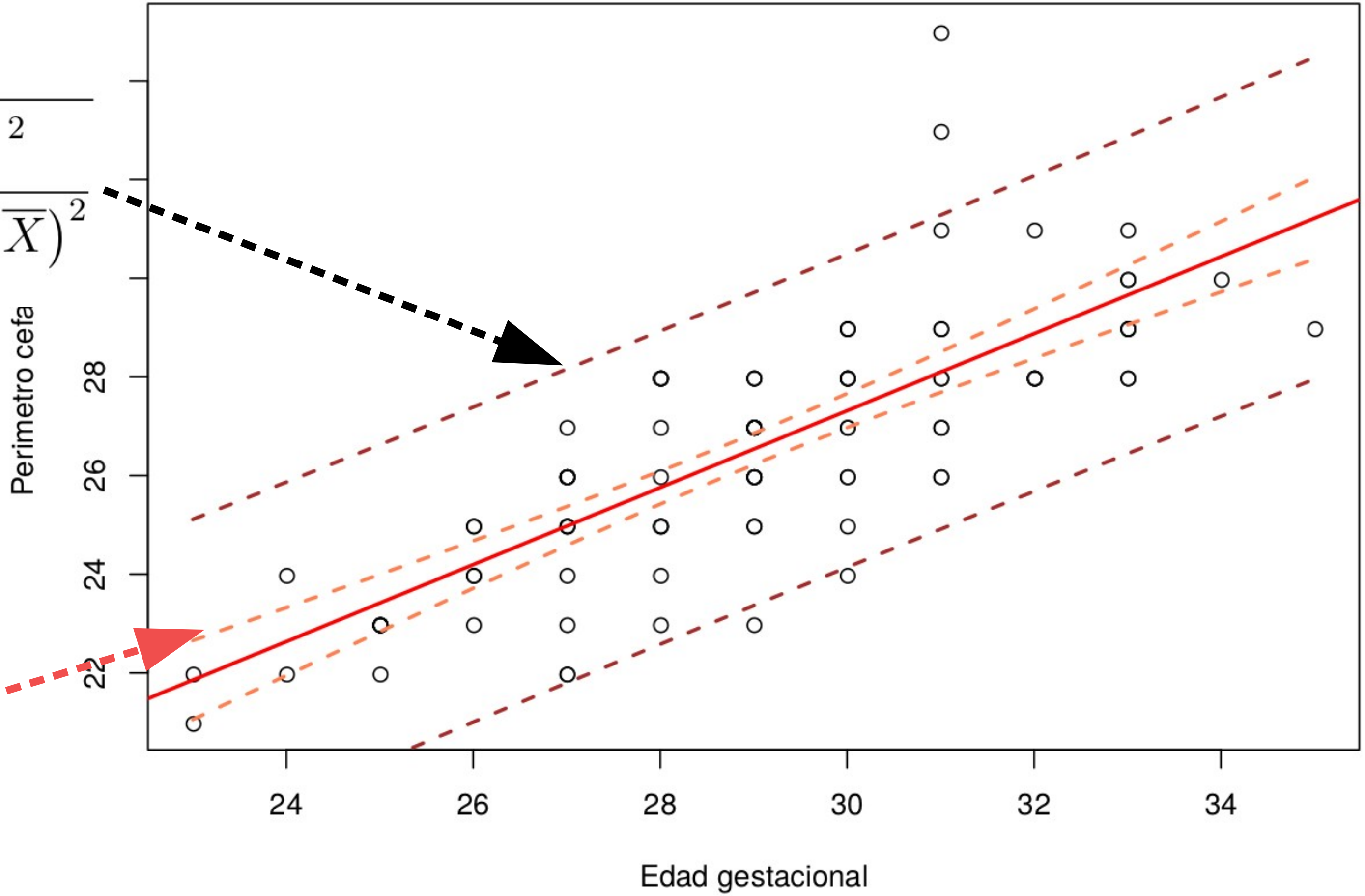
$$\begin{aligned} E\left(\widehat{Y}_h\right) &= E\left(Y_h\right) \\ Var\left(\widehat{Y}_h\right) &= \sigma^2\left[1 + \frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]. \end{aligned}$$

$$\widehat{Y}_h \pm t_{n-2;1-\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

Predecir Y versus E(Y|X=x)

$$\hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}$$

$$\hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}$$

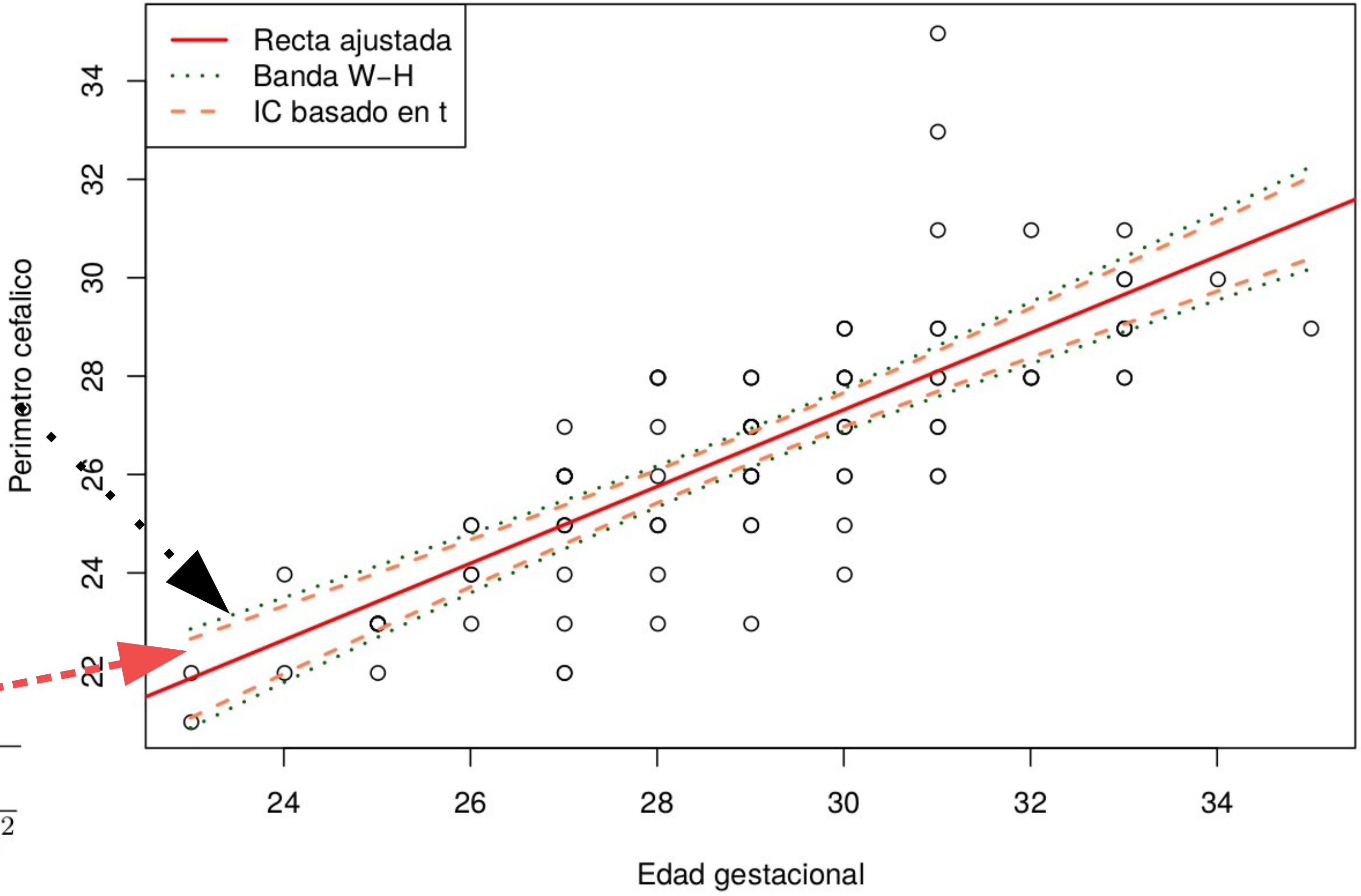


Nivel Simultaneo

$$\widehat{Y}_h \pm W \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_h - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}.$$

$$W = \sqrt{2F_{1-\alpha,2,n-2}},$$

$$\widehat{Y}_h \pm t_{n-2;\frac{\alpha}{2}} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\left(x_h - \overline{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}}$$



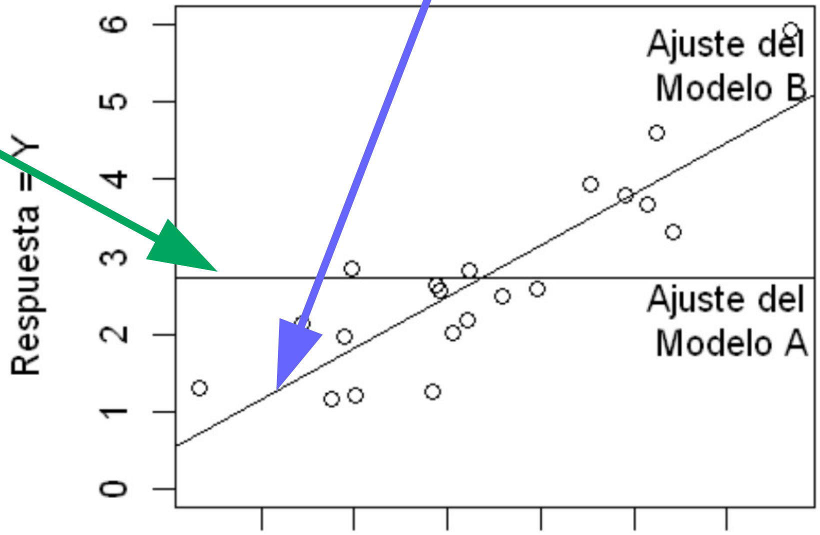
Análisis de la Varianza

Modelo A: $E(Y \mid X) = \mu$, o escrito de otro modo

Modelo A: $Y_i = \mu + u_i$ con $u_i \sim N\left(0, \sigma_Y^2\right)$, $1 \leq i \leq n$, independientes entre sí.

Modelo B: $E(Y \mid X) = \beta_0 + \beta_1 X$, o escrito de otro modo

Modelo B: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, con $\varepsilon_i \sim N\left(0, \sigma^2\right)$, $1 \leq i \leq n$, independientes entre sí.



$$\underbrace{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}_{\text{SSTo}} = \underbrace{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}_{\text{SSRes}} + \underbrace{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}_{\text{SSReg}}$$

Análisis de la Varianza: Demostración

$$\overbrace{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}^{\text{SSTo}} = \overbrace{\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2}^{\text{SSRes}} + \overbrace{\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2}^{\text{SSReg}}$$

Proof [edit]

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\bar{y}+\hat{y}_i-\hat{y}_i)^2 = \sum_{i=1}^{n}((\hat{y}_i-\bar{y})+\underbrace{(y_i-\hat{y}_i}_{\hat{\varepsilon}_i}$$

$$= \sum_{i=1}^{n}((\hat{y}_i-\bar{y})^2+2\hat{\varepsilon}_i(\hat{y}_i-\bar{y})+\hat{\varepsilon}_i^2)$$

$$= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2+2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i-\bar{y})$$

$$= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2+2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_px_{ip}-\bar{y})$$

$$= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2+2(\hat{\beta}_0-\bar{y})\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i}_{0}+2\hat{\beta}_1\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_ix_{i1}}_{0}+\cdots+2\hat{\beta}_p\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_ix_{ip}}_{0}$$

$$= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \mathrm{ESS}+\mathrm{RSS}$$

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

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

Tabla de Anova

Tabla 14: Tabla de ANOVA para el modelo de Regresión Lineal Simple

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Regresión	SSReg	1	MSReg	$\frac{\text{MSReg}}{\text{MSRes}}$	$P\left(F_{1,n-2} \geq F_{obs}\right)$
Residuos	SSRes	$n-2$	MSRes		
Total	SSTo	$n-1$			

$$\text{MSReg} = \frac{\text{SSReg}}{1}$$

$$\text{MSRes} = \frac{\text{SSRes}}{n-2}$$

$$F = \frac{\text{MSReg}}{\text{MSRes}} = \frac{\text{SSReg}(n-2)}{\text{SSRes}}$$

$$\underbrace{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}^{\text{SSTo}} = \overbrace{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}^{\text{SSRes}} + \overbrace{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}^{\text{SSReg}}$$

Coeficiente de Determinación

$$100\,\% \text{ de variabilidad} \quad \text{———} \quad \text{SSTo}$$

$$\% \text{ de variabilidad explicada} \quad \text{———} \quad \text{SSTo} - \text{SSRes}$$

Luego el porcentaje de variabilidad explicada es

$$\frac{\text{SSTo} - \text{SSRes}}{\text{SSTo}} \times 100\,\%.$$

A la cantidad

$$\frac{\text{SSTo} - \text{SSRes}}{\text{SSTo}} = \frac{\text{SSReg}}{\text{SSTo}}$$

se la denomina R^2, o **coeficiente de determinación**

- $0 \leq R^2 \leq 1$
- No depende de las unidades de medición.
- Es el cuadrado del coeficiente de correlación de Pearson para la muestra $\{(X_i, Y_i)\}_{1 \leq i \leq n}$. También es el cuadrado del coeficiente de correlación de Pearson para los pares $\left\{\left(\widehat{Y}_i, Y_i\right)\right\}_{1 \leq i \leq n}$, es decir, entre los valores de la covariable observados y los predichos por el modelo lineal.
- Mientras mayor es R^2 mayor es la fuerza de la variable regresora (X) para predecir a la variable respuesta (Y).
- Mientras mayor sea R^2 menor es la SSRes y por lo tanto, más cercanos están los puntos a la recta.
- Toma el mismo valor cuando usamos a X para predecir a Y o cuando usamos a Y para predecir a X.

Leverage

El valor predicho de un dato puede escribirse como combinación lineal de las observaciones

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i = \sum_{k=1}^{n} h_{ik} Y_k \tag{33}$$

donde

$$h_{ik} = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)\left(X_k - \overline{X}\right)}{S_{XX}}$$

$$S_{XX} = \sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2$$

y como caso particular tenemos que

$$\begin{aligned}
\sum_{k=1}^{n} h_{ik} &= 1, \quad \sum_{i=1}^{n} h_{ik} = 1 \\
\sum_{i=1}^{n} h_{ii} &= 2 \\
\frac{1}{n} &\leq h_{ii} \leq \frac{1}{s} \leq 1.
\end{aligned}$$

$$h_{ii} = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{S_{XX}}. \tag{34}$$

Residuos $e_i = Y_i - \widehat{Y}_i \approx \varepsilon_i$

3.1.2. Residuos

Dijimos en la Sección 2.8 que los residuos son cantidades observables, que representan de alguna manera el correlato empírico de los errores. Para verificar los supuestos del modelo lineal, suelen usarse métodos gráficos que involucran a los residuos. El modelo lineal

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

supone que los errores ε tienen media poblacional cero y varianza constante (que denominamos σ^2), y que son indendientes para distintas observaciones. Sin embargo, ya hemos visto que no ocurre lo mismo con los residuos. Vimos que los residuos no son independientes. Además, puede probarse que

$$\begin{aligned} E(e_i) &= 0 \\ Var(e_i) &= \sigma^2 (1 - h_{ii}) \end{aligned} \tag{37}$$

$$h_{ii} = \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{S_{XX}}$$

Residuos Estandarizados

3.1.3. Residuos estandarizados

Para hacer más comparables a los residuos entre sí, podemos dividir a cada uno de ellos por un estimador de su desvío estándar, obteniendo lo que se denominan *residuos estandarizados*:

$$rest_i = \frac{e_i}{\sqrt{\widehat{\sigma}^2 (1 - h_{ii})}}. \tag{38}$$

Recordemos que el estimador de σ^2 bajo el modelo de regresión está dado por

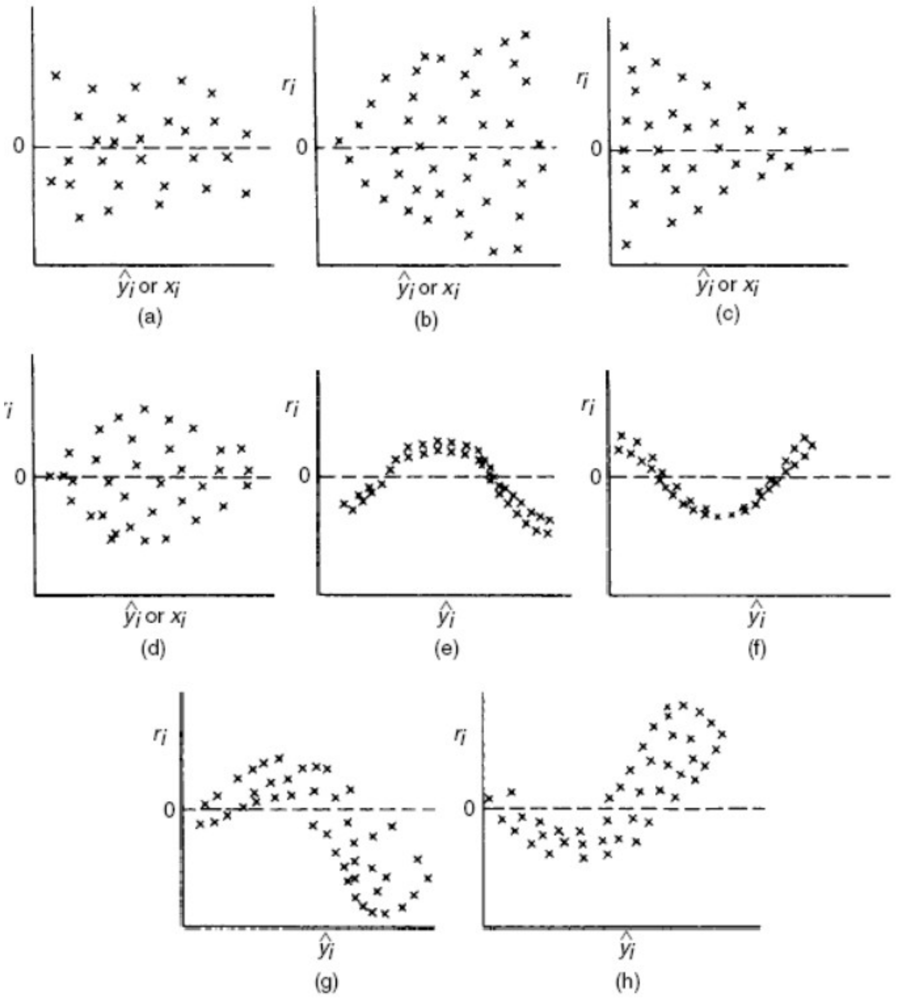
$$\widehat{\sigma}^2 = \frac{SS\text{Res}}{n-2}$$

$$\boxed{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}$$

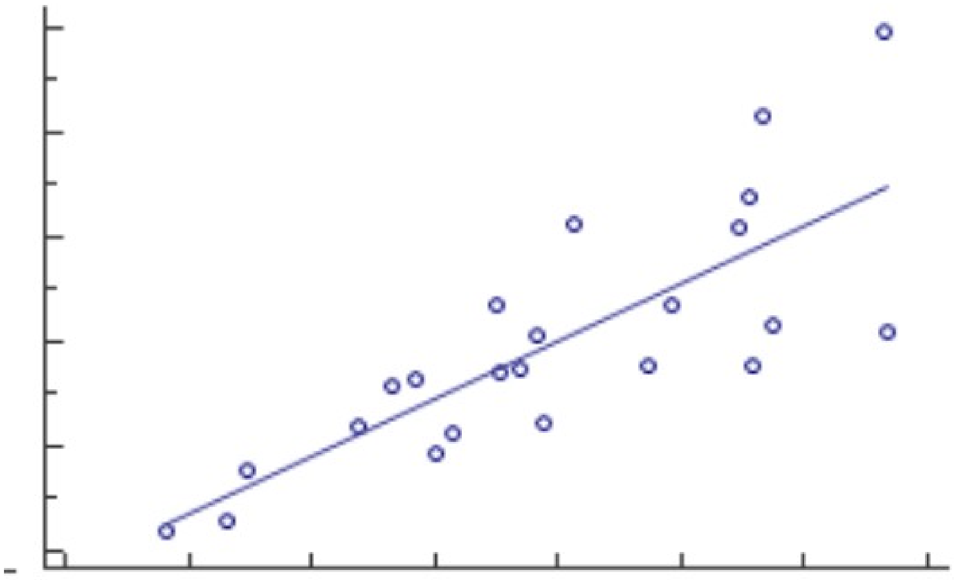
Puede probarse que los residuos estandarizados tienen media poblacional cero (igual que los residuos), y varianza poblacional igual a uno, es decir

$$\begin{aligned} E(rest_i) &= 0 \\ Var(rest_i) &= 1, \quad \text{para todo } i. \end{aligned}$$

La Estructura de los Residuos



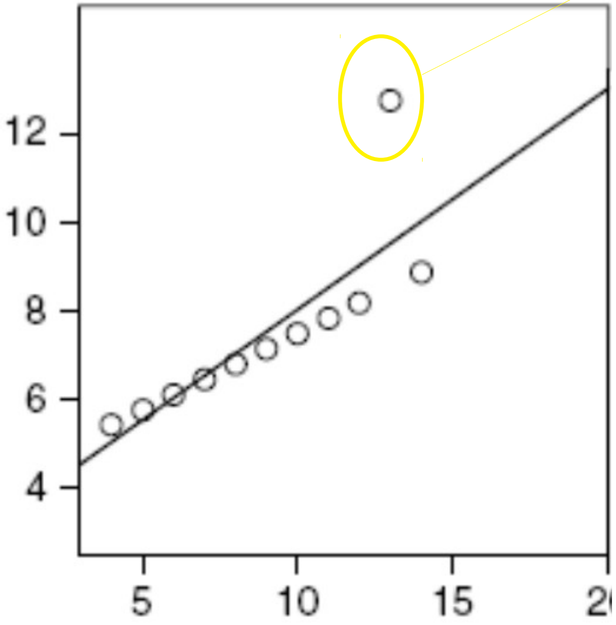
Heterocedasticidad



$$Var\,(Y \mid X = x_i) = Var\,(\varepsilon_i) = \frac{\sigma^2}{w_i}$$

$$g_{wls}\,(a, b) = \sum_{i=1}^{n} w_i\,(Y_i - (a + bX_i))^2\,.$$

Outliers



$$rest_i = \frac{e_i}{\sqrt{\widehat{\sigma}^2\left(1-h_{ii}\right)}}.$$

1. Eliminamos esa observación de la muestra, de modo que ahora tenemos una muestra con $n-1$ casos.

2. Usando el conjunto de datos reducidos volvemos a estimar los parámetros, obteniendo $\widehat{\beta}_{0(i)}, \widehat{\beta}_{1(i)}$ y $\widehat{\sigma}^2_{(i)}$ donde el subíndice (i) está escrito para recordarnos que los parámetros fueron estimados sin usar la i-ésima observación.

3. Para el caso omitido, calculamos el valor ajustado $\widehat{Y}_{i(i)} = \widehat{\beta}_{0(i)} + \widehat{\beta}_{1(i)} X_i$. Como el caso i-ésimo no fue usado en la estimación de los parámetros, Y_i y $\widehat{Y}_{i(i)}$ son independientes. La varianza de $Y_i - \widehat{Y}_{i(i)}$ puede calcularse y se estima usando $\widehat{\sigma}^2_{(i)}$.

$$t_i = \frac{Y_i - \widehat{Y}_{i(i)}}{\sqrt{\widehat{Var}\left(Y_i - \widehat{Y}_{i(i)}\right)}}$$

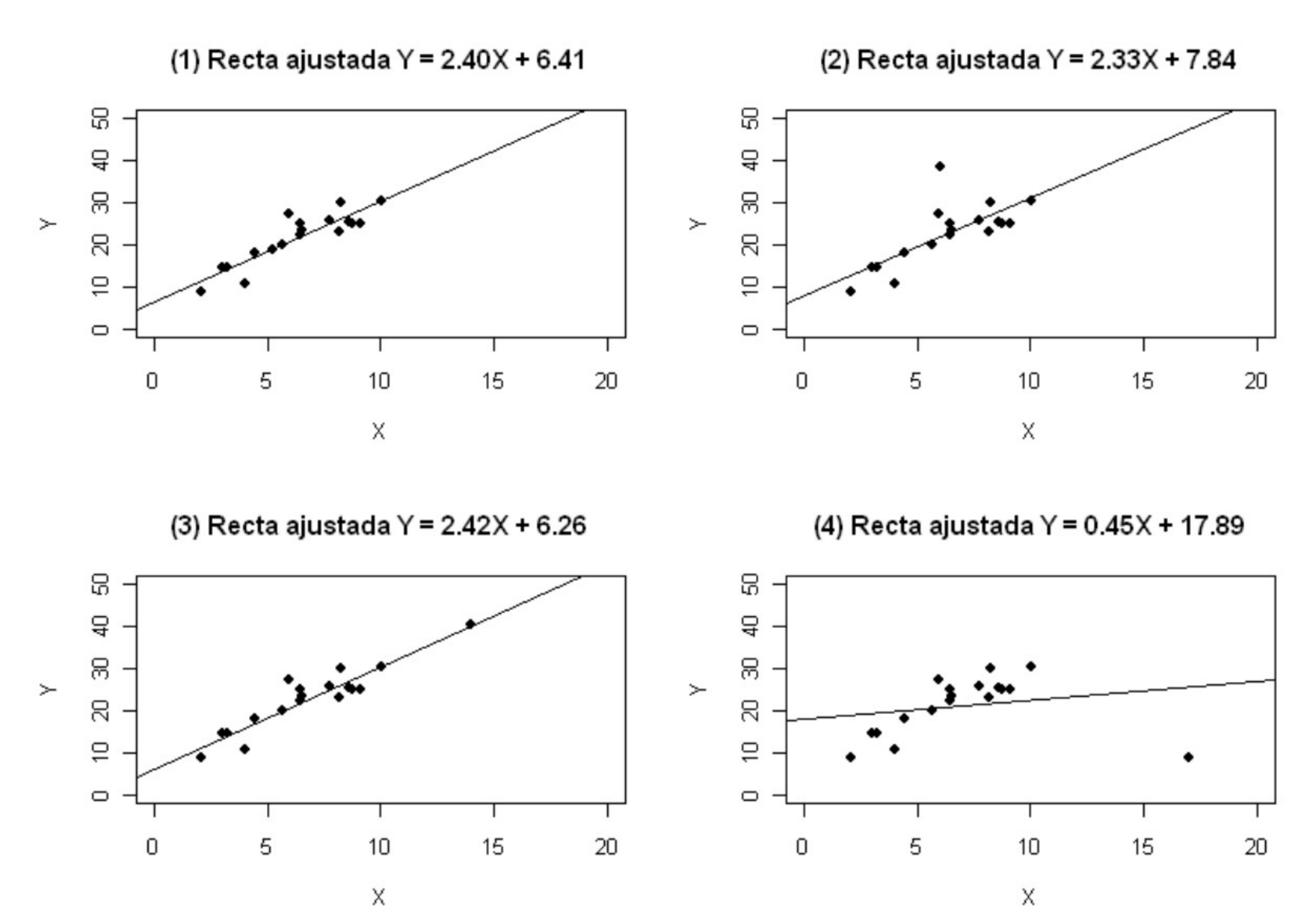
4. Escribamos

$$t_i = \frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = rest_i \sqrt{\frac{n-3}{n-2-rest_i}}$$

$$t_i \sim t_{n-3}$$

la versión estandarizada del estadístico en consideración. Si la observación i-ésima sigue el modelo, entonces la esperanza de $Y_i - \widehat{Y}_{i(i)}$ debería ser cero. Si no lo sigue, será un valor no nulo. Luego, si llamamos δ a la esperanza poblacional de esa resta, $\delta = E\left(Y_i - \widehat{Y}_{i(i)}\right)$, y asumimos normalidad de los errores, puede probarse que la distribución de t_i bajo la hipótesis $H_0 : \delta = 0$ es una t de Student con $n-3$ grados de libertad, $t_i \sim t_{n-3}$ (recordar que

Observaciones Influyentes



Distancia de Cook

$$D_i = \frac{\left(\widehat{Y}_{(i)i} - \widehat{Y}_i\right)^2}{2\widehat{\sigma}^2},$$

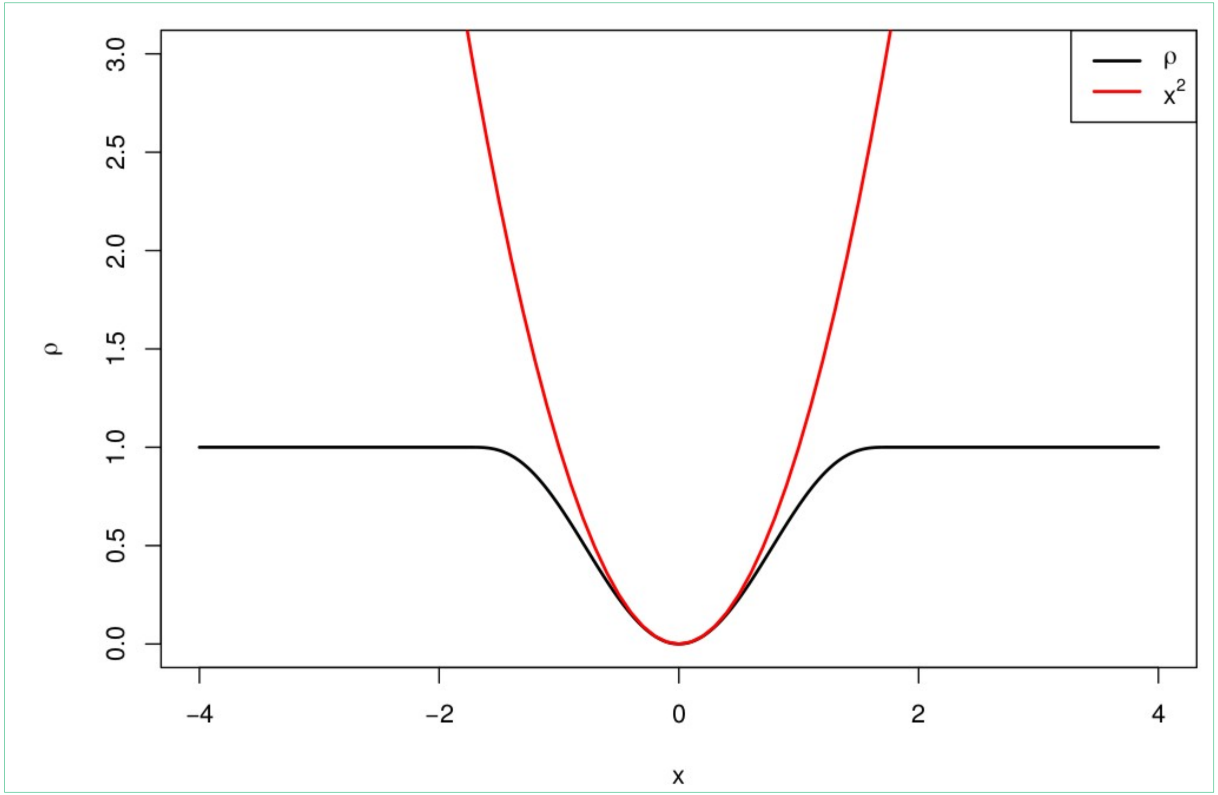
$$D_i = \frac{1}{2}\left(rest_i\right)^2 \frac{h_{ii}}{1 - h_{ii}}.$$

- Si $D_i <$ percentil 0,20 de la distribución $F_{2,n-2}$ entonces la observación no es influyente.
- Si $D_i >$ percentil 0,50 de la distribución $F_{2,n-2}$ entonces la observación es muy influyente y requerirá tomar alguna medida.
- Si D_i se encuentra entre el percentil 0,20 y el percentil 0,50 de la distribución $F_{2,n-2}$ se sugiere mirar además otros estadísticos.

Regresión Robusta

$$g\left(a,b\right) = \sum_{i=1}^{n} \left(Y_i - \left(a + bX_i\right)\right)^2 ,$$

$$g\left(a,b\right) = \sum_{i=1}^{n} \rho\left(\frac{Y_i - \left(a + bX_i\right)}{s_n}\right)$$



OLS is BLUE

$$y_i = \sum_{j=1}^{K} \beta_j X_{ij} + \varepsilon_i \quad \forall i = 1, 2, \ldots, n$$

$$\sum_{j=1}^{K} \lambda_j \beta_j$$

- They have mean zero: $\mathrm{E}[\varepsilon_i] = 0.$
- They are homoscedastic, that is all have the same finite variance: $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$ for all i and
- Distinct error terms are uncorrelated: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j.$

$$\widehat{\beta}_j = c_{1j} y_1 + \cdots + c_{nj} y_n$$

$$\mathrm{E}\left[\left(\sum_{j=1}^{K} \lambda_j \left(\widehat{\beta}_j - \beta_j\right)\right)^2\right]$$

$$\mathrm{E}\left[\widehat{\beta}_j\right] = \beta_j$$

minimizes the **sum of squares**

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{K} \widehat{\beta}_j X_{ij}\right)^2$$