Regresión lineal simple

Organización

Estructura del tema

- Planteamiento del problema. Ejemplos.
- El modelo de regresión lineal simple.
- Recta de regresión de mínimos cuadrados.
- Estimadores de los parámetros del modelo.
- Análisis de los supuestos del modelo
- Intervalos de confianza y contrastes de hipótesis para los parámetros del modelo.
- Predicción.

Planteo del problema

- Estudiar si existe un comportamiento lineal entre dos variables numéricas.
- Ver si una de las variables (a la que llamaremos variable independiente) nos ayuda a predecir o explicar el comportamiento de otra variable (que llamaremos dependiente).

Ejemplos

- Por ejemplo, podríamos interesarnos en:
  - Estudiar cómo cambia el peso de una persona de acuerdo a su edad.
  - Estudiar cómo cambia el peso de un bebé al nacer según el tiempo de gestación.
  - Estudiar cómo cambia la esperanza de vida de un país en función de su producto bruto interno.
  - Predecir el rendimiento que tiene un auto en función de la velocidad a la que circula.
  - Predecir el peso de un bebé recién nacido si se conoce cuánto mide.

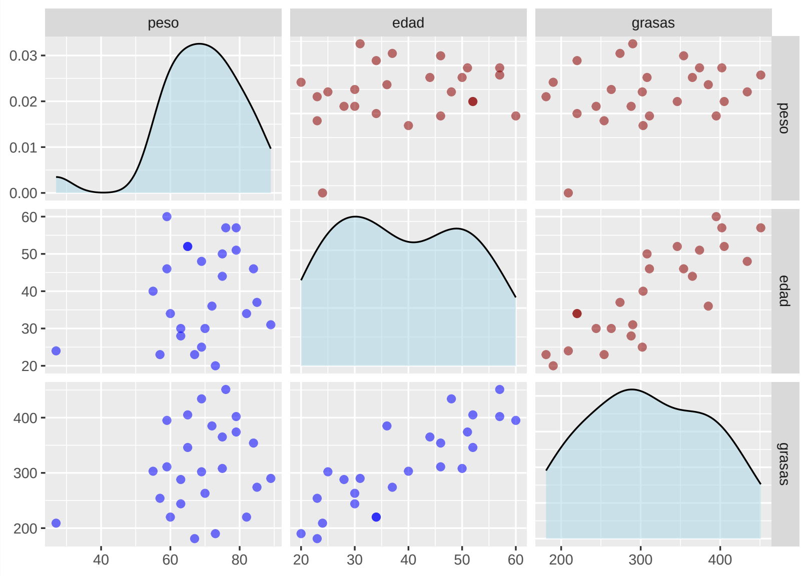
El modelo

- Supone la existencia de una relación lineal entre las variables.
- Se destaca por su simplicidad y facilidad de interpretación.
- Puede adaptarse mediante transformaciones de las variables involucradas.
- Su estudio es clave como base para comprender modelos estadísticos más complejos.

Ejemplo: archivo Grasas.csv

- El archivo *Grasas.csv* corresponden a mediciones de la edad, el peso y la cantidad de grasas en sangre, realizadas a 25 personas.
- La siguiente figura muestra un gráfico de dispersión (scatter plot) entre dichas variables.

Ejemplo: archivo Grasas.csv

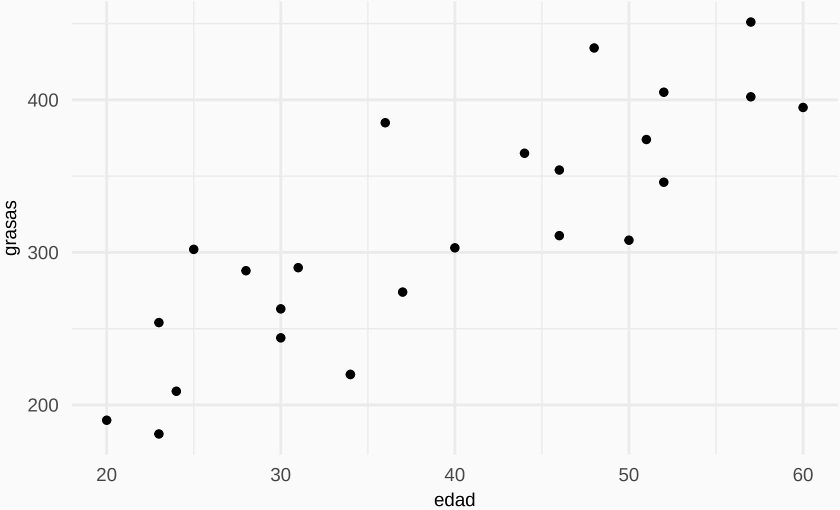


Preguntas que nos podemos hacer

- ¿Qué podemos decir sobre la relación entre estas variables?
- ¿Podemos afirmar, con un cierto nivel de significación, que existe evidencia de que el peso de una persona tiende a aumentar con la edad?
- ¿Podemos afirmar, con un cierto nivel de significación, que existe evidencia de que la cantidad de grasa en sangre de una persona tiende a aumentar con la edad?
- ¿Podemos predecir, aproximadamente, la cantidad de grasa en sangre según la edad de la persona? ¿Qué grado de confianza tiene esa predicción?

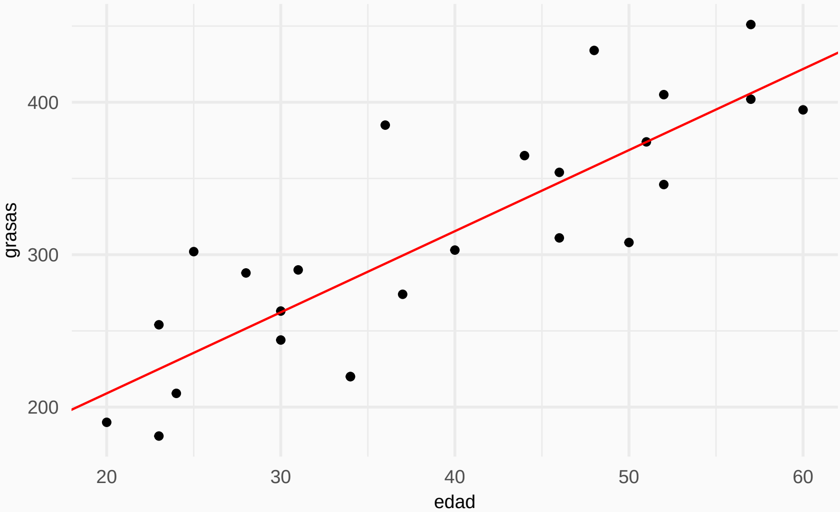
grasas vs. edad

- Objetivo: Encontrar la recta que mejor ajusta a estos datos.



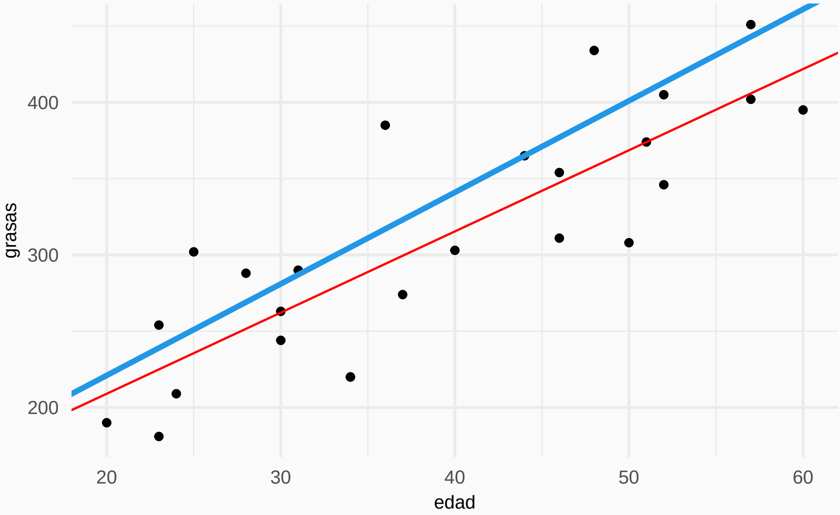
grasas vs. edad

- ¿Cuál de estas rectas ajusta mejor al conjunto de datos?



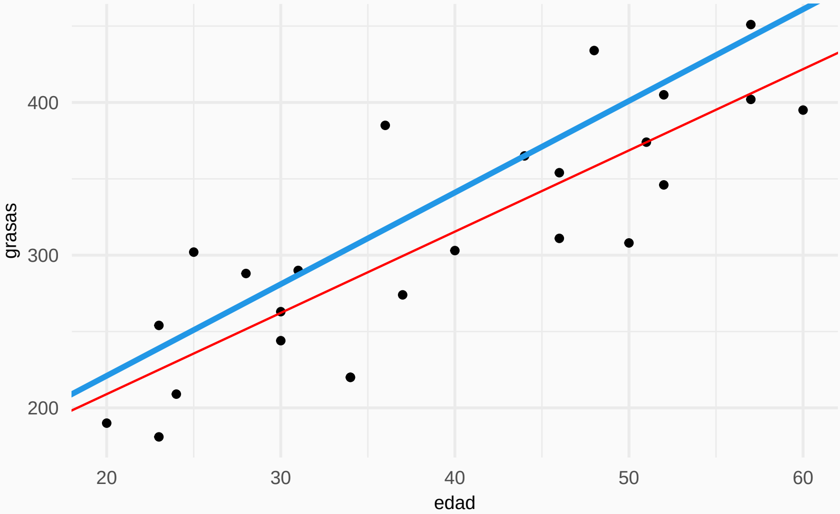
grasas vs. edad

- ¿Cuál de estas rectas ajusta mejor al conjunto de datos?



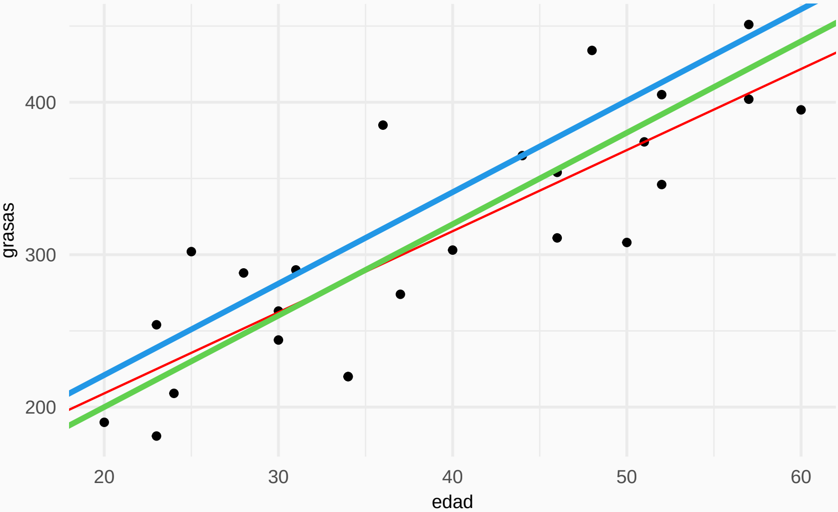
grasas vs. edad

- ¿Cuál de estas rectas ajusta mejor al conjunto de datos?



grasas vs. edad

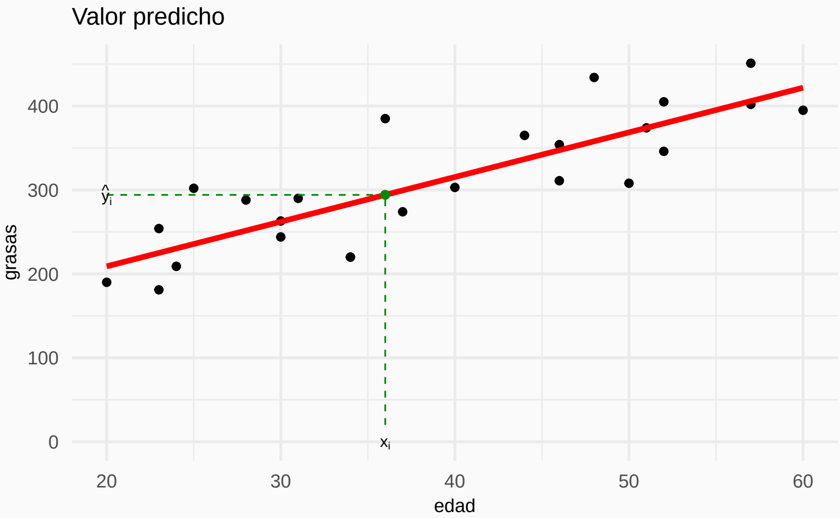
- ¿Cuál de estas rectas ajusta mejor al conjunto de datos?



grasas vs. edad

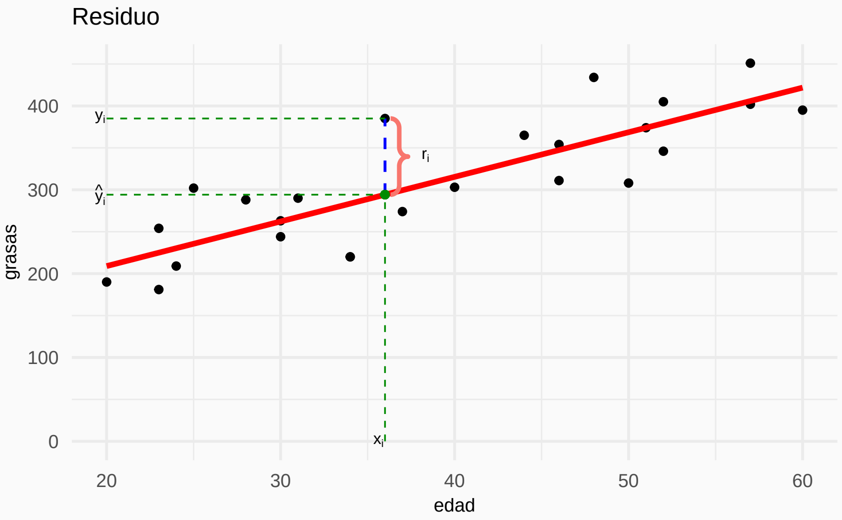
- Hay que definir un criterio para elegir la recta que mejor ajusta a los datos.
- Para esto vamos a definir el concepto de predicho y de residuo.

Predicho

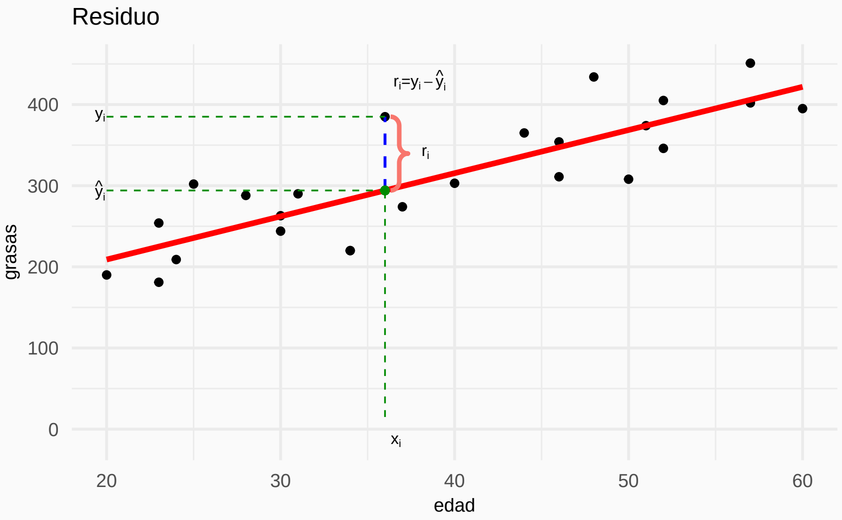


Valor predicho
grasas
400
300
200
100
0
$\hat{y}_i$
x_i
20
30
40
50
60
edad

Residuo

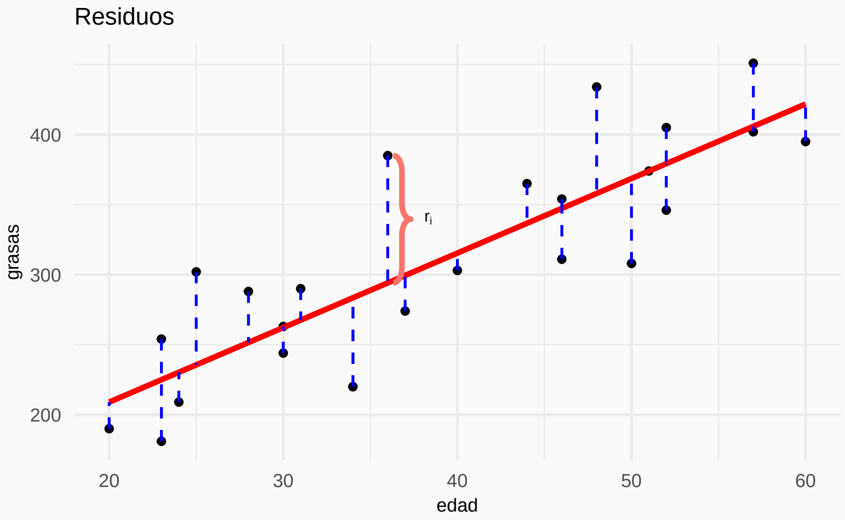

Residuo
grasas
edad
400
300
200
100
0
20
30
40
50
60
y_i
$\hat{y}_i$
r_i
x_i

Residuo

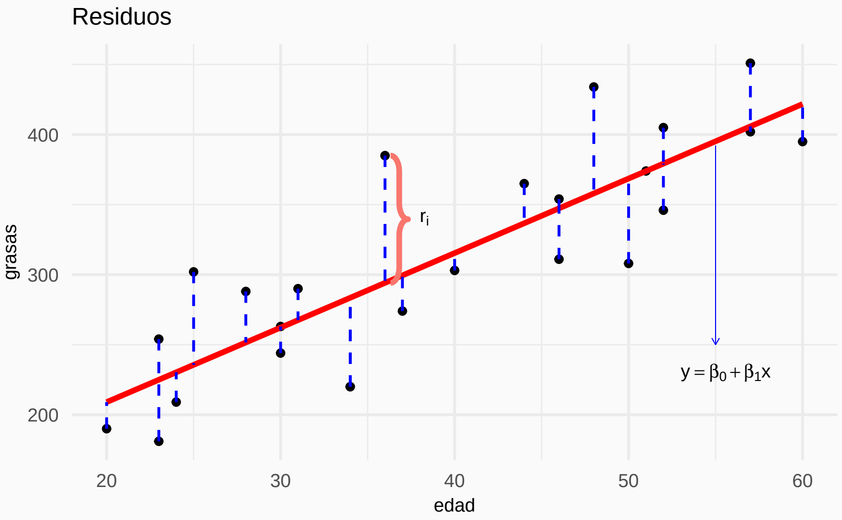


Residuo
$r_i = y_i - \hat{y}_i$
y_i
r_i
$\hat{y}_i$
x_i
400
300
200
100
0
grasas
20
30
40
50
60
edad

Criterio



Criterio



Residuos
400
300
200
grasas
20
30
40
50
60
edad
r_i
$y = \beta_0 + \beta_1 x$

Criterio

- Hay que encontrar los parámetros β_0 y β_1.
- Abordaremos este problema desde dos puntos de vista.
- Matemático, a través del método de Mínimos Cuadrados (MC).
- Estadístico: a través del modelo de regresión lineal (MRL).

Primer criterio: Método de Mínimos Cuadrados

- ¿Qué propone el método de MC?
- Minimizar la suma de los residuos al cuadrado.

$$\widehat{\beta_0}, \widehat{\beta_1} = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} r_i^2$$

$$\widehat{\beta_0}, \widehat{\beta_1} = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \widehat{y}_1)^2$$

$$\widehat{\beta_0}, \widehat{\beta_1} = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Estimadores de mínimos cuadrados

- En términos matemáticos, esta minimización se reduce a resolver el siguiente sistema de ecuaciones.

$$\begin{cases} \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = 0 \\ \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = 0 \end{cases}$$

- Derivando, obtenemos que los estimadores de mínimos cuadrados de β_0 y β_1 son los que resuelven el siguiente sistema de ecuaciones, conocidas como las ecuaciones normales.

$$\begin{cases} \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) = 0 \\ \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i) x_i = 0 \end{cases} \tag{1}$$

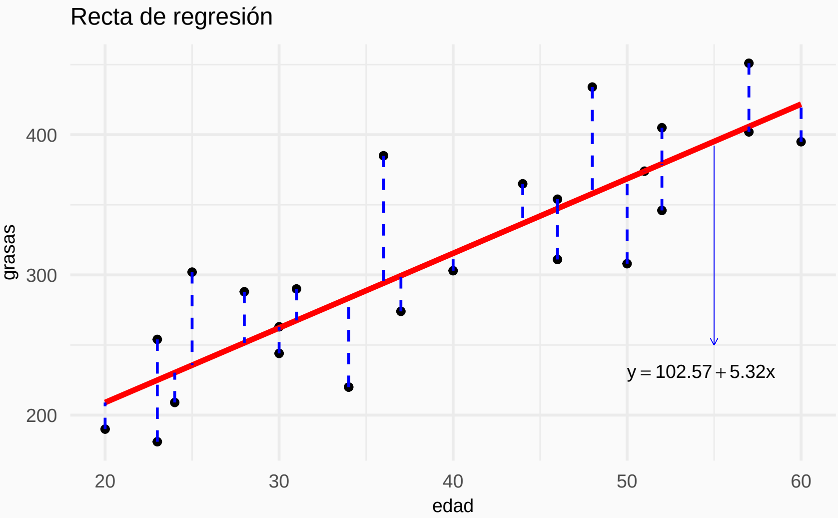
Estimadores de mínimos cuadrados

- Despejando, obtenemos las siguientes expresiones para $\widehat{\beta}_0$ y $\widehat{\beta}_1$, que los llamaremos $\widehat{\beta}_{0,\text{MC}}$ y $\widehat{\beta}_{1,\text{MC}}$.

$$\widehat{\beta}_{1,\text{MC}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

$$\widehat{\beta}_{0,\text{MC}} = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

Recta de regresión por MC



MC

- Este abordaje no nos permite hacer inferencia sobre los parámetros de la regresión.
- Esto lo resuelve el modelo de regresión lineal.
- Antes de presentar el modelo de regresión lineal daremos algunos conceptos.

Covarianza entre dos variables

Definición:

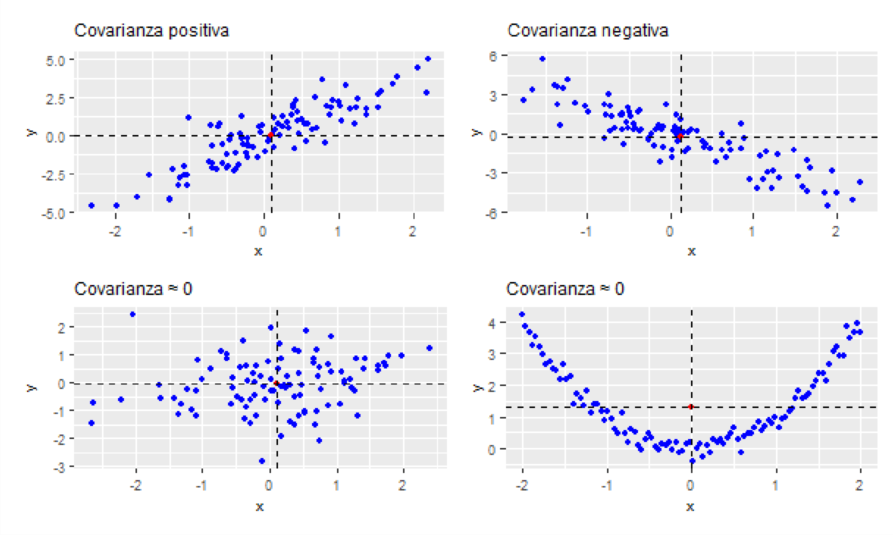
La covarianza entre dos variables aleatorias X y Y se define como:

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Interpretación:

- $\text{Cov}(X, Y) > 0$: ambas variables tienden a aumentar juntas (relación positiva).
- $\text{Cov}(X, Y) < 0$: cuando una variable aumenta, la otra tiende a disminuir (relación negativa).
- $\text{Cov}(X, Y) \approx 0$: no hay relación lineal clara.

Covarianza



Covarianza positiva
5.0
2.5
0.0
-2.5
-5.0
-2
-1
0
1
2
y
x
Covarianza negativa
6
3
0
-3
-6
-1
0
1
2
y
x
Covarianza ≈ 0
2
1
0
-1
-2
-3
-2
-1
0
1
2
y
x
Covarianza ≈ 0
4
3
2
1
0
-2
-1
0
1
2
y
x

Coeficiente de correlación

- $\text{Cov}(X, Y)$ depende de las unidades que se miden X e Y.
- Es conveniente tener una medida de la relación lineal entre las variables que no dependa de las unidades.
- Esta medida es el Coeficiente de Correlación Lineal.

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Propiedades:

- No depende de las unidades de medida de las variables.
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza.
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Coeficiente de correlación

- Cov(X, Y) depende de las unidades que se miden X e Y.
- Es conveniente tener una medida de la relación lineal entre las variables que no dependa de las unidades.
- Esta medida es el Coeficiente de Correlación Lineal.

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Propiedades:

- No depende de las unidades de medida de las variables.
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza.
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Coeficiente de correlación

- $\text{Cov}(X, Y)$ depende de las unidades que se miden X e Y.
- Es conveniente tener una medida de la relación lineal entre las variables que no dependa de las unidades.
- Esta medida es el Coeficiente de Correlación Lineal.

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Propiedades:

- No depende de las unidades de medida de las variables.
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza.
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Coeficiente de correlación

- $\mathrm{Cov}(X, Y)$ depende de las unidades que se miden X e Y.
- Es conveniente tener una medida de la relación lineal entre las variables que no dependa de las unidades.
- Esta medida es el Coeficiente de Correlación Lineal.

$$\rho_{xy} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Propiedades:

- No depende de las unidades de medida de las variables.
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza.
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Coeficiente de correlación

- $\text{Cov}(X, Y)$ depende de las unidades que se miden X e Y.
- Es conveniente tener una medida de la relación lineal entre las variables que no dependa de las unidades.
- Esta medida es el Coeficiente de Correlación Lineal.

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Propiedades:

- No depende de las unidades de medida de las variables.
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza.
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Coeficiente de correlación

- $\text{Cov}(X, Y)$ depende de las unidades que se miden X e Y.
- Es conveniente tener una medida de la relación lineal entre las variables que no dependa de las unidades.
- Esta medida es el Coeficiente de Correlación Lineal.

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Propiedades:

- No depende de las unidades de medida de las variables.
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza.
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Coeficiente de correlación

- $\mathrm{Cov}(X, Y)$ depende de las unidades que se miden X e Y.
- Es conveniente tener una medida de la relación lineal entre las variables que no dependa de las unidades.
- Esta medida es el Coeficiente de Correlación Lineal.

$$\rho_{xy} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Propiedades:

- No depende de las unidades de medida de las variables.
- Siempre toma valores entre -1 y 1.
- Su signo se interpreta igual que el de la covarianza.
- Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Aunque $r_{xy} \approx 0$, las variables x e y no son necesariamente independientes.

Modelo de regresión lineal simple

- El modelo de regresión lineal simple es un modelo que estudia la relación entre dos variables aleatorias:
    - $X =$ variable predictora, regresora o explicativa.
    - $Y =$ variable dependiente o respuesta.
- Se llama modelo lineal simple porque solamente existe una variable regresora en el modelo.

Objetivo de la regresión lineal

Construir un modelo que describa la relación entre X e Y, y permita:

- Analizar si existe una relación lineal entre ambas variables.
- Cuantificar esta relación mediante un estudio del modelo.
- Predecir el valor de Y para un valor dado de $X = x$.

Modelo lineal simple

La forma del modelo lineal simple es:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2}$$

donde $\varepsilon_i \sim N(0, \sigma^2)$ son variables aleatorias independientes.

- β_0: ordenada al origen o intercept, indica el valor medio de Y cuando $X = 0$.
- β_1: pendiente, indica la variación media de la variable respuesta cuando X varía en una unidad.

Supuestos del modelo

La ecuación (2) junto con el ítem a) nos habla de ciertos supuestos que debe cumplir el modelo para que sea válido.

- Los ε_i tienen media cero, $E(\varepsilon_i) = 0$ (homogeneidad de los errores).
- Los ε_i tienen todos la misma varianza desconocida que llamaremos σ^2, y es otro parámetro del modelo. (homoscedasticidad de los errores).
- Los ε_i tienen distribución normal (normalidad de los errores).
- Los ε_i son independientes entre sí, y son no correlacionados con las X_i (independencia de los errores).

Modelo lineal simple

- Observación: Estos cuatro supuestos pueden resumirse en la siguiente expresión

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ con } \varepsilon_i \sim N\left(0, \sigma^2\right), \quad 1 \leq i \leq n,$$

independientes entre sí.

1. Lo bueno: La incorporación del término aleatorio permite hacer inferencia sobre los parámetros estimados.
2. Lo malo: Agrega un parámetro a estimar que es σ^2.

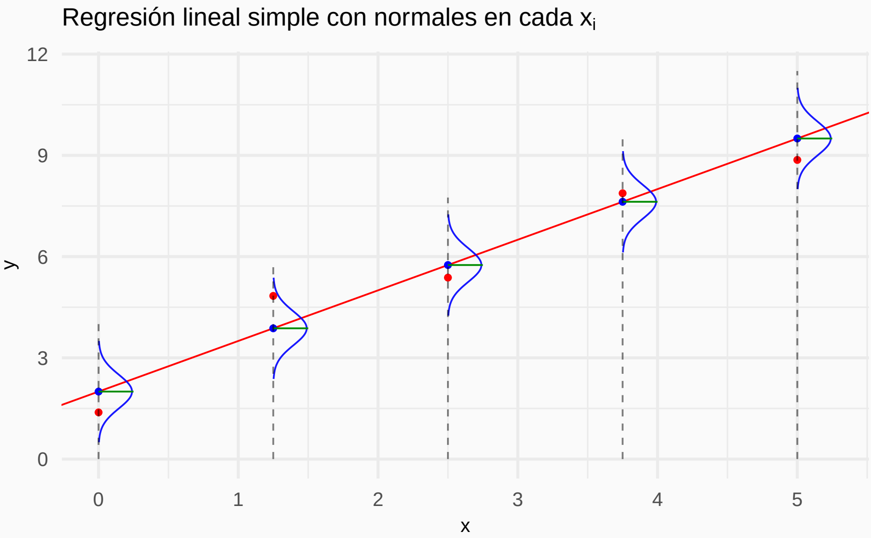
Modelo lineal simple

- Comentarios: en la ecuación (2) lo único observable son los pares (X_i, Y_i).
- Se desconocen β_0 como a β_1 y σ^2(que son números fijos). Estos parámetros hay que estimarlos.
- A los ε_i no los observamos.

Estimación de los parámetros del modelo

- El modelo es $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, donde $\varepsilon_i \sim N(0, \sigma^2)$ son variables aleatorias independientes.
- Esto indica que, para cada i, la variable respuesta $Y_i = Y|X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
- Esto se observa en el siguiente gráfico.

Estimación de los parámetros del modelo



Estimación de los parámetros del modelo

- Es decir, para cada i la variable respuesta

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

  podemos utilizar el método de Máxima Verosimilitud (MV) para estimar los parámetros del modelo β_0, β_1 y σ^2.

- Que llamaremos $\widehat{\beta}_{0,\mathrm{MV}}$, $\widehat{\beta}_{1,\mathrm{MV}}$ y $\widehat{\sigma^2}_{\mathrm{MV}}$, respectivamente.

Método de máxima verosimilitud

- El estimador de máxima verosimilitud (EMV) es un método estadístico para estimar los parámetros de un modelo de probabilidad.
- Consiste en encontrar el valor del parámetro (o los parámetros) que *maximiza la función de verosimilitud*, es decir, la probabilidad (o densidad) de observar los datos muestrales, dados esos parámetros.
- Definición:

  Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con función de densidad (o de masa) $f(x;\theta)$, donde θ es un parámetro desconocido. La función de verosimilitud es:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

  El estimador de máxima verosimilitud $\hat{\theta}$ es el valor de θ que

Método de máxima verosimilitud

- El estimador de máxima verosimilitud (EMV) es un método estadístico para estimar los parámetros de un modelo de probabilidad.
- Consiste en encontrar el valor del parámetro (o los parámetros) que *maximiza la función de verosimilitud*, es decir, la probabilidad (o densidad) de observar los datos muestrales, dados esos parámetros.

Método de máxima verosimilitud

- Definición:
  Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con función de densidad (o de masa) $f(x;\theta)$, donde θ es un parámetro desconocido. La función de verosimilitud es:

  $$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

  El estimador de máxima verosimilitud $\hat{\theta}$ es el valor de θ que maximiza $L(\theta)$

  $$\hat{\theta} = \arg\max_{\theta} \log L(\theta)$$

Método de máxima verosimilitud

- Función de densidad

$$f\left(y_i\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right\}$$

- Función de verosimilitud

$$\mathcal{L}\left(\beta_0, \beta_1, \sigma^2 / X = x_i\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right]$$

- Para maximizar esta función en función de β_0, β_1 y σ^2, tomamos logaritmos, lo cual lleva a:

$$\mathcal{L}\left(\beta_0, \beta_1, \sigma^2 / X = x_i\right) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2.$$

Método de máxima verosimilitud

- Derivamos respecto a los tres parámetros e igualamos a cero.
- Obtenemos el sistema de ecuaciones siguiente, llamado Ecuaciones normales.

$$\frac{\partial \mathcal{L}}{\partial \beta_0}\left(\beta_0, \beta_1, \sigma^2\right) = \frac{1}{2\sigma^2}\sum_i 2\left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \beta_1}\left(\beta_0, \beta_1, \sigma^2\right) = \frac{1}{2\sigma^2}\sum_i 2x_i\left(y_i - \beta_0 - \beta_1 x_i\right) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2}\left(\beta_0, \beta_1, \sigma^2\right) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2 = 0$$

Método de máxima verosimilitud

- Al resolverlo obtenemos:

$$\widehat{\beta}_{1,\text{MV}} = \frac{\text{cov}(X, Y)}{\text{var}(X)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\beta}_{0,\text{MV}} = \bar{Y} - \widehat{\beta}_{1,\text{MV}}\bar{x}$$

$$\widehat{\sigma^2}_{\text{MV}} = \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{\beta}_{0,\text{MV}} - \widehat{\beta}_{1,\text{MV}}x_i\right)^2}{n} = \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n}$$

Algunos comentarios

- Observación 1: Bajo la hipótesis de normalidad de los errores, se tienen las siguientes igualdades:

$$\widehat{\beta}_{0,\text{MV}} = \widehat{\beta}_{0,\text{MC}}$$

$$\widehat{\beta}_{1,\text{MV}} = \widehat{\beta}_{1,\text{MC}}$$

Algunos comentarios

- Observación 2: Si bien $\widehat{\sigma^2}_{MV} = \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{n}$, un estimador de $\widehat{\sigma}^2_{MV}$ con mejores propiedades es

$$\widehat{\sigma^2} = S_r^2 = \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2$$

- Y este es el estimador de σ^2 que se considera en el problema de regresión lineal simple.

Algunos comentarios

- Observación 3:
  - Para cada $i = 1, \ldots, n$:

$$Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right).$$

  - Para cada valor $X_i = x$ fijo:

$$E\left[Y \mid X = x\right] = \beta_0 + \beta_1 x.$$

Ejemplo en R

```
regresion <- lm(grasas ~ edad, data = grasas)

> regresion

Call:
lm(formula = grasas ~ edad, data = grasas)

Coefficients:
(Intercept) edad
102.575 5.321
```

Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.478 -26.816  -3.854  28.315  90.881 

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept) 102.5751    29.6376   3.461    0.00212 ** 
edad          5.3207     0.7243   7.346   1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882 
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max
-63.478 -26.816  -3.854  28.315  90.881

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 102.5751     29.6376   3.461    0.00212 **
edad          5.3207      0.7243   7.346   1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

$\hat{\beta}_0$

Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max
-63.478 -26.816  -3.854  28.315  90.881

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 102.5751    29.6376   3.461    0.00212 **
edad          5.3207     0.7243   7.346    1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

$\hat{\beta}_1$

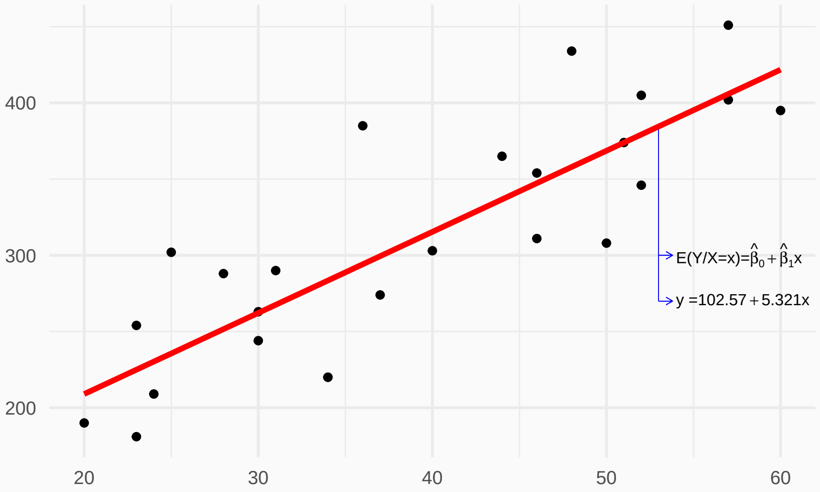
Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max
-63.478 -26.816  -3.854  28.315  90.881

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 102.5751    29.6376   3.461    0.00212 **
edad          5.3207     0.7243   7.346   1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\hat{\sigma}$

```
Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

Modelo lineal



¿Qué nos interesa resolver?

1. Verificar el cumplimiento de los supuestos usando resúmenes, gráficos y pruebas estadísticas.
2. Evaluar la adecuación del modelo a los datos mediante una medida de ajuste.
3. Estimar los parámetros del modelo a partir de las observaciones.
4. Realizar inferencias sobre los parámetros (pruebas de hipótesis e intervalos de confianza para β_0, β_1 y σ^2).

¿Qué nos interesa resolver?

1. Predecir el valor de Y para un valor de una nueva observación de X.
2. Estimar la esperanza condicional de Y para un valor de X (observado o no observado), construyendo un intervalo de confianza para dicha esperanza y evaluando el error asociado.
3. Construir un intervalo de predicción para el valor de Y correspondiente a un nuevo valor de X.

Supuestos del modelo lineal

- ¿Cuáles son los supuestos que hay que validar?

  1. $\mathrm{E}(\varepsilon_i) = 0$
  2. Independencia: los errores ε_i son independientes.
  3. Homocedasticidad: varianza constante de los ε_i.
  4. Normalidad: los ε_i se distribuyen normalmente.

- Estos supuestos se reducen a ver que $\varepsilon_i \sim N(0, \sigma^2)$ para todo $i = 1, \ldots, n$, independientes entre si.

Organización

Diagnóstico de la regresión

- Son técnicas que permiten validar el cumplimiento de los supuestos del modelo planteado.
- Se basan, en general, en:
    - el análisis de los residuos.
    - en el análisis de la existencia de puntos influyentes.

Análisis de residuos

- Notar que los ε_i no son observales.
- Lo que sí son observables son los residuos $r_i = Y_i - \widehat{Y}_i$.
- Desafortunadamente no se pueden considerar a los r_i como estimadores de los ε_i porque muchas de las cualidades de los ε_i no las heredan los r_i.

Análisis de residuos

- Por ejemplo: los residuos no son independientes, están correlacionados. Pero esta correlación es chica.
- De hecho, a partir de las ecuaciones normales (1), se puede ver que $\sum_{i=1}^{n} r_i = 0$.
- Lo que implica que, conociendo $r_1, \ldots, r_{n-1}$ se puede conocer r_n.
- Además, los ε_i tienen todos la misma varianza (son homoscedáticos), pero los r_i no.

Análisis de residuos

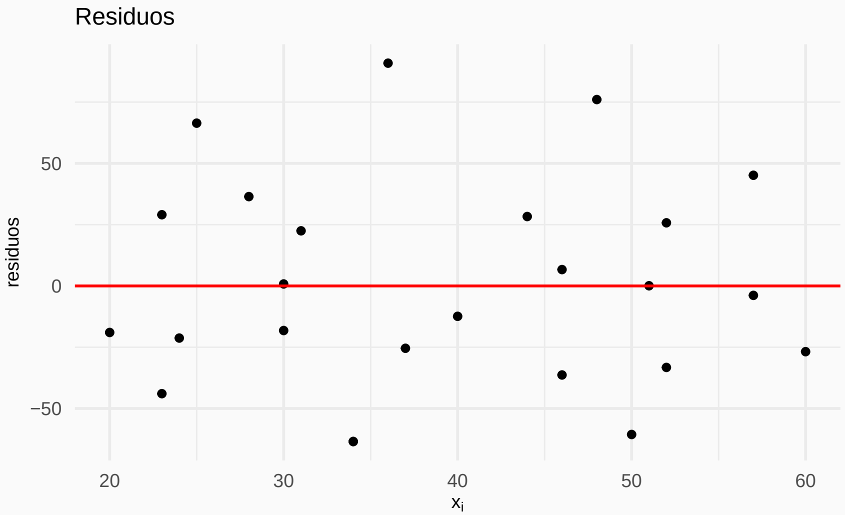
- Pero... es lo que podemos observar. Así que los usaremos para analizar la validez de los supuestos del modelo.

- Importante: Si los residuos del modelo ajustado no presentan un comportamiento razonable, puede indicar que la especificación del modelo, total o parcial, no es adecuada para los datos analizados.

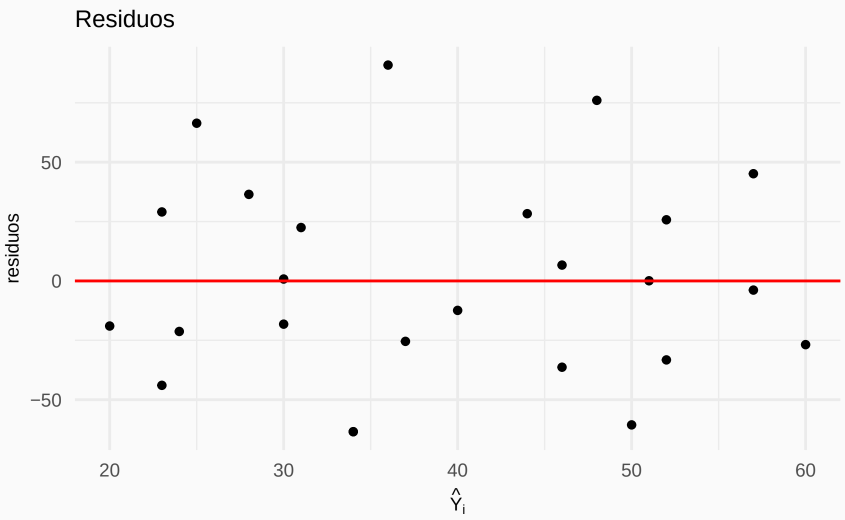
Chequeo de supuestos: gráfico de residuos

- Gráfico de residuos vs. la covariable (también llamado *residual plot*).
  - Es un scatter plot de los r_i vs. X_i (variable regresora o covariable).
  - En regresión lineal simple: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$.
  - Por ser transformación lineal, graficar r_i vs. $\widehat{Y}_i$ o vs. X_i es equivalente.
- Este gráfico permite verificar visualmente si los supuestos del modelo se cumplen, salvo el de normalidad de los residuos.

r_i vs. X_i



r_i vs. $\widehat{Y}_i$



Análisis de residuos

- Se puede probar que los r_i verifican que:

$$E(r_i) = 0$$

$$Var(r_i) = \sigma^2(1 - h_{ii}), \text{ donde} \tag{3}$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \tag{4}$$

- Si bien los r_i no son estimadores de los ε_i, se pueden considerar como su correlato empírico.

- Por este motivo analizamos los r_i para validar los supuestos del modelo.

Análisis de residuos

- Observaciones:
  - de (3) se puede ver que la varianza de los r_i no es constante.
  - La varianza del residuo de un dato depende del valor de la covariable.
  - Los residuos correspondientes a distintas observaciones tienen diferentes varianzas.
  - Cuánto mayor sea h_{ii} menor será la varianza del r_i.
  - Mientras más cercano a uno sea h_{ii}, más cercana a cero será la varianza del residuo de la observación i-ésima.

Análisis de residuos

- Entonces:

  1. Hay que definir otro concepto de residuo que permita hacer comparables a los residuos entre si.
  2. Hay que estudiar la observaciones con h_{ii} alto.

Residuos estandarizados

- Vamos a definir un concepto de residuo que permita comparar los residuos en una escala común, independientemente de su variabilidad.

- Residuo estandarizado:

$$\begin{aligned} r_{si} &= \frac{r_i}{\sqrt{\widehat{\sigma}^2\,(1-h_{ii})}} \\ &= \frac{r_i}{S_r\sqrt{(1-h_{ii})}} \end{aligned} \tag{5}$$

- Se puede probar que los r_s tienen media poblacional cero (igual que los residuos), y varianza poblacional igual a uno, es decir

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

r_{is} vs. $\widehat{Y}_i$

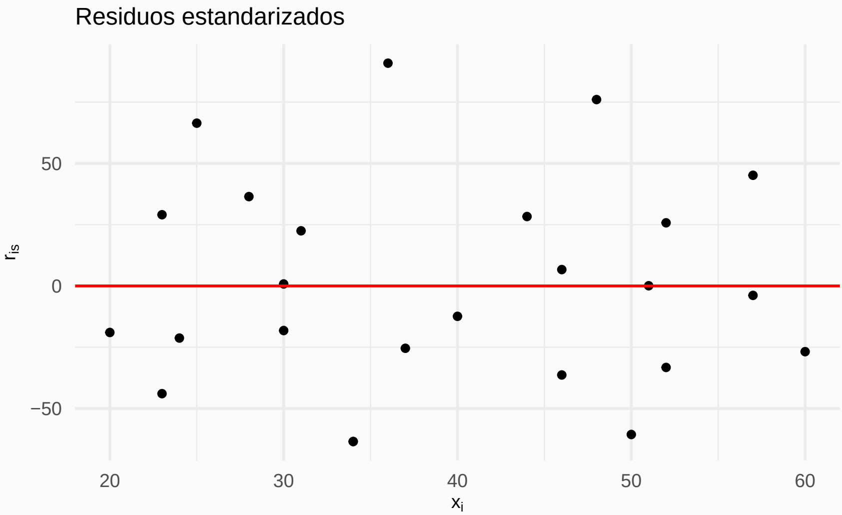


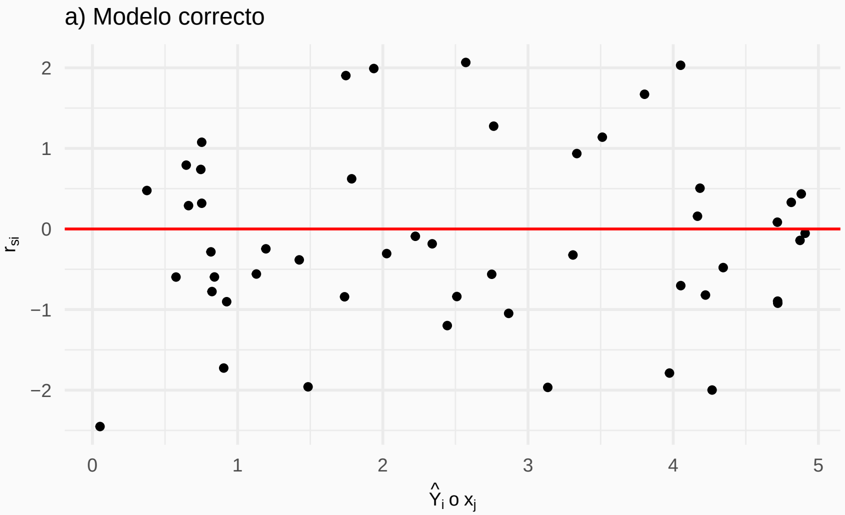
Gráfico de residuos

- Entonces: Los residuos estandarizados
  - Permiten igualar la variabilidad a lo largo de los valores de la covariable.
  - Permiten hacer comparables a los residuos entre si por llevarlos a una misma escala.
  - Si bien están correlacionados, esta correlación no es muy importante.

Gráfico de residuos

- ¿Cómo debería verse el gráfico de residuos vs. la covariable (o los predichos) si el modelo es adecuado?
- Si el modelo es correcto, el gráfico de los residuos versus predichos o versus la covariable debería lucir como una nube de puntos sin estructura, ubicada alrededor del eje horizontal.

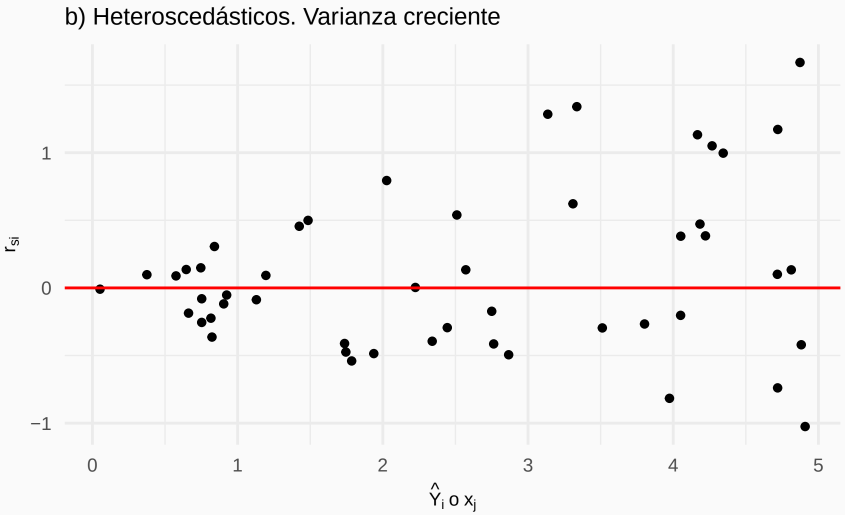
Gráficos de residuos



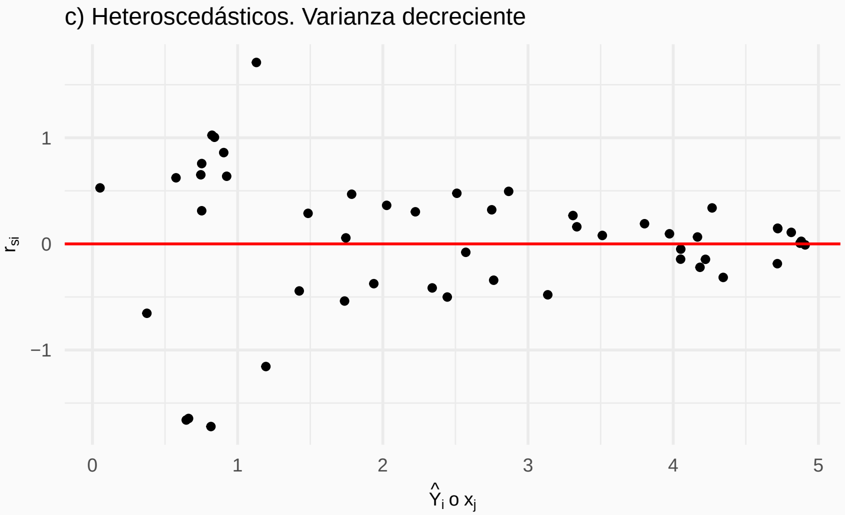
Residuos modelo incorrecto

- En este caso el gráfico de residuos estandarizados versus la variable predictora (o versus los valores predichos) suele tener algún tipo de estructura.
- Los gráficos b) y c) violan el supuesto de homoscedasticidad.
- Los gráficos d) a g) indican que se viola el supuesto de linealidad de la esperanza condicional.

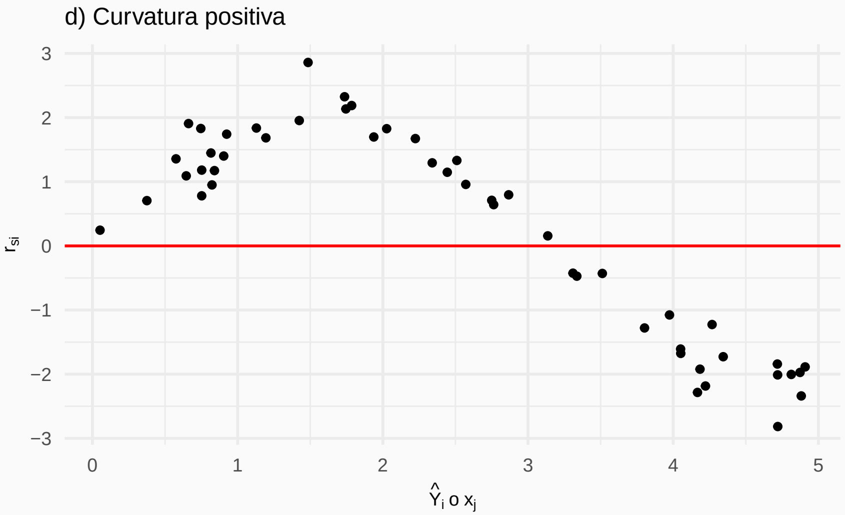
Heteroscedásticos con varianza creciente



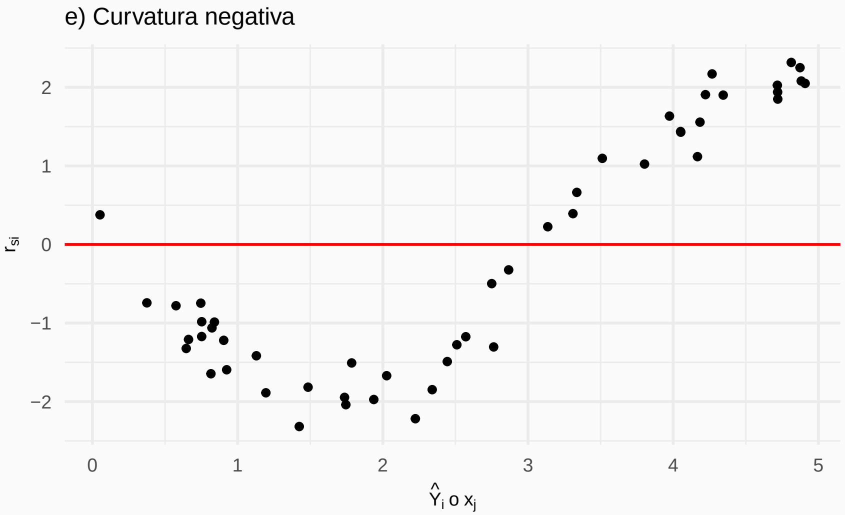
Heteroscedásticos con varianza creciente



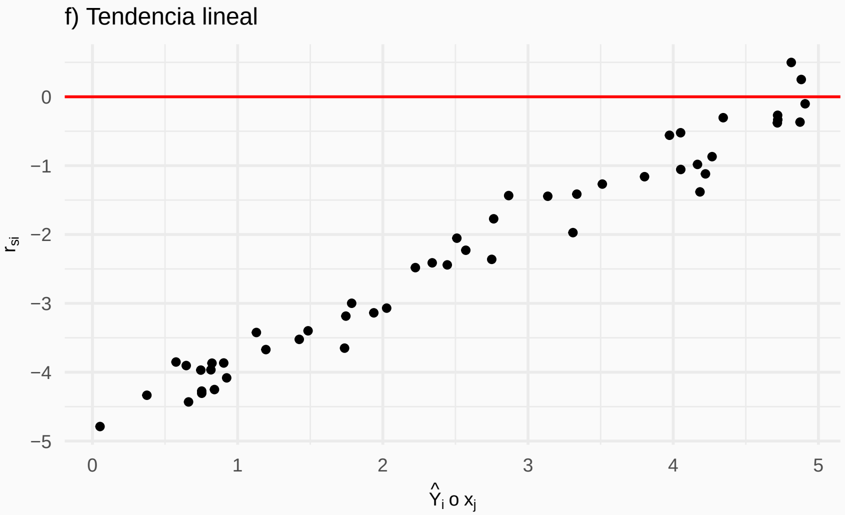
Curvatura



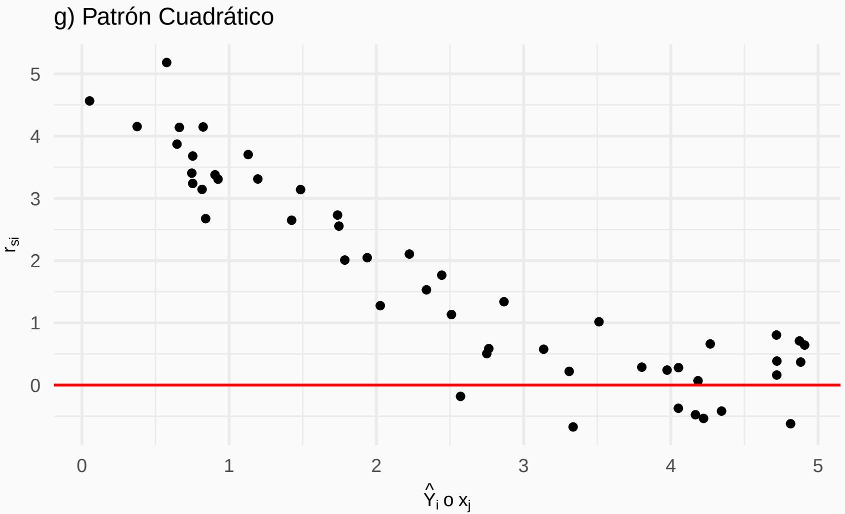
Curvatura



Heteroscedásticos con varianza creciente



Heteroscedásticos con varianza creciente



Residuos modelo incorrecto

- Los gráficos b) y c) violan el supuesto de homoscedasticidad.
- Los gráficos d) a g) indican que se viola el supuesto de linealidad de la esperanza condicional.

Diagnóstico de la regresión: Análisis de residuos

- ¿Cómo validamos el supuesto de independencia?
  - Si las observaciones provienen de una muestra aleatoria de sujetos, entonces en principio se consideran independientes.
  - Sin embargo, hay situaciones frecuentes en las que este supuesto puede no cumplirse.

Violaciones del supuesto de independencia

Ejemplo 1: Datos recolectados secuencialmente

- Cuando las mediciones se toman en orden temporal, puede existir dependencia entre observaciones consecutivas.
- Esto es común en determinaciones de laboratorio o mediciones hechas por un mismo operador.
- Puede existir un patrón relacionado con el funcionamiento del equipo o con el observador.

Detección de dependencia temporal

- Una herramienta útil para detectar dependencia es graficar los residuos contra el orden de medición.
- Esto permite visualizar posibles correlaciones temporales entre observaciones.

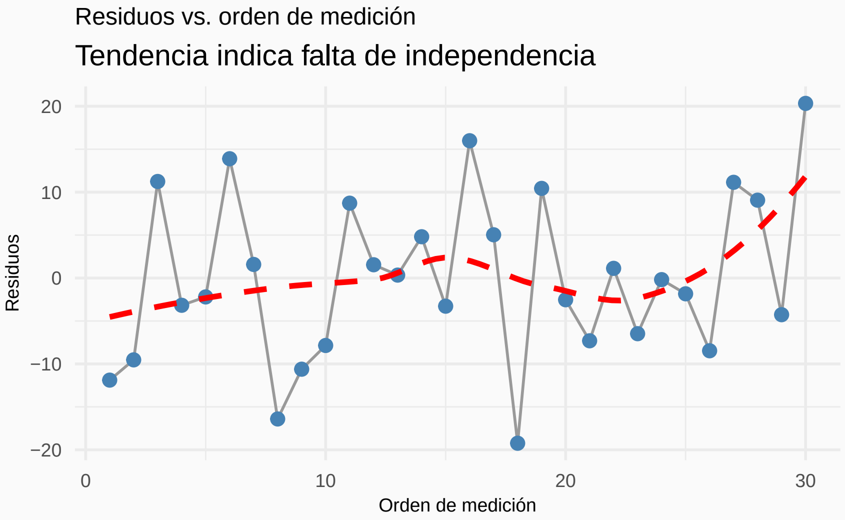
Detección de dependencia temporal

- Ejemplo: Medición de glucosa en sangre en laboratorio.
- Supongamos que un técnico de laboratorio mide la glucosa en sangre de 30 muestras en orden secuencial a lo largo de la mañana. Cada medición se realiza usando el mismo equipo, sin calibrar entre muestras.
- Puede pasar que el equipo puede tener un desvío gradual en su precisión por temperatura, tiempo de uso o batería.
- Las primeras muestras podrían estar sistemáticamente más bajas (o más altas) que las últimas, independientemente del valor real.
- Esto genera una tendencia temporal que rompe el supuesto de independencia entre observaciones.

Detección de dependencia temporal

- ¿Cómo detectarlo?
  - Graficando los residuos del modelo contra el orden en que se tomaron las muestras.
  - Si se observa un patrón (por ejemplo, residuos que crecen con el número de muestra), es un indicio claro de dependencia.

Dependencia temporal



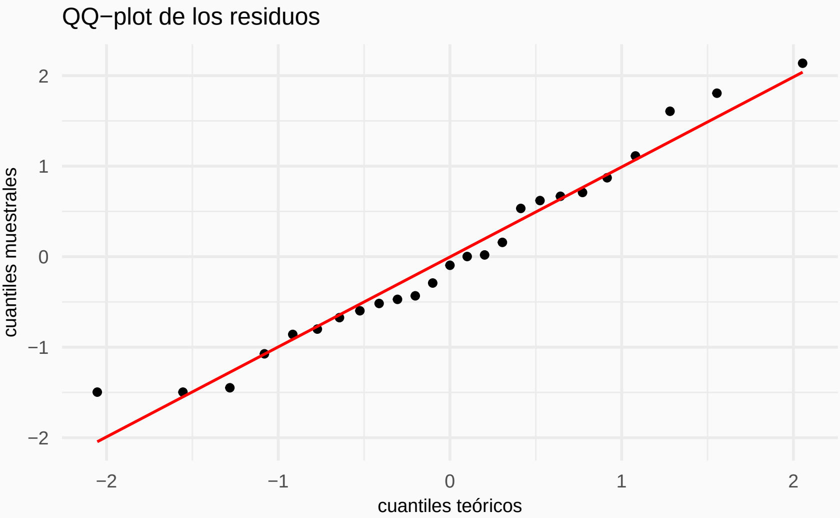
Ejemplo 2: Observaciones repetidas sobre los mismos sujetos

- Si varias mediciones se hacen sobre el mismo sujeto (o animal), las observaciones no serán independientes.
- En este caso, se puede considerar un modelo de regresión múltiple donde el sujeto entra como covariable.

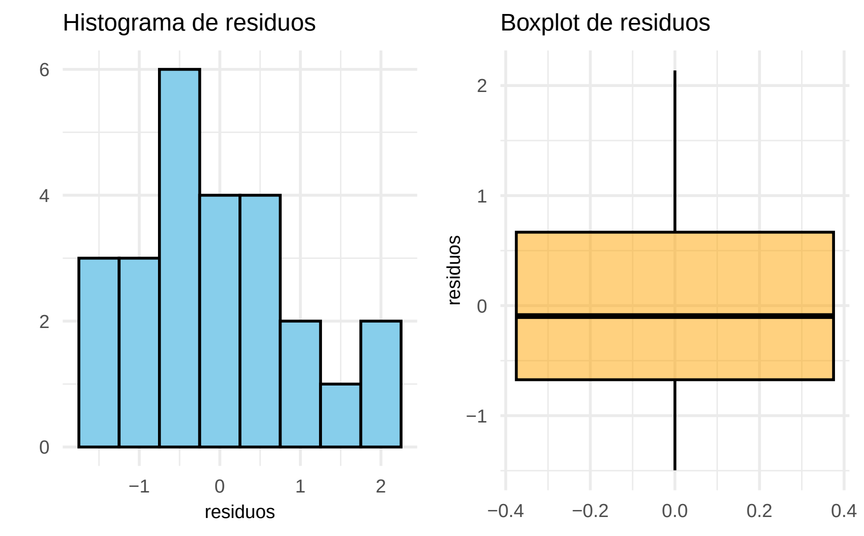
Modelos apropiados para dependencia

- En situaciones de dependencia entre observaciones, los modelos correctos son:
  - Modelos ANOVA con efectos aleatorios.
  - Modelos de efectos mixtos.
- Estos modelos exceden el contenido de estas notas.

Normalidad de los residuos

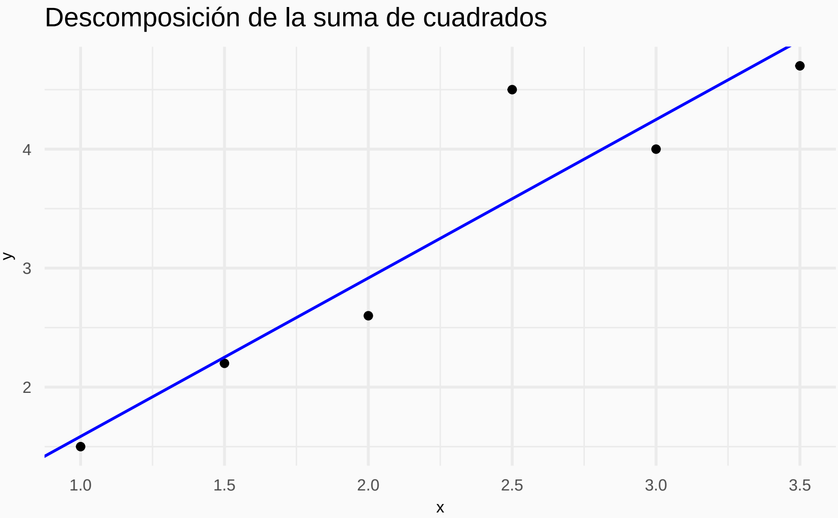


Normalidad de los residuos

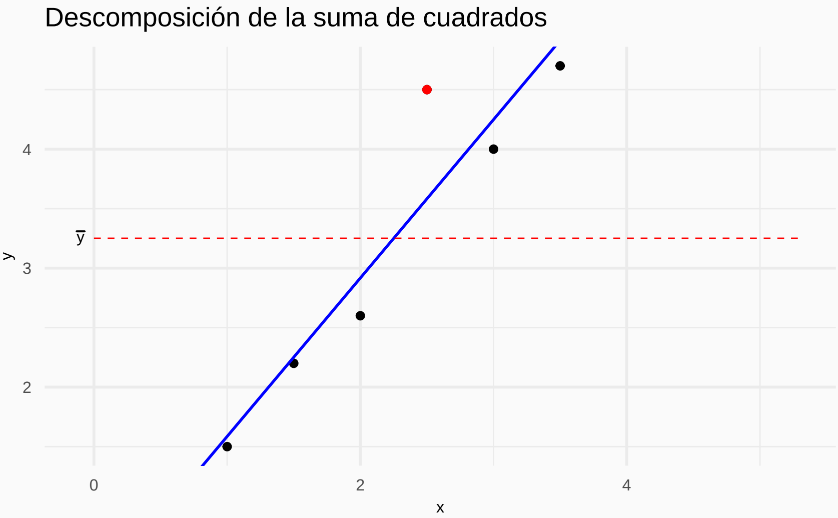


Organización

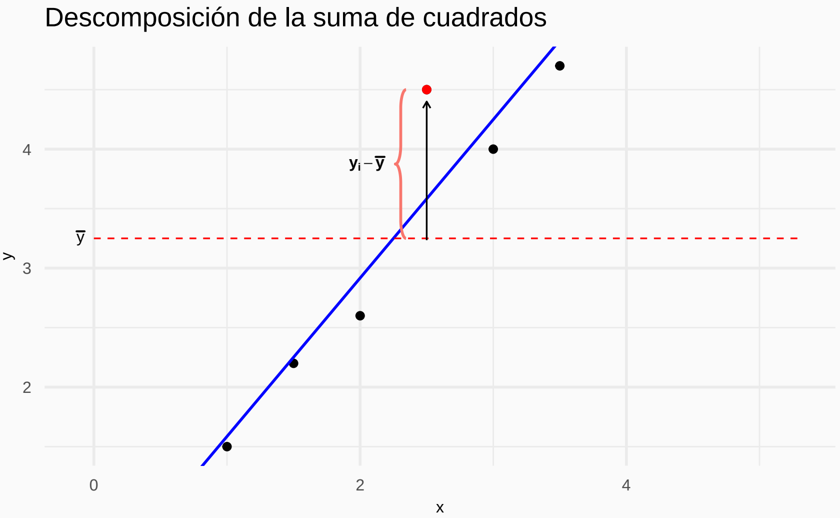
Descomposición de la suma de cuadrados



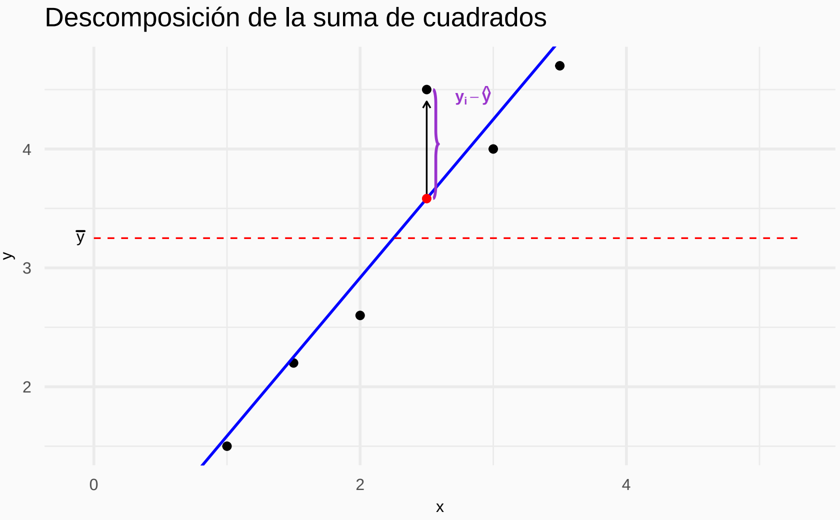
Descomposición de la suma de cuadrados



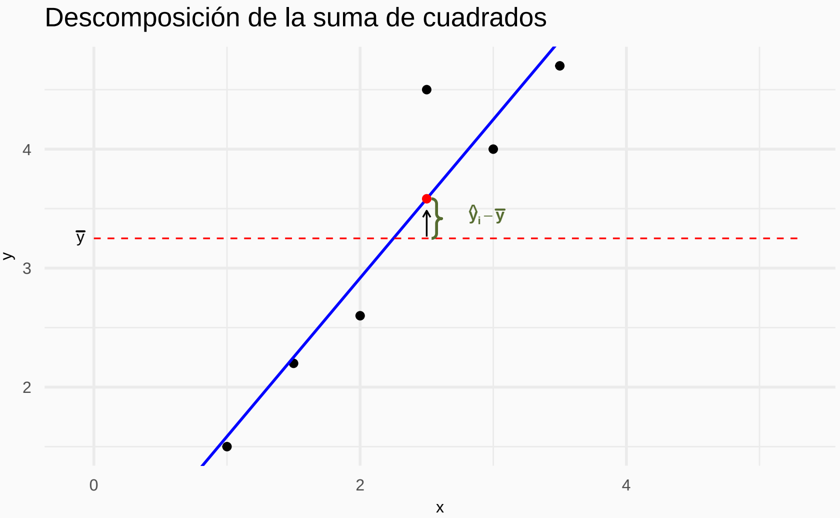
Descomposición de la suma de cuadrados



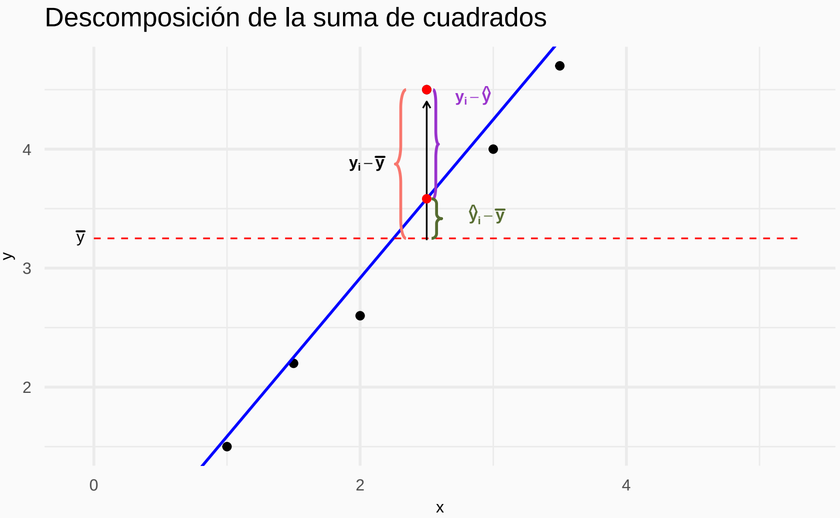
Descomposición de la suma de cuadrados



Descomposición de la suma de cuadrados



Descomposición de la suma de cuadrados



Descomposición de la suma de cuadrados

- Entonces:

$$Y_i - \bar{Y} = Y_i - \widehat{Y}_i + \widehat{Y}_i - \bar{Y}$$

- Notación:

$$\underbrace{Y_i - \bar{Y}}_{\text{desviación total}} = \underbrace{\widehat{Y}_i - \bar{Y}}_{\substack{\text{desvío de los predichos} \\ \text{respecto de la media}}} + \underbrace{Y_i - \widehat{Y}_i}_{\substack{\text{desvío alrededor} \\ \text{de la recta de regresión} \\ \text{ajustada}}}$$

Descomposición de la suma de cuadrados

- Obviamente es falsa la siguiente igualdad:

$$\left(Y_i - \bar{Y}\right)^2 = \left(\widehat{Y}_i - \bar{Y}\right)^2 + \left(Y_i - \widehat{Y}_i\right)^2.$$

- Pero, se puede probar que:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\substack{\text{Suma de cuadrados}\\ \text{totales}\\ \text{(SST)}}} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2}_{\substack{\text{Suma de cuadrados}\\ \text{debida a la regresión}\\ \text{(SSReg)}}} + \underbrace{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{\substack{\text{Suma de cuadrados}\\ \text{de los residuos}\\ \text{(SSRes)}}} \tag{6}$$

Coeficiente de determinación R^2

- Si en la ecuación (6) dividimos ambos miembros por $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, obtenemos

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

Coeficiente de determinación R^2

- Por lo tanto, obtenemos:

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

- En forma más compacta:

$$1 = \frac{\text{SSReg}}{\text{SST}} + \frac{\text{SSRes}}{\text{SST}}$$

Coeficiente de determinación R^2

- ¿Cuánto de la variabilidad total de la Y queda explicada por el modelo?

$$1 = \underbrace{\frac{\text{SSReg}}{\text{SST}}}_{\text{A}} + \underbrace{\frac{\text{SSRes}}{\text{SST}}}_{\text{B}}$$

- A: es la proporción de la variabilidad explicada por el modelo, respecto de la variablilidad total.
- B: es la proporción de la variabilidad debida a los residuos, respecto de la variablilidad total.

Coeficiente de determinación R^2

- Entonces

$$\frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{SSRes}}{\text{SST}}$$

- Se espera que un buen modelo tenga un valor chico en SSRes.
- Por lo tanto un valor alto en A, cercano a 1, indica un buen ajuste.

Coeficiente de determinación R^2

- Definición: Coeficiente de determinación R^2.
  R^2 indica qué parte de la variabilidad total de la variable dependiente Y es explicada por la variable independiente. Por lo tanto, sirve como una medida del poder predictivo del modelo.

Coeficiente de determinación R^2

- Propiedades de R^2
  - $0 \leq R^2 \leq 1$.
  - No depende de las unidades de medición.
  - 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.

Coeficiente de determinación R^2

- En R.

```
> summary(regresion)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.478 -26.816  -3.854  28.315  90.881 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.5751    29.6376   3.461  0.00212 ** 
edad          5.3207     0.7243   7.346 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,	Adjusted R-squared:  0.6882 
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

R^2

Coeficiente de determinación R^2

- El valor 0.7012 implica una relación lineal moderadamente fuerte entre la edad de la persona y la cantidad de grasas en sangre.
- Este valor indica que el modelo explica un 70.12 % de la variabilidad observada.
- Por lo tanto, un 29.88 % de la variabilidad observada no queda explicada por el modelo.

Coeficiente de determinación R^2

- Ejemplo en R y Python.

Organización

El problema

Diagnóstico de la regresión

Calidad del ajuste

Inferencia para los parámetros del modelo

Puntos influyentes

- Outliers
- Leverage

Distribucion de $\hat{\beta}_1$

Recordemos que el estimador de β_1 está dado por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Podemos reescribir $\hat{\beta}_1$ como:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

Propiedades de $\hat{\beta}_1$

Esta forma de escribir a $\hat{\beta}_1$ es útil para estudiar sus propiedades teóricas. Bajo los supuestos mencionados, se puede probar que

$$E(\hat{\beta}_1 | X = x) = \beta_1$$

$$Var(\hat{\beta}_1 | X = x) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

y de esta forma

$$\hat{\beta}_1 | X = x \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Propiedades de $\hat{\beta}_0$

Recordemos que el estimador de β_0 está dado por:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Bajo los supuestos mencionados, se puede probar que

$$\begin{aligned} E(\hat{\beta}_0 | X = x) &= \beta_0 \\ Var(\hat{\beta}_1 | X = x) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \end{aligned}$$

y, de esta forma

$$\hat{\beta}_0 | X = x \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \right).$$

Inferencia sobre los parámetros

- El conocer la distribución de $\hat{\beta}_0$ y $\hat{\beta}_1$ nos va a permitir hacer inferencia sobre ambos parámetros.
- Encontrar intervalos de confianza y hacer test de hipótesis para ambos parámetros.
- Para esto necesitamos un estadístico pivote.
- Vamos a llamar

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x}^2).$$

Inferencia para β_0

- El estadístico pivote es:

$$T = \frac{\hat{\beta}_0 - \beta_0}{S_r\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}}} = \frac{\hat{\beta}_0 - \beta_0}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$$

  donde, recordemos que, $S_r^2 = \dfrac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n-2}$,
  $\text{se}(\hat{\beta}_0) = S_r\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}}$ es el error estándar estimado de $\hat{\beta}_0$.

- El estadístico T es el pivote para hallar intervalos de confianza y hacer test de hipótesis para β_0.

Intervalo de Confianza para β_0

- Un intervalo de confianza de nivel $(1-\alpha)$ para β_0 es:

$$\left(\hat{\beta}_0 - t_{(\alpha/2,n-2)}\mathrm{se}(\hat{\beta}_0),\ \hat{\beta}_0 + t_{(\alpha/2,n-2)}\mathrm{se}(\hat{\beta}_0)\right) \tag{7}$$

donde $t_{(\alpha/2,n-2)}$ es el cuantil $(1-\alpha/2)$ de la distribución t de student con $n-2$ grados de libertad.

- Por lo tanto, la expresión de (7) es

$$\left[\hat{\beta}_0 + t_{n-2,\alpha/2} S_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}; \hat{\beta}_0 - t_{n-2,\alpha/2} S_r \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right]$$

Test de Hipótesis para β_0

- Queremos testear las hipótesis

$$H_0 : \beta_0 = \beta_0 \text{ vs. } \beta_0 \neq \beta_0.$$

- El estadístico de prueba es:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\mathrm{se}(\hat{\beta}_0)} \sim t_{n-2} \quad \text{si } H_0 \text{ es verdadera.}$$

- R provee automáticamente el valor de T y el p-valor asociado al test para estas hipótesis.

Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max
-63.478 -26.816  -3.854  28.315  90.881

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 102.5751    29.6376    3.461   0.00212 **
edad          5.3207     0.7243    7.346  1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

T valor $\hat{\beta}_0$

Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max
-63.478 -26.816  -3.854  28.315  90.881

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 102.5751    29.6376    3.461    0.00212 **
edad          5.3207     0.7243    7.346   1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

p-valor $\hat{\beta}_0$

Inferencia para β_1

- El estadístico pivote para β_1 es:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\frac{S_r}{\sqrt{S_{xx}}}} = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2}$$

donde $\mathrm{se}(\hat{\beta}_1) = \dfrac{S_r}{\sqrt{S_{xx}}}$ es el error estándar estimado de $\hat{\beta}_1$.

- El estadístico T es el pivote para hallar intervalos de confianza y hacer test de hipótesis para β_1.

Intervalo de Confianza para β_1

- Un intervalo de confianza de nivel $(1-\alpha)$ para β_1 es:

$$\left(\hat{\beta}_1 - t_{(\alpha/2,n-2)}\mathrm{se}(\hat{\beta}_1),\ \hat{\beta}_1 + t_{(\alpha/2,n-2)}\mathrm{se}(\hat{\beta}_1)\right),$$

donde $t_{(\alpha/2,n-2)}$ es el cuantil $(1-\alpha/2)$ de la distribución t de student con $n-2$ grados de libertad.

- Por lo tanto, la expresión de es:

$$\left[\hat{\beta}_1 - t_{n-2,\alpha/2}\frac{S_r}{S_{xx}};\hat{\beta}_1 + t_{n-2,\alpha/2}\frac{S_r}{S_{xx}}\right].$$

Test de Hipótesis para β_1

- Queremos testear las hipótesis

$$H_0 : \beta_1 = \beta_1 \text{ vs. } \beta_1 \neq \beta_1.$$

- El estadístico de prueba es:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\frac{S_r}{\sqrt{S_{xx}}}} = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2} \quad \text{si } H_1 \text{ es verdadera.}$$

- R provee automáticamente el valor de T y el p-valor asociado al test para estas hipótesis.

Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.478 -26.816  -3.854  28.315  90.881 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.5751    29.6376   3.461  0.00212 ** 
edad          5.3207     0.7243   7.346 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,	Adjusted R-squared:  0.6882 
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

T valor $\hat{\beta}_1$

Summary

```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max
-63.478 -26.816  -3.854  28.315  90.881

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 102.5751    29.6376   3.461    0.00212 **
edad          5.3207     0.7243   7.346   1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

p-valor $\hat{\beta}_1$

Inferencia para σ

- Se puede probar que el estadístico

$$W = \frac{(n-2)S_r^2}{\sigma^2} \sim \chi^2_{n-2}.$$

- De esta forma, un intervalo de confianza de nivel $(1-\alpha)$ para σ está dado por:

$$\left[\frac{(n-2)S_r^2}{\chi^2_{n-2,\alpha/2}}, \frac{(n-2)S_r^2}{\chi^2_{n-2,1-\alpha/2}}\right]$$

Intervalos de confianza para la respuesta media de Y para un $x = x_0$ fijo

- Queremos construir un intervalo de confianza para $E\left(Y_0 \mid X = x_0\right)$.
- Es decir, un intervalo de confianza para la respuesta media para algun valor prefijado de la variable respuesta en x_0.
- Observemos que x_0 puede ser o no ser un valor observado en la muestra. Pero siempre tiene que estar dentro del rango de valores observados para X, es decir, entre el mínimo y máximo valor observado para X.

Intervalos de confianza para la respuesta media de Y para un $X = x_0$ fijo

- El parámetro poblacional a estimar es, entonces
$$E\left(Y_0 \mid X = x_0\right) = \beta_0 + \beta_1 x_0.$$
- El estimador puntual está dado por
$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$
- Se puede probar que, un intervalo de confianza para el valor medio de Y para un $X = x_0$ fijo, es de la forma:

$$\widehat{Y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} \widehat{\sigma} \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}},$$

$$\widehat{Y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} S_r \sqrt{\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}.$$

Intervalos de predicción de una nueva observación de Y cuando $X = x_0$

- Queremos predecir una nueva observación Y correspondiente a un valor $X = x_0$ dado.
- Esta nueva observación debe ser obtenida en forma independiente de las observaciones con las que se definió el modelo.
- Tenemos ahora dos fuentes de variabilidad:
  - la incerteza en la estimación de $E(Y_0)$ alrededor de la cual estará la nueva observación.
  - la variabilidad de Y alrededor de su media, que proviene de su distribución.

Diferencia entre intervalo de confianza y de predicción

- El intervalo de confianza para la esperanza de Y condicional al valor de $X, E\left(Y_0 \mid X = x_0\right)$, es un procedimiento que nos permite encontrar, a partir de los datos, un intervalo de posibles valores que, con cierta probabilidad, puede tomar un parámetro poblacional, en este caso, la esperanza de Y cuando la variable X toma el valor x_0.
- El intervalo de predicción de una nueva observación Y_0 medida cuando $X = x_0$ es un procedimiento que nos permite encontrar, a partir de los datos, un valor que nos permita predecir el valor que puede tomar esa variable aleatoria.
- Este último procedimiento incorpora otra fuente de variablidad: la que proviene de la aleatoriedad de Y.

Intervalos de predicción de una nueva observación de Y cuando $X = x_0$

- Nuestra mejor predicción es nuevamente

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

- Pero ahora el error asociado será mayor. Estimamos el error estándar de la predicción con

$$\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}.$$

Intervalos de predicción de una nueva observación de Y cuando $X = x_0$

- A partir de este error estándar podemos construir un intervalo de predicción (que abreviaremos IP) de nivel $(1-\alpha)$ para el valor predicho de Y cuando $X = x_0$, que está dado por:

$$\widehat{Y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}},$$

$$\widehat{Y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} S_r \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}}.$$

Conclusión

- Intervalo para la respuesta media: mide la precisión en la estimación de la media de Y.
- Intervalo de predicción: incluye además la variabilidad individual de una nueva observación.

Conclusión:

Siempre el intervalo de predicción será más ancho que el intervalo para la media.

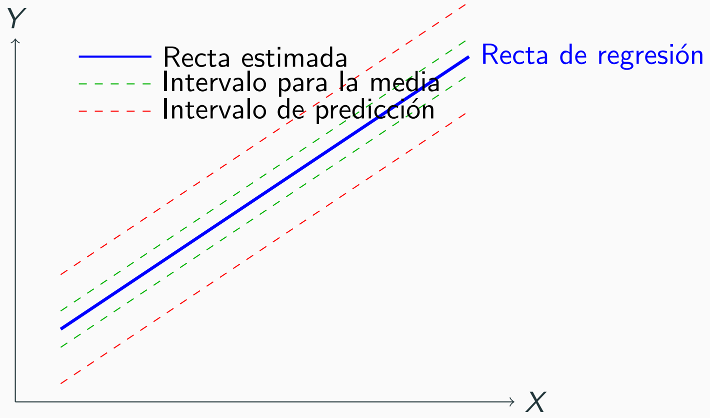
Ejemplo

- Volvamos al ejemplo de las mediciones de la edad, el peso y la cantidad de grasas en sangre, realizadas a 25 personas.
- La instrucción *summary* nos da un resumen de varios elementos de la regresión.

Ejemplo

- Veamos un ejemplo hecho en R y en python.

Comparación gráfica de los intervalos



Organización

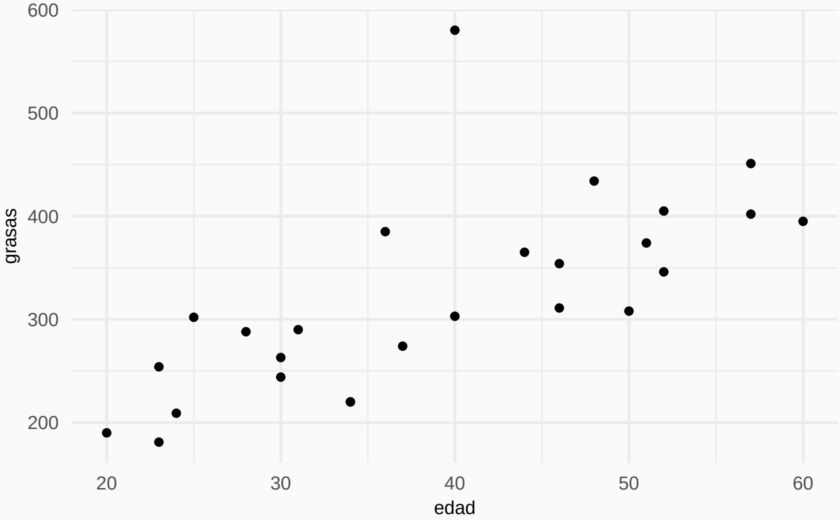
Diagnóstico del modelo: Puntos influyentes

- Es importante verificar que las estimaciones obtenidas no estén excesivamente influenciadas por una sola observación o por un pequeño grupo de datos.
- En este sentido, es necesario detectar la existencia de puntos que pueden influir en la determinación del modelo.
- Observaciones cuya presencia afecta de forma considerable las conclusiones se denominan puntos influyentes.
- Su detección es una parte fundamental del diagnóstico del modelo.

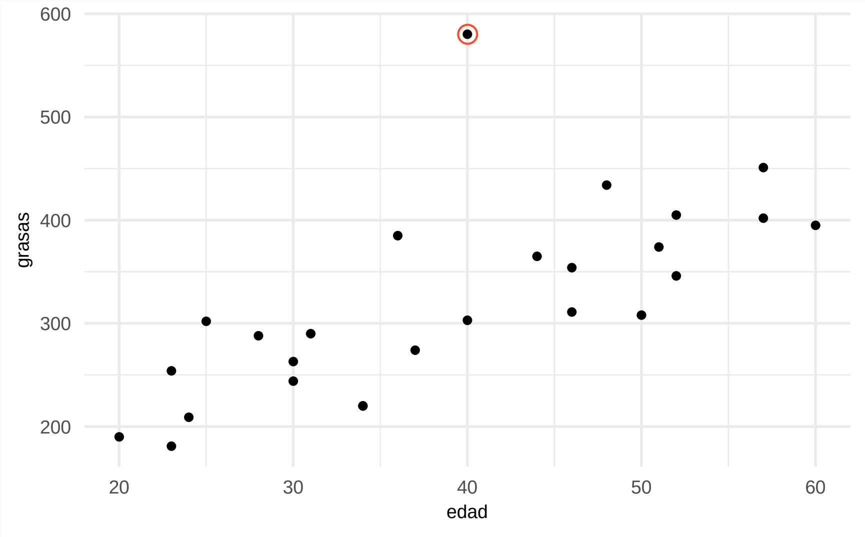
Outliers

- En algunos situaciones se observan respuestas que no se comportan como la mayoría de las observaciones.
- A este tipo de caso lo llamamos outlier o dato atípico.
- La presencia de estos puntos puede cambiar el modelo de regresión.
- Por este motivo, hay que detectarlos y estudiarlos.

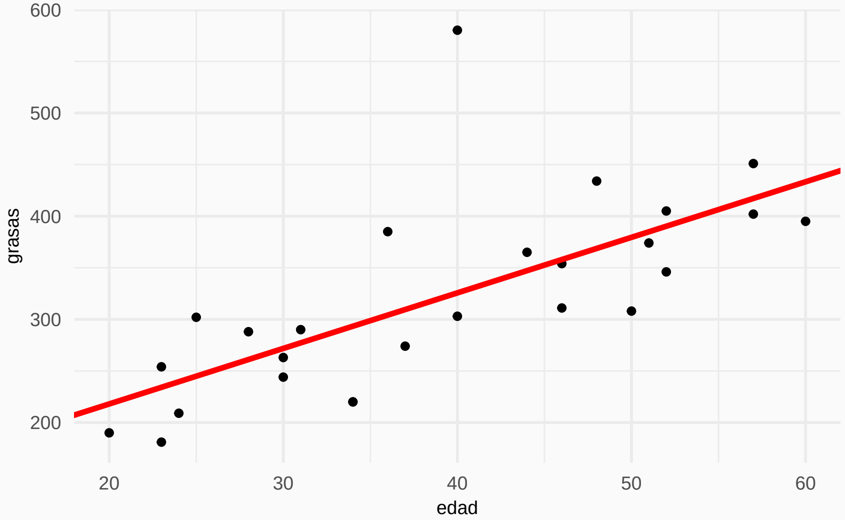
Outliers



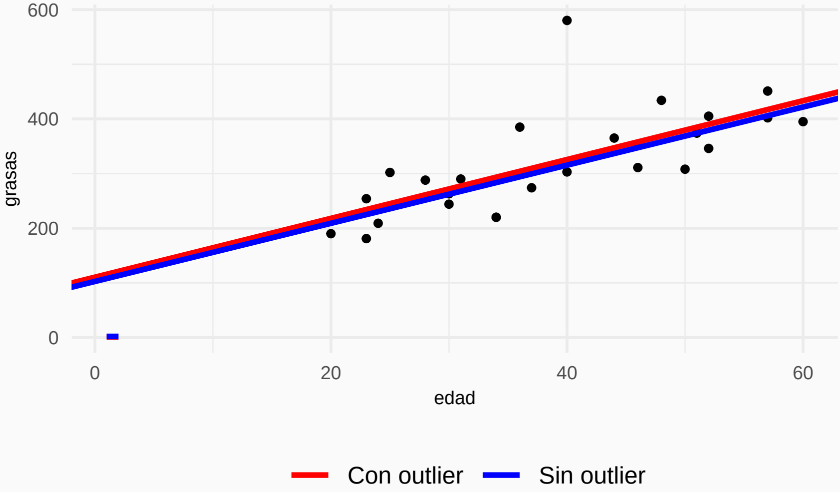
Outliers



Outliers



Outliers



Outliers

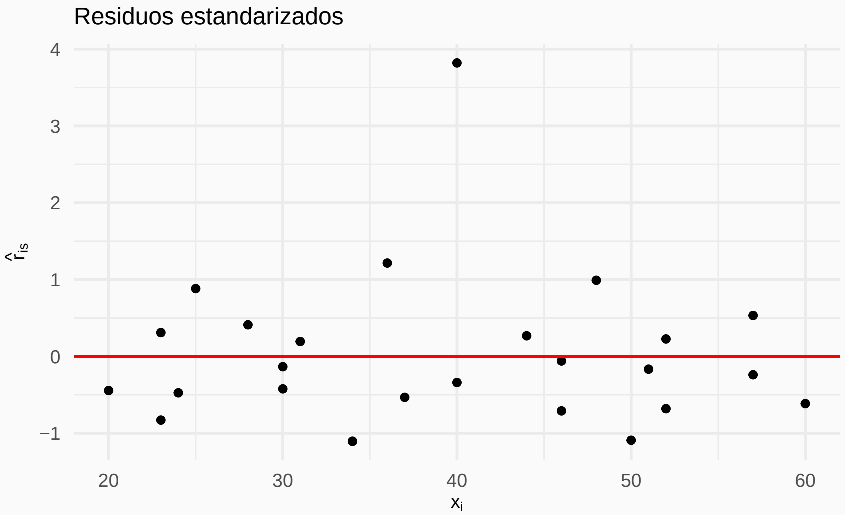
```
> summary(modelo)

Call:
lm(formula = grasas ~ edad, data = grasas)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.478 -26.816  -3.854  28.315  90.881 

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)    
(Intercept) 102.5751    29.6376   3.461    0.00212 ** 
edad          5.3207     0.7243   7.346   1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882 
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

```
> summary(regresion)

Call:
lm(formula = grasas ~ edad, data = datos)

Residuals:
   Min     1Q Median     3Q    Max 
-73.34 -37.60 -13.00  19.36 254.59 

Coefficients:
          Estimate Std. Error t value Pr(>|t|)    
(Intercept)  110.324     46.304   2.383   0.0255 *  
edad           5.383      1.133   4.753 7.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67.97 on 24 degrees of freedom
Multiple R-squared:  0.4849,    Adjusted R-squared:  0.4634 
F-statistic: 22.59 on 1 and 24 DF,  p-value: 7.784e-05
```

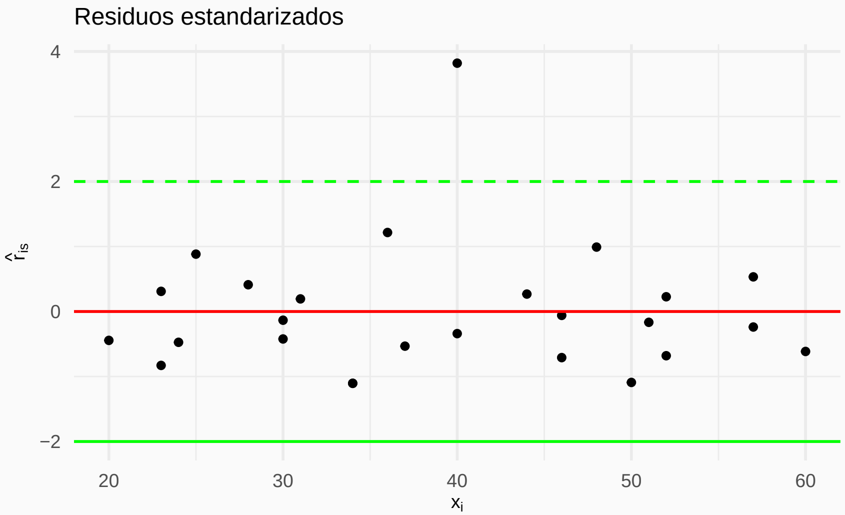
Outliers y su impacto en la regresión

- El concepto de outlier es *relativo al modelo específico considerado*.
- Si se modifica la forma del modelo, ese outlier puede dejar de serlo.
- Identificar outliers puede ser útil porque:
  - El método de cuadrados mínimos es muy sensible a observaciones alejadas.
  - Datos que se apartan mucho de la tendencia general pueden cambiar sustancialmente las estimaciones del modelo.

¿Cómo los detecto?



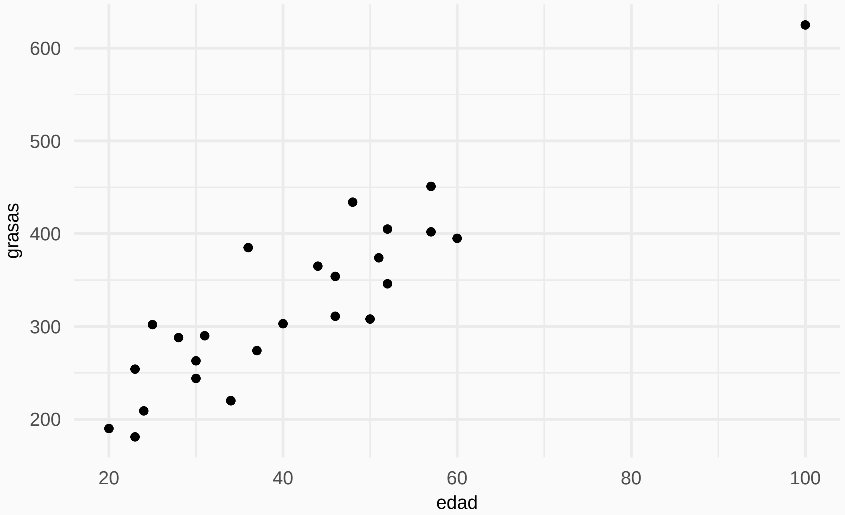
Cómo los detecto



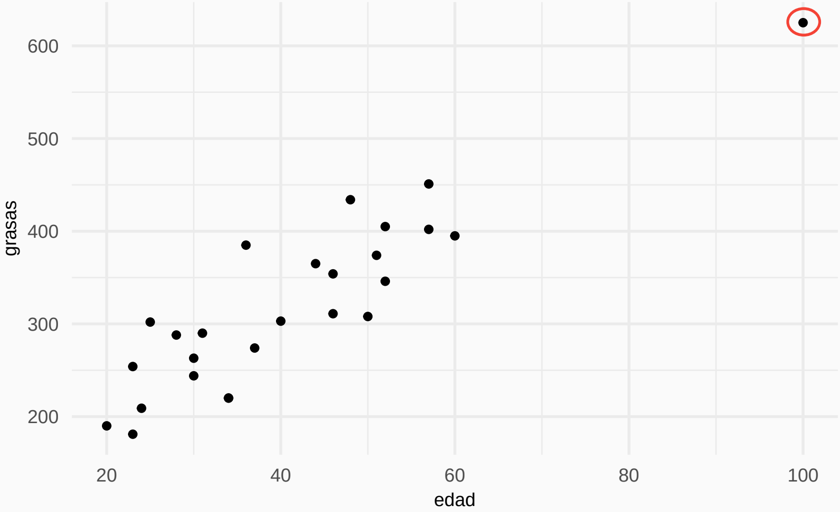
Outliers y su impacto en la regresión

- Estudiando las observaciones cuyo $|r_i| > 2$
- Si el outlier corresponde a un error de tipeo o un dato mal registrado, se lo puede eliminar o corregir.
- Si es correcto y no hay razones para eliminarlos, estudiar otros métodos como los modelos robustos.

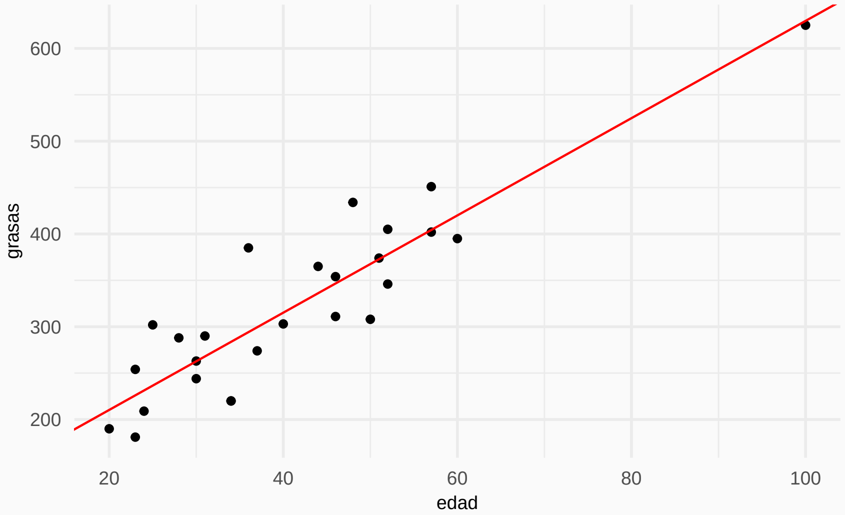
Leverage bueno



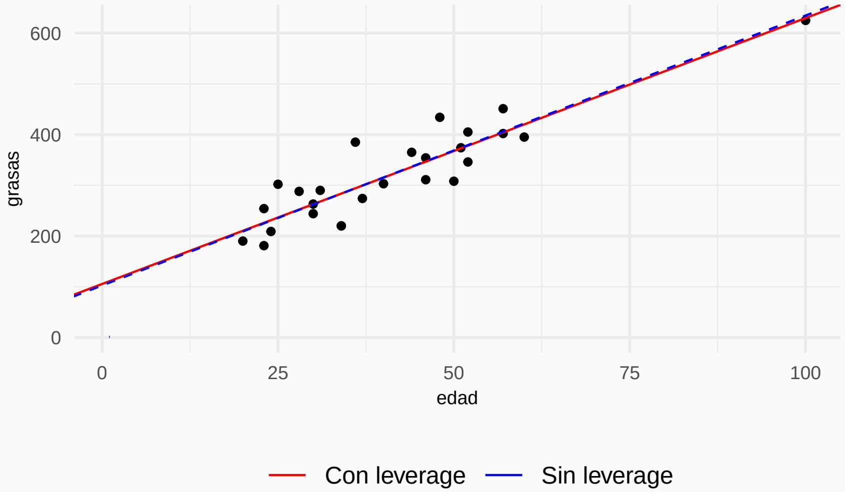
Leverage bueno



Regresión con leverage bueno



Regresión con y sin leverage bueno

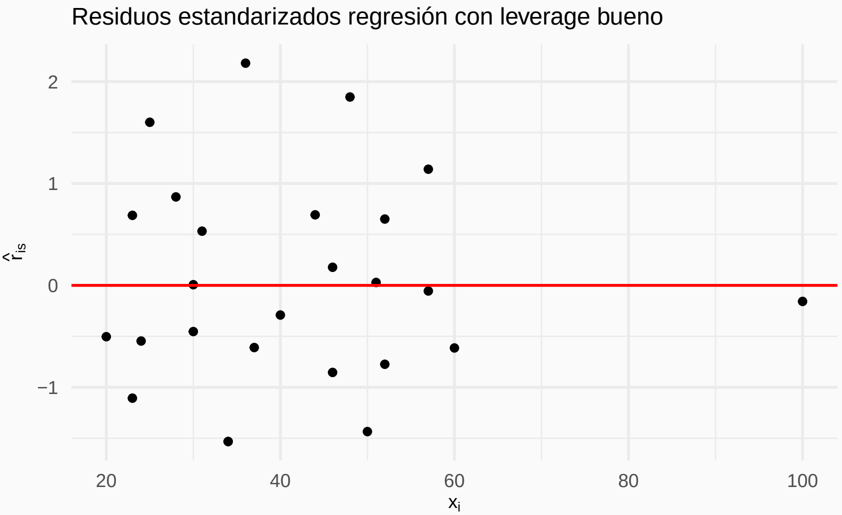


Summary regresión con y sin leverage bueno

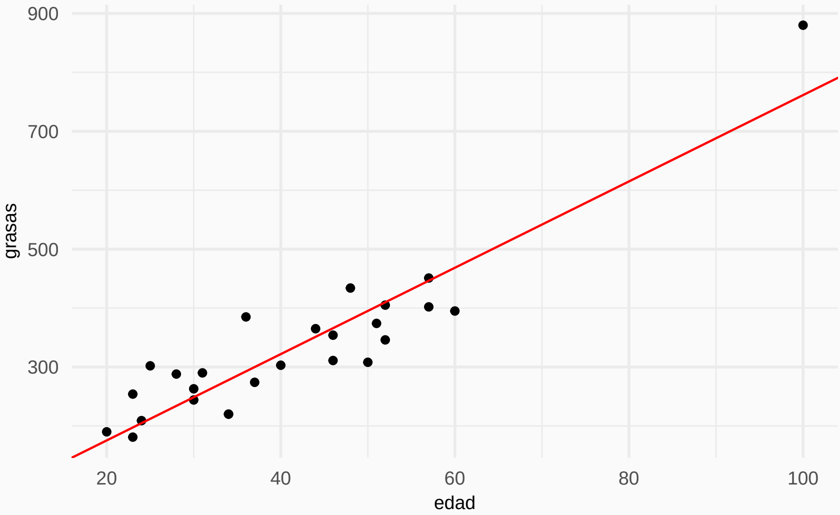
```
> summary(regresion.original) #Sin punto de alto leverage

Call:
lm(formula = grasas ~ edad, data = datos.originales)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.478 -26.816  -3.854  28.315  90.881 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.5751    29.6376   3.461  0.00212 ** 
edad          5.3207     0.7243   7.346 1.79e-07 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882 
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

```
> summary(regresion.ConLevBueno) #Con punto de alto leverage

Call:
lm(formula = grasas ~ edad, data = datos.ConLevBueno)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.695 -25.312  -3.459  27.712  90.821 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 105.4709    22.4607   4.696 9.00e-05 ***
edad          5.2419     0.5029  10.423 2.18e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 42.57 on 24 degrees of freedom
Multiple R-squared:  0.8191,    Adjusted R-squared:  0.8115 
F-statistic: 108.6 on 1 and 24 DF,  p-value: 2.175e-10
```

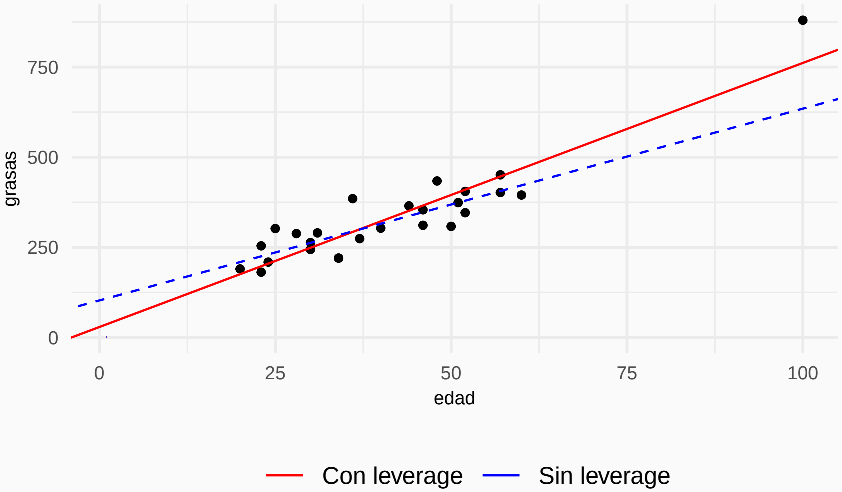
Residuos con leverage bueno



Regresión con leverage malo



Regresion con y sin leverage malo



Summary con y sin leverage malo

```
> summary(regresion.original) #Sin punto de alto leverage

Call:
lm(formula = grasas ~ edad, data = datos.originales)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.478 -26.816  -3.854  28.315  90.881 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 102.5751    29.6376   3.461  0.00212 ** 
edad          5.3207     0.7243   7.346 1.79e-07 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.46 on 23 degrees of freedom
Multiple R-squared:  0.7012,    Adjusted R-squared:  0.6882 
F-statistic: 53.96 on 1 and 23 DF,  p-value: 1.794e-07
```

```
> summary(regresion) # Con punto de alto leverage

Call:
lm(formula = grasas ~ edad, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max 
-87.163 -40.453  -4.734  29.165 118.566 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  28.8926    29.0073   0.996    0.329    
edad          7.3254     0.6495  11.279 4.46e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.98 on 24 degrees of freedom
Multiple R-squared:  0.8413,    Adjusted R-squared:  0.8347 
F-statistic: 127.2 on 1 and 24 DF,  p-value: 4.461e-11
```

¿Cómo los detecto?

- Habíamos visto que

$$Var(r_i) = \sigma^2(1 - h_{ii}), \text{ donde}$$
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

- Valores grandes de h_{ii} podrían interferir en la definición del modelo.
- Al valor h_{ii} (palanca) se llama leverage de la observación i-ésima.

Criterio leverage alto

- Analicemos a las obsevaciones con h_{ii} alto.
- ¿Qué es alto?
- Se puede probar que $\sum_{i=1}^{n} h_{ii} = p$ donde p es la cantidad de parámetros a estimar en la regresión. En regresión lineal simple $p = 2$.
- Por lo tanto, el criterio que tomaremos para detectar observaciones con leverage alto es:

$$h_{ii} > 2\bar{h} > 2\frac{p}{n} > \frac{4}{n},$$

donde $\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n}$.

Leverage

- Observaciones:
  - Cuánto mayor sea h_{ii} menor será la varianza del r_i.
  - Mientras más cercano a uno sea h_{ii} más cercana a cero será la varianza del residuo de la observación i-ésima.
  - Observaciones con gran h_{ii} implica que $\widehat{Y}_i$ tenderá a estar cerca del valor observado Y_i, sin importar cuánto sea el valor Y_i observado.

Diagnóstico de puntos de alto leverage

- El leverage mide cuánto se aleja una observación del centro de las X.
- Un punto de alto leverage tiene potencial de influir fuertemente sobre la recta de regresión.
- No todo punto con leverage alto es malo: depende también de su residuo.
  - Alto leverage bueno: alto leverage pero residuo pequeño $\Rightarrow$ no distorsiona el modelo.
  - Alto leverage malo: alto leverage y residuo grande $\Rightarrow$ influye fuertemente.

Distancia de Cook

La distancia de Cook D_i mide la influencia de la observación i sobre todos los valores ajustados del modelo.

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{Y}_j - \widehat{Y}_{(i)j} \right)^2}{2\widehat{\sigma^2}}$$

$$D_i = \frac{r_{si}^2}{2} \frac{h_{ii}}{1 - h_{ii}}$$

Distancia de Cook

Donde

- $\widehat{Y}_j$: es la predicción con todos los datos.
- $\widehat{Y}_{(i)j}$: es la predicción sin la observación i.
- pp: número de parámetros en el modelo (incluyendo el intercept).
- $\widehat{\sigma^2}$: estimación de la varianza del error.
- n: número total de observaciones.

Criterios de interpretación

- Un criterio que vamos a adoptar es:
  - Si $D_i <$ percentil 0,20 de la distribución $F_{(2,n-2)}$ $(F_{0,2(2,n-2)})$, la observación no es influyente.
  - Si $D_i >$ percentil 0,50 de la distribución $F_{(2,n-2)}$ $(F_{0,5(2,n-2)})$, la observación es muy influyente y puede requerir acción.
  - Si D_i está entre ambos percentiles, se recomienda observar también otros estadísticos de influencia.

Criterios de interpretación

- La distancia de Cook combina:
  - Leverage: qué tan lejos está un punto en el eje x respecto de la media de las observaciones en ese eje.
  - Residuo: qué tan lejos está la observación en el eje y respecto del valor predicho.
- Un punto con
  - leverage alto y residuo grande $\rightarrow$ alta distancia de Cook.
  - sólo leverage alto pero residuo pequeño $\rightarrow$ baja distancia de Cook (punto "bueno")