4 Introducción a la regresión

Métodos empíricos 2

03/05/2022

Hoy

- Intuiciones
- Líneas
- Regresión lineal

Intuiciones

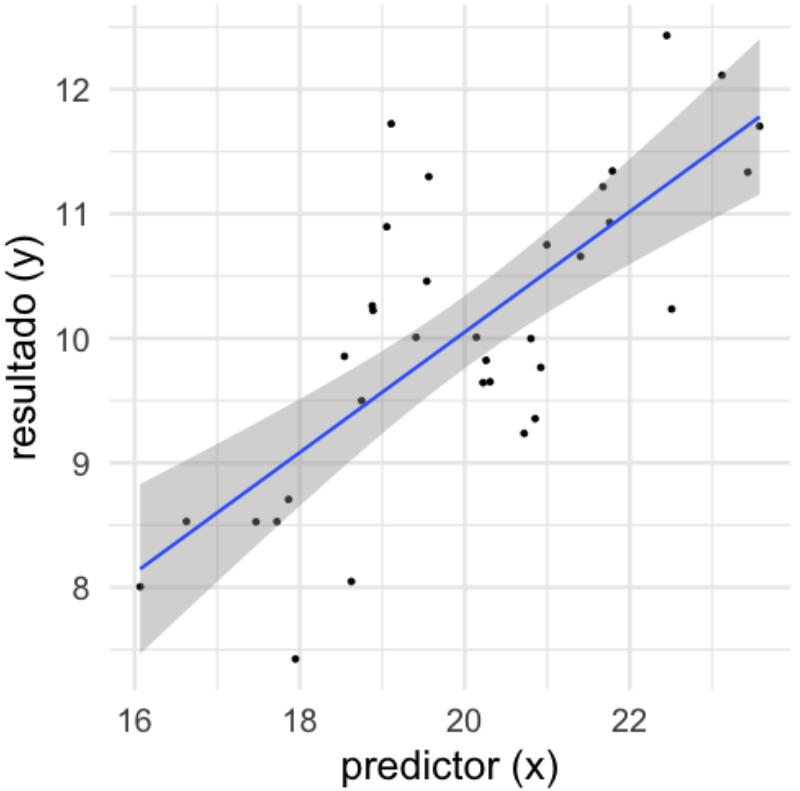
Regresión

Técnica fundamental para predecir un **resultado** a base de uno o más **predictores**

- Predicción
- Exploración de asociaciones
- Extrapolación
- Inferencia causal

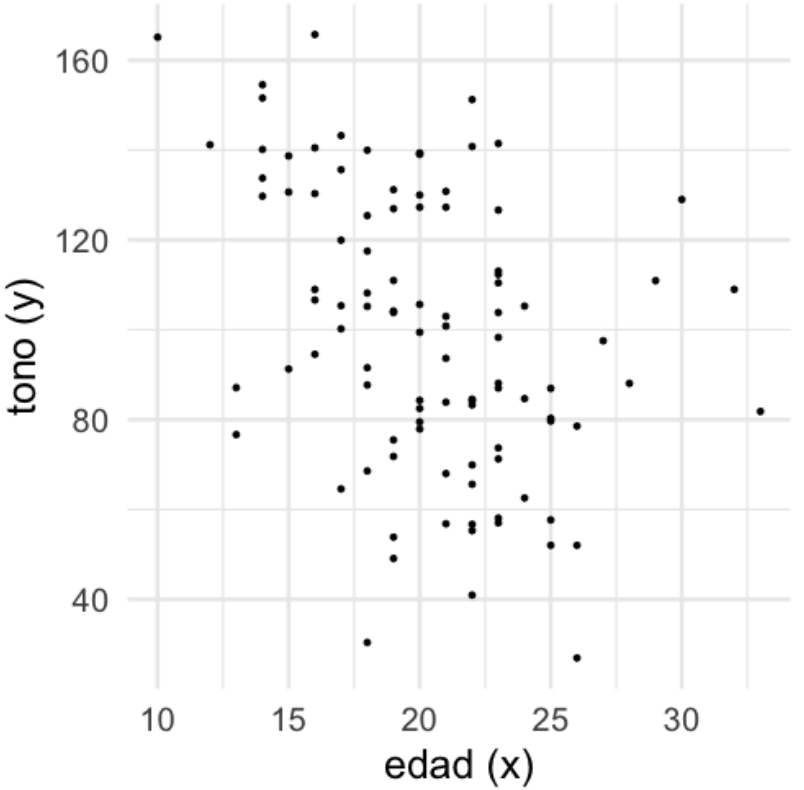
Regresión lineal

Estimación de relación **lineal** entre resultado (y) y uno o más predictores (x). Otra formulación: Estimación de predicción de y a base de x.



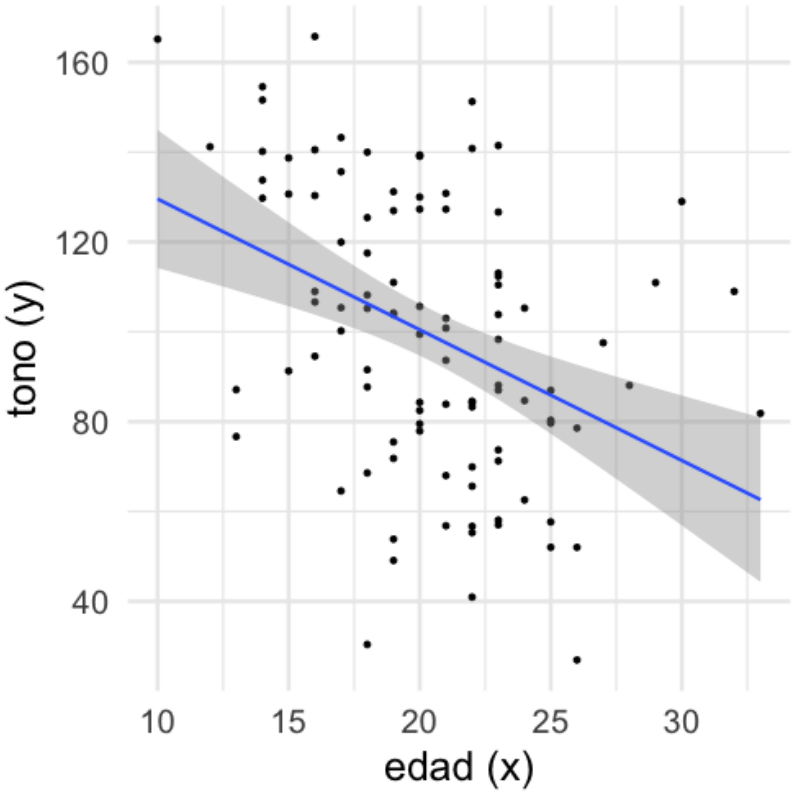
Regresión lineal (ejemplo)

Estimación de relación **lineal** entre resultado (*tono*) y uno o más predictores (*sexo*; *contexto*; *edad*). Otra formulación: Estimación de predicción de *tono* a base de *sexo* y/o *contexto* y/o *edad*.



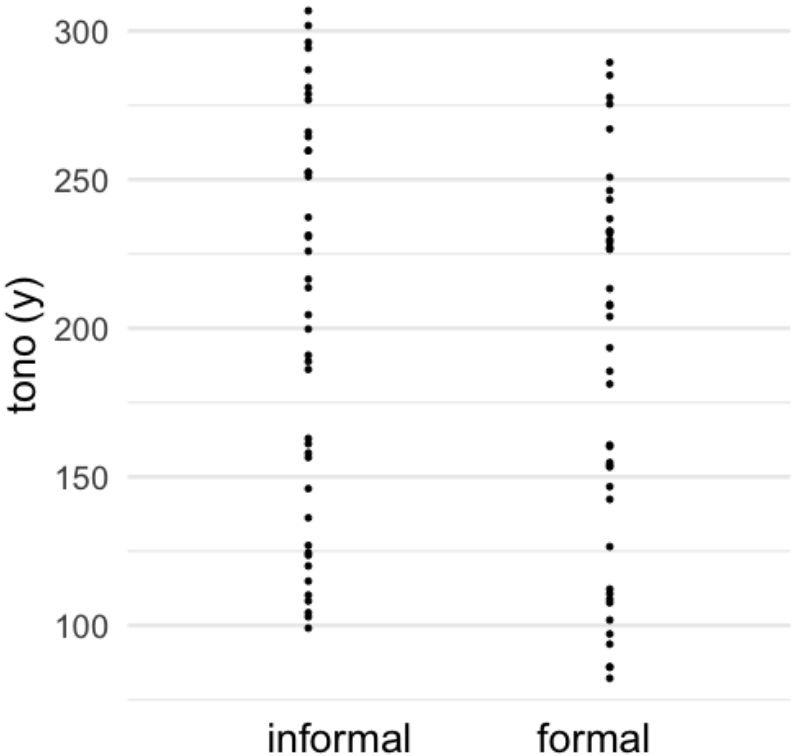
Regresión lineal (ejemplo)

Estimación de relación **lineal** entre resultado (*tono*) y uno o más predictores (*sexo*; *contexto*; *edad*). Otra formulación: Estimación de predicción de *tono* a base de *sexo* y/o *contexto* y/o *edad*.



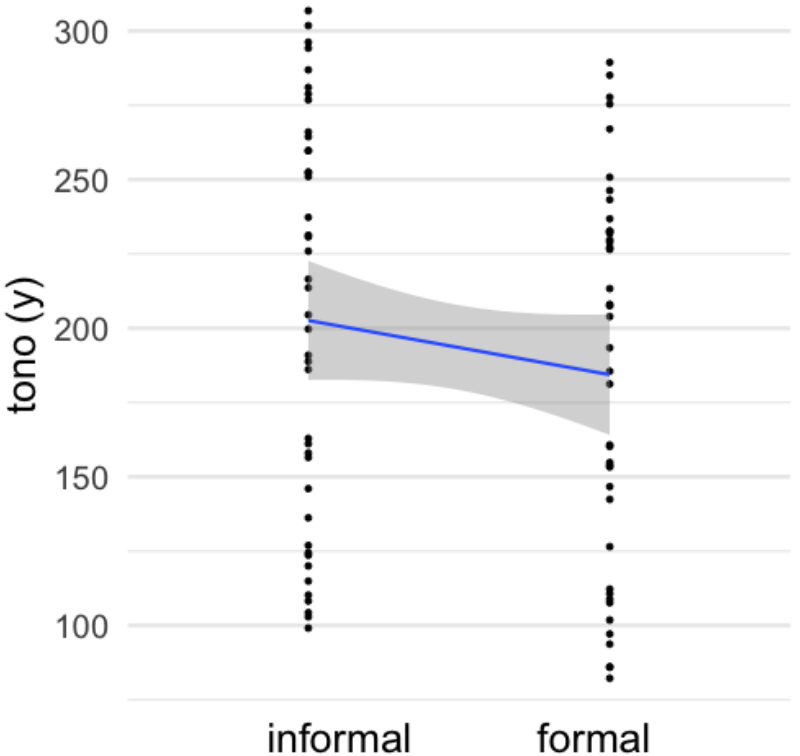
Regresión lineal (ejemplo)

Estimación de relación **lineal** entre resultado (*tono*) y uno o más predictores (*sexo*; *contexto*; *edad*). Otra formulación: Estimación de predicción de *tono* a base de *sexo* y/o *contexto* y/o *edad*.

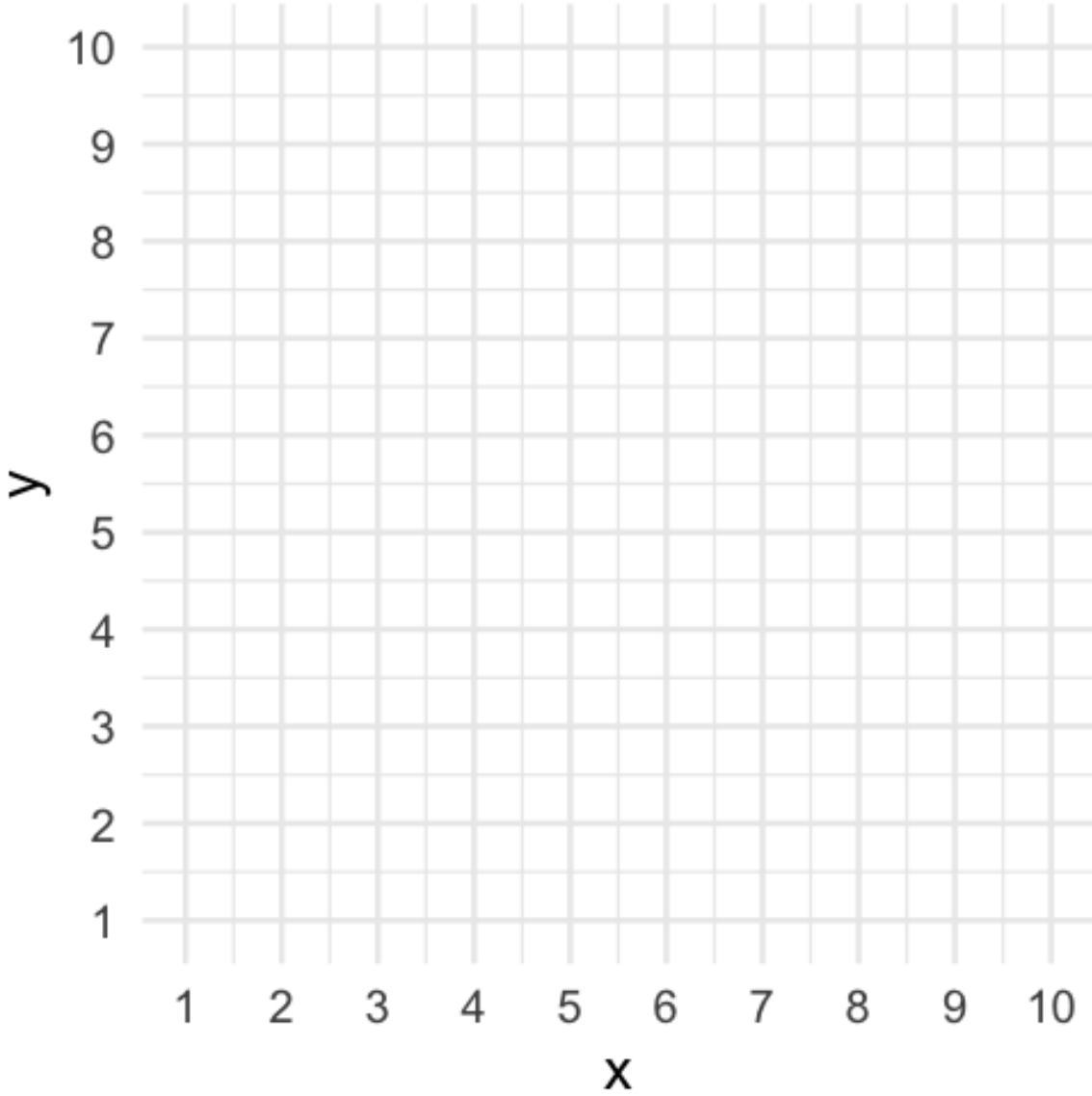


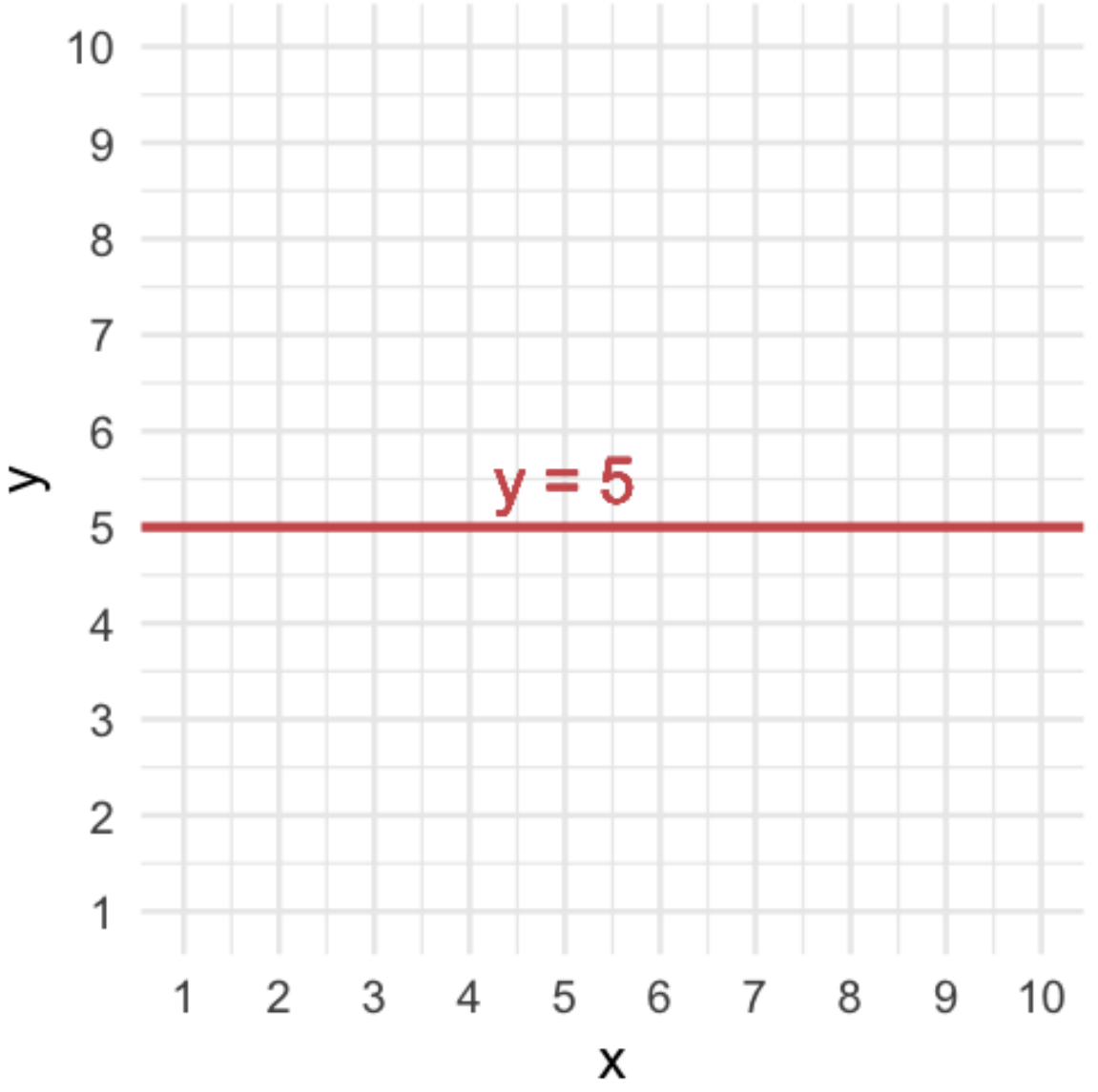
Regresión lineal (ejemplo)

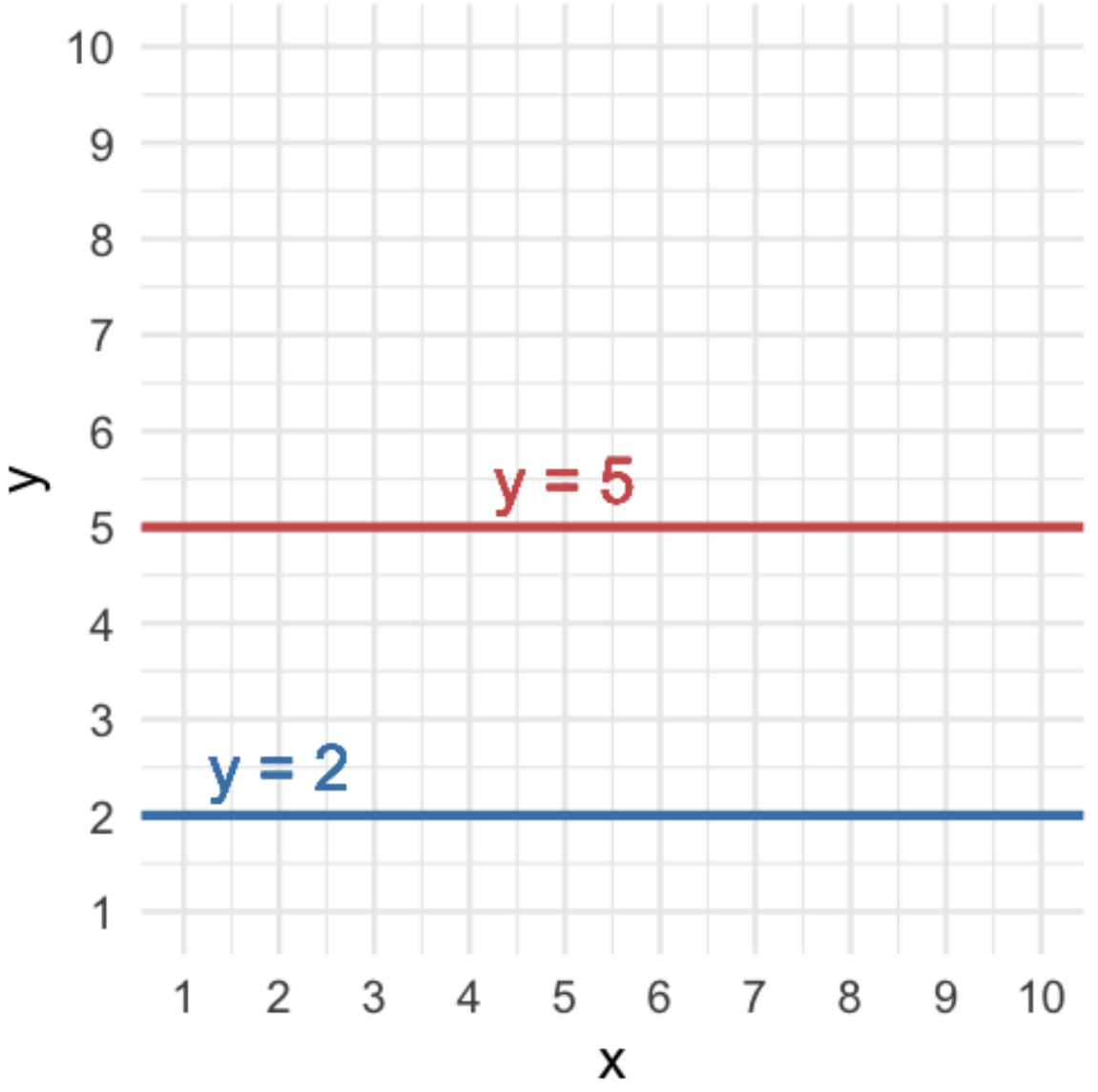
Estimación de relación **lineal** entre resultado (*tono*) y uno o más predictores (*sexo*; *contexto*; *edad*). Otra formulación: Estimación de predicción de *tono* a base de *sexo* y/o *contexto* y/o *edad*.

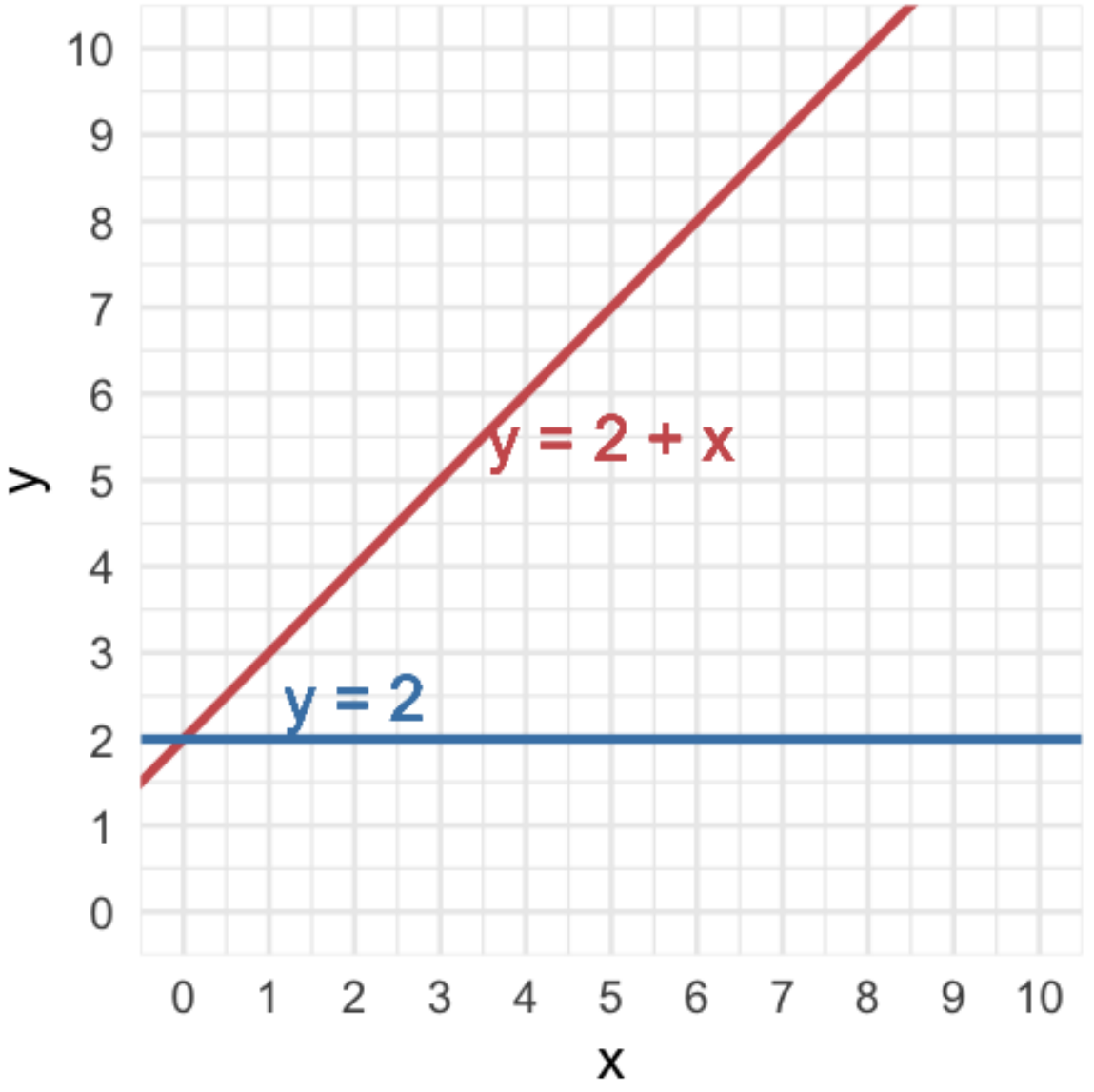


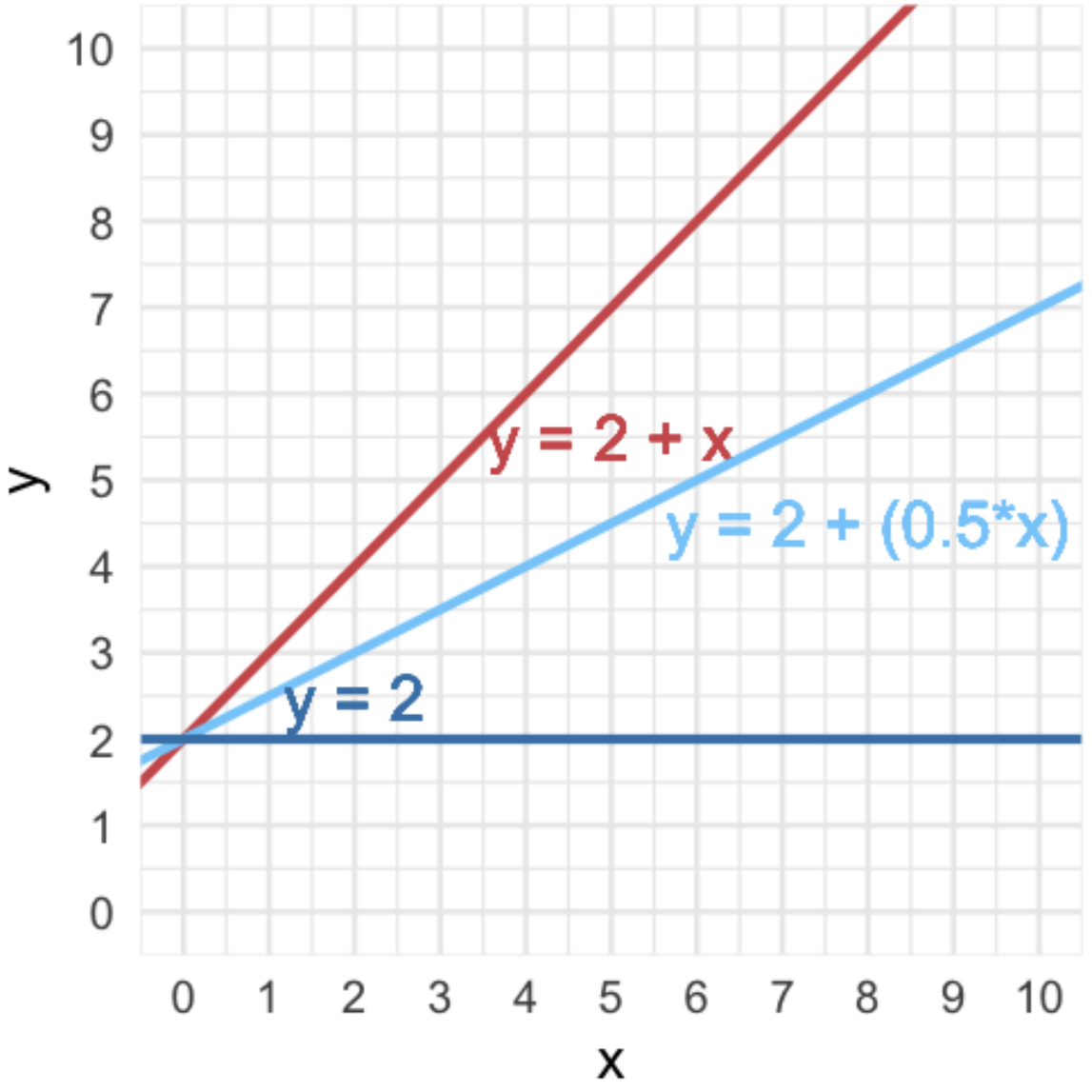
Repasando líneas


10
9
8
7
6
5
4
3
2
1
y
1
2
3
4
5
6
7
8
9
10
x

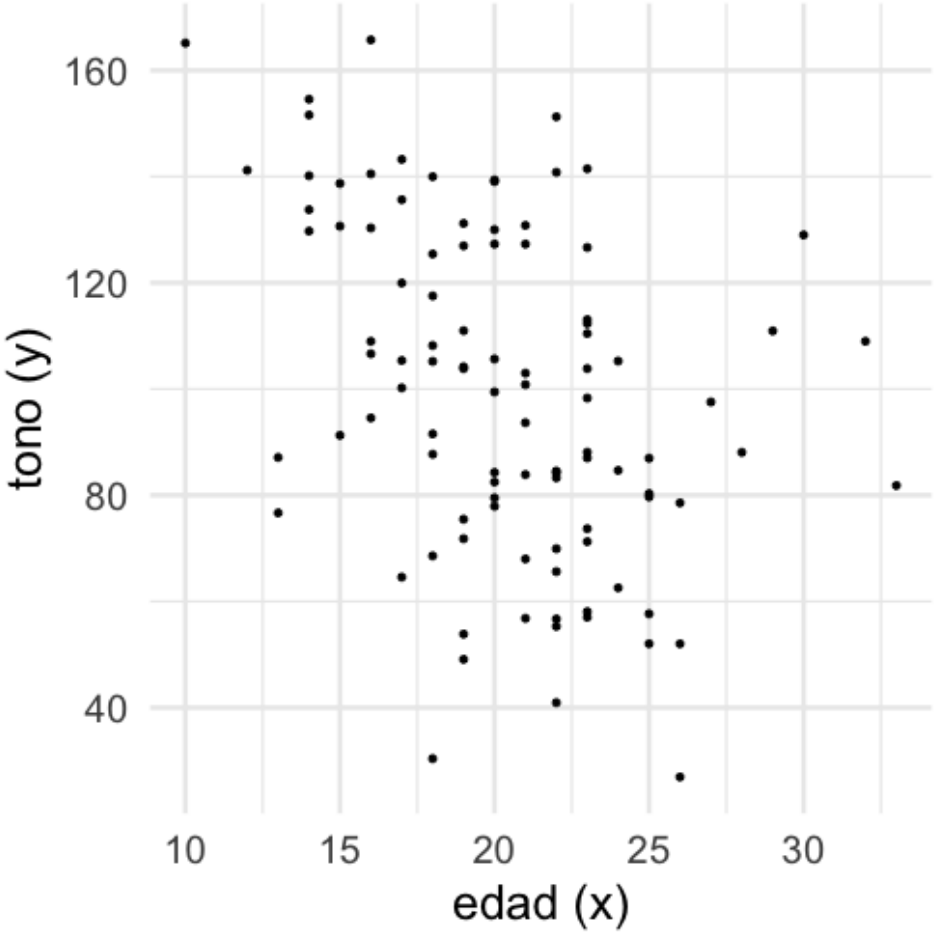

y = 5
y
x
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10


10
9
8
7
6
y
y = 5
5
4
3
y = 2
2
1
1 2 3 4 5 6 7 8 9 10
x

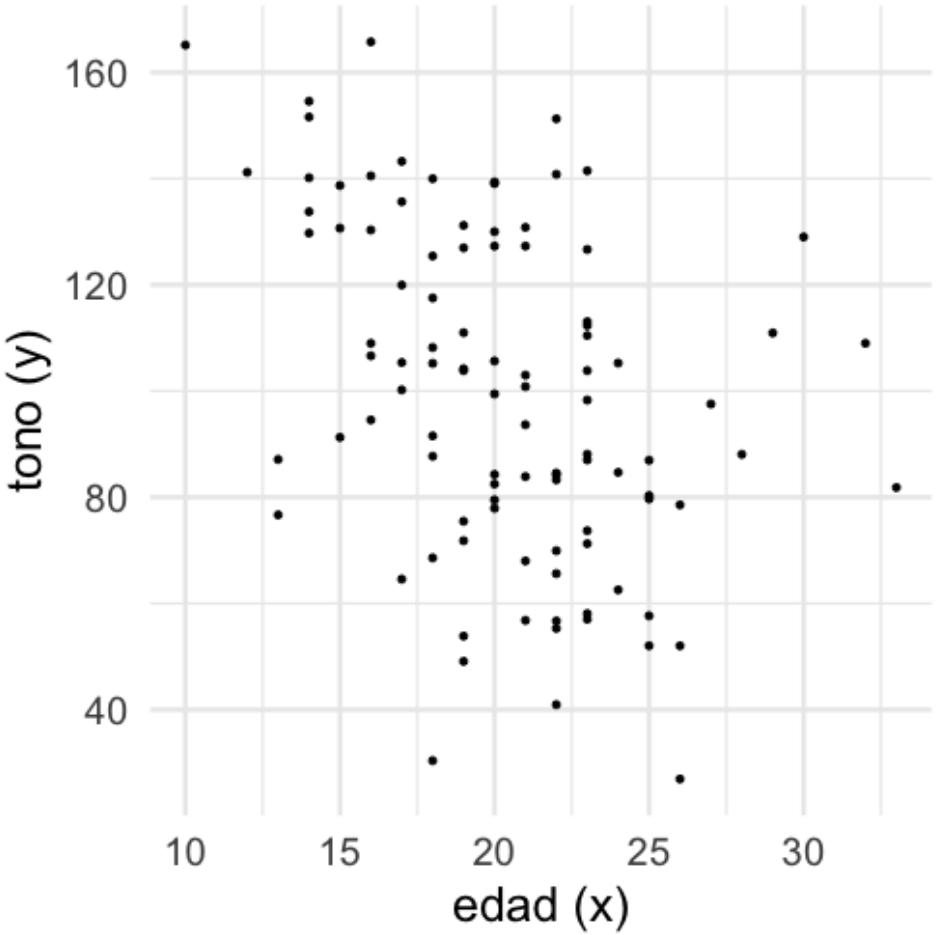

y = 2 + x
y = 2
y
x
0
1
2
3
4
5
6
7
8
9
10


y = 2 + x
y = 2 + (0.5*x)
y = 2
y
x
0
1
2
3
4
5
6
7
8
9
10

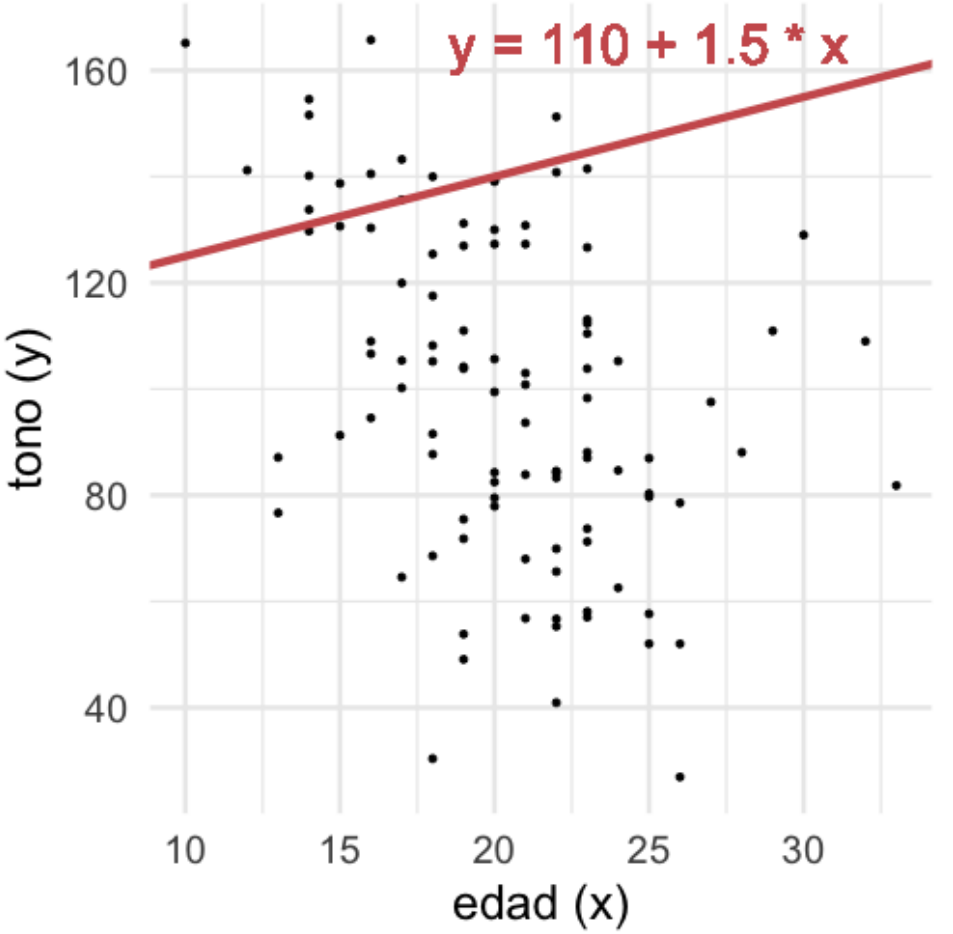
tono en función a edad



$$\text{tono} = \beta_0 + (\beta_1 \times \text{edad})$$



$\text{tono} = 110 + (1.5 \times \text{edad})$?

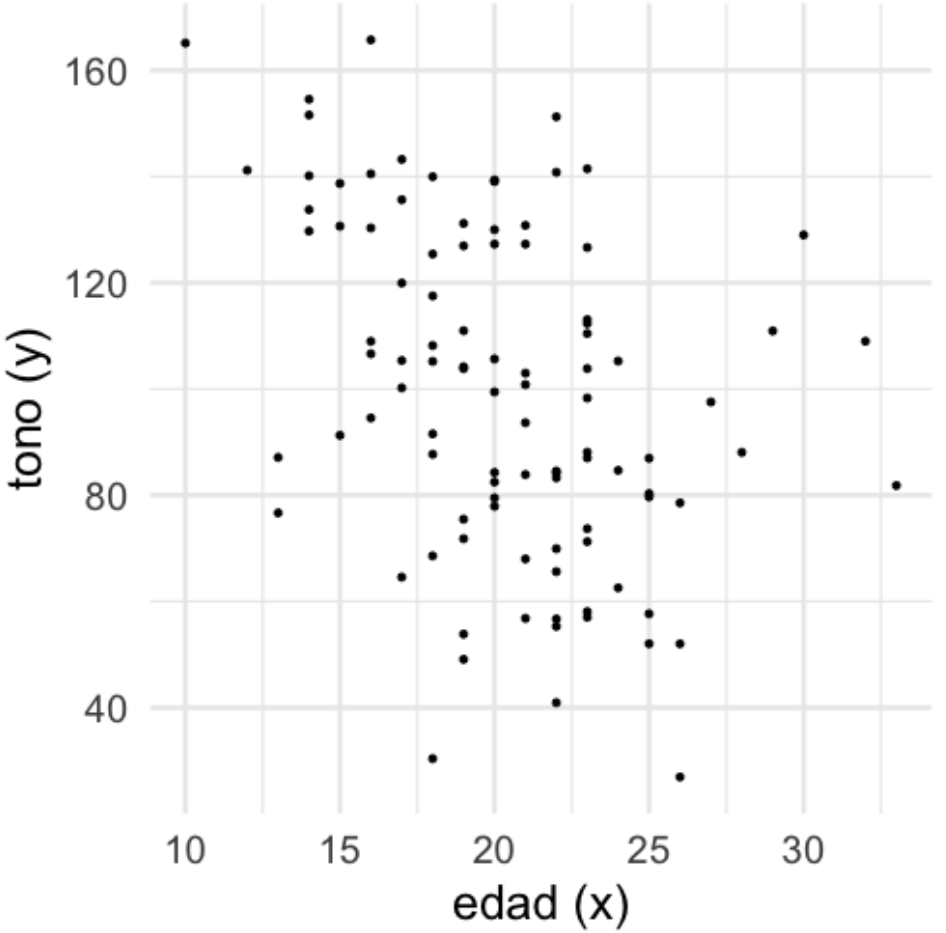


Intuición del algorítmo de regresión

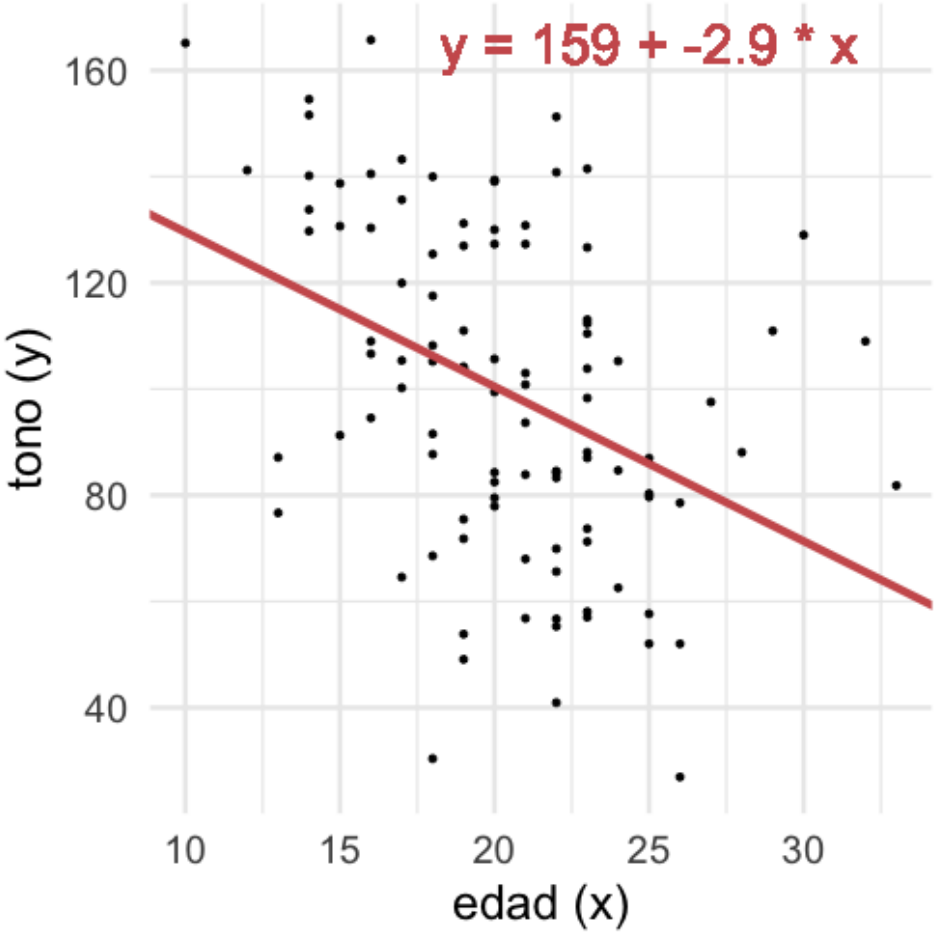
Busca la línea que minimiza el error (cuadrado)

En otras palabras: Busca la línea que, en promedio, está lo más cerca de los puntos

Dónde pondríais la línea?



$$\text{tono} = 159 + (-2.9 \times \text{edad})$$



Regresión lineal

Regresión lineal

Estimado de relación **lineal** entre resultado (y) y uno o más predictores.

Con un solo predictor x:

$$y_i \sim \text{Normal}(\mu_i, \sigma),$$
$$\mu_i = \beta_0 + \beta_1 x_i$$

Regresión lineal

Estimado de relación **lineal** entre *tono* (y) y uno o más predictores.

Con un solo predictor *edad*:

$$y_i \sim \text{Normal}(\mu_i, \sigma),$$
$$\mu_i = \beta_0 + \beta_1 \text{edad}_i$$

Regresión lineal (R)

```
head(df)
```

```
##   ages     pitch
## 1   18 139.99208
## 2   20  77.92071
## 3   20  82.48851
## 4   19 126.96205
## 5   16 108.97956
## 6   23  98.30374
```

```
nrow(df)
```

```
## [1] 100
```

Regresión lineal (R)

```
pitch_model <- lm(formula = pitch ~ 1 + ages,
                  data    = df)
pitch_model
```

```
##
## Call:
## lm(formula = pitch ~ 1 + ages, data = df)
##
## Coefficients:
## (Intercept)         ages
##     158.682       -2.911
```

$$tono_i = 159 + (-2.9 \times \text{edad}_i)$$

¿Cuál es la predicción del tono esperado de una persona de 12 años de edad? ¿Y de una de 25?

Más allá de predicciones

```
summary(pitch_model)
```

```
##
## Call:
## lm(formula = pitch ~ 1 + ages, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.888 -21.364  -0.273  21.562  57.647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.6819    14.3871  11.029  < 2e-16 ***
## ages         -2.9107     0.6925  -4.203 5.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.99 on 98 degrees of freedom
## Multiple R-squared:  0.1527,    Adjusted R-squared:  0.1441
## F-statistic: 17.67 on 1 and 98 DF,  p-value: 5.814e-05
```

Residuales

Differencia entre predicción y resultado real.

```
summary <- summary(pitch_model)
summary$residuals[1:5]
```

```
##          1          2          3          4          5
##  33.702507 -22.547484 -17.979687  23.583169  -3.131381
```

Residuales

```
plot(pitch_model$residuals)
```

Más allá de predicciones

```
summary(pitch_model)
```

```
##
## Call:
## lm(formula = pitch ~ 1 + ages, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.888 -21.364  -0.273  21.562  57.647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.6819    14.3871  11.029  < 2e-16 ***
## ages         -2.9107     0.6925  -4.203 5.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.99 on 98 degrees of freedom
## Multiple R-squared:  0.1527,    Adjusted R-squared:  0.1441
## F-statistic: 17.67 on 1 and 98 DF,  p-value: 5.814e-05
```

Error estándar (standard error)

Incertidúmbre del modelo al respecto de un parámetro

```
summary$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 158.681938 14.3871099 11.029452 7.169041e-19
## ages         -2.910687  0.6924807 -4.203275 5.814231e-05
```

Más allá de predicciones

```
summary
```

```
##
## Call:
## lm(formula = pitch ~ 1 + ages, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.888 -21.364  -0.273  21.562  57.647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.6819    14.3871  11.029  < 2e-16 ***
## ages         -2.9107     0.6925  -4.203 5.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.99 on 98 degrees of freedom
## Multiple R-squared:  0.1527,    Adjusted R-squared:  0.1441
## F-statistic: 17.67 on 1 and 98 DF,  p-value: 5.814e-05
```

Residual standard error

El valor (estimado) del error del modelo. Corresponde a σ

$$y \sim \text{Normal}(\mu, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n$$

Goodness of fit

R^2 es la proporción de la varianza del resultado que explican los predictores.

```
summary$r.squared
```

```
## [1] 0.152744
```

Más allá de predicciones

```
summary
```

```
##
## Call:
## lm(formula = pitch ~ 1 + ages, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.888 -21.364  -0.273  21.562  57.647
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.6819    14.3871  11.029  < 2e-16 ***
## ages         -2.9107     0.6925  -4.203 5.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.99 on 98 degrees of freedom
## Multiple R-squared:  0.1527,    Adjusted R-squared:  0.1441
## F-statistic: 17.67 on 1 and 98 DF,  p-value: 5.814e-05
```

Próxima sesión

- Entrega de "Assignment 4" (08:00 AM 10/05)
- Entrega de parte I de "Revisión por pares" (08:00 AM 10/05)

- **Regresión con más de un predictor**

- Assignment 5: 10/05 - 17/05
- Ejercicio práctico: 17/05 - 24/05
- Entrega parte II de "Revisión por pares": 24/05 - 31/05
- Informe final: 28/06

Transformaciones: Centrar

Centrar datos implica restar una constante a todos los valores de una variable

```
mean(df$ages)                                #promedio de edades
```

```
## [1] 20.35
```

```
df$ages.cent <- df$ages - mean(df$ages) #var. centrado
lm(df$pitch ~ df$ages.cent)             #coefficiente centrado
```

```
##
## Call:
## lm(formula = df$pitch ~ df$ages.cent)
##
## Coefficients:
##  (Intercept)  df$ages.cent
##       99.449        -2.911
```