

# 4 Introducció a la regressió

Mètodes empírics 2

02/05/2023

# Avui

- Aprofitant distribucions (R)
- Intuïcions
- Línies
- Regressió lineal

# Aprofitant distribucions (R)

**Anomena una de les tres distribucions que vam veure la setmana passada. Quants/Quins paràmetres la defineixen? I què controlen aquests paràmetres?**

# Intuïcions

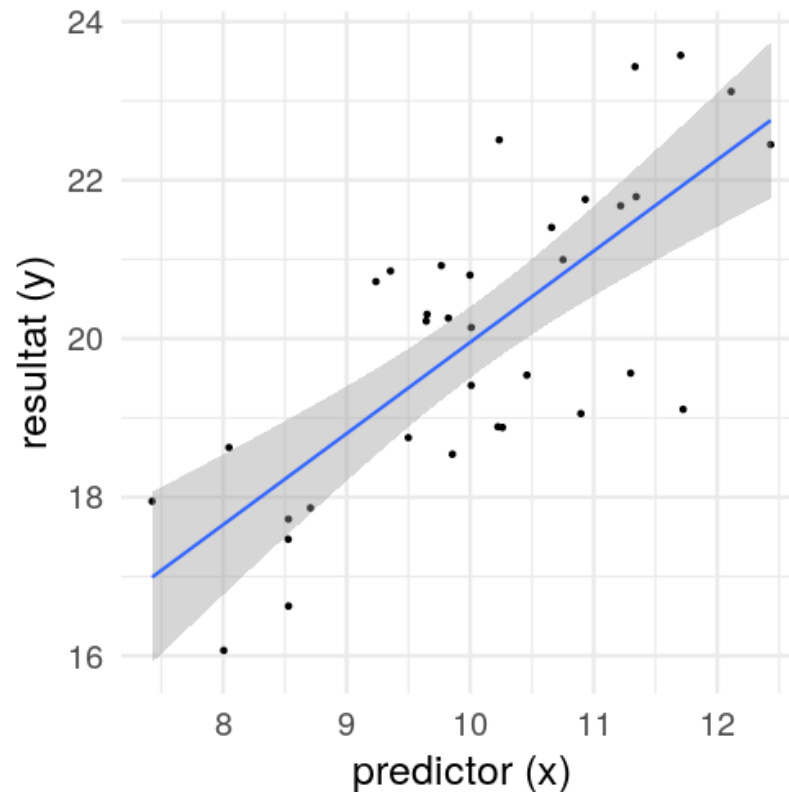
# Regressió

Tècnica fonamental per predir un **resultat** a base d'un o més **predictors**

- Predicció
- Exploració d'associacions
- Extrapolació
- Inferència causal

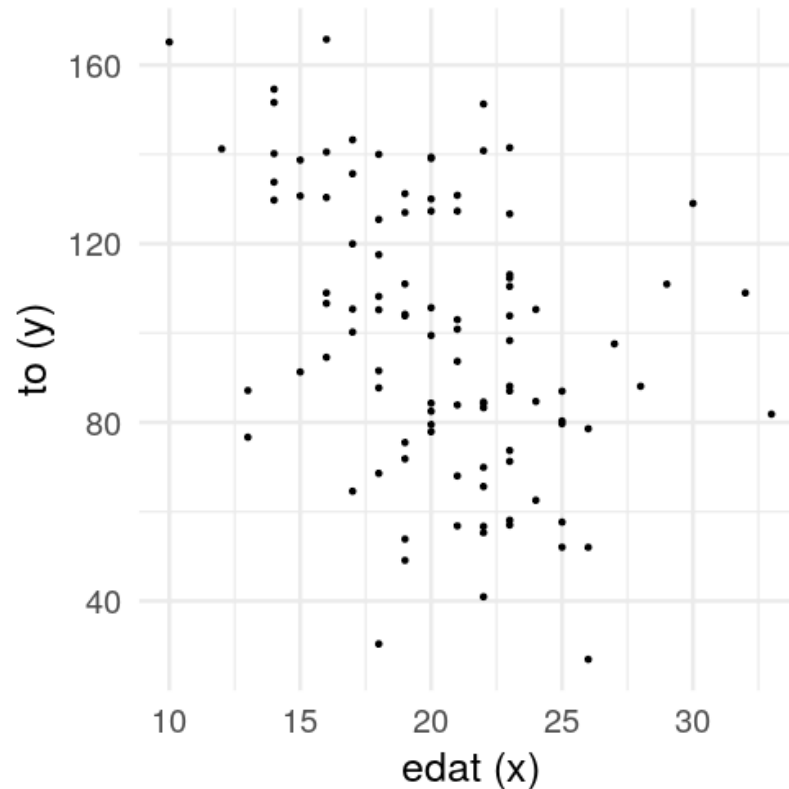
# Regressió lineal

Estimació de relació **lineal** entre resultat (  $y$  ) i un o més predictors (  $x$  ). Una altra formulació: Estimació de predicció de  $y$  a base de  $x$ .



# Regressió lineal (exemple)

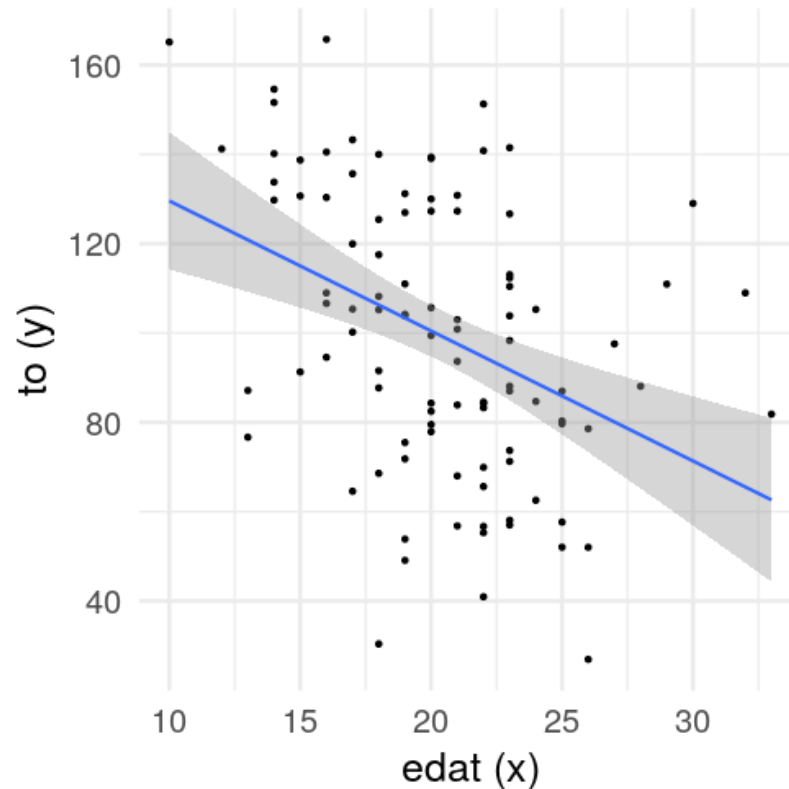
Estimació de relació **lineal** entre resultat ( *to* ) i un o més predictors ( *sexe*; *context*; *edat* ). Una altra formulació: Estimació de predicció de *to* a base de *sexe* i/o *context* i/o *edat*.





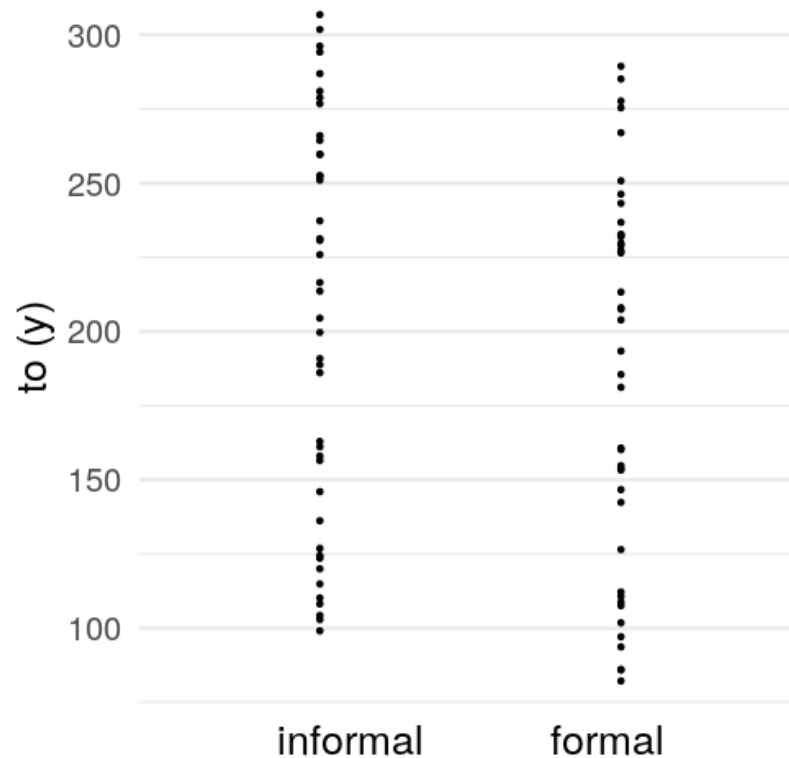
# Regressió lineal (exemple)

Estimació de relació **lineal** entre resultat ( *to* ) i un o més predictors ( *sexe*; *context*; *edat* ). Una altra formulació: Estimació de predicció de *to* a base de *sexe* i/o *context* i/o *edat*.



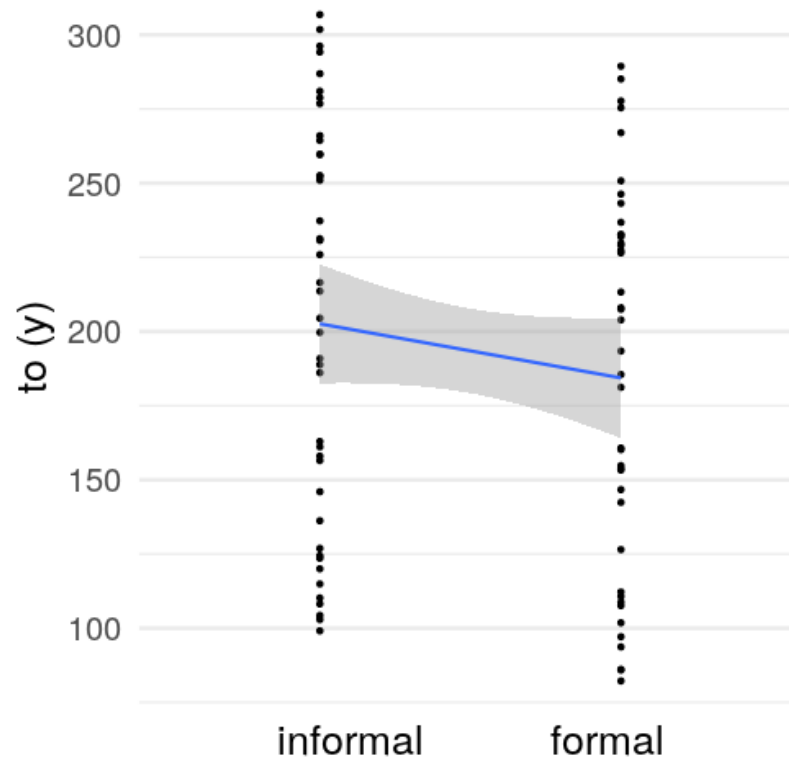
# Regressió lineal (exemple)

Estimació de relació **lineal** entre resultat ( *to* ) i un o més predictors ( *sexe*; *context*; *edat* ). Una altra formulació: Estimació de predicció de *to* a base de *sexe* i/o *context* i/o *edat*.

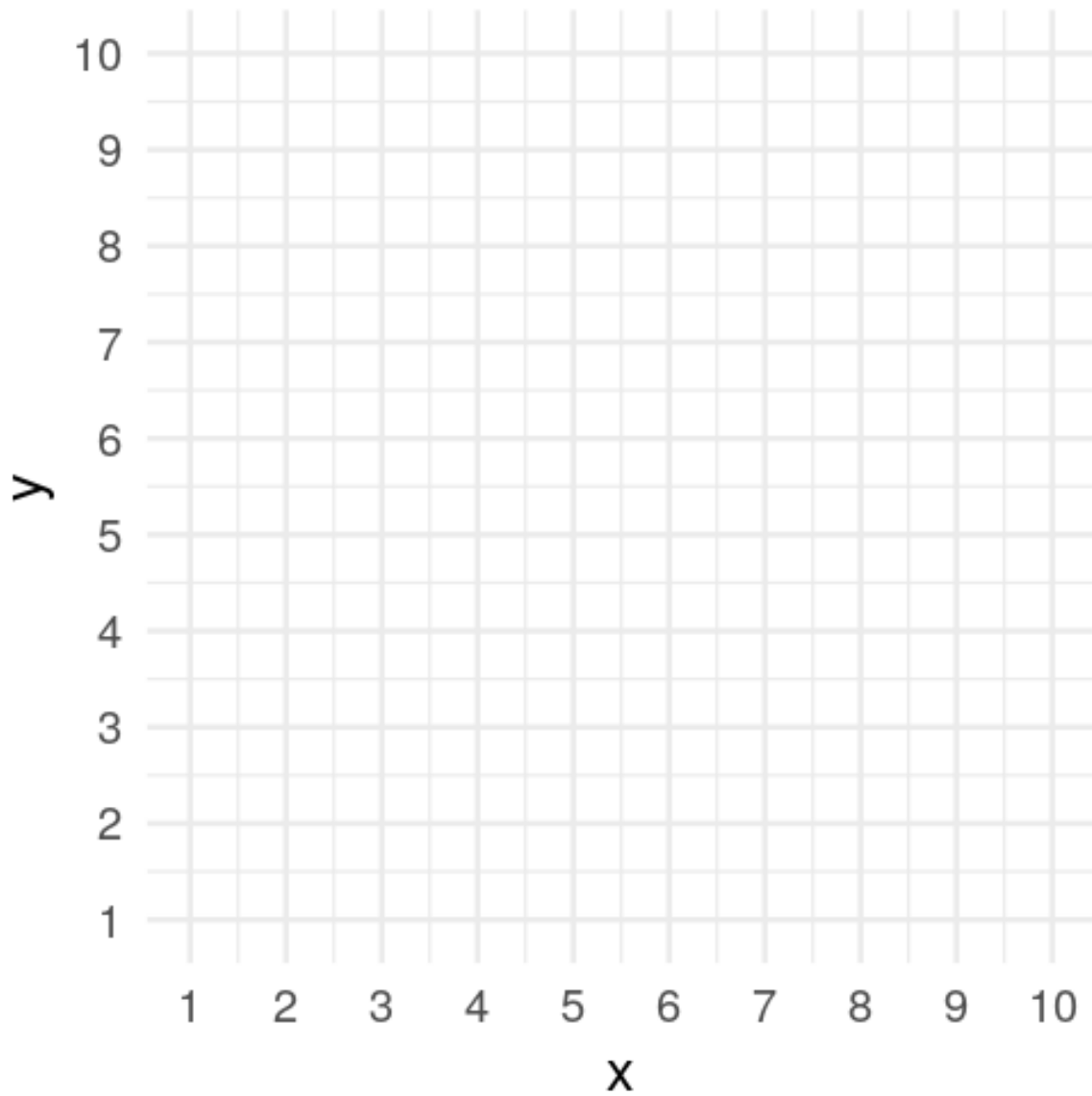


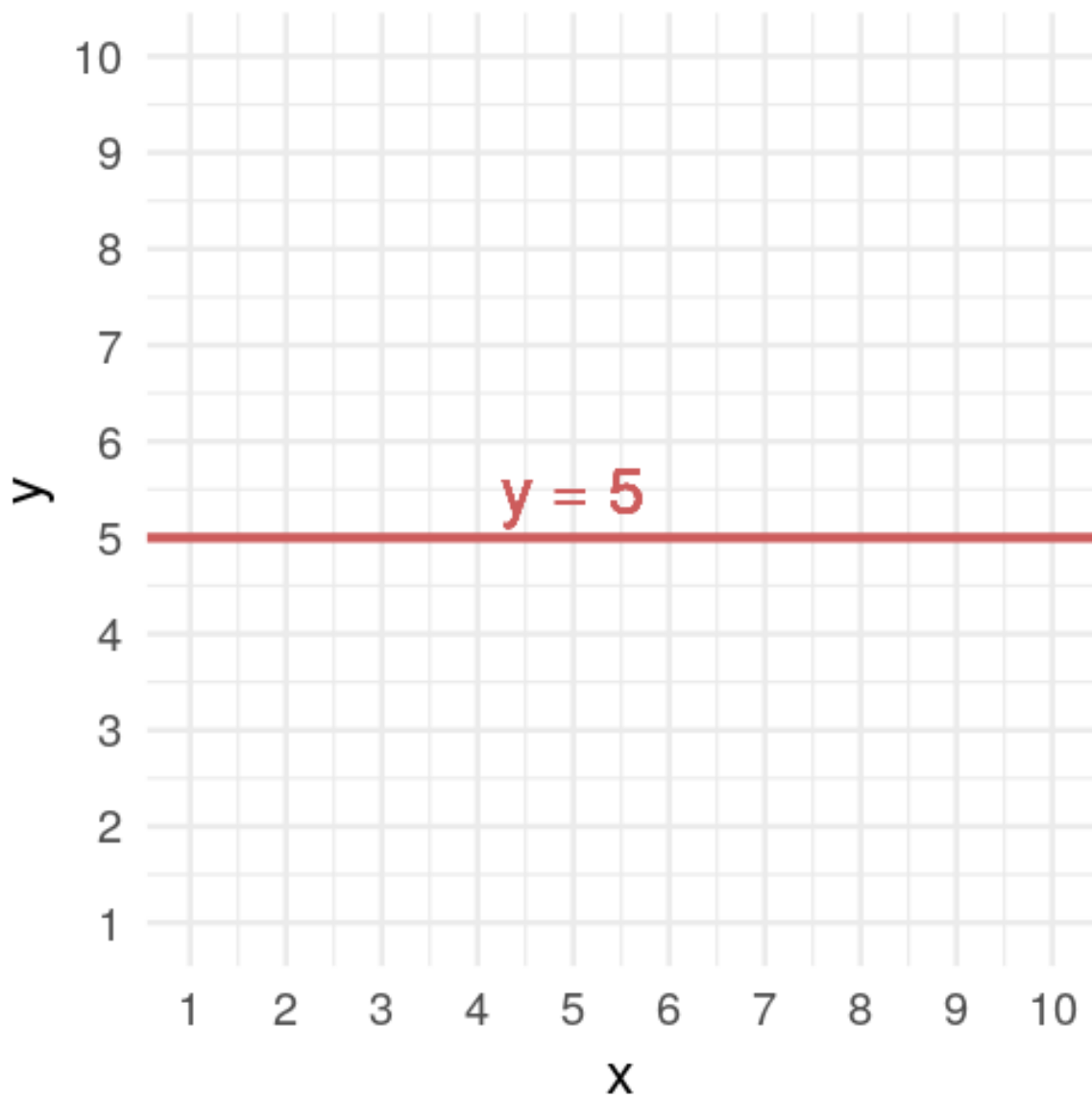
# Regressió lineal (exemple)

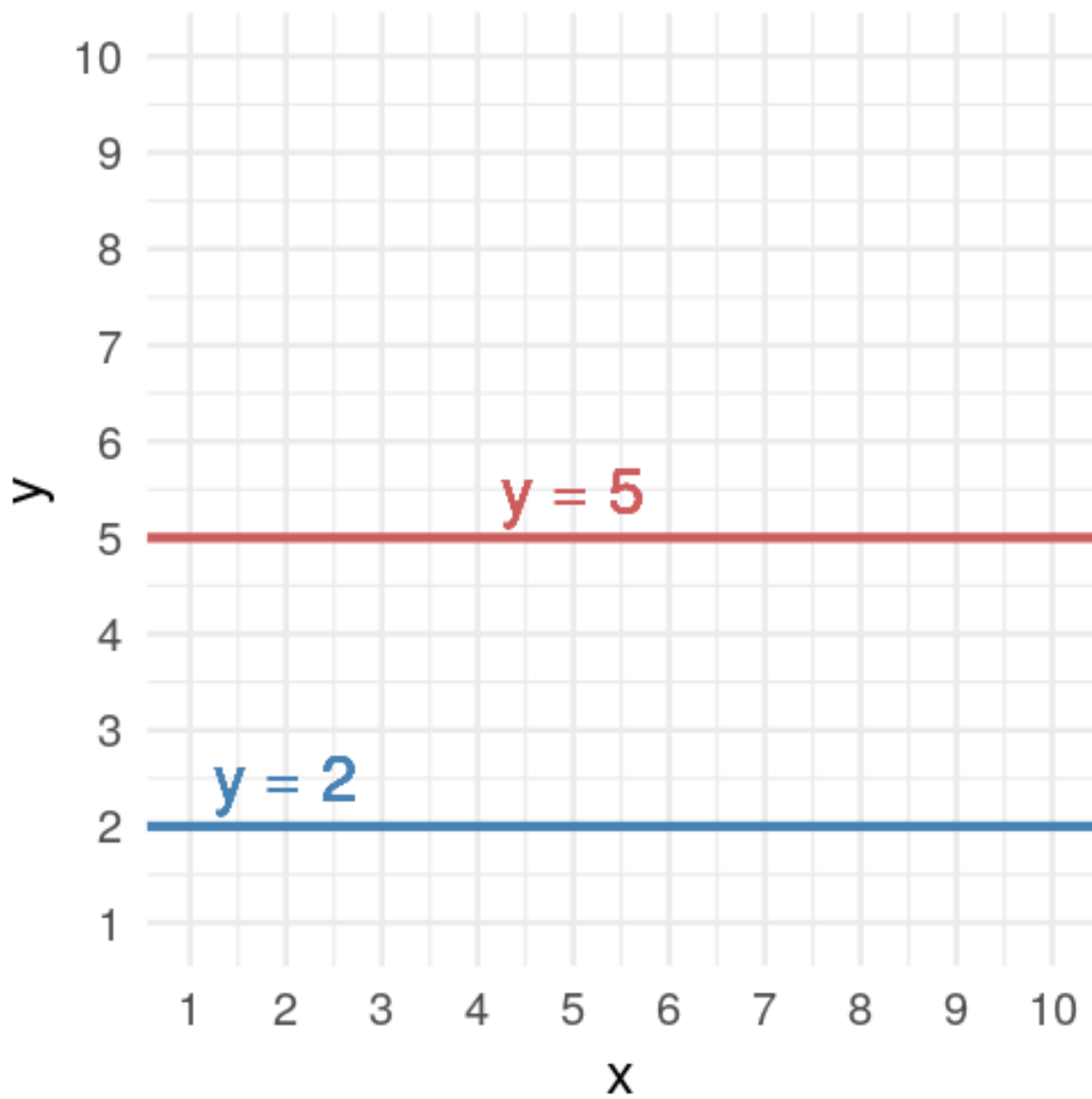
Estimació de relació **lineal** entre resultat ( *to* ) i un o més predictors ( *sexe*; *context*; *edat* ). Una altra formulació: Estimació de predicció de *to* a base de *sexe* i/o *context* i/o *edat*.

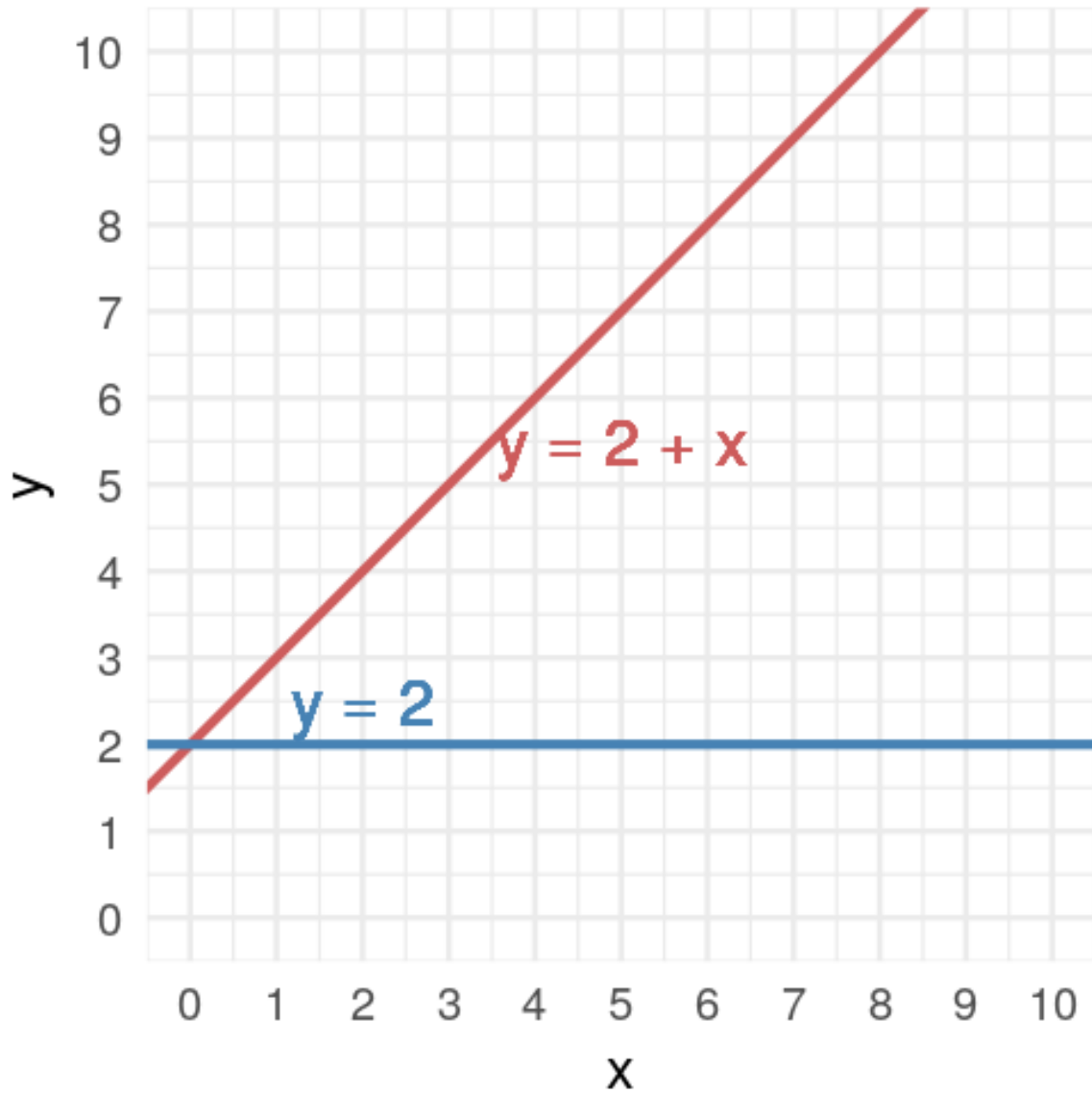


# Repasant línies

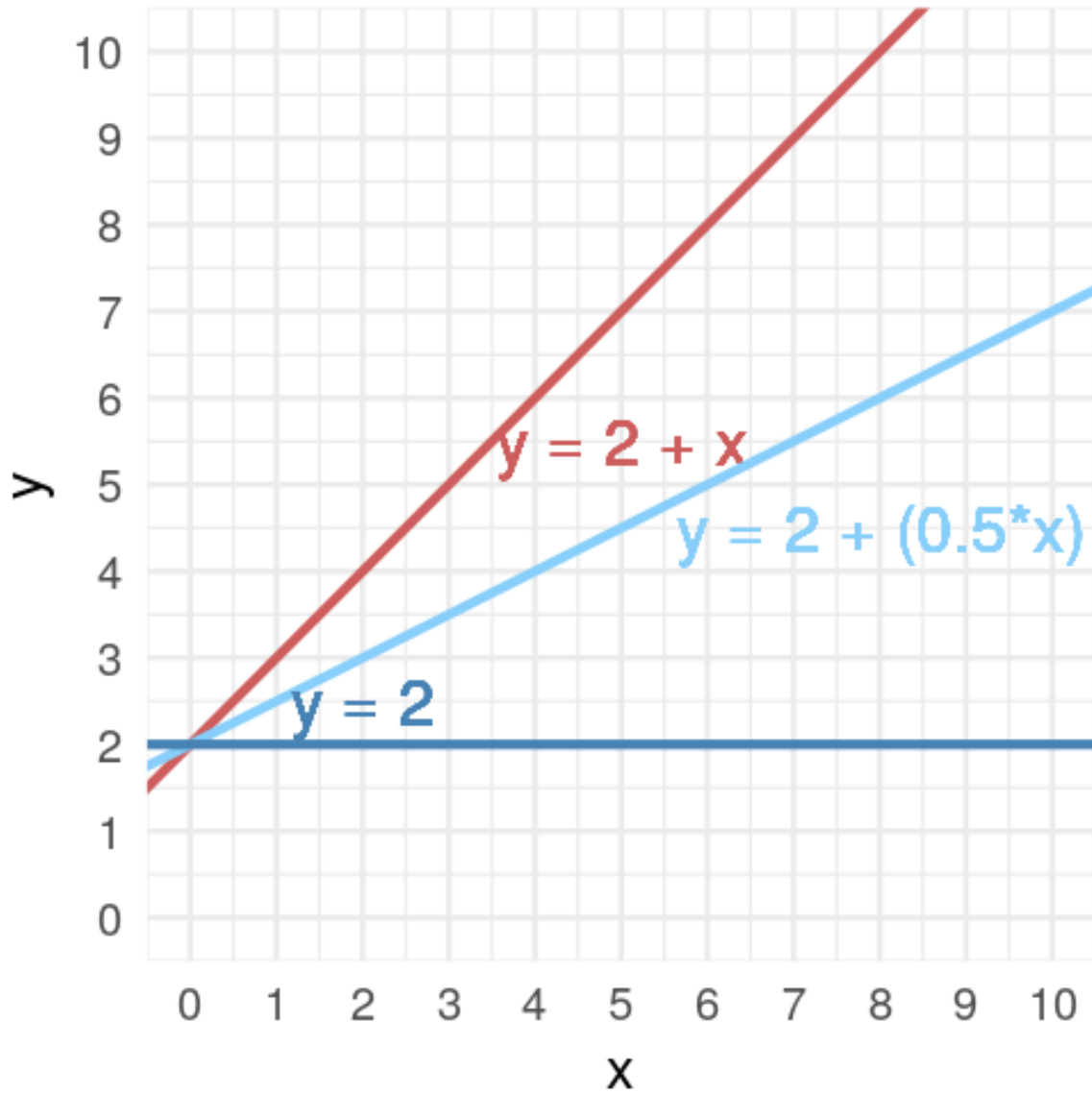




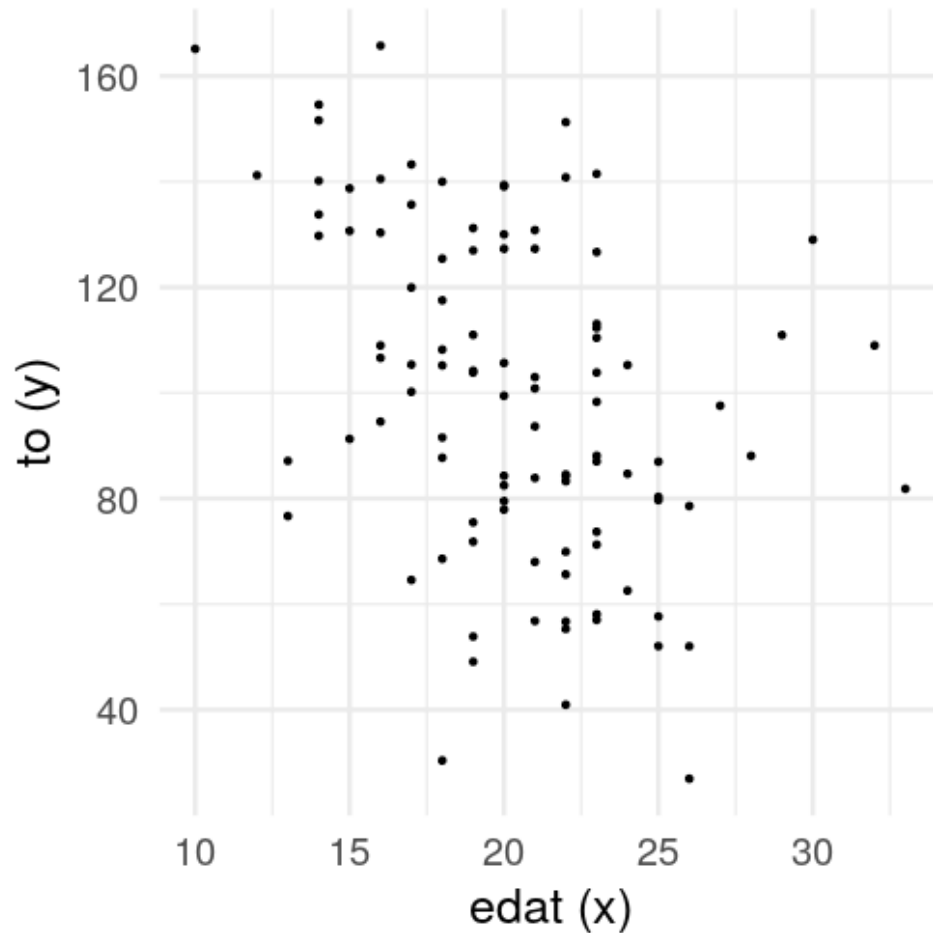




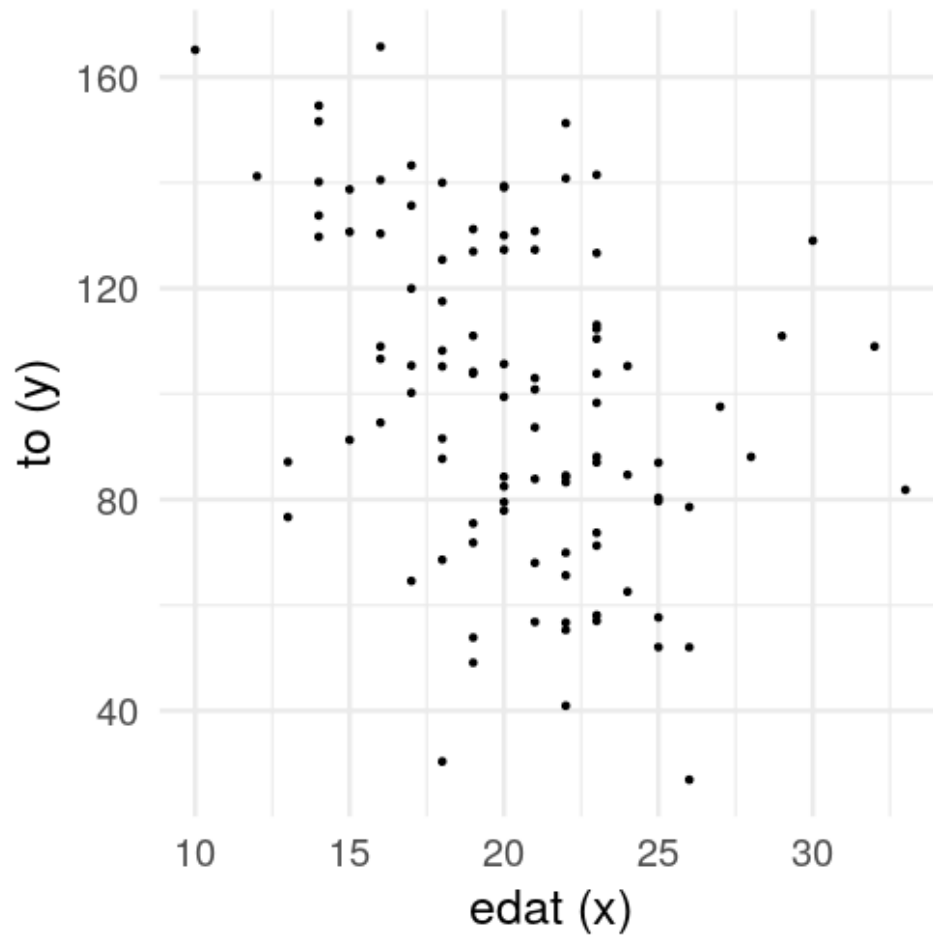




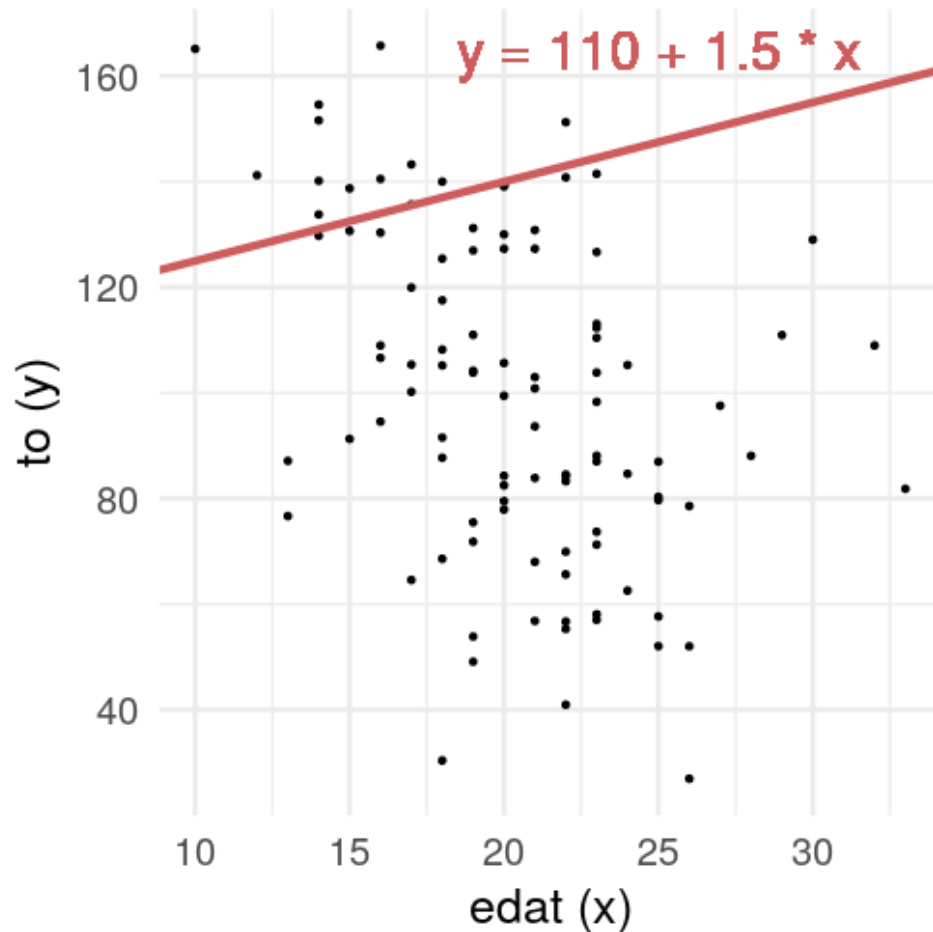
# to en funció a edat



$$to = \beta_0 + (\beta_1 \times edat)$$



$$to = 110 + (1.5 \times edat) ?$$

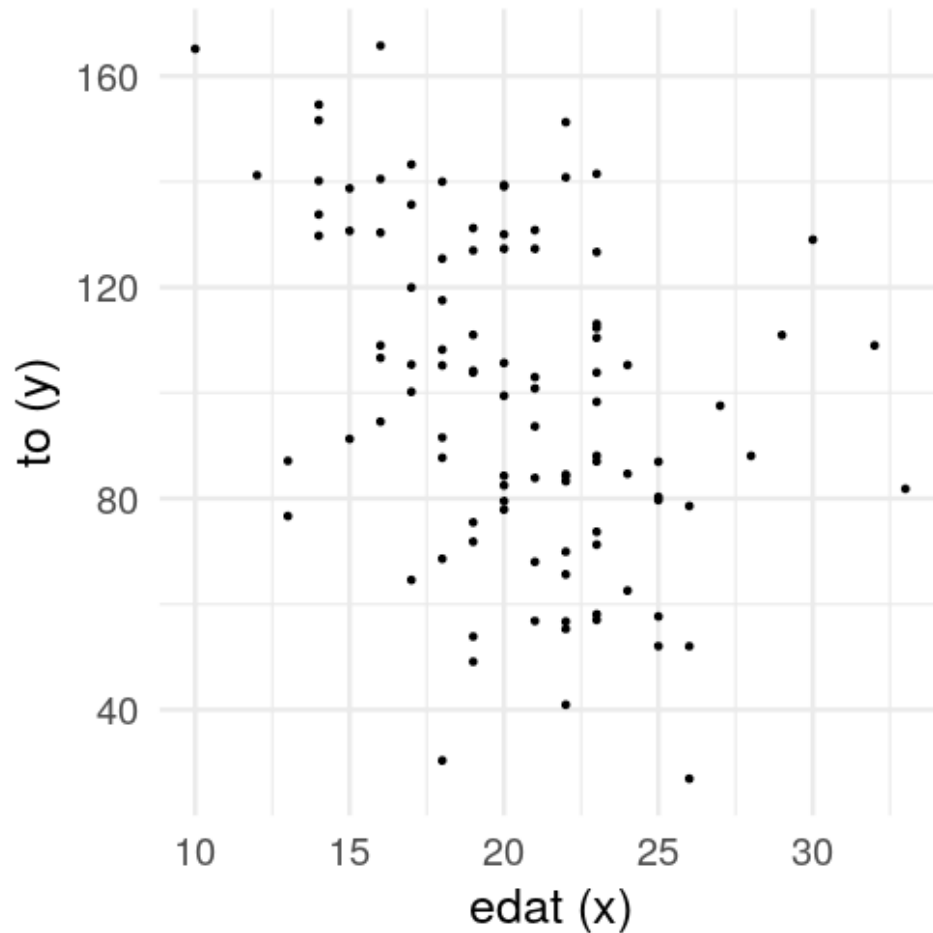


# Intuïció de l'algorisme de regressió

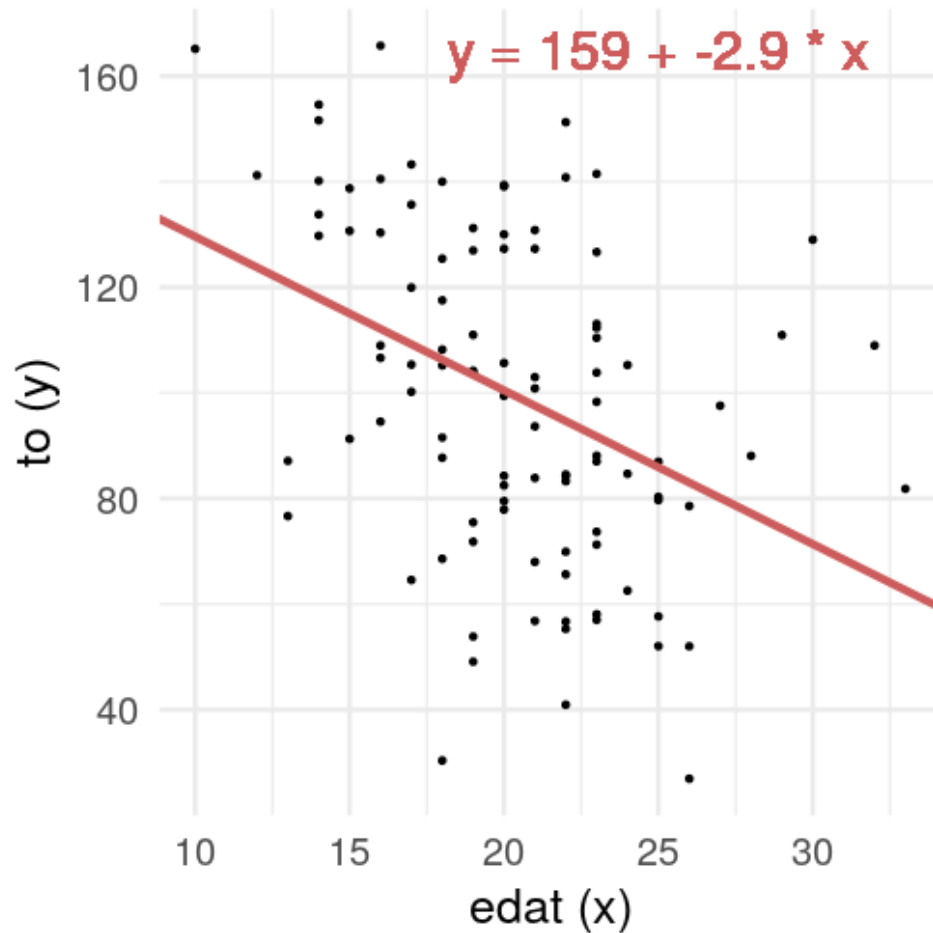
Cerca la línia que minimitza l'error (quadrat)

En altres paraules: Cerca la línia que és més a prop dels punts (i.e., de les dades)

# On posaríeu la línia?



$$to = 159 + (-2.9 \times edat)$$



# Regressió lineal



# Regressió lineal

Estimat de relació **lineal** entre resultat (  $y$  ) i un o més predictors.

Amb un sol predictor  $x$ :

$$y_i \sim \text{Normal}(\mu_i, \sigma),$$

$$\mu_i = \beta_0 + \beta_1 x_i$$

# Regressió lineal

Estimat de relació **lineal** entre *to* ( *y* ) i un o més predictors.

Amb un sol predictor *edat*:

$$y_i \sim \text{Normal}(\mu_i, \sigma),$$

$$\mu_i = \beta_0 + \beta_1 \text{edat}_i$$

# Regresió 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
```

# Regresió 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
```

---

$$to_i = 159 + (-2.9 \times edat_i)$$

Quina és la predicció del to esperat d'una persona de 12 anys? I d'una de 25?

# Més enllà de prediccions

```
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
```

# Residuals

Diferència entre predicció i resultat real.

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

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

# Residuals

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

# Més enllà de prediccions

```
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 estàndard (standard error)

Incertesa del model respecte d'un paràmetre (en funció de les dades)

```
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

# Més enllà de prediccions

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 estàndard error

El valor (estimat) de l'error del model. Correspon a  $\sigma$

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

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

# Goodness of fit

$R^2$  és la proporció de la variància del resultat que expliquen els predictors.

```
summary$r.squared
```

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

# Més enllà de prediccions

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
```

# Propera sessió

- Lliurament de "Assignment 3" (08:00 AM 09/05)
  - Lliurament de part I de "Revisió per parells" (08:00 AM 11/05)
- 

- **Regressió amb més d'un predictor**

# Transformacions: Centrar

Centrar dades implica restar una constant a tots els valors d'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) #coeficiente centrado
```

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