

| Lunes 28 | Martes 29 | Miércoles 30 | Jueves 31 | Viernes 1 |
|----------|---|--|--|-----------|
| | <ul style="list-style-type: none"> • Conceptos básicos • T-test | <ul style="list-style-type: none"> • One-way ANOVA • Two-way ANOVA | <ul style="list-style-type: none"> • LM Simples | |
| Lunes 4 | Martes 5 | Miércoles 6 | Jueves 7 | Viernes 8 |
| | | <ul style="list-style-type: none"> • LM múltiples con interacción • LM múltiples sin interacción | <ul style="list-style-type: none"> • Resolución de práctica • GLMs | |

- *ANOVA con covarianza en Statgraphics:*

ANCOVA (Análisis de Covarianzas) == modelo lineal con interacción

<http://www.differencebetween.net/science/mathematics-statistics/difference-between-ancova-and-regression/>

- *Tipos de ANOVA:*

“Importante” cuando los datos están desequilibrados.

Type I: sequential sum of squares: mira primero los efectos de los factores (var. explicativas) y su interacción independientemente.

Type II: Asume que no existe interacción entre factores

<- para cuando NO haya interacción significativa

Type III: mira primero la interacción y después el efecto de cada factor independientemente

<- para cuando la interacción sea significativa

<http://www.differencebetween.net/science/mathematics-statistics/difference-between-ancova-and-regression/>

- *Jueves 7 → ¿Datos propios?*

T-test

- Comparar **dos** grupos
H0= las medias de los dos grupos son iguales
Ha= las medias de los dos grupos son distintas

>t.test(Y ~ X)

ANOVA (One-way)

- Comparar **más de dos** grupos
H0= La media de los grupos no difiere
Ha= La media de los grupos difiere al menos entre dos grupos

>aov(Y ~ X) %>%summary()

ANOVA (Two-way)

- Comparar el efecto de la **combinación** de varios factores
H0= La media de los grupos no difiere
Ha= La media de los grupos difiere al menos entre dos grupos

>aov(Y ~ X1* X2) %>%summary()



Estadística aplicada en R

Modelos Lineales:

Regresión simple
Regresión múltiple sin interacción
Regresión múltiple con interacción

-Marzo 2022-

Carlota Solano
carlota.solano.udina@upm.es



4.1. Regresión lineal simple

1. ¿Qué es?

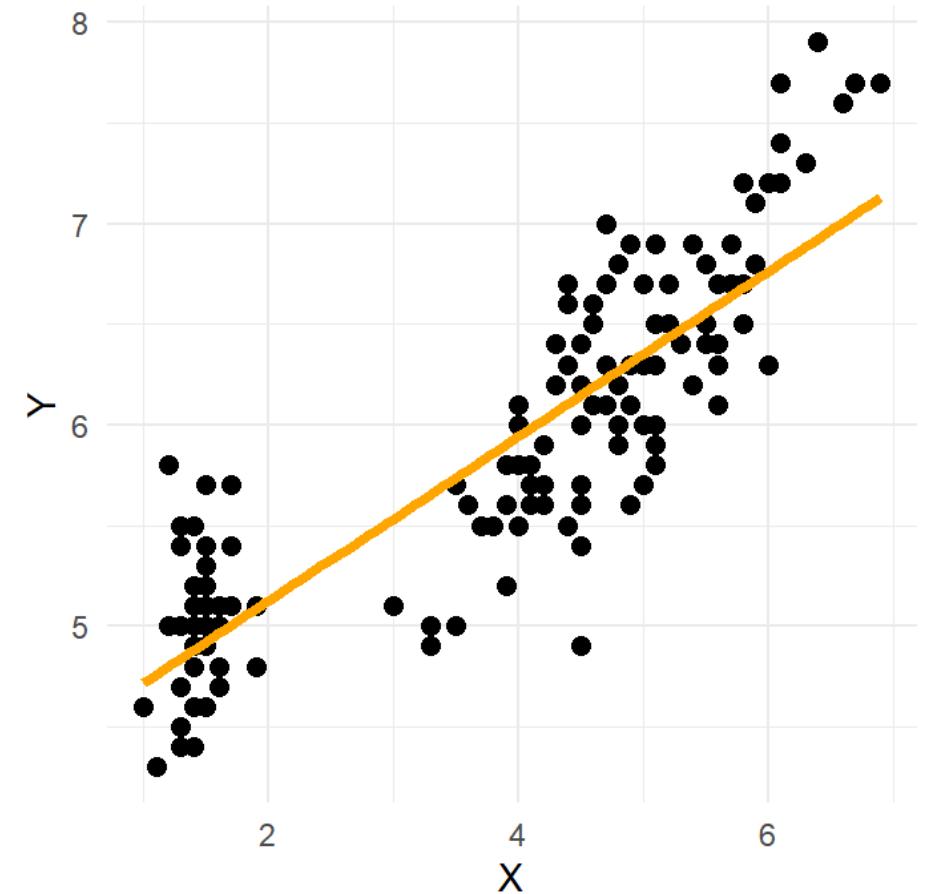
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El objetivo es encontrar la línea que mejor defina los datos

2. ¿Cuándo se puede utilizar?

Cuando quieres definir cómo se relacionan dos elementos.

Correlación no implica causalidad



4.1. Regresión lineal simple

1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

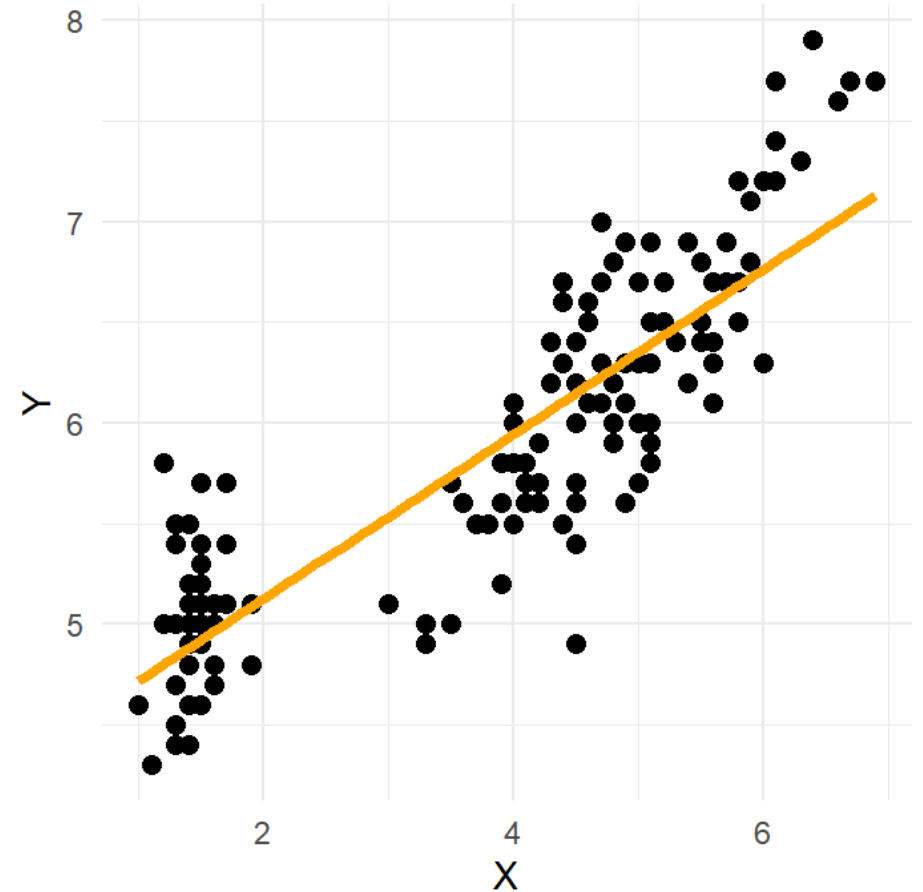
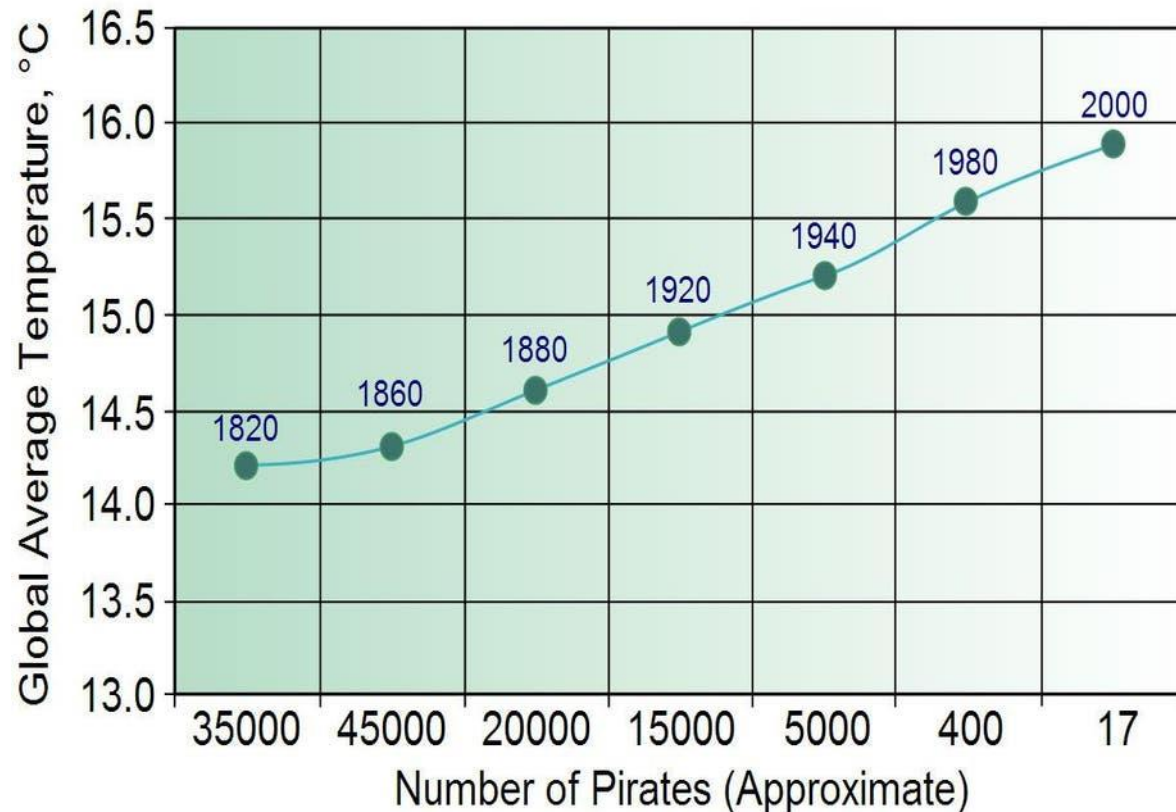
→ El objetivo es encontrar la línea que mejor defina los datos

2. ¿Cuándo se puede utilizar?

Cuando quieres definir cómo se relacionan dos elementos.

Correlación no implica causalidad

Global Average Temperature vs. Number of Pirates



4.1. Regresión lineal simple

1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

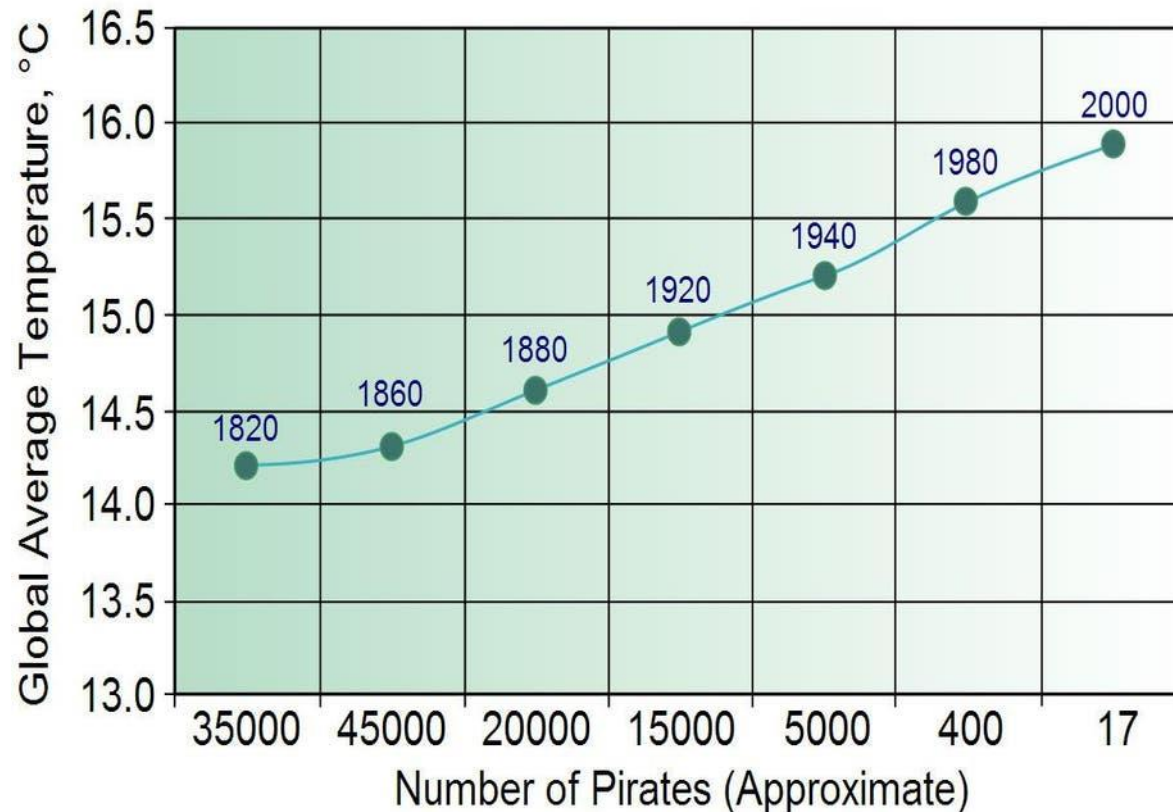
→ El objetivo es encontrar la línea que mejor defina los datos

2. ¿Cuándo se puede utilizar?

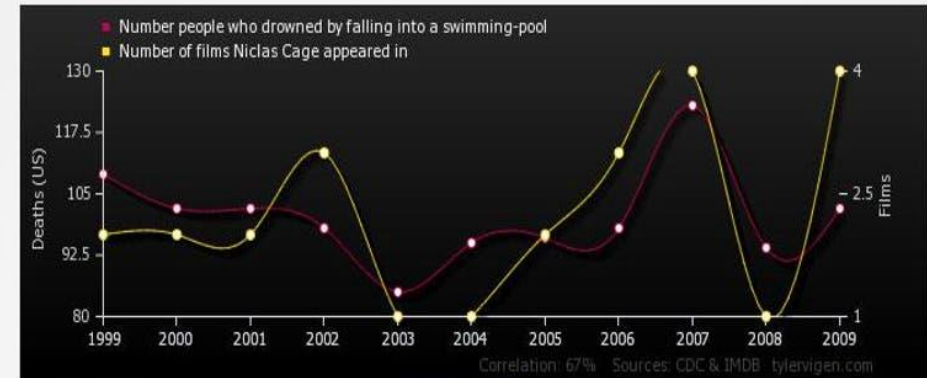
Cuando quieres definir cómo se relacionan dos elementos.

Correlación no implica causalidad

Global Average Temperature vs. Number of Pirates



Number people who drowned by falling into a swimming-pool
correlates with
Number of films Nicolas Cage appeared in



Upload this image to imgur

2

X

6

4.1. Regresión lineal simple

1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El objetivo es encontrar la línea que mejor defina los datos

2. ¿Cuándo se puede utilizar?

Cuando quieres definir cómo se relacionan dos elementos.

Correlación no implica causalidad

3. ¿Qué tipo de datos se necesitan?

Variable respuesta (dep.; y) → Numérica y continua

Variable explicativa (indep.; x) → Numérica y continua

$$y = a + m x$$

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

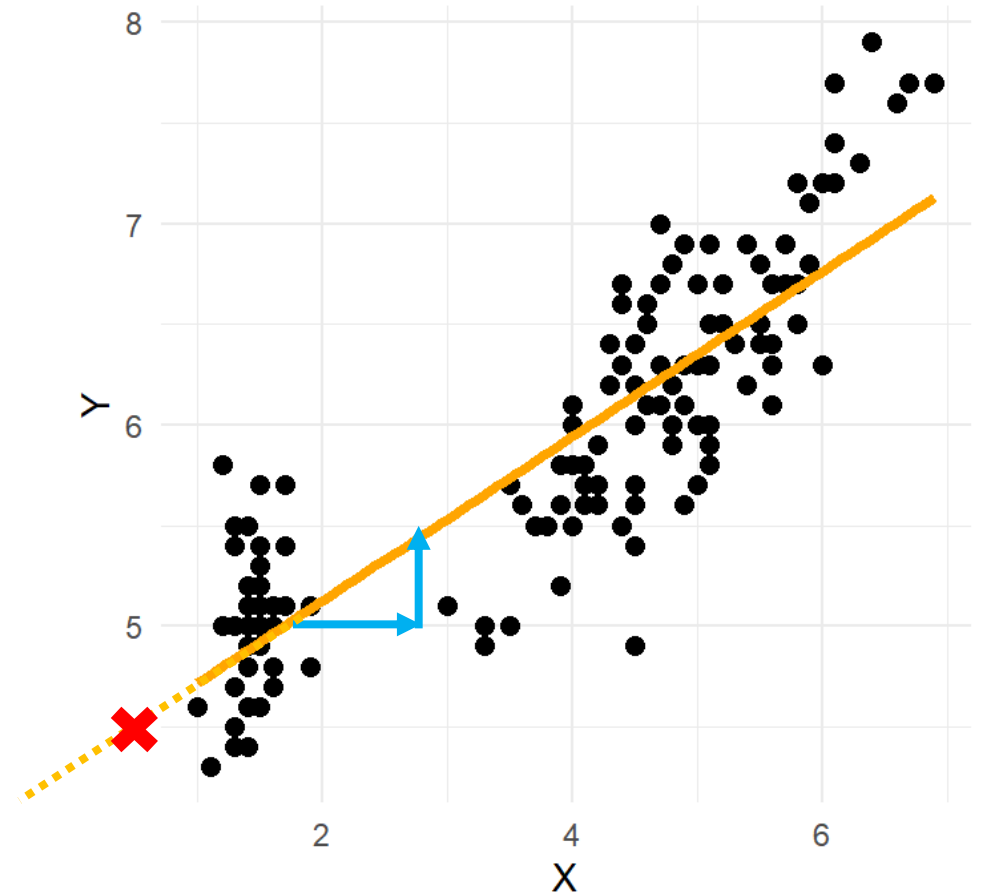
Intercepto



Pendiente



Error o residuo



4.1. Regresión lineal simple

1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El objetivo es encontrar la línea que mejor defina los datos

2. ¿Cuándo se puede utilizar?

Cuando quieres definir cómo se relacionan dos elementos.

Correlación no implica causalidad

3. ¿Qué tipo de datos se necesitan?

Variable respuesta (dep.; y) → Numérica y continua

Variable explicativa (indep.; x) → Numérica y continua

$$y = a + m x$$

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Intercepto

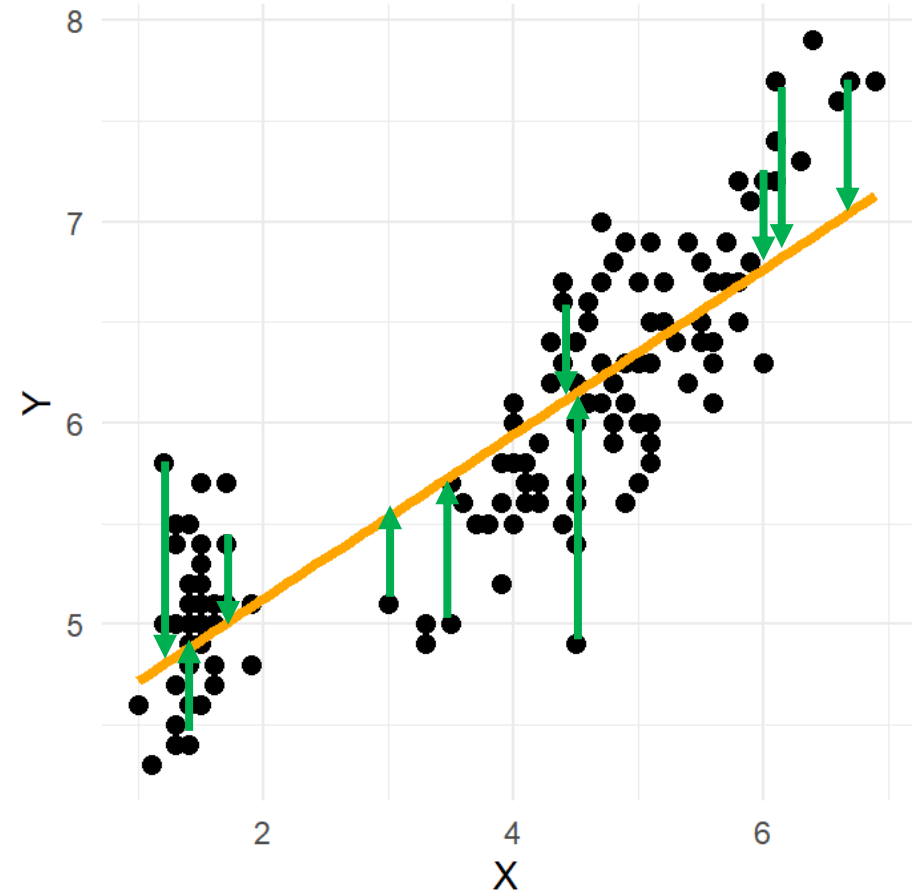


Pendiente



Error o residuo

$\varepsilon = \text{valor real} - \text{predicho por modelo}$



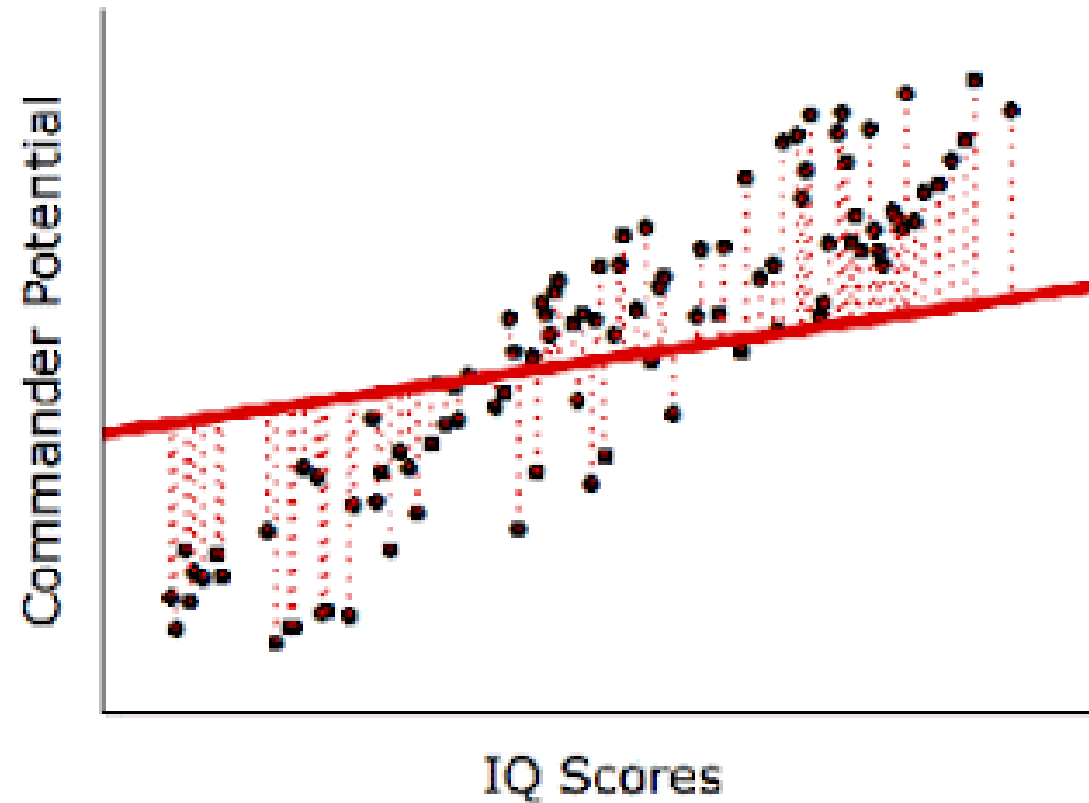
4.1. Regresión lineal simple

1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El **objetivo** es encontrar la línea que mejor defina los datos = *Encontrar los valores de b_0 y b_1 que nos permiten minimizar la suma de los cuadrados de los residuos*

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



4.1. Regresión lineal simple

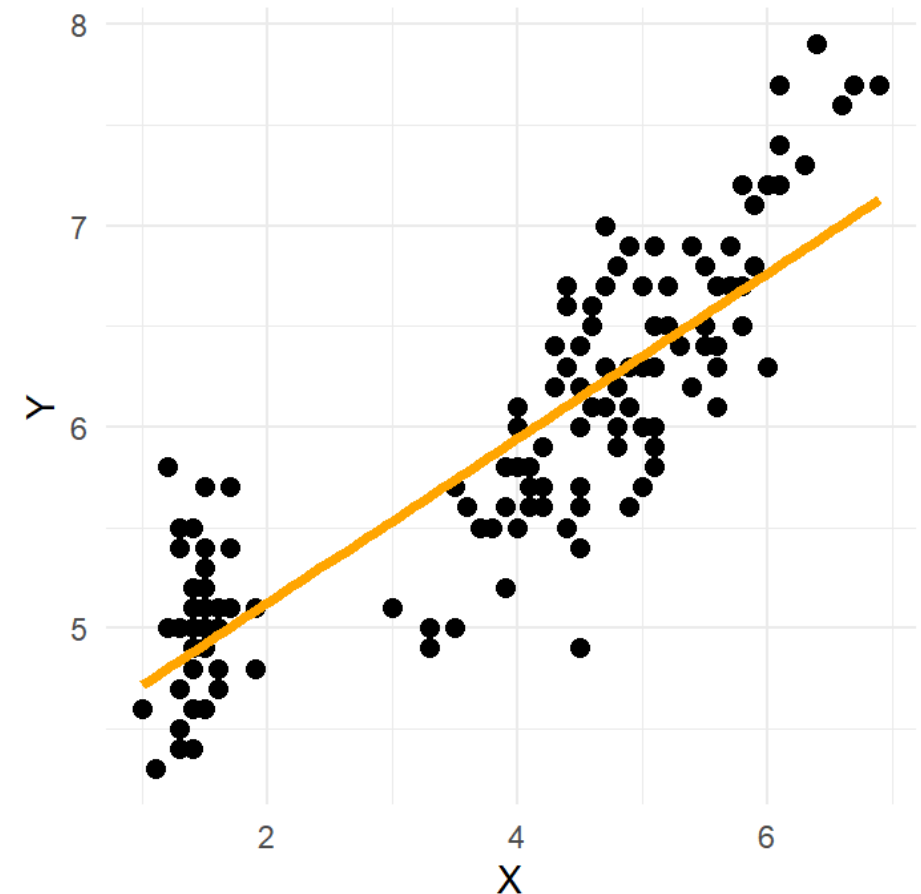
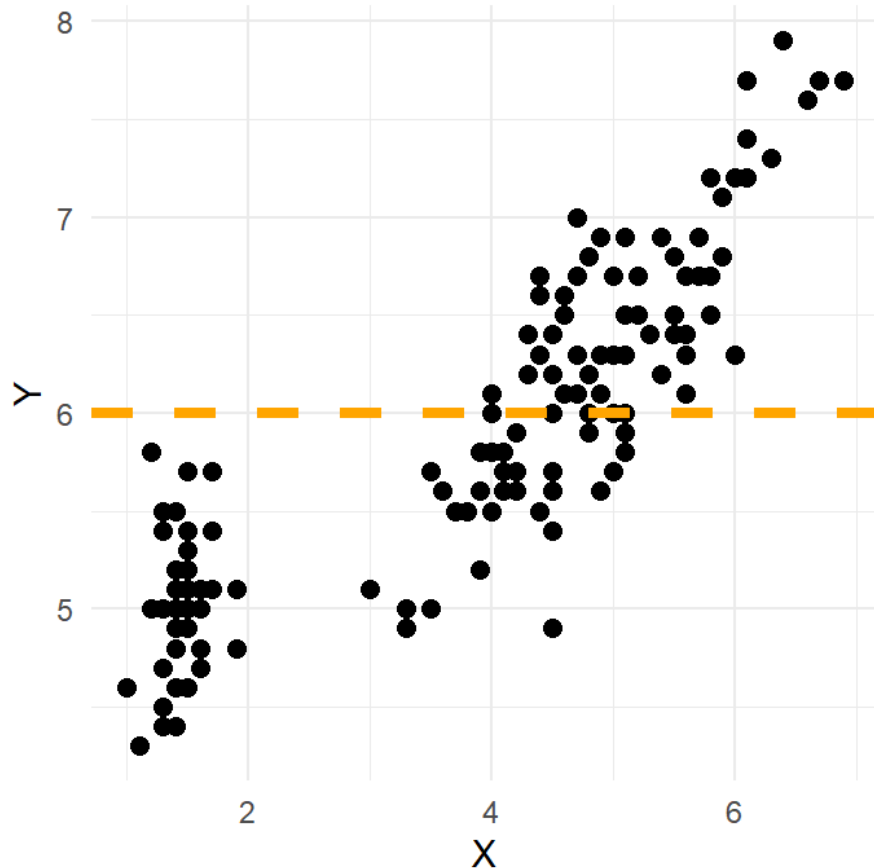
1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El **objetivo** es encontrar la línea que mejor defina los datos = **Encontrar los valores de b_0 y b_1 que nos permiten minimizar la suma de los cuadrados de los residuos**

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

$$SS_{res} = \sum_{i=1}^n (\varepsilon_i)^2$$



4.1. Regresión lineal simple

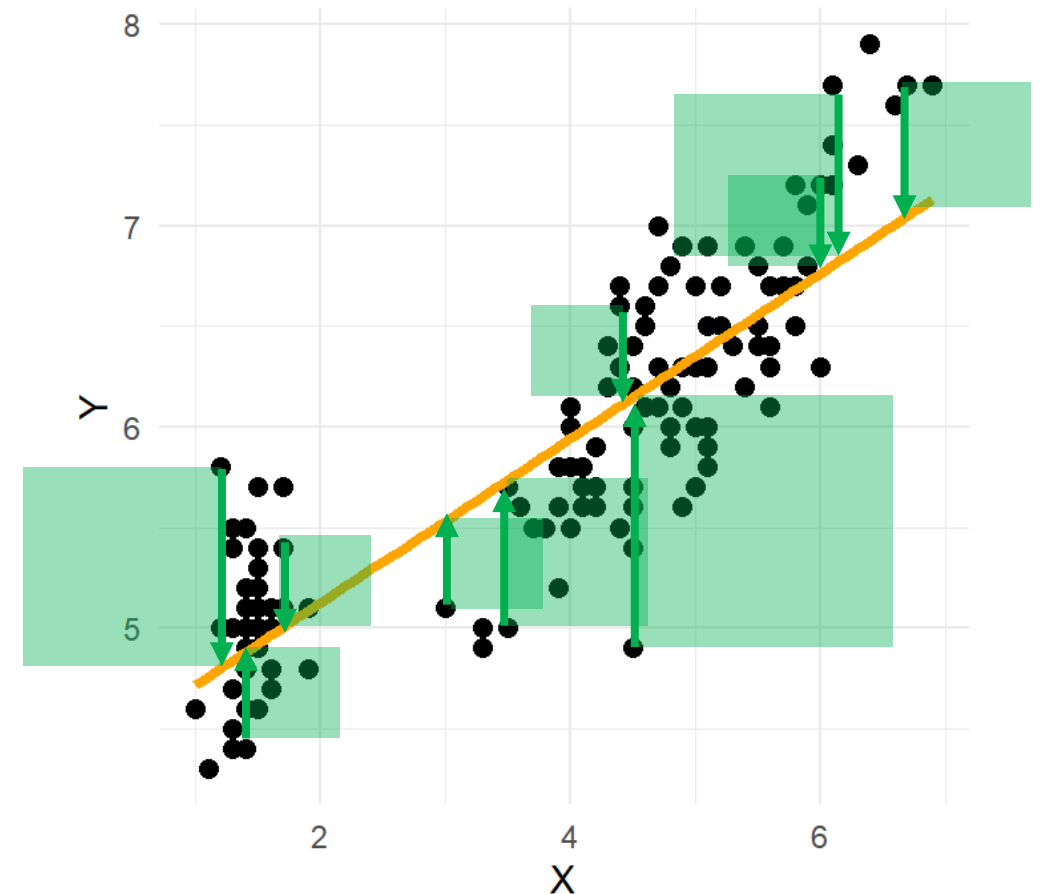
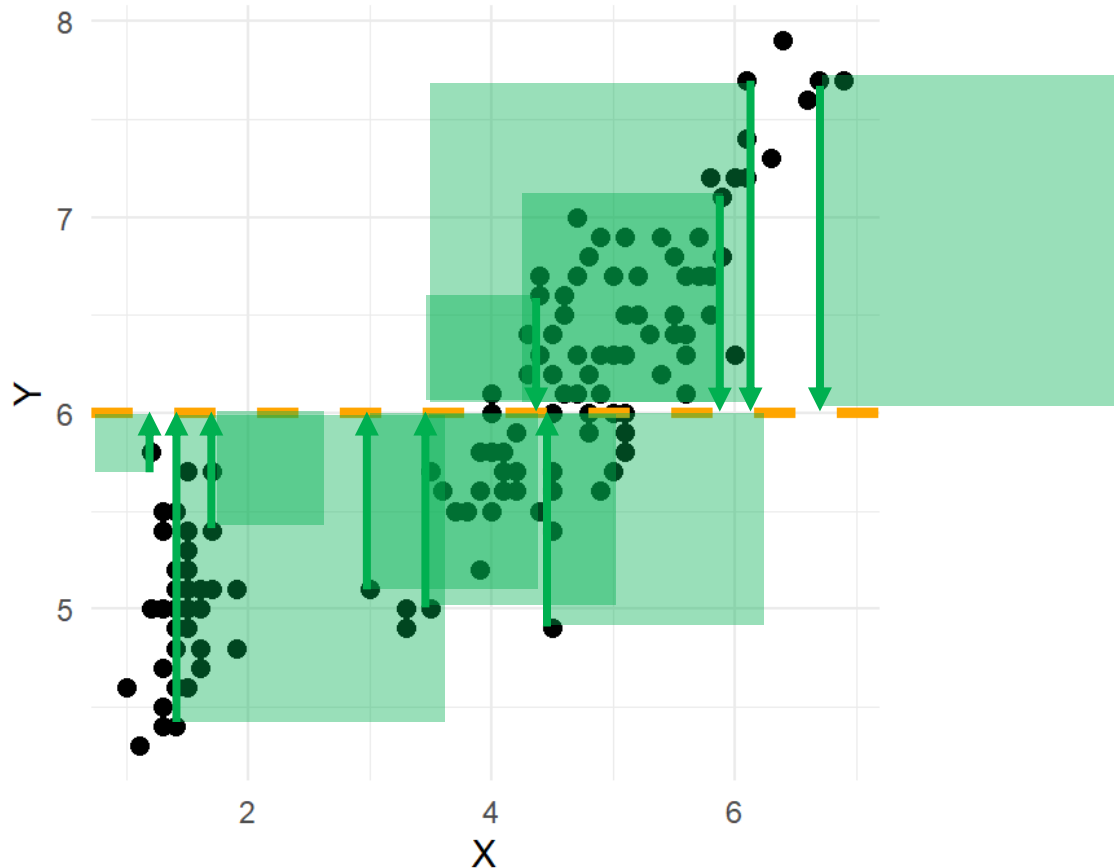
1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El **objetivo** es encontrar la línea que mejor defina los datos = **Encontrar los valores de b_0 y b_1 que nos permiten minimizar la suma de los cuadrados de los residuos**

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

$$SS_{res} = \sum_{i=1}^n (\varepsilon_i)^2$$



4.1. Regresión lineal simple

1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El **objetivo** es encontrar la línea que mejor defina los datos = **Encontrar los valores de b_0 y b_1 que nos permiten minimizar la suma de los cuadrados de los residuos**

$SS_{res} = \sum_{i=1}^n (\epsilon_i)^2 \rightarrow$ Mide la varianza no explicada por el modelo

$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2 \rightarrow$ Mide la varianza del modelo

$R^2 = 1 - \frac{SS_{res}}{SS_{total}} \rightarrow$ proporción de varianza de la var. respuesta que está explicada por el modelo

4.1. Regresión lineal simple

1. ¿Qué es?

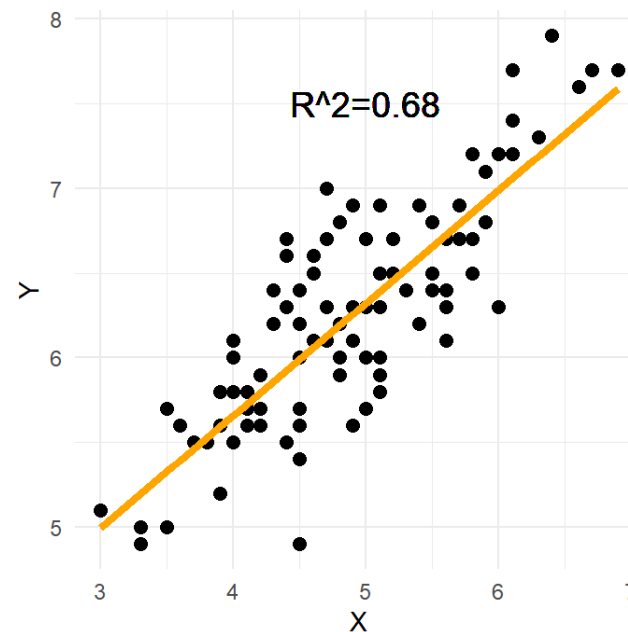
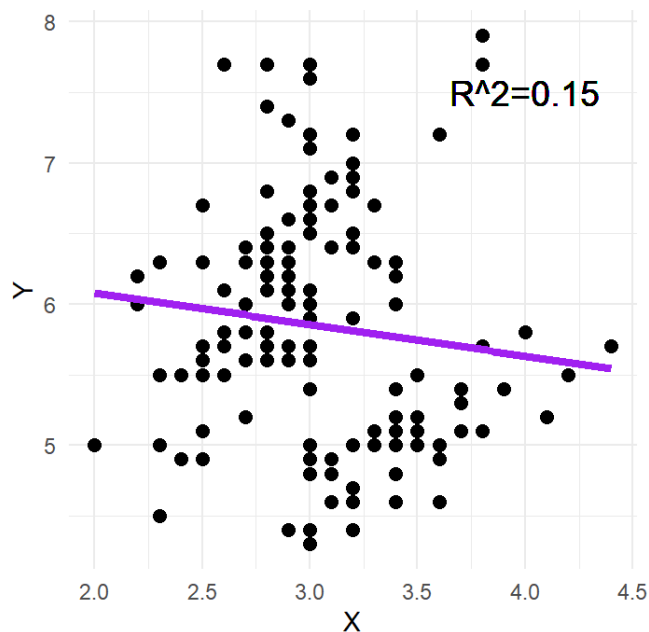
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El **objetivo** es encontrar la línea que mejor defina los datos = **Encontrar los valores de b_0 y b_1 que nos permiten minimizar la suma de los cuadrados de los residuos**

$SS_{res} = \sum_{i=1}^n (\epsilon_i)^2 \rightarrow$ Mide la varianza no explicada por el modelo

$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2 \rightarrow$ Mide la varianza del modelo

$R^2 = 1 - \frac{SS_{res}}{SS_{total}} \rightarrow$ proporción de varianza de la var. respuesta que está explicada por el modelo



4.1. Regresión lineal simple

4. *¿Qué asunciones tiene?*

- La relación entre variables es **lineal**.
- Distribución **normal** de los residuos (o de las variables) del modelo.
- Igualdad de **varianza** de los residuos en torno a la línea de la regresión.
- **Independencia** de las observaciones (i.e. de los datos).

4.1. Regresión lineal simple

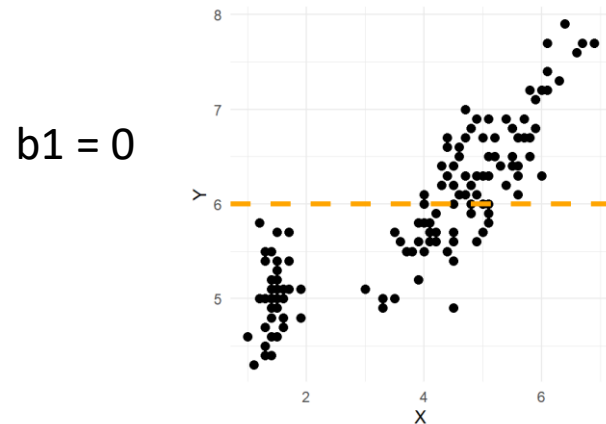
4. ¿Qué asunciones tiene?

- La relación entre variables es **lineal**.
- Distribución **normal** de los residuos (o de las variables) del modelo.
- Igualdad de **varianza** de los residuos en torno a la línea de la regresión.
- **Independencia** de las observaciones (i.e. de los datos).

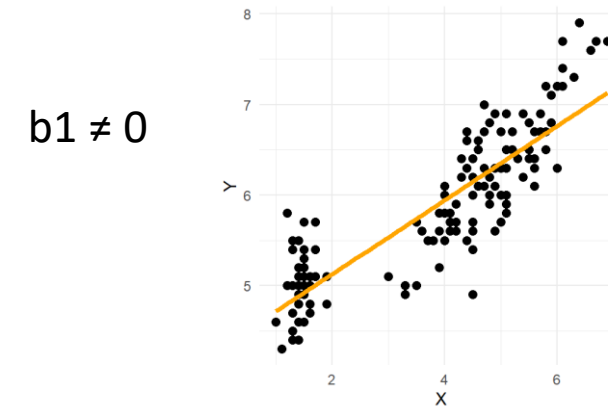
¡Ojo con los outliers → valores atípicos!

5. Matemáticamente, ¿cuál es la hipótesis?

H0: No existe una relación entre las variables estudiadas



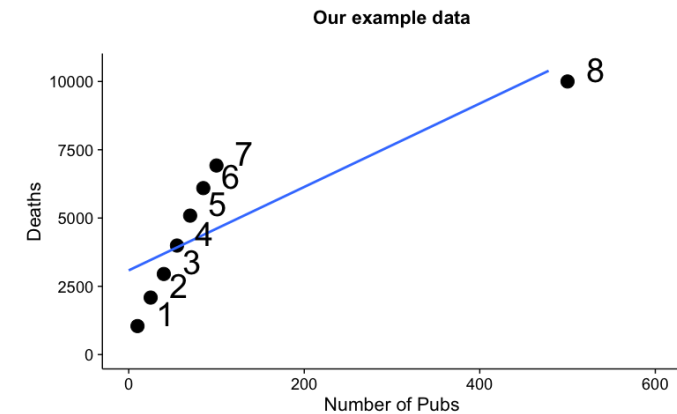
Ha: Existe una relación lineal entre las variables



6. ¿Cómo se corre en R?



```
> holi<-lm (y ~ x)  
> summary(holi)
```



4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm^3) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)
```

Call:

```
lm(formula = trees$volume ~ trees$Height)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|--------|--------|
| -21.274 | -9.894 | -2.894 | 12.068 | 29.852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|---------------|----------|------------|---------|----------|-----|
| (Intercept) | -87.1236 | 29.2731 | -2.976 | 0.005835 | ** |
| trees\$Height | 1.5433 | 0.3839 | 4.021 | 0.000378 | *** |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom

Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358

F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

Estimación de coeficientes que definen la línea de regresión

- (Intercept) = intercepto = b_0 : Valor de y cuando $x=0$

- Var. explicativa = pendiente = b_1 : Por cada incremento en una unidad en la var. explicativa, la var. respuesta varía b_1

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm^3) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)
```

Call:

```
lm(formula = trees$volume ~ trees$Height)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|--------|--------|
| -21.274 | -9.894 | -2.894 | 12.068 | 29.852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|---------------|----------|------------|---------|----------|-----|
| (Intercept) | -87.1236 | 29.2731 | -2.976 | 0.005835 | ** |
| trees\$Height | 1.5433 | 0.3839 | 4.021 | 0.000378 | *** |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358
F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

Estimación de coeficientes que definen la línea de regresión

- (Intercept) = intercepto = b_0 : Valor de y cuando $x=0$

- *Var. explicativa* = pendiente = b_1 : Por cada incremento en una unidad en la var. explicativa, la var. respuesta varía b_1

(Poco valor biológico) Cuando un árbol tiene una altura cero, su volumen es -87.12 dm^3

Por cada dm de altura más, el volumen del árbol incrementa 1.54 dm^3 .

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm^3) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)
```

Call:

```
lm(formula = trees$volume ~ trees$Height)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|--------|--------|
| -21.274 | -9.894 | -2.894 | 12.068 | 29.852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|---------------|----------|------------|---------|--------------|
| (Intercept) | -87.1236 | 29.2731 | -2.976 | 0.005835 ** |
| trees\$Height | 1.5433 | 0.3839 | 4.021 | 0.000378 *** |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom

Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358

F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

Estimación de coeficientes que definen la línea de regresión

- (Intercept) = intercepto = b_0 : Valor de y cuando $x=0$

- *Var. explicativa* = pendiente = b_1 : Por cada incremento en una unidad en la var. explicativa, la var. respuesta varía b_1

Std. Error = Error estándar: **precisión** de la media estimada
(!) $\pm 1.96 \cdot \text{s.e.} = 95\% \text{CI}$

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm^3) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)
```

Call:

```
lm(formula = trees$volume ~ trees$Height)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|--------|--------|
| -21.274 | -9.894 | -2.894 | 12.068 | 29.852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|---------------|----------|------------|---------|--------------|
| (Intercept) | -87.1236 | 29.2731 | -2.976 | 0.005835 ** |
| trees\$Height | 1.5433 | 0.3839 | 4.021 | 0.000378 *** |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom

Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358

F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

Estimación de coeficientes que definen la línea de regresión

- (Intercept) = intercepto = b_0 : Valor de y cuando $x=0$

- *Var. explicativa* = pendiente = b_1 : Por cada incremento en una unidad en la var. explicativa, la var. respuesta varía b_1

Std. Error = Error estándar: **precisión** de la media estimada (!) $\pm 1.96 * \text{s.e.} = 95\% \text{CI}$

T-value: estimación/ s.e. --> \uparrow t value = \downarrow s.e.

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm³) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)
```

Call:

```
lm(formula = trees$volume ~ trees$Height)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|--------|--------|
| -21.274 | -9.894 | -2.894 | 12.068 | 29.852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|---------------|----------|------------|---------|--------------|
| (Intercept) | -87.1236 | 29.2731 | -2.976 | 0.005835 ** |
| trees\$Height | 1.5433 | 0.3839 | 4.021 | 0.000378 *** |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358
F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

Estimación de coeficientes que definen la línea de regresión

- (Intercept) = intercepto = b0: Valor de y cuando x=0

- *Var. explicativa* = pendiente = b1: Por cada incremento en una unidad en la var. explicativa, la var. respuesta varía b1

Std. Error = Error estándar: **precisión** de la media estimada (!) $\pm 1.96 * s.e. = 95\%CI$

T-value: estimación/ s.e. --> \uparrow t value = \downarrow s.e.

Pr(>|t|)= p-valor y significancia → Valores estadísticamente **distintos (o no) de cero.**

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm^3) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)

Call:
lm(formula = trees$volume ~ trees$Height)

Residuals:
    Min       1Q   Median       3Q      Max
-21.274  -9.894  -2.894   12.068   29.852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -87.1236    29.2731  -2.976  0.005835 **
trees$Height   1.5433     0.3839   4.021  0.000378 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-squared:  0.3579,    Adjusted R-squared:  0.3358
F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

Residual standard error: desviación estándar de residuos.
Cuanto menor sea el valor, mejor es la predicción

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm^3) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)

Call:
lm(formula = trees$volume ~ trees$Height)

Residuals:
    Min       1Q   Median       3Q      Max
-21.274  -9.894  -2.894   12.068   29.852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -87.1236    29.2731  -2.976  0.005835 **
trees$Height   1.5433     0.3839   4.021  0.000378 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-squared:  0.3579,    Adjusted R-squared:  0.3358
F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

Residual standard error: desviación estándar de residuos.
Cuanto menor sea el valor, mejor es la predicción

Degrees of freedom: grados de libertad:

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm³) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)
```

Call:

```
lm(formula = trees$volume ~ trees$Height)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|--------|--------|
| -21.274 | -9.894 | -2.894 | 12.068 | 29.852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|---------------|----------|------------|---------|----------|-----|
| (Intercept) | -87.1236 | 29.2731 | -2.976 | 0.005835 | ** |
| trees\$Height | 1.5433 | 0.3839 | 4.021 | 0.000378 | *** |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom

Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358

F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

Residual standard error: desviación estándar de residuos.
Cuanto menor sea el valor, mejor es la predicción

Degrees of freedom: grados de libertad

(...) R-squared: R²: proporción de la varianza explicada por el modelo.

$$(R^2 = 1 - \frac{SS_{res}}{SS_{tot}})$$

4.1. Regresión lineal simple

7. ¿Cómo se interpreta el resultado de R?

E.g. La borrasca Filomena ha dejado muchos árboles caídos, y hemos aprovechado para medir el volumen (dm³) y la altura (dm) de unos cerezos criollos (*Prunus serotina*). ¿Existe una relación entre el volumen de un árbol y su altura? ¿A mayor altura, mayor volumen, o viceversa?

```
> lmtree<-lm(trees$volume~trees$Height)
> summary(lmtree)
```

Call:

```
lm(formula = trees$volume ~ trees$Height)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|--------|--------|
| -21.274 | -9.894 | -2.894 | 12.068 | 29.852 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|---------------|----------|------------|---------|----------|-----|
| (Intercept) | -87.1236 | 29.2731 | -2.976 | 0.005835 | ** |
| trees\$Height | 1.5433 | 0.3839 | 4.021 | 0.000378 | *** |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-squared: 0.3579, Adjusted R-squared: 0.3358
F-statistic: 16.16 on 1 and 29 DF, p-value: 0.0003784

Residual standard error: desviación estándar de residuos.
Cuanto menor sea el valor, mejor es la predicción

Degrees of freedom: grados de libertad:

(...) R-squared: R²: proporción de la varianza explicada por el modelo.

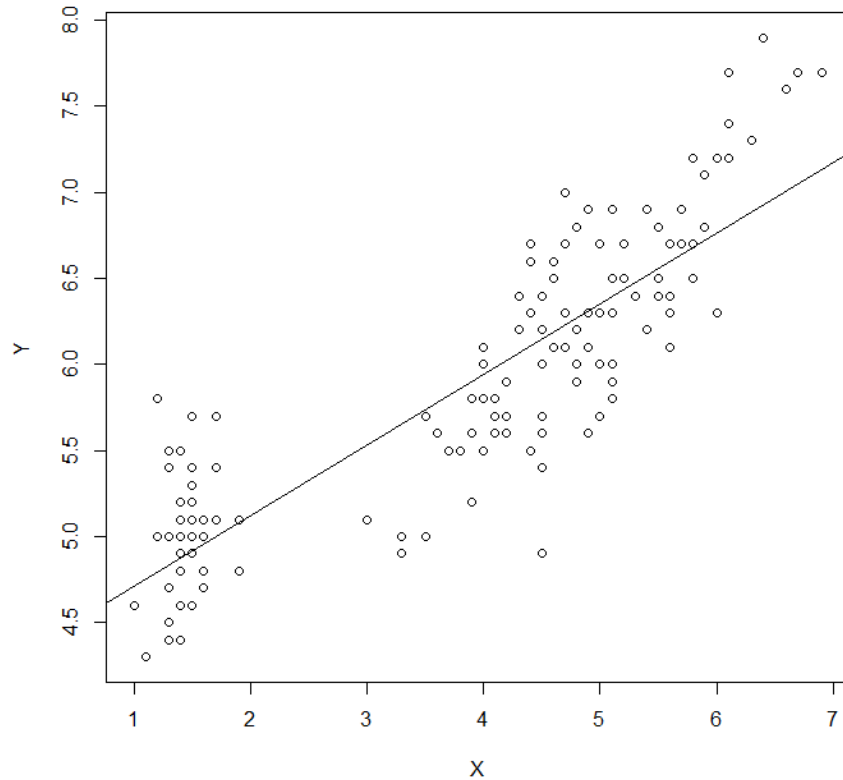
$$(R^2 = 1 - \frac{SS_{res}}{SS_{tot}})$$

F-stats & p-value: Test general para comprobar la H0 → Todos los coeficientes del modelo son igual a cero.

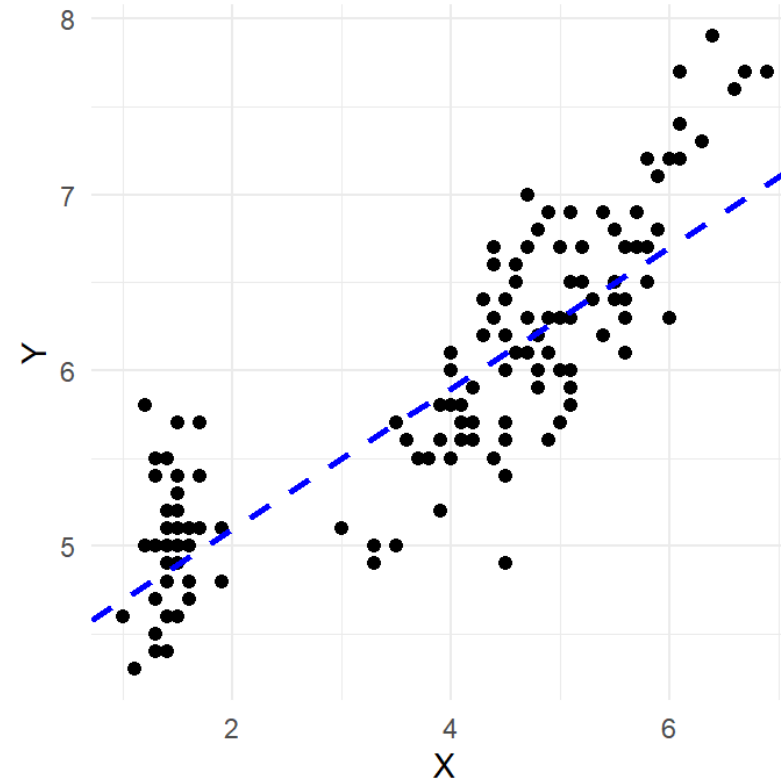
4.1. Regresión lineal simple

8. ¿Cómo se puede representar?

R



```
> plot(data$y ~ data$x)
> abline(a=intercepto, b=pendiente)
```



```
> ggplot(data, aes(x=var.indep, y=var.dep))+
  geom_point(size=4)+
  theme_minimal(base_size=22)+
  labs(x="X",y="Y")+
  geom_abline(intercept = intercepto, slope = pendiente,
              color="blue", linetype="dashed", size=2)
```

4.1. Ejemplo all together

R



¿Existe relación entre el tamaño corporal de los renacuajos y el tamaño de su intestino?

4.2. Ejercicios de Modelos Lineales

Ejercicio: 4. Ejer_LMs (primera parte)

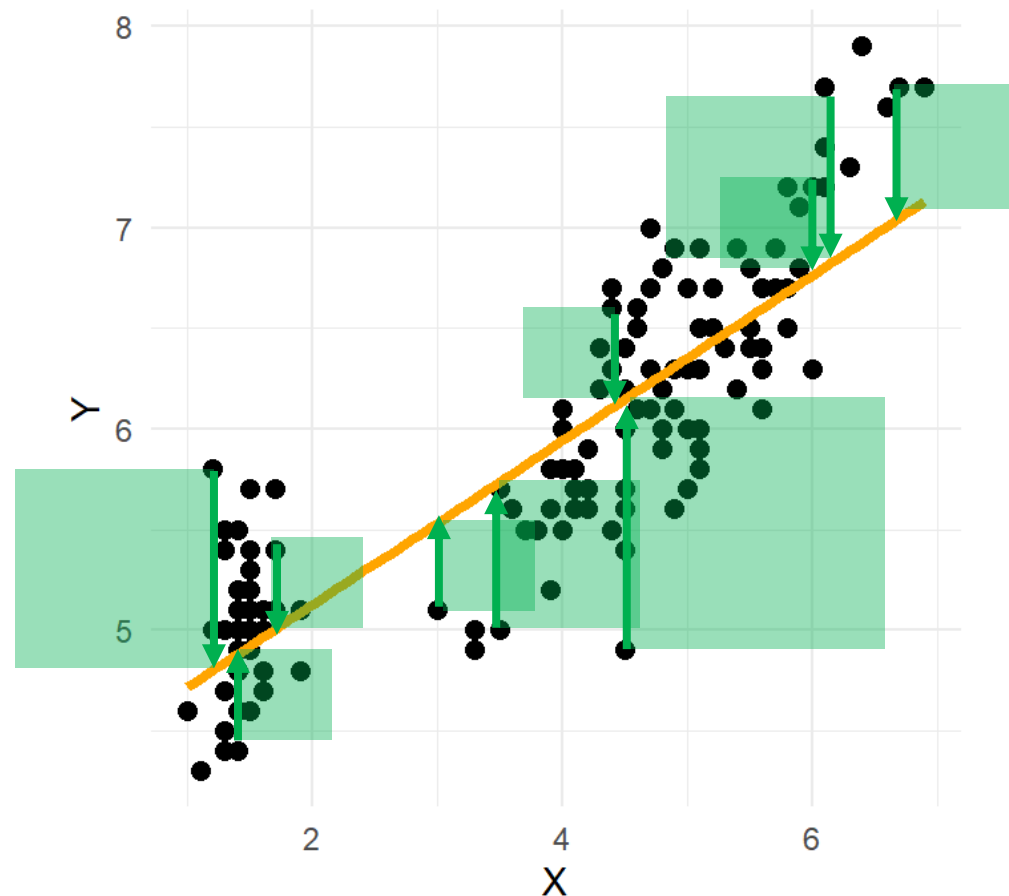
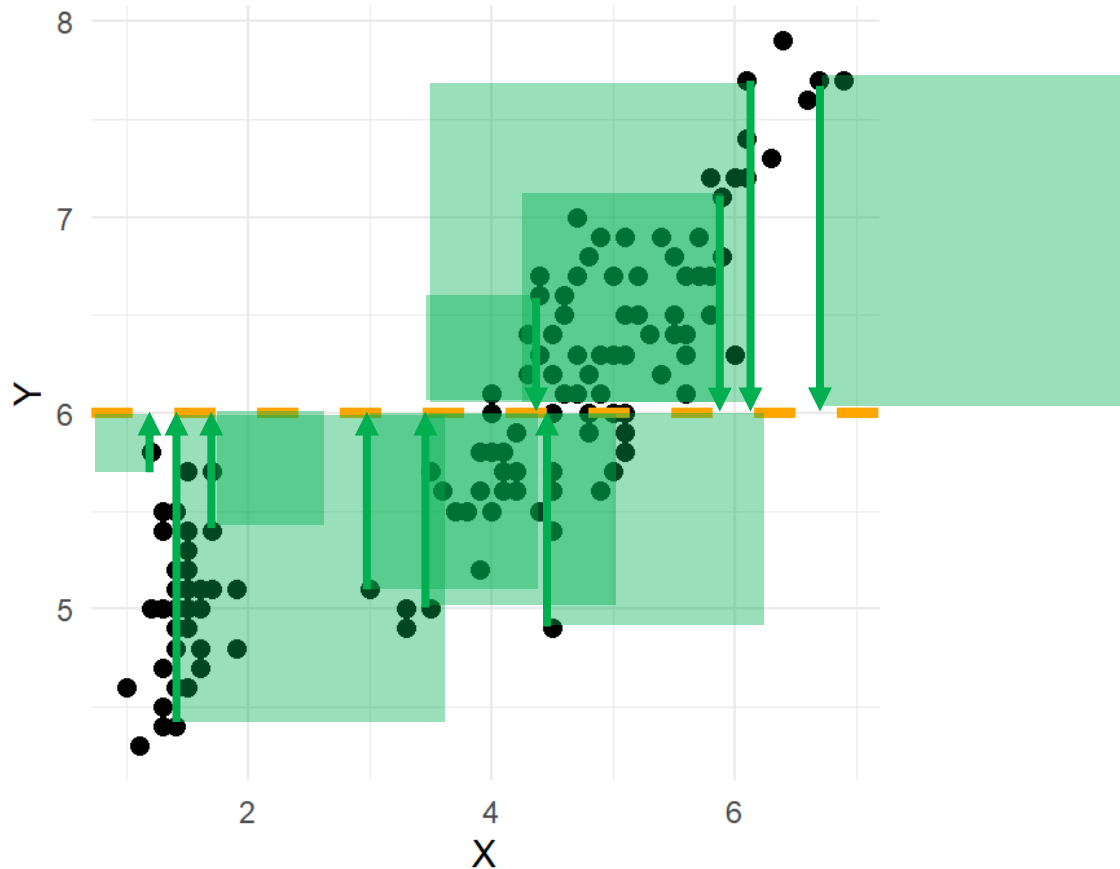
| Lunes 28 | Martes 29 | Miércoles 30 | Jueves 31 | Viernes 1 |
|----------|---|--|--|-----------|
| | <ul style="list-style-type: none"> • Conceptos básicos • T-test | <ul style="list-style-type: none"> • One-way ANOVA • Two-way ANOVA | <ul style="list-style-type: none"> • LM Simples | |
| Lunes 4 | Martes 5 | Miércoles 6 | Jueves 7 | Viernes 8 |
| | | <ul style="list-style-type: none"> • LM múltiples con interacción • LM múltiples sin interacción | <ul style="list-style-type: none"> • Resolución de práctica • GLMs | |

Earlier this week...

Modelos lineales → El **objetivo** es encontrar la línea que mejor defina los datos = **Encontrar los valores de b_0 y b_1 que nos permiten minimizar la suma de los cuadrados de los residuos**

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

$$SS_{res} = \sum_{i=1}^n (\varepsilon_i)^2$$



Earlier this week...

Modelos lineales → El **objetivo** es encontrar la línea que mejor defina los datos = ***Encontrar los valores de b_0 y b_1 que nos permiten minimizar la suma de los cuadrados de los residuos***

$SS_{res} = \sum_{i=1}^n (\epsilon_i)^2$ → Mide la varianza no explicada por el modelo

$SS_{total} = \sum_{i=1}^n (y_i - \bar{y})^2$ → Mide la varianza del modelo

$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$ → proporción de varianza de la var. respuesta que está explicada por el modelo

- **Asunciones en Modelos Lineales: CON LOS RESIDUOS**

Para obtener residuos necesitamos tener un modelo

```
>Modelito<-lm(y~x)  
>resid(Modelito)  
>hist(resid(Modelito))
```


Earlier this week...

T-test

- Comparar **dos** grupos
H0= las medias de los dos grupos son iguales
Ha= las medias de los dos grupos son distintas

>t.test(Y ~ X)

ANOVA (One-way)

- Comparar **más de dos** grupos
H0= La media de los grupos no difiere
Ha= La media de los grupos difiere al menos entre dos grupos

>aov(Y ~ X) %>%summary()

ANOVA (Two-way)

- Comparar el efecto de la **combinación** de varios factores
H0= La media de los grupos no difiere
Ha= La media de los grupos difiere al menos entre dos grupos

>aov(Y ~ X1* X2) %>%summary()

Modelo Lineal Simple

- Determinar cómo se relacionan dos variables continuas
H0= No existe relación entre variables $\rightarrow b1 = 0$
Ha= Existe relación entre variables $\rightarrow b1 \neq 0$

>lm(y ~ x) %>%summary()



Estadística aplicada en R

Modelos Lineales:

Regresión simple
Regresión múltiple sin interacción
Regresión múltiple con interacción

-Marzo 2022-

Carlota Solano

carlota.solano.udina@upm.es



4.2. Modelo lineal múltiple con interacción

1. ¿Qué es?

Es un modelo que relaciona una variable dependiente con varias variables independientes.

$$y = a + mx$$

→ El objetivo es encontrar la línea que mejor defina los datos = **Encontrar los valores de *intercepto* y *pendiente* que nos permiten minimizar la suma de los cuadrados de los residuos**

2. ¿Cuándo se puede utilizar?

Cuando quieres evaluar la influencia que tienen múltiples variables independientes sobre una variable respuesta.

Correlación no implica causalidad

¿Cómo difiere del modelo lineal simple?

Incluimos más de una variable explicativa para estudiar cómo todas afectan a nuestra variable respuesta

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \varepsilon_i$$

Intercepto



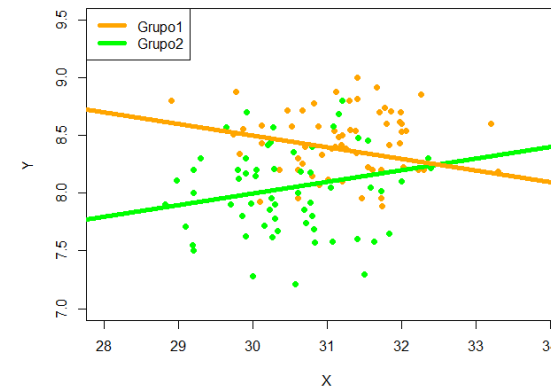
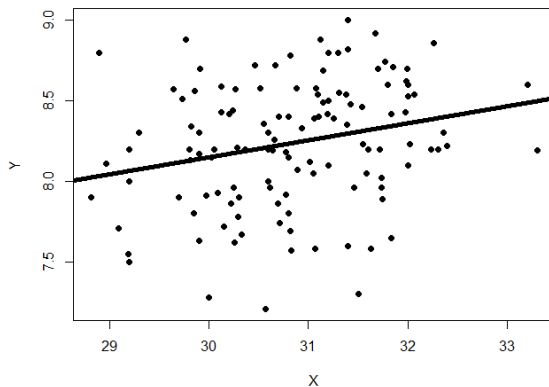
Efecto de x1



Efecto de x2



Residuo



4.2. Modelo lineal múltiple con interacción

3. ¿Qué tipo de datos se necesitan?

Variable respuesta (dep.; y) → Numérica continua

Variables explicativas (indep.; x) → Continuas y categóricas

← **Variables continuas** → **pendiente**: efecto del incremento de una unidad de x sobre y
Variables categóricas → efecto del cambio de grupo de x sobre y

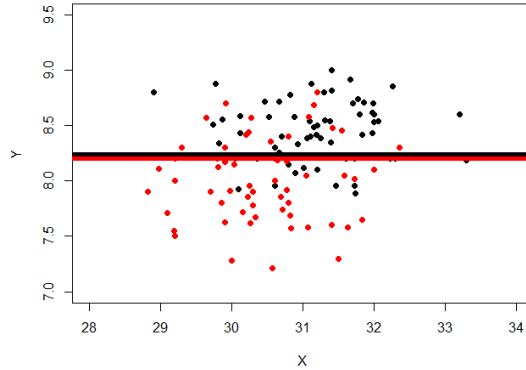
4. ¿Qué asunciones tiene?

- Variables explicativas no deben ser colineales → **No correlación** entre variables explicativas
- **Principio de parsimonia** (Navaja de Ockham)
- **Relación lineal** entre variables respuesta y explicativas
- Distribución **normal** de los residuos del modelo.
- Igualdad de **varianza** de los residuos en torno a la línea de la regresión.
- **Independencia** de las observaciones (i.e. de los datos).
- ¡Ojo con los **outliers**!

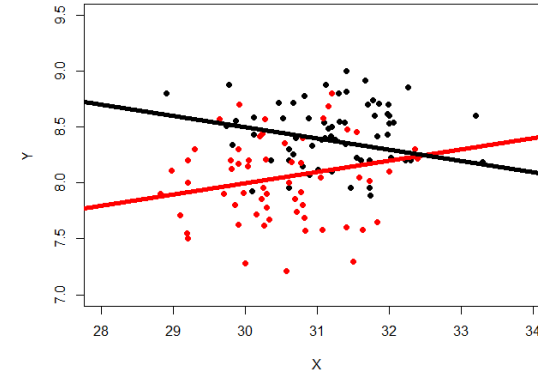
4.2. Modelo lineal múltiple con interacción

5. ¿Cuál es la hipótesis?

H0: Las vars. explicativas no afectan a la var. respuesta



Ha: Las vars. explicativas afectan a la var. respuesta



6. ¿Cómo se corre en R?



```
> aguacate<-lm (y ~ x1 * x2)  
> summary(aguacate)
```

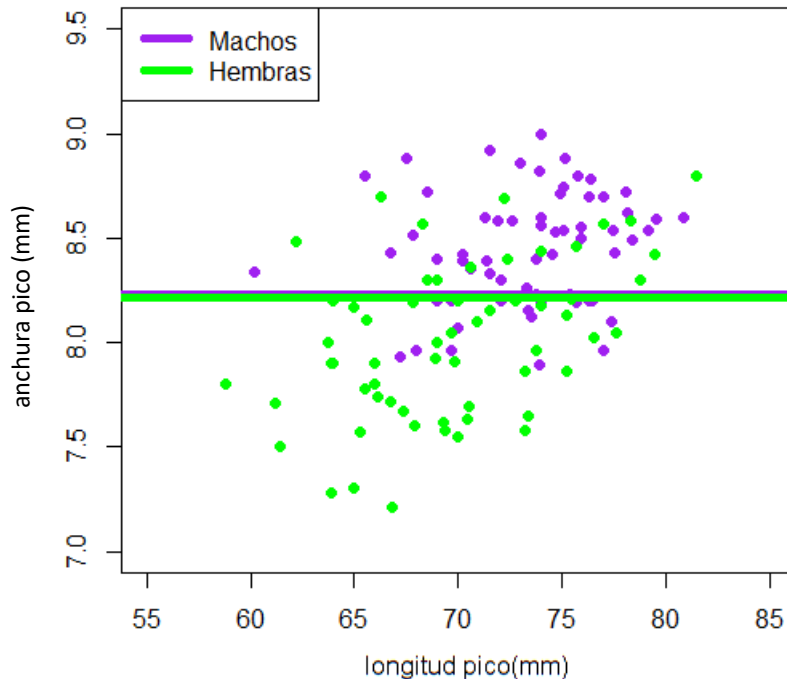
4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

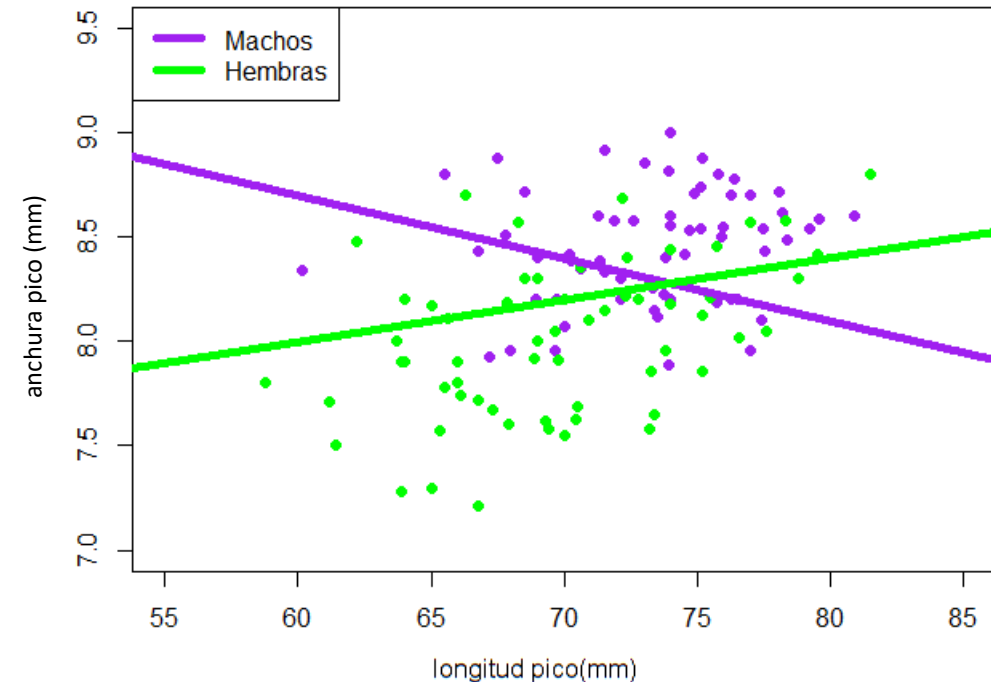
E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



H0: La anchura y la longitud del pico no están relacionado ni en hembras ni en machos



Ha: La anchura y la longitud del pico están relacionadas, y esta relación difiere entre machos y hembras



4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



Grupos: Female & Male (Hembras y machos)

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.67412 | -0.19634 | 0.00585 | 0.23708 | 0.65073 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|------------------------------|----------|------------|---------|----------|-----|
| (Intercept) | 3.54664 | 0.98028 | 3.618 | 0.000437 | *** |
| BJ\$BillLength | 0.18465 | 0.04050 | 4.559 | 1.25e-05 | *** |
| BJ\$KnownSexM | 3.38540 | 1.37280 | 2.466 | 0.015086 | * |
| BJ\$BillLength:BJ\$KnownSexM | -0.12525 | 0.05534 | -2.263 | 0.025438 | * |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom

Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076

F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



Grupos: Female & Male (Hembras y machos)

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.67412 | -0.19634 | 0.00585 | 0.23708 | 0.65073 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|------------------------------|----------|------------|---------|----------|-----|
| (Intercept) | 3.54664 | 0.98028 | 3.618 | 0.000437 | *** |
| BJ\$BillLength | 0.18465 | 0.04050 | 4.559 | 1.25e-05 | *** |
| BJ\$KnownSexM | 3.38540 | 1.37280 | 2.466 | 0.015086 | * |
| BJ\$BillLength:BJ\$KnownSexM | -0.12525 | 0.05534 | -2.263 | 0.025438 | * |

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom

Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076

F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

Estimación de coeficientes que definen las líneas de regresión

- **Intercepto de grupo de referencia**: Valor de "y" cuando x=0

4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



Grupos: Female & Male (Hembras y machos)

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.67412 | -0.19634 | 0.00585 | 0.23708 | 0.65073 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|------------------------------|----------|------------|---------|----------|-----|
| (Intercept) | 3.54664 | 0.98028 | 3.618 | 0.000437 | *** |
| BJ\$BillLength | 0.18465 | 0.04050 | 4.559 | 1.25e-05 | *** |
| BJ\$KnownSexM | 3.38540 | 1.37280 | 2.466 | 0.015086 | * |
| BJ\$BillLength:BJ\$KnownSexM | -0.12525 | 0.05534 | -2.263 | 0.025438 | * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom

Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076

F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

Estimación de coeficientes que definen las líneas de regresión

- **Intercepto de grupo de referencia:** Valor de “y” cuando x=0

- **Var. Explicativa continua = pendiente de grupo de referencia:** efecto del incremento de una unidad de “x” sobre “y” para el grupo de referencia

4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



Grupos: Female & Male (Hembras y machos)

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.67412 | -0.19634 | 0.00585 | 0.23708 | 0.65073 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|------------------------------|----------|------------|---------|--------------|
| (Intercept) | 3.54664 | 0.98028 | 3.618 | 0.000437 *** |
| BJ\$BillLength | 0.18465 | 0.04050 | 4.559 | 1.25e-05 *** |
| BJ\$KnownSexM | 3.38540 | 1.37280 | 2.466 | 0.015086 * |
| BJ\$BillLength:BJ\$KnownSexM | -0.12525 | 0.05534 | -2.263 | 0.025438 * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom

Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076

F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

Estimación de coeficientes que definen las líneas de regresión

- **Intercepto de grupo de referencia:** Valor de “y” cuando x=0

- **Var. Explicativa continua** = pendiente de grupo de referencia: efecto del incremento de una unidad de “x” sobre “y” para el grupo de referencia

- **Var. Explicativa categórica** = intercepto del segundo grupo: efecto del cambio de grupo respecto al grupo de referencia

4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



Grupos: Female & Male (Hembras y machos)

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.67412 | -0.19634 | 0.00585 | 0.23708 | 0.65073 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|------------------------------|----------|------------|---------|----------|-----|
| (Intercept) | 3.54664 | 0.98028 | 3.618 | 0.000437 | *** |
| BJ\$BillLength | 0.18465 | 0.04050 | 4.559 | 1.25e-05 | *** |
| BJ\$KnownSexM | 3.38540 | 1.37280 | 2.466 | 0.015086 | * |
| BJ\$BillLength:BJ\$KnownSexM | -0.12525 | 0.05534 | -2.263 | 0.025438 | * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom

Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076

F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

Estimación de coeficientes que definen las líneas de regresión

- **Intercepto de grupo de referencia:** Valor de “y” cuando $x=0$

- **Var. Explicativa continua** = pendiente de grupo de referencia: efecto del incremento de una unidad de “x” sobre “y” para el grupo de referencia

- **Var. Explicativa categórica** = intercepto del segundo grupo: efecto del cambio de grupo respecto al grupo de referencia

- **Interacción entre vars. explicativas**= pendiente del segundo grupo: efecto del incremento de una unidad de “x” sobre “y” para el segundo grupo

4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



Grupos: Female & Male (Hembras y machos)

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.67412 | -0.19634 | 0.00585 | 0.23708 | 0.65073 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|------------------------------|----------|------------|---------|----------|-----|
| (Intercept) | 3.54664 | 0.98028 | 3.618 | 0.000437 | *** |
| BJ\$BillLength | 0.18465 | 0.04050 | 4.559 | 1.25e-05 | *** |
| BJ\$KnownSexM | 3.38540 | 1.37280 | 2.466 | 0.015086 | * |
| BJ\$BillLength:BJ\$KnownSexM | -0.12525 | 0.05534 | -2.263 | 0.025438 | * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076
F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

- Las **hembras** tienen una anchura de pico de 3.54 mm cuando tienen un pico de longitud de 0 mm
- El pico de las **hembras** incrementa 0.18 mm anchura al incrementar 1 mm de longitud
- Los **machos** tienen una anchura de pico de 3.54+3.38 mm cuando tienen un pico de longitud de 0 mm
- El pico de los **machos** incrementa en 0.18-0.12 mm de anchura al incrementar 1 mm de longitud

Ecuaciones

Anchura pico hembras= $3.54 + 0.18 \cdot \text{Long. pico}$

Anchura pico machos= $(3.54+3.38) + (0.18-0.12) \cdot \text{Long. pico}$

4.2. Modelo lineal múltiple con interacción

7. ¿Cómo se interpreta el resultado de R?

E.g. Queremos estudiar la morfología del pico de los arrendajos azules, concretamente si cuanto más ancho es el pico, su longitud también es mayor. Además, queremos ver si esta relación morfológica difiere entre machos y hembras.



Grupos: Female & Male (Hembras y machos)

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.67412 | -0.19634 | 0.00585 | 0.23708 | 0.65073 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|------------------------------|----------|------------|---------|----------|-----|
| (Intercept) | 3.54664 | 0.98028 | 3.618 | 0.000437 | *** |
| BJ\$BillLength | 0.18465 | 0.04050 | 4.559 | 1.25e-05 | *** |
| BJ\$KnownSexM | 3.38540 | 1.37280 | 2.466 | 0.015086 | * |
| BJ\$BillLength:BJ\$KnownSexM | -0.12525 | 0.05534 | -2.263 | 0.025438 | * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076
F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

P-valor:

Significancia de los valores en el grupo de ref: Valor distinto de cero

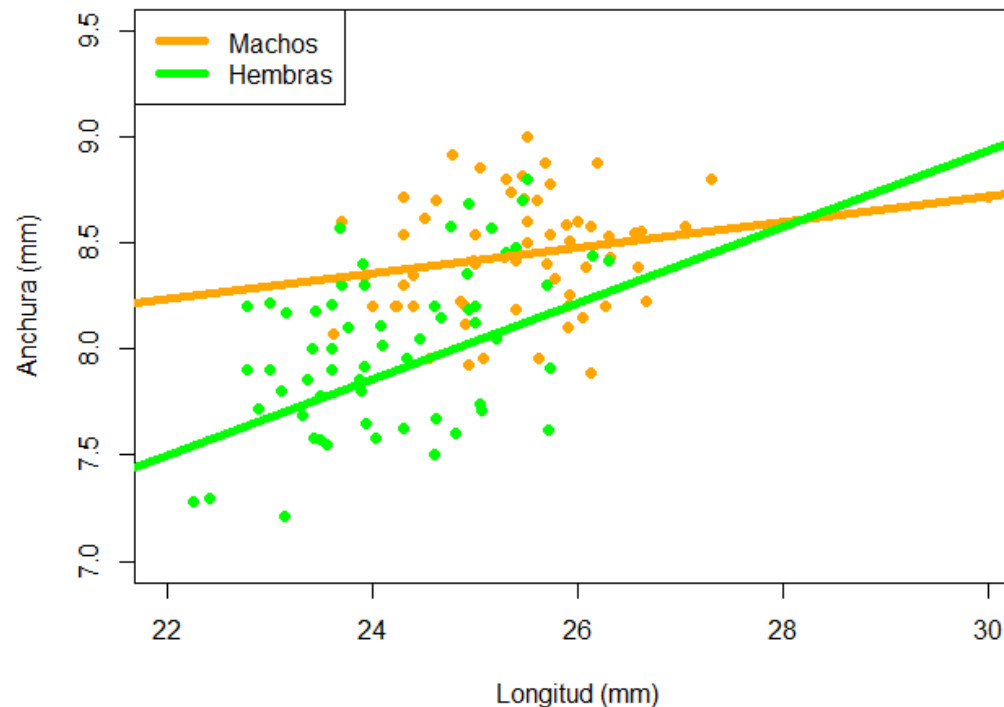
Significancia de los valores en otros grupos: Valor distinto del de grupo de referencia

El intercepto de las líneas de regresión es distinto

La pendiente de las líneas de regresión es distinta == La relación entre las variables estudiadas difiere entre grupos

4.2. Modelo lineal múltiple con interacción

8. ¿Cómo se puede representar?



```
>plot(BJ$BillDepth[BJ$KnownSex=="M"]~BJ$BillLength[BJ$KnownSex=="M"],  
      xlim=c(22,30),ylim=c(7,9.5),col="orange",pch=19,  
      xlab="Longitud(mm)",ylab="Anchura (mm)")  
>points(BJ$BillDepth[BJ$KnownSex=="F"]~BJ$BillLength[BJ$KnownSex=="F"],col="green",pch=19)  
  
>abline(a=3.54,b=0.18,col="green",lwd=5)  
>abline(a=3.54+3.38,b=0.18-0.12,col="orange",lwd=5)  
  
>legend("topleft",legend = c("Machos","Hembras"), col=c("orange","green"),lwd=5)
```

4.1. Ejercicios de Modelos Lineales

Ejercicio: 4. Ejer_LMs (segunda parte)

Ejercicio: 5. Ejer_LMConInteracción

4.2. Regresión lineal múltiple (aditiva)

1. ¿Qué es?

Es un método de estimación de la relación entre una variable dependiente y varias variables independientes.

→ El objetivo es encontrar la línea que mejor defina los datos

2. ¿Cómo difiere del modelo lineal simple?

Incluimos más de una variable explicativa para estudiar cómo todas afectan a nuestra variable respuesta

$$y_i = \underset{\substack{\text{Intercepto} \\ \uparrow}}{b_0} + \underset{\substack{\text{Efecto de } x_1 \\ \uparrow}}{b_1}x_{i1} + \underset{\substack{\text{Efecto de } x_2 \\ \uparrow}}{b_2}x_{i2} + \cdots + \underset{\text{Error o residuo}}{\varepsilon_i}$$

3. ¿Qué tipo de datos se necesitan?

Variable respuesta (dep.; y) → Numérica y continua

Variables explicativas (indep.; x) → Continuas y/o categóricas

← Variables continuas → pendiente: efecto del incremento de una unidad de x sobre y
Variables categóricas → intercepto: efecto del cambio de grupo de x sobre y

4.2. Regresión lineal múltiple (aditiva)

4. ¿Qué asunciones tiene?

- Variables indeps. (x) **no correlacionadas**
- **Principio de parsimonia**
- **Relación lineal** entre vars. respuesta y explicativas
- Distribución normal de los residuos del modelo (o de las variables numéricas)
- Igualdad de **varianza** de los residuos en torno a la línea de la regresión
- **Independencia** de las observaciones

¡Ojo con los **outliers**!

$$y_i = b_0 + b_1 \text{cont. } x_{i1} + b_2 \text{cat. } x_{i2} + \varepsilon_i$$

5. Matemáticamente, ¿cuál es la hipótesis?

H0: No existe una relación entre las variables estudiadas

$$b_1 = 0$$

H0: Los distintos grupos de la var. categórica no difieren en la variable respuesta

$$b_0 = b_2$$

Ha: Existe una relación lineal entre las variables

$$b_1 \neq 0$$

Ha: Los distintos grupos de la var. categórica no difieren en la variable respuesta

$$b_0 \neq b_2$$

6. ¿Cómo se corre en R?



```
> guisante<-lm(data=db, y ~ xcont + xcat)
> summary(guisante)
```

4.2. Ejercicios de Modelos Lineales

Ejercicio: 4. Ejer_LMs (segunda parte)