Lunes	Martes	Miércoles	Jueves	Viernes
• Conceptos básicos • T-test	• One-way ANOVA • Two-way ANOVA	• LM Simples • LM múltiples sin interacción	• LM múltiples con interacción	• Resolución de práctica (?) ¿Cambio de hora? LM mixtos GLM Cervezas

- **Asunciones**: Todos los análisis estadísticos asumen ciertas características de los datos.
  Se deben comprobar antes de llevar a cabo el modelo

Normalidad

```
>shapiro.test(db$Mass)

Shapiro-Wilk normality test

data: db$Mass
W = 0.98599, p-value = 0.2366
```

H0: Distribución normal
Ha: Distribución no normal

Homogeneidad de varianza (Homocedasticidad)

```
> leveneTest(db$Mass~db$KnownSex)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   1  4.4591 0.03677 *
      121
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H0: Homogeneidad en varianza
Ha: Heterogeneidad en varianza

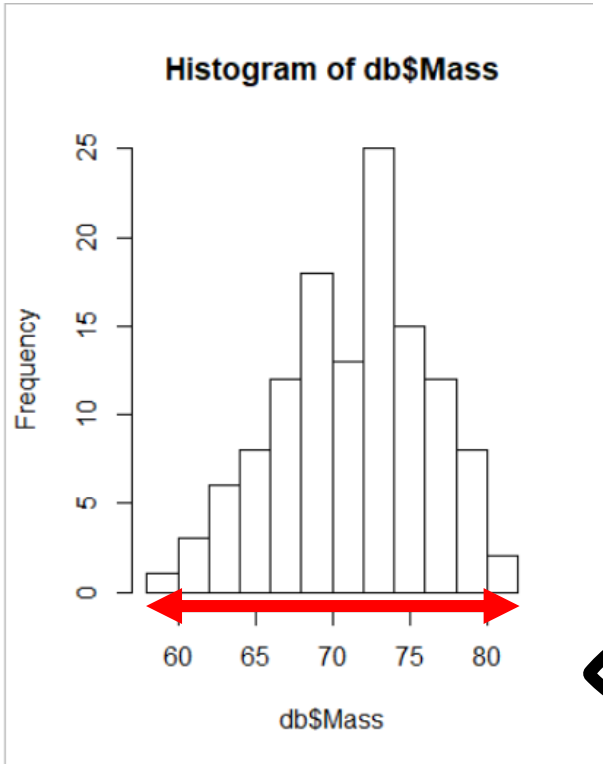
Ambos son test estadísticos PERO **NO NOS PERMITEN COMPROBAR NUESTRA HIPÓTESIS**, SOLO LAS ASUNCIONES DEL MODELO A UTILIZAR

- **Varianza** vs. **Rango** : Ambas son medidas de dispersión, pero…

Varianza:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Rango:
mín y máx de variable numérica



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

-Febrero 2021-

Carlota Solano
Álvaro Arredondo

4.1. Regresión lineal simple

1. ¿Qué es?

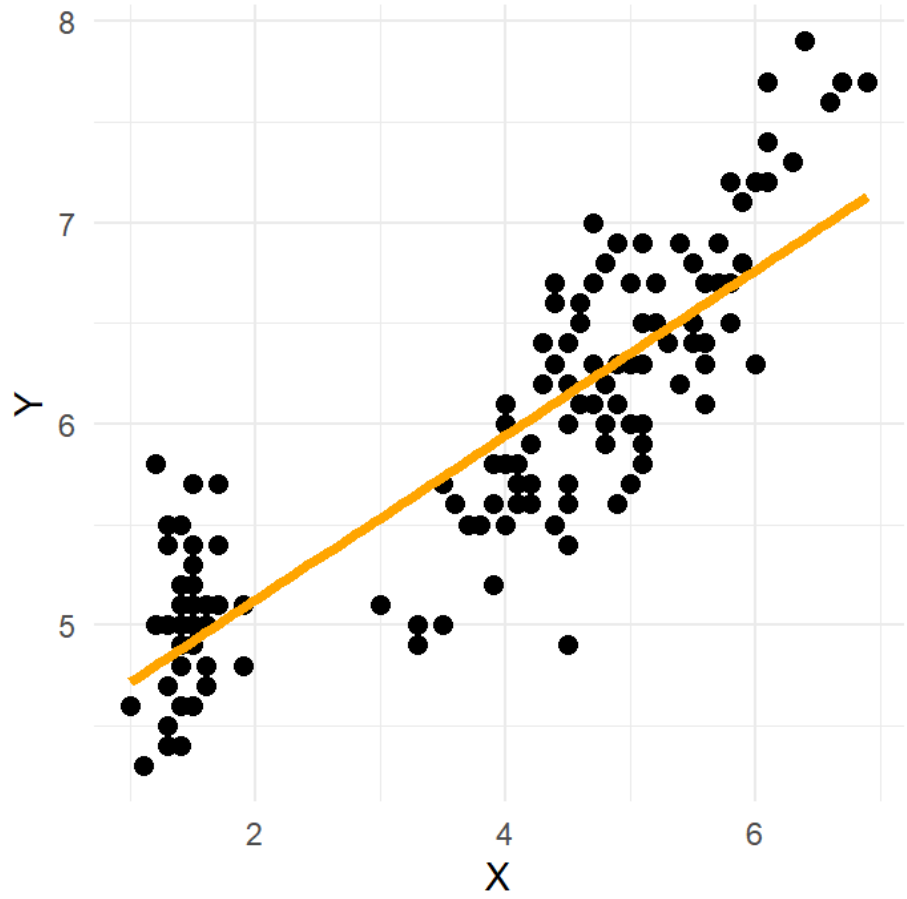
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ 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?

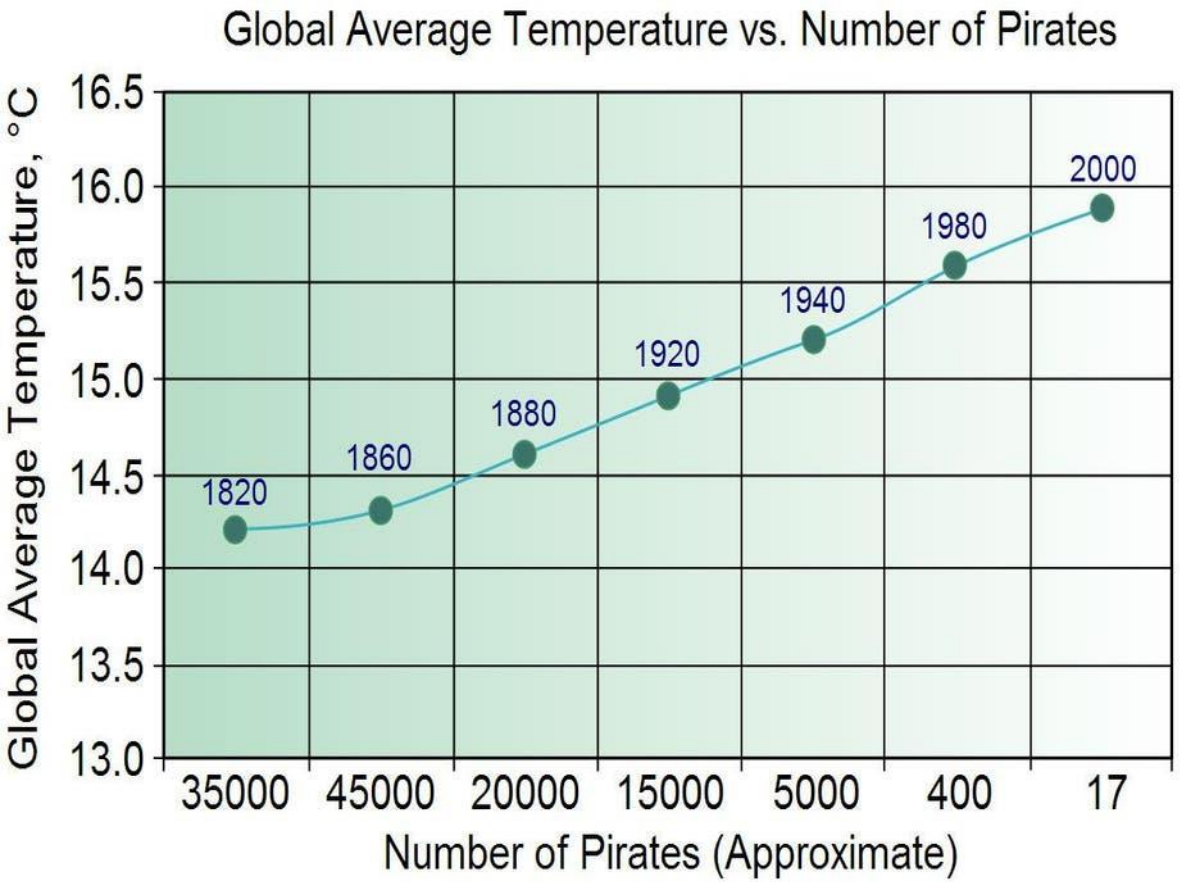
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

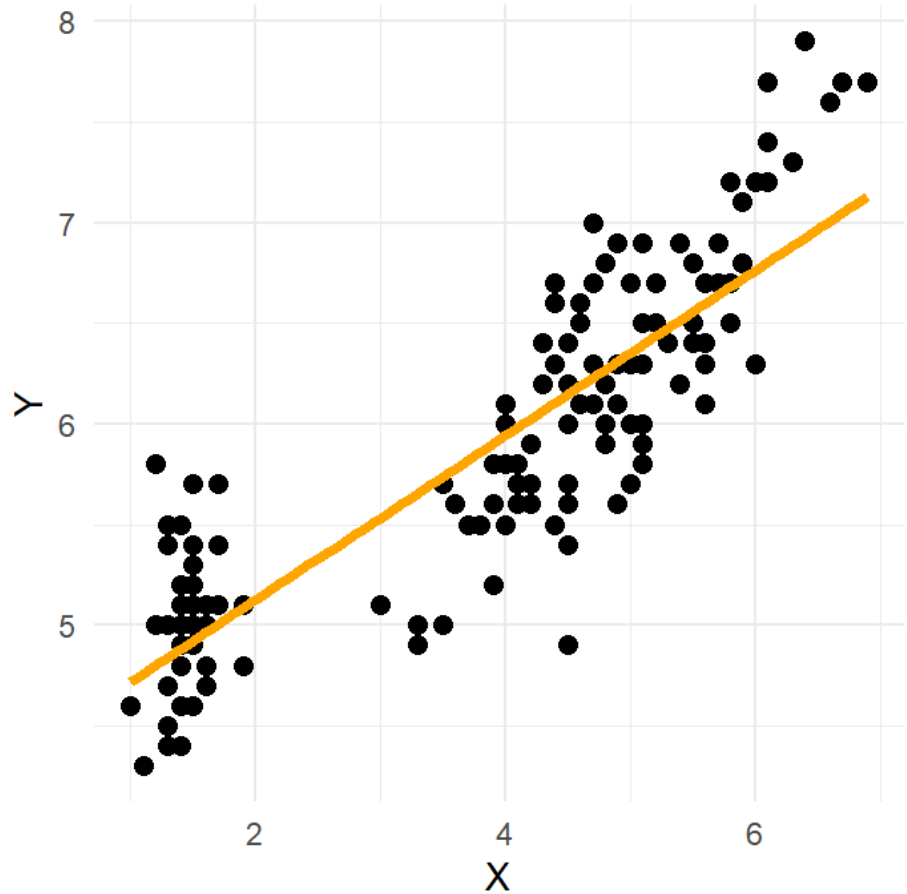
→ El objetivo es encontrar la línea que mejor defina los datos

2. ¿Cuándo se puede utilizar?

Cuando quieres definir cómo se relacionan dos elementos.

Correlación no implica causalidad





4.1. Regresión lineal simple

1. ¿Qué es?

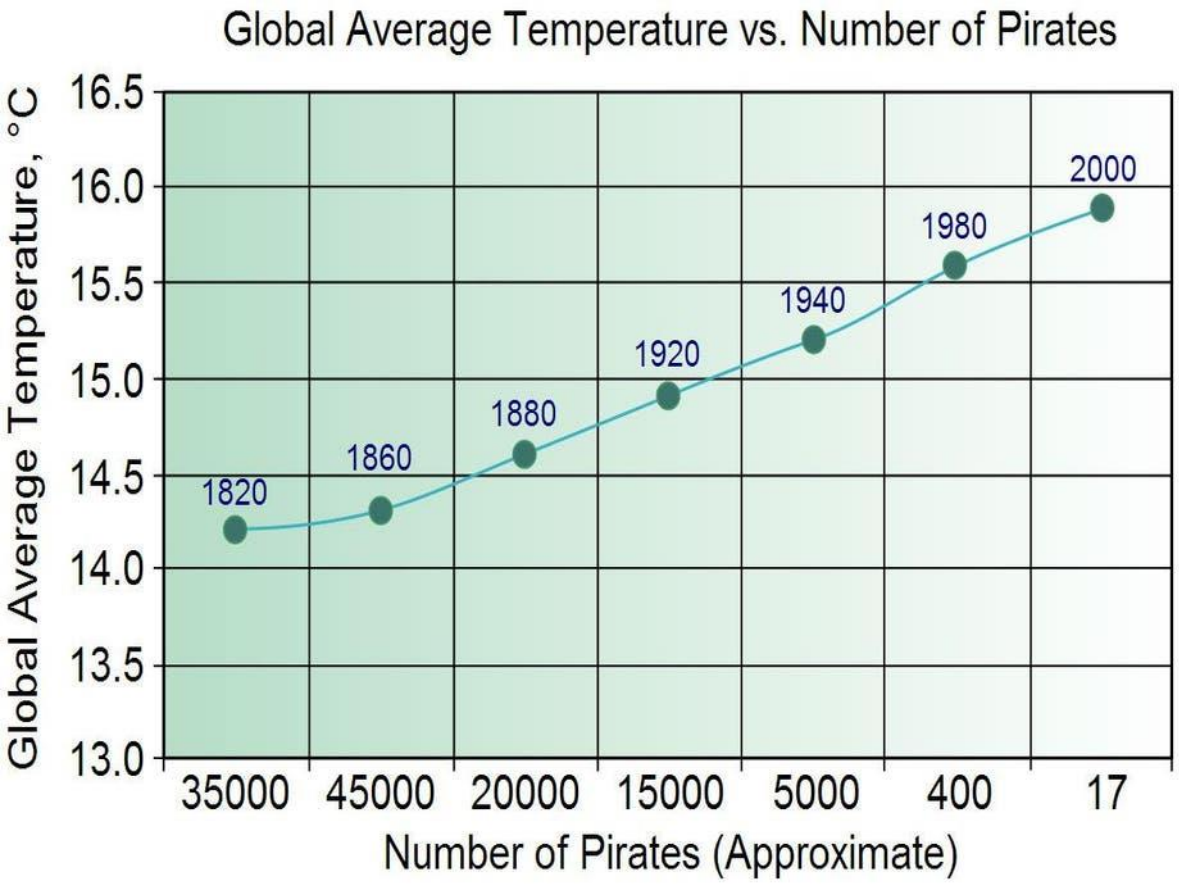
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

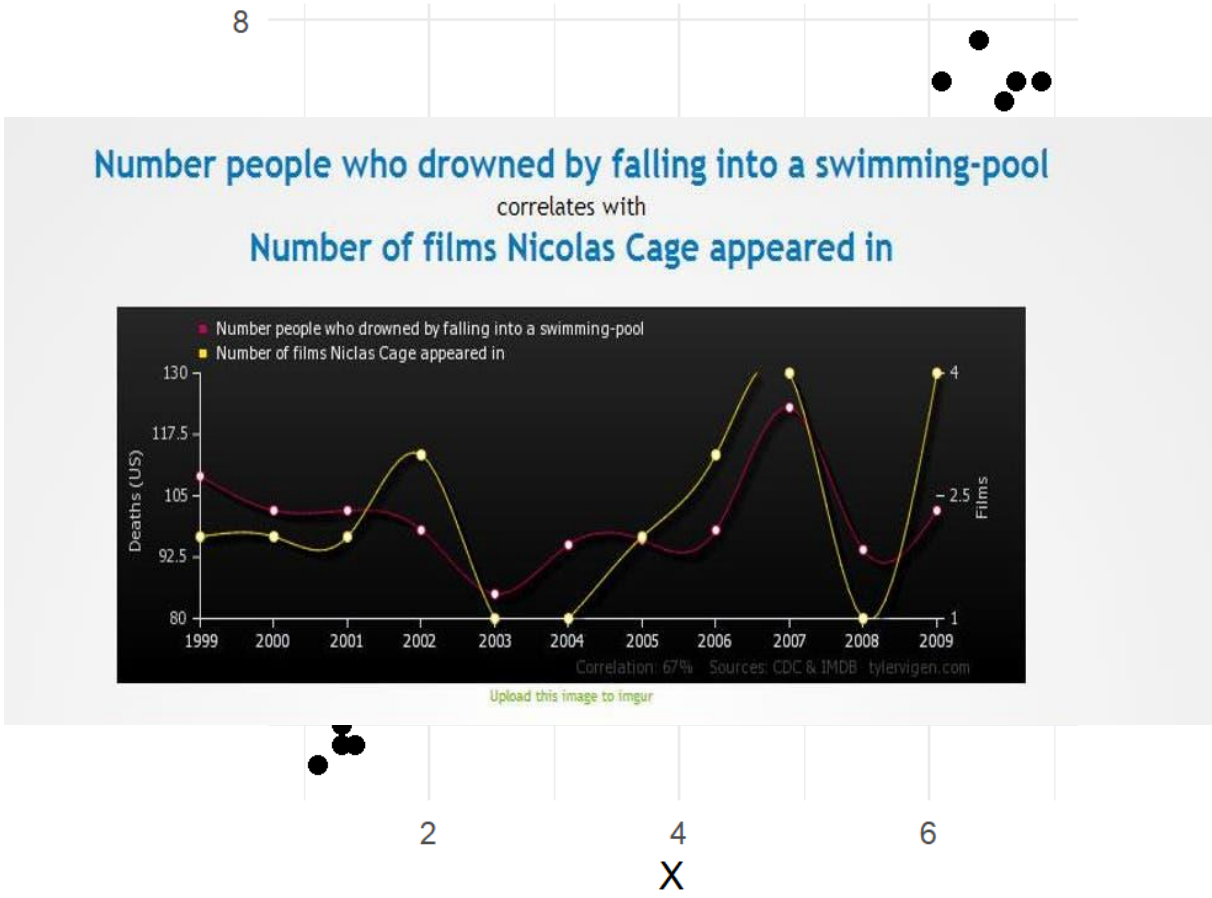
→ El objetivo es encontrar la línea que mejor defina los datos

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

3. ¿Qué tipo de datos se necesitan?

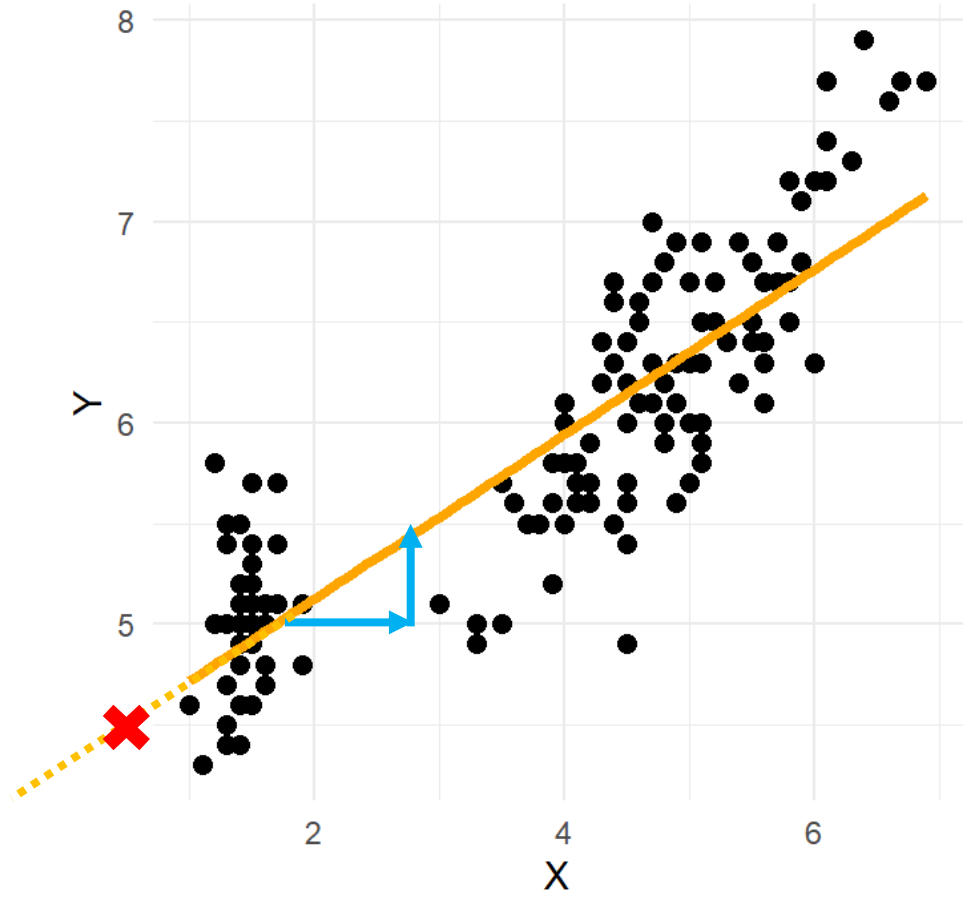
Variable respuesta (dep.; y) → Numérica y continua

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

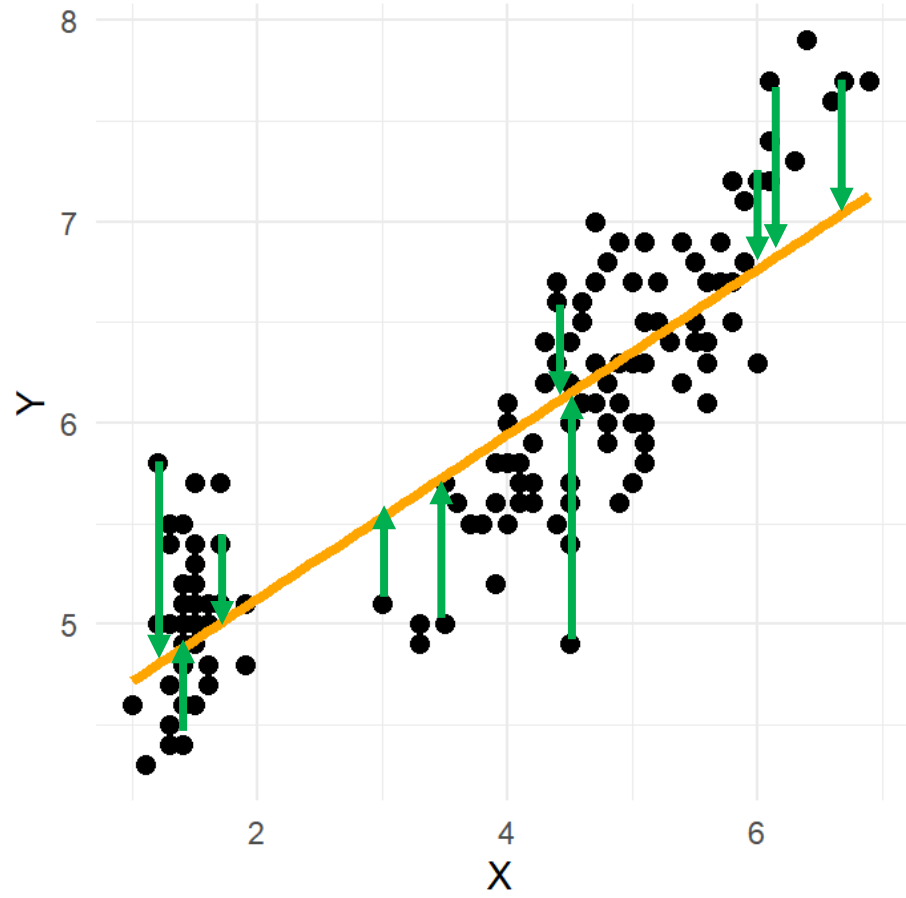
Variable explicativa (indep.; x) → Numérica y continua

$$y = a + m\,x$$

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

Intercepto Pendiente Error o residuo

$$\varepsilon = valor\ real - predicho\ por\ modelo$$



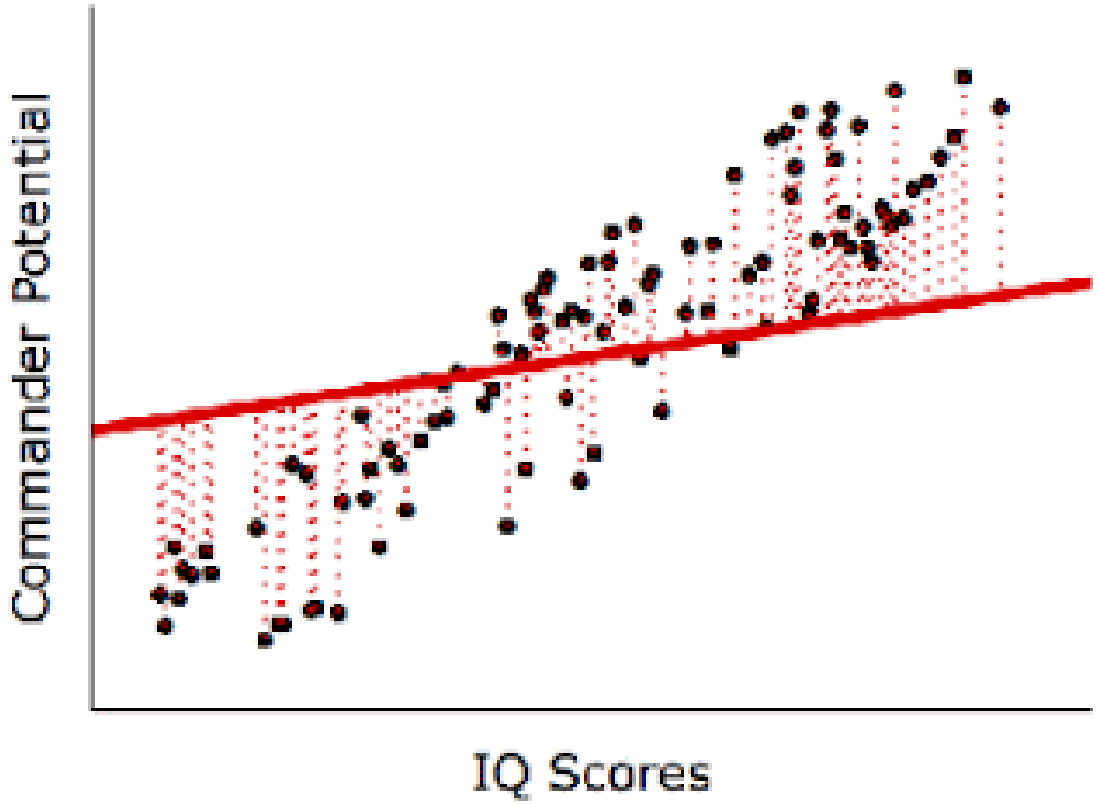
4.1. Regresión lineal simple

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 b0 y b1 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

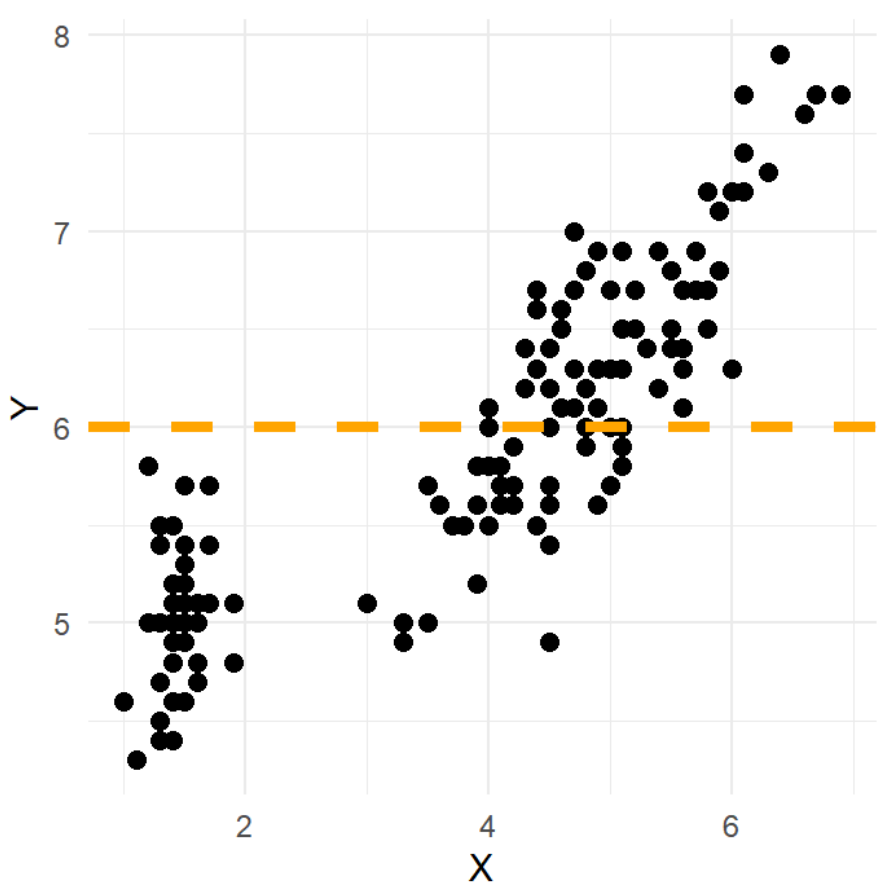
1. ¿Qué es?

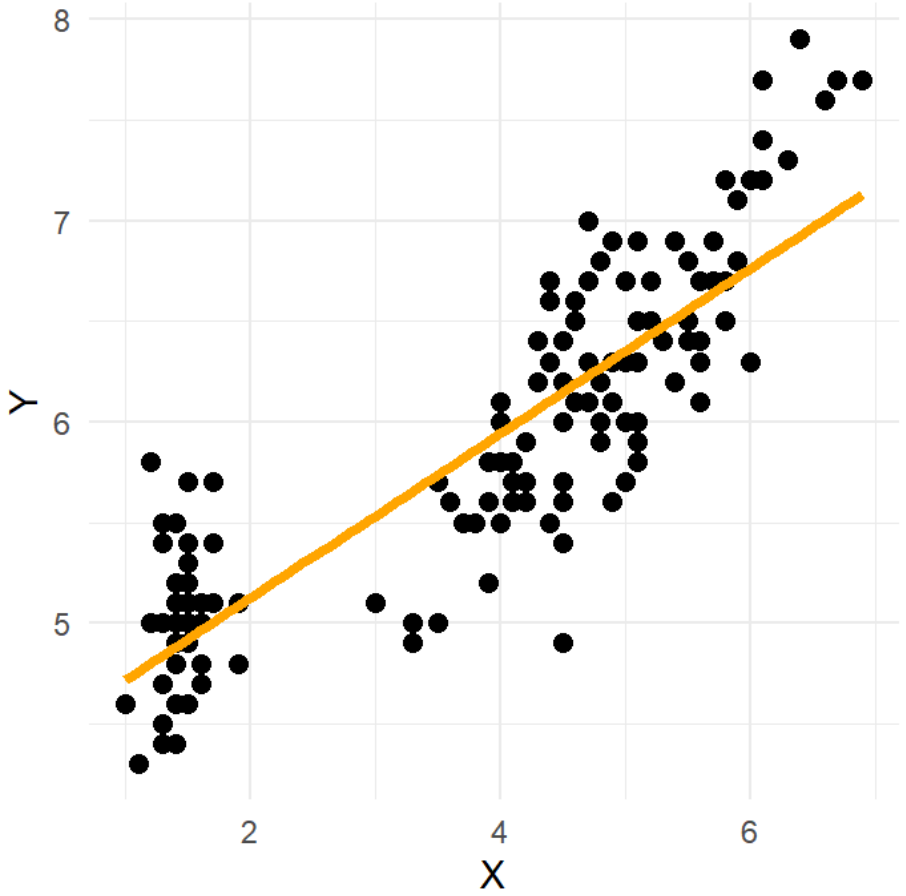
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El **objetivo** es encontrar la línea que mejor defina los datos = ***Encontrar los valores de b0 y b1 que nos permiten minimizar la suma de los cuadrados de los residuos***

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

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





4.1. Regresión lineal simple

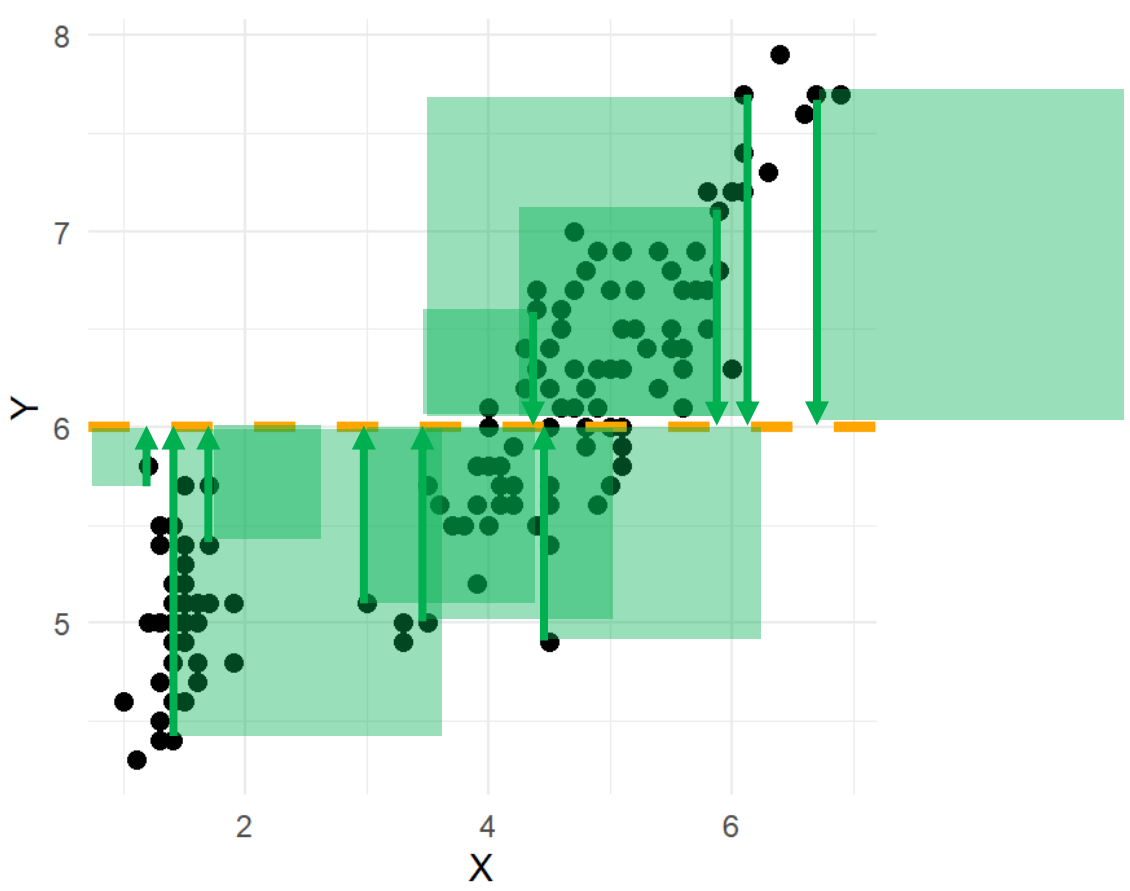
1. ¿Qué es?

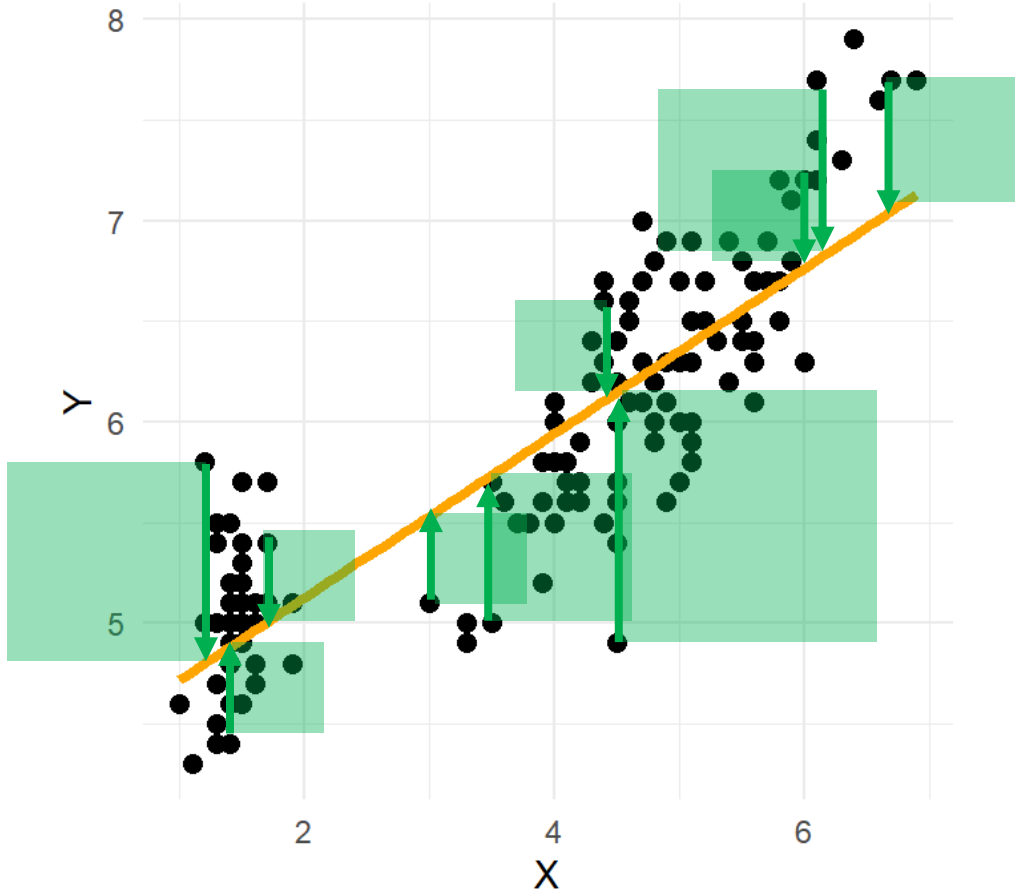
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

→ El **objetivo** es encontrar la línea que mejor defina los datos = ***Encontrar los valores de b0 y b1 que nos permiten minimizar la suma de los cuadrados de los residuos***

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

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





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 b0 y b1 que nos permiten minimizar la suma de los cuadrados de los residuos***

$SSres = \sum_{i=1}^{n}(\boldsymbol{\varepsilon_i})^2$ →Mide la varianza no explicada por el modelo

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

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

4.1. Regresión lineal simple

1. ¿Qué es?

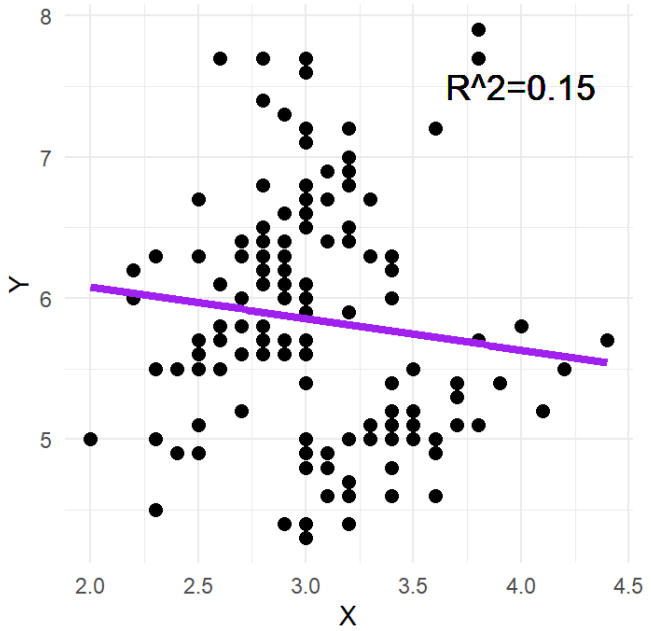
Es un método de estimación de la relación entre una variable dependiente y otra independiente.

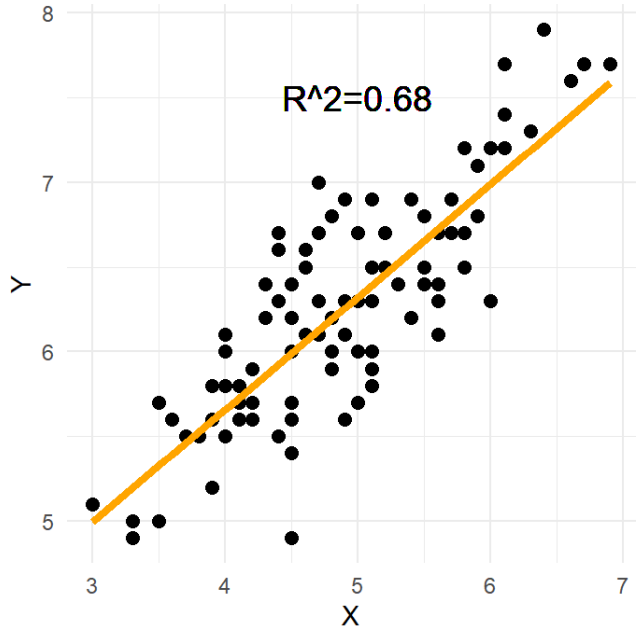
→ El **objetivo** es encontrar la línea que mejor defina los datos = ***Encontrar los valores de b0 y b1 que nos permiten minimizar la suma de los cuadrados de los residuos***

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

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

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





4.1. Regresión lineal simple

4. ¿Qué asunciones tiene?

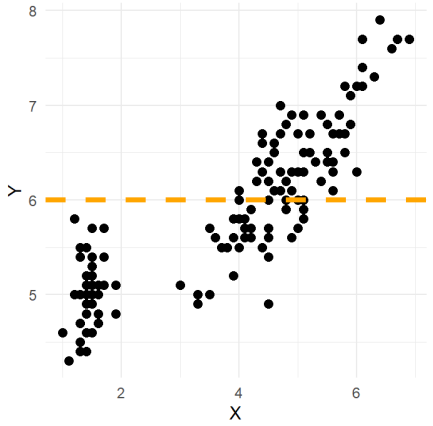
- La relación entre variables es **lineal.**
- Distribución **normal** de los residuos (o de las variables) del modelo.
- Igualdad de **varianza** de los residuos en torno a la línea de la regresión.
- **Independencia** de las observaciones (i.e. de los datos).

5. Matemáticamente, ¿cuál es la hipótesis?

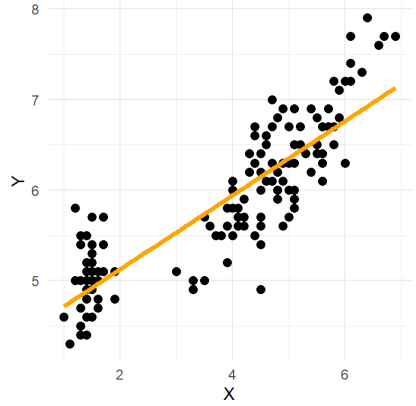
H0: No existe una relación entre las variables estudiadas

Ha: Existe una relación lineal entre las variables

b1 = 0



b1 ≠ 0



6. ¿Cómo se corre en R?

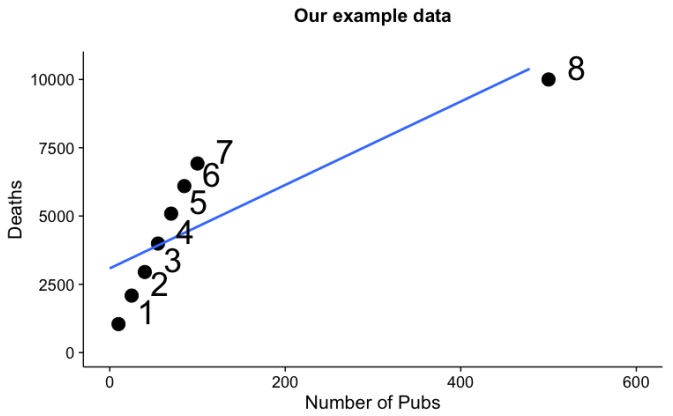
```
> holi<-lm (y ~ x)
> summary(holi)
```

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!

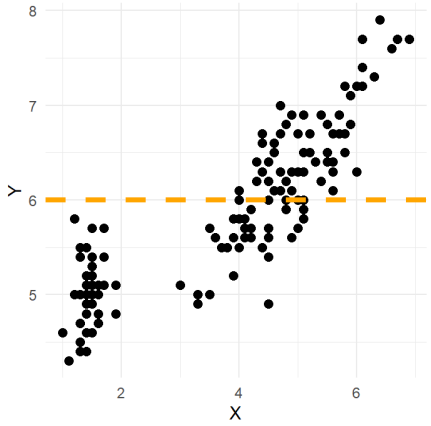


5. Matemáticamente, ¿cuál es la hipótesis?

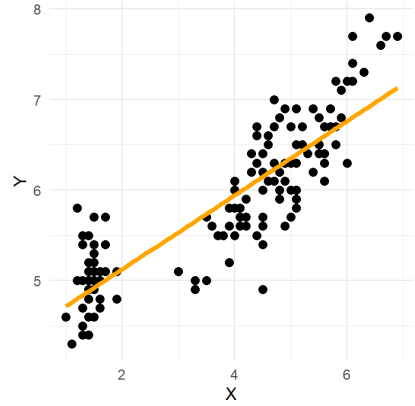
H0: No existe una relación entre las variables estudiadas

Ha: Existe una relación lineal entre las variables

b1 = 0



b1 ≠ 0



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 = 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

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 = 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

(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 = 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
(!) ±1.96* s.e.= 95%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 = 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
(!) ±1.96* s.e.= 95%CI

T- value: estimación/ s.e. --> ↑ t value= ↓ 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 (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 = 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
(!) ±1.96* s.e.= 95%CI

T- value: estimación/ s.e. --> ↑ t value= ↓ 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 (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

(…) R-squared: R^2: 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 (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

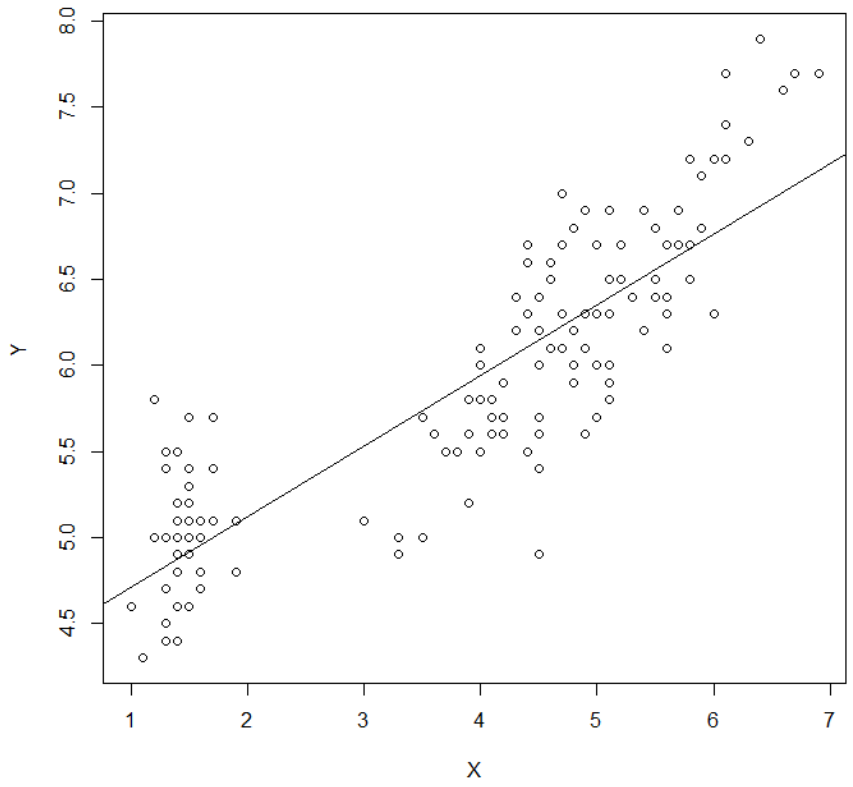
Degrees of freedom: grados de libertad:

(…) R-squared: R^2: proporción de la varianza explicada por el modelo.

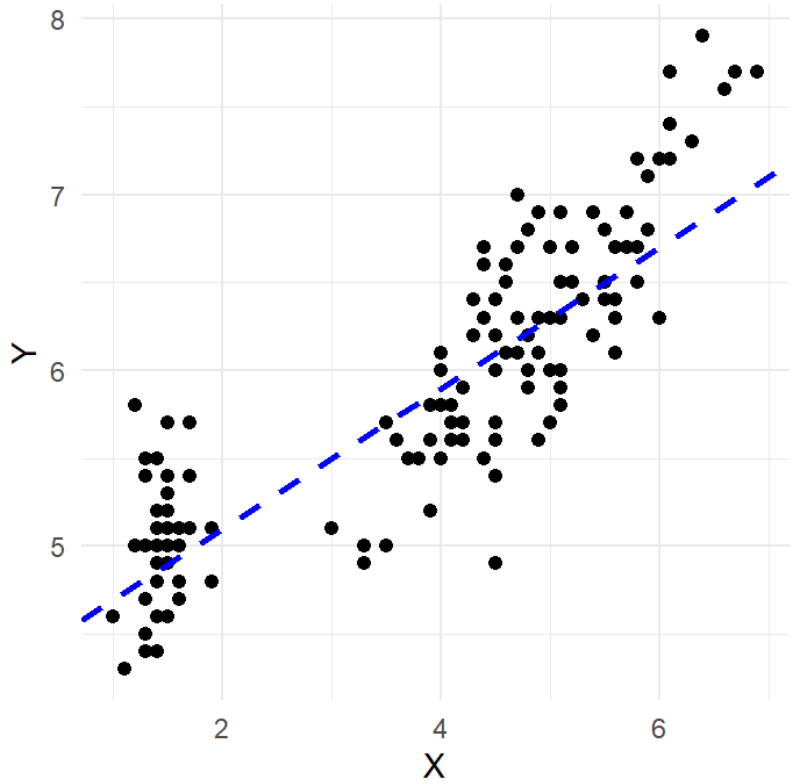
$$(R^2 = 1 - \frac{SS_{res}}{SS_{tot}})$$

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

8. ¿Cómo se puede representar?




```
> 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. Ejercicios de Modelos Lineales

Ejercicio: 4. Ejer_LMs (primera parte)

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

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + \varepsilon_i$$

Intercepto | Efecto de x1 | Efecto de x2 | Error o residuo

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**!

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

5. Matemáticamente, ¿cuál es la hipótesis?

H0: No existe una relación entre las variables estudiadas

b1 = 0

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

b0 = b2

Ha: Existe una relación lineal entre las variables

b1 ≠ 0

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

b0 ≠ b2

6. ¿Cómo se corre en R?

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

4.2. Regresión lineal múltiple (aditiva)

7. ¿Cómo se interpreta el resultado de R?

E.g. El arrendajo azul requiere de un pico largo para alcanzar a sus presas (larvas dentro de troncos). Queremos estudiar si aquellos individuos con pico más largo, tienen una mayor masa corporal, y si esta masa difiere entre machos y hembras. ¿Presentan una mayor masa aquellos individuos que tienen un pico más largo, y difiere esta entre machos y hembras?

4.2. Regresión lineal múltiple (aditiva)

7. ¿Cómo se interpreta el resultado de R?

E.g. El arrendajo azul requiere de un pico largo para alcanzar a sus presas (larvas dentro de troncos). Queremos estudiar si aquellos individuos con pico más largo, tienen una mayor masa corporal, y si esta masa difiere entre machos y hembras. ¿Existe una relación entre la masa y la longitud del pico en arrendajos azules, y difieren machos y hembras en su morfología?

```
> m1<-lm(data=BJ, Mass~BillLength+KnownSex)
> summary(m1)

Call:
lm(formula = Mass ~ BillLength + KnownSex, data = BJ)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1547  -2.9219   0.2954   2.8162  10.0722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.0083     9.6042   4.166 5.88e-05 ***
BillLength    1.2321     0.3965   3.108  0.00235 **
KnownSexM     1.8436     0.9287   1.985  0.04940 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 4.314 on 120 degrees of freedom
Multiple R-squared:  0.1944,    Adjusted R-squared:  0.181
F-statistic: 14.48 on 2 and 120 DF,  p-value: 2.326e-06
```

Estimación de coeficientes que definen la línea de regresión

- (Intercept) = intercepto de grupo de referencia: Valor de y cuando x=0
- *Var. Explicativa continua* = pendiente: efecto del incremento de una unidad de Head sobre Mass
- Var. Explicativa categórica = intercepto: efecto del cambio de grupo respecto al grupo de referencia

(Poco valor biológico) Cuando un arrendajo HEMBRA no tiene pico, su masa es de 40.01 g.
Por cada mm de pico más, la masa de los arrendajos HEMBRA Y MACHO incrementa 1.23 g.
Cuando un arrendajo MACHO no tiene pico, su masa es de (40.01+1.84) g.

4.2. Regresión lineal múltiple (aditiva)

7. ¿Cómo se interpreta el resultado de R?

E.g. El arrendajo azul requiere de un pico largo para alcanzar a sus presas (larvas dentro de troncos). Queremos estudiar si aquellos individuos con pico más largo, tienen una mayor masa corporal, y si esta masa difiere entre machos y hembras. ¿Existe una relación entre la masa y la longitud del pico en arrendajos azules, y difieren machos y hembras en su morfología?

```
> m1<-lm(data=BJ, Mass~BillLength+KnownSex)
> summary(m1)

Call:
lm(formula = Mass ~ BillLength + KnownSex, data = BJ)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1547  -2.9219   0.2954   2.8162  10.0722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.0083     9.6042   4.166 5.88e-05 ***
BillLength    1.2321     0.3965   3.108  0.00235 **
KnownSexM     1.8436     0.9287   1.985  0.04940 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 4.314 on 120 degrees of freedom
Multiple R-squared:  0.1944,    Adjusted R-squared:  0.181
F-statistic: 14.48 on 2 and 120 DF,  p-value: 2.326e-06
```

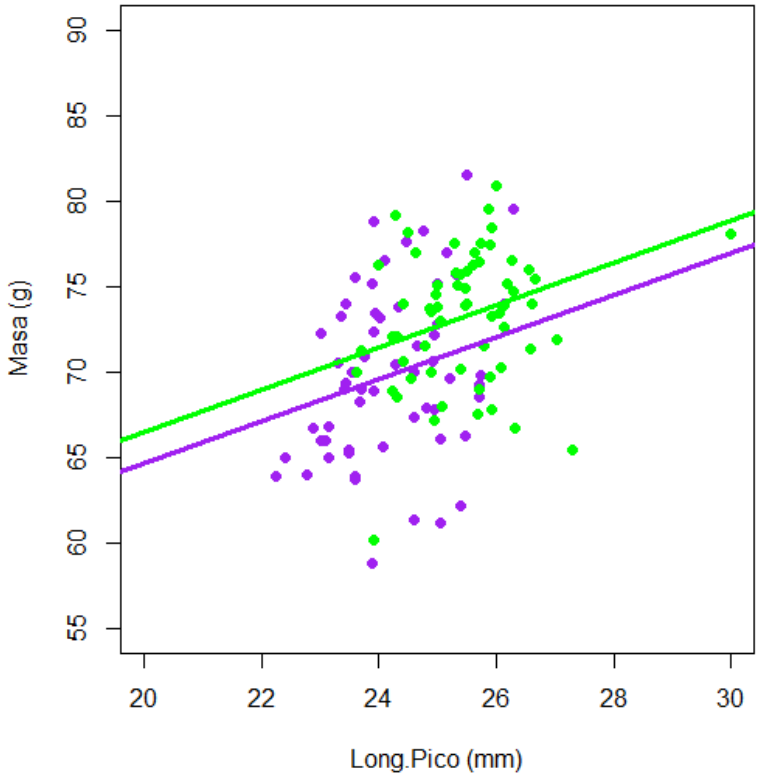
Significancia de coeficientes que definen la línea de regresión

- Intercepto de grupo de referencia difiere significativamente de cero

-La pendiente (var. explicativa continua) difiere significativamente de cero, i.e. existe una relación entre variables

-Intercepto del segundo grupo alfanumérico difiere significativamente del intercepto del grupo de referencia

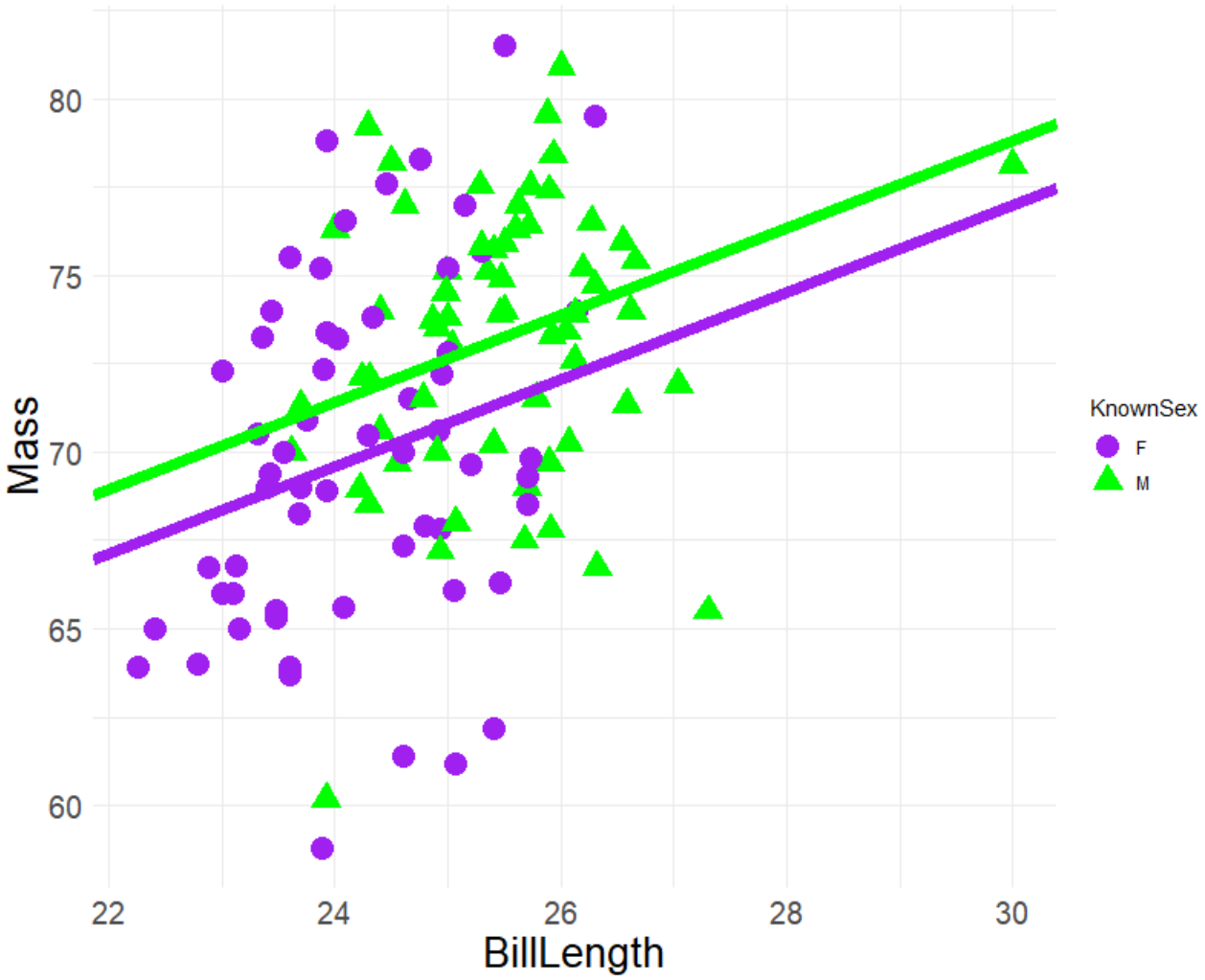
8. ¿Cómo se puede representar?




```
> plot(BJ$Mass[BJ$KnownSex=="F"]~BJ$BillLength[BJ$KnownSex=="F"],
       col="purple",xlim=c(20,30),ylim=c(55,90),
       xlab="Long.Pico (mm)",ylab="Masa (g)")
> points(BJ$Mass[BJ$KnownSex=="M"]~BJ$BillLength[BJ$KnownSex=="M"],col="green")
> abline(a=intercepto de grupo F, b=pendiente, col="purple”)
> abline(a=intercepto de grupo M, b=pendiente ,col="green")
```

4.2. Regresión lineal múltiple (aditiva)

8. ¿Cómo se puede representar?




```
>ggplot(BJ, aes(x=BillLength,y=Mass,
                color=KnownSex,shape=KnownSex))+
  geom_point(size=5)+
  theme_minimal()+
  theme(axis.text=element_text(size=15),
        axis.title = element_text(size=20))+
  scale_color_manual(values=c("purple","green"))+
  scale_size_manual(values=c(5))+
  geom_abline(intercept=40.0083,
              slope=1.2321,col="purple",size=3)+
  geom_abline(intercept=40.0083+1.8436,
              slope=1.2321,col="green",size=3)
```

4.2. Ejercicios de Modelos Lineales

Ejercicio: 4. Ejer_LMs (segunda parte)