Estadística y modelos predictivos

Santiago Caño Muñiz

All models are wrong, but some are useful

George Box

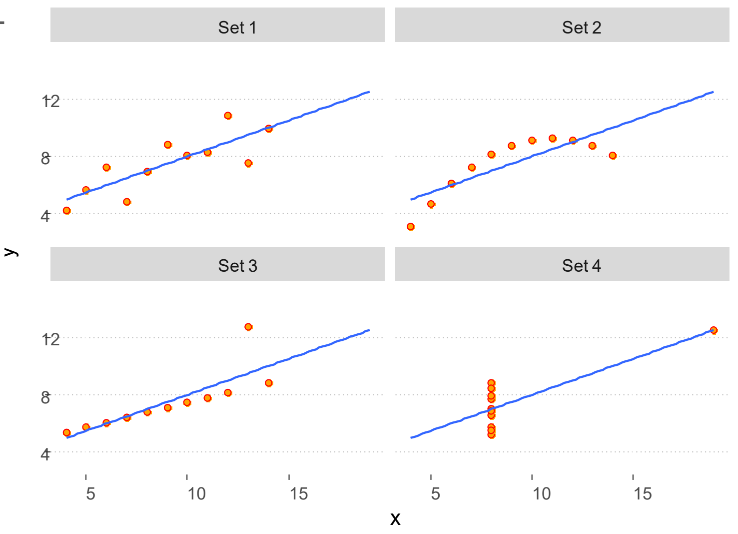
El ciclo investigador

El primer paso para tener una intuición es observer los datos

La medida

Estadísticos descriptivos

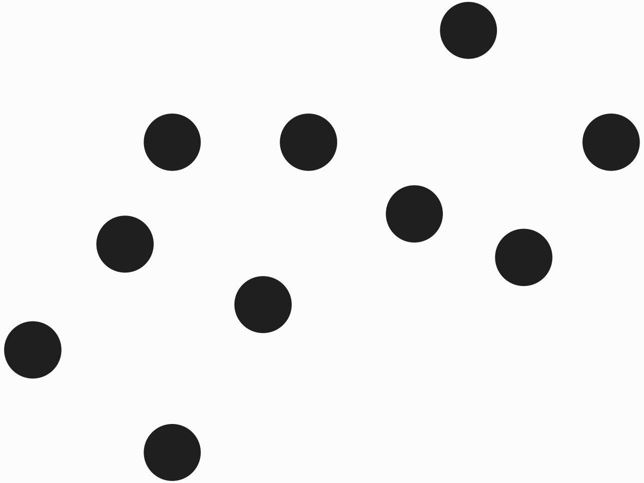
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



La medida

El primer paso para tener una intuición es observer los datos

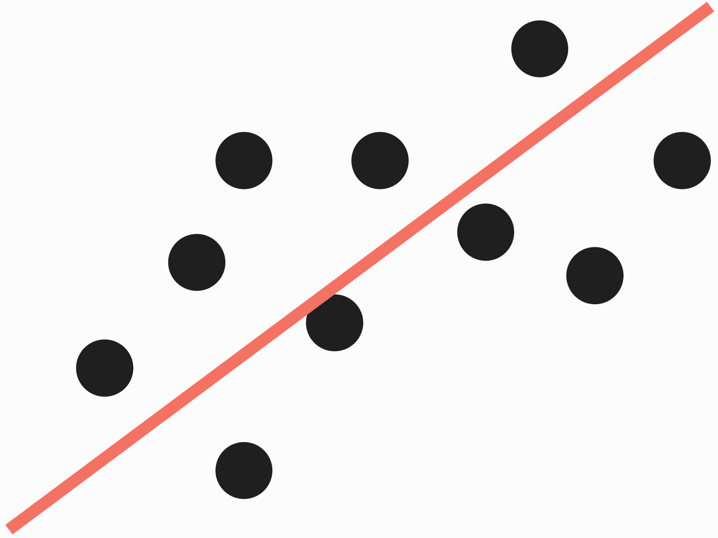
```
        x   y grupo  lugar
  1:    2   7     b Mexico
  2:    2  13     b Mexico
  3:    4  22     a España
  4:    4  25     b Mexico
  5:    5  14     a Mexico
 ---
296:   96 578     b Mexico
297:   98 295     a Mexico
298:   98 598     b España
299:   99 297     a Mexico
300:  100 604     b Mexico
```



La medida

El primer paso para tener una intuición es observer los datos

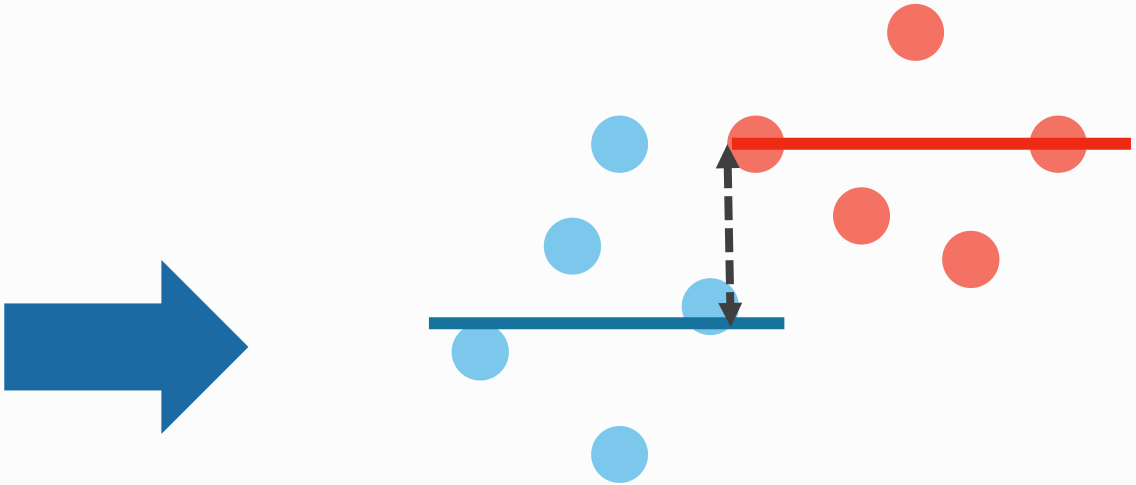
```
       x   y grupo  lugar
  1:   2   7     b Mexico
  2:   2  13     b Mexico
  3:   4  22     a España
  4:   4  25     b Mexico
  5:   5  14     a Mexico
 ---
296:  96 578     b Mexico
297:  98 295     a Mexico
298:  98 598     b España
299:  99 297     a Mexico
300: 100 604     b Mexico
```



La medida

El primer paso para tener una intuición es observer los datos

```
        x   y grupo  lugar
  1:    2   7     b Mexico
  2:    2  13     b Mexico
  3:    4  22     a España
  4:    4  25     b Mexico
  5:    5  14     a Mexico
 ---
296:   96 578     b Mexico
297:   98 295     a Mexico
298:   98 598     b España
299:   99 297     a Mexico
300:  100 604     b Mexico
```



Proyectar una idea

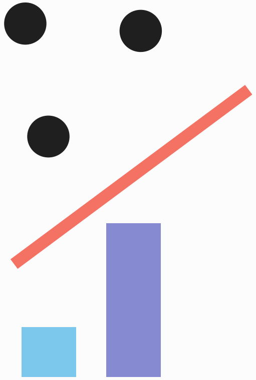
La gramática de los gráficos

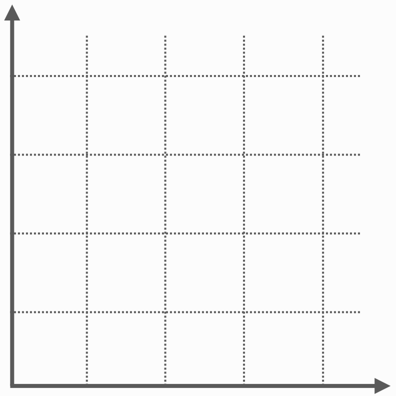
ggplot2

```
         x   y grupo  lugar
  1:     2   7     b Mexico
  2:     2  13     b Mexico
  3:     4  22     a España
  4:     4  25     b Mexico
  5:     5  14     a Mexico
 ---
296:    96 578     b Mexico
297:    98 295     a Mexico
298:    98 598     b España
299:    99 297     a Mexico
300:   100 604     b Mexico
```

Aes(

- X =
- Y =
- col =
- fill =
- size =
- label =
- Linetype =





Datos | Proyección | Geometría | Estética

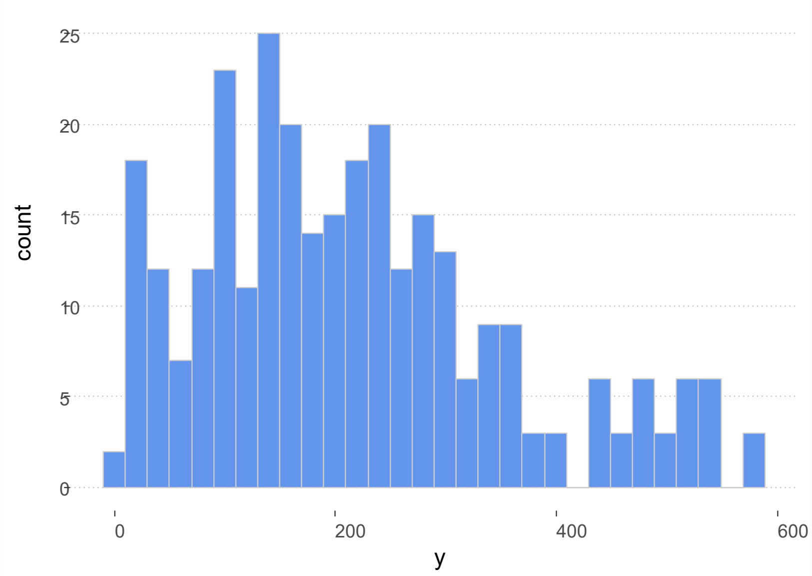
Proyectar una idea

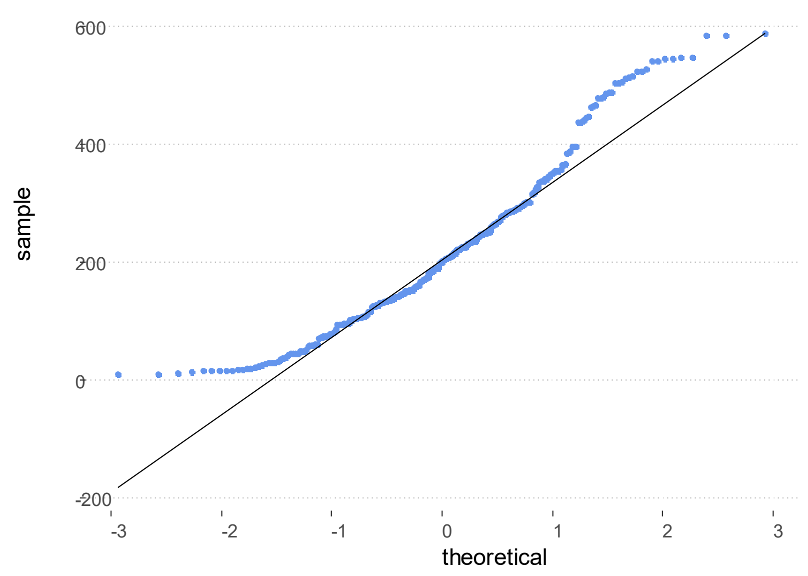
ggplot, la gramática de los gráficos

```
ggplot(data = …,                    La información que queremos representar
       mapping = aes(…)) +          Las coordenadas de representación (x, y…)

       geom_*() +                   La forma (puntos, líneas, polígonos..)

       stat_*() +                   Transformaciones estadísticas

       facet_*()                    Como los datos se dividen en subgrupos
```

Proyectar una idea

Representaciones univariantes



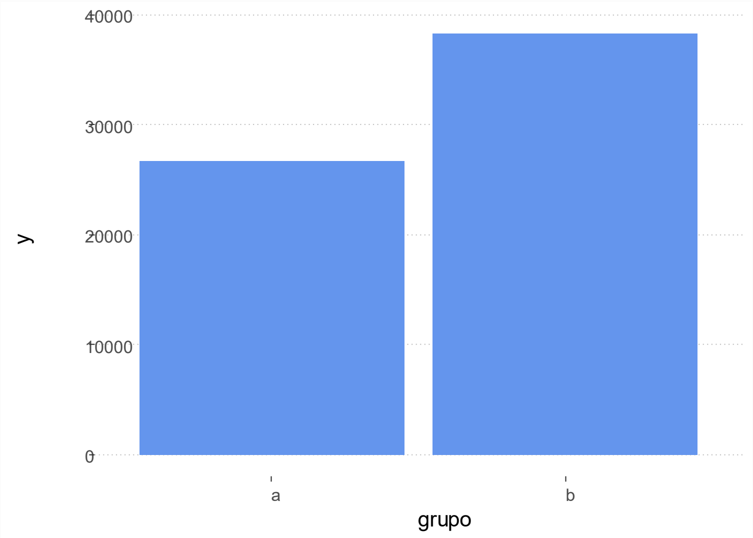


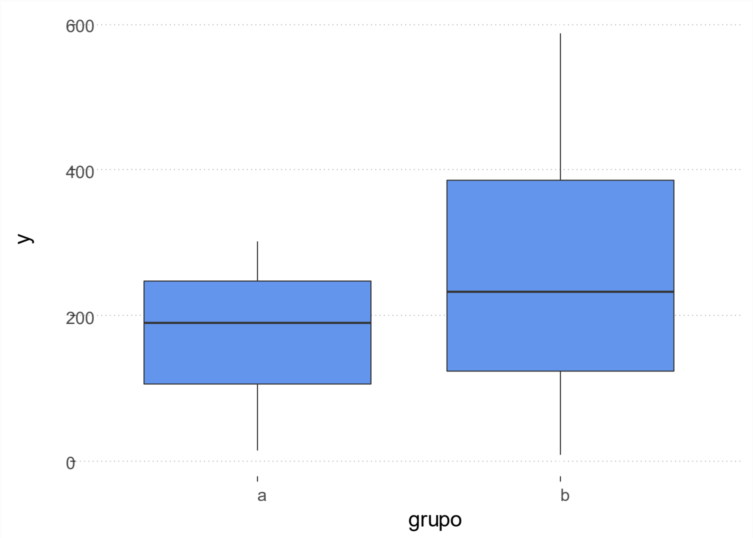

```
ggplot(d, aes(x = y)) +
       geom_histogram(fill = "cornflowerblue",
                      col = "gray80")
```

```
ggplot(d, aes(sample = y)) +
      geom_qq(col = "cornflowerblue") +
      geom_qq_line(distribution = qnorm)
```

Proyectar una idea

Variables categóricas, divisiones por grupos





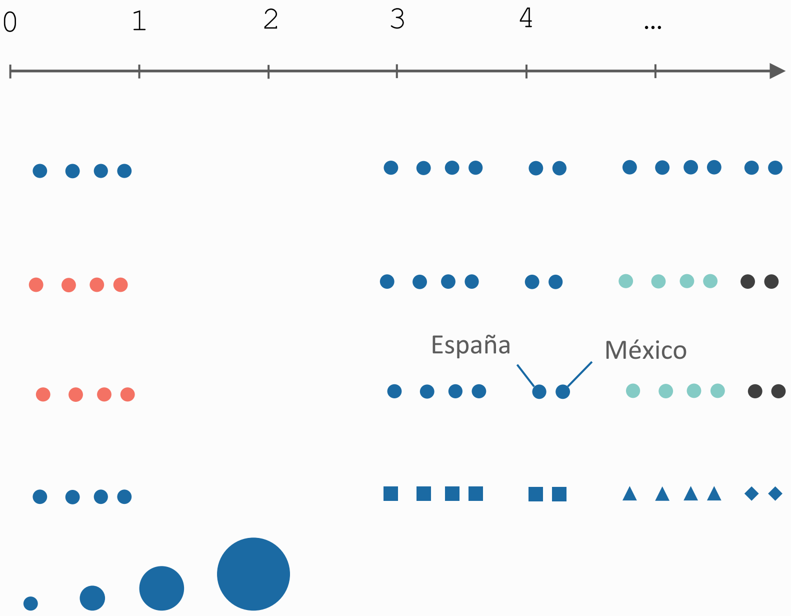

```
ggplot(d, aes(x = grupo, y = y)) +
       geom_bar(stat = "identity",
                fill = "cornflowerblue")
```

```
ggplot(d, aes(x = grupo, y = y)) +
       geom_boxplot(fill = "cornflowerblue")
```

Proyectar una idea

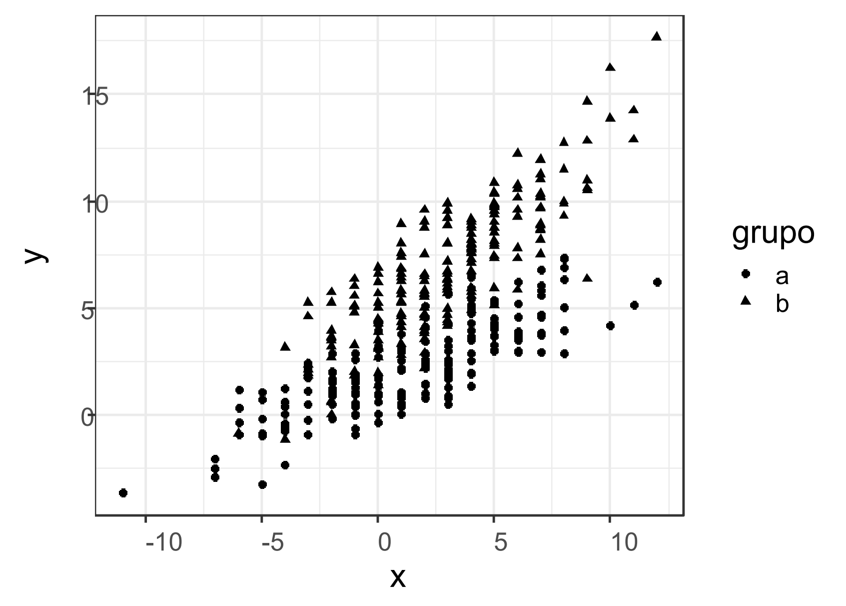
Percepción visual

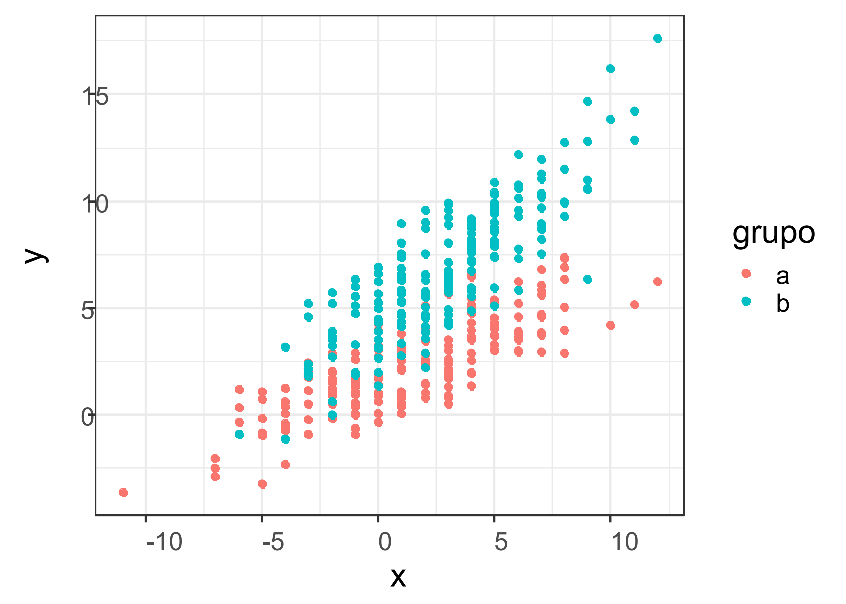
- **Posición**
- **Color**
- **Texto/Etiquetas**
- **Forma**
- **Tamaño/Área**



Proyectar una idea

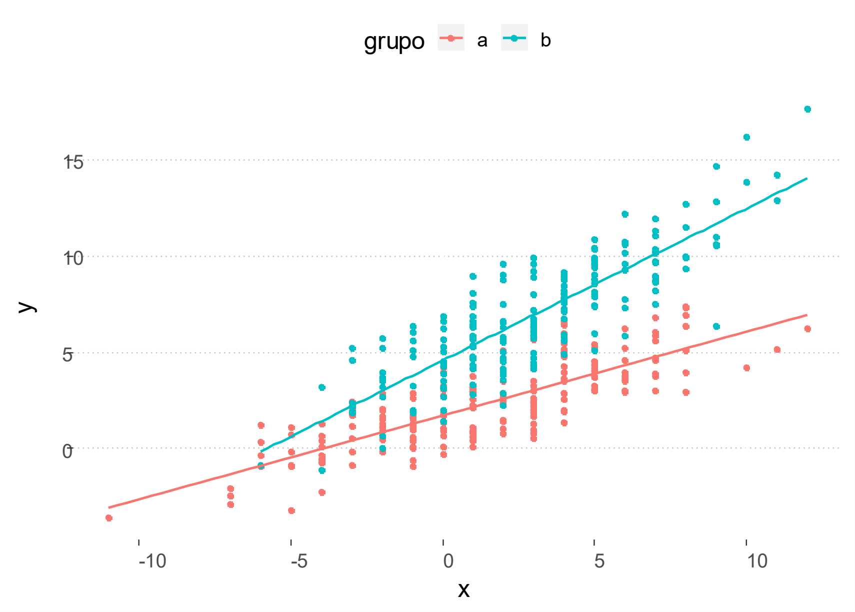
Percepción de contraste





Proyectar una idea

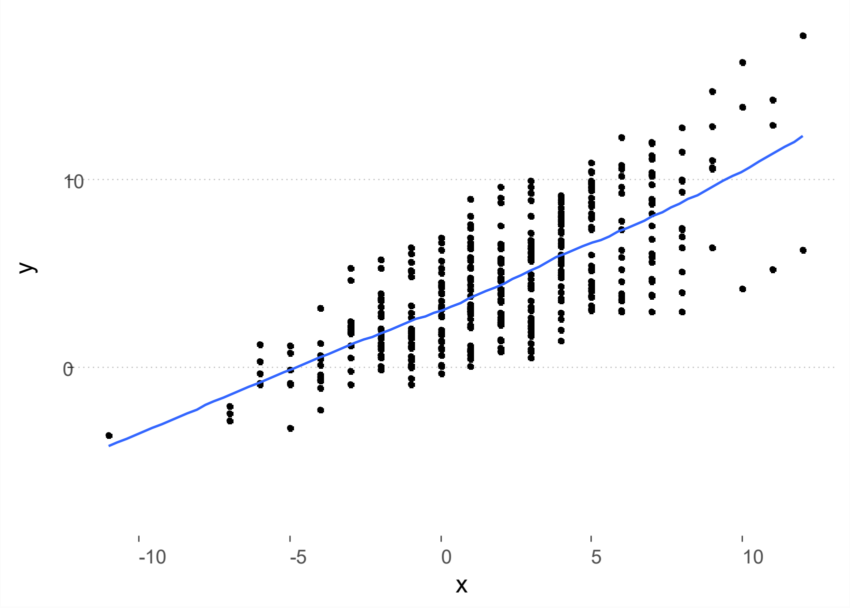
Gráfico cartesiano



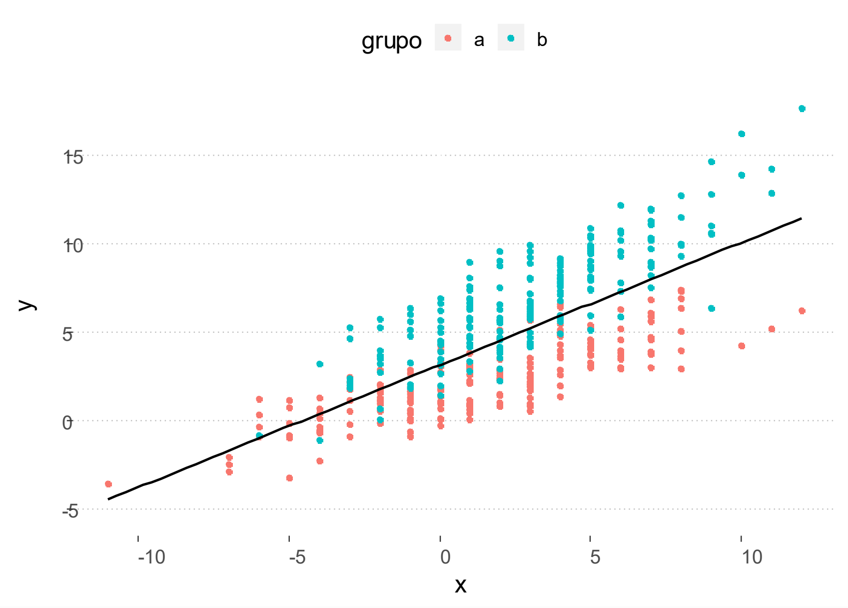

```
ggplot(d,

# Los parametros en aes() representan variables

              aes(x = x, y = y, col = grupo)) +

# Geometria de puntos
       geom_point(

# Los parámetros fuera de aes() quedan fijos

# Tamaño            size = 2,
# Forma             shape = 16,
# Transparencia      alpha = 0.5,
                     show.legend = TRUE) +
 # Regresión simple
             stat_smooth(method = "lm") +

 labs(title = "Mi titulo") +
       theme_bw(base_size = 20)
```

Proyectar una idea

Regresión simple



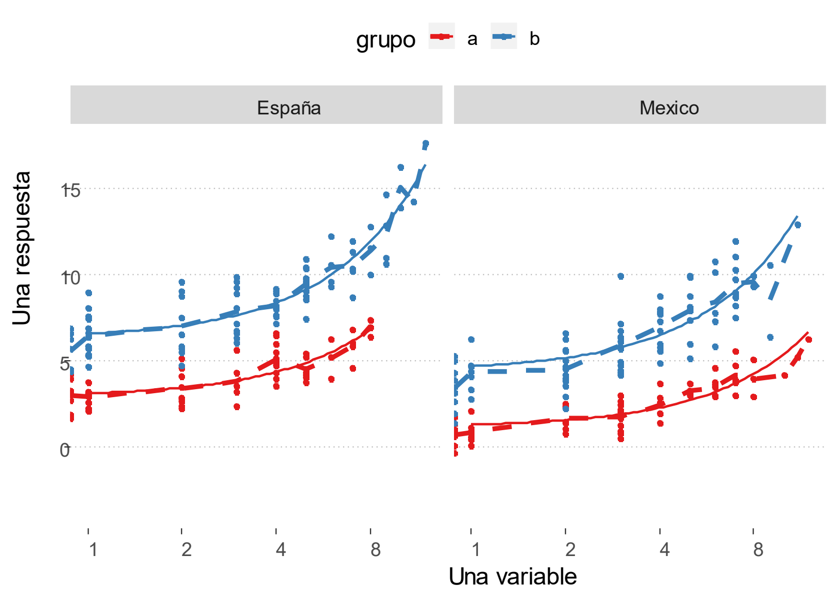

```
ggplot(d, aes(x = x, y = y)) +
      geom_point() +
      stat_smooth()
```




```
ggplot(d, aes(x = x, y = y,) +
            geom_point(aes( col = grupo) ) +
            stat_smooth(method = "lm")
```

Proyectar una idea

Gráfico X-Y + Color + Regresión




```r
ggplot(d, aes(x = x, y = y, col = grupo)) +
# Representar puntos
              geom_point() +
# Añadir lineas
              geom_line(
data = d[, .(meanY = mean(y)), # Nuevos datos
              by = .(x, grupo, lugar)],
              aes(x = x, y = meanY),
              linetype = 2, size = 1.5) +

# Regresión sencilla
              stat_smooth(method = "lm",
              formula = y ~ exp(x)) +
# Dividir datos por lugar
              facet_grid(~ lugar,
                            scales = "free_y") +
# Transformar los ejes
              scale_x_continuous(trans = "log2",
                            limits = c(1, 180)) +
# Nombrar los ejes
labs(x = "Una variable", y = "Una respuesta") +
# Elegir colores
scale_color_brewer(palette = "Set1")
```

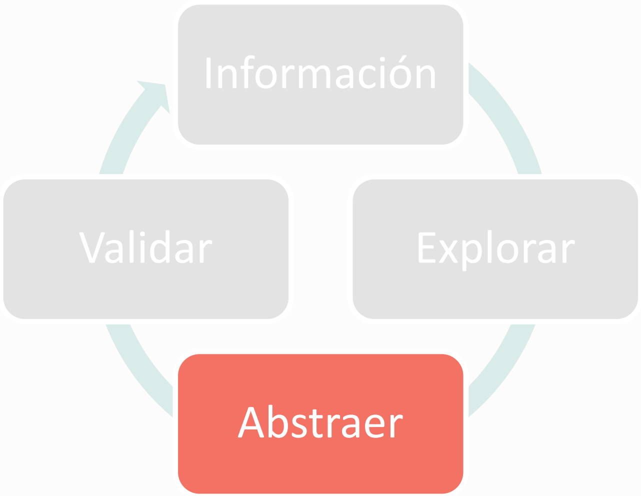
A programar

Por ejemplo

```
library(data.table)
library(ggplot2)
d <- data.table(iris)

ggplot(d, aes(x = Sepal.Length, y = Petal.Width, col = Species)) +
            geom_point() +
            labs(x = "Longitud del Sépalo", y = "Longitud del Sépalo") +
            theme_bw()
```

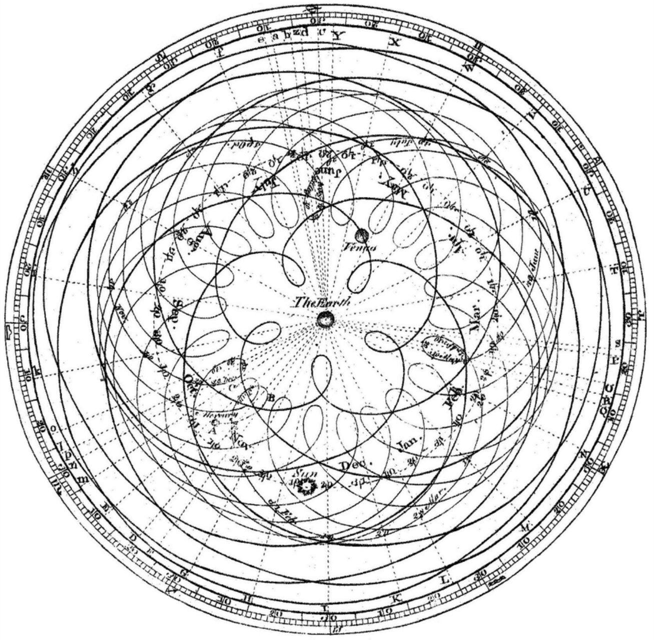
El ciclo investigador

Una vez explorados los datos, es el momento de abstraer, de ignorar las distracciones

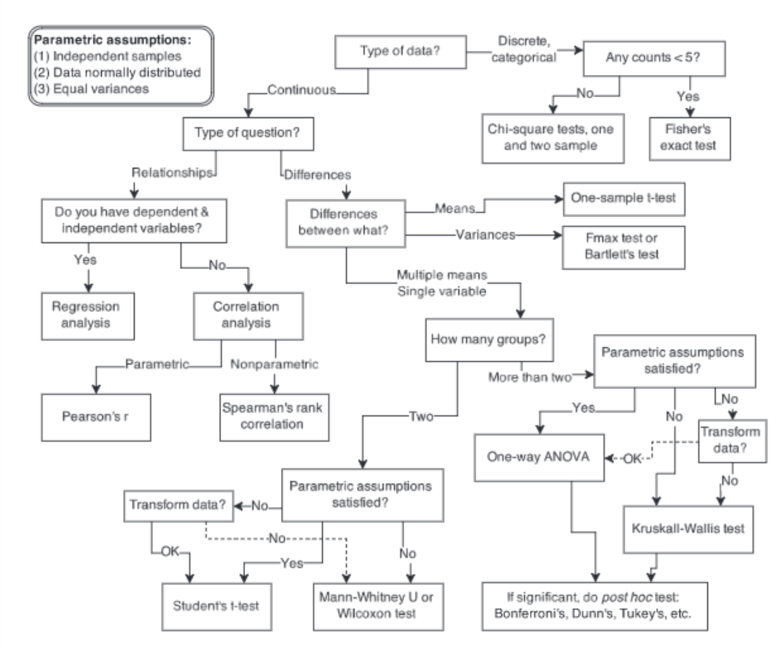


El método científico

Modelar el mundo



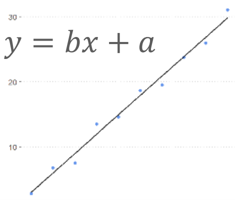
Modelo Ptolemaico del cielo con la tierra en el centro. Jean Dominique Cassini.



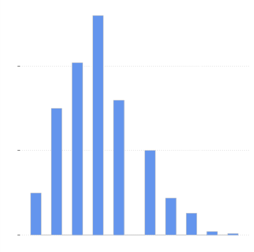
Mapa de los horrores estadísticos. R. McElreath, Rethinking Statistics.

Construyendo modelos

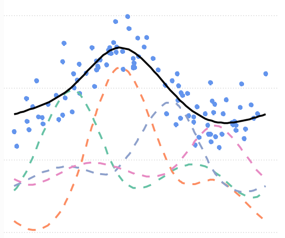
La escuela de modelos **lineares**



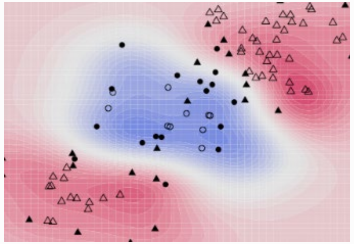
LM



GLM



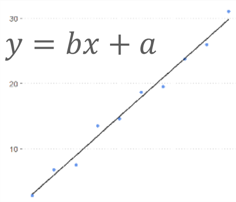
GAM



ML

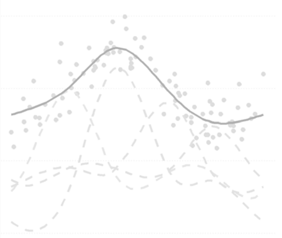
El modelo

Modelos lineares

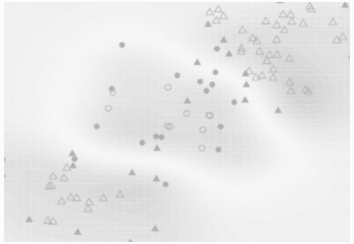


LM

GLM



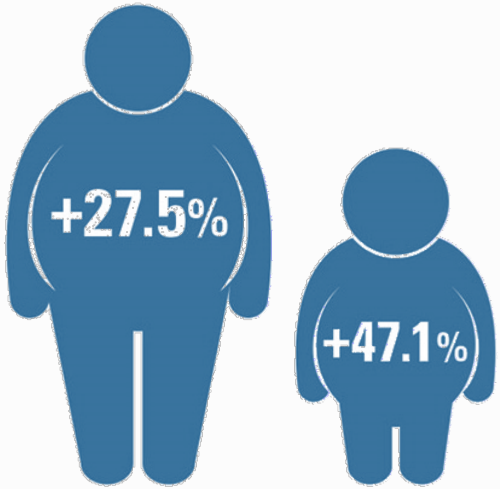
GAM



ML

La pregunta

La obesidad como problema



La pregunta inicia

- Un cuarto de la población adulta y casi la mitad de los niños padece obesidad
- Perdida de calidad de vida
- Probabilidad de **infarto**
- Causa directa de **diabetes**

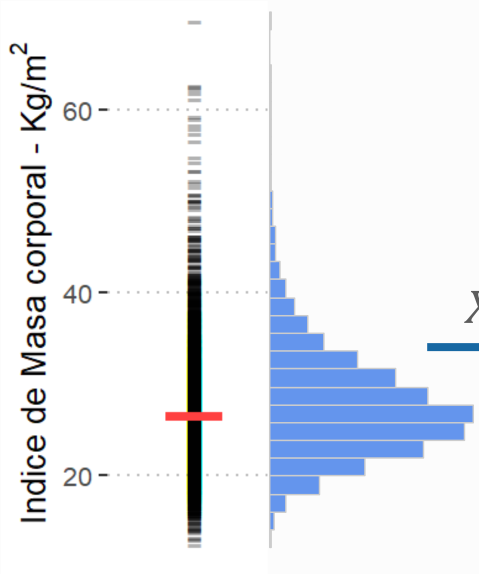
La medida

La base de la inferencia

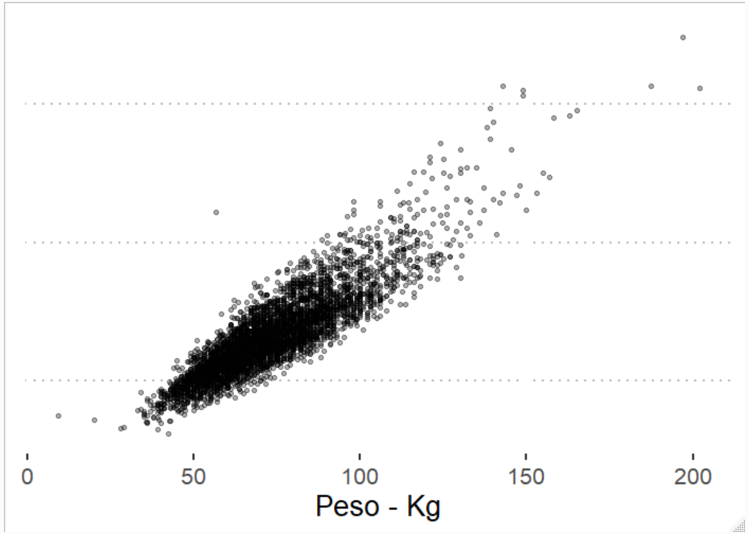
$$\boldsymbol{Y} = N(\mu, \sigma)$$

$$X_i \xrightarrow{f(x)} Y_i$$

$$\boldsymbol{Y} = [y_i]$$



X_i



El modelo

Explicaciones sencillas para relaciones complicadas

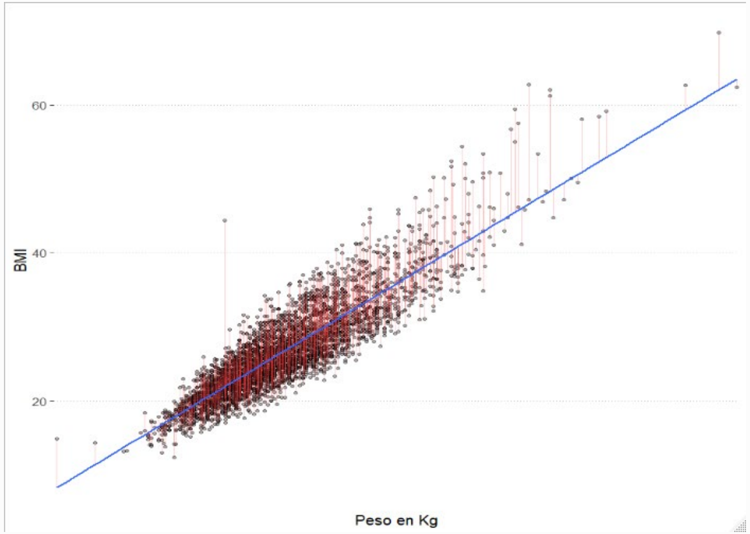
Hipótesis

- El peso está relacionado con la obesidad

$$X_i \xrightarrow{f(x)} Y_i$$

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \beta_1 X_i + \beta_0 + e_{ij} \end{cases}$$

```
library(data.table)

d <- fread("Datasets/brownfat")
m <- lm(BMI ~ Weight, # Formula Y ~ X
        data = d)
```



El modelo

Explicaciones sencillas para relaciones complicadas

Hipótesis

- El peso está relacionado con la obesidad

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \beta_1 X_i + \beta_0 + e_{ij} \end{cases}$$

```
library(data.table)

d <- fread("Datasets/brownfat")
m <- lm(BMI ~ Weight, # Formula Y ~ X
        data = d)

summary(m)

plot(m)
```

```
Family: gaussian
Link function: identity

Formula:
BMI ~ weight

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.677964   0.167303   33.94   <2e-16 ***
weight      0.286155   0.002218  129.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1


Residual standard error: 2.803 on 4840 df
Multiple R-squared:  0.7747,Ajd. R-squared:  0.7747
F-statist: 1.665e+04 on 1 and 4840 DF,p-val: < 2.2e-16
```

P-valor

Data in an uncertain world, perfect knowledge of the uncertainty

Definicion

- Definición estricta:
  - Probabilidad correspondiente al estadístico de ser posible bajo la hipótesis nula. Si cumple con la condición de ser menor al nivel de significancia impuesto arbitrariamente, entonces la hipótesis nula será, eventualmente, rechazada. (valor del estadístico calculado). (Wikipedia, extraído en 2019)

- Definición humana
  - Verificación por contradicción

$$\mathbb{P}(x|H_0)$$

Cual es la probabilidad de que mis datos sean pura casualidad

Supongamos que me equivoco, y no existe un efecto

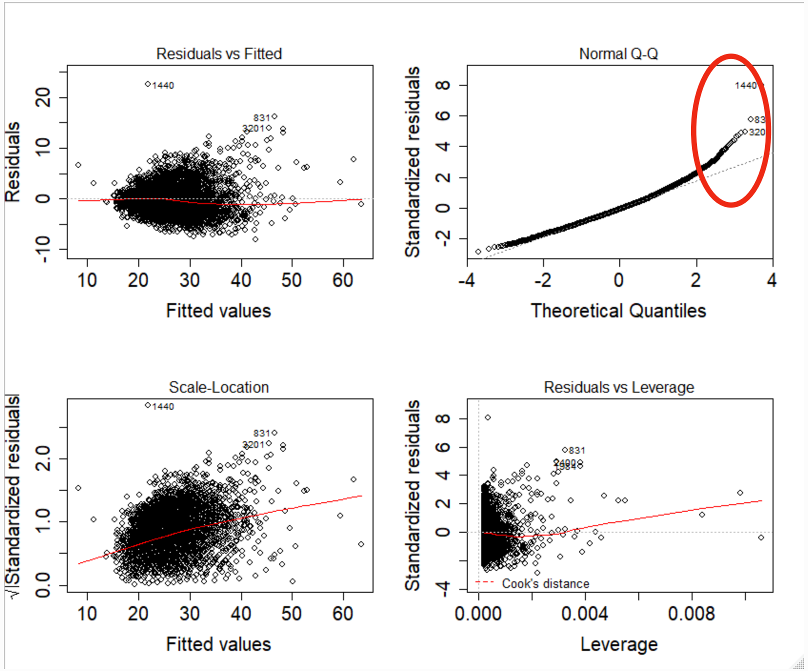
Diagnosis de un modelo

Data in an uncertain world, perfect knowledge of the uncertainty

Valor residual

$$\epsilon_i = (\hat{y} - y_i)$$

- **Valor ajustado vs residuo:** Muestra si existe curvatura en nuestro modelo.
- **Quartiles:** Muestra los residuos del modelo siguen una distribución normal
- **Escala-Localización:** Muestra si la varianza (sigma) es constante
- **Apalancamiento y residuos:** Muestra los puntos con mayor influencia en el modelo



Variables categóricas

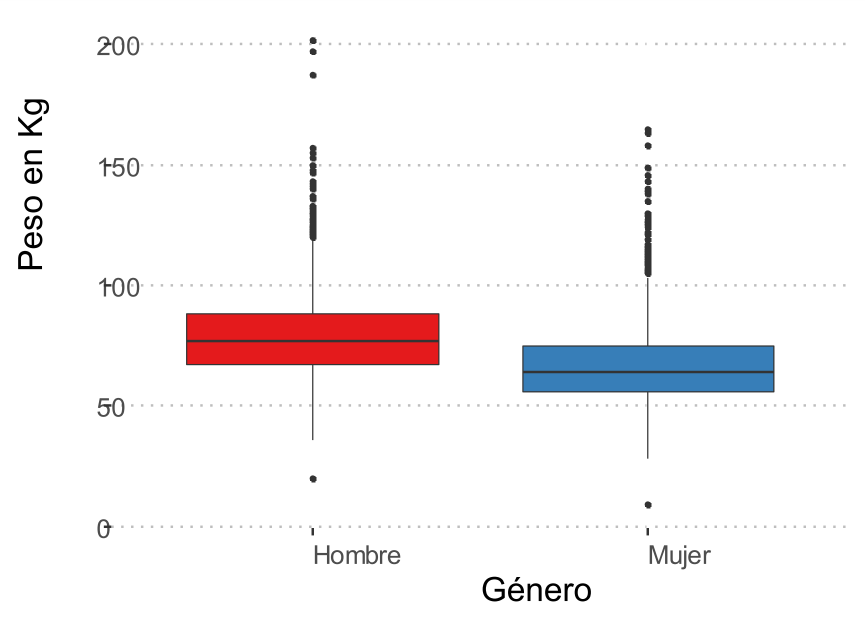
Comparar dos grupos

Hipótesis

- El género está relacionado con el peso

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \boldsymbol{\beta_2 X_{mujer}} + \boldsymbol{\beta_1 X_{hombre}} + e_{ij} \end{cases}$$

```
d[,Sex_c := ifelse(Sex == 1, "Mujer",
                   "Hombre")]
d[, Sex_c := relevel(as.factor(Sex_c),
                     ref = "Hombre")]

m1 <- lm(Weight ~ Sex_c, # Formula Y ~ X
         data = d)

summary(m1)

plot(m1)
```



Variables categóricas

Comparar dos grupos

Hipótesis

- El género está relacionado con el peso

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \boldsymbol{\beta_2 X_{mujer}} + \boldsymbol{\beta_1 X_{hombre}} + e_{ij} \end{cases}$$

```
d[,Sex_c := ifelse(Sex == 1, "Mujer",
                        "Hombre")]
d[, Sex_c := relevel(as.factor(Sex_c),
                        ref = "Hombre")]

m1 <- lm(Weight ~ Sex_c, # Formula Y ~ X
         data = d)

summary(m1)

plot(m1)
```

```
Call:
lm(formula = Weight ~ Sex_c, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-59.065 -11.226  -2.109   8.866 122.935

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.0653     0.3450  229.21   <2e-16 ***
Sex_cMujer  -11.9558     0.4931  -24.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 17.15 on 4840 degrees of freedom
Multiple R-squared:  0.1083,Adjusted R-squared:  0.1081
F-statistic:   588 on 1 and 4840 DF,  p-value: < 2.2e-16
```

A programar

Por ejemplo

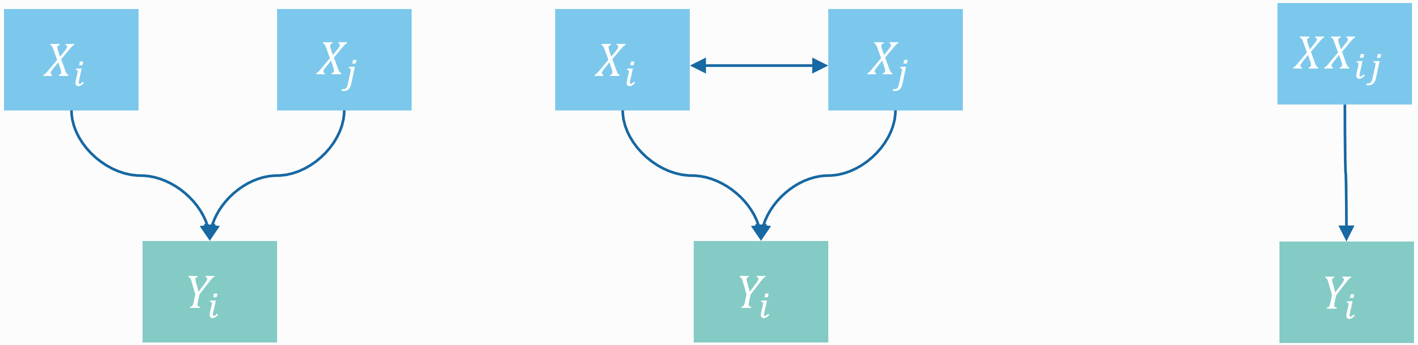
```
d <- fread("Datasets/BrownFat_2011.csv")

m <- lm(BMI ~ Weigth, # Formula Y ~ X
        data = d, )

summary(m)

plot(m)
```

Regresión múltiple

La familia crece

Hipótesis

Sospechamos que ha más parametros (*Xi, Xj)* que condicionan la variabilidad de *Y*



Añadiendo variables

La familia crece

Hipótesis

Los parámetros *Xi e Xj* controla la variabilidad de la variable *Y*

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \beta_2 X_j + \beta_1 X_i + \beta_0 + e_{ij} \end{cases}$$

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \beta_3 X_i X_j + \beta_2 X_j + \beta_1 X_i + \beta_0 + e_{ij} \end{cases}$$

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \beta_2 X_i X_j + \beta_0 + e_{ij} \end{cases}$$

```
update()                    # La función update nos permite actualizar un modelo ya inicializado!
m2 <- update(m, ~ . + Age)             # Dos parametros
m3.1 <- update(m, ~ Weigth * Age)      # Dos parámetros con interacción
m3.2 <- update(m, ~ Weigth:Age)        # Dos parámetros con interacción
```

Añadiendo variables

Entendiendo las interacciones

```
summary(m2)

>

Formula:
BMI ~ Weight + Age

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.447050   0.260057   13.26   <2e-16 ***
weight      0.288767   0.002203  131.08   <2e-16 ***
age         0.032805   0.002953   11.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 2.768 on 4839 degrees of freedom
Multiple R-squared:  0.7803,            Adjusted R-
squared:  0.7802
F-statistic:  8595 on 2 and 4839 DF,  p-value: < 2.2e-16
```

```
summary(m3.2)


>

Formula:
BMI ~ weight:age

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.369e+01  2.134e-01   64.15   <2e-16 ***
weight:age  2.858e-03  4.508e-05   63.40   <2e-16 ***
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
Residual standard error: 4.364 on 4840 degrees of freedom
Multiple R-squared:  0.4537,            Adjusted R-squared:
0.4536
F-statistic:  4020 on 1 and 4840 DF,  p-value: < 2.2e-16
```

Comparar modelos/divergencia

Regularización y Criterios de información

- Coeficiente de determinación:
  - Solo para modelos normo-lineales
  - No descuenta el número de parámetros

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Añadiendo variables

Entendiendo las interacciones

```
summary(m2)

>

Formula:
BMI ~ Weight + Age

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.447050   0.260057   13.26   <2e-16 ***
weight      0.288767   0.002203  131.08   <2e-16 ***
age         0.032805   0.002953   11.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

Residual standard error: 2.768 on 4839 degrees of freedom
Multiple R-squared:  0.7803,          Adjusted R-
squared:  0.7802
F-statistic:  8595 on 2 and 4839 DF,  p-value: < 2.2e-16
```

```
summary(m3.2)


>

Formula:
BMI ~ weight:age

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.369e+01  2.134e-01   64.15   <2e-16 ***
weight:age  2.858e-03  4.508e-05   63.40   <2e-16 ***
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
Residual standard error: 4.364 on 4840 degrees of freedom
Multiple R-squared:  0.4537,          Adjusted R-squared:
0.4536
F-statistic:  4020 on 1 and 4840 DF,  p-value: < 2.2e-16
```

El problema del sobre-ajuste

Memorizar los datos no es entenderlos

Hipótesis

- ¿Qué otras variables pensáis que influyen en el BMI?
  - ¿Tal vez el género?
  - ¿Tener o no diabetes?
  - ¿Tal vez la altura?
  - ¿La temporada y el día de observación?
- Recordad la navaja de Ockham
  - *Non sunt multiplicanda entia sine necessitate*
  - *Una explicación no debe complicarse sin necesidad*

Willem of Ockham, Iglesia de Surrey

El problema del sobre-ajuste

Memorizar los datos no es conocerlos

```
m4 <- lm(BMI ~ Weigth * Age + Height + Sex + Day + Season + Ext_Temp, data = d, )

m5 <- lm(BMI ~ Weigth * Age + Height + Sex, data = d)
```

```
summary(m4)

…
…

Multiple R-squared:0.9846,Adj. R-squared: 0.9846
F-st: 2.8e+04 on 255 and 4586 DF, p-value: < 2.2e-16
```

```
summary(m5)

…
…

Multiple R-squared: 0.9846, Adj R-squared:  0.9846
F-st: 1.3e+05 on 4 and 4837 DF, p-value: < 2.2e-16
```

Comparar modelos/divergencia

Regularización y Criterios de información

- Coeficiente de determinación:
  - Solo para modelos normo-lineales
  - No descuenta el número de parámetros

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Criterio de Información Akaike (AIC por sus siglas en inglés):
  - Probabilidad de los valores medidos respecto al modelo teórico
  - Penaliza modelos complejos

$$\text{AIC} = -\log(\mathbb{P}(\Theta|\boldsymbol{Y}) + \mathrm{k}\tau$$

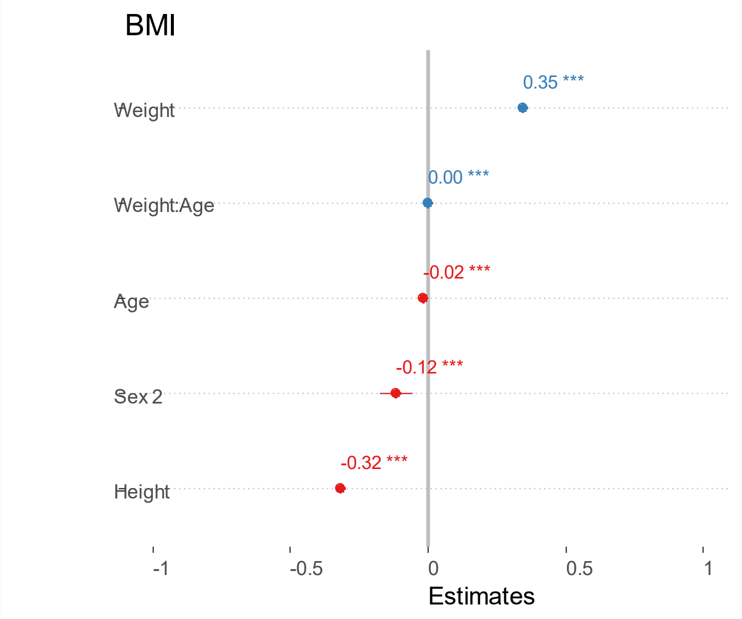
```
AIC(m4, m5, k = log(nrow(d))) %$% .[order(AIC), ]          # Regla no escrita K ~ log(n)

>     df       AIC
m5     7      10720.20
m4    10      10740.47
```

Presentar resultados

Tablas vs imágenes

Coef.	2.50%	97.50%	Estimate
(Intercept)	52.84821	54.08622	53.46721
Weight	0.341727	0.35174	0.346734
Age	-0.02391	-0.01199	-0.01795
Height	-0.31998	-0.31354	-0.31676
Sex2	-0.17377	-0.05999	-0.11688
Weight:Age	0.000225	0.000387	0.000306

```
result <- confint(m5) %>%
            data.table(., keep.rownames = T)
result[, Estimate := coef(m5)]

library(sjPlot)
plot_model(m5, show.values = TRUE, sort.est =
TRUE, value.offset = .3)
```



Otras funciones importantes

Nunca hay timpo para hablar de todo

```
summary(data)                              # Informe sumario de la tabla
cor(x, y)                                  # Correlación entre dos variables

GGally::ggpairs(data)                      # Gráfica de pares para todas las variables
GGally::ggcorr(data)                       # Gráfica de correlaciones para todas las variables

model <- lm(y ~ x, data = d)               # Modelo simple
summary(model)                             # Informe sumario del modelo
coef(model)                                # Extraer coeficientes
confint(model)                             # Extraer intervalos de confianza
plot(model)                                # Representar modelo
predict(model, newdata = )                 # Predecir nuevos datos no vistos por el modelo
fitted(model)                              # Estraer valores ajustados
resid(model)                               # Extraer residuos
allEffects(model)                          # Extraer todos los effectos del modelo
prcomp()                                   # Analisis de componentes principal para reducir variables
```

Canales de apoyo

Recursos educativos del sXXI

Dot CSV
https://www.youtube.com/channel/UCy5znSnfMsDwaLlROnZ7Qbg

Seeing Theory
https://seeing-theory.brown.edu/

3Blue1brown
https://www.youtube.com/channel/UCYO_jab_esuFRV4b17AJtAw

Stats of DOOM
https://www.youtube.com/channel/UCMdihazndR0f9XBoSXWqnYg

¡Gracias por vuestro tiempo!

¿Preguntas?