Statistics and predictive models

Santiago Caño Muñiz

All models are wrong, but some are useful

George Box

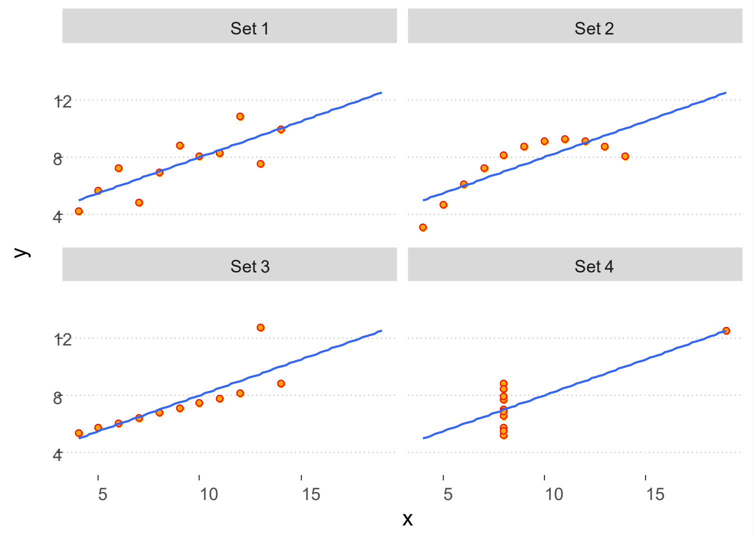
The research cycle

The first step for an intuition is to observe the data

The measure

Descriptive statistics

A		B		C		D	
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

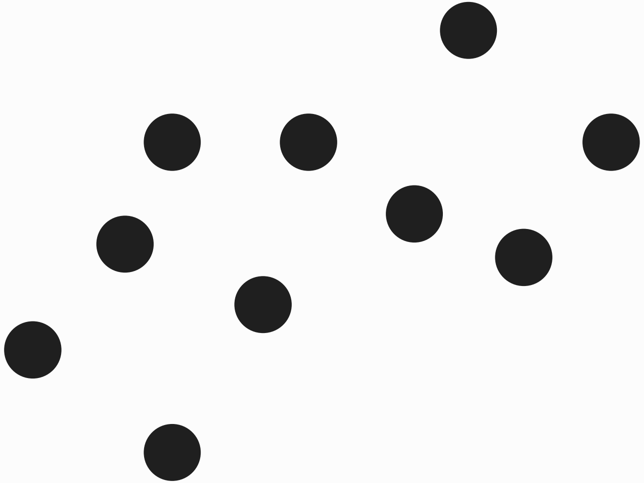


The measure

The first step for an intuition is to observe the data

	cell_type	treatment	size	intake
1:	Gram +	control	7	10
2:	Gram -	control	8	8
3:	Gram +	control	7	10
4:	Gram -	control	5	9
5:	Gram -	control	7	8

396:	Gram +	treat	4	10
397:	Gram -	treat	9	10
398:	Gram +	treat	6	11
399:	Gram -	treat	6	10
400:	Gram +	treat	7	11

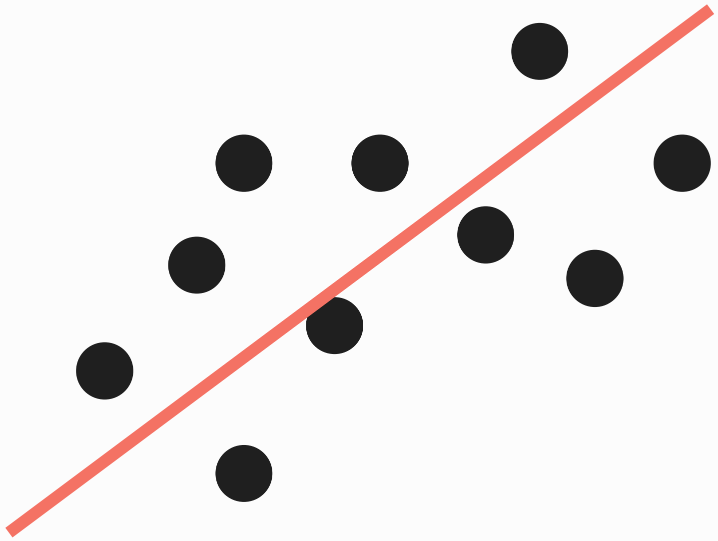


The measure

The first step for an intuition is to observe the data

	cell_type	treatment	size	intake
1:	Gram +	control	7	10
2:	Gram -	control	8	8
3:	Gram +	control	7	10
4:	Gram -	control	5	9
5:	Gram -	control	7	8

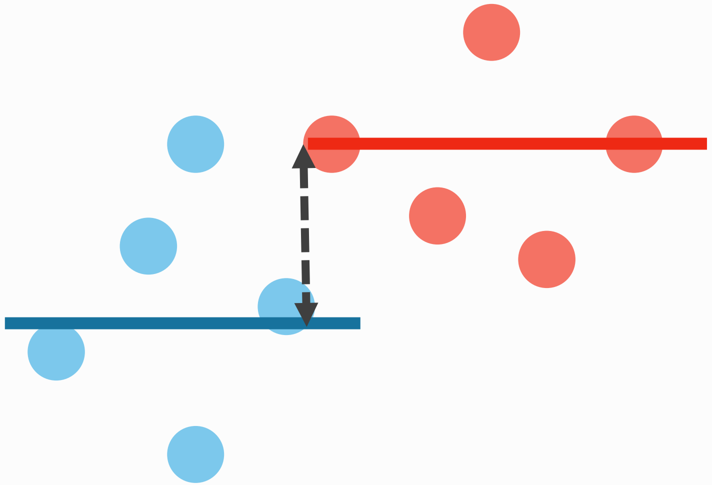
396:	Gram +	treat	4	10
397:	Gram -	treat	9	10
398:	Gram +	treat	6	11
399:	Gram -	treat	6	10
400:	Gram +	treat	7	11



The measure

The first step for an intuition is to observe the data

```
     cell_type treatment size intake
  1:    Gram +    control    7     10
  2:    Gram -    control    8      8
  3:    Gram +    control    7     10
  4:    Gram -    control    5      9
  5:    Gram -    control    7      8
 ---
396:    Gram + treat         4     10
397:    Gram - treat         9     10
398:    Gram + treat         6     11
399:    Gram - treat         6     10
400:    Gram + treat         7     11
```



Projecting an idea

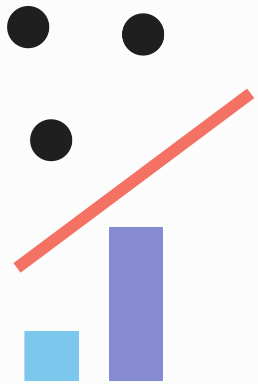
The grammar of graphs

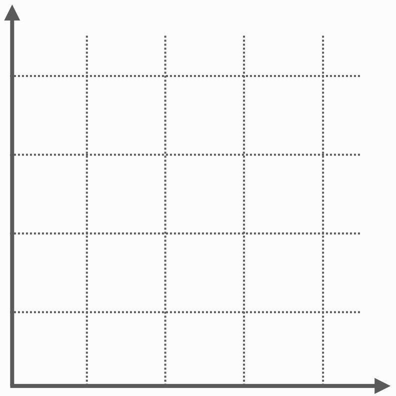
ggplot2

```
    cell_type treatment size intake
  1:   Gram +   control    7     10
  2:   Gram -   control    8      8
  3:   Gram +   control    7     10
  4:   Gram -   control    5      9
  5:   Gram -   control    7      8
 ---
396:   Gram + treat        4     10
397:   Gram - treat        9     10
398:   Gram + treat        6     11
399:   Gram - treat        6     10
400:   Gram + treat        7     11
```

Aes(

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





Data

Projection

Geometry

Aesthetic

Projecting an idea

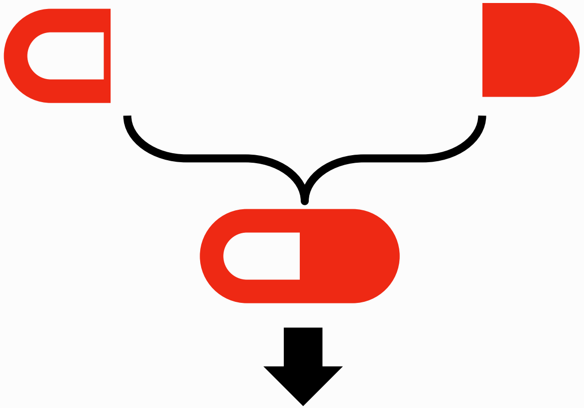
ggplot, the grammar of graphs

```
ggplot(data = ...,
       mapping = aes(...)) +
       geom_*() +
       stat_*() +
       facet_*()
```

- `ggplot(data = ...,` → The information we want to represent
- `mapping = aes(...)) +` → The representation coordinates (x, y...)
- `geom_*() +` → The shape (points, lines, polygons..)
- `stat_*() +` → Statistical transformations
- `facet_*()` → As the data is divided into subgroups

Projecting an idea

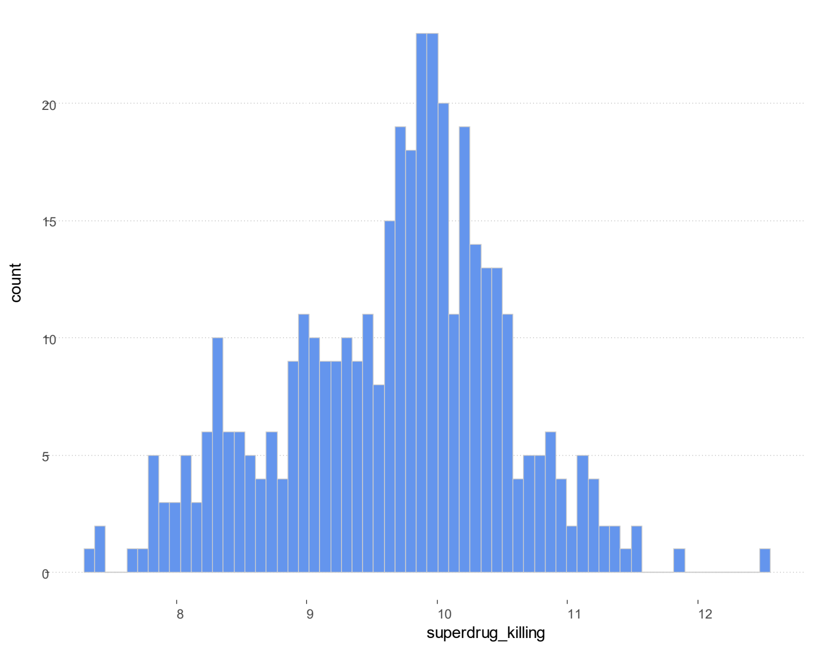
A simple example



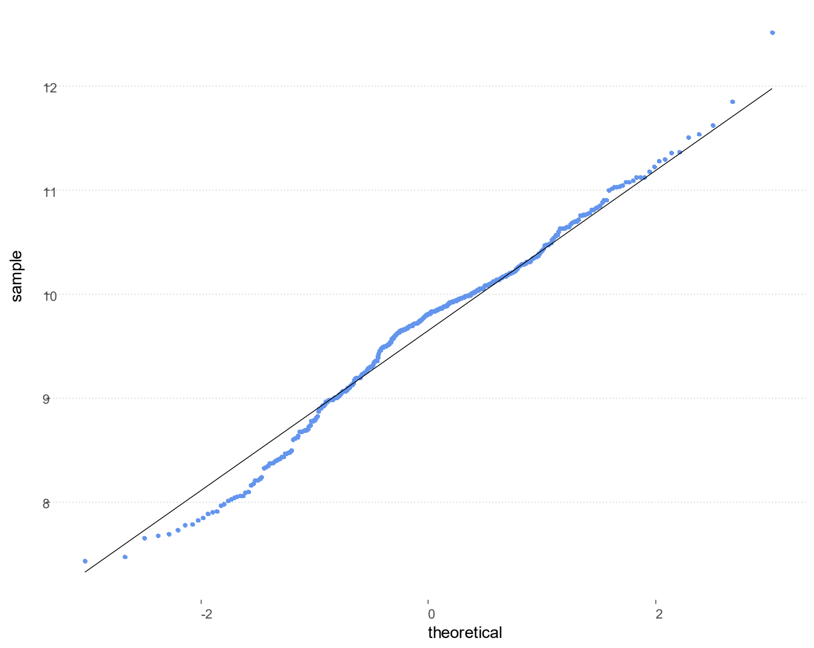
Kill bacteria

Projecting an idea

Univariate representations



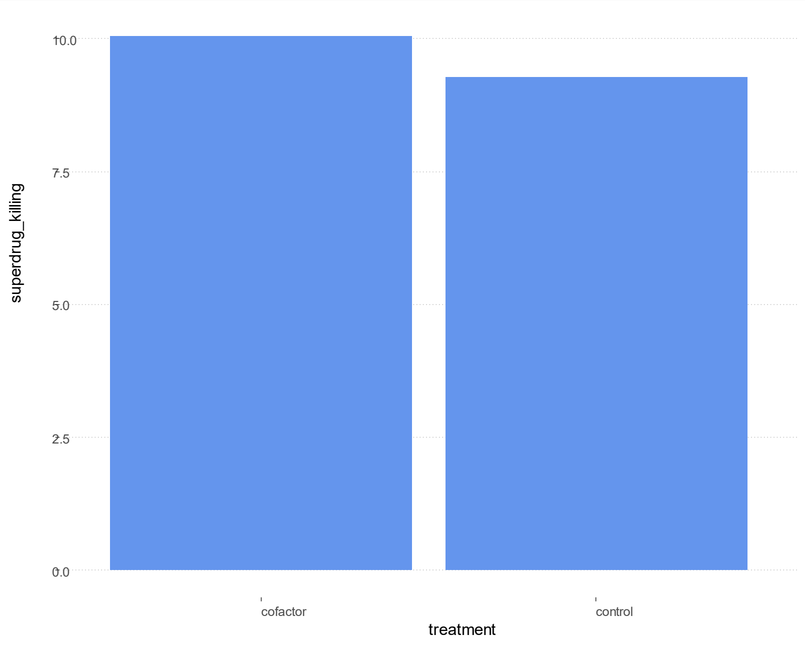

```r
ggplot(d, aes(x = drug_intake), bins = 64) +
       geom_histogram(fill = "cornflowerblue",
                      col = "gray80")
```

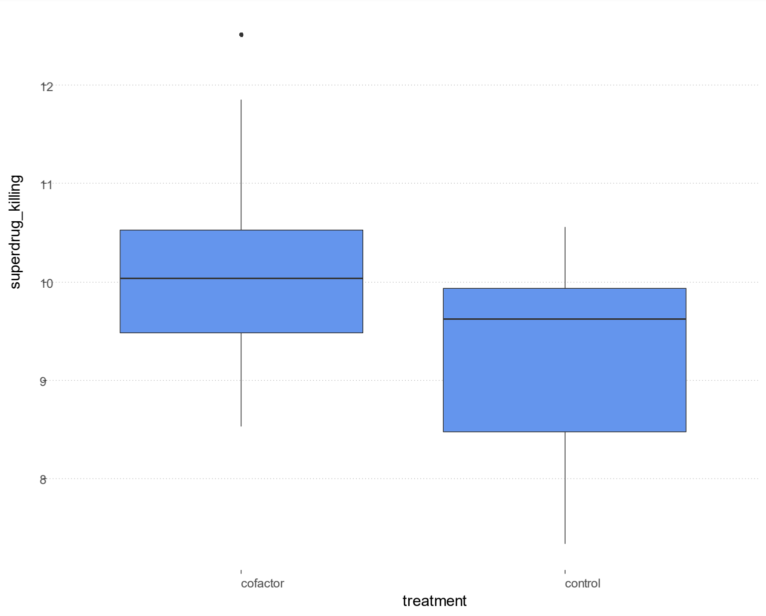



```r
ggplot(d, aes(sample = drug_intake)) +
      geom_qq(col = "cornflowerblue") +
      geom_qq_line(distribution = qnorm)
```

Projecting an idea

Categorical variables, grups divisions



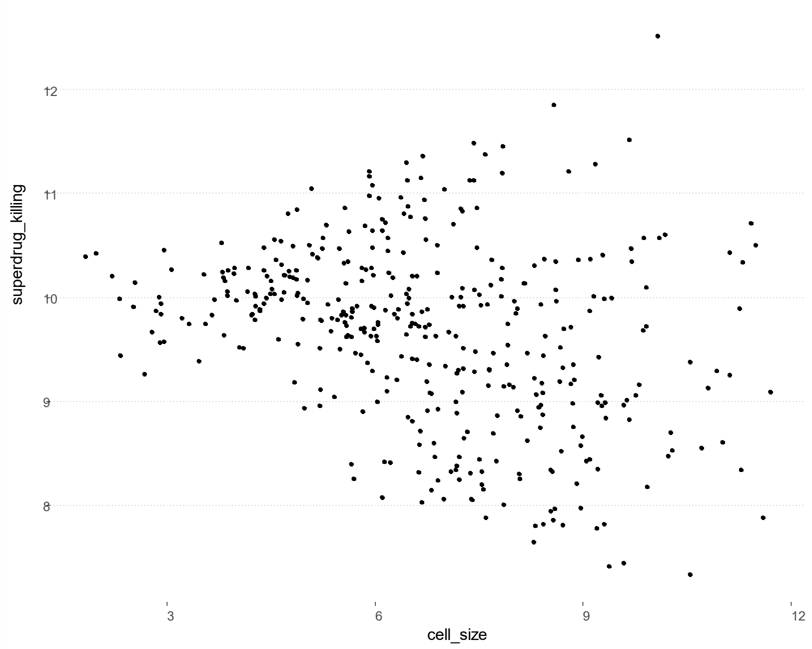



```
ggplot(d, aes(x = grupo, y = y)) +
       geom_bar(stat = "identity",
                fill = "cornflowerblue")
```

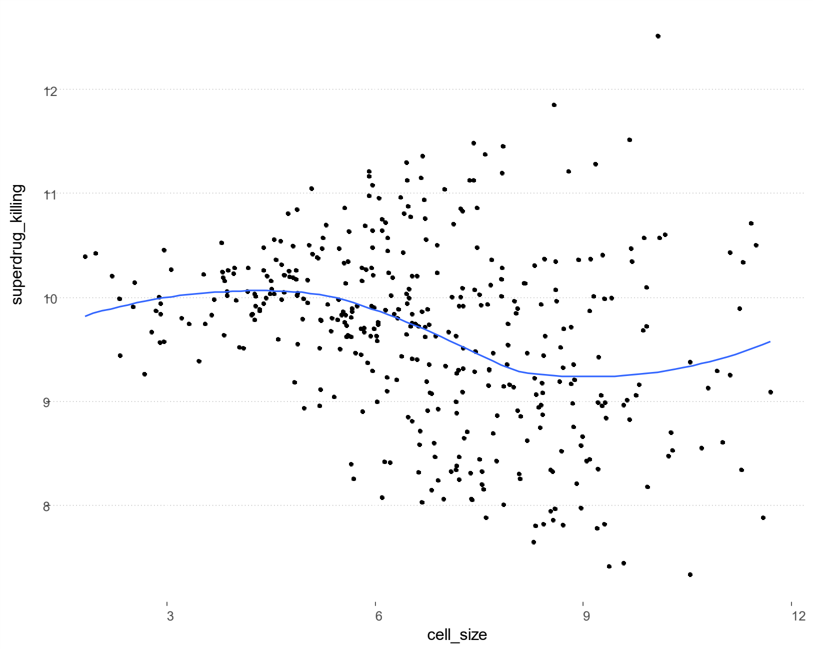
```
ggplot(d, aes(x = grupo, y = y)) +
       geom_boxplot(fill = "cornflowerblue")
```

Projecting an idea

Bivariant relationships



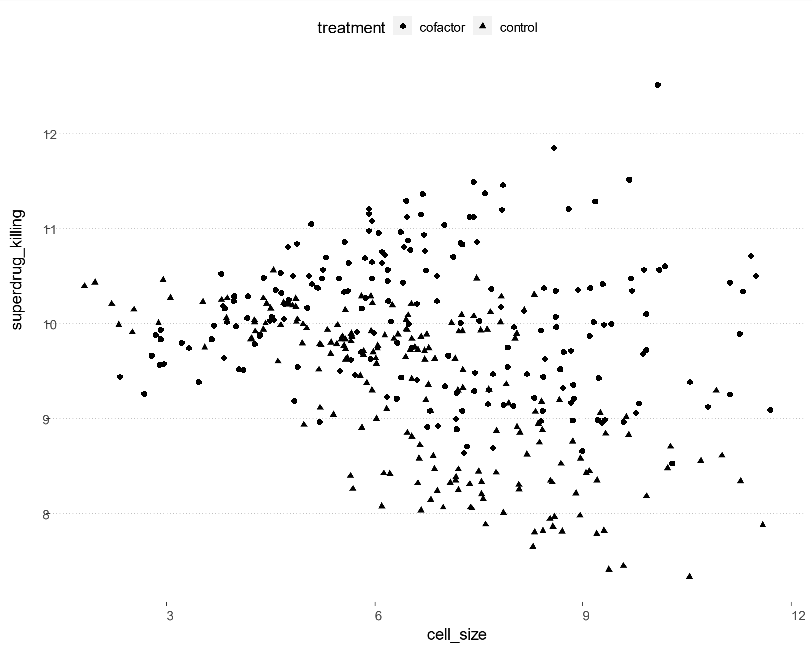

```r
ggplot(d, aes(x = x, y = y)) +
      geom_point()
```




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

Projecting an idea

Bivariant relationships, contrast of perception



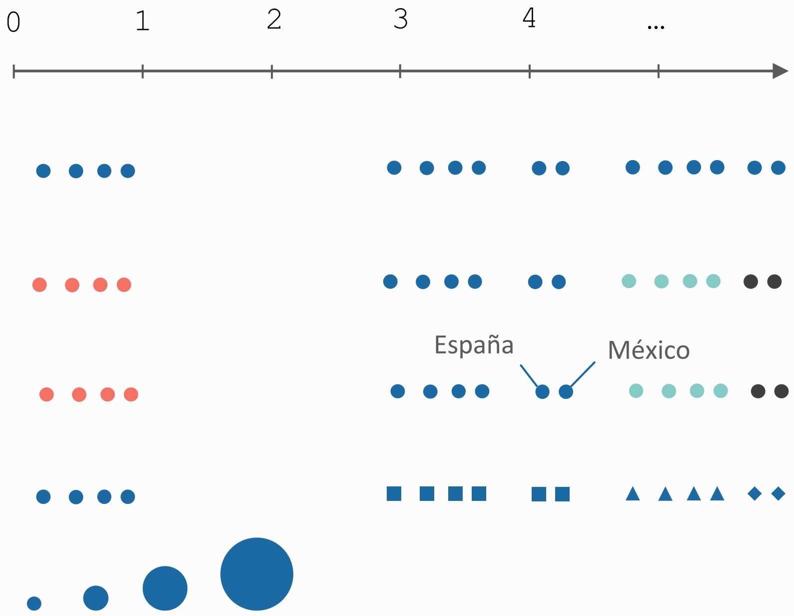

```
ggplot(d, aes(x = cell_size,
          y = superdrug_killing, shape = treatment)) +
  geom_point()
```


```
ggplot(d , aes(x = cell_size,
           y = superdrug_killing, shape = treatment)) +
          geom_point() +
          stat_smooth()
```

Projecting an idea

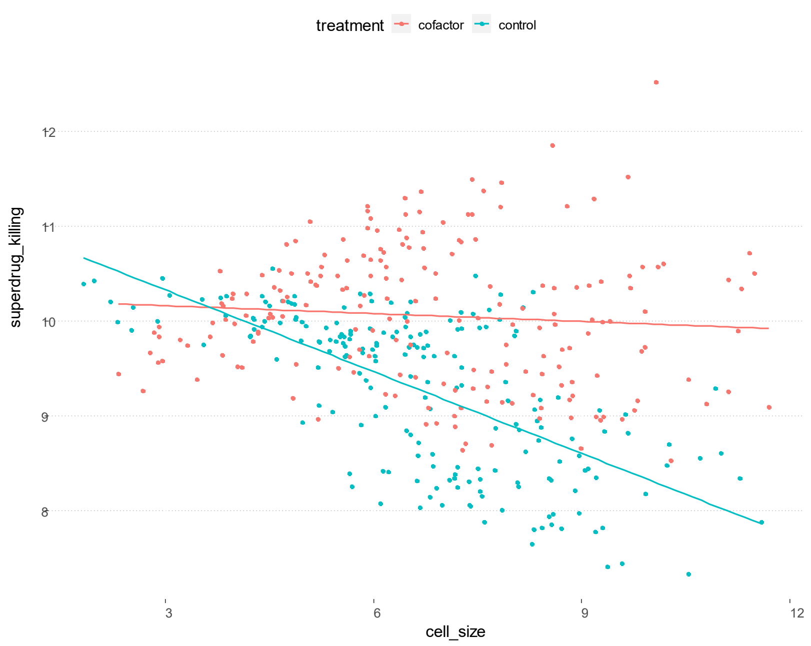
Visual perception

- **Position**
- **Color**
- **Text/Tags**
-
- **Form**
- **Size/Area**
-



Projecting an idea

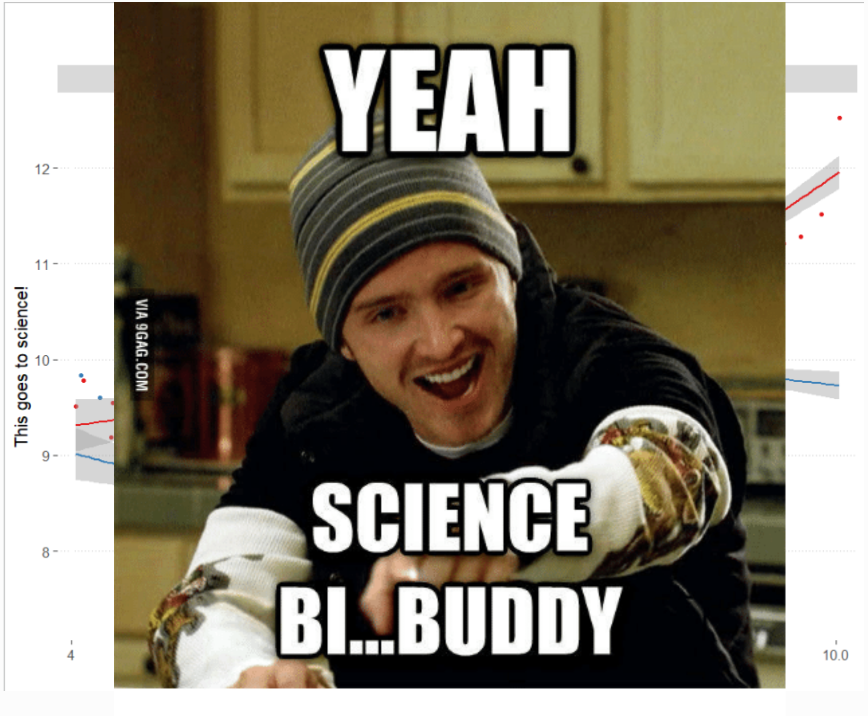
Cartesian graphic




```
ggplot(d,

# The parameters in aes() represent variables

             aes(x = cell_size,
                 y = superdrug_killing,
                 col = treatment)) +

# Point geometry
       geom_point(

# Parameters outside aes() are fixed

# Size               size = 2,
# Form               shape = 16,
# Transparency alpha = 0.5,
                      show.legend = TRUE) +
 # Add simple regression
             stat_smooth(method = "lm")
```

Projecting an idea

X-Y plot with Color and Regression Chart




```
ggplot(d, aes(x = cell_size,
              y = superdrug_killing,
              col = treatment)) +
# Represent points

              geom_point() +

# Simple regression

              stat_smooth(method = "lm",
                          fullrange = TRUE) +

# Divide data by location
              facet_grid(. ~ cell_type,
                         scales = "free_x") +
# Name the axes

labs(x = "Look at that slope",
     y = "This goes to science") +

# Choose colors
scale_color_brewer(palette = "Set1")
```

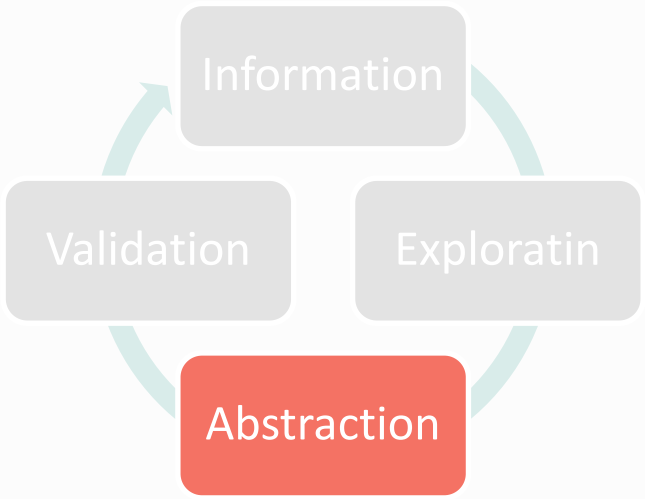
Time to program

For example

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

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

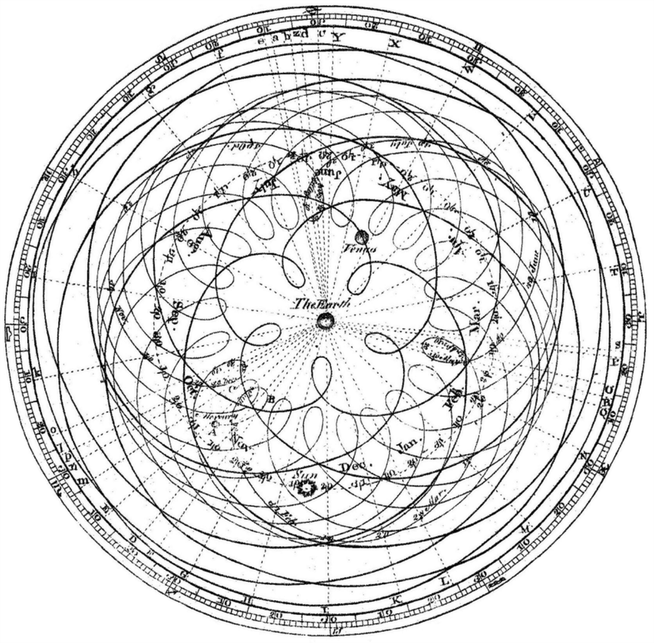
El ciclo investigador

Once the data is explored, it's time to abstract, ignore distractions

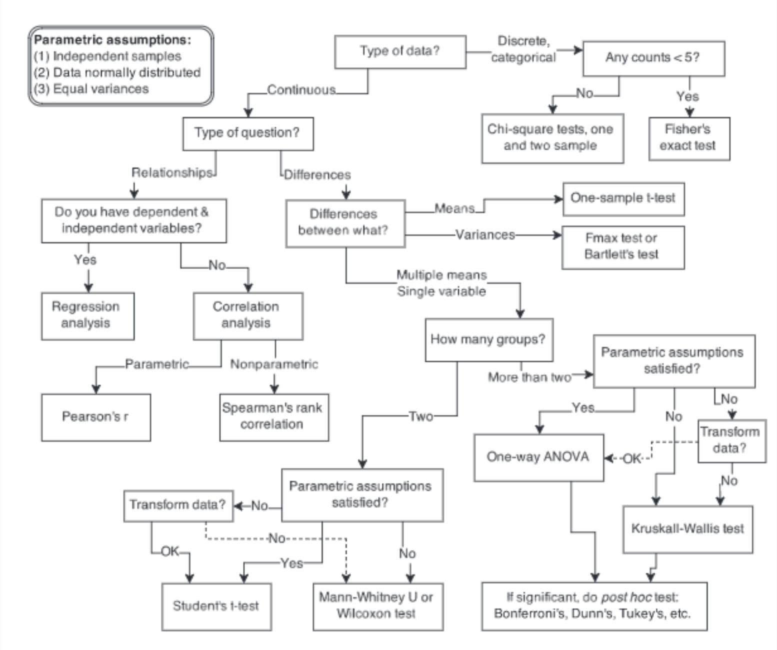


The scientific method

Modeling the world



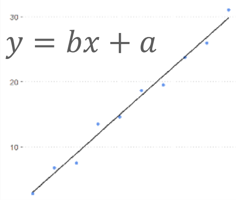
Ptolemaic model of the sky with the earth in the center. Jean Dominique Cassini.



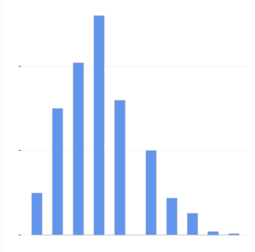
Map of statistical horrors. A. McElreath, Rethinking Statistics.

Building models

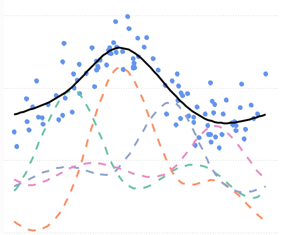
The school of linear models



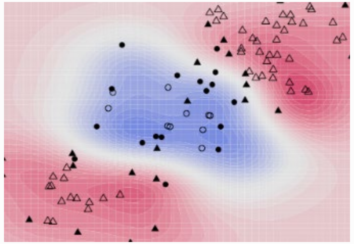
LM



GLM



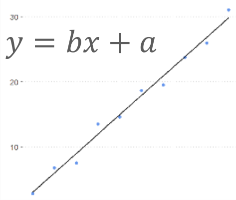
GAM



ML

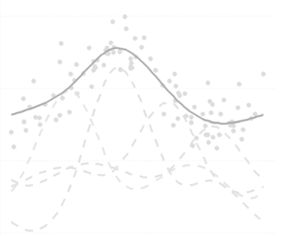
The model

Linear models

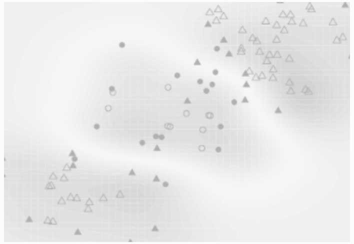


LM

GLM



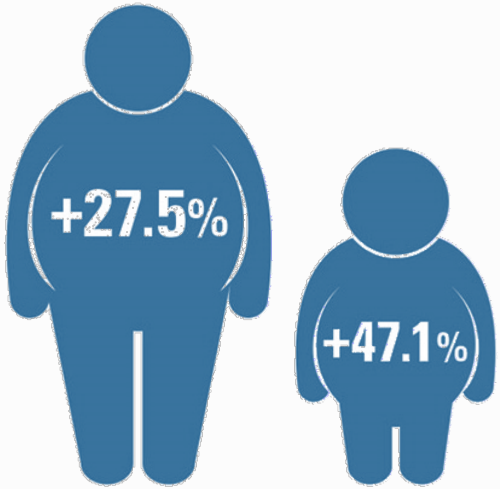
GAM



ML

The question

Obesity as a problem



The why of our research question

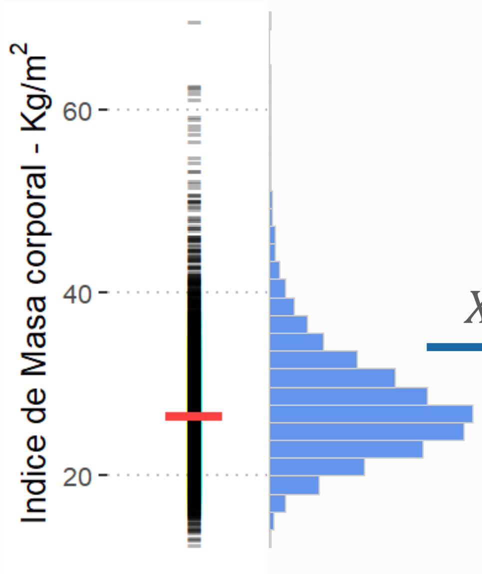
- One-quarter of the adult population and nearly half of children are obese
- Loss of quality of life
- Chance of **heart attack**
- Direct cause of **diabetes**

The measure

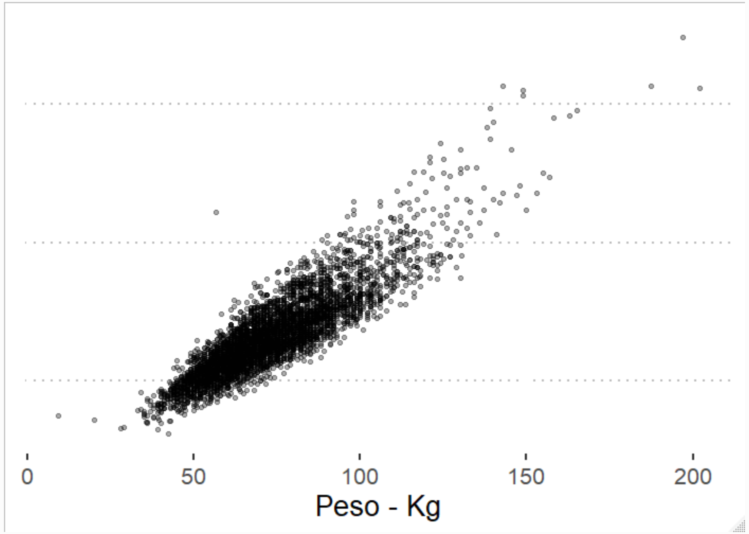
The basis of inference

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

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



$$X_i \longrightarrow$$



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

The model

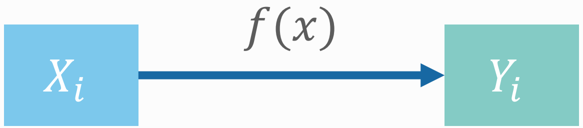
Simple explanations for complicated relationships

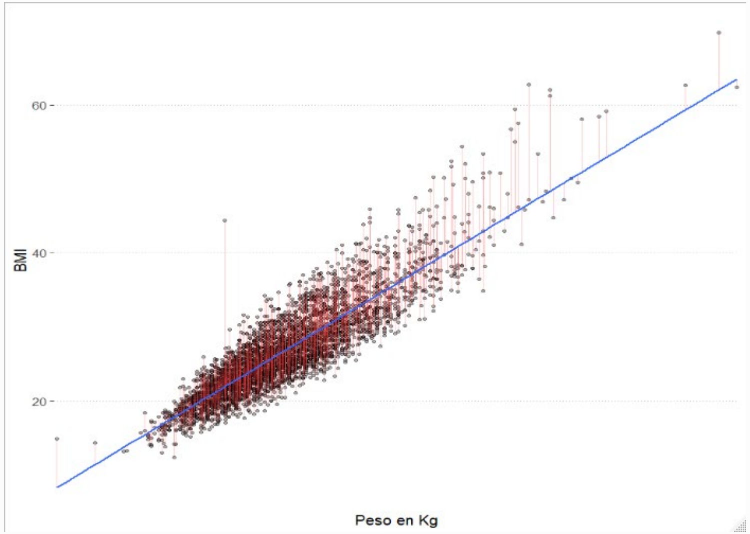
Hypothesis

- Weight is linked to obesity

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

```r
library(data.table)

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





The model

Simple explanations for complicated relationships

Hypothesis

- Weight is linked to obesity

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

```
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-value

Data in an uncertain world, perfect knowledge of the uncertainty

Definition

- *Sensu stricto*:
  - Probability corresponding to the statistic if possible under the null hypothesis. If it meets the condition of being less than the level of significance arbitrarily imposed, then the null hypothesis will eventually be rejected. (value of the calculated statistic). (Wikipedia, extracted in 2019)
  -
- Human definition
  - Proof by contradition

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

Then, what is the probability that my data will be pure chance?

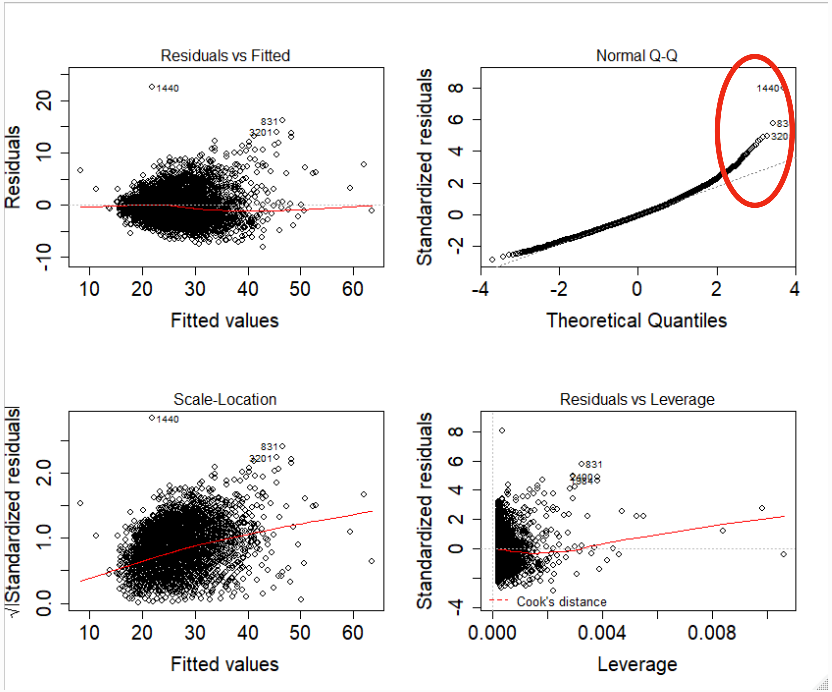
Suppose I'm wrong, and there's no effect

Diagnosis of a model

Data in an uncertain world, perfect knowledge of the uncertainty

Residual value

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



- **Adjusted value vs residue:** Shows if there is curvature in our model.
- **Quartiles:** Shows model waste follows a normal distribution
- **Scale-Location:** Shows if variance (sigma) is constant
- **Leverage and residue:** Shows the points with the greatest influence on the model

Categorical variables

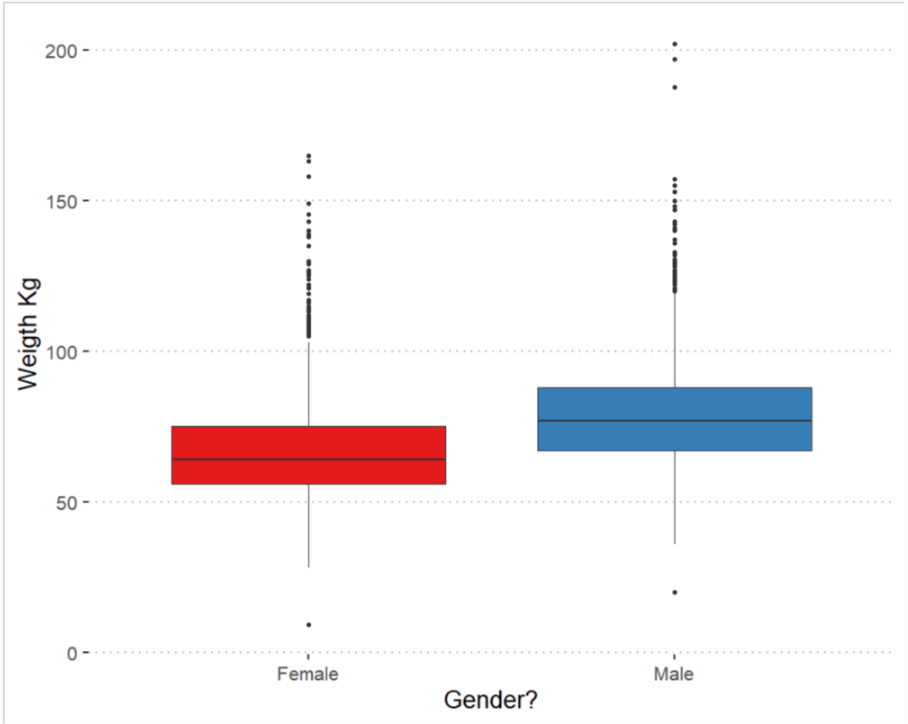
Groups comparison

Hypothesis

- Gender is related to weight

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

```
d[,Sex_c := ifelse(Sex == 1, "Female",
                              "Male")]


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

summary(m1)

plot(m1)
```



Categorical variables

Compare groups

Hypothesis

- Gender is related to weight

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

```
d[,Sex_c := ifelse(Sex == 1, "Female",
                              "Male")]


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)  67.1095     0.3523  190.49   <2e-16 ***
Sex_cMale    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
```

Categorical variables

Compare multiple groups

Hypothesis

- Cancer is related to weight

$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \boldsymbol{\beta_2 X_{C1}} + \boldsymbol{\beta_1 X_{C2}} + \cdots + \boldsymbol{\beta_1 X_{Cn}} + e_{ij} \end{cases}$$

```
cancer_t <- c ("No cancer", "lung", "digestive", "ORL", "breast", "gynaecological (female)",
            "genital (male)", "urothelial", "kidney", "brain", "skin", "thyroid", "prostate",
            "non-Hodgkin lymphoma", "Hodgkin", "kaposi", "Myelona", "leukemia", "other")

d[, Cancer_t_cat := cancer_t[Cancer_Type + 1] %>% as.factor() ]

d[, Cancer_t_cat := relevel(Cancer_t_cat, ref = "No cancer") ]

m1.2 <- lm(Weight ~ Cancer_t_cat, data = d)

summary(m1.2)
```

Categorical variables

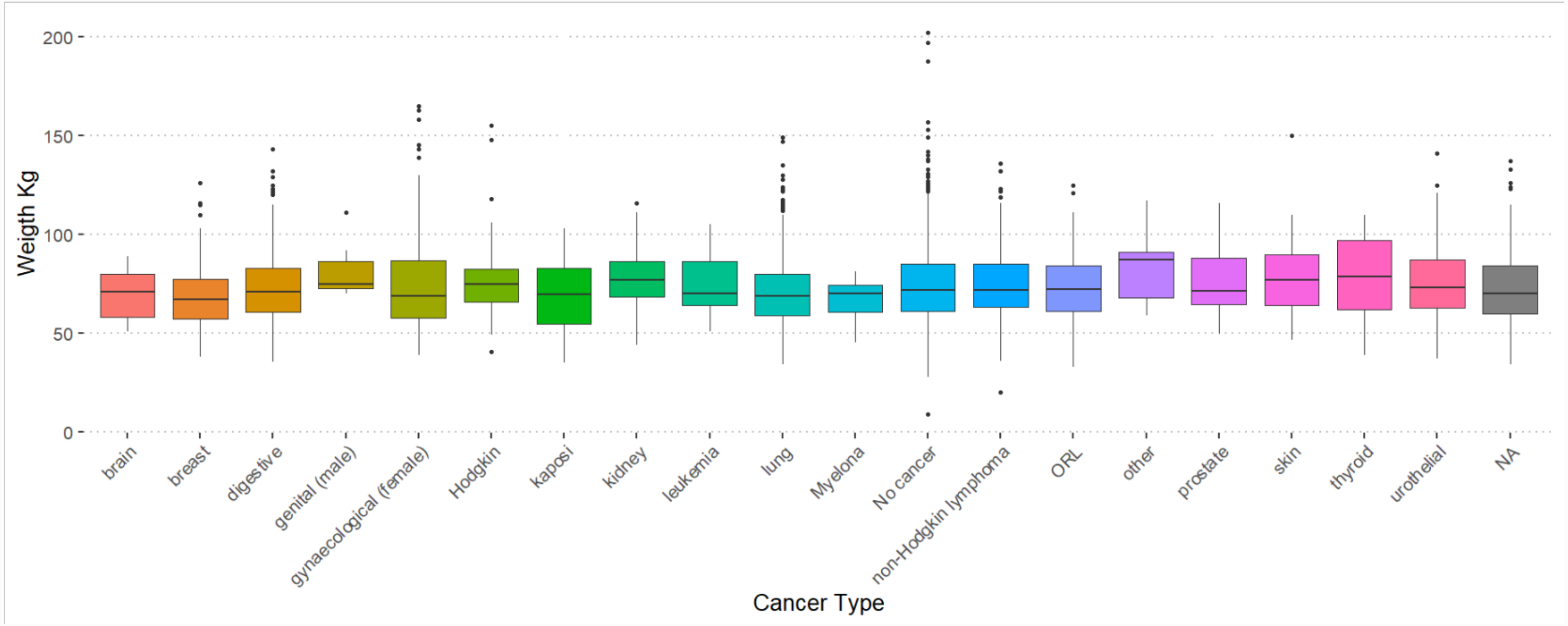
Compare groups

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

Residuals:
    Min      1Q  Median      3Q     Max
-65.464 -12.604  -2.300   9.819 127.396

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          69.1846     5.0183  13.786   <2e-16 ***
Cancer_t_catbreast                   -0.4923     5.2232  -0.094   0.9249
Cancer_t_catdigestive                 3.5625     5.0877   0.700   0.4838
.                                          .          .       .        .
Cancer_t_catprostate                  7.3554     6.4461   1.141   0.2539
Cancer_t_catskin                     10.1125     5.8768   1.721   0.0854 .
Cancer_t_catthyroid                  10.7106     6.3854   1.677   0.0935 .
Cancer_t_caturothelial                6.1154     5.4254   1.127   0.2597
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.09 on 4454 degrees of freedom
  (369 observations deleted due to missingness)
Multiple R-squared:  0.01568,	Adjusted R-squared:  0.0117
F-statistic: 3.942 on 18 and 4454 DF,  p-value: 3.687e-08
```

Categorical variables

Compare groups

Hypothesis



Time to program

For example

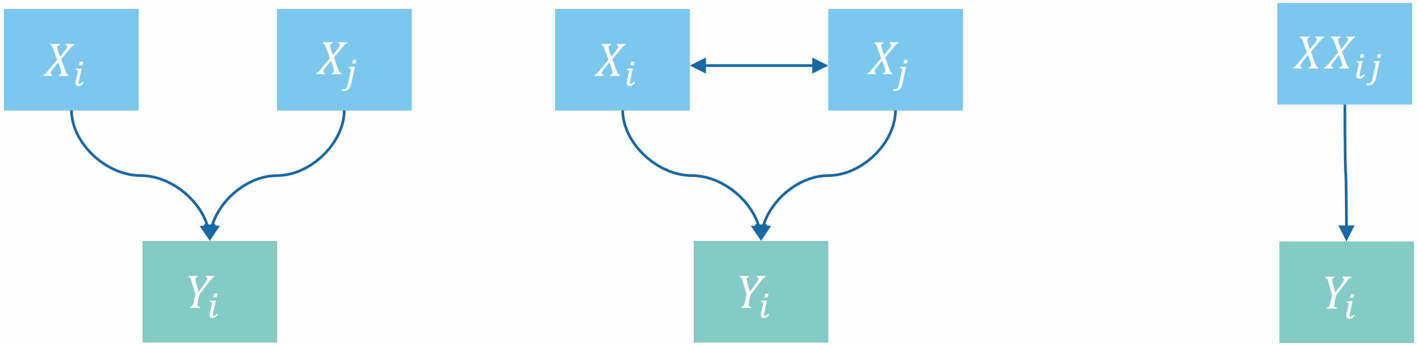
```
d <- fread("Datasets/BrownFat_2011.csv")

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

summary(m)

plot(m)
```

Multiple regression

Family grows

Hypothesis

We suspect there's more parameters (*Xi, Xj*) that might condition the variability of *Y*



Adding variables

Family grows

Hypothesis

The Xi and Xj parameters control the variability of the Y variable

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

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

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

```
update()                    # The update function allows us to update an already initialized model!
m2 <- update(m, ~ . + Sex_c)        # From the parameters
m3.1 <- update(m, ~ Sex_c*Weight)   # Two parameters with interaction
m3.2 <- update(m, ~ Sex_c:Weight)   # Two parameters with interaction
```

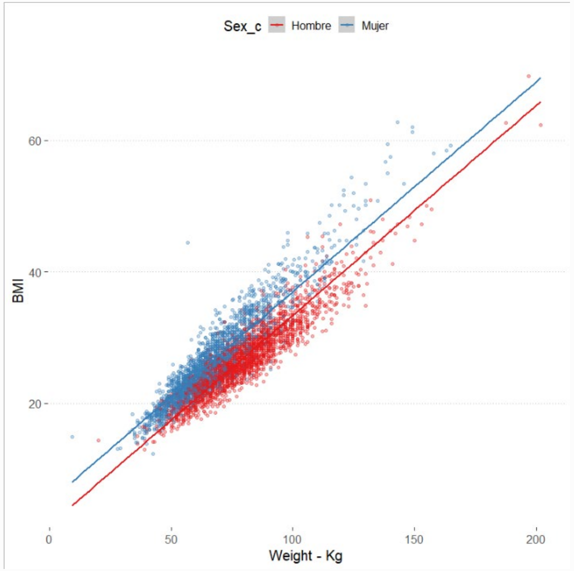
Adding variables

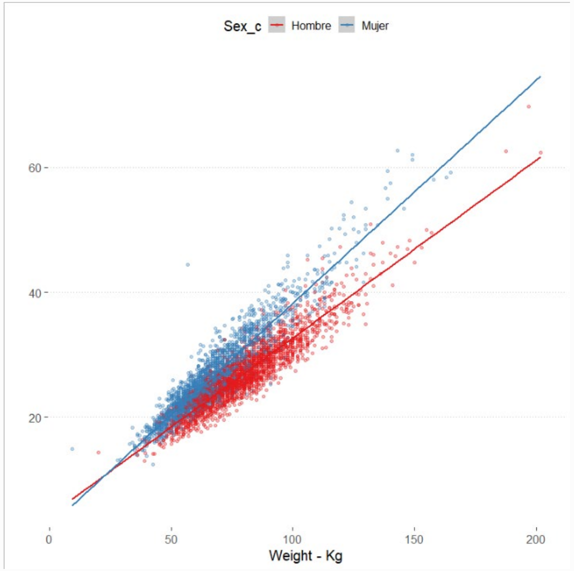
Family grows

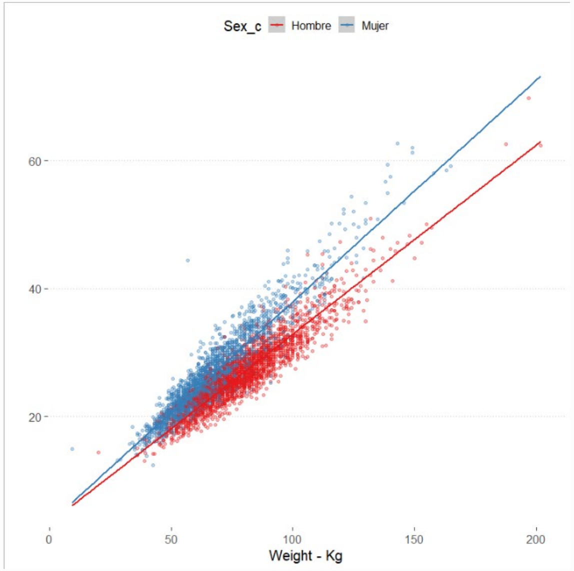
$$\begin{cases} Y \sim N(\mu, \sigma^2) \\ \mu \sim \beta_2 X_j + \beta_1 X_i + \beta_0 + e_{ij} \end{cases}$$

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

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







Adding variables

Understanding interactions

```
summary(m2)

>

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

Residuals:
   Min     1Q Median     3Q    Max 
-8.924 -1.487 -0.087  1.340 21.199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.530807   0.154388   9.915   <2e-16 ***
Weight      0.318749   0.001868 170.592   <2e-16 ***
Sex_cMujer  3.597444   0.067875  53.001   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.229 on 4839 degrees of freedom
Multiple R-squared:  0.8575,   Adjusted R-squared:  0.8574 
F-statistic: 1.456e+04 on 2 and 4839 DF,  p-value: < 2.2e-16
```

```
summary(m3.2)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6288 -1.3829 -0.1024  1.2826 21.4415 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3.319602   0.134588   24.66   <2e-16 ***
Weight:Sex_cHombre 0.295583   0.001709  172.95   <2e-16 ***
Weight:Sex_cMujer  0.346364   0.001991  173.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.15 on 4839 degrees of freedom
Multiple R-squared:  0.8675,   Adjusted R-squared:  0.8674 
F-statistic: 1.584e+04 on 2 and 4839 DF,  p-value: < 2.2e-16
```

Compare models/divergence

Regularization and Information Criteria

- Determination coefficient (R^2):
  - Proportion of variance explained
  - For normo-linear models only
  - Don't discount the number of parameters

Explained variance

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

Sample variance

Adding variables

Understanding interactions

```
summary(m2)

>

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

Residuals:
   Min     1Q Median     3Q    Max 
-8.924 -1.487 -0.087  1.340 21.199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.530807   0.154388   9.915   <2e-16 ***
Weight      0.318749   0.001868 170.592   <2e-16 ***
Sex_cMujer  3.597444   0.067875  53.001   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.229 on 4839 degrees of freedom
Multiple R-squared:  0.8575,	Adjusted R-squared:  0.8574 
F-statistic: 1.456e+04 on 2 and 4839 DF,  p-value: < 2.2e-16
```

```
summary(m3.2)

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

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6288 -1.3829 -0.1024  1.2826 21.4415 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3.319602   0.134588   24.66   <2e-16 ***
Weight:Sex_cHombre 0.295583   0.001709  172.95   <2e-16 ***
Weight:Sex_cMujer  0.346364   0.001991  173.93   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.15 on 4839 degrees of freedom
Multiple R-squared:  0.8675,	Adjusted R-squared:  0.8674 
F-statistic: 1.584e+04 on 2 and 4839 DF,  p-value: < 2.2e-16
```

The problem of over-adjustment

Memorizing data is not understanding it

Hypothesis

- What other variables do you think influence the BMI?
  - Maybe the gender?
  - Diabetes?
  - Height?
  - The season and day of observation?
- Remember Ockham's knife
  - *Non sunt multiplicanda entia sine necessitate*
  - *An explanation should not be complicated without need*

Willem of Ockham, Iglesia de Surrey

The problem of over-fitting

Memorizing data is not knowledge

```
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
```

Compare models/divergence

Regularization and Information Criteria

- Determination coefficient:
  - For normo-linear models only
  - Don't discount the number of parameters

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

- Akaike Information Criterion (AIC):
  - Probability of measured values relative to the theoretical model
  - Penalizes complex models

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

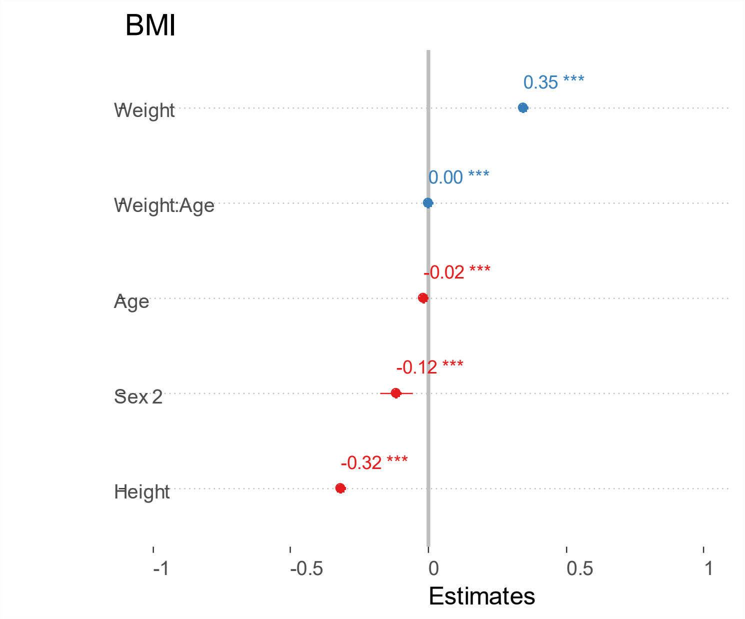
```
AIC(m4, m5, k = log(nrow(d))) %$% .[order(AIC), ]            # If K ~ log(n), then AIC = BIC

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

Present results

Tables vs images

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 tiempo para hablar de todo

```
summary(data)                       # Summary report of the table
cor(x, y)                           # Correlation between two variables
GGally::ggpairs(data)               # Pair plot for all variables
GGally::ggcorr(data)                # Correlation plot for all variables

model <- lm(y ~ x, data = d)        # Simple model
model <- lm(y ~ ., data = d)        # Model with all the variables

summary(model)                      # Summary report of the model
coef(model)                         # Extract coefficients
confint(model)                      # Extract confidence intervals
plot(model)                         # Represent model
predict(model, newdata = )          # Predict new data not seen by the model
fitted(model)                       # Strate tight values
resid(model)                        # Extracting residual error
allEffects(model)                   # Extract all effects from the model
```

Time to program

For example

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

# Find the best possible model

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

summary(m)

plot(m)
```

Support channels

Interactive support

StatQuest
https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw

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?