MEDAlics Centro di Ricerca per le Relazioni Mediterranee

Università per Stranieri "Dante Alighieri" Reggio di Calabria

Istituto di Ricerca per l'Innovazione e la Tecnologia nel Mediterraneo

Città Metropolitana di Reggio Calabria

Machine Learning

Mariachiara Fortuna | 16 - 17 Marzo 2018

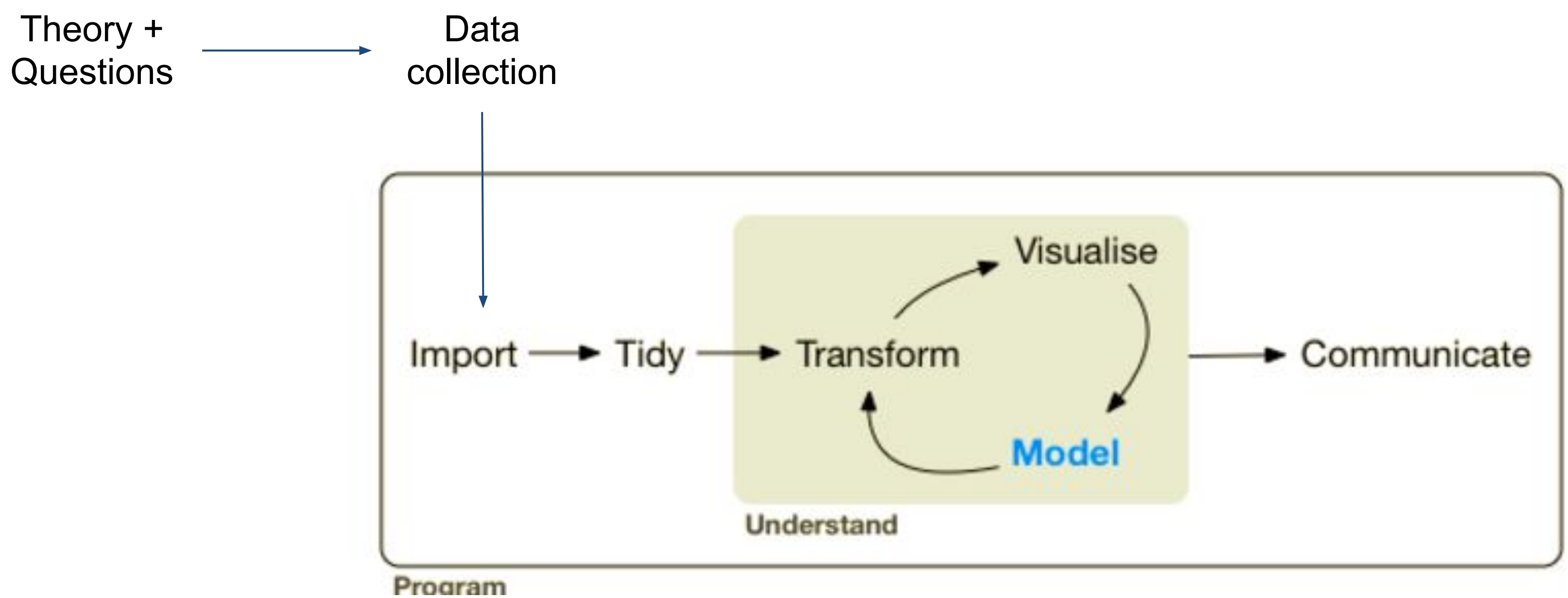
Class Materials

https://github.com/mariachiarafortuna/machineLearningClass

We will work on comics data, using R !

Data source: https://github.com/fivethirtyeight/data/tree/master/comic-characters
Inspiration: https://fivethirtyeight.com/features/women-in-comic-books/

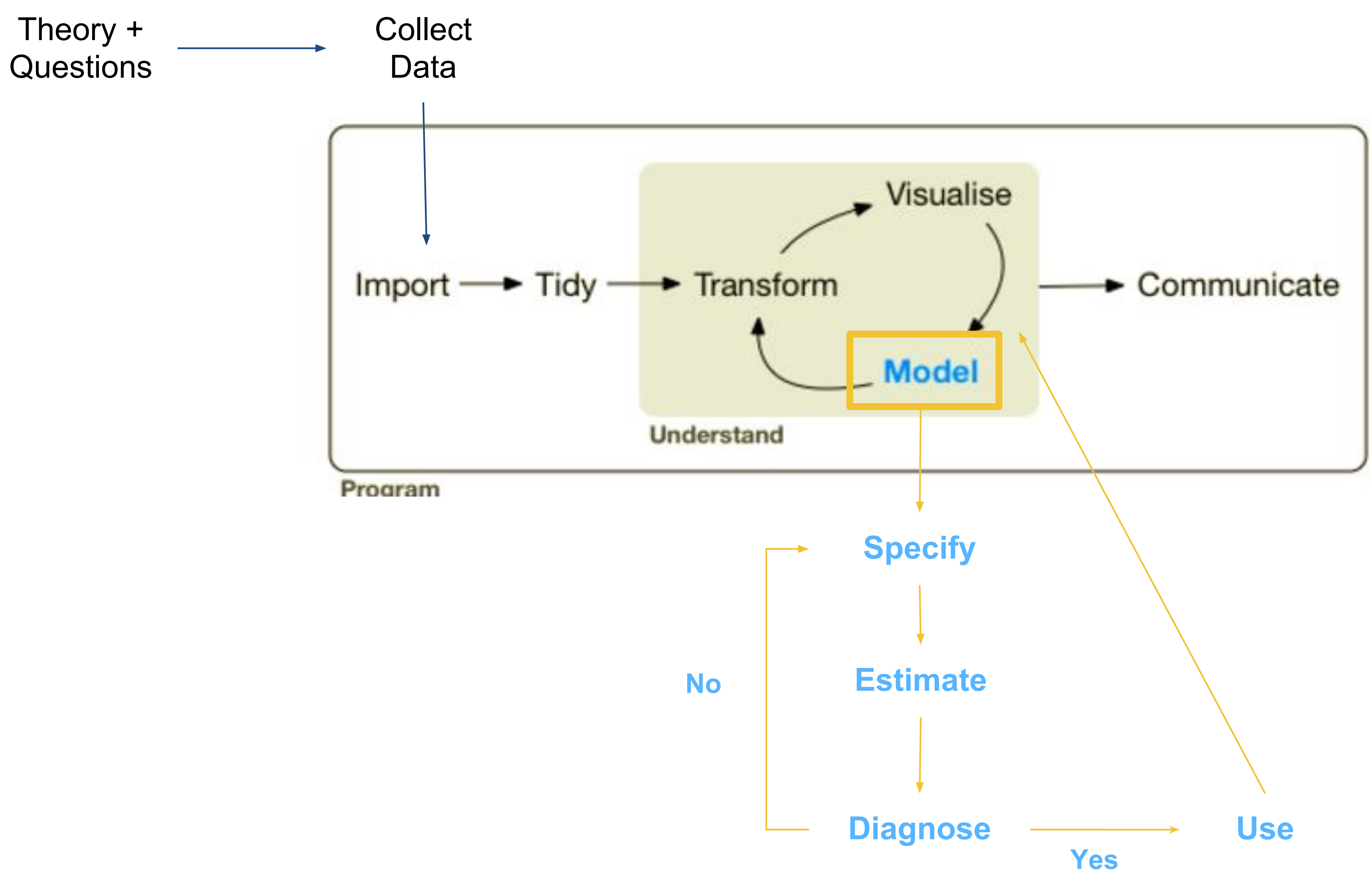
The data analysis workflow



https://rviews.rstudio.com/2017/06/08/what-is-the-tidyverse/

Forecasting Numerical Output: Linear Regression

A focus on modeling



Focus on modeling

Specify

No

Estimate

Diagnose

Yes

Use

$$Y = f(X_1, X_2, \dots, X_k) + \varepsilon$$

Y : dependent variable or response

$X_1, X_2, \dots, X_k$ **:** independent variables

f : the functional relationship

ε : error term (everything that is not explained by the model)

Simple linear regression

$$Y = \alpha + \beta X + \varepsilon$$

Y : dependent variable or response

X: independent variable

f : linear relationship

ε : error term

```
m1 <- lm(data=data, y ~ x1)
```

```
m1 <- lm(data=cars, speed ~ dist)
```

Simple linear regression

Specify

No

Estimate

Diagnose

Yes

Use

```
Call:
lm(formula = speed ~ dist, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.5293 -2.1550  0.3615  2.4377  6.4179 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.28391    0.87438   9.474 1.44e-12 ***
dist         0.16557    0.01749   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.156 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
summary(m1)
```

Simple linear regression

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.28391    0.87438   9.474 1.44e-12 ***
dist         0.16557    0.01749   9.464 1.49e-12 ***
---
```

Intercept (α): "When X is 0, we expect Y to be α"

Slope (β): "If X increases by one, we expect Y to increase by β1"

We test the null hypothesis that α and β are equal to zero (no linear relationship) *under some regularity conditions.*

P-value: probability of obtaining an effect at least as extreme as the one in our sample data, assuming the truth of the null hypothesis

~ "Probability that the relationship that we observe is due to chance"

Simple linear regression

```
Residual standard error: 3.156 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

R^2: Explained variation / Total variation

Show the percentage of the response variable variation that is explained by a linear model

We test the null hypothesis that R^2 is equal to zero (the proportion of variance explained by the model is zero), *under some regularity conditions.*

~ "Probability that the relationship expressed by the model is due to chance"

Simple linear regression

$$Y = \alpha + \beta X + \varepsilon$$

Predictions : $y_i^* = \alpha + \beta x_i$

Residuals : $e_i = y_i - y_i^* = y_i - \alpha - + \beta x_i y_i^*$

```
y_pred <- predict(m1)
ei <- residuals(m1)
```

Simple linear regression

Specify

No

Estimate

Diagnose

Yes

Use

Regularity conditions
(required for inference)

1. ***Errors have zero mean***
   (implies that $E(Yi) = \alpha + \beta Xi$)

2. ***Errors are uncorrelated***
   (implies that $Cov(Yi, Yj) = 0$)

3. ***Errors have a constant variance***
   (implies that $Var(Yi) = \sigma^2$)

[extra: Errors are normally distributed
*(gives extra inferential properties)****]***

Estimates and p-values are reliable if the regularity conditions hold

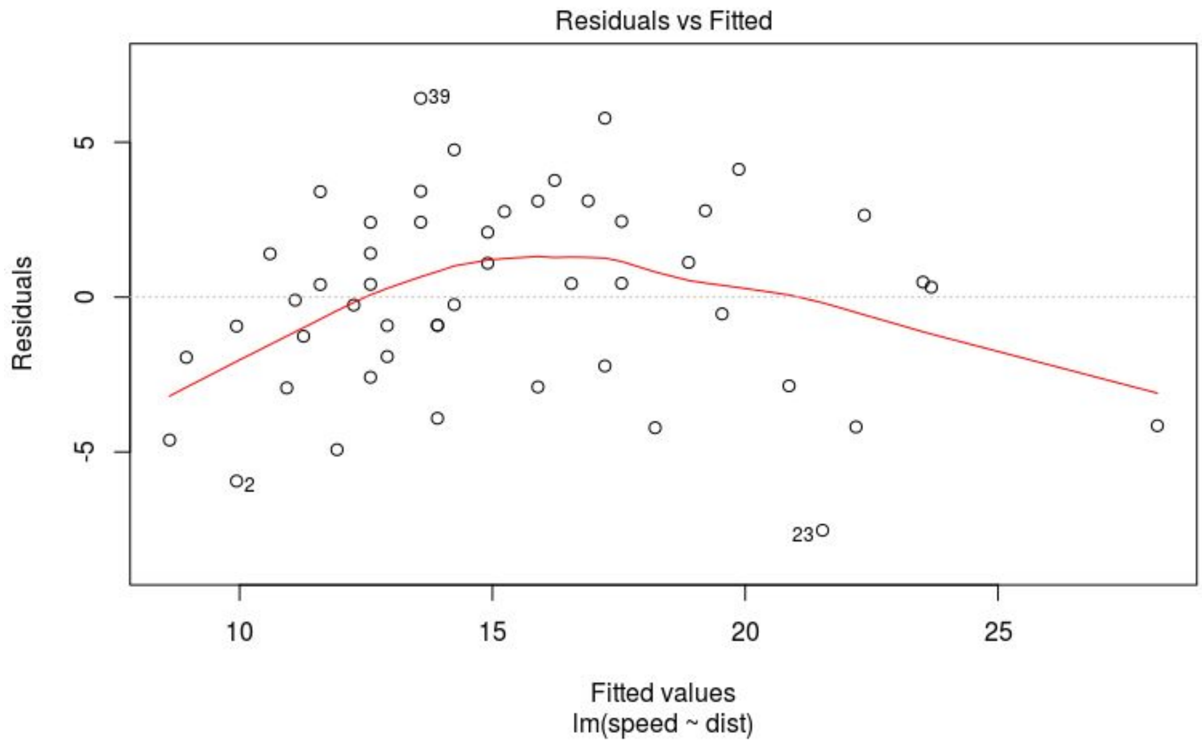
Simple linear regression

We can check if the regularity conditions hold looking at the “errors” of our model, the residuals

```
plots(m1)
```

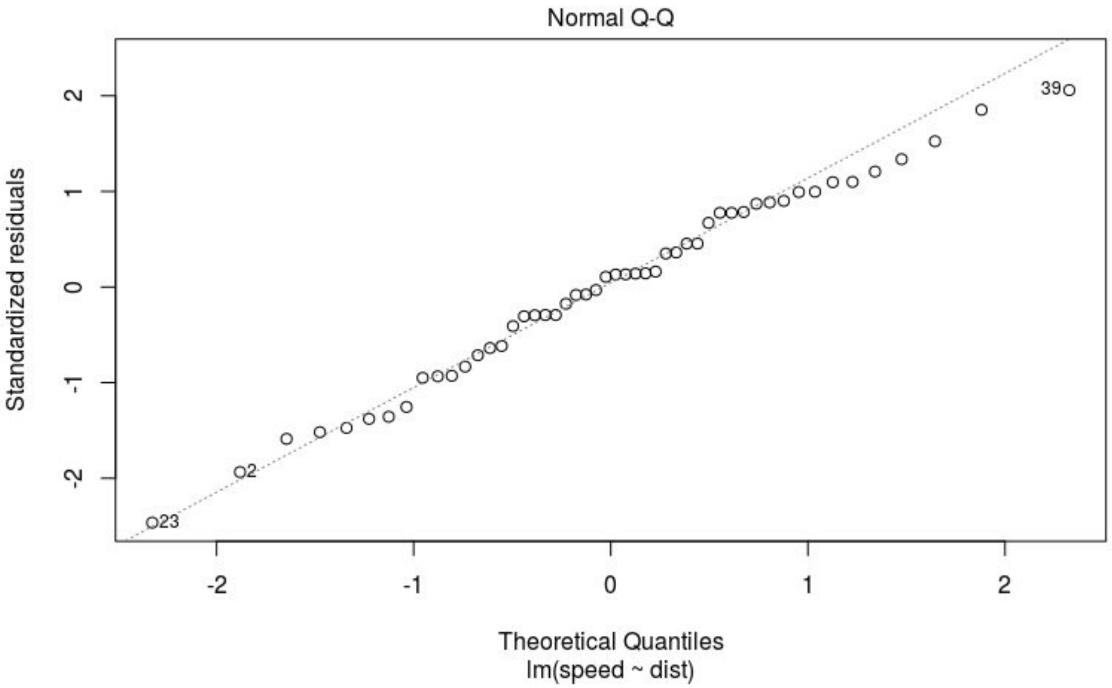
External resources:
http://data.library.virginia.edu/diagnostic-plots/

Simple linear regression



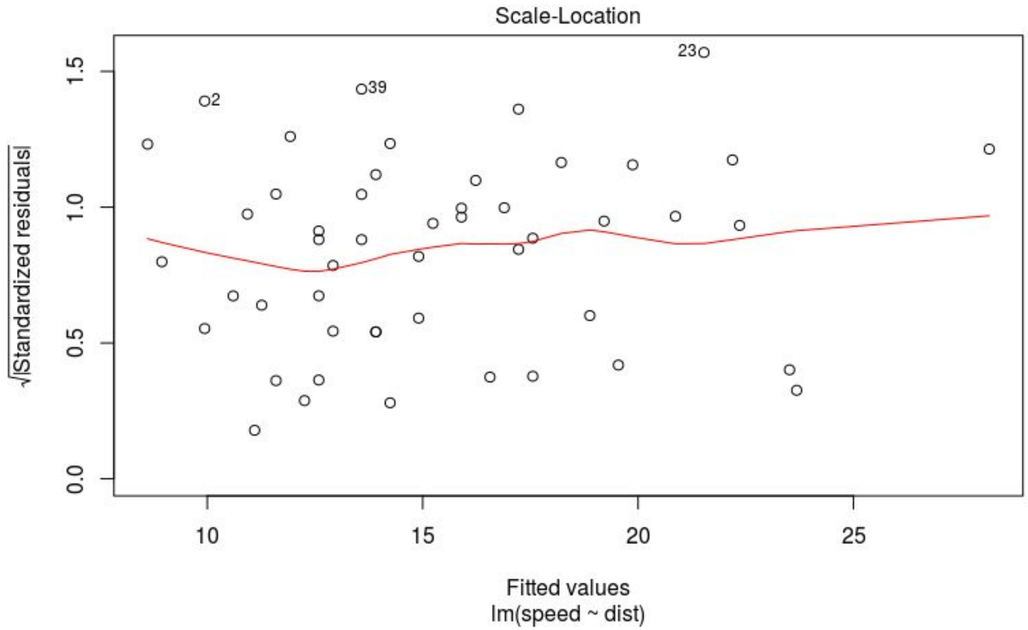
Residual vs fitted plot: If the 3 conditions hold, the residuals are fairly linear around 0, and equally "spread" around 0 when the x value increases. In particular, this plot shows if residuals have non-linear patterns.

Simple linear regression



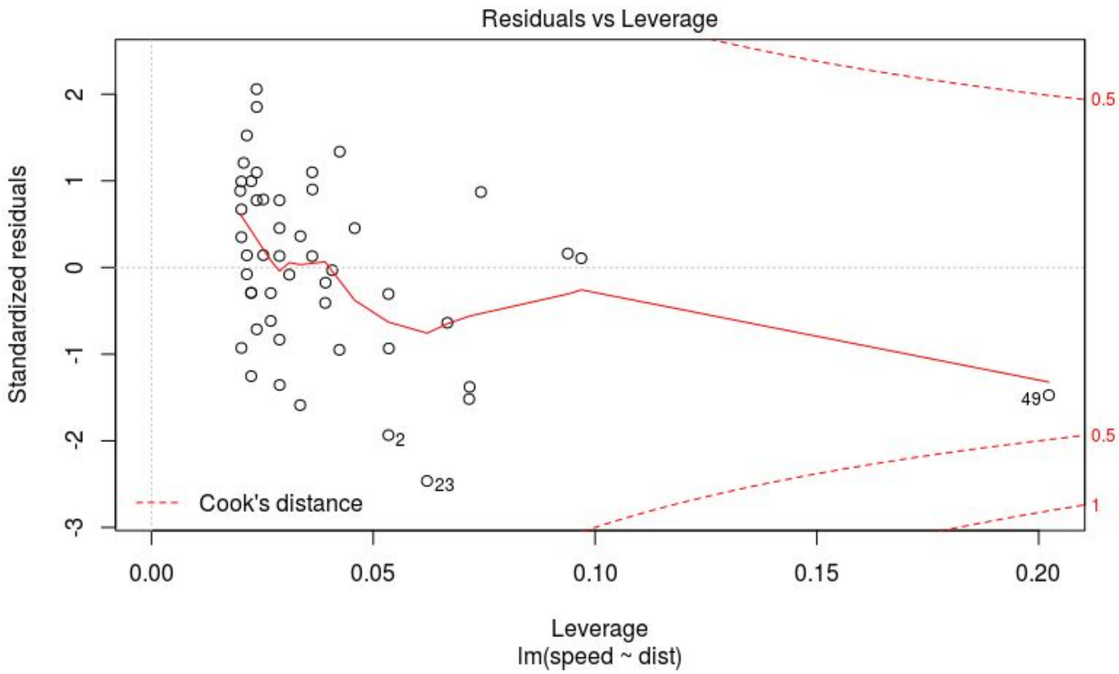
Q-q plot: If the normality condition holds, the residuals lie on the diagonal

Simple linear regression



Scale-Location plot: Shows if residuals are spread equally along the ranges of predictors: it checks if the 3' hypothesis holds.

Simple linear regression



Residual vs Leverage plot: Show if there are outliers that strongly influence the regression. We can consider to exclude influential points to increase the regression fit to the data.

Extra topic: log-level regression

Specify

No

Estimate

Diagnose

Yes

Use

$$\ln(Y) = \alpha + \beta_1 X_1$$

$\%\Delta y = 100 \cdot \beta 1 \cdot \Delta x$

"if we change x by 1 (unit), we'd expect our y variable to change by $100 \cdot \beta 1$ percent"

Technically, the interpretation is the following:

$$\%\Delta y = 100 \cdot \left(e^{\beta_1} - 1\right)$$

but the quoted interpretation is approximately true for values $-0.1 < \beta 1 < 0.1$ (and it's much easier to remember.)

External resources:
http://www.cazaar.com/ta/econ113/interpreting-beta

Multiple linear regression

Specify

No

Estimate

Diagnose

Yes

Use

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

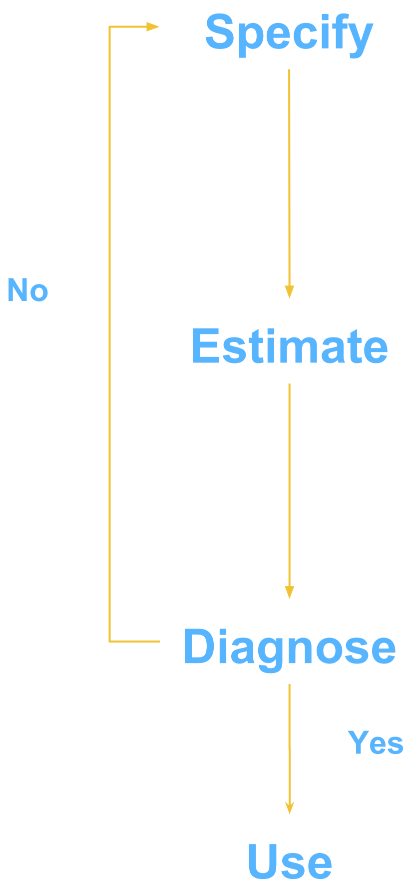
Y : dependent variable or response

$X_1, \ldots, X_k$: independent variables

- Estimation and test for each parameter
- Similar interpretation and procedure

```
m2 <- lm(data=data, y ~ x1 + x2)
```

Multiple linear regression



What if X is qualitative?

We will have a different intercept for each level of X

What if there is an interaction?

We will have a different intercept and a different slope for each level of X

```
m2 <- lm(data=data, y ~ x1 + x2 + x1*x2)
```

Find similar observations: Cluster Analysis

Cluster Analysis

Clustering: task of grouping observations in a way that objects in the same group (called a cluster) are *more similar* to each other than to those in other groups (clusters).

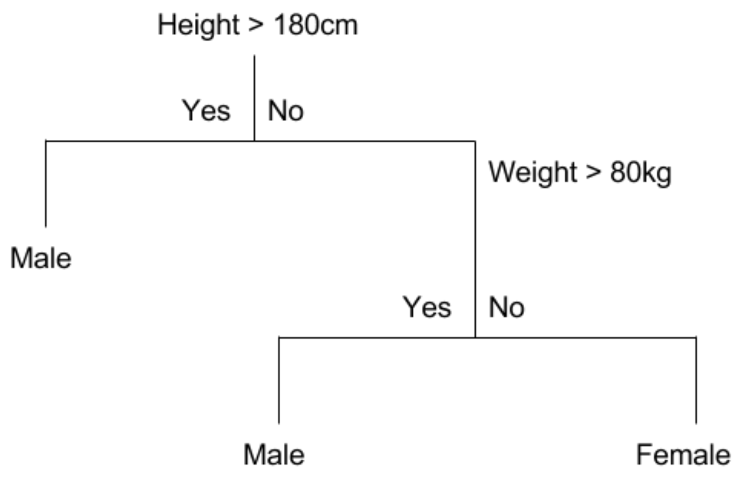
Unsupervised learning

Several methods to solve the task: we will see only **k-means** (centroid based method)

```
kmeans(scaled_data, n)
```

Classification: Decision Tree

Classification tree



Classification Trees are used to predict a qualitative response.

The method follows a hierarchical approach and the output is particularly easy to understand.

At each node (step) one independent variable and one of its values is used to partition the predictors space into simple regions.

Classification tree

Growing the Tree

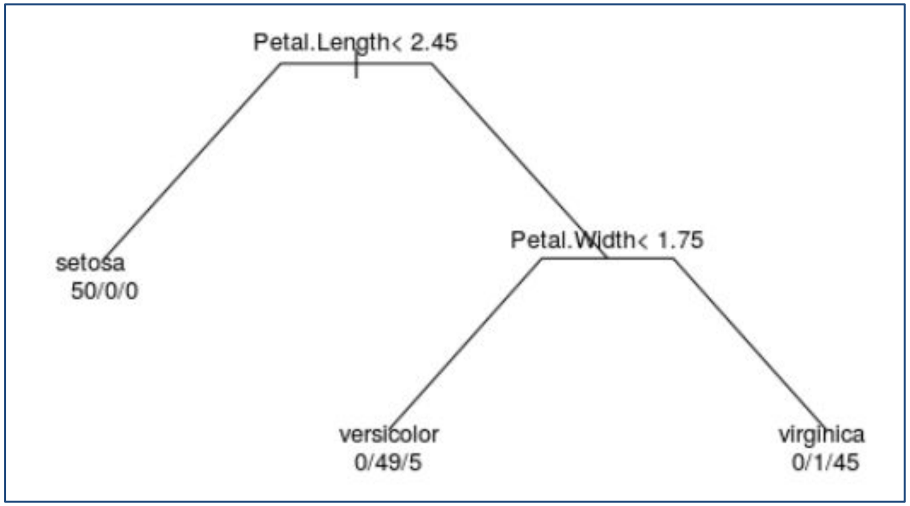
To grow a Classification Tree, the Tree algorithm searches the partition that produces the minimum “within node variability”.

```
library(rpart)

iris_tree <- rpart(Species ~ ., method = "class",
data = iris)
```

Classification tree

Growing the Tree

```
plot(iris_tree)
text(iris_tree)

print(iris_tree)
```




```
n= 150

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 150 100 setosa (0.33333333 0.33333333 0.33333333)
  2) Petal.Length< 2.45 50   0 setosa (1.00000000 0.00000000 0.00000000) *
  3) Petal.Length>=2.45 100  50 versicolor (0.00000000 0.50000000 0.50000000)
    6) Petal.Width< 1.75 54   5 versicolor (0.00000000 0.90740741 0.09259259) *
    7) Petal.Width>=1.75 46   1 virginica (0.00000000 0.02173913 0.97826087) *
```

Classification tree

Pruning the tree

When we build the tree we want to avoid **overfitting** (~ we don't want to add random variation into prediction).

Prune a Tree means choose the *right* number of split.

A way is to choose the number of split with the lowest cross-validation error.

```
Variables actually used in tree construction:
[1] Petal.Length Petal.Width

Root node error: 100/150 = 0.66667

n= 150

    CP nsplit rel error xerror     xstd
1 0.50      0      1.00   1.20 0.048990
2 0.44      1      0.50   0.70 0.061101
3 0.01      2      0.06   0.08 0.027520
```

Cross-validation error

```
printcp(iris_tree)
```

Classification: Logistic Regression

Logistic regression

Specify

No

Estimate

Diagnose

Yes

Use

$$\text{logit}(Y|x) \;=\; \alpha + \beta X \;+\; \varepsilon$$

Y : dependent variable or response

X: independent variable

f : linear relationship

ε : error term

```
m1 <- glm(data = data, y ~ x,
family = "binomial")
```

Logistic regression

Example: Probability of passing an exam versus hours of study [edit]

Suppose we wish to answer the following question:

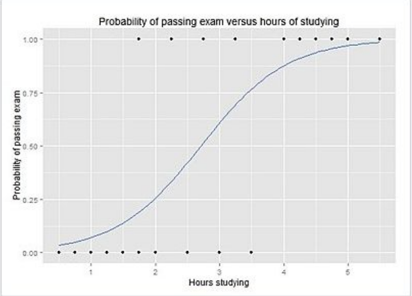
> A group of 20 students spend between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability that the student will pass the exam?

The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used.

The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.



Graph of a logistic regression curve showing probability of passing an exam versus hours studying

The logistic regression analysis gives the following output.

	Coefficient	Std.Error	z-value	P-value (Wald)
Intercept	−4.0777	1.7610	−2.316	0.0206
Hours	1.5046	0.6287	2.393	0.0167

The output indicates that hours studying is significantly associated with the probability of passing the exam ($p = 0.0167$, Wald test). The output also provides the coefficients for $\text{Intercept} = -4.0777$ and $\text{Hours} = 1.5046$. These coefficients are entered in the logistic regression equation to estimate the probability of passing the exam:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp\left(-\left(1.5046 \cdot \text{Hours} - 4.0777\right)\right)}$$

For example, for a student who studies 2 hours, entering the value $\text{Hours} = 2$ in the equation gives the estimated probability of passing the exam of 0.26:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp\left(-\left(1.5046 \cdot 2 - 4.0777\right)\right)} = 0.26$$

Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp\left(-\left(1.5046 \cdot 4 - 4.0777\right)\right)} = 0.87$$

https://en.wikipedia.org/wiki/Logistic_regression
https://datascienceplus.com/perform-logistic-regression-in-r/

Mariachiara Fortuna

mariachiara.fortuna@quantide.com
mariachiara.fortuna1@gmail.com

www.milanor.net
https://www.facebook.com/MilanoRcommunity/
https://www.meetup.com/it-IT/R-Lab-Milano/

https://github.com/mariachiarafortuna/
@maryclary