Linear Regression

Part 1

The aim of this course is to give you most of the information you need to understand Linear regressions, as well as give you all tools necessary to put it in place. We will focus on the intuition of how it works without ignoring the maths beyond. Each part ends with several exercices to do. Some of them are very easy and other harder. I highly recommend to verify all demonstration presents during the course, for 2 main reasons: verify there is no mistake and help you to understand the objects manipulated. I truly believe that equations manipulation is very helpful in order to have a good intuition of why they are the way they are.

> *"Statistics is about reducing the amount of data."* **R. Fisher**

Part 1

Intuition behind linear regression

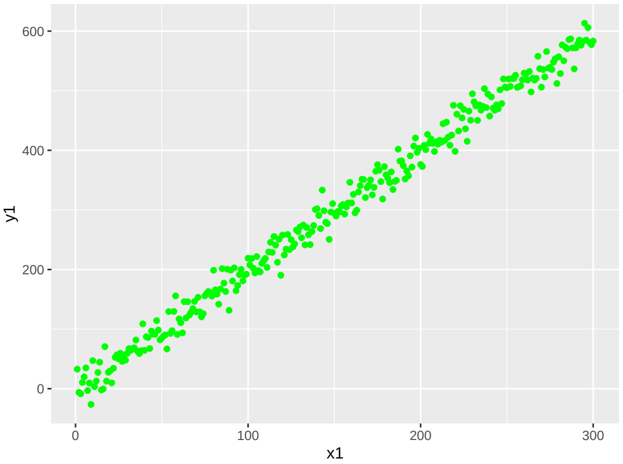
Correlation analysis **Linear regression is mostly correlation analysis**. This course will not talk about causality because its way more complex and not the point here. You just have to keep in mind that correlation does not allow to infer causality: **this is neither a necessary nor a sufficient condition**. Correlation is about relation between different variables. We want to know if 2 variables (or more) are correlated. But, if we ask ourselves what this formally means, it's not that easy to define. We will use the following definition: **2 variables are correlated if they tends to not be independent**. Are the following variables correlated or not?

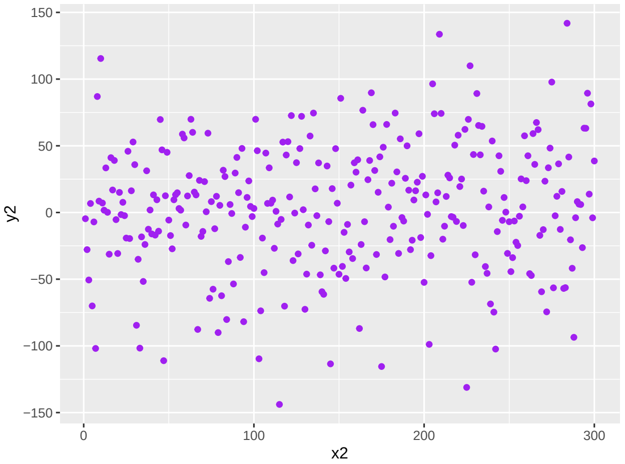
```r
#sample1
n = 300
x1 = seq(1, n)
y1 = 2*x1 + rnorm(n=n, sd=18)
data1 = data.frame(cbind(x1, y1))

#sample2
n = 300
x2 = seq(1, n)
y2 = rnorm(n=n, sd=50)
data2 = data.frame(cbind(x2, y2))

#plot both relationship in order to make it more intuitive
library(ggplot2)
plot1 = ggplot(data1, aes(x=x1, y=y1)) + geom_point(col="green")
plot2 = ggplot(data2, aes(x=x2, y=y2)) + geom_point(col="purple")

library(ggpubr)
ggarrange(plot1, plot2)
```





With plots, it's pretty obvious to detect when variables are correlated. So the question we want to ask is: **how can we objectively qualify correlation** (i.e: without using plots)?
For this, we generally use **Pearson correlation coefficient**:

$$r = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

The function $cov(x, y)$ is the covariance between x and y. The latter tends to tell us if, in a finite sample, when x_i is above (or below) the mean of its sample, y_i tends to also be above (or below) the mean of its sample. **The higher the covariance is, the more when x_i is above its mean, y_i also is**. To the contrary, the lower the covariance is, the more when x_i is above its mean, y_i is below. More formally, we can describe it this way:

$$cov(x, y) = \frac{1}{n} \sum (x_i - \overline{x})(y_i - \overline{y})$$

Covariance actually measures correlation between 2 variables. But **Pearson correlation coefficient is better because it normalizes the covariance between -1 and 1**. In fact, covariance unit is the product of the units of x and y: doesn't make lot of sense for us. That's why we divide it by the product of standard deviation.

Pearson correlation coefficient can be interpret as follow: the closer it is to 1 (resp. -1), the more there is a positive (resp. negative) correlation. If it's near 0, it seems that there is no correlation between variables (that is not completely true, but we will do like it is). If we take back our 2 last sample back, we can compare their correlation coefficient.
Ask yourself before checking the results: which one will be the highest? Why?

```r
#sample1
n = 100
x1 = seq(1, n)
y1 = -2*x1 + rnorm(n=n, sd=18)
cat("Correlation between variables from our first sample: r =", cor(x1,y1))
```

```
## Correlation between variables from our first sample: r = -0.9487655
```

```r
#sample2
n = 100
x2 = seq(1, n)
y2 = rnorm(n=n, sd=10)
cat("Correlation between variables from second first sample: r =", cor(x2,y2))
```

```
## Correlation between variables from second first sample: r = 0.05280643
```

The first 2 variables are very (negatively) correlated (r is near -1).
The other 2 variables are much less correlated (r is near 0).

Non-linear correlation

Since the Pearson coefficient correlation only measures linear correlation, **it will miss correlation that are not linear** (e.g. exponential or logarithmic). In order to solve this issue, statisticians have invented other measure of correlation. We will discuss the main one: Spearman coefficient correlation.
It's defined as follow:

$$s = \frac{\frac{1}{n}\sum(rx_i - \overline{rx}_n)(ry_i - \overline{ry}_n)}{r_\sigma(x)r_\sigma(y)}$$

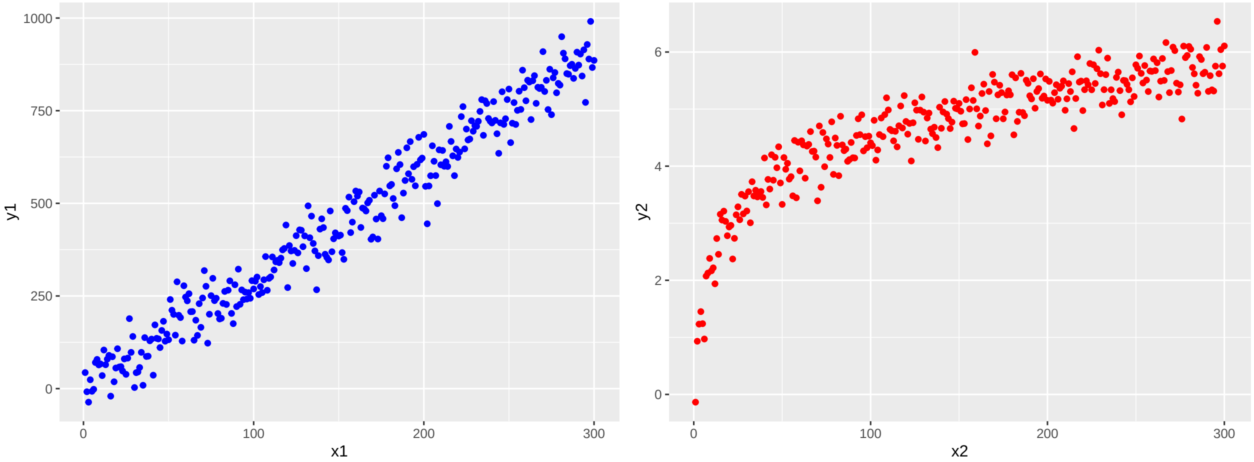
Where r design the rank of the individual compared to other individuals. Using this statistic compared to the Pearson's one allow us to measure non-linear relationship. Spearman's coefficient is considered as non-parametric because it doesn't use the value of x_i but its rank. Let's see an example of when this might be useful.

```r
#sample with linear relationship between x and y
n = 300
x1 = seq(1, n)
y1 = 3*x1 + rnorm(n=n, sd=45)
data1 = data.frame(x1,y1)

#sample with logarithmic relationship between x and y
n = 300
x2 = seq(1, n)
y2 = log(x2) + rnorm(n=n, sd=0.3)
data2 = data.frame(x2,y2)

#plot both relationship in order to make it more intuitive
library(ggplot2)
plot1 = ggplot(data1, aes(x=x1, y=y1)) + geom_point(col="blue")
plot2 = ggplot(data2, aes(x=x2, y=y2)) + geom_point(col="red")

library(ggpubr)
ggarrange(plot1, plot2)
```



Compute both Pearson and Spearman coefficients correlation

```
## Pearson correlation between variables from our first sample: r = 0.9826558
## Spearman correlation between variables from our first sample: r = 0.9831665
```

```
## Pearson correlation between variables from our second sample: r = 0.8573567
## Spearman correlation between variables from our second sample: r = 0.9177044
```

Spearman coefficient in higher for the sample with a logarithmic (non-linear) relationship. This tool is useful to illustrate that 2 variables can be considered as correlated but their relationship is more complex that a straight line. We say that Spearman coefficient is *more robust* than Pearson's.

Why are we interested in correlation?

Depending of what you're interested in (descriptive or inferential statistics), you will not use correlation for the same reason and the same way. **Correlation is an important statistic tool used in a daily basis in lots of field**. For example, if we observe a strong correlation between a genome and breast cancer, we might do some prevention for people with this genome and easily attenuate the cancer impact, especially if it's not already present.
However, **you should always have in mind that correlation has not so much to do with causality**. The latter is way more complex and will not be discussed here.

In our case, we will mainly focus on inferential statistics. The latter is used when we want to infer our sample results to the rest of the population, whether for prediction or other.

Prediction and explanation We can use linear regression for 2 things: make some predictions and explains phenomenon.

Prediction: when the model we have created is good enough, we can be confident saying that we can generalize it to new cases. Example: I have a very good model that predicts the height of people based on their birth height and the height of their parents. It can be useful for doctors to get an idea of how tall a baby will be later.

Explanation: describing statistically a variable with others. Thanks to my model, I can say that 20% of income after college is explained by the income of your parents. The more we are able to explain a variable, the more we might find what are its most important determinants.

Exercises

1. Prove that $cov(x, x) = var(x)$
2. Prove that $cor(x, x) = 1$
3. Find an example of variable prediction and of variable explanation.
4. With R, create two samples: one where the Pearson coefficient correlation is equal to 1 (but with 2 distinct variables) and one where it's between [-0.1 ; 0.1]. Use a sample size >100.
5. What is the condition on the type of the variables used in order to compute the Pearson (or Spearman) coefficient correlation (it's implicitly said in the course)?
6. *With your own words*, make a short description of what correlation is.
7. Using internet (or other), find a statistic that measures the amount of variance explained (there are multiple ones).