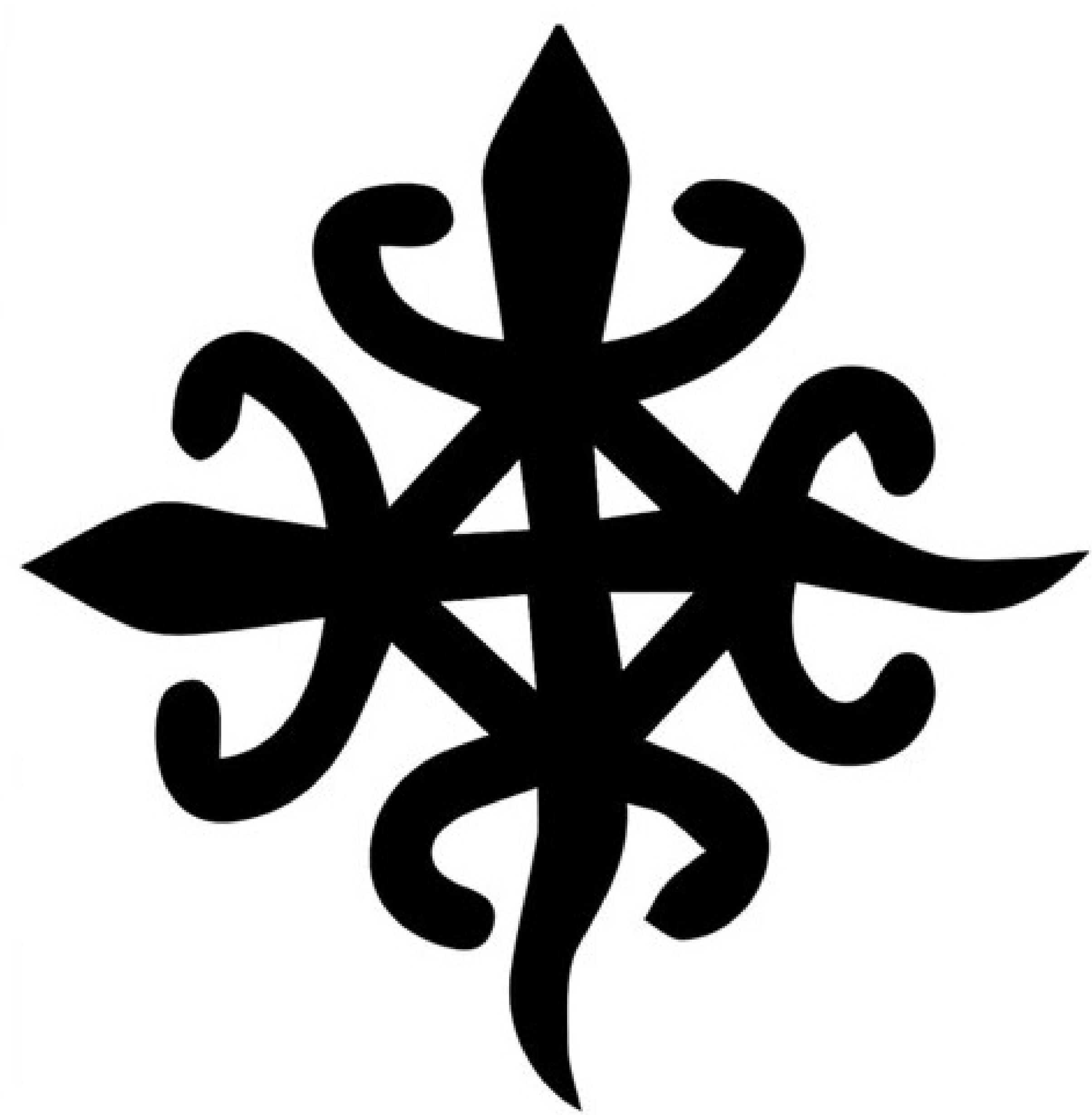


Chapter 3: About associations

Prelude: ***Hypotheses non fingo*?**

I have not yet been able to discover the reason for these properties of gravity , and I make no assumptions. *Anything that is not deduced from the phenomenon can be called a hypothesis; and hypotheses, whether metaphysical or physical, or based on hidden qualities, or mechanical, have no place in experimental philosophy. In this philosophy, particular propositions are inferred from the phenomenon, and then generalized by induction.*

The rationale presented in the previous chapter is directly related to the hypothetical-deductive method and its philosophical principles. Although suitable for this scenario, the interpretation of the p-value is not very intuitive. It involves *measuring how unlikely observations are in a hypothetical scenario under the null* *hypothesis*. His most popular (wrong) translation is that it represents *“the chance that the result of this study is wrong”*.

The framework described in the previous chapter is sufficient to produce a cryptic scientific work for laypeople.

When following pre-defined recipes (formulation of H_0 and H_1, calculation of statistics and p-values), a text seems to conform to academic standards, even if the elementary hypothesis around the research object is simplistic. Thus, inadvertently, we prioritize the form and relegate the core of scientific proposals to the background.

Another side effect is the search for p-values that reject H_0, disregarding theoretical precedents and probabilistic assumptions (multiple tests).

The difficult interpretability of the p-value and the frequent pitfalls involved in the inference process led the scientific community to question the hegemony of this parameter. There is a present tendency to abandon the p value and the limit $p < 0.05$ as canonical criteria.

We will learn about formal arguments against the hypothetical deductive method in science. For now, just know that it is always advantageous to obtain other information, complementary or alternative.

In this chapter, we will learn how to estimate (1) the magnitude of the difference between two samples and (2) how related are paired values (e.g. weight and height).

I have not as yet been able to discover the reason for these properties of gravity from phenomena, and I do not feign hypotheses. For whatever is not deduced from the phenomena must be called a hypothesis; and hypotheses, whether metaphysical or physical, or based on occult qualities, or mechanical, have no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterwards rendered general by induction. *Isaac Newton (1726). Philosophiae Naturalis Principia Mathematica, General Scholium. Third edition, page 943 of I. Bernard Cohen and Anne Whitman’s 1999 translation, University of California Press ISBN 0-520-08817-4, 974 pages.*

Effect size

The effect size helps us to express magnitudes. Returning to the previous example, what is the use of a significant difference between the size of the birds' beaks, if it is 0.00001 mm?

Still, there are cases in which small studies suggest important effects, but the sample size does not provide enough statistical power to reject the null hypothesis.

In addition to knowing how unlikely the difference is observed, it is natural to imagine how big it is.

A very popular measure is Cohen's *D (Cohen's D).*

It is a parameter that expresses the magnitude of the difference without using units of measurement.
A soccer fan tells (happily) to a friend that her favorite team won with a score of 4 × 1 (goals).However, this friend accompanies basketball and is used to scores like 102 × 93 (baskets). How is it possible to compare goals with baskets? Which win represents the most disparate scores: 4 × 1 or 102 × 93?

The problem here is that scores behave differently between sports. Basketball scores have much higher averages and dispersions. Cohen's D consists of expressing this difference in standard deviations. Simple enough:

$$D_{cohen} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

Using the * effects * library, we can directly calculate:

```
library(effects)
# O dataset galapagos_birds was created in chapter 1
>cohen.d(galapagos_birds$X1,galapagos_birds$X2)

Cohen's d

d estimate: -5.460017 (large)
95 percent confidence interval:
    lower     upper 1
-5.954047 -4.965987
```

Cohen proposed some tracks to classify the magnitude of these effects:

	Small	Medium	Big
Cohen's D	0-0.2	0.2-0.5	0.5 - 0.8

Thus, we can update our previous results, also reporting the effect size of the difference and its confidence interval. If the distributions are from the same family, we have a comparable estimate between contexts.

Correlations

In the scientific endeavor, we don't just stick to comparisons. A more noble objective is to describe exactly how the relationship between studied entities occurs.

As we know, there are many classes of functions to express relationships between variables / sets. In the previous chapters, we used some functions, such as $y = \sqrt{x}$ and $y = e^x$.

Several natural laws have become particularly known, such as the relationship between force, mass and acceleration, elucidated by Newton:

$$\vec{F} = m\vec{a}$$

And the relationship between mass and energy for an object at rest, discovered by Einstein:

$$E = mc^2; c^2 \sim 8.988 * 10^{16} \frac{m^2}{s^2}$$

The above equations describe a linear relationship between quantities.

Linear relations

A linear relationship between two variables indicates that they are correlated in a constant proportion for any interval.

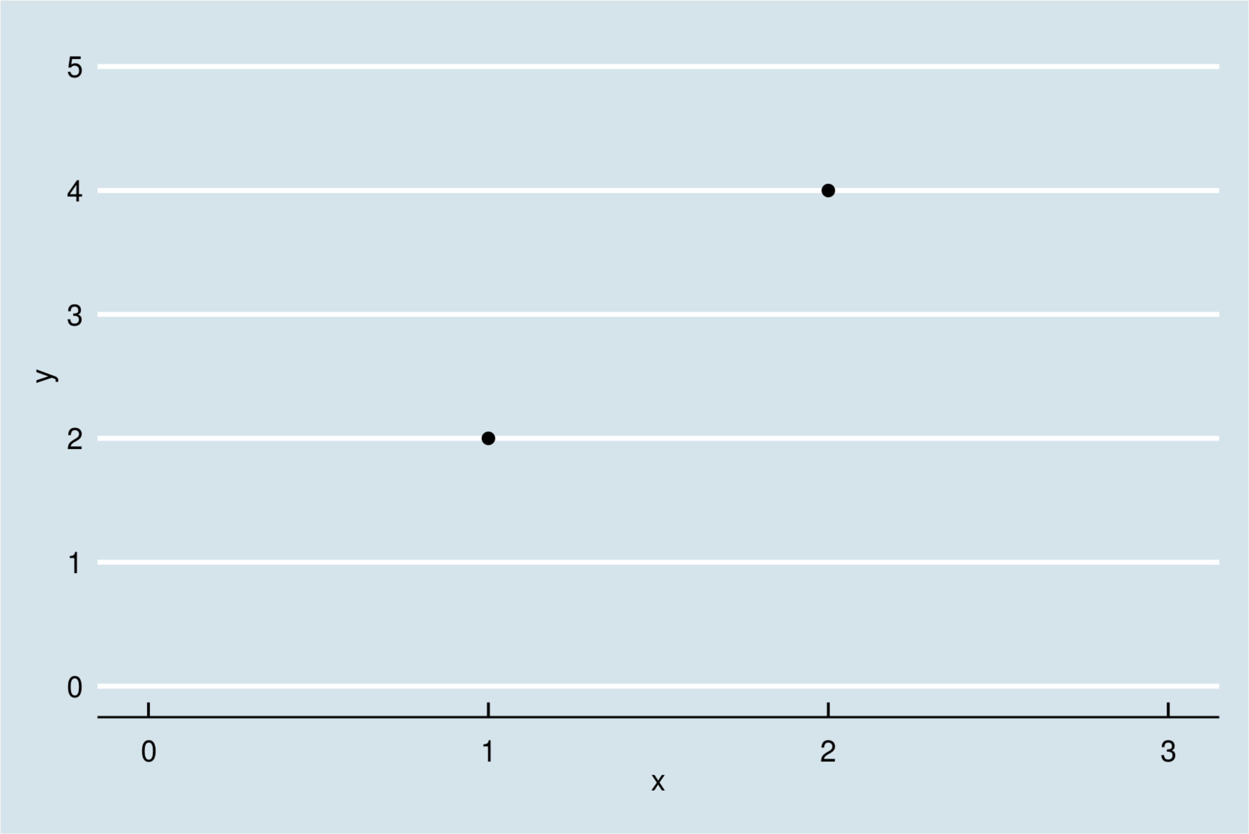
That is, higher mass values correspond to a proportional increase in energy. The value of c^2 expresses this constant proportion.

Example: a water molecule weighs approximately $m_{H_2O} = 2.992 \times 10^{-23} g$. Therefore, the associated energy is $E_{H_2O} = 2.992 \times 10^{-23} \times 8.988 \times 10^{16} \sim 2.689^{-6} J$. If we triple the number of water molecules, the same will happen with the associated energy: $E_{3H_2O} = 3 \times E_{H_2O}$.

If the correlation is positive, increments in x will be proportional to increments in y. If the correlation is negative, increments in x will be proportional to decreases in y.

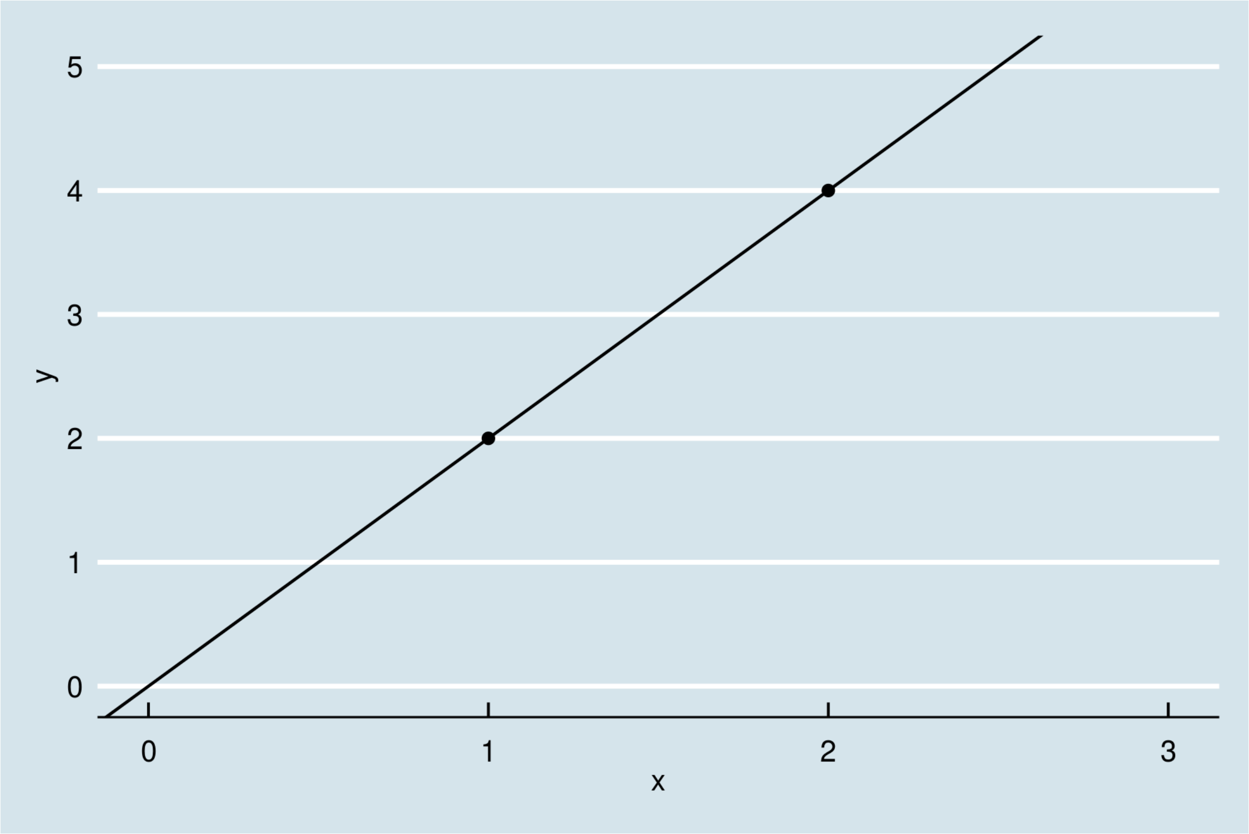
In a perfect scenario, if we know that there is a linear relationship between variables, we need only two observations to find out the proportion between them. This problem is identical to that of finding the slope of the line that passes through two points. It is easy to solve using elementary techniques.

```
>library(ggplot2)
>ggplot()+
geom_point(mapping=aes(x=1,y=2))+
geom_point(mapping=aes(x=2,y=4))+
xlim(0,3)+ylim(0,5)+
theme_economist()
```



$y = \beta * x$
$a = (1, 2); b = (2, 4) \rightarrow \beta = 2$

```
    >ggplot()+
    geom_point(mapping=aes(x=1,y=2))+
    geom_point(mapping=aes(x=2,y=4))+
    xlim(0,3)+ylim(0,5)+
    geom_abline(slope = 2)+
    theme_economist()
```



Errors and randomness

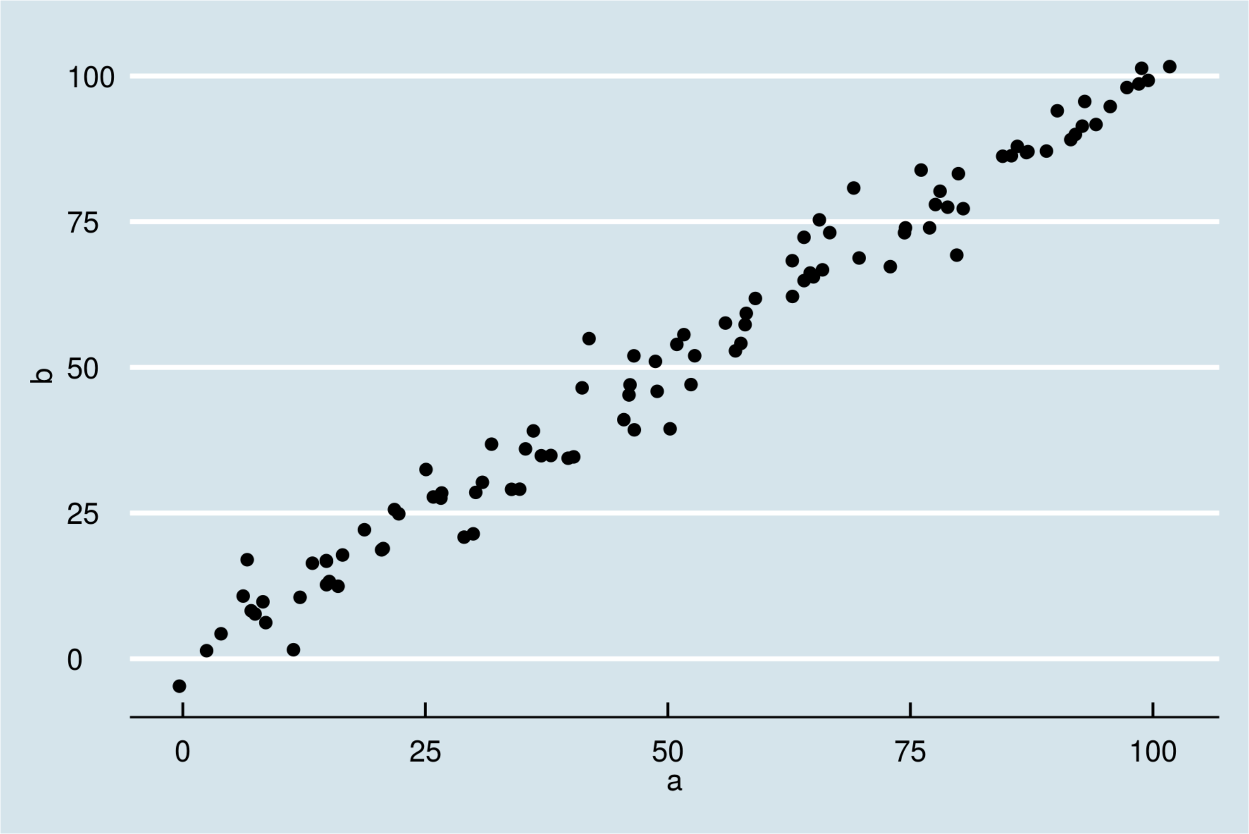
Controlling experimental factors, the relationships described are quite accurate. In a scenario without friction with surfaces and air, the measurement errors obtained with $\vec{F} = m\vec{a}$ are very low. However, this is not always true. First, we may experience interference from unknown variables.

Imagine a set of anthropometric measures, such as the height and weight of individuals. A human's height is expected to be related to his weight. However, other unmeasured characteristics, such as the percentage of total fat, may interfere with the final values. We normally treat these fluctuations as random errors [ˆ 11].

We can simulate this scenario starting from identical variables and adding random noise.

```
>set.seed(2600)
>a <- seq(1:100)+rnorm(n=100,sd=3)
>b <- seq(1:100)+rnorm(n=100,sd=3)

>cor_data <- data.frame(a,b)
>ggplot(cor_data,aes(x=a,y=b))+
geom_point()+theme_economist()
```



The result suggests that there is a strong linear relationship between x and y. On the other hand, we note that it is impossible for a line to cross all points. Next, we will investigate how to quantify the linear correlation, as well as find the line that minimizes the distance for all observations.

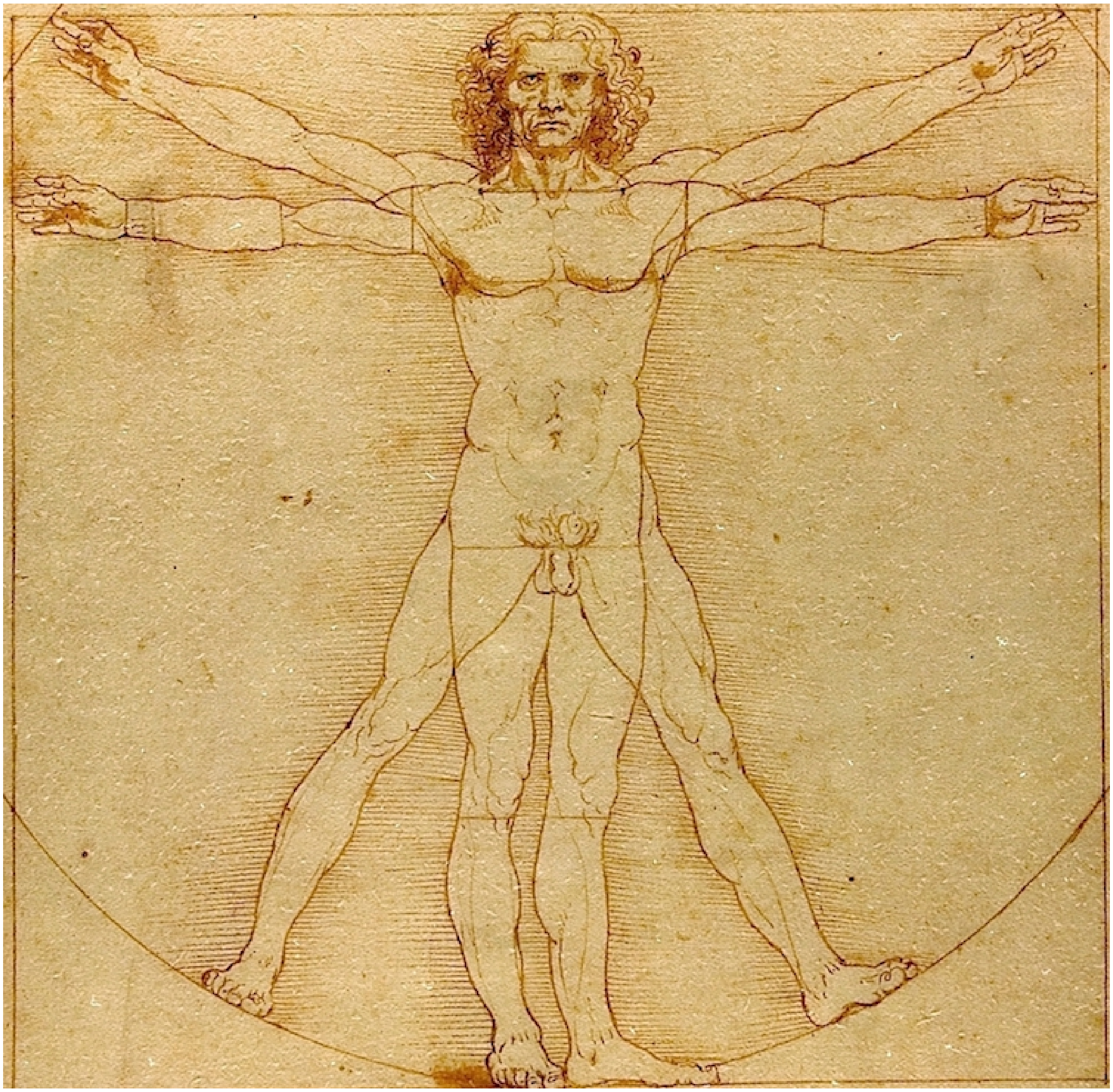
With these tools, we can extend our inferences. In addition to comparisons, we will have notions about the magnitude of a relationship, as well as we can predict the expected value for new observations.

[^ 11]: The nature of randomness is a philosophical question. Ultimately, we can imagine that it would be possible to explain random fluctuations through unknown variables (*hidden variables*). This is true of most natural phenomena. However, recent experimental findings in quantum physics (*Bell's inequality experiment*) suggest that hidden variables cannot explain the probabilistic nature of observations.

Pearson's product-moment correlation coefficient, or simply Pearson's (ρ).

Pearson's (*rho*) correlation coefficient is a real number guaranteed [^ 12] between -1 and 1. Expresses the magnitude and direction of a linear relationship, with -1 being a perfect inverse relationship and 1 being a direct relationship perfect.

For the data we generate, the correlation is almost perfect: $\rho = 0.989$. The coefficient has *product-moment* in its name, because it uses an abstraction originally used in physics, which we studied in the previous chapter: the moment (torque).



Calculating linear correlations

The notion of **distance**or* *deviation **was repeated many times. In fact, the linear correlation coefficient was born when Francis Galton (1888) numerically studied two apparently distinct problems in anthropometry [1]: 1.** Anthropology: **If we recovered from an ancient tomb only one bone of an individual's thigh (femur), what could we say about its height? 2.** Forensic science:** In order to identify criminals, what can be said about different measures by the same person?

Galton realized that he was actually dealing with the same problem. Given paired measures, (x_i, x_i'), what does the x_i deviation tell you about the x_i' deviation?

The femur recovered from a pharaoh's skeleton is 5 cm larger than the average. How far from the average do we expect your height to be? Naively, we can think that if one of the measures is 1% higher than the average,

[1] Francis Galton's account of the invention of correlation. Stephen M. Stigler. Statistical Science. 1989, Vol. 4, No. 2, 73-86.

the other will also be 1% higher. Galton realized that there was a trap in that thought.

Although there is a relationship between the measures, there are also random fluctuations: part of the deviation results from this. We need to understand the degree of correlation to make a good guess.

Then, he proposed a coefficient measuring the relationship between deviations of variables. If femur size and height are closely related, a large femur suggests an equally tall individual. Otherwise (low correlation), a large femur (high deviation) does not imply great stature.

To quantify the relationship, we multiply the deviations for each pair of measures:

$$Cov(X, X') = \sum_{i=1}^{N}(x_i - \mu_x)(x'_i - \mu_{x'})$$

The above formula expresses **covariance** between X and X' and will be useful in other contexts. The expression resembles the calculation of the first moment, but each deviation is multiplied by the corresponding deviation of the paired measure. Hence the name product-moment correlation coefficient.

Note that if both deviations agree in the direction (sign), the result of the multiplication will be positive. Consistently matching pairs increase the value of the final sum. If both deviations disagree in the direction (sign), the result will be negative. Consistently discordant pairs decrease the value of the final sum.

Thus, we can have highly correlated variables positively or negatively, as long as the sense of the association is constant. On the other hand, if the measures are at times inconsistent and at other times concordant, the values tend to cancel each other out in the sum and the result approaches zero.

Observing only the covariance is dangerous, as the values depend on the unit of measurement and data dispersion.

We calculated Pearson's correlation coefficient, normalizing [^ 17] the covariance by dividing it by the product of standard deviations:

$$\rho_{XX'} = \frac{cov(X, X')}{\sigma_X \sigma_{X'}}$$

Extensively:

$$\rho_{XX'} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(x'_i - \mu_{x'})}{\sqrt{\sum_i^N (x_i - \mu_x)^2}\sqrt{\sum_i^N (x'_i - \mu_{x'})^2}}$$

Uma boa notícia: ρ follows a known distribution, the t distribution, with n-2 degrees of freedom. We can use the previous tools to test hypotheses.

Practical example

The following example was a happy find. At the time, the Brazilian government was discussing the need to increase the number of doctors to improve health care. Some argued that it was the right decision, while others advocated that investments should be made in other areas of health.

Out of curiosity, I accessed the WHO (World Health Organization) and World Bank (World Bank) data on the number of doctors per country and health indicators. My expectation was to find at least a timid relationship between indicators. More than that, understand the location of Brazil in relation to other countries. I was surprised by a strong correlation, which we will explore next.

We adopted countries as an observational unit with measures x, the number of doctors 1,000 inhabitants, and y, the expected life expectancy at birth. Using data obtained from the WHO and World Bank portals, we plot the points on the Cartesian plane.

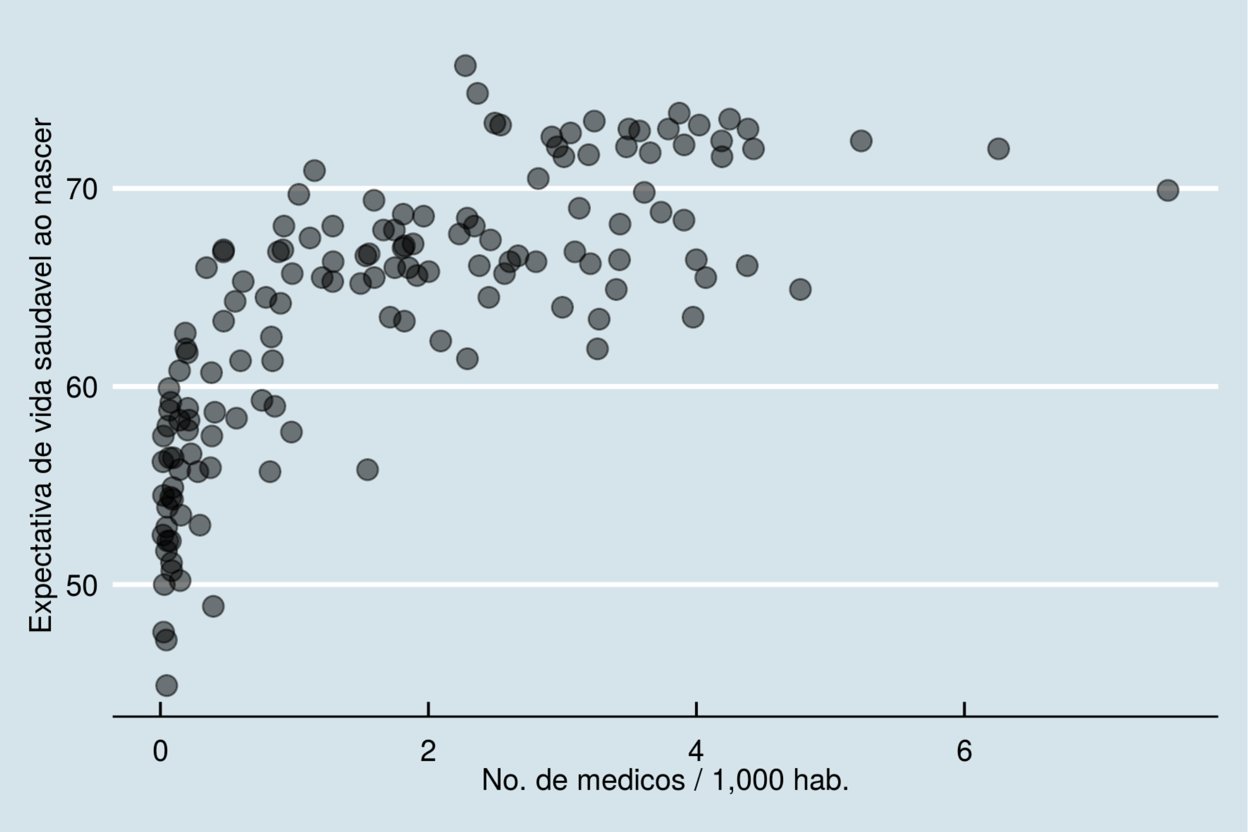
```
# http://apps.who.int/gho/data/view.main.HALEXv
# https://data.worldbank.org/indicator/SH.MED.PHYS.ZS
>library(magrittr)
>library(ggplot2)
>library(dplyr)

>worldbank_df <- read.csv("data/API_SH.MED.PHYS.ZS_DS2_en_csv_v2_10227587.csv",
                          header = T,skip = 3)
>colnames(worldbank_df)[1] <- "Country"

>worldbank_df$n_docs <- sapply(split(worldbank_df[,53:62], #lists of values
                                    seq(nrow(worldbank_df))),
    function(x) tail(x[!is.na(x)],1)) %>% #last non-null values
  as.numeric

>who_df <- read.csv("data/who_lifeexpect.csv",skip=2)
>who_df$hale <- who_df$X2016
>uni_df <- left_join(worldbank_df[,c("Country","n_docs")],
                     who_df[,c("Country","hale")],by="Country")

>ggplot(uni_df,aes(x=n_docs,y=hale))+
  geom_point(alpha=0.5,size=3) +
  xlab("No. of doctors / 1,000 inhab.")+
  ylab("Healthy life expectancy at birth")+
  theme_economist()
```



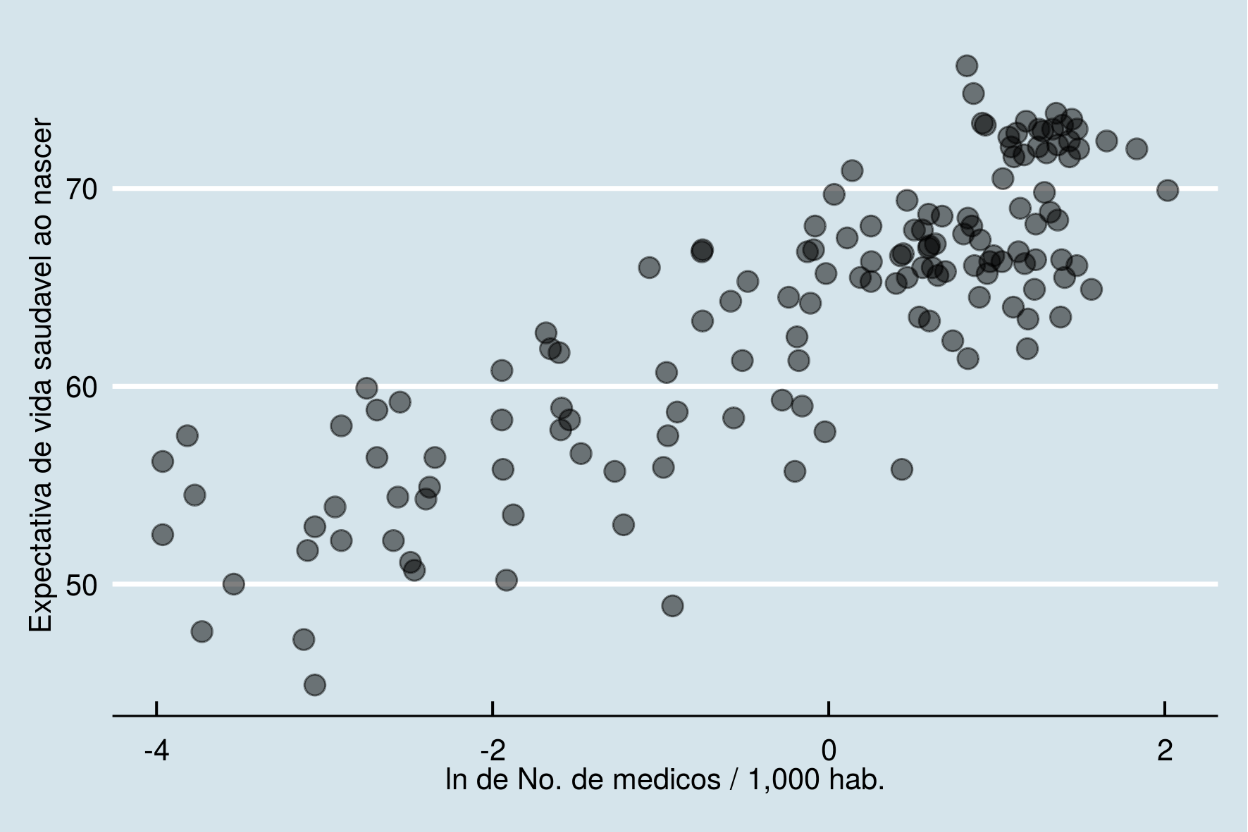
It is clear that the pattern is not random. Visually, we noticed that the value of life expectancy increases with a greater number of doctors. Still, we noticed an initially rapid increase until it reached a plateau. The pattern is similar to that of a logarithmic curve.

$y = log(x)$ or $HALE = log(N_{médicos})$

If this hypothesis is true, transforming the number of doctors using a logarithmic function will make the relationship linear with the transformed variable:
If $y = log(x)$, we do the replacement $x' = log(x)$ to get $y = x'$.

Then life expectancy becomes linearly correlated with the logarithm of the number of doctors.

```
>uni_df$log_docs <- log(uni_df$n_docs)
>ggplot(uni_df,aes(x=log_docs,y=hale))+
  geom_point(alpha=0.5,size=3) +
  xlab("ln No. of doctors / 1,000 inhab.")+
  ylab("Healthy life expectancy at birth")+
  theme_economist()
```



In fact, we see a notable linear trend for points.

Using the native implementation in R for Pearson's coefficient:

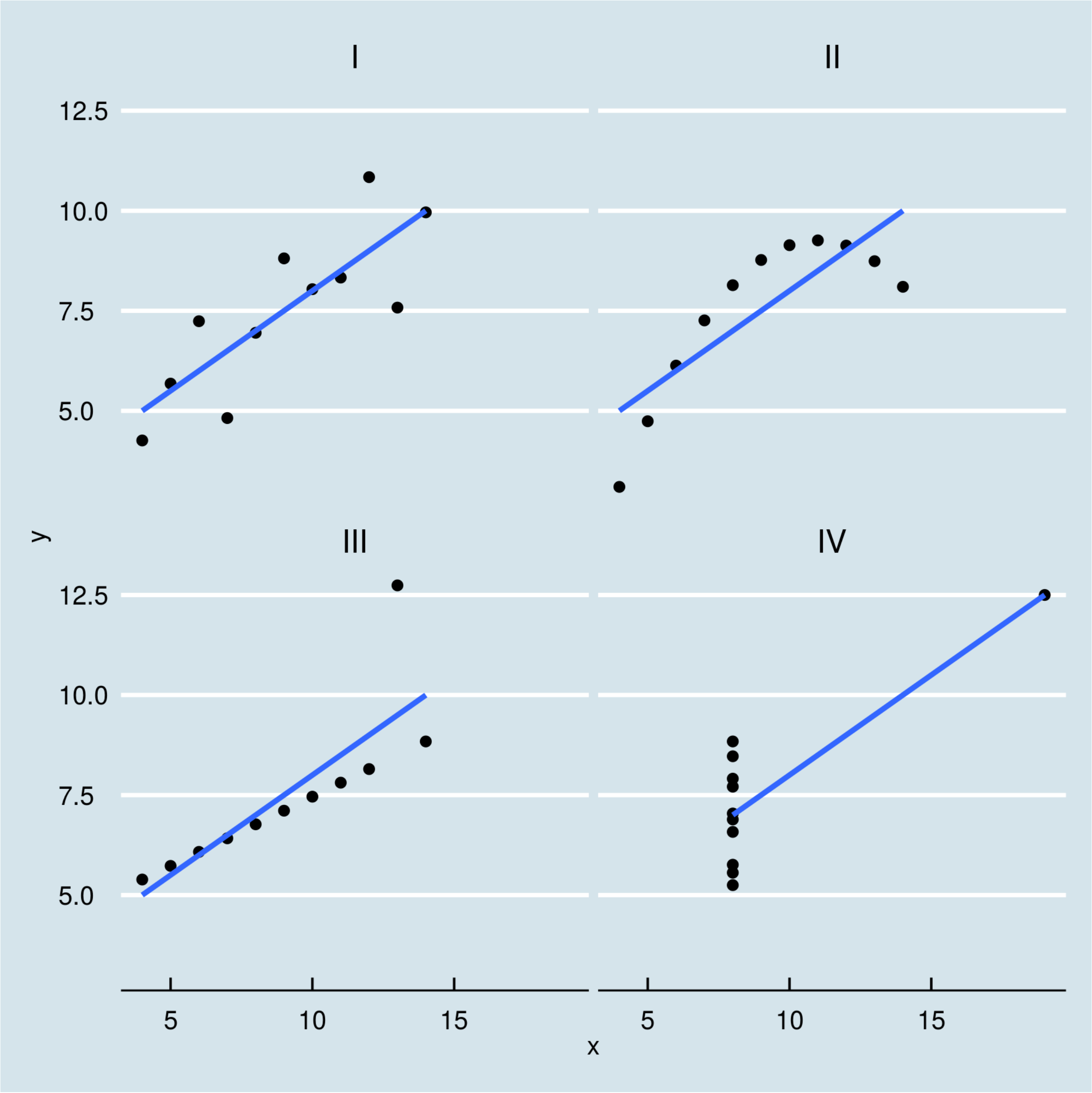
```
>cor.test(uni_df$log_docs,uni_df$hale)
Pearson's product-moment correlation
data:  uni_df$log_docs and uni_df$hale
t = 18.572, df = 143, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.7854248 0.8828027
sample estimates:
    cor
    0.8407869
```

The linear correlation obtained for our sample of countries is surprisingly large, as suggested by the visualization ($\rho \sim 0.841$).

The p value is low ($p < 0.001$) considering the null hypothesis H_0 of $\rho = 0$. We conclude then that there is a significant linear relationship of strong magnitude between the logarithm of the number of doctors and the life expectancy of the countries in our sample.

It is really curious that there is such an evident mathematical relationship between tenuously connected constructs. The average time that an organism takes between birth and death and the number of professionals working. It is virtually impossible to spell out each causal relationship behind that relationship, which manifests itself robustly through the sum of many related factors.

Note *It is customary to state that there is no relationship between variables if the relationship coefficient does not prove to be important. As we have seen, this indicator reports only on linear relationships between variables. Data visualization can be of great help in inferring the nature of relationships. Data with very different distributions can result in equal coefficients, as shown by the classic Anscombe quartet. The 4 samples below show the same correlation coefficient.*

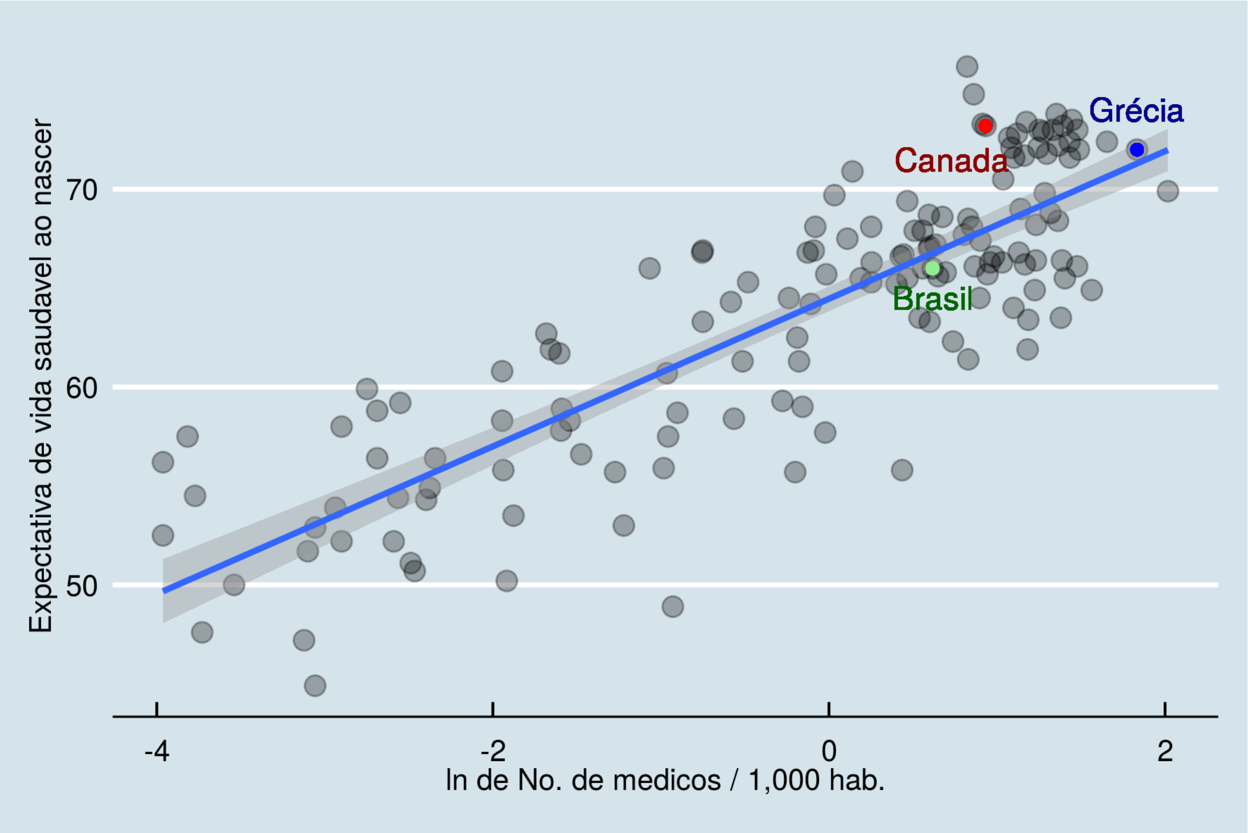


##Forecasts

We now know that it is reasonable to assume a linear relationship between these variables. As stated before, we can then find the line that minimizes the distance for observations.

The equation that describes this line tells us the expected value for life expectancy given the number of doctors.

```
>uni_df$log_docs <- log(uni_df$n_docs)
>ggplot(uni_df,aes(x=log_docs,y=hale))+
   geom_point(alpha=0.3,size=3) +  geom_smooth(method="lm")+
   geom_point(y=66.0,x=0.61626614,color="light green")+
   geom_text(y=64.5,x=0.61626614,label="Brazil",color="dark green")+
   geom_point(y=73.2,x=0.93177030,color="red")+
   geom_text(y=71.5,x=0.73177030,label="Canada",color="dark red")+
   geom_point(y=72.0,x=1.833381,color="blue")+
   geom_text(y=74.0,x=1.833381,label="Greece",color="dark blue")+
   xlab("ln No. of doctors / 1,000 inhab.")+
   ylab("Healthy life expectancy at birth")+
   theme_economist()
```



Biases must be addressed before conclusions are reached, but the model is sufficiently interpretable to make decisions. z A good policy can compare the investment value by sectors with other countries under similar conditions and different results. Assuming that there is really a linear relationship, we see that Brazil is quite close to what was expected for the number of doctors [ˆ 18]. If the strategy is to hire more people, we can look at programs in countries with more doctors per capita and positive results (e.g. Greece). If the strategy is to save on payroll and prioritize investment in structure, we can use countries with high life expectancy for the expected number of professionals (e.g. Canada).

[ˆ 18]: It is practically a consensus among specialists that Brazil has a problem with the distribution of professionals, with a shortage of doctors in poorer and less populated areas.

Predictions with linear models

How to guess one measure based on the other? Considering the linear relationship previously discovered, we can create a function that receives as input the value of a variable (number of doctors) and returns the expected value for life expectancy as an output.

Finding the equation that describes this function consists of finding the line that best fits the point cloud, as in the previous figure.

For this, we calculate the slope (β_1) and the vertical adjustment (β_0) that minimize the sum of the distances between the line and the observations. The term ϵ corresponds to errors, with normal distribution of mean 0 and standard deviation σ.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

We adjust the model using the R lm (linear model) function:

```
# log_docs : x' = log(x)
>lm(hale ~ log_docs, data=uni_df)

Call:
lm(formula = hale ~ log_docs, data = uni_df)

Coefficients:
(Intercept)     log_docs
      64.46         3.73
```

We have $\beta_0 \sim 64.46$ and $\beta_1 \sim 3.73$.
Our estimate for healthy life expectancy "starts" at 64.46 years and increases with the number of doctors in the country. Specifically, it increases by 3.73 for each unit of our transformed variable ($log(x)$).
In our dataset, Brazil has 1,852 doctors / 1,000 inhabitants. Our prediction then is:
$\hat{y}_{Brasil} = log 1.852 * 3.73 + 64.46 \sim 66.8$, which is very close to the real number (66).

Estimators

There is more than one way to estimate these parameters. One of particular interest, which will also serve in other contexts, is that of Maximum likelihood.

First, we determine a function that describes the probability of observation on the target variable (y_i) measurements of the predictor variables occur (x_i) and a set of parameters (β_k).

We can adopt as a likelihood function *(likelihood function)* for the values y_i a Gaussian probability distribution whose mean is given by the line $\mu_{yi} = \beta_0 + \beta_1 * x_i$. Thus, the probability of each value y_i is given by a Gaussian, according to the deviation to the value predicted by the line.

$$L \sim N(\mu_{yi}, \sigma^2)$$

.

Assuming that the observations are independent, the probability of the set of observations is given by their product.

$$L = \prod_{i=1}^{n} P(y_i | x_i; \beta_0, \beta_1, \sigma^2)$$

Replacing the values of μ for the Gaussian by the line's predictions:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}}$$

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y_i - (\beta_0 + \beta_1 x_i)^2}{2\sigma^2}}$$

This is our likelihood function and expresses the probability of observing the measures y_i given the measures x_i and considering a set of parameters (β_0, β_1).

The objective then is to find parameters that maximize this function. For convenience, we apply a logarithmic transformation to this function ($log \quad likelihood \quad function$). This transforms our product into a summation and we pass the counterdomain of the interval $[0; 1]$ for $[-\infty, 0)$.

$$\begin{aligned}\text{log likelihood}(\beta_0, \beta_1, \sigma^2) &= log \prod_{i=1}^{n} P(y_i|x_i; \beta_0, \beta_1, \sigma^2) \\ &= \sum_{i=1}^{n} logP(y_i|x_i; \beta_0, \beta_1, \sigma^2) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2\end{aligned}$$

The parameters that maximize the likelihood function (max. Likelihood, ML) are the same as those that maximize the logarithm of the likelihood function (log-likelihood).

We introduce the rationale of the ML estimator as it will be useful in the future. In fact, it is easy to understand the closed formulas for our parameters, as they only express the linear relationships explored [2]:

$\hat{\beta_1}$ expresses the magnitude of the correlation between X and Y.It is natural that its value is the covariance normalized by the variance of the predictor.

$$\hat{\beta_1} = \frac{cov(XY)}{\sigma_x^2}$$

$\hat{\beta_0}$ is our intercept, so it's the difference between predicted averages and predictions considering the average value in X.

$$\hat{\beta_0} = \mu_y - \hat{\beta_1}\mu_x$$

Finally, the variance of errors $\hat{\sigma^2}$ is given by the square of the deviations from the predictions in relation to the measures.

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta_0} + \hat{\beta_1} x_i))^2$$

The solutions above provide the best estimates we can obtain by minimizing the distance from the line to the points. We must then be concerned with whether the linear model found is good in predicting the data.

[2] Detalhes das deduções dos estimadores OLS and Max. Likelihood: https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/05/lecture-05.pdf ; https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/06/lecture-06.pdf

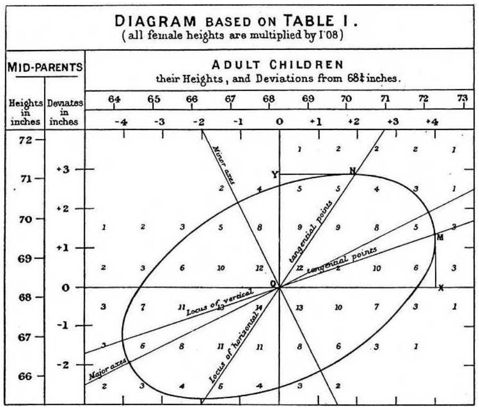


Figure 1: The first linear regression graph. Illustration by Francis Galton (1875) relationship between height of parents and children.

Evaluating performance There are different parameters to evaluate the performance of a model. In general, they seek to quantify how far the model results differ from ideal results.

For linear regression, the R^2 (coefficient of determination) is a widely used coefficient. Express the proportion between **(1)** variance explained by the model and **(2)** total variation. We call residual (or error) the difference between predicted values and real values.

(1) To capture the magnitude of model errors, we add the square of all residuals *(sum of squared residuals, SSR)* in relation to the predicted values. Be y_i the observations and $\hat{y}_i$ predictions:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(2) The total variability is quantified by adding the squared deviations from the mean *(total sum of squares, TSS)*, a term we saw in the variance calculation (second moment):

$$TSS = \sum_{i=1}^{n} (y_i - \mu_y)^2$$

So the fraction $\frac{SSR}{TSS}$ is the desired proportion. We define R^2 like:

$$R^2 = 1 - \frac{SSR}{TSS}$$

An intuitive view of SSR and TSS:

```
>source("aux/multiplot.R")
>doc_lmfit <- lm(hale ~ log_docs, data=uni_df)
>uni_df$preds[complete.cases(uni_df)] <- predict(doc_lmfit)
>uni_df$hale_mean <- mean(uni_df$hale,na.rm = T)
>ssr_res <- ggplot(uni_df,aes(x=log_docs,y=hale))+
    geom_point(alpha=0.5,size=3) +
    geom_segment(aes(xend = log_docs, yend = preds)) +
    geom_smooth(method="lm")+
    xlab("")+
    ylab("Healthy life expectancy at birth")+
    ggplot2::ggtitle("SSR") + theme_economist()
```

```
>tss_res <- ggplot(uni_df,aes(x=log_docs,y=hale))+
    geom_point(alpha=0.5,size=3) +
    geom_segment(aes(xend = log_docs, yend = hale_mean)) +
    geom_abline(slope = 0,intercept = 63.28165)+
    xlab("ln No. of doctors / 1,000 inhab.")+
    ylab("Healthy life expectancy at birth")+
    ggplot2::ggtitle("TSS")+theme_economist()

>multiplot(ssr_res,tss_res)
```

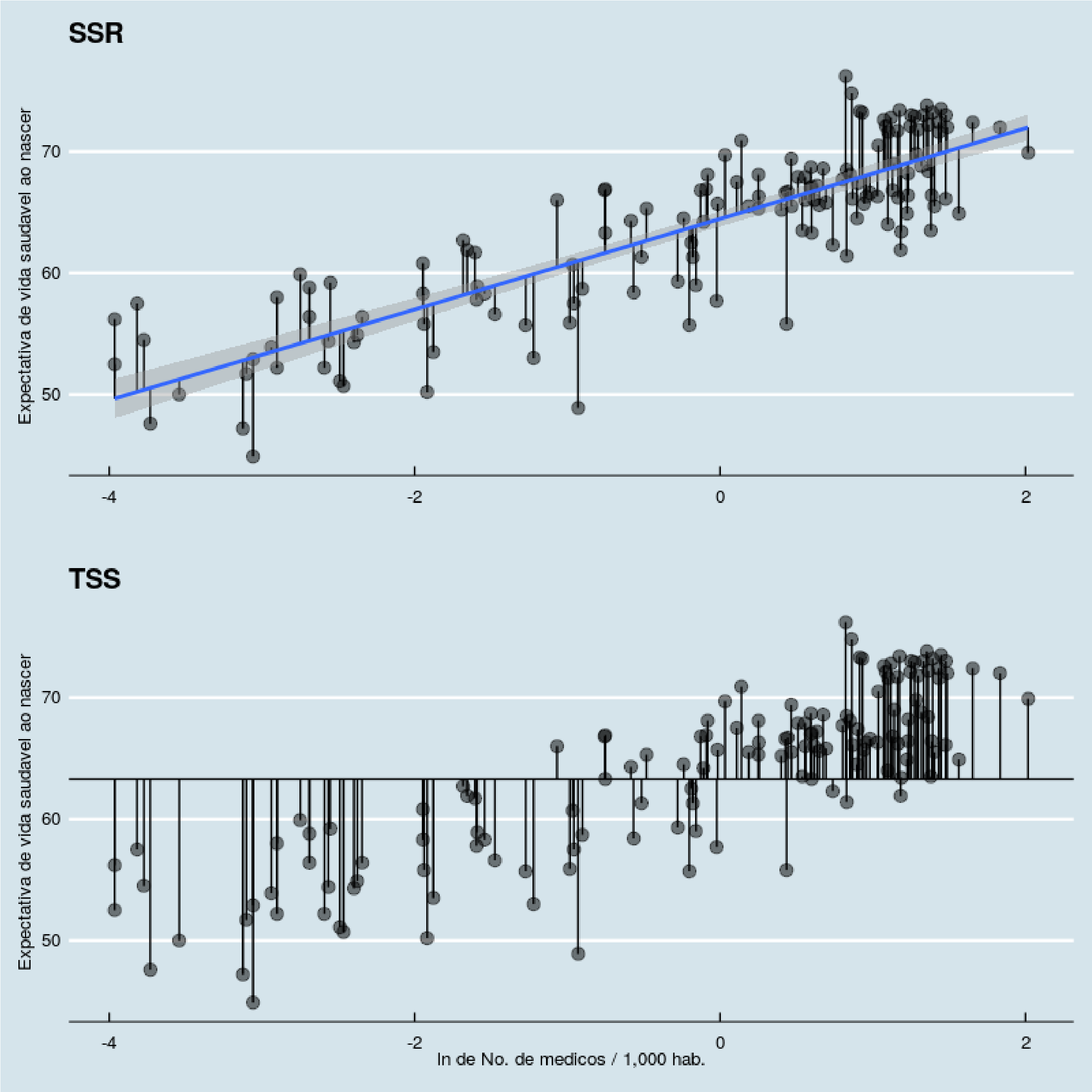


Figure 2: The square of the distance between a point and the line corresponds to a residue. We obtain SSR and TSS by adding all the residues in the upper and lower figures, respectively.

Values of R^2 close to 1 indicate residue sum (SSR) similar to 0. Using the line as a guide accumulates almost zero errors. Values of R^2 close to 0 indicate $\frac{SSR}{TSS} \sim 1$ and the predictions obtained by the model are as good as kicking the average for all cases.

```
>lm(hale ~ log_docs, data=uni_df) %>% summary
Call:
lm(formula = hale ~ log_docs, data = uni_df)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0964  -2.3988   0.3233   2.8229   8.6708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.4613     0.3162  203.84   <2e-16 ***
log_docs      3.7303     0.2009   18.57   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.779 on 143 degrees of freedom
  (119 observations deleted due to missingness)
Multiple R-squared:  0.7069,    Adjusted R-squared:  0.7049
F-statistic: 344.9 on 1 and 143 DF,  p-value: < 2.2e-16
```

To obtain the predicted values, we use the *predict* method:

```
>head(predict(doc_lmfit))

       2        3        4        7        8        9
59.90747 57.23226 65.39962 66.11533 69.54483 68.30608
```

It is also possible to obtain predictions for new values by specifying the *newdata* argument. For a country with 1.5 doctors / 1,000 inhabitants:

```
>predict(doc_lmfit,newdata = data.frame(log_docs=log(1.5)))
       1
65.97381
```

Assumptions There are some auxiliary procedures to check for possible flaws and points in the model that need attention. For example, residues can be asymmetrical. This indicates that performance changes at different intervals (heteroscedacity). Different violations require different attitudes, such as treating outliers or changing the model type. A complete list of premises, along with the R codes to test them, is available in the auxiliary material (*lm-asssumptions.R*)

Correlations and nonparametric tests

We thoroughly verified analyzes involving normal distribution, t distribution and linear relationships. However, measures often do not follow a defined distribution. Thus, making inferences using the ** parameters ** described $(\mu, \sigma, t ...)$ nos levaria a direitos erradas. Para lidar com distribuições arbitrárias, vamos abrir mão deles e conhecer ferramentas *não-paramétricas* : the rank correlation coefficient ρ Spearman's test and Mann Whitney's U test.

Ranks and Spearman's ρ

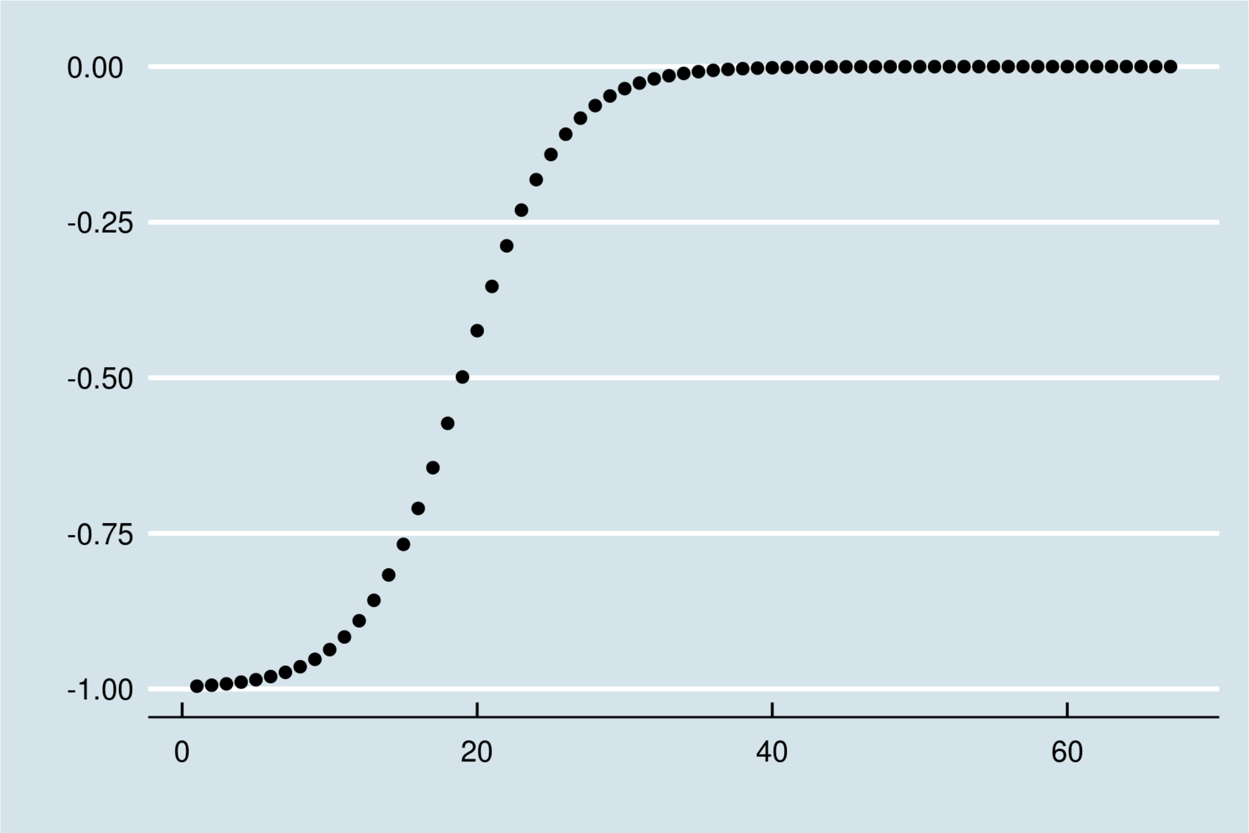
Linear relationships maintain constant proportions and we learn how to quantify them. On the other hand, two variables can have relations of other types, non-linear. In particular, if the measures have very extreme values * (outliers) * a calculation like the previous one suffers a lot with biases. A simple solution to this problem is to rank the values. Thus, the items in the set are treated by their position in relation to other items, regardless of the associated values. Example:

$$S = (1, 3, 89, 89, 39, 209) \rightarrow S_{ranked} = (1, 2, 4, 4, 3, 5)$$

Spearman's ρ is that Pearson's product-moment coefficient applied to ranks. Thus, we measure the degree to which two variables increase (or decrease) in magnitude by observing only the order of observations. That is: **greater than** , **equal** or **less than** . Specifically, we investigate whether there is a * monotonicity * relationship between them.

For the (sigmoid) relationship, between x and y below:

```
>set.seed(2600)
>sig_data <- data.frame(y_vals = -(1 / (1 + exp(seq(-10,10,by =0.3) )*100 ) ),
                        x_vals = 1:67)
>ggplot(sig_data,aes(x=x_vals,y=y_vals))+
geom_point()+theme_economist()+xlab("")+ylab("")
```



Pearson's coefficient is $\rho \sim 0.850$[3] :

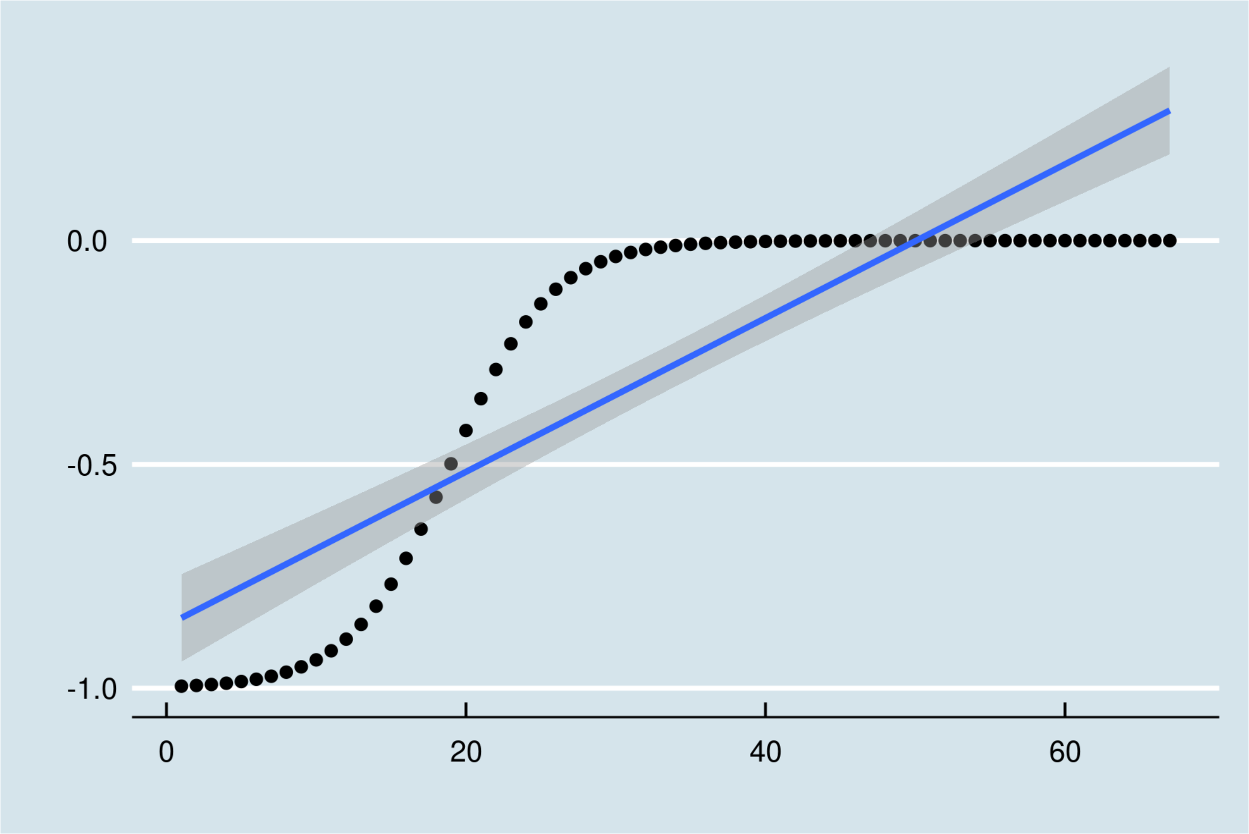
```
>cor.test(sig_data$y_vals,
+          sig_data$x_vals)

    Pearson's product-moment
    correlation

data:  sig_data$y_vals and +sig_data$x_vals
t = 12.993, df = 65, p-value <
2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7658181 0.9051711
sample estimates:
      cor
0.8497162

>ggplot(sig_data,aes(x=x_vals,y=y_vals))+
  geom_point()+ geom_smooth(method="lm")+
  theme_economist()+xlab("")+ylab("")
```

[3] As noted in the graph, the linear correlation is not that high. The coefficient approaches 1 $\rho \sim 0.850$ because the upper deviations symmetrically compensate for the lower ones. The example reinforces the importance of plotting the data for better understanding (see Anscombe Quartet).



Since the relationship is perfectly monotonic, the ordered pairs (x_i, y_i) always have the same rank. The fifth highest value in x is also the fifth highest value in y. Therefore, Spearman's coefficient is 1:

```
>cor.test(sig_data$y_vals,
+          sig_data$x_vals,method = "spearman")

    Spearman's rank correlation rho

data:  sig_data$y_vals and sig_data$x_vals
S = 0, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
  1
```

The coefficient ρ Spearman's is preferable when the measurements appear to differ greatly in terms of the family of the distribution of origin. Especially, when the average does not seem to correspond well to the center of the distributions. Remember that Pearson's coefficient is based on deviations from the mean in both samples.

Mann-Whitney U test

The Mann-Whitney U test uses U statistics to make inferences. The rationale is identical to the Student's t test. We establish null hypothesis $H_0$and alternative hypothesis H_1.
Then, we calculate the probability that our observations will happen if the null hypothesis is true. This time, we will use the U statistic. Remember that the t statistic was calculated based on parameters extracted from the sample:

$$t = Z/s = (\mu' - \mu)/\frac{\sigma}{\sqrt{n}}$$

The U statistic does not depend on parameters (e.g. μ, σ), being calculated based on each observation.

First, we calculate the ranks for each measure r_i joining the observations of samples A and B, of sample sizes n_a and n_b in just one set ($N_{tot} = n_a + n_b$).

Then, we separate the samples again and calculate the sum of the ranks in each group, called R_a and R_b. The U statistic is given by the following expression:

$$U_a = R_a - \frac{n_a(n_a + 1)}{2}$$
$$U_b = R_b - \frac{n_b(n_b + 1)}{2}$$

We use the smallest value of U to query the corresponding probability (p-value) for the null hypothesis.

The term $\frac{n(n+1)}{2}$ corresponds to the minimum sum of ranks for the sample. Ranks are a regular sequence $(1, 2, 3, ...)$, so that the sum of all values is identical to the sum of an arithmetic progression of N terms.

$$\Sigma_{ranks} = \frac{N(N + 1)}{2}$$

While R_i corresponds to the sum of the ranks calculated with the two samples, the term above would correspond to the minimum sum of the ranks for a sample, if the ranks occupied the initial sequence $A = (1, 2, 3, 4, ..., n_a)$in the joint sample.

The definition for the test is not unanimous in the literature, so that some authors and software (e.g. R) implement the calculation with the above subtraction and others (e.g. S-PLUS) do not. In R, the functions **dwilcox (x, m, n)** and **pwilcox (q, m, n)** return the cumulative distribution and density for the U statistic corresponding to samples with sizes m and n. **wilcox.test (x, y, . . .)** is the basic implementation of the Mann Whitney test. The Mann Whitney test is the Wilcoxon test for two samples.

Exercises

1. Pearson's product-moment coefficient describes which types of relationship?
    - Is it useful for modeling quadratic relationships between variables?
    - We cite non-linear relationships, such as $E = mc^2$.Cite another example of a natural phenomenon with a non-linear profile where the ρ Pearson's does not work.
2. Create a function that calculates the nth moment for a sample:

- `n_moment <- function(x,n) {sum((x- mean(x))^n)/length(x)}`
- Calculate the skewness value. As mentioned in the chapter, it is the 3rd moment normalized [by the 2nd moment to the exponent 3/2].

$$\frac{\mu_3}{\mu_2^{3/2}}$$

- Calculate the value of kurtosis. As mentioned, it is the 4th standardized moment [by the square of the 2nd moment minus 3].

$$\frac{\mu_4}{\mu_2^2 - 3}$$

- Values can be checked with implementations `e1071::skewness` and `e1071::kurtosis`

3. Using the * iris * dataset, compare the 4 numerical variables (*Sepal / Petal Lenght / Width*) between species (*Species*) using Student's t test and U Mann Whitney test. In any case, do the methods differ regarding the rejection of the null hypothesis?
    - Get the effect size (Cohen's D) for the differences.
4. Using the * iris * dataset:
    - Make a scatterplot between two measurements. The pairs function can help.
        - Check for significant linear correlation between variables.
        - If present, adjust a linear regression model.
        - Adjust a regression model for each species.
        - Note the values of R^2 for each model. What is your impression of the performance changes?