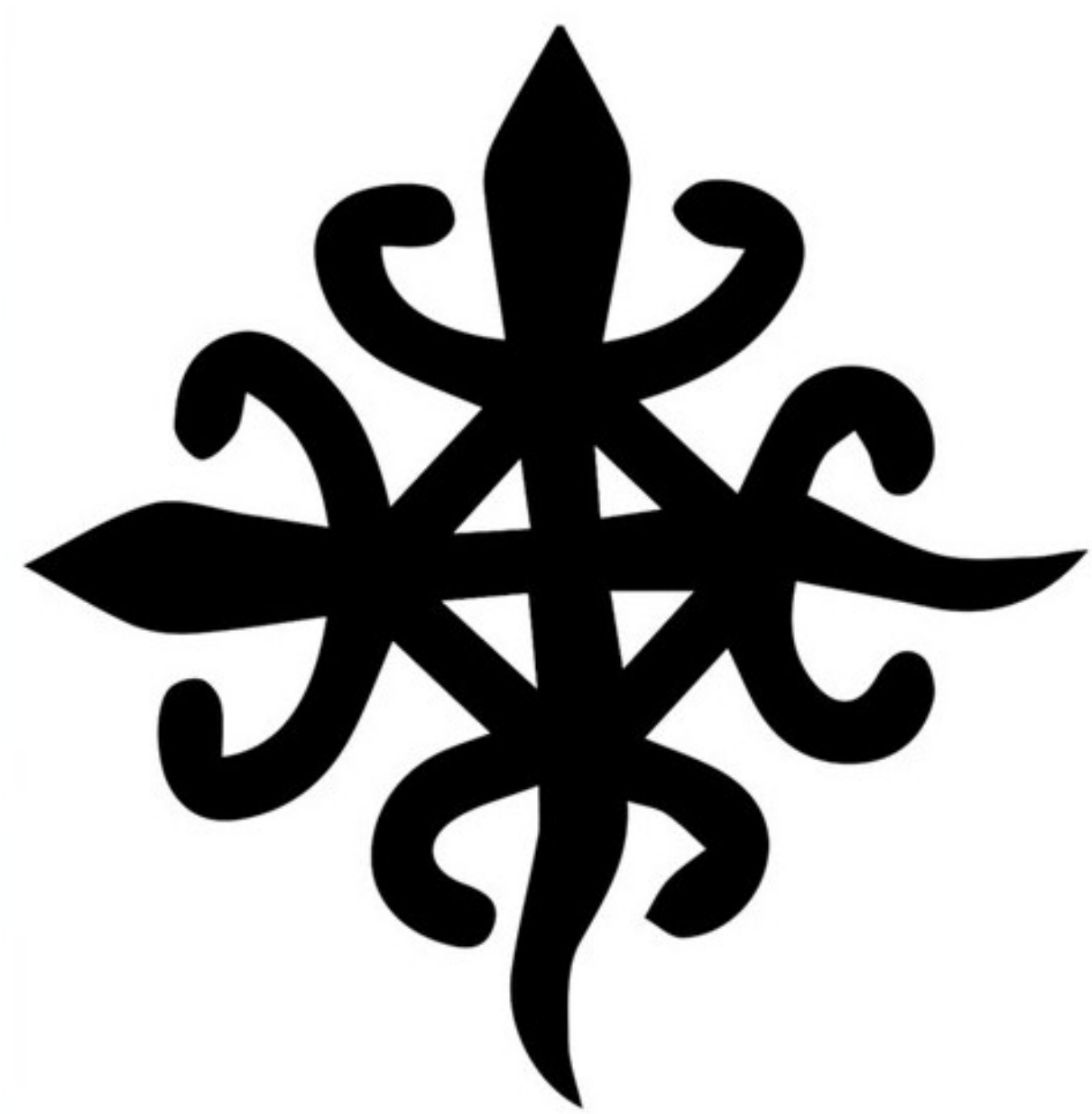


Chapter 3: About associations

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.

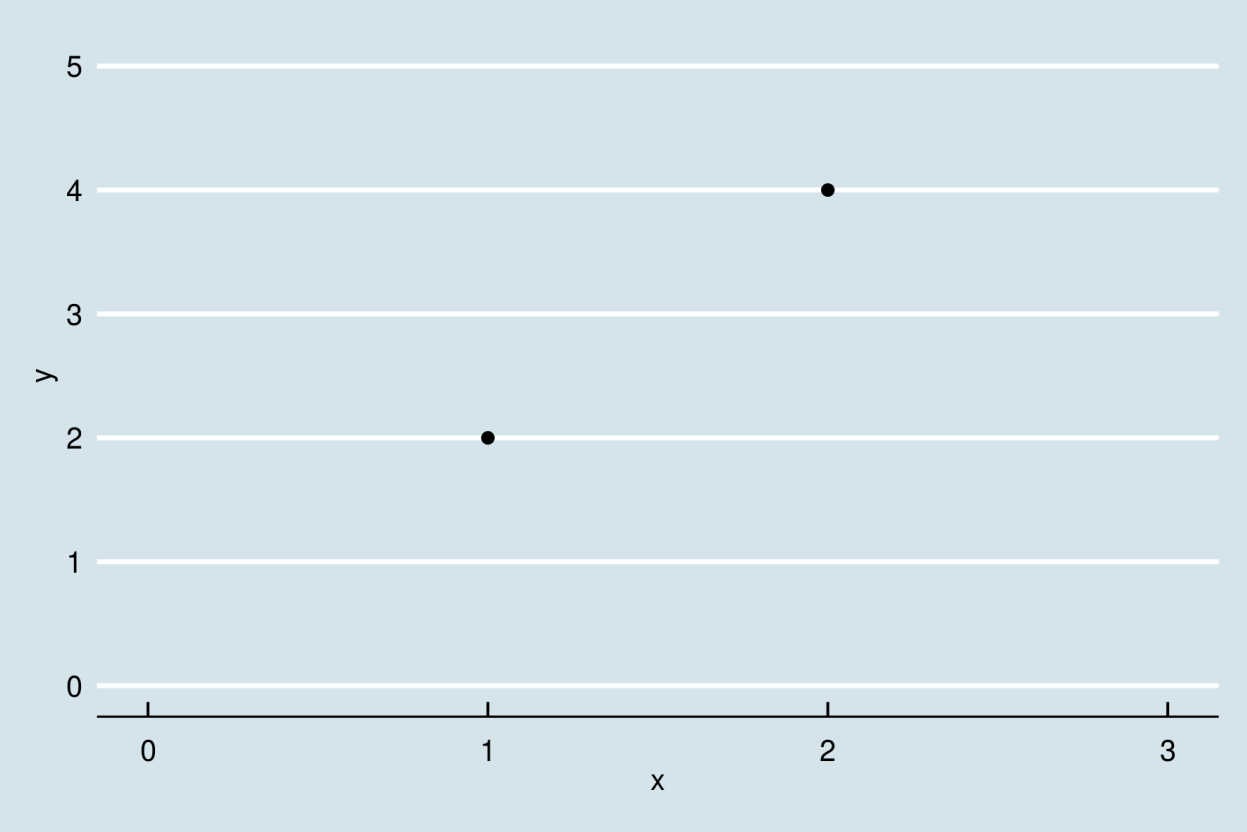
That is, higher mass values correspond to a proportional increase in energy. The value of c^2 expresses this constant proportion.

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.

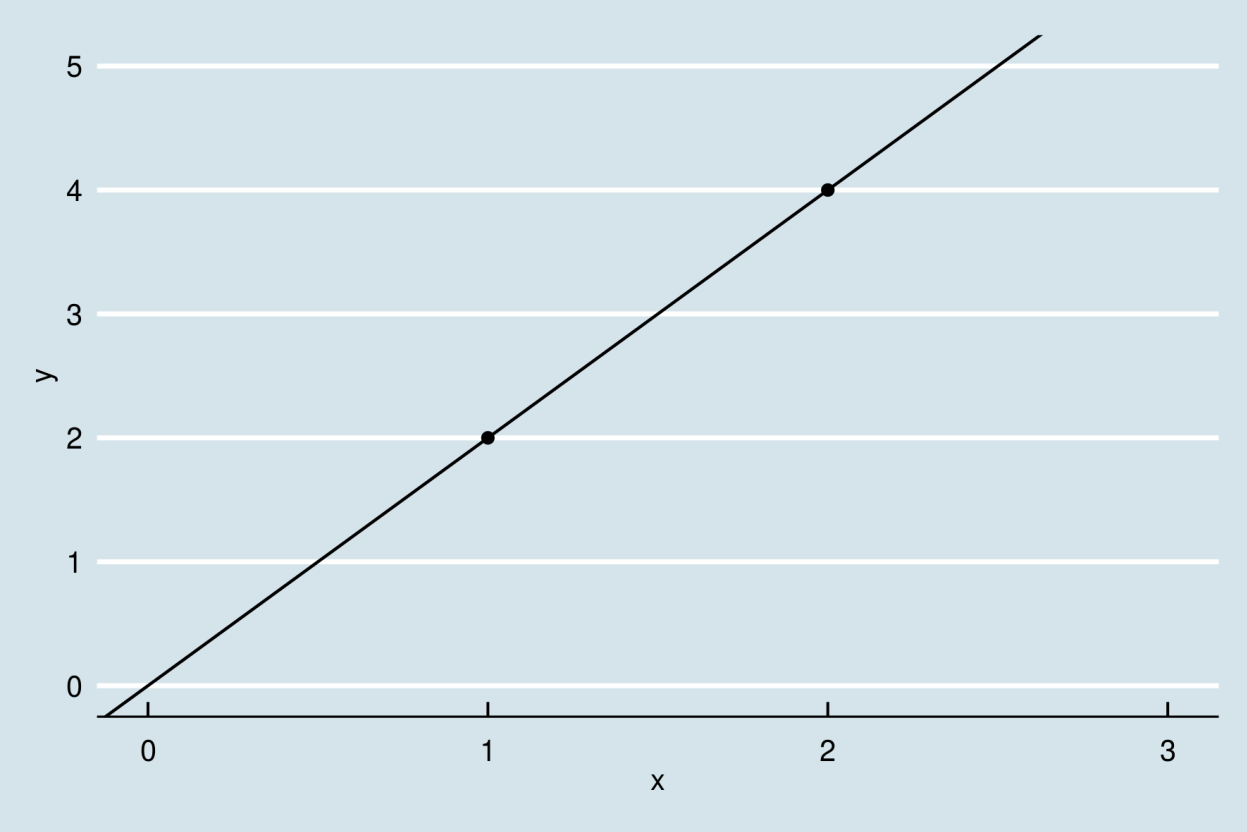
In a perfect scenario, if we know that there is a linear relationship between variables, we need only two observations to find out the proportion between them. This problem is identical to that of finding the slope of the line that passes through two points. It is easy to solve using elementary techniques.

```
>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

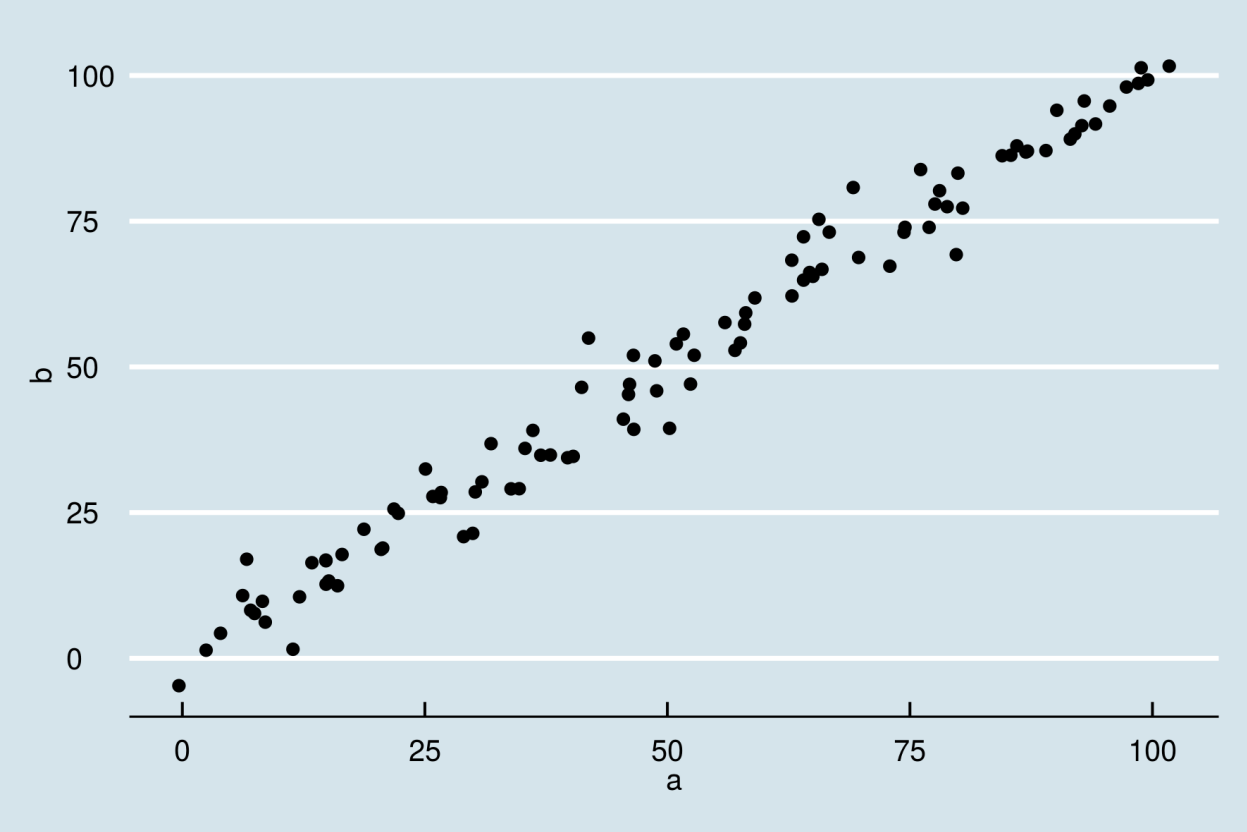
Controlling experimental factors, the relationships described are quite accurate. In a scenario without friction with surfaces and air, the measurement errors obtained with $\vec{F} = m\vec{a}$ are very low. However, this is not always true. First, we may experience interference from unknown variables.

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 [^ 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.

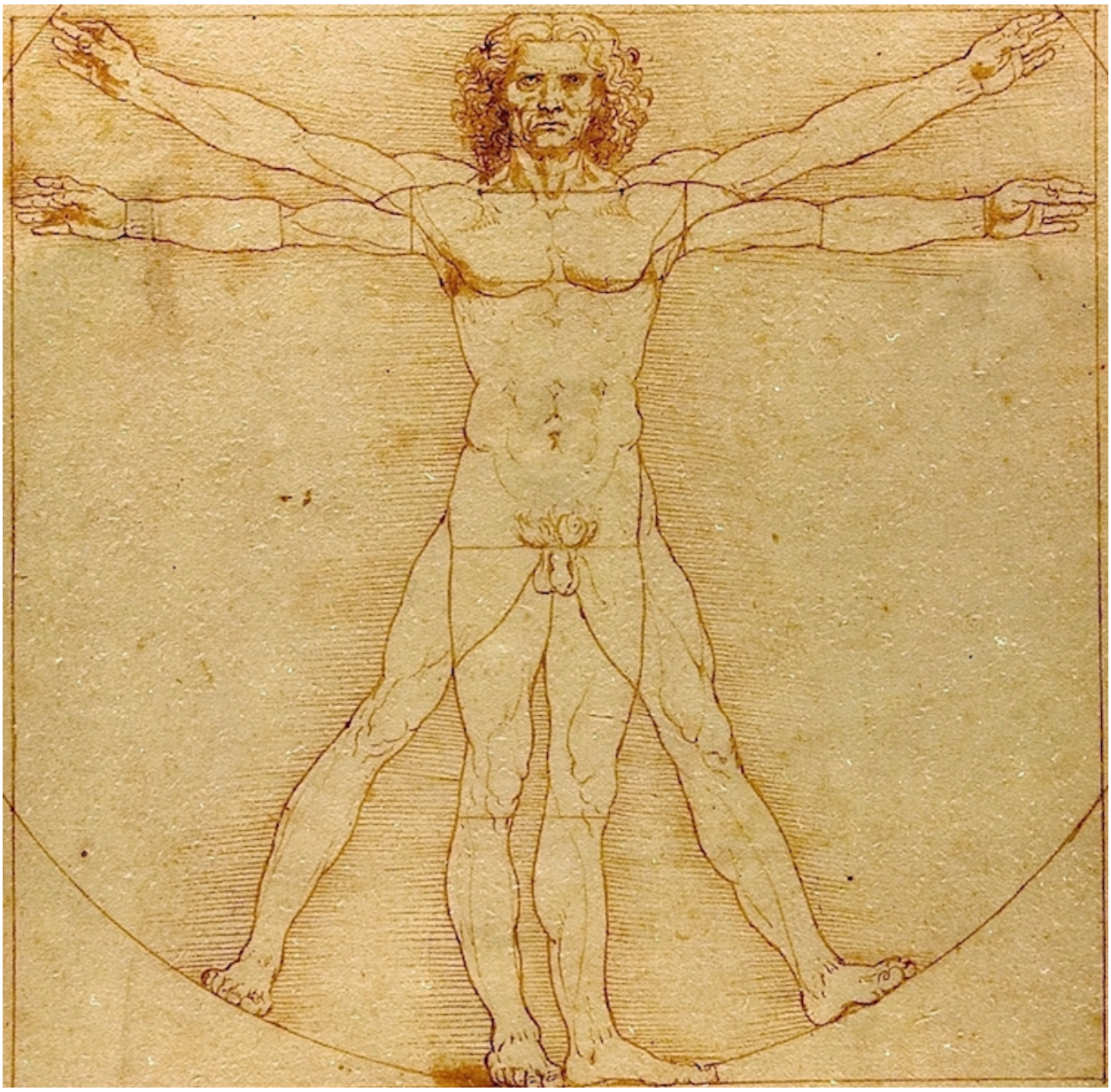
With these tools, we can extend our inferences. In addition to comparisons, we will have notions about the magnitude of a relationship, as well as we can predict the expected value for new observations.

[^ 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. **Antropologia:** Se recuperássemos de um túmulo antigo apenas um osso da coxa (fêmur) de um indivíduo, o que poderíamos dizer sobre sua altura?
2. **Ciência forense:** Com o intuito de identificar criminosos, o que pode ser dito sobre medidas diferentes de uma mesma pessoa?

Galton percebeu que, na verdade, estava lidando com o mesmo problema. Dadas medidas pareadas, (x_i, x_i'), o que o desvio de x_i informa sobre o desvio de x_i'?

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

O fêmur recuperado do esqueleto de um faraó é 5 cm maior que a média. Quão distante da média esperamos que seja sua altura? Ingenuamente, podemos pensar que se uma das medidas é 1% maior que a média, a outra também será 1% maior. Galton percebeu que havia um armadilha nesse pensamento.

Apesar de haver uma relação entre as medidas, há também flutuações aleatórias: parte do desvio é resultante disso. Precisamos entender o grau de correlação pra fazer um bom palpite.

Então, propôs um coeficiente mensurando a relação entre desvios de variáveis. Se tamanho do fêmur e altura estão muito relacionadas, um fêmur grande sugere indivíduo igualmente alto. Caso contrário (baixa correlação), um fêmur grande (desvio alto) não implica grande estatura.

Para quantificar a relação, multiplicamos os desvios de cada par de medidas:

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

A fórmula acima expressa a **covariância** entre X e X' e será útil em outros contextos. A expressão lembra o cálculo do primeiro momento, porém cada desvio é multiplicado pelo desvio correspondente da medida pareada. Daí o nome coeficiente de correlação *produto-momento*.

Note que, se ambos os desvios concordam em sentido (sinal), o resultado da multiplicaçao será positivo. Pares consistentemente concordantes aumentam o valor da soma final. Se ambos os desvios discordam em sentido (sinal), o resultado será negativo. Pares consistentemente discordantes diminuem o valor da soma final.

Assim, podemos ter variáveis altamente correlacionadas positiva ou negativamente, desde que o sentido da associação seja constante. Em contrapartida, se as medidas são ora discordantes e ora concordantes, os valores tendem a se anular na soma e o resultado se aproxima de zero.

Observar apenas a covariância é perigoso, pois os valores dependem da unidade de medida e da dispersão dos dados.

Calculamos o coeficiente de correlação de Pearson, normalizando[2] a covariância ao dividí-la pelo produto dos desvios-padrão:

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

De forma extensa:

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

Uma boa notícia: ρ segue uma distribuição conhecida, a distribuição t, com n-2 graus de liberdade. Podemos usar as ferramentas anteriores para testar hipóteses.

Exemplo prático

O exemplo a seguir foi um feliz achado. Na época, o governo brasileiro discutia a necessidade da ampliar número de médicos para melhorar a assistência à saúde. Alguns defendiam ser uma decisão acertada, enquanto outros advogavam que os investimentos deveriam ser feitos em outras áreas da saúde.

Por curiosidade, acessei os dados da WHO (World Health Organization) e do banco mundial (World Bank) sobre quantidade de médicos por país e indicadores de saúde. Minha expectativa era encontrar pelo menos uma tímida relação entre indicadores. Mais do que isso, entender qual a localização do Brasil em relação a outros países. Fui surpreendido por uma forte correlação, que exploraremos a seguir.

[2] Aqui, normalização tem o sentido de ajustar a escala das medidas. Não confundir com transformações para que os dados passem a ter distribuição gaussiana.

Adotamos países como unidade observacional com medidas x, o número de médicos 1,000 habitantes, e y, a expectativa de vida saudável ao nascer.
Usando dados obtidos dos portais da WHO e do World Bank, plotamos os pontos no plano cartesiano.

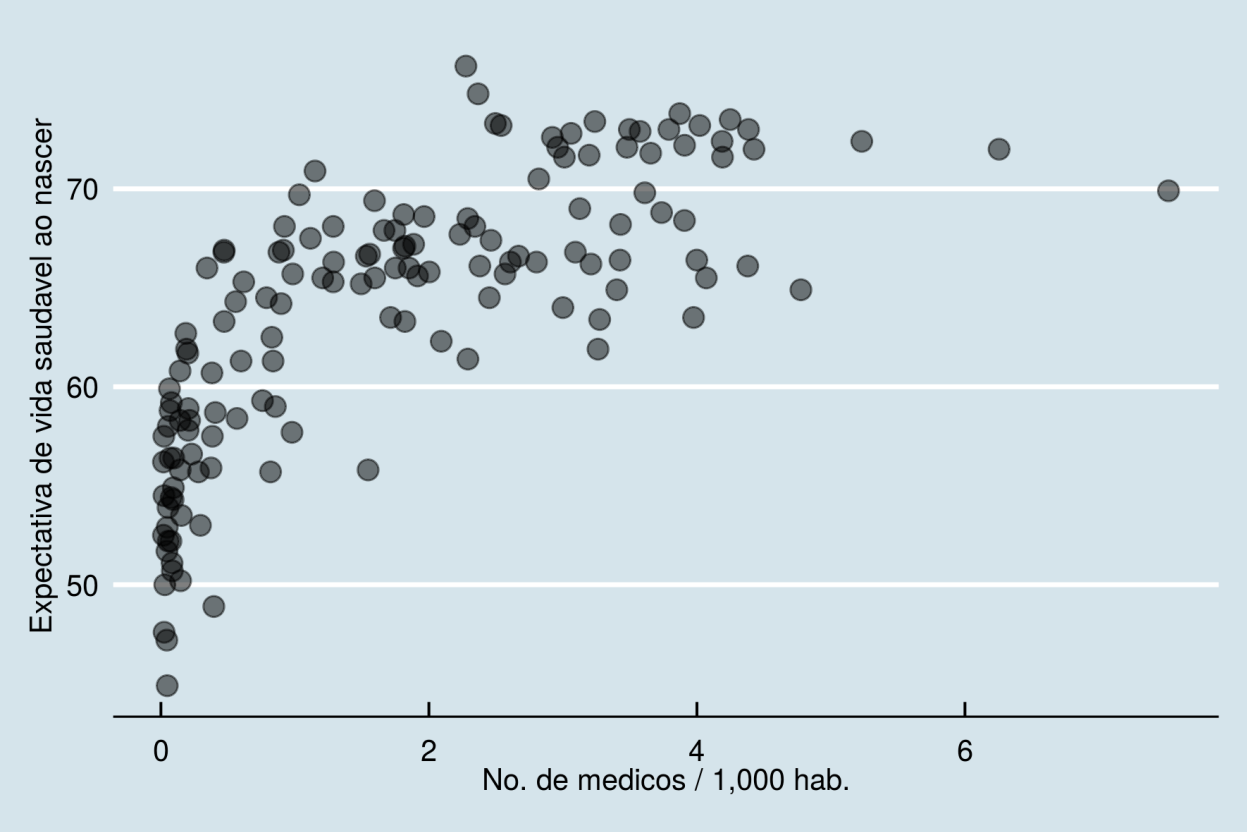
```
# http://apps.who.int/gho/data/view.main.HALEXv
# https://data.worldbank.org/indicator/SH.MED.PHYS.ZS
>library(magrittr)
>library(ggplot2)
>library(dplyr)

>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)) %>% #ultimos valores não nulos
  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. de medicos / 1,000 hab.")+
  ylab("Expectativa de vida saudavel ao nascer")+
  theme_economist()
```



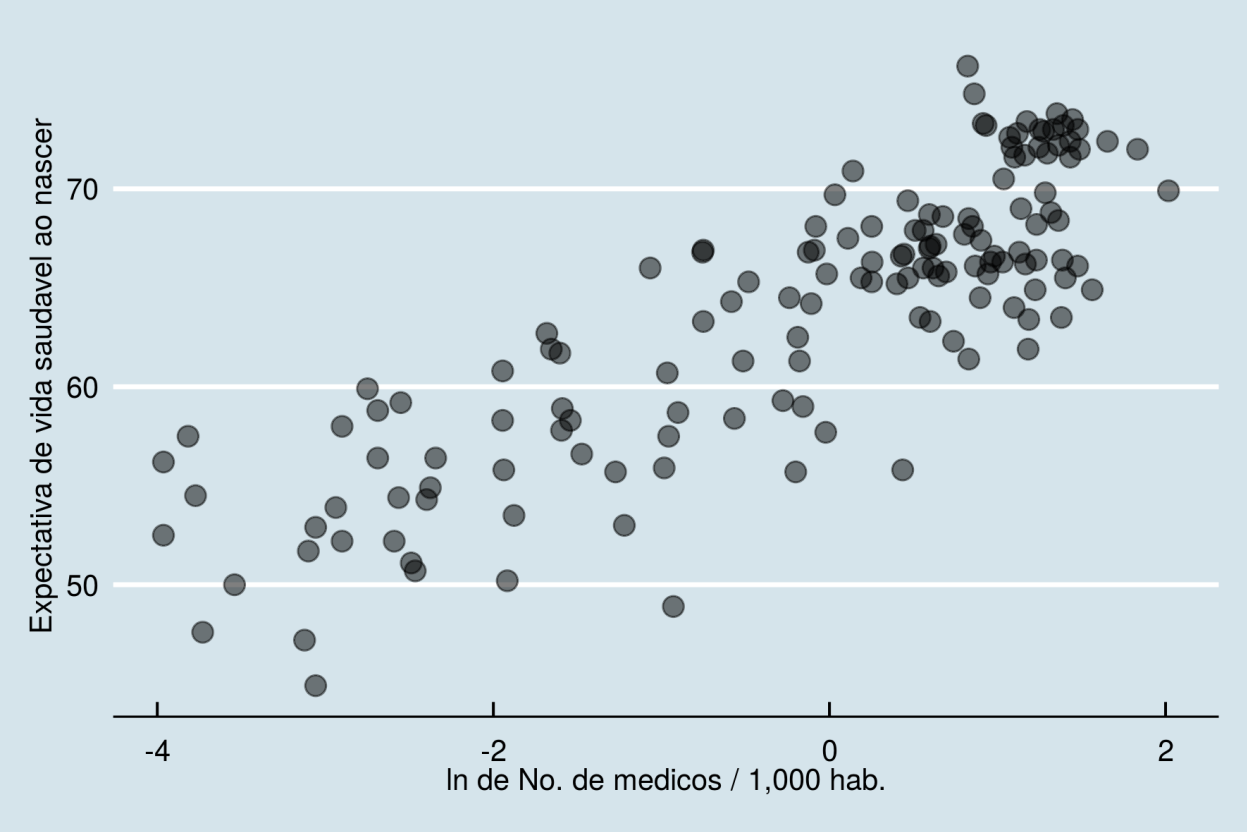
É evidente que o padrão não é aleatório. Visualmente, notamos que o valor da expectativa de vida aumenta com maior Nº de médicos. Ainda, notamos um aumento inicialmente rápido até atingir um platô. O padrão é semelhante ao de uma curva logarítmica.

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

Se essa hipótese for verdade, transformar o número de médicos usando função logaritmica tornará a relação linear com a variável transformada:
Se $y = log(x)$, fazemos a substituição $x' = log(x)$ para obtermos $y = x'$.

Então a expectativa de vida se torna linearmente correlacionada ao logaritmo do número de médicos.

```
>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 de No. de medicos / 1,000 hab.")+
  ylab("Expectativa de vida saudavel ao nascer")+
  theme_economist()
```



De fato, verificamos uma notável tendência linear para os pontos.

Usando a implementação nativa em R para o coeficiente de Pearson:

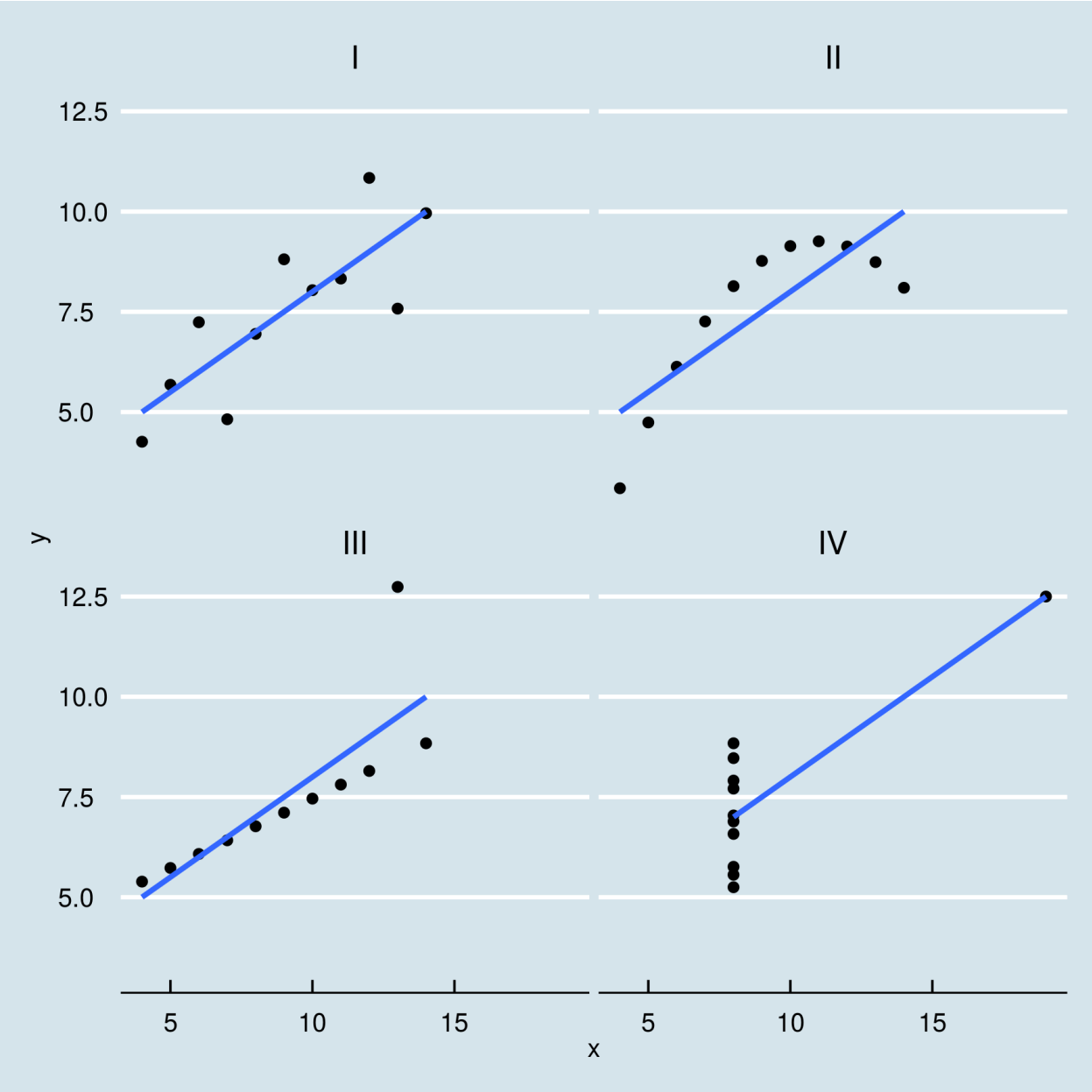
```
>cor.test(uni_df$log_docs,uni_df$hale)
Pearson's product-moment correlation
data:  uni_df$log_docs and uni_df$hale
t = 18.572, df = 143, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.7854248 0.8828027
sample estimates:
    cor
0.8407869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

A correlação linear obtida para nossa amostra de países é surpreendentemente grande, como sugeria a visualização ($\rho \sim 0.841$).

O valor p é baixo ($p < 0.001$) considerando a hipótese nula H_0 de $\rho = 0$. Concluímos então que há uma relação linear significativa de forte magnitude entre o logaritmo do número de médicos e a expectativa de vida dos países em nossa amostra.

É realmente curioso que exista uma relação matemática tão evidente entre construtos tenuamente conectados. O tempo médio que um organismo leva entre nascimento e morte e o número de profissionais atuantes. É virtualmente impossível explicitar cada relação causal por trás dessa relação, que se manifesta de forma robusta através da soma de muitos fatores relacionados.

Nota *É costumaz afirmar que não existe relação entre variáveis caso o coeficiente de relação não se mostre importante. Como vimos, esse indicador informa apenas sobre relações lineares entre variáveis. A visualização dos dados pode ser de grande ajuda na inferência sobre a natureza de relações.*
Dados com distribuições bastante diferentes podem resultar em coeficientes iguais, como mostra o clássico quarteto de Anscombe. As 4 amostras abaixo apresentam o mesmo coeficiente de correlação.

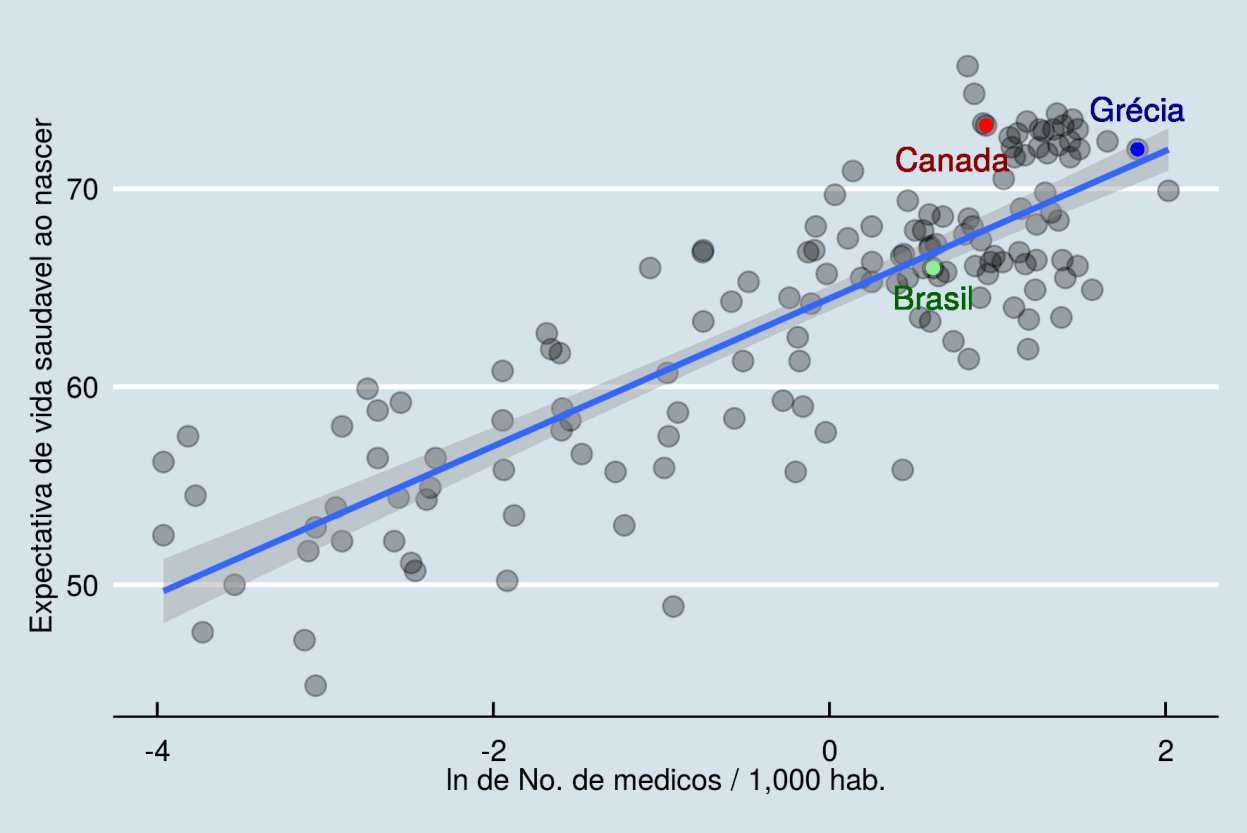


Previsões

Agora, sabemos que é razoável assumir uma relação linear entre essas variáveis. Como dito antes, podemos então encontrar a reta que minimiza a distância para as observações.

A equação que descreve essa reta nos informa o valor esperado para expectativa de vida dado o número de médicos.

```
>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="Brasil",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="Grécia",color="dark blue")+
   xlab("ln de No. de medicos / 1,000 hab.")+
   ylab("Expectativa de vida saudavel ao nascer")+
   theme_economist()
```



Vieses devem ser enderaçados antes de conclusões, mas o modelo é suficientemente interpretável para tomar decisões.
Uma boa política pode comparar o valor de investimento por setores com outros países em condições semelhantes e resultados diferentes.
Assumindo que realmente há uma relação linear, vemos que o Brasil está bastante próximo do esperado para o número de médicos[3]. Caso a estratégia seja contratar mais pessoas, podemos nos espelhar em programas de países com mais médicos por habitante e resultados positivos (e.g. Grécia).
Se a estratégia for economizar com a folha de pagamentos e priorizar investimento em estrutura, podemos usar países com expectativa de vida alta para o número de profissionais esperado (e.g. Canada).

[3]É praticamente consenso entre especialistas que o Brasil possui problema de distribuição de profissionais, com déficit de médicos em áreas mais pobres e pouco populosas.

Predições com modelos lineares

Como adivinhar uma medida com base na outra? Considerando a relação linear descoberta anteriormente, podemos criar uma função que receba como input o valor de uma variável (número de médicos) e retorne como output o valor esperado para a expectativa de vida.

Descobrir a equação que descreve esta função consiste em encontrar a reta que melhor se ajusta à nuvem de pontos, como na figura anterior.

Para isso, calculamos a inclinação (β_1) e o ajuste vertical (β_0) que minimizam a soma das distâncias entre a reta e as observações. O termo ϵ corresponde aos erros, com distribuição normal de média 0 e desvio padrão σ.

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

Ajustamos o modelo usando a função lm(linear model) do R:

```
# 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
```

Temos $\beta_0 \sim 64.46$ e $\beta_1 \sim 3.73$.
Nossa estimativa para a expectativa de vida saudável "começa" em 64.46 anos e aumenta com o número de médicos no país. Especificamente, aumenta em 3.73 para cada unidade de nossa variável transformada ($log(x)$).
Em nosso dataset, o Brasil possui 1.852 médicos/1,000 hab. Nossa predição então é:
$\hat{y}_{Brasil} = log1.852 * 3.73 + 64.46 \sim 66.8$, o que está bastante próximo do número real(66).

Estimadores

Existe mais de uma maneira de estimar esses parâmetros.
Uma de particular interesse, que também servirá em outros contextos, é a de Maximum likelihood (máxima verossimilhança).

Primeiro, determinamos uma função que descreve a probabilidade da observação na variável alvo (y_i) ocorrer dadas medidas das variáveis preditoras (x_i) e um conjunto de parâmetros (β_k).

Podemos adotar como função de verossimilhança *(likelihood function)* para os valores y_i uma distribuição de probabilidades gaussiana cuja média é dada pela reta $\mu_{yi} = \beta_0 + \beta_1 * x_i$. Assim, a probabilidade de cada valor y_i é dada por uma gaussiana, de acordo com o desvio para o valor previsto pela reta.

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

.

Assumindo que as observações são independentes, a probabilidade do conjunto de observações é dada pelo produto delas.

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

Subsitituindo os valores de μ para a gaussiana pelas previsões da reta:

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

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

Essa é nossa função de verossimilhança e expressa a probabilidade de observarmos as medidas y_i dadas as medidas x_i e considerando um conjunto de parâmetros (β_0, β_1).

O objetivo então é encontrar parâmetros que maximizem essa função. Por conveniência, aplicamos uma transformação logaritmica nesta função ($log \quad likelihood \quad function$). Isso transforma nosso produtório em um somatório e passamos o contradomínio do intervalo $[0; 1]$ para $[-\infty, 0)$.

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

Os parâmetros que maximizam a função de verossimilhança (max. likelihood, ML) são os mesmos que maximizam a o logaritmo da função de verossimilhança (log-likelihood).

Introduzimos o racional do estimador ML pois ele será útil futuramente. Em verdade, é fácil entender as fórmulas fechadas para nossos parâmetros, pois apenas expressam as relações lineares exploradas [4]:

$\hat{\beta}_1$ expressa a magnitude da correlação entre X e Y. É natural que seu valor seja a covariância normalizada pela variância do preditor.

$$\hat{\beta}_1 = \frac{cov(XY)}{\sigma_x^2}$$

$\hat{\beta}_0$ é nosso intercepto, então é a diferença entre médias preditas e predições considerando o valor médio em X.

$$\hat{\beta}_0 = \mu_y - \hat{\beta}_1 \mu_x$$

Por fim, a variância dos erros $\hat{\sigma^2}$ é dada pelo quadrado dos desvios das predições em relação às medidas.

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

As soluções acima fornecem as melhores estimativas que podemos obter minimizando a distância da reta aos pontos.
Devemos então nos preocupar em saber se o modelo linear encontrado é bom na predição dos dados.

[4] 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

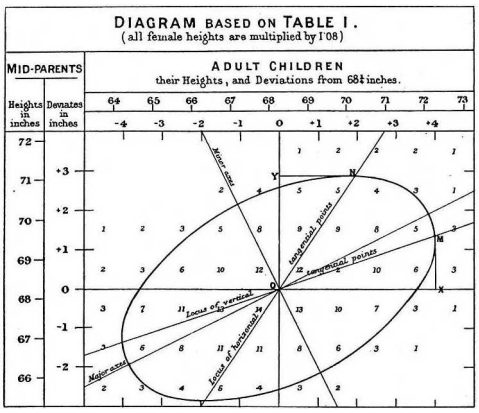


Figure 1: O primeiro gráfico de regresão linear. Ilustração de Francis Galton (1875) relação entre altura de pais e filhos.

Avaliando performance Existem diferentes parâmetros para avaliar a performance de um modelo. Em geral, eles buscam quantificar o quanto os resultados do modelo se distanciam de resultados ideais.

Para regressão linear, o R^2 (coeficiente de determinação) é um coeficiente bastante usado. Expressa a proporção entre **(1)** variância explicada pelo modelo e **(2)** variação total. Chamamos de resíduo(ou erro) a diferença entre valores preditos e valores reais.

(1) Para capturar a magnitude dos erros do modelo, somamos o quadrado de todos os resíduos *(sum of squared residuals, SSR)* em relação aos valores preditos. Sejam y_i as observações e $\hat{y}_i$ as predições:

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

(2) A variabilidade total é quantificada pela soma do quadrado dos desvios em relação à média *(total sum of squares, TSS)*, um termo que vimos no cálculo da variância (segundo momento):

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

Então a fração $\frac{SSR}{TSS}$ é a proporção desejada. Definimos R^2 como:

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

Uma visualização intuitiva de SSR e 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("Expectativa de vida saudavel ao nascer")+
    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 de No. de medicos / 1,000 hab.")+
    ylab("Expectativa de vida saudavel ao nascer")+
    ggplot2::ggtitle("TSS")+theme_economist()

>multiplot(ssr_res,tss_res)
```

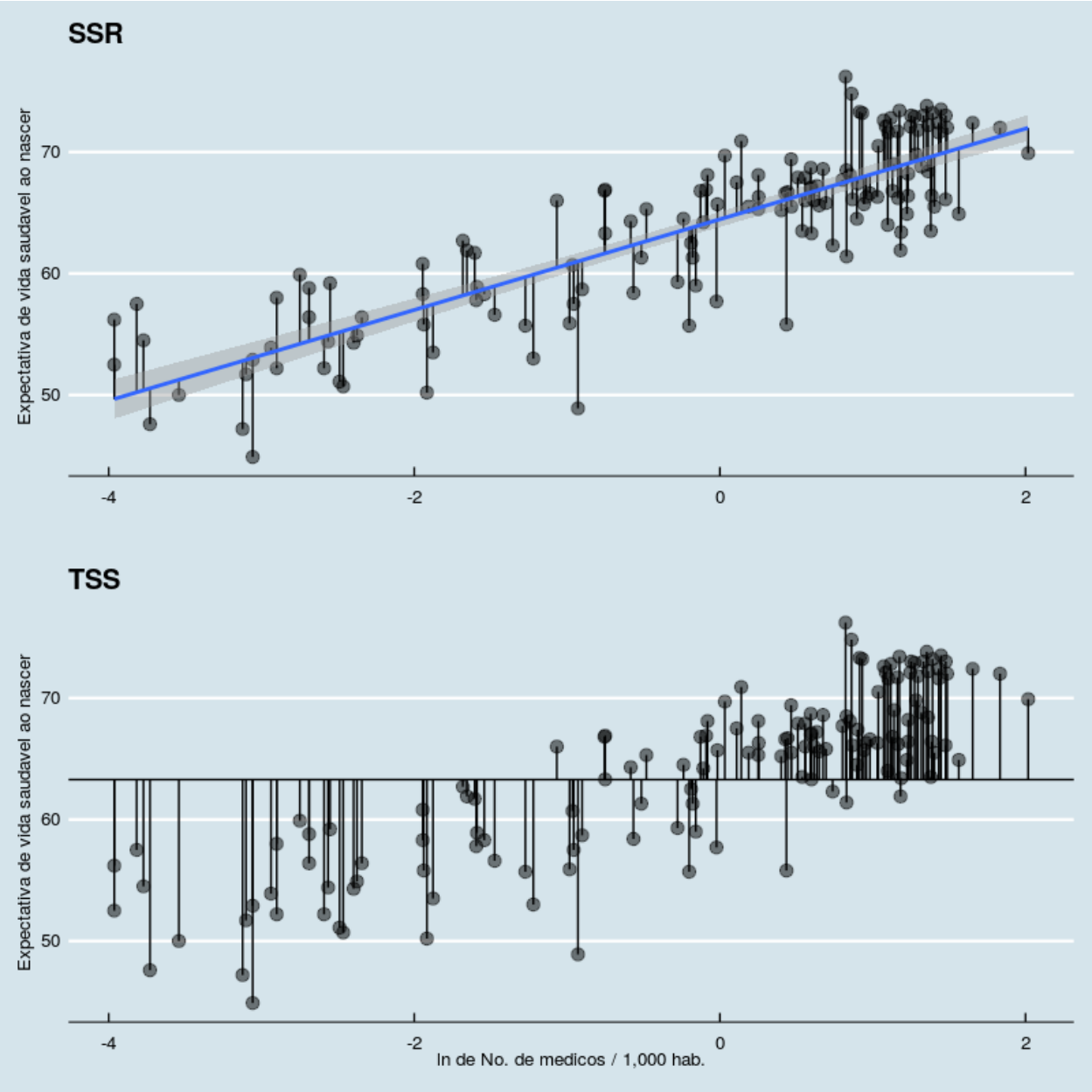


Figure 2: O quadrado da distância entre um ponto e a reta corresponde a um resíduo. Obtemos SSR e TSS somando todos os resíduos nas figuras superior e inferior, respectivamente.

Valores de R^2 próximos a 1 indicam soma de resíduos (SSR) similar a 0. Usar a reta como guia acumula erros quase nulos. Valores de R^2 próximos a 0 indicam $\frac{SSR}{TSS} \sim 1$ e as predições obtidas pelo modelo são tão boas quanto chutar a média para todos os casos.

```
>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
```

Para obter os valores preditos, usamos o método *predict*:

```
>head(predict(doc_lmfit))

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

É possível também obter predições para novos valores especificando o argumento *newdata*. Para um país com 1.5 médicos/1,000 habitantes:

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

Premissas Existem alguns procedimentos auxiliares para checar possíveis falhas e pontos no modelo que precisam de atenção. Por exemplo, os resíduos podem ser assimétricos. Isso indica que o desempenho muda em diferentes intervalos (heteroscedacidade). Diferentes violações necessitam de atitudes diferentes, como tratar outliers ou mudar tipo do modelo. Uma lista completa de premissas, junto aos códigos em R para testá-las, está disponível no material auxiliar (*lm-asssumptions.R*)

Correlações e testes não paramétricos

Verificamos minuciosamente análises envolvendo a distribuição normal, a distribuição t e relações lineares. Entretanto, muitas vezes as medidas não seguem uma distribuição definida. Assim, realizar inferências usando os **parâmetros** descritos ($\mu, \sigma, t...$) nos levaria a conclusões erradas.
Para lidar com distribuições arbitrárias, vamos abrir mão deles e conhecer ferramentas *não-paramétricas*: o coeficiente de correlação de ranks ρ de Spearman e o teste U de Mann Whitney.

Ranks e o ρ de Spearman

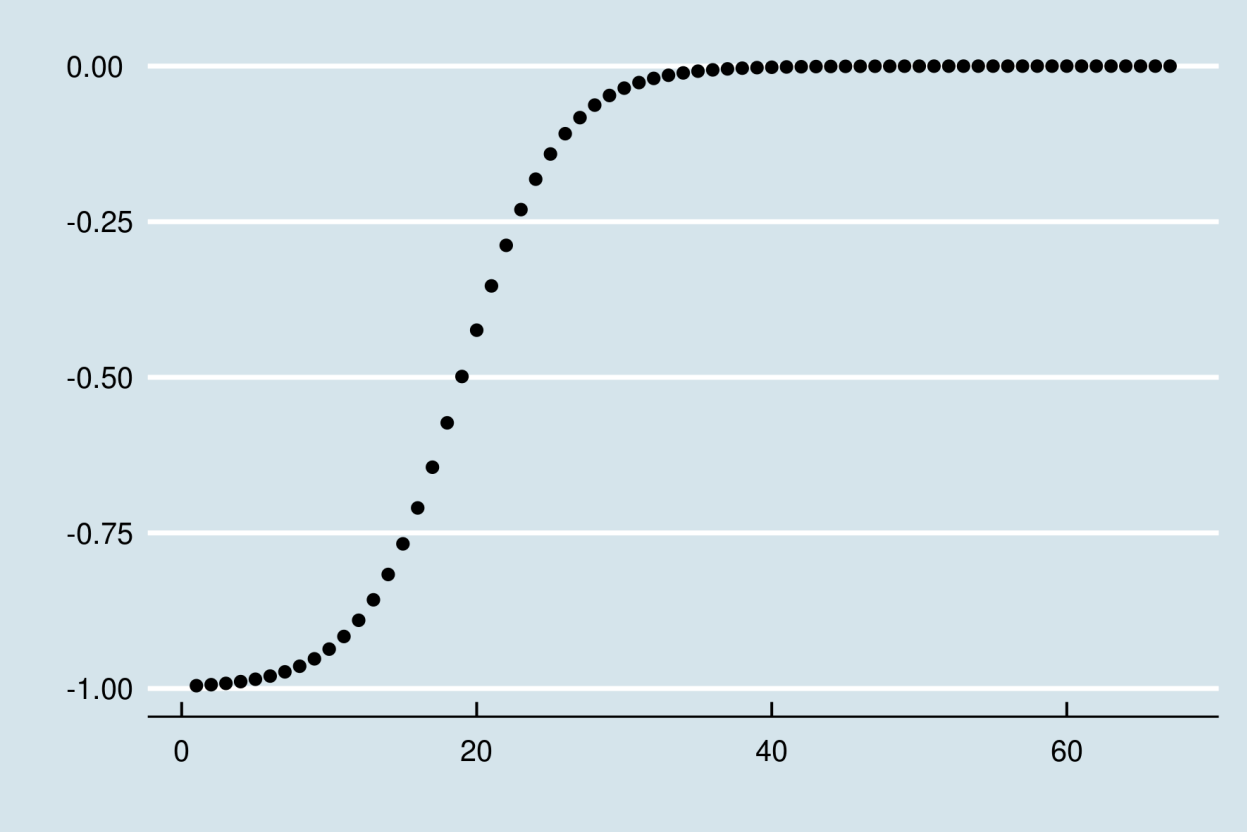
Relações lineares mantêm proporções constantes e aprendemos como quantificá-las. Por outro lado, duas variáveis podem ter relações de outros tipos, não lineares. Em especial, se as medidas apresentam valores muito extremos *(outliers)* um cálculo como o anterior sofre bastante com vieses.
Uma simples solução para esse problema é ranquear os valores. Assim, os itens do conjunto são tratados pela sua posição em relação a outros itens, de forma independente dos valores associados. Exemplo:

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

O ρ de Spearman é que o coeficiente produto-momento de Pearson aplicado aos ranks. Assim, medimos o grau em que duas variáveis aumentam (ou diminuem) em magnitude observando apenas a ordem das observações. Isto é: **maior que**, **igual** ou **menor que**. Especificamente, investigamos se há uma relação de *monotonicidade* entre elas.

Para a relação (sigmoide), entre x e y abaixo:

```
>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("")
```



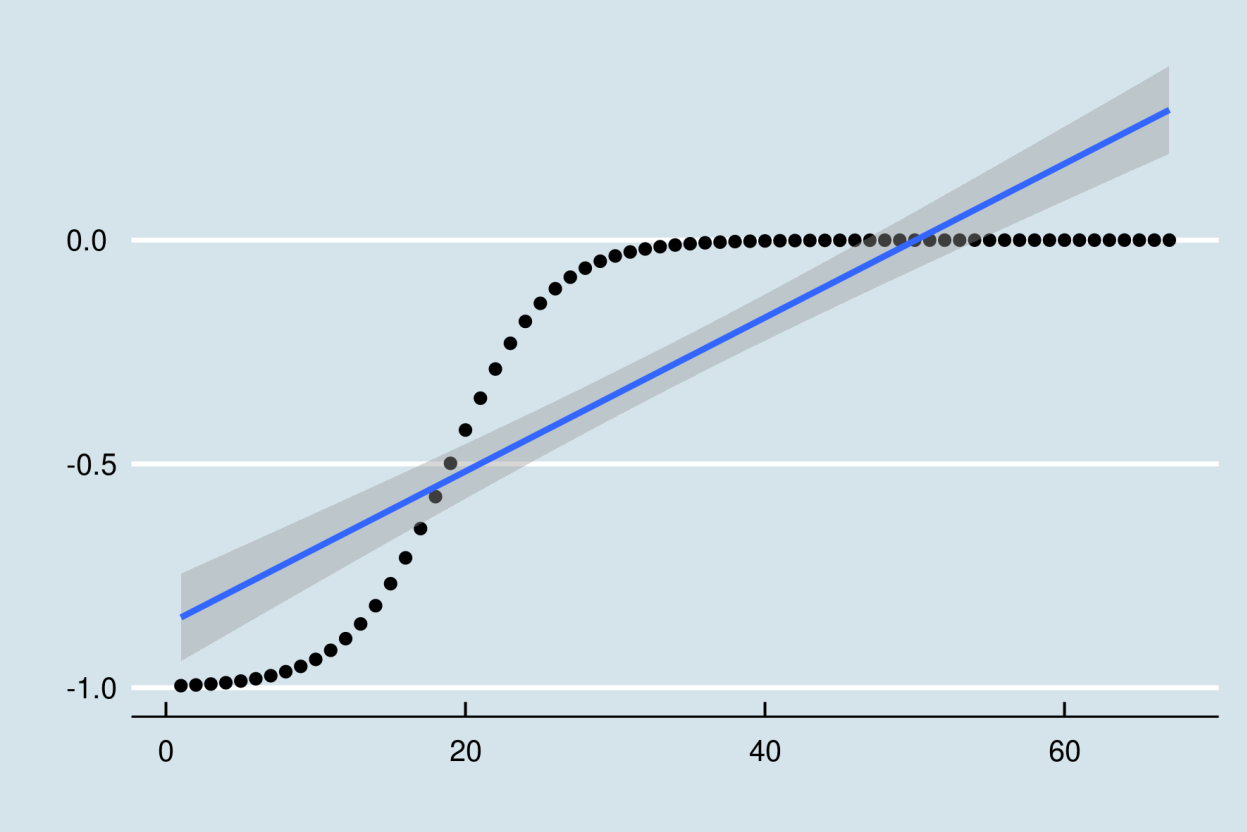
O coeficiente de Pearson é $\rho \sim 0.850$[ˆ20] :

```
>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("")
```



Como a relação é perfeitamente monotônica, os pares ordenados (x_i, y_i) sempre possuem o mesmo rank. O quinto valor mais alto em x é também o quinto valor mais alto em y. Portanto, o coeficiente de Spearman é 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.

[^ 20]: 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).

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?