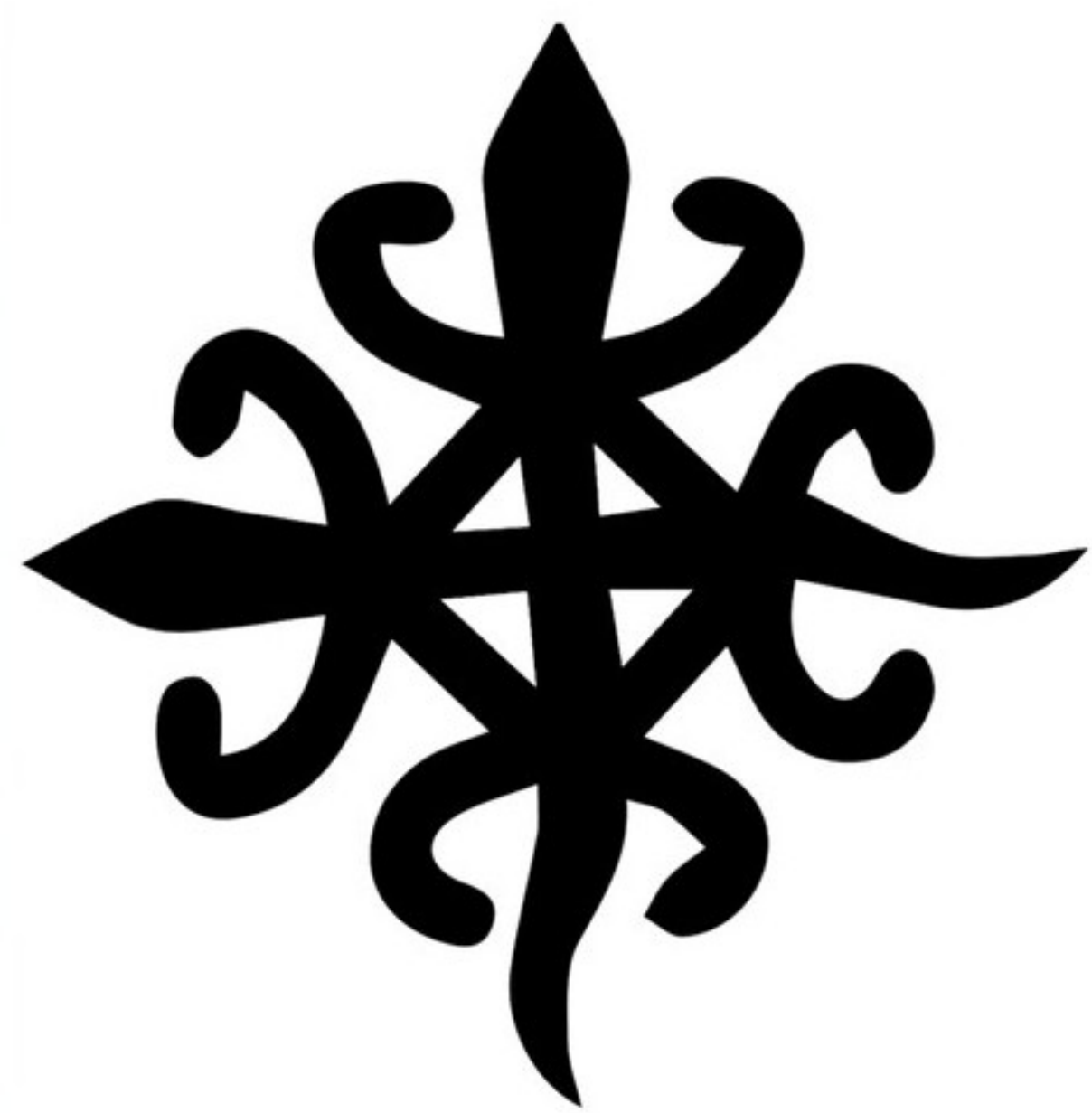


Capítulo 3 : Sobre associações

Prelúdio: *Hypotheses non fingo*?

Eu ainda não fui capaz de descobrir a razão para essas propriedades da gravidade, e não faço hipóteses. *Tudo aquilo que não é deduzido do fenômeno pode ser chamado de hipótese; e hipóteses, sejam metafísicas ou físicas, ou baseadas em qualidades ocultas, ou mecânicas, não têm lugar na filosofia experimental. Nesta filosofia, as proposições particulares são inferidas a partir do fenômeno, e então generalizadas por indução.*

O racional apresentado no capítulo anterior é diretamente relacionado ao método hipotético-dedutivo e seus princípios filosóficos. Apesar de adequado a este cenário, a interpretação do valor p não é muito intuitiva. Envolve *mensurar quão improváveis são as observações em um cenário hipotético na vigência da hipótese*

nula.
Sua tradução (errada) mais popular é de que representa *"a chance de o resultado deste estudo estar errado".*

O arcabouço descrito no capítulo anterior é suficiente para produzir um trabalho científico críptico para leigos.

Ao seguir receitas pré-definidas (formulação de H_0 e H_1, cálculo de estatísticas e valores p), um texto parece estar em conformação com os padrões acadêmicos, mesmo que a hipótese elementar em torno do objeto de pesquisa seja simplória. Assim, inadvertidamente, priorizamos a forma e relegamos a segundo plano o miolo de propostas científicas.

Outro efeito colateral é a busca por valores p que rejeitem H_0, desprezando precedentes teóricos e premissas probabilísticas (múltiplos testes).

A difícil interpretabilidade do valor p e as armadilhas frequentes envolvidas no processo de inferência levaram a comunidade científica a questionar a hegemonia desse parâmetro. Há uma presente tendência a abandonar o valor p e o limite $p < 0.05$ como critérios canônicos.

Vamos conhecer argumentos formais contra o método hipotético dedutivo nas ciências. Por enquanto, basta sabermos que é sempre vantajoso obter outras informações, complementares ou alternativas.

Neste capítulo, vamos aprender a estimar (1) a magnitude da diferença entre duas amostras e (2) quão relacionados são valores pareados (e.g. peso e altura).

Ainda não fui capaz de descobrir a razão dessas propriedades da gravidade dos fenômenos, e não finjo hipóteses. Pois tudo o que não é deduzido dos fenômenos deve ser chamado de hipótese; e as hipóteses, sejam metafísicas ou físicas, ou baseadas em qualidades ocultas ou mecânicas, não têm lugar na filosofia experimental. Nessa filosofia, proposições particulares são inferidas dos fenômenos e posteriormente tornadas gerais por indução. *Isaac Newton (1726). Philosophiae Naturalis Principia Mathematica, General Scholium. Terceira edição, página 943 da tradução de I. Bernard Cohen e Anne Whitman de 1999, University of California Press ISBN 0-520-08817-4, 974 páginas.*

Tamanho de efeito

O tamanho de efeito nos ajuda a expressar magnitudes.
Retomando o exemplo anterior, de que adianta uma diferença significativa entre o tamanho dos bicos dos pássaros, se ela for de 0.00001 mm?

Ainda, existem casos em que estudos pequenos sugerem efeitos importantes, porém o tamanho amostral não fornece poder estatístico suficiente para rejeição da hipótese nula.

Além de saber quão improvável é a diferença observada, é natural imaginarmos o quão grande ela é.

Uma medida bastante popular é o *D de Cohen (Cohen's D)*.

É um parâmetro que expressa a magnitude da diferença sem usar unidades de medida.
Uma torcedora de futebol conta (feliz) a um amigo que seu time favorito venceu com placar de 4×1 (gols).
Porém, esse amigo acompanha basquetebol e está acostumado a placares como 102×93 (cestas).
Como é possível comparar gols com cestas? Qual vitória representa pontuações mais discrepantes: 4×1 ou 102×93?

O problema aqui é que as pontuações se comportam de maneiras diferentes entre os esportes. Os placares no basquete possuem médias e dispersões muito maiores.
O D de Cohen consiste em expressar essa diferença em desvios-padrão. Bastante simples:

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

Usando a biblioteca *effects*, podemos calcular diretamente:

```
library(effects)
# O dataset galapagos_birds foi criado no capitulo 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 propôs algumas faixas para classificar a magnitude desses efeitos:

	Pequeno	Médio	Grande
Cohen's D	0-0.2	0.2-0.5	0.5 - 0.8

Assim, podemos atualizar nossos resultados anteriores, reportando também o tamanho de efeito da diferença e seu intervalo de confiança. Se as distribuições forem da mesma família, temos uma estimativa comparável entre contextos.

Correlações

Na empreitada científica, não nos atemos apenas a comparações. Um objetivo mais nobre é descrever exatamente como se dá a relação entre entidades estudadas.

Como sabemos, existem muitas classes de funções para expressar relações entre variáveis/conjuntos. Nos capítulos anteriores, usamos algumas funções, como $y = \sqrt{x}$ e $y = e^x$.

Diversas leis naturais tornaram-se particularmente conhecidas, como a relação entre força, massa e aceleração, elucidada por Newton:

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

E a relação entre massa e energia para um objeto em repouso, descoberta por Einstein:

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

As equações acima descrevem uma relação linear entre grandezas.

Relações lineares

Uma relação linear entre duas variáveis indica que elas estão correlacionadas em uma proporção constante para qualquer intervalo.

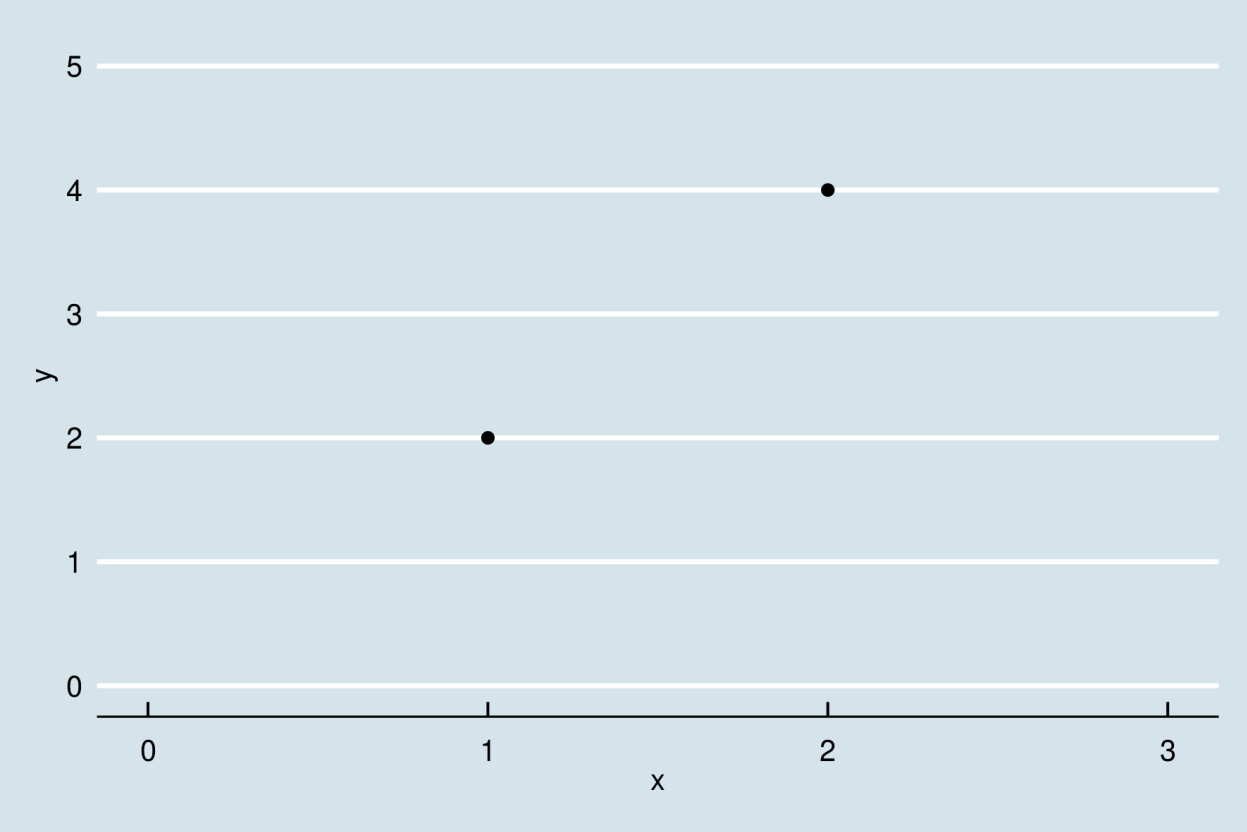
Isto é, valores maiores de massa correspondem a um aumento proporcional em energia. O valor de c^2 expressa essa proporção constante.

Exemplo: uma molécula de água pesa aproximadamente $m_{H_2O} = 2.992 \times 10^{-23} g$. Portanto, a energia associada é $E_{H_2O} = 2.992 \times 10^{-23} \times 8.988 \times 10^{16} \sim 2.689^{-6} J$. Se triplicarmos o número de moléculas de água, o mesmo acontecerá com a energia associada: $E_{3H_2O} = 3 \times E_{H_2O}$.

Se a correlação é positiva, incrementos em x serão proporcionais a incrementos em y. Se a correlação é negativa, incrementos em x serão proporcionais a decréscimos em y.

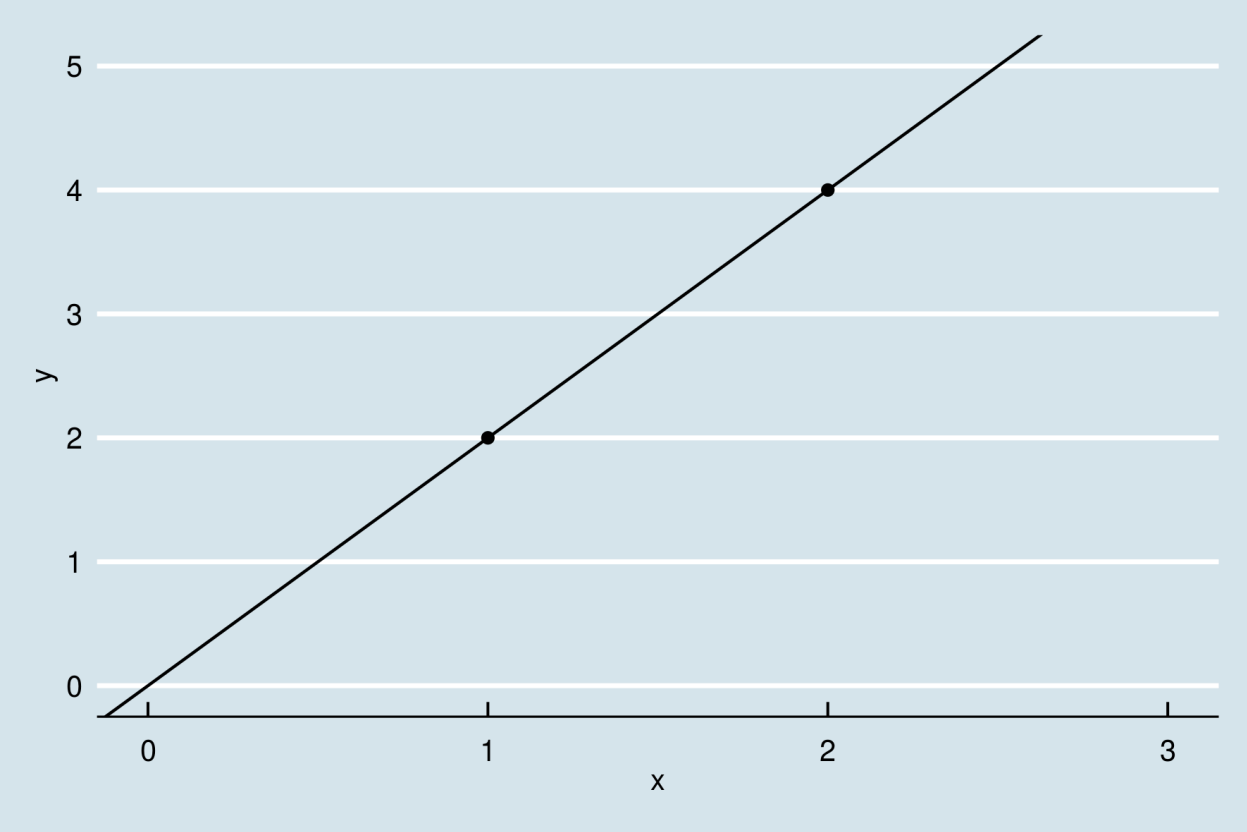
Num cenário perfeito, se sabemos que há uma relação linear entre variáveis, precisamos de apenas duas observações para descobrir proporção entre elas. Esse problema é idêntico ao de encontrar a inclinaçaõ da reta que passa por dois pontos. É de fácil resolução usando técnicas elementares.

```
>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()
```



Erros e aleatoriedade

Controlando fatores experimentais, as relações descritas são bastante precisas. Em um cenário sem atrito com superfícies e com o ar, os erros de medida obtidos com $\vec{F} = m\vec{a}$ são muito baixos.
Entretanto, nem sempre isso é verdadeiro.
Primeiro, podemos sofrer interferência de variáveis desconhecidas.

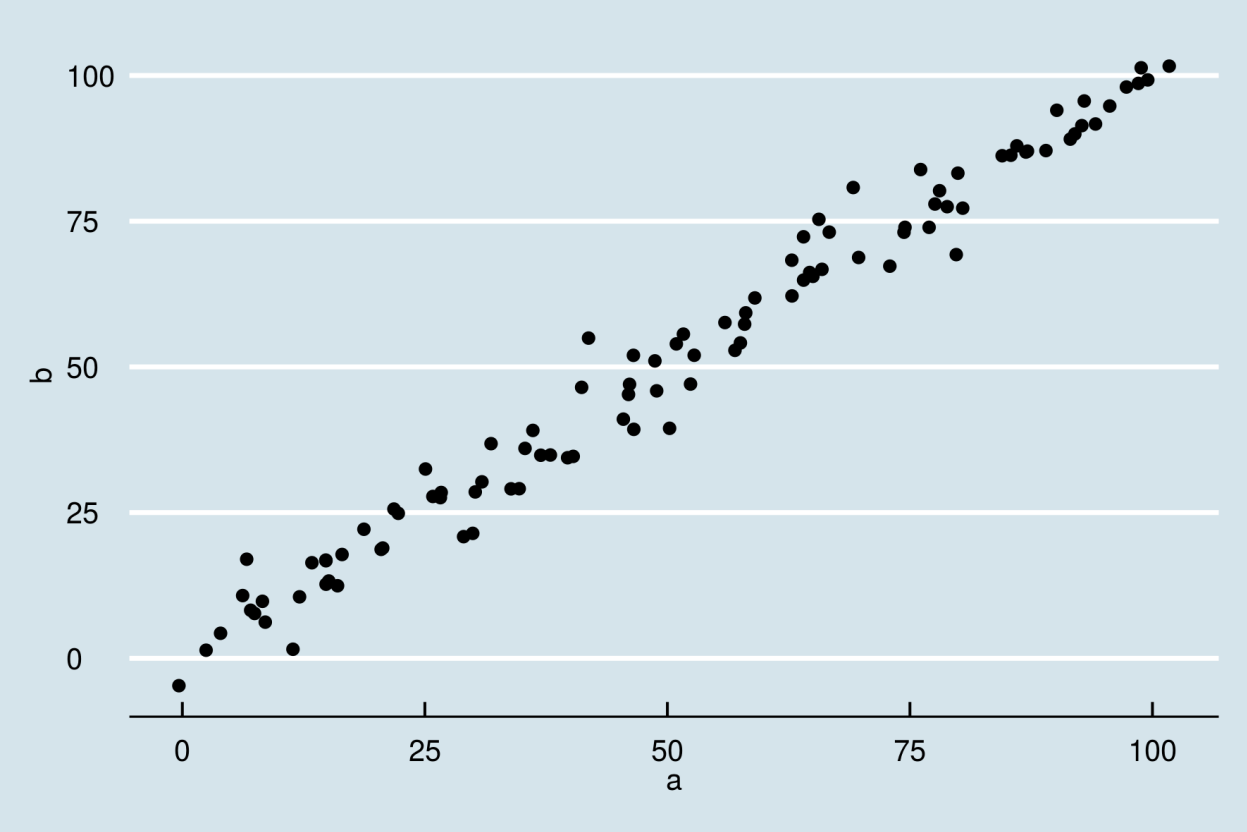
Imaginemos um conjunto de medidas antropométricas, como altura e peso de indivíduos.
É esperado que a altura de um ser humano esteja relacionada com seu peso. Entretanto, outras características não medidas, como percentual de gordura total, podem interferir nos valores finais. Normalmente, tratamos essas flutuações como erros aleatórios[1].

Podemos simular este cenário partindo de variáveis idênticas e adicionando ruído aleatório.

```
>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()
```

[1] A natureza da aleatoriedade é uma questão filosófica. Em última instância, podemos imaginar que seria possível explicar flutuações randômicas através de variáveis desconhecidas (*hidden variables*). Isso é verdade para a maioria dos fenômenos naturais. Entretanto, descobertas experimentais recentes em física quântica (*Bell's inequality experiment*) sugerem que variáveis ocultas não podem explicar a natureza probabilística das observações.



O resultado sugere que há uma forte relação linear entre x e y. Por outro lado, notamos que é impossível para uma reta cruzar todos os pontos. A seguir, vamos investigar como quantificar a correlação linear, assim como encontrar a reta que minimiza a distancia para todas as observações.

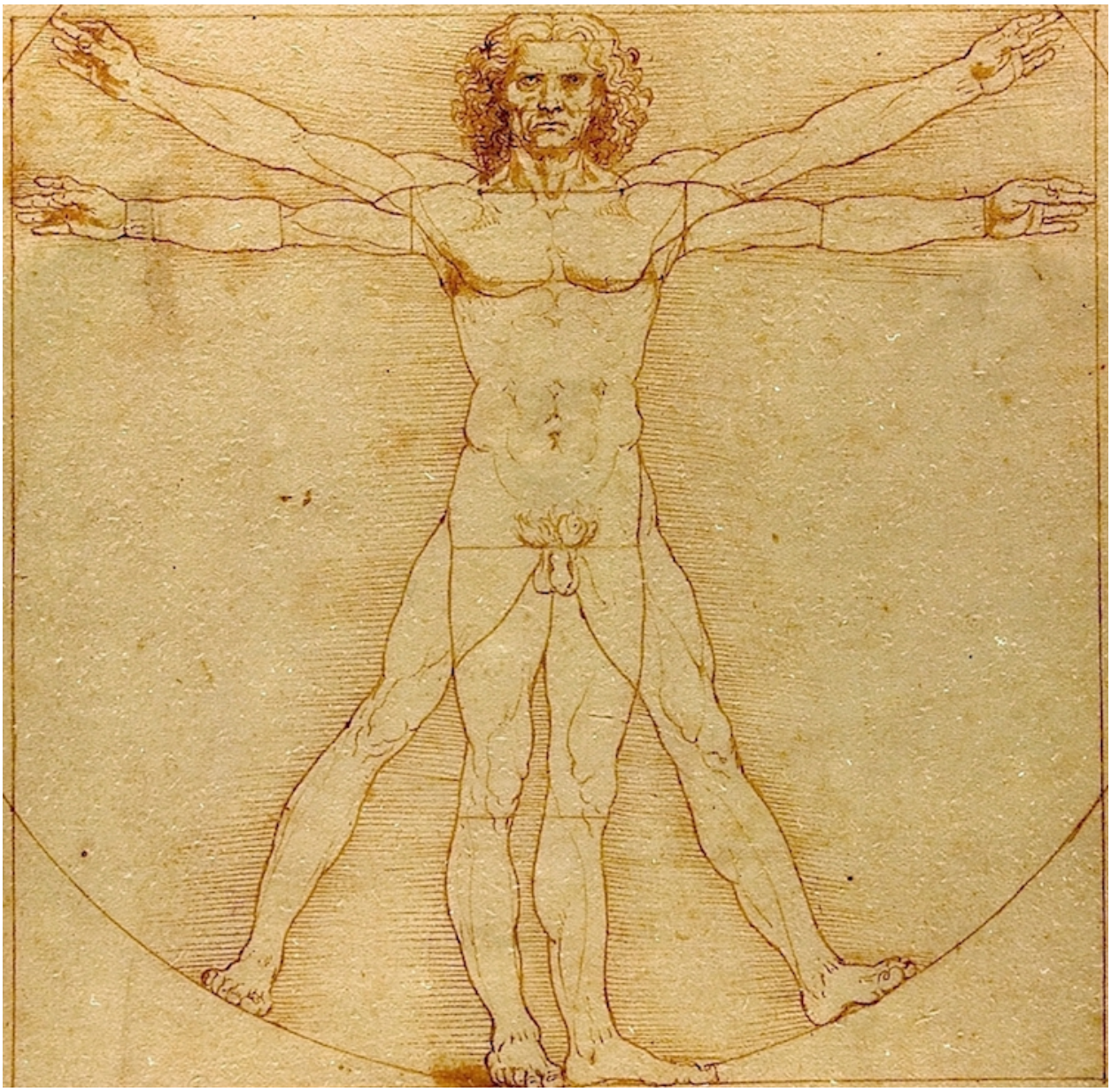
Com essas ferramentas, podemos estender nossas inferências. Além de comparações, teremos noções sobre a magnitude de uma relação, assim como poderemos prever o valor esperado para novas observações.

O coeficiente de correlação produto-momento de Pearson, ou, simplesmente, ρ de Pearson.

O coeficiente de correlação (ρ) de Pearson é um número real garantidamente[2] entre -1 e 1. Expressa a magnitude e o sentido de uma relação linear, sendo -1 uma relação inversa perfeita e 1 uma relação direta perfeita.

Para os dados que geramos, a correlação é quase perfeita: $\rho = 0.989$.
O coeficiente possui *produto-momento* em seu nome, pois usa uma abstração originalmente empregada na física, que estudamos no capítulo anterior: o momento(torque).

[2]Inequalidade de Cauchy–Schwarz. $\left(\sum_{i=1}^{n} u_i v_i\right)^2 \leq \sum_{i=1}^{n} u_i^2 \sum_{i=1}^{n} v_i^2$



Calculando correlações lineares

A noção de **distância** ou **desvio** se repetiu muitas vezes.
De fato, o coeficiente de correlação linear nasceu quando Francis Galton (1888) estudava numericamente dois problemas aparentemente distintos em antropometria [3] :

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'?

[3] 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[4] 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.

[4] 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.

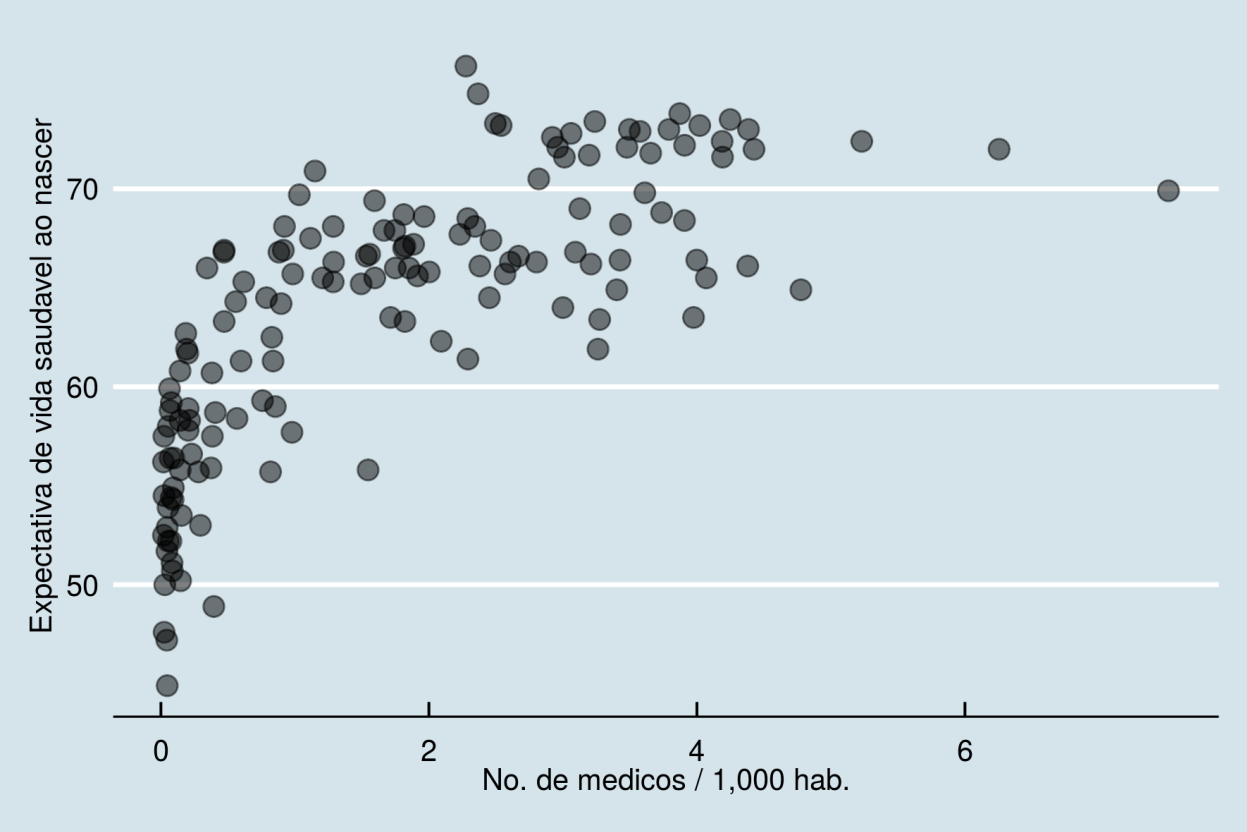
```
# http://apps.who.int/gho/data/view.main.HALEXv
# https://data.worldbank.org/indicator/SH.MED.PHYS.ZS
>library(magrittr)
>library(ggplot2)
>library(dplyr)

>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()
```



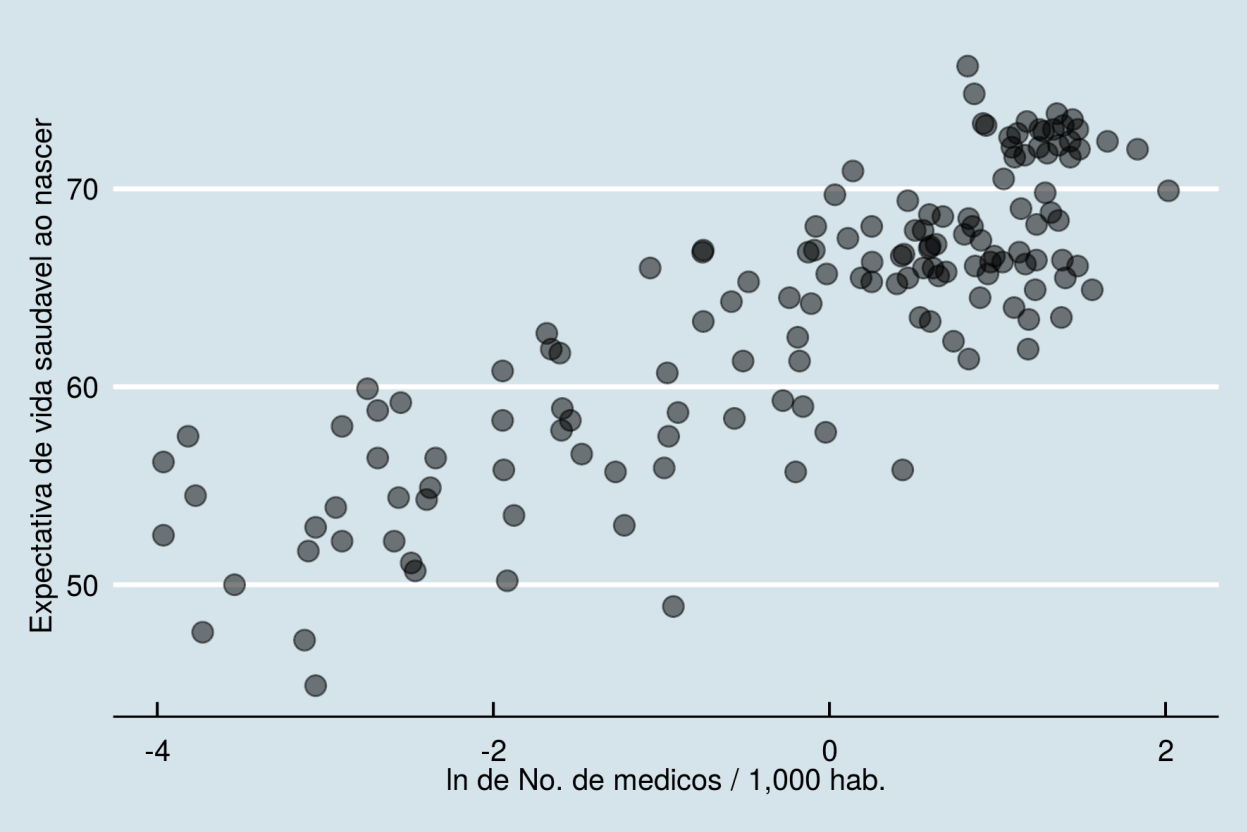
É evidente que o padrão não é aleatório. Visualmente, notamos que o valor da expectativa de vida aumenta com maior Nº de médicos. Ainda, notamos um aumento inicialmente rápido até atingir um platô. O padrão é semelhante ao de uma curva logarítmica.

$y = log(x)$ 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:

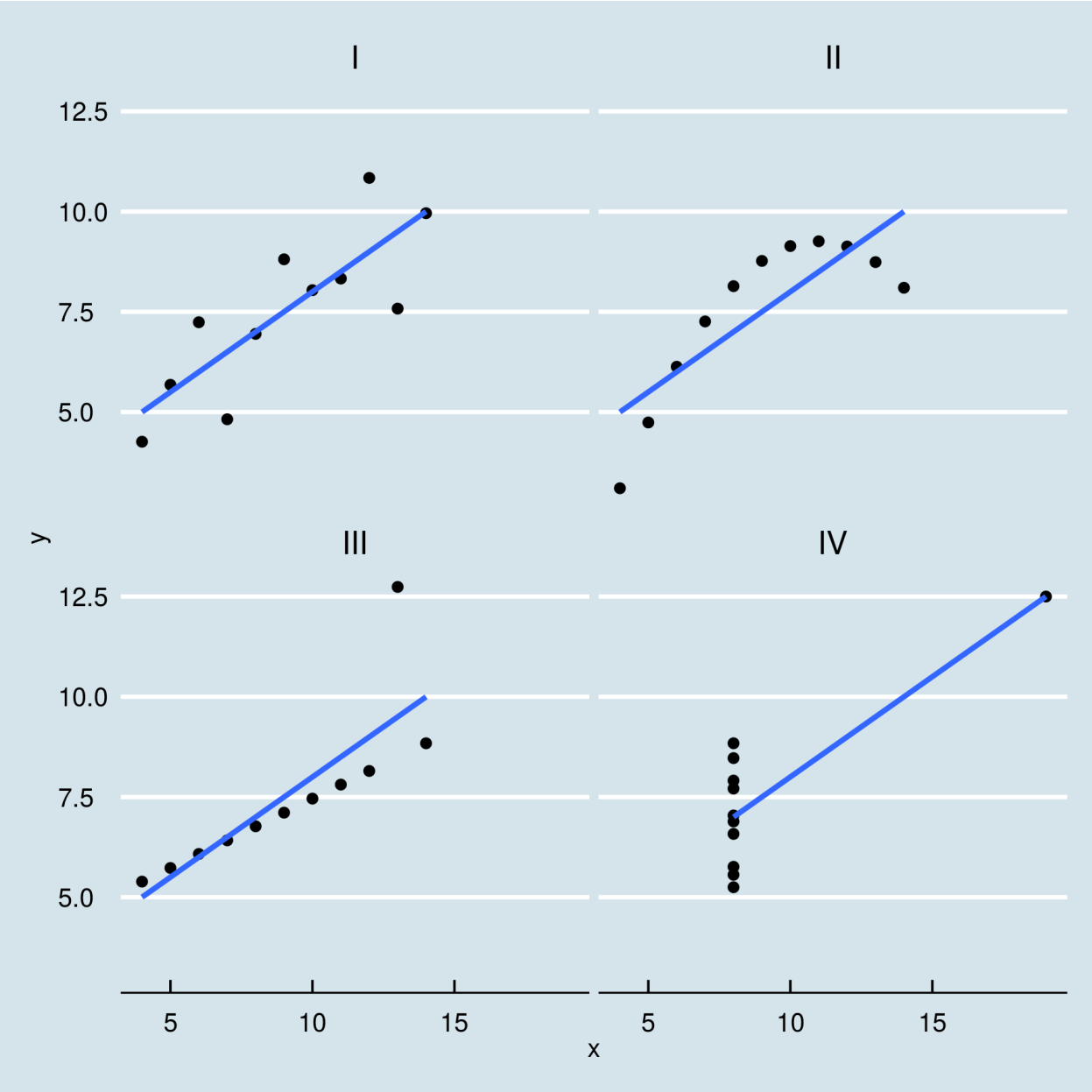
```
>cor.test(uni_df$log_docs,uni_df$hale)
Pearson's product-moment correlation
data:  uni_df$log_docs and uni_df$hale
t = 18.572, df = 143, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval: 0.7854248 0.8828027
sample estimates:
    cor
0.8407869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

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.

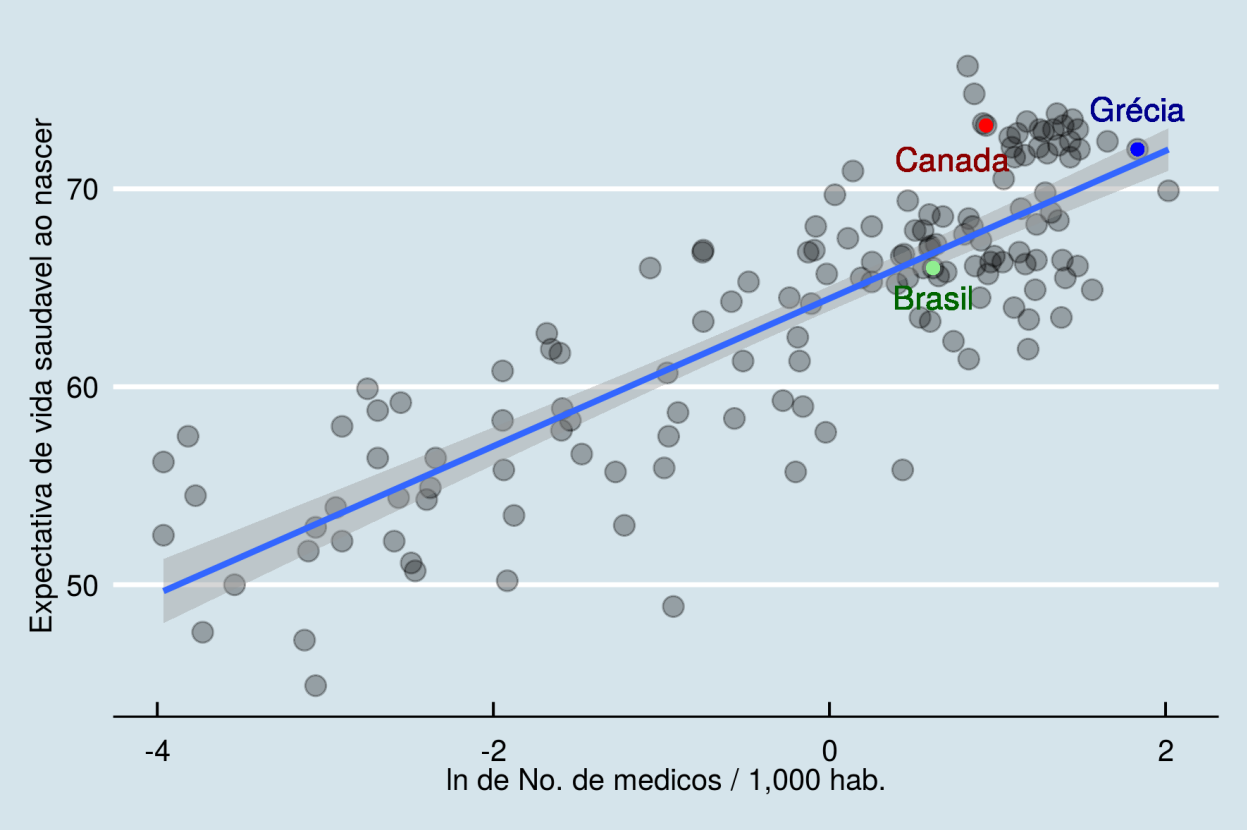


Previsões

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[5]. 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).

[5]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 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 [6]:

$\hat{\beta_1}$ 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.

[6] 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

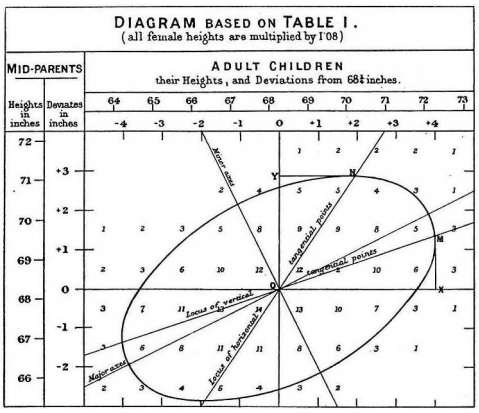


Figure 1: O primeiro gráfico de regresão linear. Ilustração de Francis Galton (1875) relação entre altura de pais e filhos.

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)
```

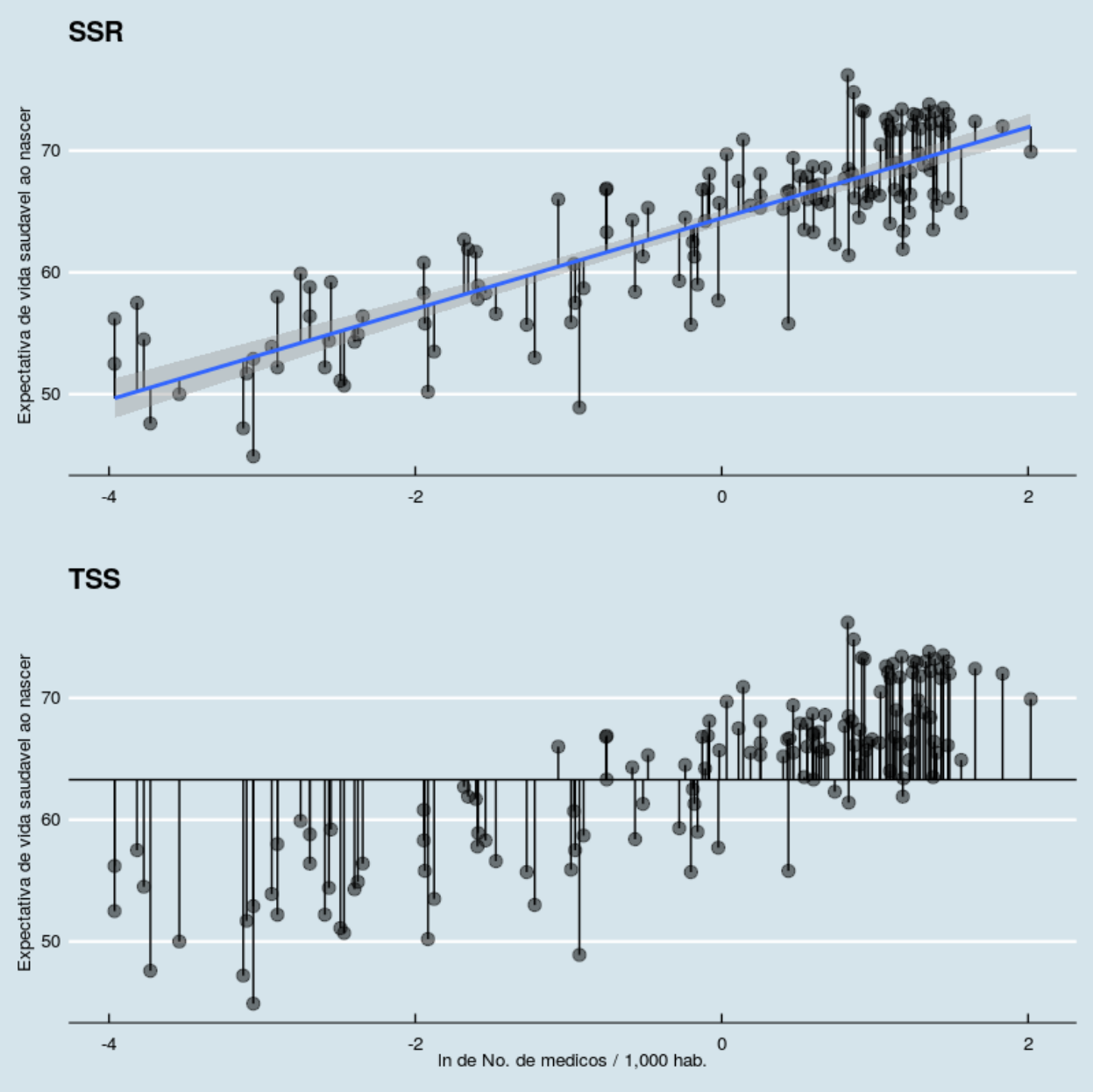


Figure 2: O quadrado da distância entre um ponto e a reta corresponde a um resíduo. Obtemos SSR e TSS somando todos os resíduos nas figuras superior e inferior, respectivamente.

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

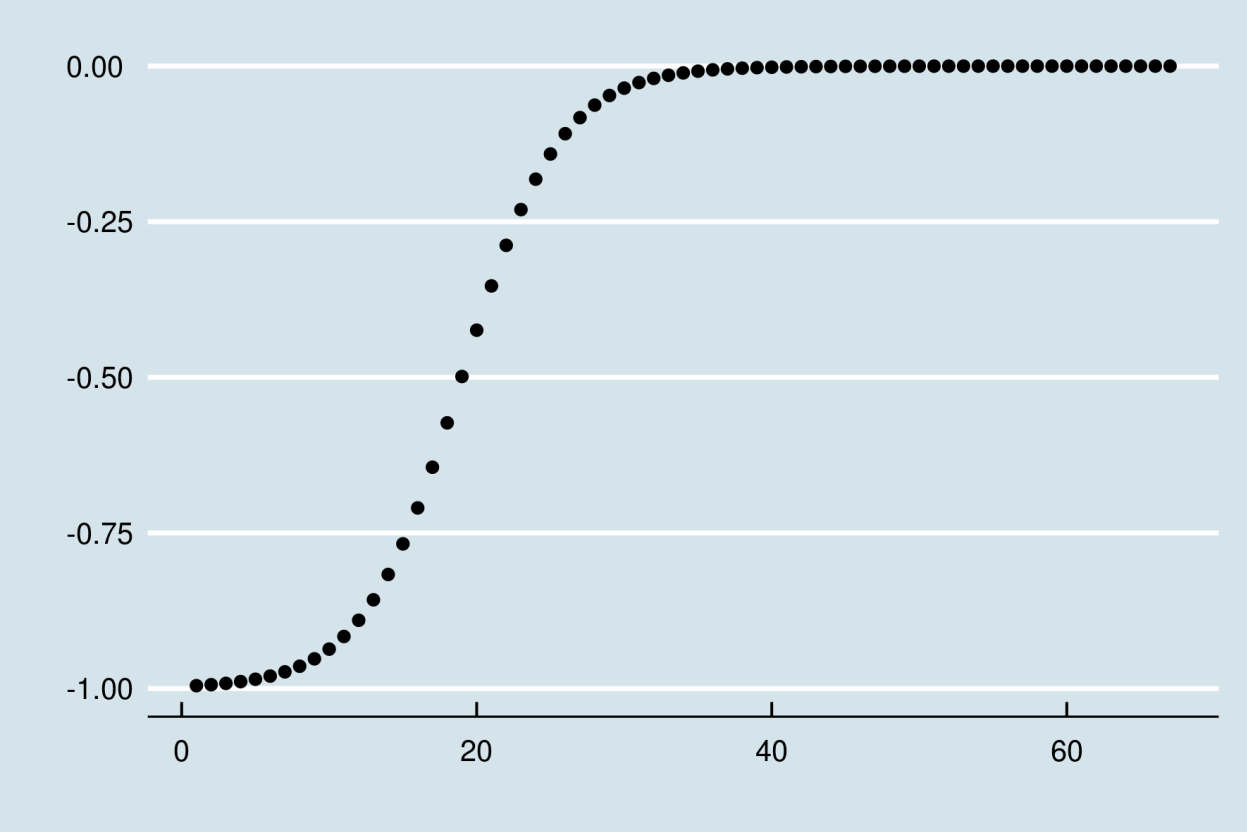
Relações lineares mantêm proporções constantes e aprendemos como quantificá-las. Por outro lado, duas variáveis podem ter relações de outros tipos, não lineares. Em especial, se as medidas apresentam valores muito extremos *(outliers)* um cálculo como o anterior sofre bastante com vieses.
Uma simples 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 $\rho \sim 0.850$[7] :

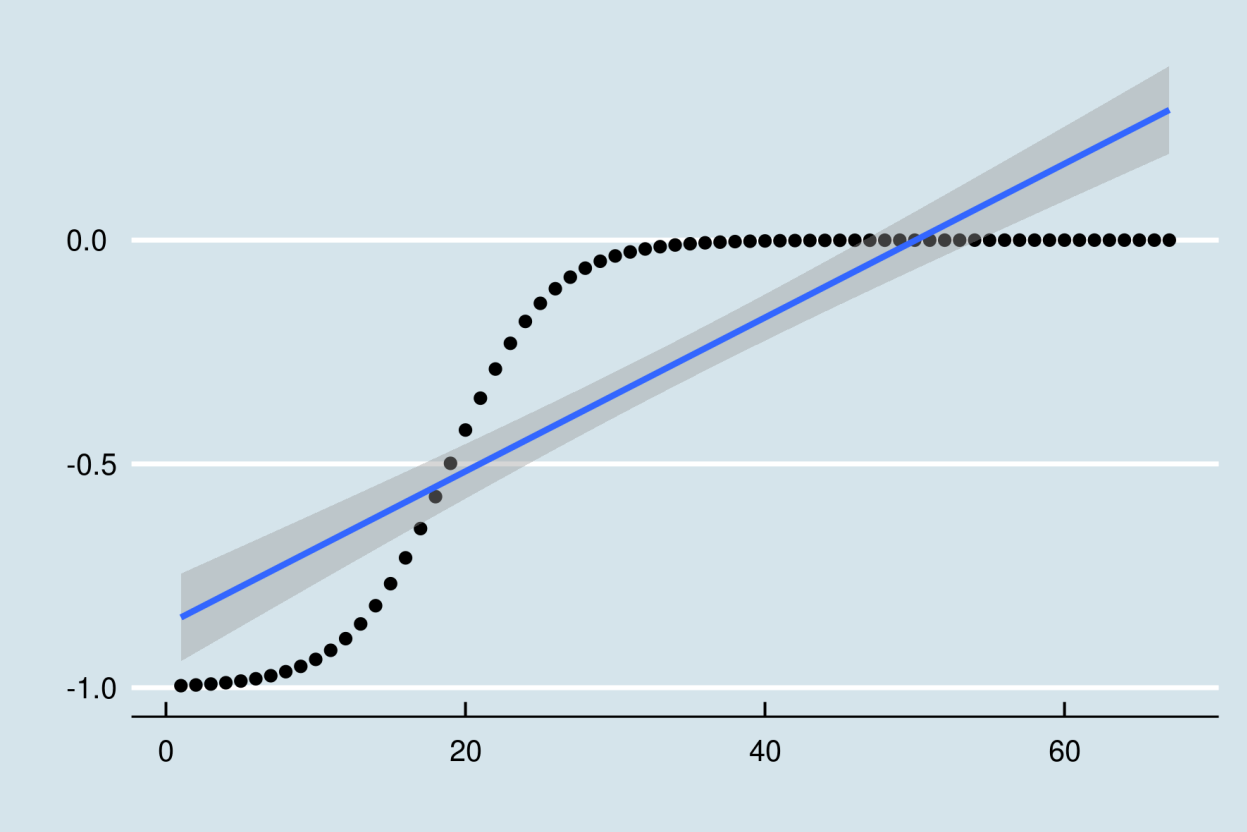
```
>cor.test(sig_data$y_vals,
+          sig_data$x_vals)

    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("")
```

[7]Como observamos no gráfico, a correlação linear não é tão alta. O coeficiente se aproxima de 1 $\rho \sim 0.850$ pois os desvios superiores compensam simetricamente os inferiores. O exemplo reforça a importância de plotar os dados para um melhor entendimento (ver Quarteto de Anscombe).



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
```

O coeficiente ρ de Spearman é preferível quando as medidas parecem diferir muito quanto à família da distribuição de origem. Especialmente, quando a média não parece corresponder bem ao centro das distribuições. Lembre-se que de o coeficiente de Pearson é baseado nos desvios em relação à média em ambas as amostras.

Teste U de Mann-Whitney

O teste U de Mann-Whitney faz uso da estatística U para fazer inferências. O racional é idêntico ao do teste t de Student.
Estabelecemos hipótese nula H_0 e hipótese alternativa H_1.
Então, calculamos a probabilidade de nossas observações acontecerem caso a hipótese nula seja verdadeira.
Desta vez, usaremos a estatística U. Lembremos que a estatística t era calculada com base em parâmetros extraídos da amostra:

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

A estatística U não depende de parâmetros (e.g. μ, σ), sendo calculada com base em cada observação.

Primeiro, calculamos os ranks de cada medida r_i unindo as observações das amostras A e B, de tamanhos amostrais n_a e n_b em apenas um conjunto ($N_{tot} = n_a + n_b$).

Depois, separamos novamente as amostras e calculamos a soma dos ranks em cada grupo, chamadas R_a e R_b.
A estatística U é dada pela seguinte expressão:

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

Usamos o menor valor de U para consultar a probabilidade (valor p) correspondente para a hipótese nula.

O termo $\frac{n(n+1)}{2}$ corresponde à soma mínima dos ranks para a amostra.
Os ranks são uma sequência regular $(1, 2, 3, ...)$, de forma que a soma de todos os valores é idêntica à soma de uma progressão aritmética de N termos.

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

Enquanto R_i corresponde à soma dos ranks calculados com as duas amostras, o termo acima corresponderia à soma mínima dos ranks para uma amostra, caso os ranks ocupassem a sequência inicial $A = (1, 2, 3, 4, ..., n_a)$ na amostra conjunta.
A definição para o teste não é unânime na literatura, de forma que alguns autores e softwares (e.g. R) implementam o cálculo com a subtração acima e outros (e.g. S-PLUS) não o fazem.
Em R, as funções **dwilcox(x,m,n)** e **pwilcox(q,m,n)** retornam a distribuição e a densidade cumulativa para a estatística U correspondente a amostras com tamanhos m e n. **wilcox.test(x,y,. . .)** é a implementação base do teste de Mann Whitney. O teste de Mann Whitney é o teste de Wilcoxon de duas amostras.

Exercícios

1. O coeficiente produto-momento de Pearson descreve quais tipos de relação?

- Ele é útil para modelar relações quadráticas entre variáveis?

- Citamos relações não lineares, como $E = mc^2$. Cite um outro exemplo de fenômeno natural de perfil não-linear em que o ρ de Pearson não funciona.

2. Crie uma função que calcula o n-ésimo momento para uma amostra:

- `n_moment <- function(x,n) {sum((x- mean(x))^n)/length(x)}`
- Calcule o valor de skewness. Como citado no capítulo, é o 3o momento normalizado [pelo 2o momento ao expoente 3/2].

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

- Calcule o valor de kurtosis. Como citado, é o 4 momento noramlizado [pelo quadrado do 2o momento menos 3].

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

- Os valores podem ser conferidos com as implementações `e1071::skewness` e `e1071::kurtosis`

3. Usando o dataset *iris*, compare as 4 variáveis numéricas (*Sepal/Petal Lenght/Width*) entre espécies (*Species*) usando teste t de Student e teste de U Mann Whitney. Em algum caso os métodos divergem quanto à rejeição da hipótese nula?

- Obtenha o tamanho de efeito (D de Cohen) para as diferenças.

4. Usando o dataset *iris*:

- Faça um scatterplot entre duas medidas. A função `pairs` pode ajudar.
  - Verifique se há correlação linear significativa entre as variáveis.
  - Se existir, ajuste um modelo de regressão linear.
  - Ajuste um modelo de regressão para cada espécie.
  - Observe os valores de R^2 para cada modelo. Qual a sua impressão sobre as mudanças de performance?