# Capítulo 2

# Expectativa de vida e assistência em saúde

## Correlações, coeficiente *r* de Pearson e regressão linear.

No exemplo anterior, estudamos a diferença entre dois conjuntos de medidas independentes. Estudamos as medidas dos bicos de pássaros, medidas unidimensionais de números reais.

Os cientistas costumam usar esse racional para comparar quantias vindas de grupos qualitativamente distintos, porém equiparáveis.

Em seguida, vamos tratar de outra abstração: correlações. Agora, tratamos de medidas bidimensionais, pontos representados por pares ordenados (xi, yi) em R². Adotando países como unidade experimental: (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)) %>% #last non missing value

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")],

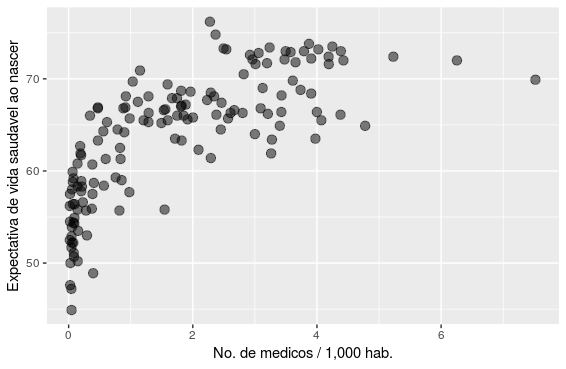
who\_df[,c("Country","hale")],by="Country")

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")



É evidente que o padrão não é aleatório. Visualmente, notamos que o valor da expectativa de vida aumenta parece aumentar nos pontos com maior Nº de médicos.

Ainda, notamos um aumento inicialmente rápido, com um platô em seguida. O padrão é semelhante ao de uma curva logarítmica.

y = log(x) ou HALE = log(No.Medicos)

Se isso for verdade, transformar o Nº de Médicos usando escala logarítmica tornará a relação linear. Se, y = log (x), fazemos a transformação *x’ = log(xlog)*, então:  
y (x’) = log(x’) = (log(log(xlog)) = xlog.  
y (x’) = xlog

Então y (expectativa de vida) é linearmente porcional a xlog(logaritmo da Nº de médicos).

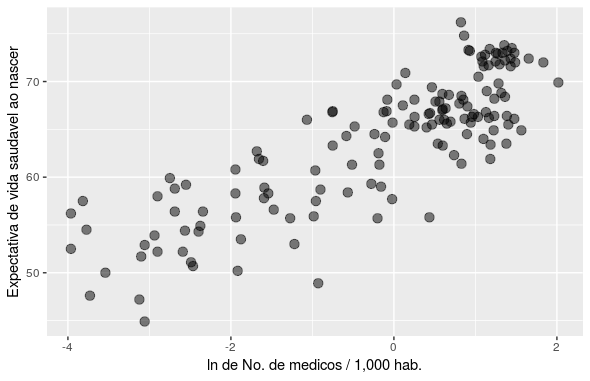
>uni\_df$log\_docs <- log(uni\_df$n\_docs)

>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")



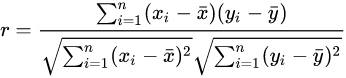
De fato, verificamos uma notável tendência linear para os pontos.

Uma forma de medir o grau de correlação em medidas bidimensionais é através do coeficiente *r* de Pearson. Ele é dado pela covariância entre variáveis dividida pelo produto dos desvios.



A covariância pode ser dada pelo valor esperado do produto dos desvios de ambas as amostras em relação às respectivas médias. Se os desvios concordam em sinal, a correlação aumenta. Se as medidas são discrepantes, a correlação de diminui.

Se temos padrões consistentemente concordantes, teremos uma correlação final próxima de 1. Se o padrão for consistentemente discordante, teremos uma correlação final próxima de -1. Se não houver um padrão consistente, a correlação é próxima de 0.



Então:

> 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

**0.840**7869

A correlação linear obtida para nossa amostra de países é grande, como sugeria a visualização (r = 0.840). Supondo uma distribuição t, obtemos um valor p baixo (p <0.001) considerando a hipótese nula H0 de r = 0. Concluímos então que há uma relação linear entre o logaritmo do número de médicos e a expectativa de vida dos países em nossa amostra.

Do ponto de vista Popperiano, a hipótese de que há uma relação positiva entre número de médicos de um país e a expectativa de vida é científica. Ela é testável, como demonstramos acima. Um resultado sugerindo que *r* = 0 é plausível invalidaria a ideia.

### Predições

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 (nº 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 da figura anterior.

yi = β0 + β1x*i + ε*

Para isso, calculamos a inclinação (β1) e o ajuste vertical(β0) 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.

Ajustamos o modelo usando o método *lm(linear model)* do R:

# log\_docs : x’ = log(x)

> lm(log\_docs ~ hale, data=uni\_df) %>% summary

Call:

lm(formula = log\_docs ~ hale, data = uni\_df)

Residuals:

Min 1Q Median 3Q Max

-2.44037 -0.47861 -0.01513 0.55290 2.13418

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -12.2731 0.6542 -18.76 <2e-16 \*\*\*

hale 0.1895 0.0102 18.57 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8517 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

Temos β0= -12.27 e β1=0.18. Existe mais de uma maneira de estimar esses parâmetros.

Uma de particular interesse em muitos contextos é a de *Maximum likelihood* (máxima verossimilhança)[[1]](#footnote-2).

Determinamos uma função que descreve a probabilidade da observação na variável alvo (yi) ocorrer dadas medidas das variáveis preditoras (xi) e um conjunto de parâmetros (bk).

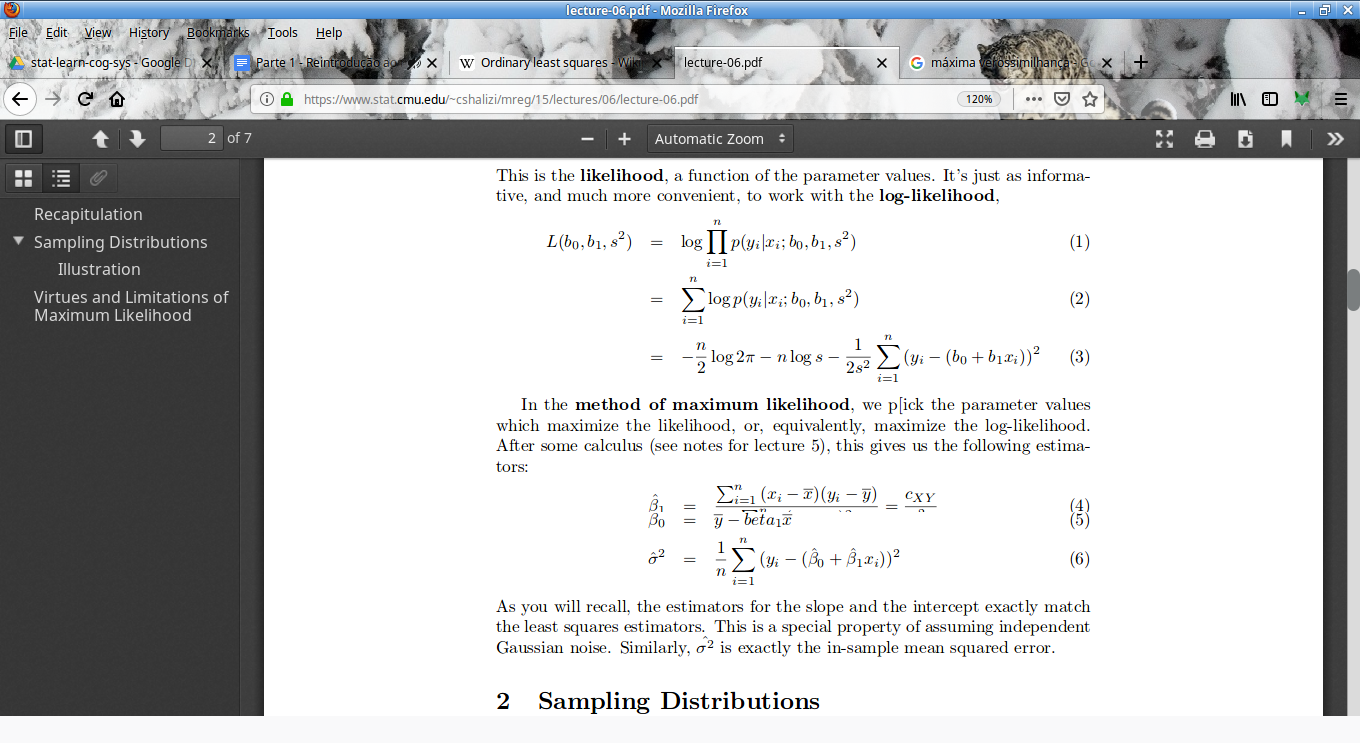
Podemos supor que os valores yi são dados por uma distribuição normal cuja média é uyi = β0 + β1\*xi.

P(yi | xi *;* β0,β*1*,σ²) =

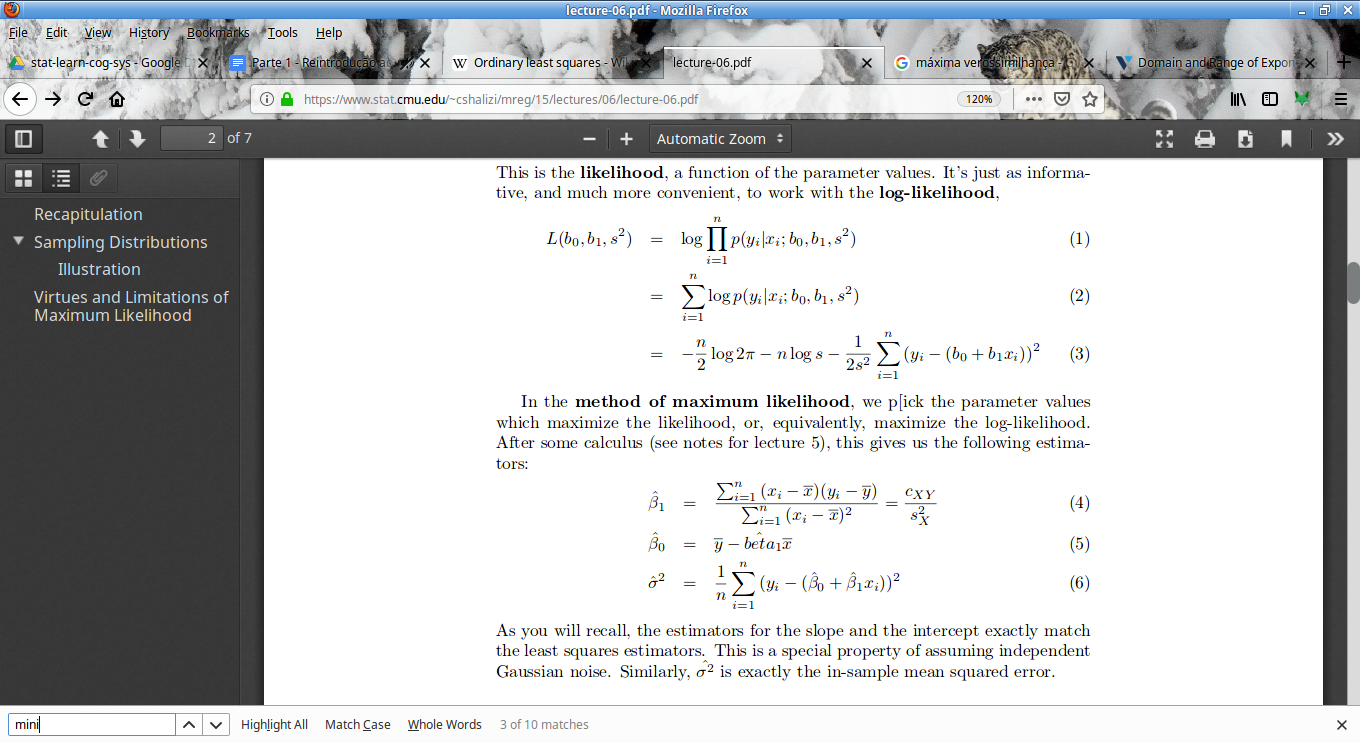


Essa é nossa função de verossimilhança (*likelihood function*) e dá a probabilidade observarmos as medidas yi dadas as medidas x*i* econsiderando um conjunto de parâmetros.

O objetivo então é maximizar essa função. Por conveniência, aplicamos uma transformação logaritmica nesta função (*log likelihood function*). Isso transforma nosso produtório em um somatório e passamos o contradomínio do intervalo [0;1] para [0;+∞):



Usando cálculo, chegamos a:



Devemos então nos preocupar em saber se a solução encontrada é de fato acurada aos dados. Existem diferentes parâmetros para avaliar a qualidade de nosso modelo:

@R² (~ 0,70) @RMSEA @AIC/BIC

1. Dedução completa:

   *Max. Likelihood*: <https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/06/lecture-06.pdf> ;

   OLS: <https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/05/lecture-05.pdf> [↑](#footnote-ref-2)