Escola de Serviço Público do Espírito Santo - Esesp

TRILHAS DE CAPACITAÇÃO - SEFAZ

Introdução à Econometria

GOVERNO DO ESTADO DO ESPÍRITO SANTO
Secretaria de Gestão e Recursos Humanos

CONTEÚDO PROGRAMÁTICO

R
Revisão estatística
Tratamento de dados
Tipos de dados
Regressão simples
A hipótese da normalidade
Inferência estatística
Logaritmos
Regressão múltipla
Variáveis qualitativas

ANTES DE COMEÇAR

O que é a econometria?

ANTES DE COMEÇAR

O que é a econometria?

“A econometria consiste na aplicação da estatística aos dados econômicos para dar suporte empírico aos modelos construídos pela matemática econômica e para obter resultados numéricos”

“A econometria pode ser definida como a análise quantitativa dos fenômenos econômicos baseada no desenvolvimento concorrente da teoria e da observação”

“A econometria pode ser definida como uma ciência social na qual as ferramentas da teoria econômica, da matemática e da inferência estatística são aplicadas na análise dos fenômenos econômicos”

“A econometria está preocupada com a determinação empírica das leis econômicas”

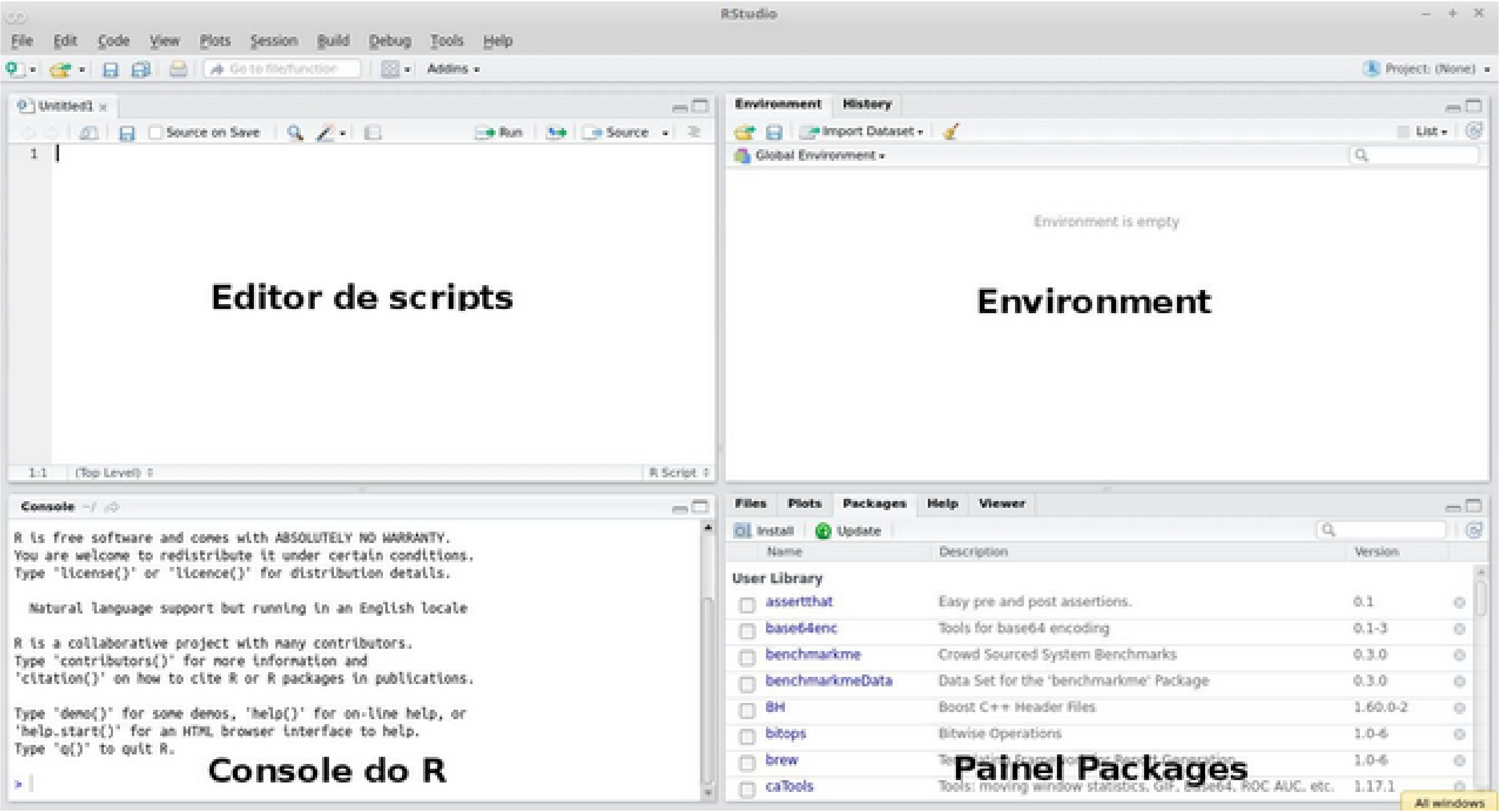
Fonte: as citações estão no livro Basic Econometrics, de Damodar Gujarati.

ANTES DE COMEÇAR

Material e recursos

Nosso curso será baseado nos livros “Basic Econometrics”, de Damodar Gujarati, “Principle of Econometrics”, de R. Carter-Hill, William E. Griffiths e Guay C. Lim, e “Introductory Econometrics”, de Jeffrey M. Wooldridge. A maioria dos dados e exercícios são retirados destes livros, outros foram elaborados exclusivamente para este curso. O enfoque do curso é aplicado, mas também teremos exposições teóricas. O software que utilizaremos é o R, através do R Studio. Embora haja uma diversidade de softwares econométricos, o R foi escolhido por ser amplamente empregado na comunidade acadêmica e por profissionais do mercado (além de ser livre).

R



Baseado no livro “Processamento e modelagem de dados financeiros com o R”, de Marcelo Perlin.

R

Objeto numérico:
> a <- 1
> a
[1] 1
> class(a)
[1] "numeric"

Objeto de texto:
> b <- "1"
> b
[1] "1"
> class(b)
[1] "character"

Vetores:
> x <- c(1,2,3,4,5)
> x
[1] 1 2 3 4 5
> class(x)
[1] "numeric"

Vetores:
> y <- c(1,2,3,4,"5")
> y
[1] "1" "2" "3" "4" "5"
> class(y)
[1] "character"

Operações sobre um mesmo vetor:
> soma <- x+1
> soma
[1] 2 3 4 5 6

> subtracao <- x-1
> subtracao
[1] 0 1 2 3 4

> multiplicacao <- x*2
> multiplicacao
[1] 2 4 6 8 10

> divisao <- x/2
> divisao
[1] 0.5 1.0 1.5 2.0 2.5

Operações sobre diferentes vetores:
> v1 <- c(20:25)
> v2 <- c(rep(5:6,2),7,8)

> v1+v2
[1] 25 27 27 29 31 33

> v1-v2
[1] 15 15 17 17 17 17

> v1*v2
[1] 100 126 110 138 168 200

> v1/v2
[1] 4.000000 3.500000 4.400000 3.833333 3.428571 3.125000

R

```
> class(v1)
[1] "integer"
> class(v2)
[1] "numeric"
> class(v1+v2)
[1] "numeric"
> is.vector(v1)
[1] TRUE
> is.vector(v2)
[1] TRUE
```

Data Frames

```
> conjunto <- data.frame(v1,v2)
> conjunto
  v1 v2
1 20  5
2 21  6
3 22  5
4 23  6
5 24  7
6 25  8
```

```
> class(conjunto)
[1] "data.frame"
> class(conjunto[,1])
[1] "integer"
> class(conjunto[,2])
[1] "numeric"

> conjunto[3,1]
[1] 22

> conjunto[,1]
[1] 20 21 22 23 24 25
> conjunto[,2]
[1] 5 6 5 6 7 8
```

```
> conjunto[,3] <- conjunto[,1]
+ conjunto[,2]

> conjunto
  v1 v2 V3
1 20  5 25
2 21  6 27
3 22  5 27
4 23  6 29
5 24  7 31
6 25  8 33
```

R

```
> conjunto[1,]
  v1 v2 V3
1 20  5 25
> conjunto[6,]
  v1 v2 V3
6 25  8 33

> conjunto[7,] <- conjunto[6,] - conjunto[1,]
> conjunto
  v1 v2 V3
1 20  5 25
2 21  6 27
3 22  5 27
4 23  6 29
5 24  7 31
6 25  8 33
7  5  3  8

> head(conjunto,2)
  v1 v2 V3
1 20  5 25
2 21  6 27
```

```
> tail(conjunto,2)
  v1 v2 V3
6 25  8 33
7  5  3  8

> conjunto_original <- conjunto
> conjunto_A <- conjunto[-1,]
> conjunto_B <- conjunto[,-1]

> length(conjunto_original)
[1] 3
> length(conjunto_A)
[1] 3
> length(conjunto_B)
[1] 2
> length(conjunto_A[,1])
[1] 6
> length(conjunto_B[,1])
[1] 7
```

```
> conjunto_A
  v1 v2 V3
2 21  6 27
3 22  5 27
4 23  6 29
5 24  7 31
6 25  8 33
7  5  3  8

> conjunto_B
  v2 V3
1  5 25
2  6 27
3  5 27
4  6 29
5  7 31
6  8 33
7  3  8
```

R

```
> colnames(conjunto_original) <- c("vetor1","vetor2","vetor3")
> colnames(conjunto_A) <- c("vetor1","vetor2","vetor3")
> colnames(conjunto_B) <- c("vetor2","vetor3")

> conjunto_original
  vetor1 vetor2 vetor3
1     20      5     25
2     21      6     27
3     22      5     27
4     23      6     29
5     24      7     31
6     25      8     33
7      5      3      8
```

R

```
> conjunto_original <- ts(conjunto_original, start=c(2011), frequency=1)

> conjunto_original
Time Series:
Start = 2011
End = 2017
Frequency = 1
     vetor1 vetor2 vetor3
2011     20      5     25
2012     21      6     27
2013     22      5     27
2014     23      6     29
2015     24      7     31
2016     25      8     33
2017      5      3      8
```

```
> start(conjunto_original)
[1] 2011    1
> end(conjunto_original)
[1] 2017    1
```

```
> class(conjunto_original)
[1] "mts"    "ts"     "matrix"
> class(conjunto_original[,1])
[1] "ts"
> class(conjunto_original[,2])
[1] "ts"
> class(conjunto_original[,3])
[1] "ts"
```

R

```
> min(conjunto_original)
[1] 3
> max(conjunto_original)
[1] 33
> mean(conjunto_original)
[1] 17.14286
> median(conjunto_original)
[1] 21

> summary(conjunto_original)
    vetor1         vetor2          vetor3
 Min.   : 5.0   Min.   :3.000   Min.   : 8.00
 1st Qu.:20.5   1st Qu.:5.000   1st Qu.:26.00
 Median :22.0   Median :6.000   Median :27.00
 Mean   :20.0   Mean   :5.714   Mean   :25.71
 3rd Qu.:23.5   3rd Qu.:6.500   3rd Qu.:30.00
 Max.   :25.0   Max.   :8.000   Max.   :33.00

> median(conjunto_original[,2])
[1] 6
```

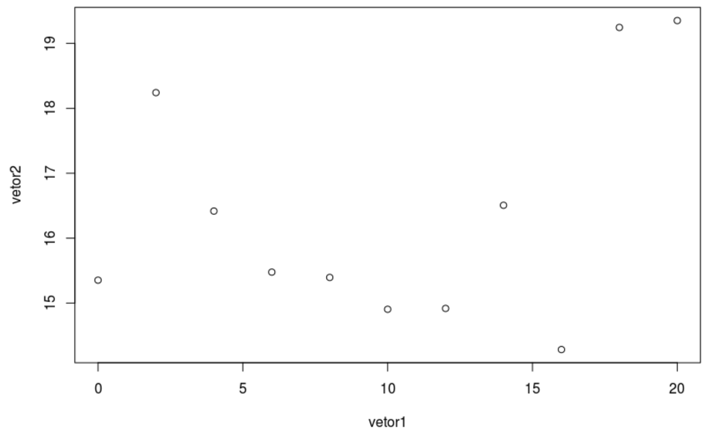
```
> colunas_original <-
c(conjunto_original[,1],conjunto_original[,2],con
junto_original[,3])

> colunas_original
 [1] 20 21 22 23 24 25  5  5  6  5  6  7  8  3 25
27 27 29 31 33  8

> min(colunas_original)
[1] 3
> max(colunas_original)
[1] 33
> mean(colunas_original)
[1] 17.14286
> median(colunas_original)
[1] 21
```

R

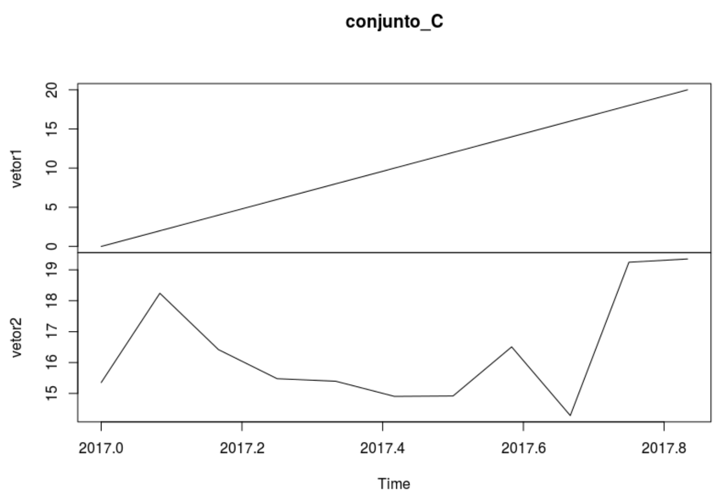
```
> vetor1 <- seq(from=0, to=20, by=2)
> vetor1
 [1]  0  2  4  6  8 10 12 14 16 18 20
> vetor2 <- rnorm(11,15,3)
> vetor2
  [1] 15.35296 18.24183 16.41732 15.47680 15.39484 14.90453 14.91763 16.50673 14.28406
19.24582 19.35191

> conjunto_C <- ts(data.frame(vetor1,vetor2), start=c(2017,1), frequency=12)
> conjunto_C
         vetor1   vetor2
Jan 2017      0 15.35296
Feb 2017      2 18.24183
Mar 2017      4 16.41732
Apr 2017      6 15.47680
May 2017      8 15.39484
Jun 2017     10 14.90453
Jul 2017     12 14.91763
Aug 2017     14 16.50673
Sep 2017     16 14.28406
Oct 2017     18 19.24582
Nov 2017     20 19.35191
```

R

> plot(vetor1,vetor2)

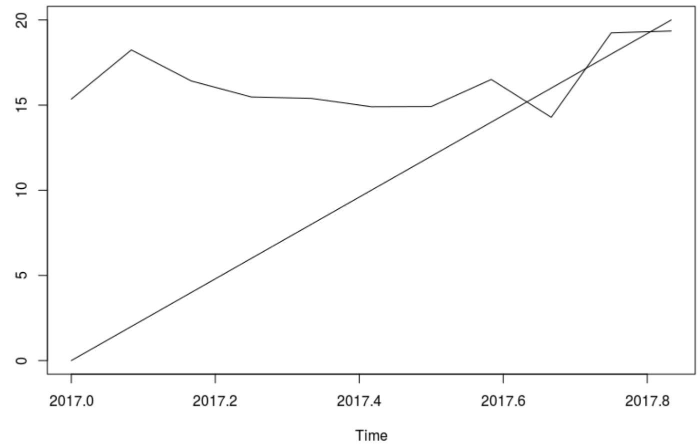


> plot.ts(conjunto_C)

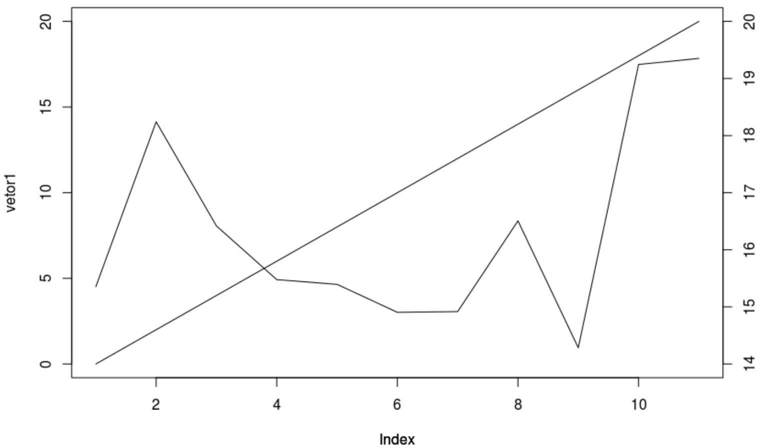


R

> ts.plot(conjunto_C)



> plot(vetor1, type="l", ylim=c(0,20))
> par(new=T)
> plot(vetor2, type="l", ylim=c(14,20), axes=FALSE, ylab="")
> axis(4,ylim=c(14,20))



R

```
> lista <- list(conjunto_original, conjunto_A, conjunto_B, conjunto_C)
> lista
[[1]]
Time Series:
Start = 2011
End = 2017
Frequency = 1
     vetor1 vetor2 vetor3
2011     20      5     25
2012     21      6     27
2013     22      5     27
2014     23      6     29
2015     24      7     31
2016     25      8     33
2017      5      3      8

[[2]]
  vetor1 vetor2 vetor3
2     21      6     27
3     22      5     27
4     23      6     29
5     24      7     31
6     25      8     33
7      5      3      8

[[3]]
  vetor2 vetor3
1      5     25
2      6     27
3      5     27
4      6     29
5      7     31
6      8     33
7      3      8

[[4]]
         vetor1   vetor2
Jan 2017      0 15.35296
Feb 2017      2 18.24183
Mar 2017      4 16.41732
Apr 2017      6 15.47680
May 2017      8 15.39484
Jun 2017     10 14.90453
Jul 2017     12 14.91763
Aug 2017     14 16.50673
Sep 2017     16 14.28406
Oct 2017     18 19.24582
Nov 2017     20 19.35191

> length(lista)
[1] 4

> class(lista)
[1] "list"

> lista[[3]]
  vetor2 vetor3
1      5     25
2      6     27
3      5     27
4      6     29
5      7     31
6      8     33
7      3      8

> lista[[2]][5,2]
[1] 8
> lista[[2]][5,2]*5
[1] 40
```

R

```
> names(lista) <- c("Conjunto Original","Conjunto A", "Conjunto B", "Conjunto C")
> lista
$`Conjunto Original`
Time Series:
Start = 2011
End = 2017
Frequency = 1
     vetor1 vetor2 vetor3
2011     20      5     25
2012     21      6     27
2013     22      5     27
2014     23      6     29
2015     24      7     31
2016     25      8     33
2017      5      3      8

$`Conjunto A`
  vetor1 vetor2 vetor3
2     21      6     27
3     22      5     27
4     23      6     29
5     24      7     31
6     25      8     33
7      5      3      8

$`Conjunto B`
  vetor2 vetor3
1      5     25
2      6     27
3      5     27
4      6     29
5      7     31
6      8     33
7      3      8
```

```
$`Conjunto C`
         vetor1   vetor2
Jan 2017      0 15.35296
Feb 2017      2 18.24183
Mar 2017      4 16.41732
Apr 2017      6 15.47680
May 2017      8 15.39484
Jun 2017     10 14.90453
Jul 2017     12 14.91763
Aug 2017     14 16.50673
Sep 2017     16 14.28406
Oct 2017     18 19.24582
Nov 2017     20 19.35191

> conjunto_D <- log(lista$`Conjunto A`[1:3,])
> conjunto_D
    vetor1   vetor2   vetor3
2 3.044522 1.791759 3.295837
3 3.091042 1.609438 3.295837
4 3.135494 1.791759 3.367296

> lista[[5]] <- conjunto_D

> lista[[5]]
    vetor1   vetor2   vetor3
2 3.044522 1.791759 3.295837
3 3.091042 1.609438 3.295837
4 3.135494 1.791759 3.367296
```

Excluindo o quinto elemento da lista:

```
lista <- lista[-5]
```

R

Exercícios

1. Construa:
a) um vetor A com 10 elementos;
b) um vetor B com 10 elementos, de modo que o vetor A não seja uma combinação linear do vetor B;
c) um vetor C que seja uma combinação linear dos dois primeiros;
d) um data frame com os três vetores.

2. Construa, a partir do primeiro data frame, um segundo data frame que contenha apenas as três primeiras linhas e as três primeiras colunas. Transforme este segundo data frame em uma matriz (use o comando as.matrix).

3. Use o comando length para verificar o tamanho da matriz; use o comando dim para verificar as dimensões da matriz. Armazene os resultados em um único vetor (vetor D, que terá tamanho 3).

4. Inverta a matriz construída no exercício 2 usando o comando solve. Aparecerá uma mensagem de erro. Tente solucionar o problema invertendo a submatriz 2x2 correspondente. Armazene a nova matriz.

5. Crie uma lista com 5 elementos (lista1): a) o data frame criado no exercício 1; b) a matriz criada no exercício 2; c) o vetor D criado no exercício 3; d) a matriz criada no exercício 4; e) um vetor com o seu nome.

6. Crie uma nova lista (lista 2) em que a ordem dos elementos seja a inversa à da primeira lista. Para criar a lista 2 vazia, utilize o comando list().

7. Crie um vetor que armazene o primeiro elemento (linha, coluna) de cada componente da segunda lista.

R

Importação e exportação de arquivos

A forma mais simples de importar um conjunto de dados consiste em digitá-los diretamente no R na forma de um vetor, um data frame, uma matriz etc. Porém, nem sempre isto é conveniente devido à quantidade de dados, ao modo como se encontram organizados etc.

Geralmente, os arquivos de dados estão disponíveis externamente em forma de planilhas excel, arquivos CSV, tabelas HTML, tabelas de bancos de dados ou arquivos tabulares (texto ou outro formato), podendo ser acessados local ou remotamente.

R

Importação e exportação de arquivos

Antes de começar a importar ou exportar arquivos, é necessário definir o diretório (pasta) de trabalho que será utilizado.

Primeiramente, verificar qual é o diretório de trabalho.

```
> getwd()
```

Depois, setar o diretório de trabalho apropriado.

```
> setwd("C:/seunome/entrada")
```

Digitar novamente getwd() para conferir se o diretório foi mudado.

Dica importante: usar sempre barra simples / no lugar da barra invertida apresentada pelo Windows. Por exemplo, o diretório C:\pasta teria que ser digitado no R como C:/pasta.

R

Importação e exportação de arquivos

O comando list.files() mostra todos os arquivos ou pastas que estão em nosso diretório de trabalho.

Exemplo:

```
> list.files("./gujarati/CSV")
[1] "Table_1.1.csv" "Table_1.2.csv" "Table_1.3.csv" "Table_1.4.csv"
```

Notem que se o diretório de trabalho foi definido como C:/pasta, e se os arquivos estão em C:/pasta/gujarati/CSV, basta colocar um ponto e indicar o restante do caminho como argumento para os comandos do R. No caso, ./gujarati/CSV, o que equivale ao diretório C:/pasta/gujarati/CSV.

R

Importação e exportação de arquivos

<u>Lendo arquivos CSV (comma-separated values):</u>

O formato CSV é um dos mais populares porque é bastante simples e muitos pacotes estatísticos fazem importações e exportações com esse formato.

Um exemplo de importação local desse arquivo seria:
samp <- read.csv("C:/seunome/entrada/CSV/Table_1.1.csv)

Se sua pasta de trabalho já está definida no R como C:/seunome, pode ser utilizado o comando:

samp <- read.csv("./entrada/CSV/Table_1.1.csv)

Podemos definir outros argumentos no comando read.csv. Por exemplo, se a primeira linha do conjunto de dados indica os nomes das variáveis (header=TRUE), se o decimal é ponto ou vírgula (dec=”.” ou dec=”,”) e qual é o separador (geralmente, no CSV o separador é a vírgula.)

Também é possível fazer uma importação remota de um arquivo:
Site ← ‘https://github.com/mfsalomao/IntroducaoEconometria/tree/master/wooldridge/CSV’
volat <- read.csv(paste0(site,”/volat.csv”))

Para exportar arquivos, utilizar o comando write.csv(x, file="filename", row.names=FALSE).

R

Importação e exportação de arquivos

Lendo de tabelas HTML:

Às vezes é conveniente importarmos uma tabela diretamente em formato HTML para o R. Existe uma função que faz essa leitura: a função readHTMLTable, do pacote XML.

A primeira coisa é carregar o pacote:
> library('XML')

Caso não esteja instalado no R, instalá-lo a partir do comando install.packages e depois carregá-lo:
> install.packages('XML')
> library('XML')

Feito isso, definir uma URL:
Url <- 'http://en.wikipedia.org/wiki/World_population'

Usar o comando:
tbls <- readHTMLTable(url)

Há várias tabelas no site acima. Se quisermos especificar o número de tabelas, devemos utilizar o which. No caso, para importar apenas a terceira tabela:
tbl <- readHTMLTable(url, which=3)

R

Importação e exportação de arquivos

Lendo arquivos de texto tabulados:

Dizemos que um arquivo é tabulado quando existe uma estrutura definida para delimitar linhas e colunas. A função read.table importa esse tipo de arquivo.

Podemos fazer uma importação via web:

```
tbl <- read.table("ftp://ftp.example.com/download/data.txt")
```

Ou em arquivos localizados em alguma pasta no computador:

```
tbl <- read.table(“.entrada/TXT/exemplo.txt”)
```

Assim como o read.csv, o read.table tem como argumentos o header (TRUE ou FALSE, dependendo se a primeira linha vem ou não com os rótulos das variáveis), o sep (separador), o dec (qual a notação de decimal utilizada etc).

O comando write.table pode ser utilizado para exportar arquivos em formato txt a partir da seguinte estrutura:

```
> write.table(dados, file="dados.txt")
```

R

Importação e exportação de arquivos

Lendo arquivos excel:

Existem funções para importar arquivos XLS ou XLSX. Uma delas é a função read_excel do pacote readxl. Exemplo:

```
> library('readxl')
> rcl <- read_excel(".entrada/Excel/rcl_estados.xlsx", sheet=1)
> rcl <- as.data.frame(rcl)
```

Para exportar arquivos excel, pode ser utiizada a função write.xlsx

```
> library(xlsx)
> write.xlsx(dados, file = "rcl.xlsx", sheetName = "planilha1", row.names = FALSE)
```


esesp

R

Importação e exportação de arquivos

Lendo arquivos excel:

Desafio

Vimos anteriormente que o comando read_excel lê planilha por planilha. Como armazenar em um mesmo objeto um arquivo excel com várias planilhas?

REVISÃO ESTATÍSTICA

Variáveis aleatórias

Lançamento de um dado. Qual a média e a variância amostral se lançarmos um dado 2 vezes?

dado ← sample(1:6, 2, replace=TRUE)
mean(dado)
var(dado)

REVISÃO ESTATÍSTICA

Variáveis aleatórias

Variável aleatória: variável cujo valor é desconhecido até a sua observação. O valor de uma variável aleatória é resultado de um experimento (controlado ou não controlado); não pode ser perfeitamente predito com exatidão. As variáveis aleatórias podem ser classificadas em discretas ou contínuas.

Variável aleatória discreta: só pode tomar um número finito de valores, os quais podem ser contados utilizando-se os inteiros positivos. São enumeráveis.

Variável aleatória contínua: pode tomar qualquer valor real (e não apenas números inteiros) em um intervalo de números reais.

REVISÃO ESTATÍSTICA

Regras de somatório

Sejam **x** e **y** variáveis aleatórias e **a** e **b** constantes:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

$$\sum_{i=1}^{n} a = na$$

$$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

REVISÃO ESTATÍSTICA

Regras de somatório

Sejam **x** e **y** variáveis aleatórias e **a** e **b** constantes:

$$\sum_{i=1}^{n}(a + bx_i) = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n}(ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times x_3 \times \ldots \times x_n$$

REVISÃO ESTATÍSTICA

Regras de somatório

Exercícios

1) Sejam x e y vetores inteiros com 10 elementos com distribuição de probabilidade uniforme, tais que n = 10, limite inferior da distribuição = 15, limite superior da distribuição = 25. Sejam a e b constantes inteiras quaisquer. Faça:

a) somatório de x;
b) somatório de y;
c) somatório de x+y;
d) somatório de x + somatório de y;
e) somatório de a*x;
f) somatório de a*x + somatório de y;
g) somatório de a + somatório de b*y
h) somatório de a*x + somatório de b*y;
i) produtório de x;
j) produtório de 1/y.

Nota: para gerar os valores, utilize o comando runif. Para obter as partes inteiras, utilize o comando floor. Para obter os somatórios, utilize o comando sum (cuidado com as constantes). Para produtórios, prod. Armazene os resultados em variáveis do tipo letraA, letraB etc.

REVISÃO ESTATÍSTICA

Medidas de localização e de dispersão

Média aritmética: $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

Mediana: é o valor que está no meio de uma série ordenada. Por exemplo, em x = {1,2,3,4,5}, a mediana é 3. Se a série for y = {1,2,3,4,5,6}, a mediana é 3,5 (média entre 3 e 4).

Variância amostral: $$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Erro-padrão: a raiz quadrada da variância amostral.

Covariância amostral: $$\mathrm{cov}(\mathrm{x,y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

REVISÃO ESTATÍSTICA

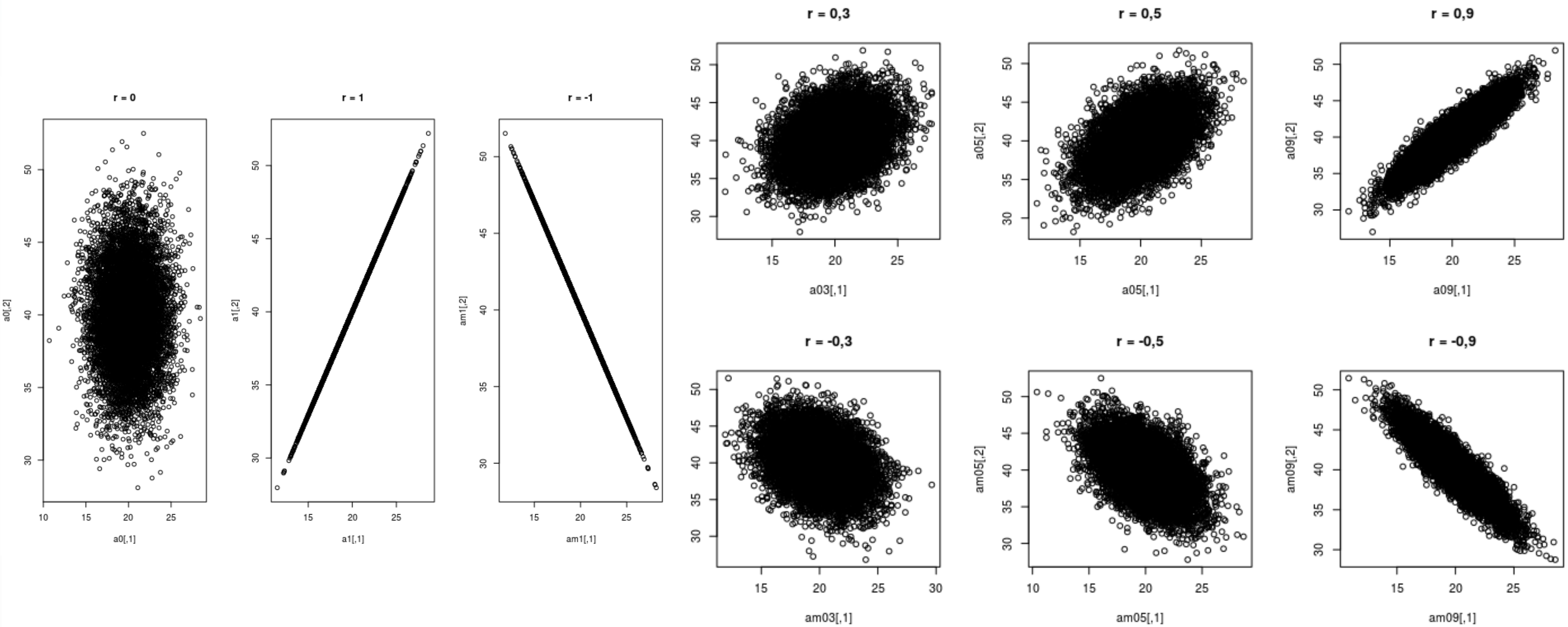
Coeficiente de correlação

Nos revela o grau de associação linear entre duas variáveis. É uma medida adimensional. Varia entre 1 e -1. Se o resultado é positivo, dizemos que as duas variáveis tendem a variar no mesmo sentido. Isto será tanto mais forte quanto o valor estiver próximo de 1. Caso contrário, se for negativo, as variáveis tendem a variar em sentidos opostos. Isto será tanto mais forte quanto o valor estiver próximo de -1. Valores próximos a zero revelam pouco ou nenhum grau de associação linear. Representado por r ou corr(x,y).

$$corr(x,y) = \frac{cov(x,y)}{s_x s_y}$$

REVISÃO ESTATÍSTICA

Coeficiente de correlação



REVISÃO ESTATÍSTICA

Medidas de loc., disp. e coef. de corr.

Exercícios:
Criaremos três variáveis com dados fornecidos pela turma: altura (metros), peso (quilos) e idade.

Obtenham:
a) a média e a mediana de cada variável
b) a variância e o erro-padrão de cada variável
c) correlação entre altura e peso (plotar gráfico)
d) correlação entre altura e idade (plotar gráfico)
e) o IMC de cada integrante da turma.
Dica: IMC = peso/altura²

REVISÃO ESTATÍSTICA

Medidas de loc., disp. e coef. de corr.

Tabela do IMC

Resultado	Situação
Abaixo de 17	Muito abaixo do *peso*
Entre 17 e 18,49	Abaixo do *peso*
Entre 18,5 e 24,99	*Peso* normal
Entre 25 e 29,99	Acima do *peso*
Entre 30 e 34,99	*Obesidade* I
Entre 35 e 39,99	*Obesidade* II (severa)
Acima de 40	*Obesidade* III (mórbida)

No R:

Média: mean(variável) ou sum(variável)/length(variável)

Mediana: median(variável)

Variância: var(variável) ou sum((variável-mean(variável))^2)/(length(variável)-1)

Covariância: cov(variável1,variável2) ou sum((variavel1-mean(variável1))*((variável2-mean(variável2))))/(length(variável1)-1)

Correlação: corr(variável1,variável2) ou cov(variável1,variável2)/(sqrt(var(variável1))*sqrt(var(variável2)))

REVISÃO ESTATÍSTICA

Variáveis aleatórias

Variável aleatória: variável cujo valor é desconhecido até a sua observação. O valor de uma variável aleatória é resultado de um experimento (controlado ou não controlado); não pode ser perfeitamente predito com exatidão. As variáveis aleatórias podem ser classificadas em discretas ou contínuas.

Variável aleatória discreta: só pode tomar um número finito de valores, os quais podem ser contados utilizando-se os inteiros positivos. São enumeráveis.

Variável aleatória contínua: pode tomar qualquer valor real (e não apenas números inteiros) em um intervalo de números reais.

REVISÃO ESTATÍSTICA

Variáveis aleatórias

Uma pesquisa com 1000 indivíduos levantou suas preferências sobre combinações entre tipos de vinho e queijos. Eis os resultados:

	Vinho Branco	Vinho Tinto	Total por Queijo
Gorgonzola	200	270	470
Brie	300	100	400
Outro	60	70	130
Total por Vinho	560	440	1000

No R: abra o arquivo queijos_vinhos.RData

Qual a probabilidade conjunta da combinação vinho tinto e queijo gorgonzola?

REVISÃO ESTATÍSTICA

Variáveis aleatórias

Para calcular as probabilidades, vamos dividir o número de ocorrências favoráveis a cada evento pelo tamanho do espaço amostral.

	Vinho Branco	Vinho Tinto
Gorgonzola	0.20	0.27
Brie	0.30	0.10
Outro	0.06	0.07

Qual a probabilidade conjunta da combinação vinho tinto e queijo gorgonzola? Resposta: 27%.

REVISÃO ESTATÍSTICA

Variáveis aleatórias

	Vinho Branco	Vinho Tinto
Gorgonzola	0.20	0.27
Brie	0.30	0.10
Outro	0.06	0.07

Probabilidades conjuntas f(x,y)

Vinho branco e queijo gorgonzola = 20%
Vinho branco e queijo brie = 30%
Vinho branco e outro tipo de queijo = 6%
Vinho tinto e queijo gorgonzola = 27%
Vinho tinto e queijo brie = 10%
Vinho tinto e outro tipo de queijo = 7%
Note que a soma de todas as probabilidades condicionais é igual a 100%

REVISÃO ESTATÍSTICA

Variáveis aleatórias

	Vinho Branco	Vinho Tinto
Gorgonzola	0.20	0.27
Brie	0.30	0.10
Outro	0.06	0.07

Probabilidades Marginais f(x) ou f(y)

Agora que já sabemos o conceito de probabilidade conjunta, vamos responder às perguntas abaixo:

Qual a probabilidade marginal (incondicional) de encontrarmos combinações de vinho branco? E de vinho tinto? E de queijo gorgonzola? E de queijo brie? E de outro tipo de queijo?

REVISÃO ESTATÍSTICA

Variáveis aleatórias

	Vinho Branco	Vinho Tinto	Total por Queijo
Gorgonzola	0.20	0.27	0.47
Brie	0.30	0.10	0.40
Outro	0.06	0.07	0.13
Total por Vinho	0.56	0.44	1.00

Respostas:

Qual a probabilidade marginal de encontrarmos combinações de vinho branco? 20% + 30% + 6% = 56%

E de vinho tinto? 27% + 10% + 7% = 44%

E de queijo gorgonzola? 20%+ 27% = 47%

E de queijo brie? 30% + 10% = 40%

E de outro tipo de queijo? 6% + 7% = 13%

Note que a soma de todas as probabilidades marginais é igual a 100%

REVISÃO ESTATÍSTICA

Variáveis aleatórias

	Vinho Branco	Vinho Tinto	Total por Queijo
Gorgonzola	0.20	0.27	0.47
Brie	0.30	0.10	0.40
Outro	0.06	0.07	0.13
Total por Vinho	0.56	0.44	1.00

Probabilidades Condicionais

$f(x|y) = f(x,y)/f(y)$ ou $f(y|x) = f(x,y)/f(x)$

Frequentemente, a chance de ocorrência de um evento está condicionada à ocorrência de outro evento. Assim sendo, qual a probabilidade de escolhermos queijo gorgonzola DADO QUE temos somente vinho tinto?

REVISÃO ESTATÍSTICA

Variáveis aleatórias

	Vinho Branco	Vinho Tinto	Total por Queijo
Gorgonzola	0.20	0.27	0.47
Brie	0.30	0.10	0.40
Outro	0.06	0.07	0.13
Total por Vinho	0.56	0.44	1.00

Respostas

1) Qual a probabilidade de escolhermos queijo gorgonzola DADO QUE temos somente vinho tinto? 0,27/0,44 = 0,614 ou 61,4%.

2) Qual a probabilidade de escolhermos queijo brie DADO QUE temos somente vinho tinto? 0,10/0,44 = 0,227 ou 22,7%.

3) Qual a probabilidade de escolhermos outro tipo de queijo DADO QUE temos somente vinho tinto? 0,07/0,44 = 0,159 ou 15,9%.
Note que a soma de todas as probabilidades condicionais é igual a 100%

REVISÃO ESTATÍSTICA

Variáveis aleatórias

As variáveis aleatórias Vinho e Queijo são estatisticamente independentes? Para responder essa pergunta, vamos olhar a tabela abaixo. Ela representa o lançamento simultâneo de duas moedas não viciadas.

	Cara	Coroa	Total Moeda 2
Cara	0.25	0.25	0.5
Coroa	0.25	0.25	0.5
Total Moeda 1	0.50	0.50	1.0

Não é difícil perceber que as probabilidades conjuntas são iguais a 25% e que as probabilidades marginais de se tirar Cara ou Coroa são iguais a 50% cada.

REVISÃO ESTATÍSTICA

Variáveis aleatórias

	Cara	Coroa	Total Moeda 2
Cara	0.25	0.25	0.5
Coroa	0.25	0.25	0.5
Total Moeda 1	0.50	0.50	1.0

Quais são as probabilidades condicionais?

Se a probabilidade condicional $f(x|y)$ é igual à probabilidade marginal $f(x)$, a implicação disso é clara. Dado que a fórmula da probabilidade condicional é $f(x|y) = f(x,y)/f(y)$, substituindo, temos: $f(x) = f(x,y)/f(y)$, o que nos dá que $f(x,y) = f(x)f(y)$, ou seja, que a probabilidade conjunta é igual ao produto das probabilidades marginais. Podemos ver que isto não se aplica no caso dos queijos e vinhos, de modo que estas variáveis NÃO SÃO estatisticamente independentes.

REVISÃO ESTATÍSTICA

Variáveis aleatórias

	Cara	Coroa	Total Moeda 2
Cara	0.25	0.25	0.5
Coroa	0.25	0.25	0.5
Total Moeda 1	0.50	0.50	1.0

Quais são as probabilidades condicionais?

Dado que tiramos Cara na primeira moeda, qual a probabilidade de tirar Cara na segunda moeda? 0,25/0,50 = 0,5 ou 50%. E de tirar Coroa na segunda moeda, dado que tiramos Cara na primeira moeda? Igualmente, 0,25/0,50 = 0,5 ou 50%.

Podemos concluir que a probabilidade condicional de tirar Cara na segunda moeda, dado que tiramos Cara na primeira, é igual à probabilidade de tirarmos Cara na segunda moeda independentemente do que tirarmos na primeira. O mesmo vale para Coroa. Em outras palavras, as probabilidades condicionais são iguais às probabilidades marginais (incondicionais), ou seja, o que tiramos em uma moeda não afeta em nada o que tiramos na outra.

REVISÃO ESTATÍSTICA

Esperança matemática

Seja X uma variável aleatória discreta que consiste no lançamento simultâneo de dois dados. Seja f(X) as probabilidades de cada um desses lançamentos:

$$x = \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12$$

$$f(x) = \left(\frac{1}{36}\right)\left(\frac{2}{36}\right)\left(\frac{3}{36}\right)\left(\frac{4}{36}\right)\left(\frac{5}{36}\right)\left(\frac{6}{36}\right)\left(\frac{5}{36}\right)\left(\frac{4}{36}\right)\left(\frac{3}{36}\right)\left(\frac{2}{36}\right)\left(\frac{1}{36}\right)$$

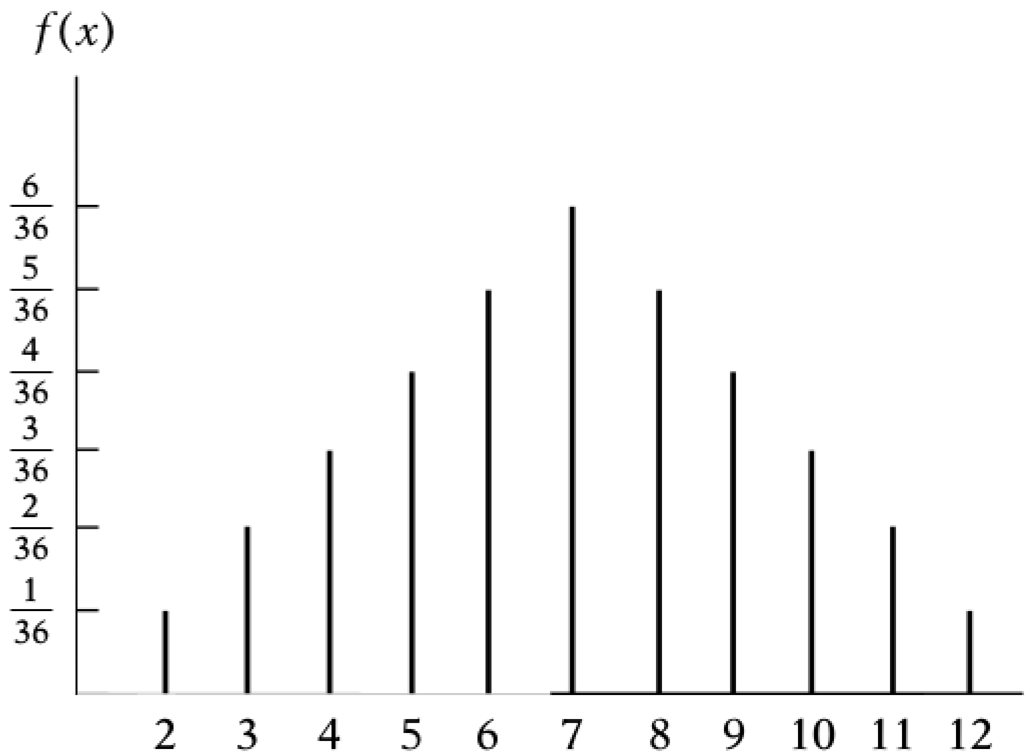
Qual o valor esperado de X? Representamos a esperança matemática ou o valor esperado de X por E(X).

$E(X) = \sum_{i=1}^{n} X_i f(X_i)$. Fazendo os cálculos, encontramos E(X)=7.

REVISÃO ESTATÍSTICA

Esperança matemática

O gráfico abaixo ilustra a Função Densidade de Probabilidade (FDP) de X.



REVISÃO ESTATÍSTICA

Esperança matemática

Algumas propriedades da esperança matemática:
1) se b é uma constante, $E(b) = b$

2) se a e b são constantes, $E(aX+b) = aE(X) + b$

3) se X e Y são variáveis aleatórias independentes, então $E(XY) = E(X)E(Y)$

4) seja X uma variável aleatória e seja $E(X) = \mu$.
Então $var(X) = E(X - \mu)^2 = E(X^2) - [E(X)]^2$

5) Propriedades da variância:
5.1) a variância de uma constante é zero.
5.2) sejam a e b constantes, $var(aX + b) = a^2 var(X)$ (demonstre)
5.3) se X e Y são va independentes, então:
$var(X + Y) = var(X) + var(Y)$ e $var(X - Y) = var(X) + var(Y)$
$var(aX + bY) = a^2 var(X) + b^2 var(Y)$

6) $cov(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - \mu_x \mu_y$

7) se X e Y forem independentes, $cov(X, Y) = 0$

8) O coeficiente de autocorrelação populacional é $\rho = \frac{\operatorname{cov}(X, Y)}{\sqrt{\{\operatorname{var}(X)\operatorname{var}(Y)\}}} = \frac{\operatorname{cov}(X, Y)}{\sigma_x \sigma_y}$

REVISÃO ESTATÍSTICA

Esperança matemática

Para variáveis aleatórias discretas, a esperança matemática é simbolizada pelo somatório:

$$E(X) = \sum xf(x)$$

Para variáveis aleatórias contínuas, o somatório é substituído pela integral definida:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

REVISÃO ESTATÍSTICA

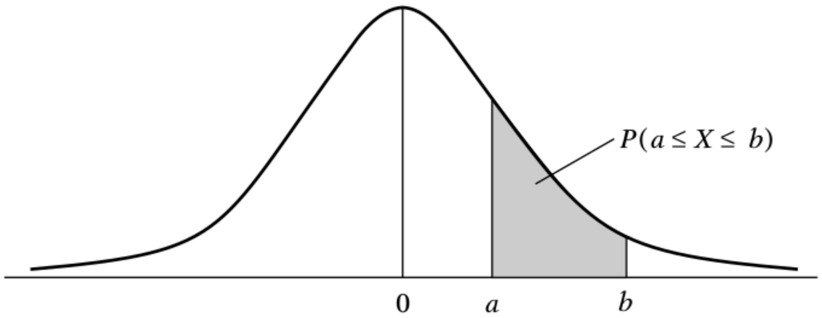
FDP

Seja f(x) a função densidade de probabilidade (FDP) de uma variável aleatória. As seguintes propriedades são válidas para as variáveis aleatórias contínuas:

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$\int_{a}^{b} f(x)\,dx = P(a \leq x \leq b)$$



REVISÃO ESTATÍSTICA

FDP

Considere que uma distribuição de probabilidade contínua seja definida pela seguinte função densidade de probabilidade (FDP):

$$f(x) = \frac{1}{9}x^2 \qquad 0 \leq x \leq 3$$

Dado o intervalo definido acima, a probabilidade de que x se encontre nesse intervalo é 100%. Ou seja:

$\int_0^3 \frac{1}{9}x^2 dx = 1$ o que significa que $\frac{1}{27}x^3|_0^3 = 1$,ou seja,

$$\frac{1}{27}x^3|_0^3 = \frac{1}{27}3^3 - \frac{1}{27}0^3 = 1$$

REVISÃO ESTATÍSTICA

FDP

Considere que uma distribuição de probabilidade contínua seja definida pela seguinte função densidade de probabilidade (FDP):

$$f(x) = \frac{1}{9}x^2 \qquad 0 \leq x \leq 3$$

Se quisermos calcular a probabilidade em um intervalo menor, como por exemplo 0 e 1, a probabilidade também deve ser menor do que 1.

$\int_0^1 \frac{1}{9}x^2 dx = \frac{1}{27}$ o que significa que $\frac{1}{27}x^3|_0^1 = \frac{1}{27}$,ou seja,

$$\frac{1}{27}x^3|_0^1 = \frac{1}{27}1^3 - \frac{1}{27}0^3 = \frac{1}{27}$$

REVISÃO ESTATÍSTICA

FDP

Considere que uma distribuição de probabilidade contínua seja definida pela seguinte função densidade de probabilidade (FDP):

$$f(x) = \frac{1}{9}x^2 \qquad 0 \leq x \leq 3$$

Qual o valor esperado da PDF acima? Expressamos o valor esperado pela fórmula:

$$E(X) = \int_0^3 x\left(\frac{x^2}{9}\right)dx$$

Calculado a integral, obtemos:

$$\frac{1}{9}\left[\left(\frac{x^4}{4}\right)\right]_0^3 = \frac{9}{4} = 2.25$$

REVISÃO ESTATÍSTICA

FDP

Considere que uma distribuição de probabilidade contínua seja definida pela seguinte função densidade de probabilidade (FDP):

$$f(x) = \frac{1}{9}x^2 \qquad 0 \leq x \leq 3$$

Qual a variância da PDF acima? Lembre-se da fórmula da variância: $\text{var}(X) = E(X^2) - [E(X)]^2$
Logo:

$$E(X^2) = \int_0^3 x^2 \left(\frac{x^2}{9}\right) dx$$
$$= \int_0^3 \frac{x^4}{9} dx$$
$$= \frac{1}{9}\left[\frac{x^5}{5}\right]_0^3$$
$$= 243/45$$

Como $E(X) = \frac{9}{4}$, temos que:

$$\text{var}(X) = 243/45 - \left(\frac{9}{4}\right)^2$$
$$= 243/720 = 0.34$$

REVISÃO ESTATÍSTICA

Esperança Condicional

Seja f(x,y) a FDP conjunta das variáveis aleatórias X e Y. O valor esperado condicional de X, dado Y=y é definido como:

$$E(X \mid Y = y) = \sum x f(x \mid Y = y)$$ para V.A. discretas.

$$= \int_{-\infty}^{\infty} x f(x \mid Y = y) dx$$ para V.A. contínuas.

Note que E(X|Y) é uma V.A., mas E(X|Y=y) é uma constante, uma vez que y é um valor específico de Y.

REVISÃO ESTATÍSTICA

Esperança Condicional

Exercicio

Considere a tabela com as probabilidades conjuntas das variáveis aleatórias discretas X e Y abaixo:

		X			
		−2	**0**	**2**	**3**
Y	**3**	0.27	0.08	0.16	0
	6	0	0.04	0.10	0.35

Obtenha E(Y|X=2). Dica: aplique a fórmula da esperança matemática utilizando como FDP as probabilidades condicionais de Y quando X=2.

Obtenha E(Y) utilizando a lei das expectativas iteradas, que estabelece a seguinte relação entre esperanças condicionais e incondicionais: $E(Y) = E_X[E(Y \mid X)]$

REVISÃO ESTATÍSTICA

Distribuição Normal

Uma variável aleatória contínua X é normalmente distribuída se sua PDF tem a seguinte forma:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \qquad -\infty < x < \infty$$

Os únicos parâmetros da distribuição normal (ou gaussiana) são a média e a variância. Uma vez especificados, é possível encontrar qualquer probabilidade desta distribuição. A distribuição normal apresenta algumas características:

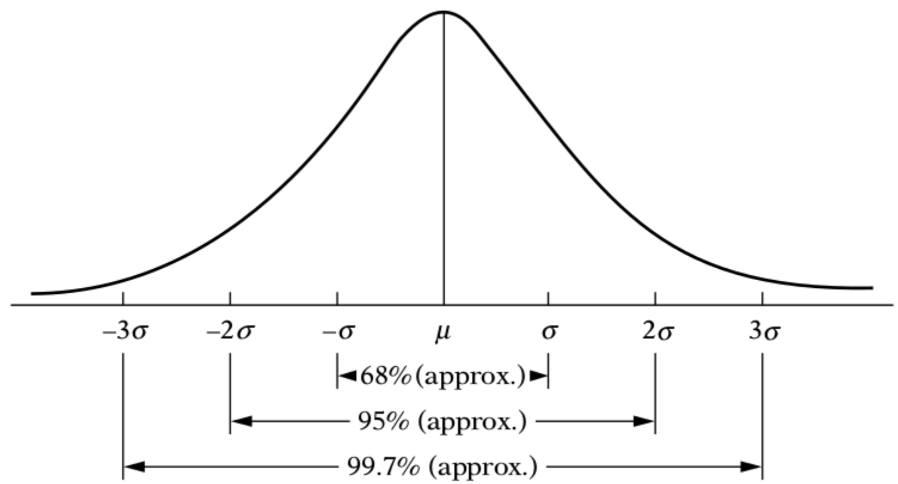
1 – é simétrica em relação à média;

2 – aproximadamente 68% da área da distribuição normal está entre 1 desvio-padrão da média, cerca de 95% da área está entre 2 DP da média e cerca de 99,7% da área está entre 3 DP da média.

REVISÃO ESTATÍSTICA

Distribuição Normal

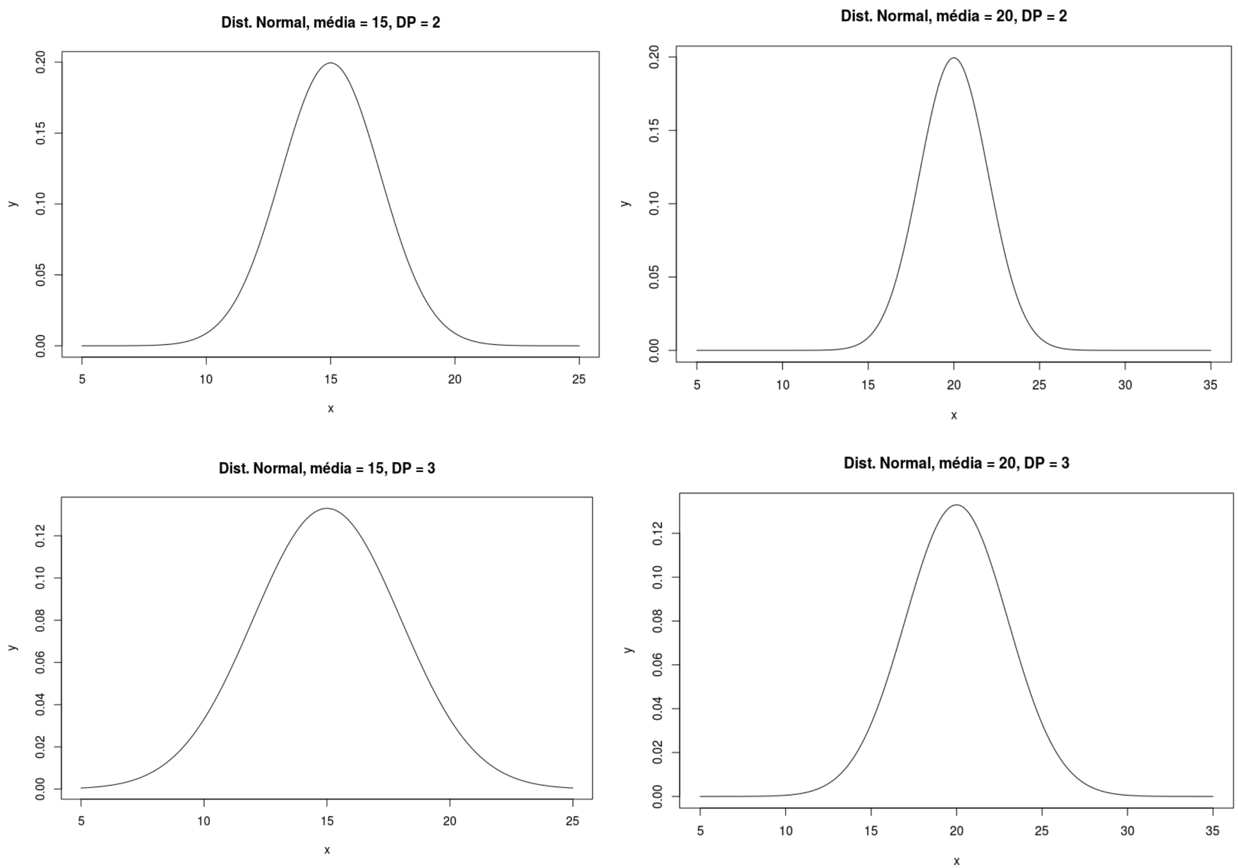
Graficamente, a distribuição normal apresenta uma conhecida forma de sino em torno da média:



Uma V.A. normal X é representada como $X \sim N(\mu, \sigma^2)$, o que significa que X é uma V.A normalmente distribuída Com média μ e desvio-padrão σ.

REVISÃO ESTATÍSTICA

Distribuição Normal



No R:

Gráfico direito superior:

x <- seq(5,35,length=1000)

y <- dnorm(x,mean=20, sd=2)

plot(x,y, type="l", lwd=1, main="Dist. Normal, média = 20, DP = 2")

REVISÃO ESTATÍSTICA

Distribuição Normal

Podemos perceber que existem várias possibilidades de construção de uma distribuição normal dependendo dos parâmetros informados: média e variância. Para obter determinada probabilidade em um dado intervalo seria necessário calcular a integral da FDP da distribuição normal:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \qquad -\infty < x < \infty$$

REVISÃO ESTATÍSTICA

Distribuição Normal

No entanto, existem formas mais fáceis de se obter qualquer probabilidade sem ter de calcular a integral da FDP da distribuição normal. Basta padronizar a variável aleatória de interesse, digamos X~N(8,4) para a V.A Z~N(0,1), que já está tabulada. A fórmula de conversão é: $Z = \frac{x-\mu}{\sigma}$

Por exemplo, qual a probabilidade de que X assuma um valor entre X_1=4 e X_2=12? Queremos Pr(4≤X≤12):

$$Z_1 = \frac{X_1-\mu}{\sigma} = \frac{4-8}{2} = -2$$

$$Z_2 = \frac{X_2-\mu}{\sigma} = \frac{12-8}{2} = +2$$

Olhando a tabela da distribuição normal, obtemos $\Pr(0 \leq Z \leq 2) = 0.4772$. Por simetria: $\Pr(-2 \leq Z \leq 0) = 0.4772$. Logo:
0,4772 + 0,4772 = 0,9544.
Pergunta: e a probabilidade de que X exceda 12?

REVISÃO ESTATÍSTICA

Distribuição Normal

Exercício

Suponha que o tempo necessário para atendimento de clientes na fila de um banco seja normalmente distribuído com média igual a 8 minutos e variância de 4 minutos. Qual a probabilidade de que um atendimento dure:

a) menos do que 5 minutos?
b) mais do que 10 minutos?
c) entre 7 e 9 minutos?
d) 75% dos atendimentos requerem no mínimo quanto tempo de atendimento?

Respostas: a) Pr = 6,68%; b) Pr = 15,87%; c) Pr = 38,3%; d) X = 6,66 (Pr=25%,Z=-0,67)

REVISÃO ESTATÍSTICA

Distribuição Normal

Exercício

Suponha que o tempo necessário para atendimento de clientes na fila de um banco seja normalmente distribuído com média igual a 8 minutos e variância de 4 minutos. Qual a probabilidade de que um atendimento dure:

a) menos do que 5 minutos? pnorm(5,8,2)
b) mais do que 10 minutos? 1-pnorm(10,8,2)
c) entre 7 e 9 minutos? pnorm(9,8,2)-pnorm(7,8,2)
d) 75% dos atendimentos requerem no mínimo quanto tempo de atendimento? qnorm(.25,8,2)

Respostas: a) Pr = 6,68%; b) Pr = 15,87%; c) Pr = 38,3%; d) X = 6,66 (Pr=25%,Z=-0,67)

esesp

REVISÃO ESTATÍSTICA

Teorema Central do Limite

Se Y1,…,YN forem variáveis aleatórias idêntica e independentemente distribuídas com média μ e desvio-padrão σ, e $\bar{Y} = \frac{\sum Y_i}{N}$ então:

$$Z_N = \frac{\bar{Y} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

possui uma distribuição de probabilidade que converge para a normal padronizada N(0,1) quando N →∞.

REVISÃO ESTATÍSTICA

Teorema Central do Limite

Exemplo de aplicação prática do Teorema Central do Limite (simulação):

1) construa um vetor com a distribuição uniforme com 2000 observações, limite mínimo de 0 e limite máximo de 100.

2) repita o experimento 2000 vezes e armazene o resultado num data-frame. Veja o histograma.

3) construa um vetor com as médias de cada experimento (cada coluna do data-frame).

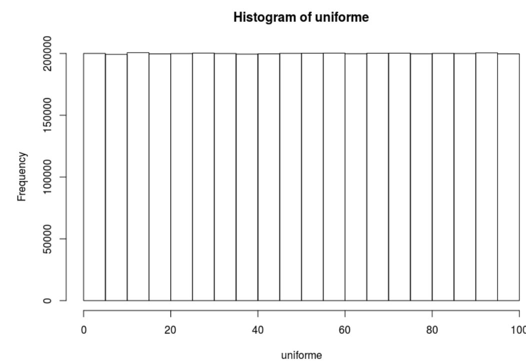
4) note que este vetor tenderá a apresentar uma distribuição normal com média μ e desvio-padrão $\frac{\sigma}{\sqrt{N}}$ da distribuição uniforme. Veja o histograma.

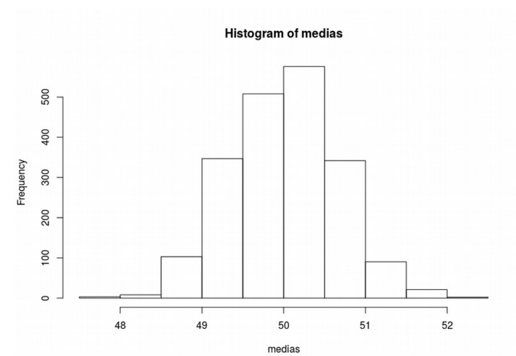
REVISÃO ESTATÍSTICA

Teorema Central do Limite

Exemplo de aplicação prática do Teorema Central do Limite (simulação):

```r
# constroi data frame de 2000 colunas com 2000 obs da dist. uniforme (lim min=0, lim max=100)
uniforme <- data.frame()
uniforme <- runif(2000, 0,100)
i = 1
while (i <= 1999) {
  uniforme <- cbind(uniforme,runif(2000, 0,100))
  i <- i+1
}

# constroi vetor de medias de cada coluna da distribuição uniforme
medias <- vector()
medias <- mean(uniforme[,1])
for (i in 2:2000) {
  medias[i] <- mean(uniforme[,i])
}

# constroi vetor de desvios-padrão de cada coluna da distribuição uniforme
dp <- vector()
dp <- sd(uniforme[,1])
for (i in 2:2000) {
  dp[i] <- sd(uniforme[,i])/sqrt(length(uniforme[,i]))
}

# a média do vetor de médias tende a ser igual a média da distribuição uniforme
mean(medias)
mean(uniforme[,1]) # ou qualquer outra coluna do data.frame

# o desvio-padrão do vetor de médias tende a ser dp/raiz(dp) da distribuição uniforme
sd(medias)
mean(dp)

# o histograma do vetor de médias tende a se aproximar de uma distribuição normal
hist(medias)
```



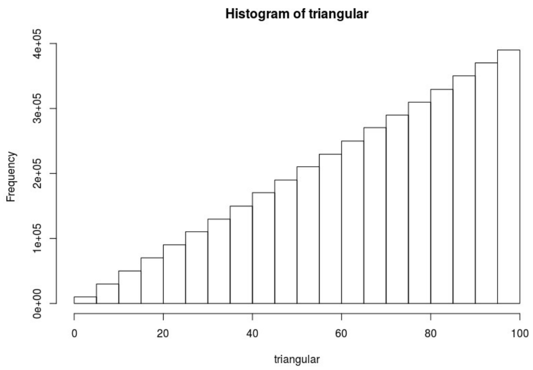


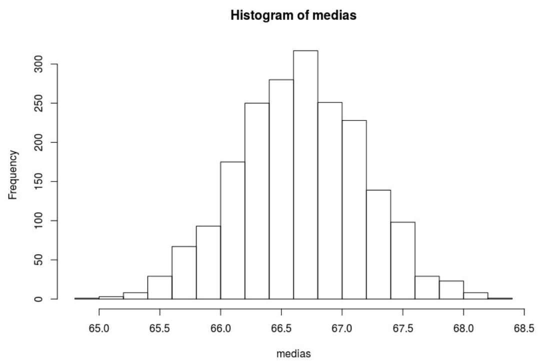
REVISÃO ESTATÍSTICA

Teorema Central do Limite

Outro exemplo, agora a partir da distribuição triangular (simulação):

```
library('triangle')
# constroi data frame de 2000 colunas com 2000 obs da dist. triangular (lim min=0, lim max=100)
triangular <- data.frame()
triangular <- rtriangle(2000, 0,100,100)
i = 1
while (i <= 1999) {
  triangular <- cbind(triangular,rtriangle(2000, 0,100,100))
  i <- i+1
}

# constroi vetor de medias de cada coluna da distribuição triangular
medias <- vector()
medias <- mean(triangular[,1])
for (i in 2:2000) {
  medias[i] <- mean(triangular[,i])
}

# constroi vetor de desvios-padrão de cada coluna da distribuição triangular
dp <- vector()
dp <- sd(triangular[,1])
for (i in 2:2000) {
  dp[i] <- sd(triangular[,i])/sqrt(length(triangular[,i]))
}

# a média do vetor de médias tende a ser igual a média da distribuição triangular
mean(medias)
mean(triangular[,1]) # ou qualquer outra coluna do data.frame

# o desvio-padrão do vetor de médias tende a ser dp/raiz(dp) da distribuição triangular
sd(medias)
mean(dp)

# o histograma do vetor de médias tende a se aproximar de uma distribuição normal
hist(medias)
```





REVISÃO ESTATÍSTICA

Outras Distribuições de Probabilidade

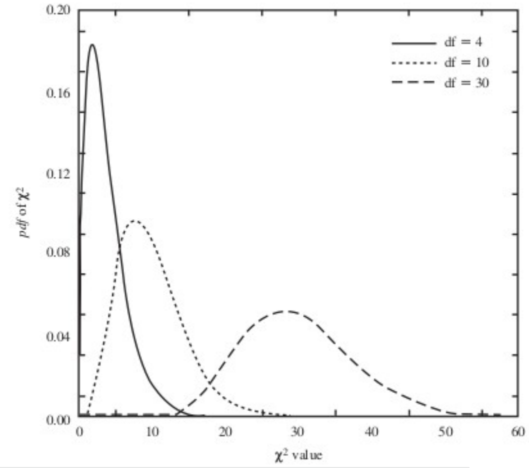
Distribuição Qui-Quadrado (chi-square)

Considere n variáveis aleatórias normais padronizadas Z. Seja: $V = Z_1^2 + Z_2^2 + \ldots + Z_m^2 \sim \chi^2_{(m)}$
A notação $V \sim \chi^2_{(m)}$ significa que a variável aleatória V possui distribuição qui-quadrado com m graus de liberdade. O parâmetro de graus de liberdade m indica o número de variáveis N(0,1) que são somadas e elevadas ao quadrado para compor V. O parâmetro que define a forma da distribuição qui-quadrado é m. À medida que m aumenta, a distribuição qui-quadrado tende a se aproximar da normal.

Propriedades:

$$E[V] = E\left[\chi^2_{(m)}\right] = m$$

$$\mathrm{var}[V] = \mathrm{var}\left[\chi^2_{(m)}\right] = 2m$$



REVISÃO ESTATÍSTICA

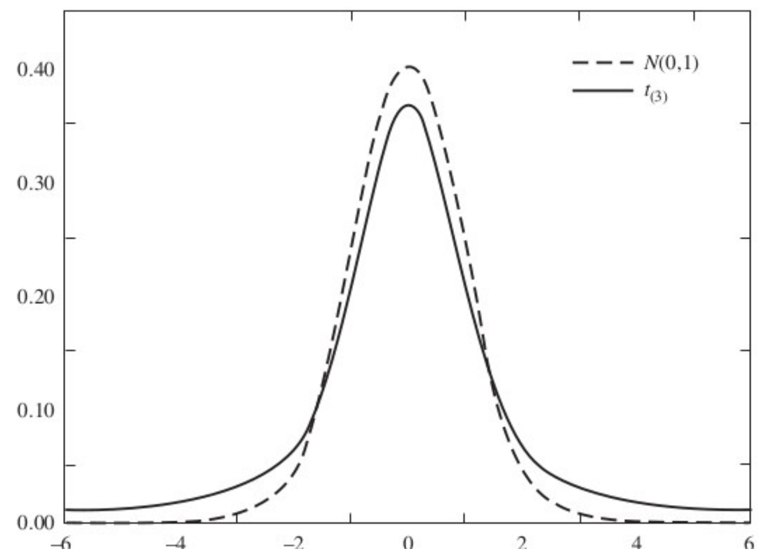
Outras Distribuições de Probabilidade

Distribuição t de Student

Uma variável aleatória com distribuição de probabilidade t é formada pela divisão de uma variável aleatória normal padronizada Z~N(0,1) pela raiz quadrada da razão de uma variável aleatória independente $V \sim \chi^2_{(m)}$ por seus graus de liberdade.

Formalmente: $t = \dfrac{Z}{\sqrt{V/m}} \sim t_{(m)}$

Propriedades: E[t(m)] = 0
Var [t(m)] = m/(m-2)



REVISÃO ESTATÍSTICA

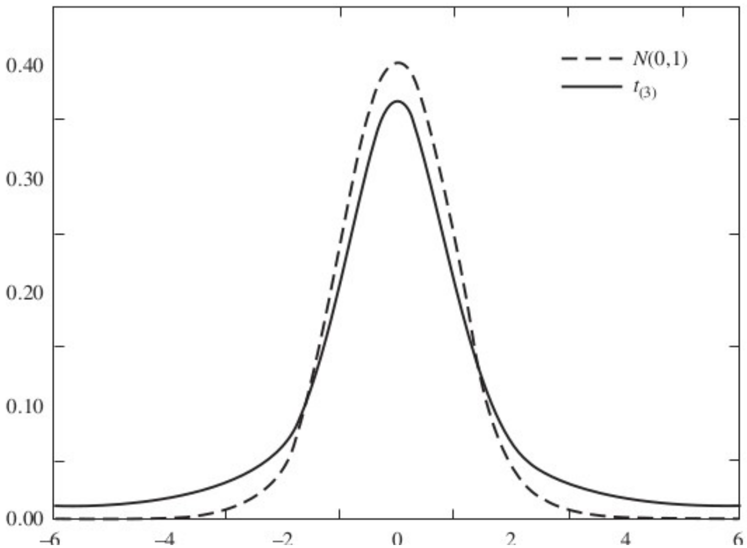
Outras Distribuições de Probabilidade

Distribuição F

Uma variável aleatória com distribuição de probabilidade F é formada pela razão de duas variáveis aleatórias independentes qui-quadrado divididas por seus respectivos graus de liberdade.

Formalmente: $$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1,m_2)}$$

A forma de F é determinada por seus graus de liberdade no numerador e no denominador.



REVISÃO ESTATÍSTICA

Outras Distribuições de Probabilidade

Distribuições no R

```
# Qui-quadrado com 1000 observações e 10 df
x <- rchisq(1000,10)
y <- dchisq(x,10)
plot(x,y)

# Qui-quadrado com 1000 observações e 100 df
x <- rchisq(1000,100)
y <- dchisq(x,100)
plot(x,y)


# t com 1000 observações e 10 df
x <- rt(1000,10)
y <- dt(x,10)
plot(x,y)

# t com 1000 observações e 100 df
x <- rt(1000,100)
y <- dt(x,100)
plot(x,y)

# F com 1000 observações e 10 df no numerador e 20 df no denominador
x <- rf(1000,10,20)
y <- df(x,10,20)
plot(x,y)

# F com 1000 observações e 1000 df no numerador e 200 df no denominador
x <- rf(1000,100,200)
y <- df(x,100,200)
plot(x,y)
```

REVISÃO ESTATÍSTICA

Outras Distribuições de Probabilidade

Exercícios

Sejam X1, X2,…,Xn variáveis aleatórias independentes com a mesma distribuição de probabilidade, com média μ e desvio-padrão σ. A média amostral dessas variáveis é representada por $\frac{\sum_{i=1}^{N} X_i}{n} = \bar{X}$

Utilizando as propriedades de valor esperado, mostre que:

1) O valor esperado de $\bar{X}$ é μ.

2) O desvio-padrão de $\bar{X}$ é $\frac{\sigma}{\sqrt{N}}$

TRATAMENTO DE DADOS

Deflacionamento

Séries econômicas têm um efeito preço (variação nominal) e um efeito quantidade (variação real). Muitas vezes é conveniente analisarmos apenas a variação real da série, ou seja, o quanto a série aumentou ou diminuiu sem considerarmos o efeito da inflação. A equação de Fisher nos diz o seguinte:

$$r = \frac{1+i}{1+\pi} - 1$$ ou, rearranjando-se os termos: $$\pi = \frac{1+i}{1+r} - 1$$

Onde r é a taxa de variação real, i é a taxa de variação nominal e π é a taxa de inflação do período.

esesp

TRATAMENTO DE DADOS

Deflacionamento

Números-Índices são medidas criadas para captar a variação relativa a uma data-base. Seja Δ uma taxa. Se a data-base é t0, e a fixamos em 100, então em t1 teremos o número-índice (1+Δt0)*100. E, assim, sucessivamente:

$$t_2 = (1 + \Delta_{t_1}) \times (1 + \Delta_{t_0}) \times 100$$

$$t_3 = (1 + \Delta_{t_2}) \times (1 + \Delta_{t_1}) \times (1 + \Delta_{t_0}) \times 100$$

$$t_n = (1 + \Delta_{t_{n-1}}) \times \ldots \times (1 + \Delta_{t_2}) \times (1 + \Delta_{t_1}) \times (1 + \Delta_{t_0}) \times 100$$

Ou, na notação de produtório:

$$t_n = 100 \times \prod_{i=0}^{n-1} (1 + \Delta_{t_i})$$

Exemplo do IPCA numa planilha:

A	B	C	D
Ano.Mês	IPCA – Taxa	IPCA – Núm. Ind.	Fórmula
1993.12	36,84	100,00	100,00
1994.01	41,31	141,31	C2*(1+B3/100)
1994.02	40,27	198,22	C3*(1+B4/100)
1994.03	42,75	282,96	C4*(1+B5/100)
1994.04	42,68	403,73	C5*(1+B6/100)

TRATAMENTO DE DADOS

Deflacionamento

Vamos construir os números-índices do IPCA no R. Para tanto, iremos utilizar a série ipca_var, retirada do site IPEADATA, que mostra a variação % mensal do IPCA de dezembro de 1993 a agosto de 2017. Nossa data-base será dezembro de 1993.

Podemos fazer o seguinte procedimento: criar um vetor nulo chamado ipca_ind e colocar como primeiro valor o número 100 (esta será a data-base). Em seguida, aplicar um laço for para iterar o produtório mencionado no slide anterior. Eis o código, que pode ser colocado num script do R:

```r
ipca_ind <- NULL   #    Cria vetor nulo
ipca_ind[1] <- 100 #    Mês-base: dez-1993 igual a 100


for(i in 2:length(ipca_var)) {  # Início do laço for
  ipca_ind[i] <- ipca_ind[i-1]*(1+ipca_var[i]/100)
} #fim do laço for

ipca_ind <- ts(ipca_ind, start=c(1993,12), frequency=12)  #Transforma em série temporal
```

TRATAMENTO DE DADOS

Deflacionamento

Um exemplo prático: o ICMS do ES em dezembro de 2014 era de R$ 7.653.548,96. Um ano depois, em dezembro de 2015, a arrecadação foi de ICMS registrou o valor de R$ 7.469.915,35. Percebemos que houve uma queda de -2,39%. Porém, em 2015, a inflação oficial, medida pelo IPCA, foi de 10,67%. Portanto, em termos reais a queda foi muito maior. Pela equação de Fisher, i = 0,02399326 e π = 0,106735. Então:

r = (1-0,02399326) / (1+0,106735) – 1
r = - 0,1181207

Isto significa que, em termos reais, o ICMS de 2015 foi 11,81% menor do que o ICMS de 2014.

Vamos verificar isto no R?

TRATAMENTO DE DADOS

Deflacionamento

Série mensal do IPCA no site IPEADATA, dez=100. Esta série já está no R com o nome ipca_ind, de dezembro de 1993 a agosto de 2017.

Série mensal do ICMS do Espírito Santo, valores nominais, conforme relatório do SIGEFES, período de janeiro de 2014 a agosto de 2017. Esta série se encontra no R com o nome de ICMS.

Veja que os períodos são diferentes. Vamos restringir a série do IPCA ao mesmo período da série do ICMS?
> ipca_restrito ← window(ipca_ind, start=c(2014,1))

Agora vamos deflacionar a série do ICMS a preços de dezembro de 2014.

Usaremos a fórmula: $ICMS_r = ICMS_i \times \frac{ipca_{[dez2014]}}{ipca_i}$

TRATAMENTO DE DADOS

Deflacionamento

No R:

> icms_real ← icms*(ipca_restrito[12]/ipca_restrito)   (deflaciona o ICMS)

> var_icms ← (icms[24]/icms[12]) – 1   (variação nominal do ICMS)

> var_icms_real ← (icms_real[24]/icms_real[12]) – 1 (var. real do ICMS)

O que encontraremos se fizermos:
> (1+var_icms) / (1+var_icms_real) – 1   (este número é familiar?)

Como exercício, encontre os itens abaixo para 2016, utilizando como base para o deflacionamento o valor de dezembro de 2015:
a) série deflacionada do ICMS
b) variação nominal do ICMS
c) variação real do ICMS
d) taxa de inflação de 2016 (IPCA)
e) gráfico com as duas séries de ICMS, real e nominal (dica: ts.plot)

TRATAMENTO DE DADOS

Deflacionamento

Exercícios:

Abra a série IR no Rstudio. Analise-a graficamente.

1) Importe a série do IGP-M (taxas) no IPEADATA utilizando o comando read.csv2.

2) Construa números-índices para a série do IGP-M.

3) Deflacione a série IR pelo IGP-M utilizando dezembro de 2016 como ano-base.

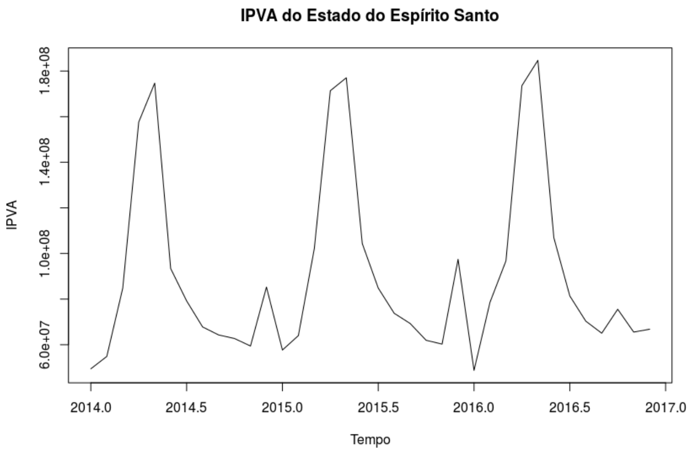
4) Plote um gráfico da série nominal e outro da série real (deflacionada).

TRATAMENTO DE DADOS

Dessazonalização

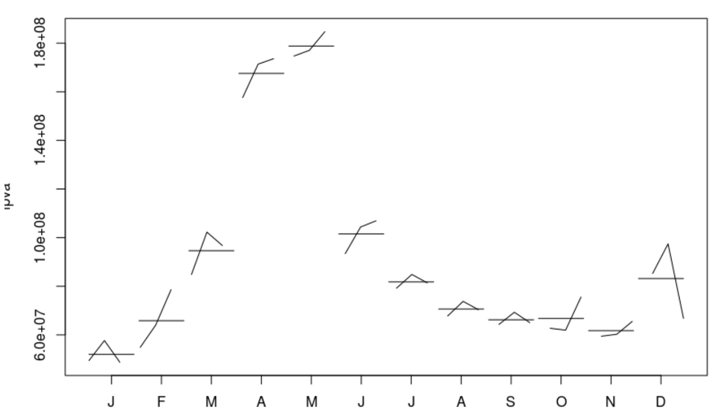
Observe o comportamento da série do IPVA do ES no período entre janeiro de 2014 a dezembro de 2016:

>plot.ts(ipva, xlab="Tempo", ylab="IPVA", main="IPVA do Estado do Espírito Santo")



Note que, recorrentemente, nos meses de abril e maio há um pico de arrecadação. Isto pode ser confirmado por um outro gráfico muito útil:

>monthplot(ipva)



TRATAMENTO DE DADOS

Dessazonalização

Por que dessazonalizar? Suponha que você queira avaliar o impacto de determinada política econômica sobre a taxa de desemprego e das vendas no comércio. O comportamento favorável destas variáveis em dezembro, por exemplo, pode decorrer em virtude de características da época (empregos temporários ou consumo de natal), não necessariamente da política econômica em si.

Há diversos métodos de dessazonalização. Em tese, qualquer série temporal pode ser decomposta nos seguintes elementos:

1) tendência
2) aleatórios
3) sazonais

A dessazonalização consiste na remoção dos elementos sazonais.

TRATAMENTO DE DADOS

Dessazonalização

Aplicaremos no R o método de decomposição aditiva de médias móveis, tomando como exemplo a série do IPVA. É muito simples e prático. Para melhorar a efetividade do método, o aplicaremos sobre a série logaritmizada (discutiremos logaritmos em detalhes mais adiante).

> ipva_decomposto ← decompose(log(ipva), type=”additive”)

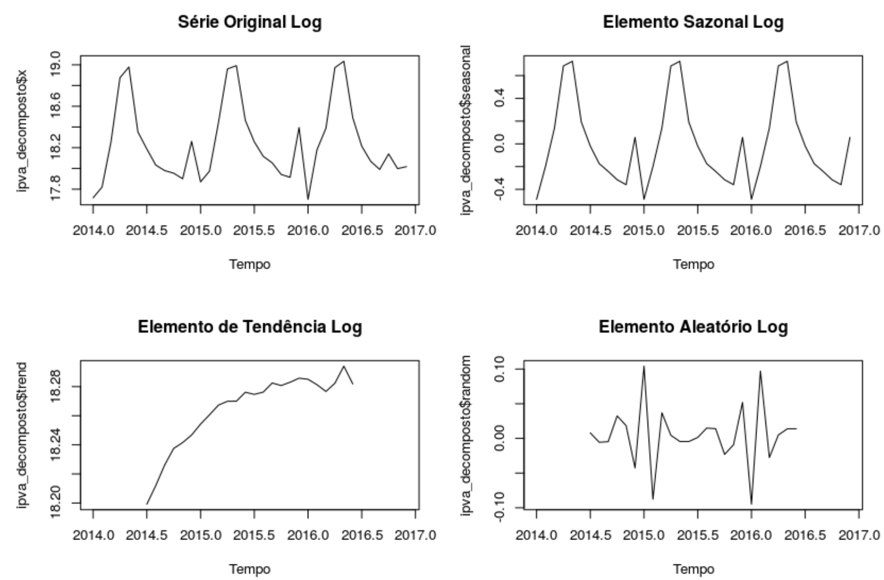
```
$x
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2014 17.71634 17.82027 18.25627 18.87580 18.97866 18.35243 18.18746 18.03191 17.97897 17.95399 17.90028 18.26130
2015 17.87016 17.97436 18.44307 18.95945 18.99196 18.46348 18.25635 18.11657 18.05393 17.94167 17.91423 18.39482
2016 17.70187 18.17980 18.38783 18.97230 19.03424 18.48754 18.21459 18.06863 17.99051 18.14030 17.99819 18.01722

$seasonal
             Jan         Feb         Mar         Apr         May         Jun         Jul         Aug         Sep         Oct
2014 -0.48829164 -0.19855581  0.13880684  0.68511367  0.72641734  0.19187043 -0.01970775 -0.17456158 -0.24258084 -0.31595346
2015 -0.48829164 -0.19855581  0.13880684  0.68511367  0.72641734  0.19187043 -0.01970775 -0.17456158 -0.24258084 -0.31595346
2016 -0.48829164 -0.19855581  0.13880684  0.68511367  0.72641734  0.19187043 -0.01970775 -0.17456158 -0.24258084 -0.31595346
             Nov         Dec
2014 -0.35969694  0.05713973
2015 -0.35969694  0.05713973
2016 -0.35969694  0.05713973

$trend
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug      Sep      Oct      Nov      Dec
2014       NA       NA       NA       NA       NA       NA 18.19922 18.21205 18.22625 18.23752 18.24156 18.24674
2015 18.25424 18.26063 18.26728 18.26989 18.26996 18.27611 18.27466 18.27621 18.28247 18.28070 18.28300 18.28576
2016 18.28502 18.28129 18.27665 18.28228 18.29405 18.28182       NA       NA       NA       NA       NA       NA

$random
              Jan          Feb          Mar          Apr          May          Jun          Jul          Aug          Sep          Oct
2014           NA           NA           NA           NA           NA           NA  0.007952117 -0.005575522 -0.004693472  0.032421961
2015  0.104213257 -0.087722075  0.036973602  0.004439142 -0.004415497 -0.004497451  0.001398376  0.014926016  0.014043965 -0.023071467
2016 -0.094862764  0.097072569 -0.027623109  0.004911352  0.013765991  0.013847944           NA           NA           NA           NA
              Nov          Dec
2014  0.018421154 -0.042573293
2015 -0.009070660  0.051923787
2016           NA           NA

$figure
 [1] -0.48829164 -0.19855581  0.13880684  0.68511367  0.72641734  0.19187043 -0.01970775 -0.17456158 -0.24258084 -0.31595346
[11] -0.35969694  0.05713973

$type
[1] "additive"

attr(,"class")
[1] "decomposed.ts"
```

TRATAMENTO DE DADOS

Dessazonalização

O método retorna as seguintes séries:
x: a série original logaritmizada
seasonal: elemento sazonal logaritmizado
trend: elemento de tendência logaritmizado
random: elemento aleatório logaritmizado

À direita, os gráficos correspondentes a cada elemento:




```
> par(mfrow=c(2,2))
> plot.ts(ipva_decomposto$x, xlab="Tempo", main="Série Original Log")
> plot.ts(ipva_decomposto$seasonal, xlab="Tempo", main="Elemento Sazonal Log")
> plot.ts(ipva_decomposto$trend, xlab="Tempo", main="Elemento de Tendência Log")
> plot.ts(ipva_decomposto$random, xlab="Tempo", main="Elemento Aleatório Log")
```

TRATAMENTO DE DADOS

Dessazonalização

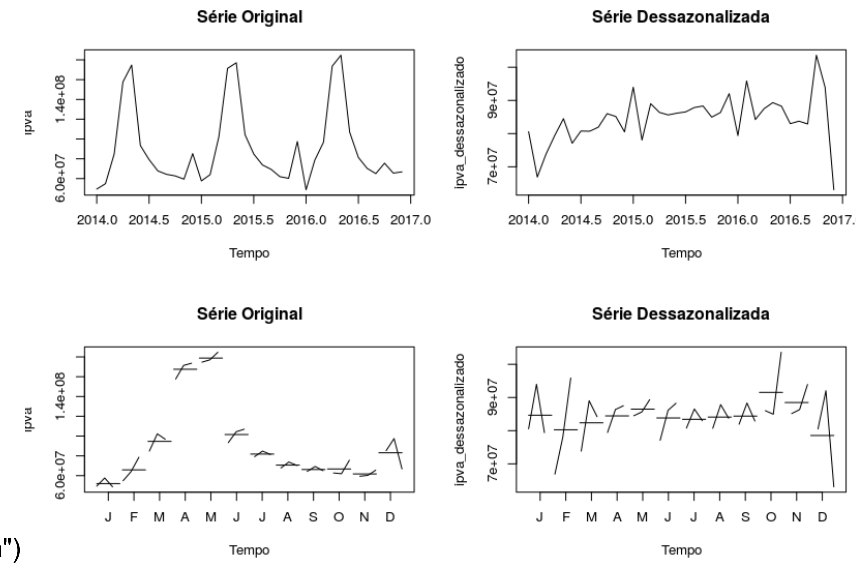
A série foi decomposta em seus elementos básicos, mas ainda não foi dessazonalizada. Para isso, devemos subtrair os fatores sazonais.

> ipva_dessazonalizado <- ipva_decomposto$x – ipva_decomposto$seasonal

Como a série está em logaritmo natural, é necessário retornar os valores aos níveis originais, aplicando um antilog:

> ipva_dessazonalizado ← exp(ipva_dessazonalizado)

Veja os gráficos à direita. A série foi dessazonalizada.




```
> par(mfrow=c(2,2))
> plot.ts(ipva, xlab="Tempo", main="Série Original")
> plot.ts(ipva_dessazonalizado, xlab="Tempo", main="Série Dessazonalizada")
> monthplot(ipva, xlab="Tempo", main="Série Original")
> monthplot(ipva_dessazonalizado, xlab="Tempo", main="Série Dessazonalizada")
```

TRATAMENTO DE DADOS

Dessazonalização

Exercícios

1) Abra no RStudio a série ar_condicionado (fonte: IPEADATA). Esta série apresenta a quantidade de aparelhos de ar condicionado vendidos no Brasil entre janeiro de 1994 e dezembro de 1999. Analise-a graficamente e comente o que há de particular nela.

2) Decomponha a série em seus elementos constitutivos: tendência, sazonal e aleatório.

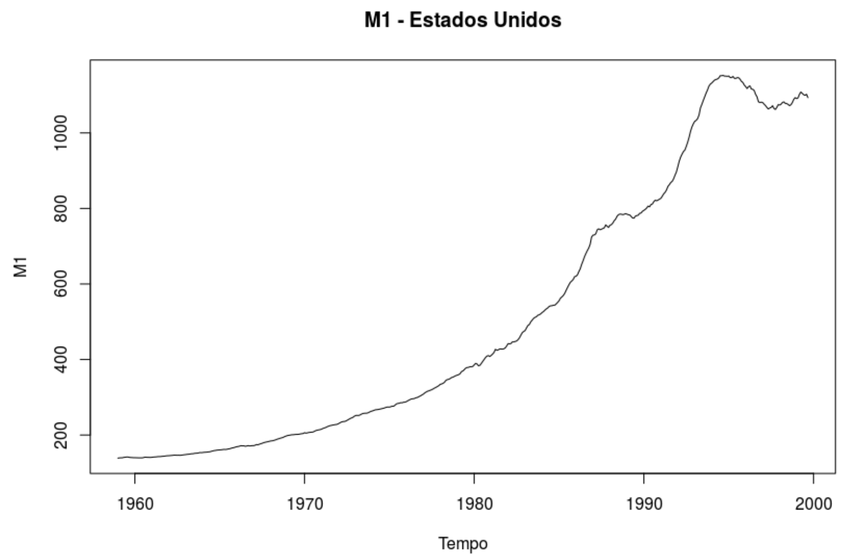
3) Caso necessário, dessazonalize a série pelo método das médias móveis e plote os gráficos da série original e da série dessazonalizada.

TIPOS DE DADOS

Séries temporais (time series)

Conjunto de observações de valores que uma ou mais variáveis assumem em diferentes momentos do tempo. Uma hipótese fundamental para regressão de duas ou mais séries temporais é que ambas sejam estacionárias ou cointegradas.

```
> plot.ts(tabela_1.4, xlab="Tempo", ylab="M1", main="M1 - Estados Unidos")
```



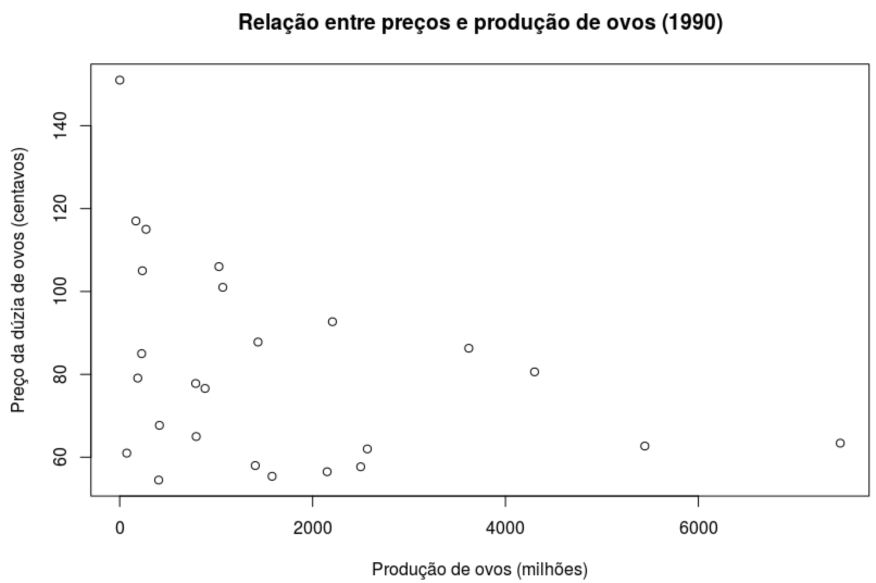
esesp

TIPOS DE DADOS

Corte (cross-section)

Dados de uma ou mais variáveis coletados <u>no mesmo ponto</u> do tempo.

```
> plot(tabela_1.1$Y1, tabela_1.1$X1, xlab="Produção de ovos (milhões)", ylab="Preço da dúzia de ovos (centavos)", main="Relação entre preços e produção de ovos (1990)")
```



TIPOS DE DADOS

Combinados (pooled)

Mistura de elementos de séries temporais e dados de corte.

> tabela_1.2

	Canada	France	Germany	Italy	Japan	UK	US
1973	40.8	34.6	62.8	20.6	47.9	27.9	44.4
1974	45.2	39.3	67.1	24.6	59.0	32.3	49.3
1975	50.1	43.9	71.1	28.8	65.9	40.2	53.8
1976	53.9	48.1	74.2	33.6	72.2	46.8	56.9
1977	58.1	52.7	76.9	40.1	78.1	54.2	60.6
1978	63.3	57.5	79.0	45.1	81.4	58.7	65.2
1979	69.2	63.6	82.2	52.1	84.4	66.6	72.6
1980	76.1	72.3	86.7	63.2	90.9	78.5	82.4
1981	85.6	81.9	92.2	75.4	95.3	87.9	90.9
1982	94.9	91.7	97.1	87.7	98.1	95.4	96.5
1983	100.4	100.4	100.3	100.8	99.8	99.8	99.6
1984	104.7	108.1	102.7	111.5	102.1	104.8	103.9
1985	109.0	114.4	104.8	121.1	104.1	111.1	107.6
1986	113.5	117.3	104.7	128.5	104.8	114.9	109.6
1987	118.4	121.1	104.9	134.4	104.8	119.7	113.6
1988	123.2	124.4	106.3	141.1	105.6	125.6	118.3
1989	129.3	128.7	109.2	150.4	108.1	135.3	124.0
1990	135.5	133.0	112.2	159.6	111.4	148.2	130.7
1991	143.1	137.2	116.3	169.8	115.0	156.9	136.2
1992	145.3	140.5	122.1	178.8	116.9	162.7	140.3
1993	147.9	143.5	127.6	186.4	118.4	165.3	144.5
1994	148.2	145.8	131.1	193.7	119.3	169.4	148.2
1995	151.4	148.4	133.5	204.1	119.1	175.1	152.4
1996	153.8	151.4	135.5	212.0	119.3	179.4	156.9
1997	156.3	153.2	137.8	215.7	121.3	185.0	160.5

> tabela_1.2[,4] (é uma série temporal - Itália)
> tabela_1.2[4,] (é um dado de corte - 1976)

TIPOS DE DADOS

Exercícios

Pesquise no site do IBGE séries dos tipos série temporal, corte e combinados.

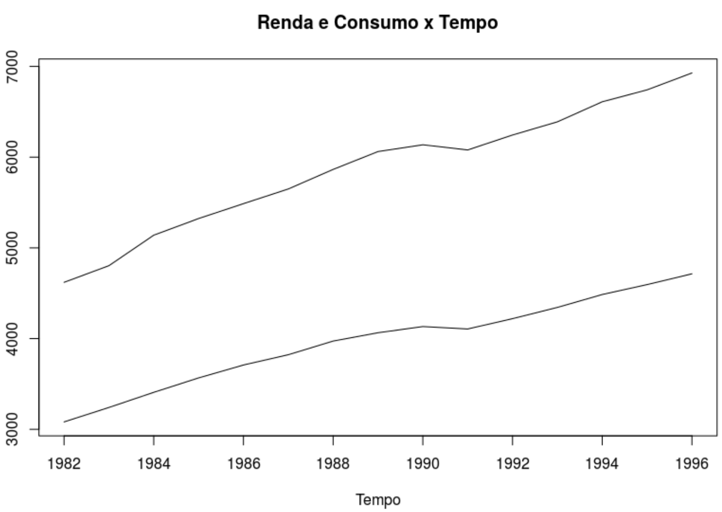
REGRESSÃO SIMPLES

O que é regressão?

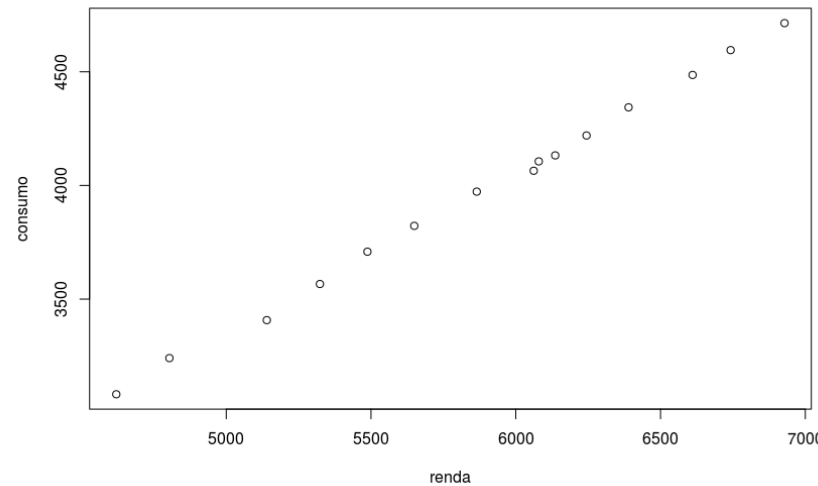
Qual a relação entre as variáveis de consumo(Y) e renda(X) ?

```
> tabela_I.1
Time Series:
Start = 1982
End = 1996
Frequency = 1
          Y      X
1982 3081.5 4620.3
1983 3240.6 4803.7
1984 3407.6 5140.1
1985 3566.5 5323.5
1986 3708.7 5487.7
1987 3822.3 5649.5
1988 3972.7 5865.2
1989 4064.6 6062.0
1990 4132.2 6136.3
1991 4105.8 6079.4
1992 4219.8 6244.4
1993 4343.6 6389.6
1994 4486.0 6610.7
1995 4595.3 6742.1
1996 4714.1 6928.4
```

```
> ts.plot(tabela_I.1, xlab="Tempo", main="Renda e
Consumo x Tempo")
```




```
> consumo <- as.vector(tabela_I.1[,1])
> renda <- as.vector(tabela_I.1[,2])
> plot(renda, consumo)
```




```
Agora, digite:
> abline(lm(consumo ~ renda))

O que acontece?
```

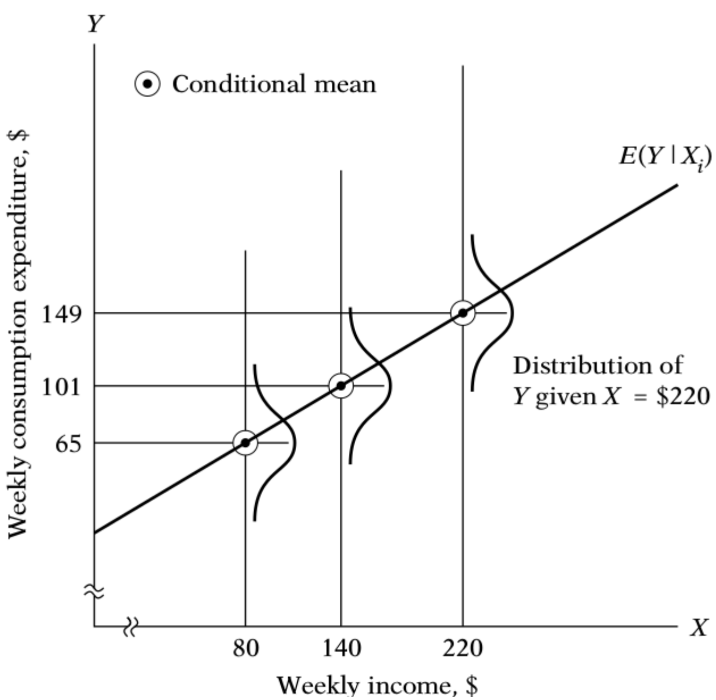
REGRESSÃO SIMPLES

O que é regressão?

Qual a relação entre as variáveis de consumo(Y) e renda(X) ?

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$



Fonte: Gujarati, Basic Econometrics.

Onde:

Beta 1 é o intercepto.
Beta 2 é o coeficiente angular.
u é o termo de perturbação estocástica, ou termo de erro.

E(Y|Xi) é o valor esperado (média) condicional da variável dependente (consumo) para valores fixos da variável independente (renda). É por onde passa a reta de regressão. Note que E(Y|Xi) é diferente do valor esperado incondicional de Y, E(Y), que é simplesmente a média populacional de Y (consumo), que não guarda relação nenhuma com X (renda).

REGRESSÃO SIMPLES

O que é regressão?

Já temos elementos para responder a pergunta acima. A análise de regressão consiste no estudo da dependência de uma variável, a variável dependente (ou endógena, ou explicada), em uma ou mais variáveis independentes (ou exógenas ou explicativas) com o propósito de estimar ou prever os valores médios das primeiras em termos de valores fixos (em amostragem repetida) ou conhecidos das últimas.

Regressão x causação: regressão não implica necessariamente em causação, que pode ser aferida por testes de causalidade (ex: Granger).

Regressão x correlação: são conceitos diferentes. Na correlação, ambas as variáveis são aleatórias (estocásticas) e tratadas igualmente. Na regressão, a variável dependente é aleatória, mas as variáveis independentes são fixas ou não estocásticas.

REGRESSÃO SIMPLES

O que é regressão?

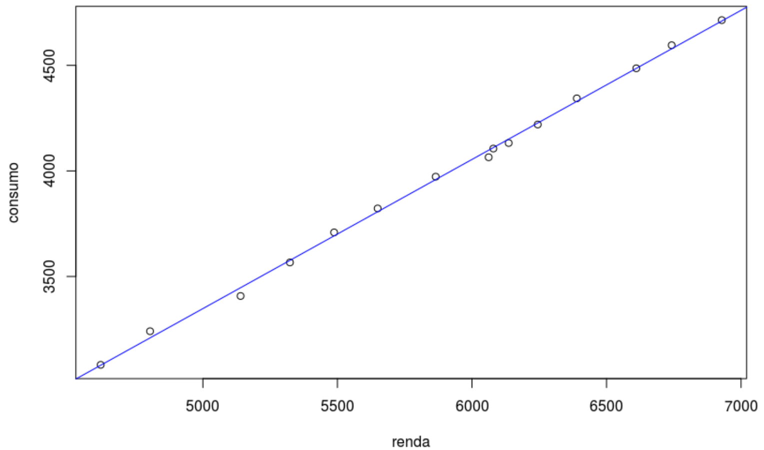
Retornemos à nossa Tabela I.1, que traz as variáveis de renda e consumo. Vamos estimar um modelo econométrico, com o consumo como variável dependente e a renda como variável independente.

```
> modelo1 <- lm(consumo ~ renda)
> modelo1
Call:
lm(formula = consumo ~ renda)
Coefficients:
(Intercept)        renda
  -184.0780       0.7064
```

$$Y_i = -184,0780 + 0,7064X_i + \hat{u}_i$$

Repare no ûi acima. Por que o chapéu?

```
> plot(renda,consumo)
> abline(modelo1, col="blue")
```

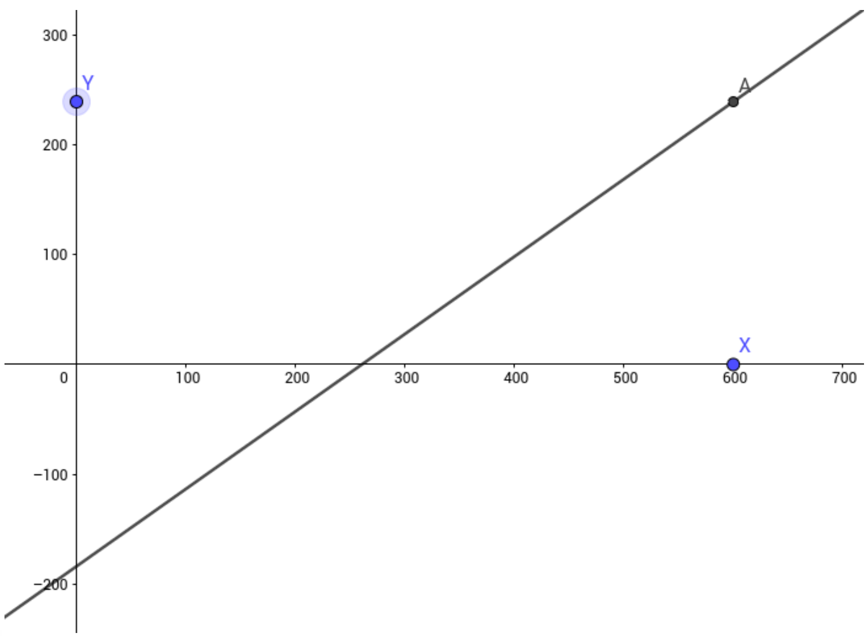


Nosso beta1 é igual a -184,0780, nosso beta2 é igual a 0,7064. Ok, mas o que isto significa? E como essas estimativas foram obtidas?

REGRESSÃO SIMPLES

O que é regressão?

O coeficiente angular, beta2, significa que, em média, um aumento de R$ 1,00 na renda implica um aumento de R$ 0,71 no consumo. Isto é coerente com a teoria O intercepto, beta1, nem sempre possui significado econômico ou prático. No caso, ele significaria o consumo quando a renda é zero.



Vamos imaginar apenas a reta de regressão à esquerda. Temos a equação:

$$Y_i = -184,0780 + 0,7064X_i + \hat{u}_i$$

Isto significa que:

$$E(Y|X_i) = -184,0780 + 0,7064X_i$$

Qual seria, pois, o valor de Y quando X for igual a, digamos, 600? Basta substituir na equação. E(Y|600) = -184,0780+0,7064(600), o que resulta em Y = 239,76, ou seja, quando a renda for R$ 600,00, o consumo será, em média, de R$ 239,76.

REGRESSÃO SIMPLES

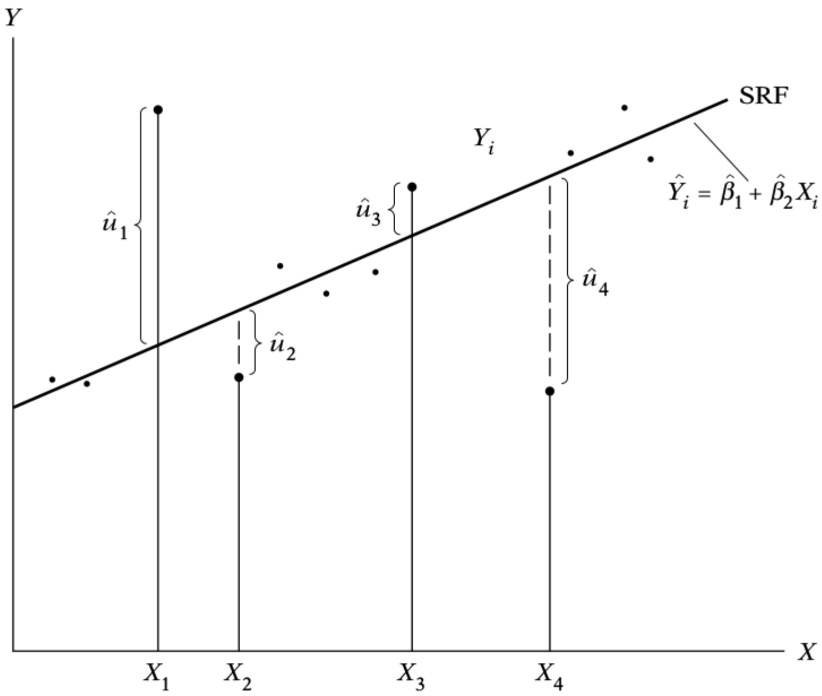
Estimativa dos parâmetros

Existem várias formas de se estimar os parâmetros beta1 e beta2. Duas metodologias são bem conhecidas: 1) o método dos mínimos quadrados ordinários (MQO); 2) a estimação por máxima verissimilhança (maximum likelihood). Veremos apenas a primeira delas, a estimação por MQO.

REGRESSÃO SIMPLES

Estimativa por MQO

Repare no gráfico abaixo:



Trata-se de uma reta de regressão (Sample Regression Function), onde ûi = Yi-Ŷi. Cada ûi é um resíduo da regressão, ou seja, o que sobra após retirarmos de cada Yi seus valores esperados condicionados a cada Xi. Os resíduos costumam representar todos os demais fatores que afetam a regressão além de Xi. O método dos MQO consiste em estimar os betas minimizando os quadrados dos resíduos.

REGRESSÃO SIMPLES

Estimativa por MQO

Estimando beta1 e beta2, sabendo que: $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$

$$\frac{\partial(\sum_{i=1}^{n} \hat{u}_i^2)}{\partial\hat{\beta}_2} = 0 \qquad \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\frac{\partial(\sum_{i=1}^{n} \hat{u}_i^2)}{\partial\hat{\beta}_1} = 0 \qquad \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$$

Igualando as derivadas parciais (acima e à esquerda) igual a zero, ou seja, minimizando os quadrados dos resíduos, obtemos os estimadores (acima e à direita) de beta2 e beta1.

REGRESSÃO SIMPLES

Soma dos quadrados

Conceitos importantes:

TSS: soma total dos quadrados.

ESS: soma dos quadrados da parte explicada da regressão.

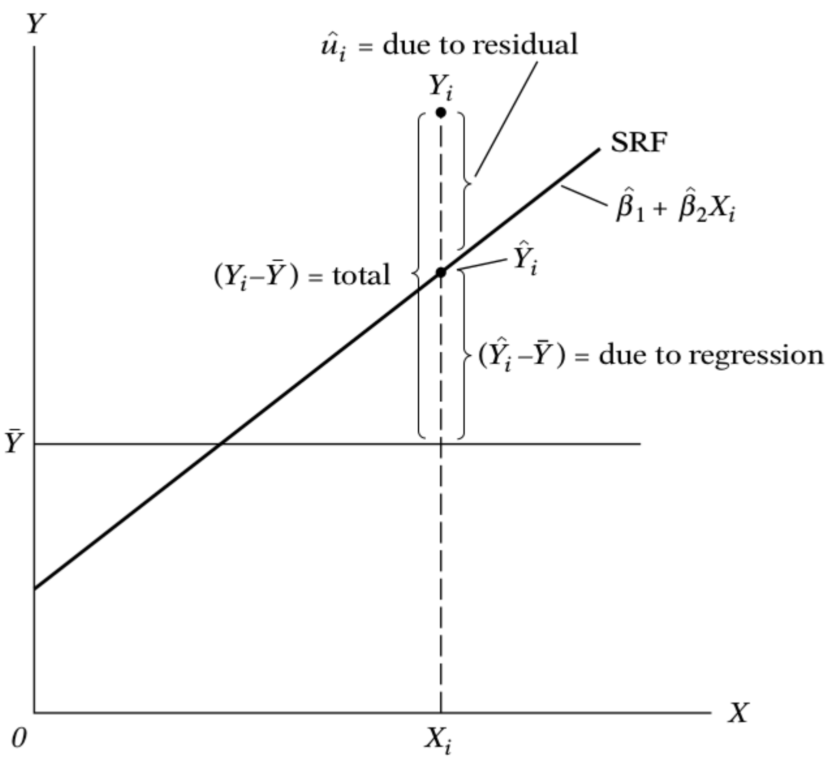
RSS: soma dos quadrados dos resíduos.

TSS = ESS + RSS, ou seja a soma total dos quadrados pode ser dividida em duas parcelas: a primeira parcela, ESS, diz respeito à linha de regressão, ao passo que a segunda parcela, RSS, aos fatores que não podem ser explicados pela linha de regressão (resíduos).

REGRESSÃO SIMPLES

Soma dos quadrados

Conceitos importantes:



$$\mathrm{TSS} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^{\,2} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$\mathrm{RSS} = \sum_{i=1}^{n} (\hat{u}_i)^2$$

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{\mathrm{TSS}} = \underbrace{\hat{\beta}_2^{\,2} \sum_{i=1}^{n} (X_i - \bar{X})^2}_{\mathrm{ESS}} + \underbrace{\sum_{i=1}^{n} (\hat{u}_i)^2}_{\mathrm{RSS}}$$

REGRESSÃO SIMPLES

Coeficiente de determinação

Também conhecido como r^2. É o quadrado do coeficiente de correlação. O coeficiente de determinação mede a proporção da porcentagem da variação total de Y explicada pelo modelo de regressão. Sua fórmula é dada por: $r^2 = ESS/TSS$. O coeficiente de determinação varia entre 0 e 1, sendo que quanto mais próximo de 1, melhor o ajuste. Em outras palavras, o r^2 nos diz o quanto nosso X, por meio do beta2, explica o Y. Obviamente, se beta2 for igual a zero, r^2 também será zero. Analogamente, se r for zero, r^2 também será zero, pois as variáveis não estão correlacionadas.

REGRESSÃO SIMPLES

Hipóteses do MCRL

O Modelo Clássico de Regressão Linear (MCRL) possui as seguintes hipóteses:

H1: linearidade dos parâmetros
H2: X não é estocástico
H3: $E(U_i|X_i) = 0$
H4: homocedasticidade
H5: não há autocorrelação dos resíduos
H6: $E(U_iX_i) = 0$
H7: número de observações > número de parâmetros
H8: $var(X) \neq 0$
H9: o modelo deve estar especificado corretamente
H10: não há multicolinearidade perfeita

REGRESSÃO SIMPLES

Hipóteses do MCRL

Teorema de Gauss-Markov: respeitadas as hipóteses do MCRL, os estimadores obtidos por MQO serão os melhores (mais eficientes) estimadores lineares não viesados.

Lineares: os estimadores serão funções lineares de uma função estocástica, como Y.

Não viesados: $E(\hat{\beta}_2) = \beta_2$ ou seja, o valor esperado da estimativa de beta2 será igual ao verdadeiro beta2.

Eficientes: a variância do estimador é a menor possível.

REGRESSÃO SIMPLES

Exercícios

Abra a tabela 3.8 no R Studio. Ela apresenta duas séries, PIB nominal (NGDP) e PIB real (RGDP) dos EUA.

a) plote as duas séries contra o tempo;

b) crie uma série chamada “tempo” que corresponda aos anos da tabela. 1959 será 1, 1960 será 2, e assim por diante. Ajuste um modelo em que o PIB nominal é variável dependente e o tempo é a independente. Estime os parâmetros por MQO utilizando e sem utilizar o comando lm.

c) como você interpreta os coeficientes beta1 e beta2?

d) qual o PIB em 2017?

A HIPÓTESE DA NORMALIDADE

Propriedades

Assumimos que os termos de perturbação estocástica seguem a distribuição normal (gaussiana), onde $u_i \sim \mathrm{NID}(0, \sigma^2)$, o que significa que ui é normal e identicamente distribuído. Isto implica que:

$$E(u_i) = 0$$

$$E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma^2$$

$$E\{[(u_i - E(u_\mathrm{i})][u_j - E(u_j)]\} = E(u_i\, u_j) = 0 \quad i \neq j$$

1) O valor esperado de ui é zero (os erros acima e abaixo da reta de regressão se anulam reciprocamente); 2) A variância de ui é constante (homocedasticidade); 3) Não existe correlação serial entre os erros.

A HIPÓTESE DA NORMALIDADE

Propriedades

De acordo com as propriedades dos estimadores de MQO, os betas são não viesados, possuem variância mínima (eficientes), são consistentes, isto é, assim que o tamanho da amostra aumenta, convergem para seus valores populacionais, e os betas são funções lineares dos termos de erro, ou seja, são também normalmente distribuídos. Em resumo:

$$E(\hat{\beta}_1) = \beta_1$$

$$\sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2$$

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$$

$$E(\hat{\beta}_2) = \beta_2$$

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum x_i^2}$$

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$$

$$Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}}$$

Onde: $\sum x_i^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$

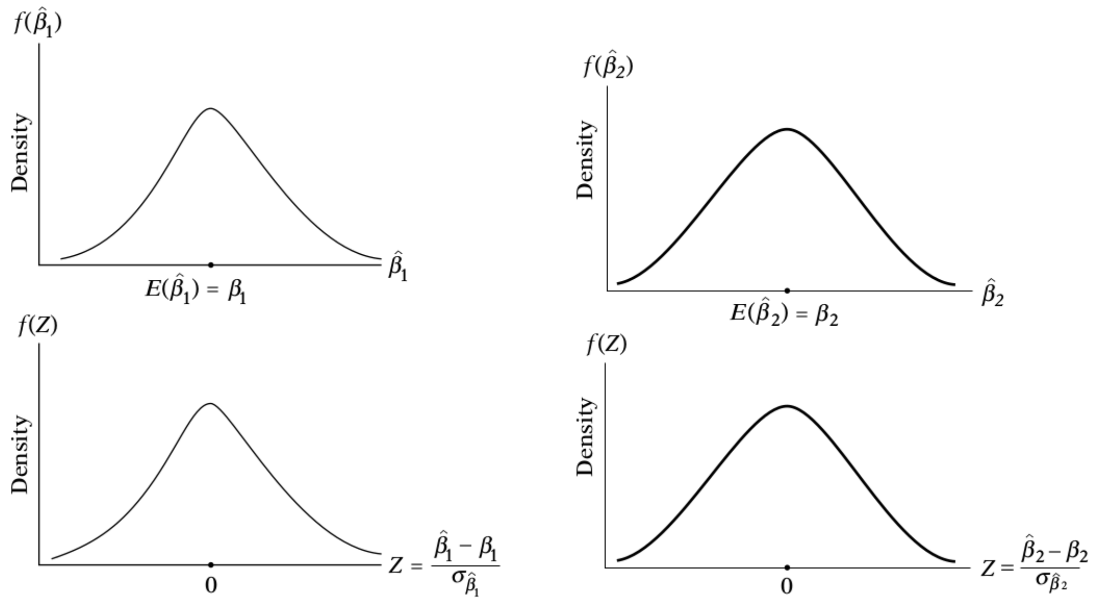
E Z é uma variável com distribuição normal padronizada com média zero e variância um.

$$Z \sim \mathrm{N}(0,1)$$

A HIPÓTESE DA NORMALIDADE

Propriedades

Os gráficos abaixo sintetizam as propriedades dos estimadores sob a distribuição normal:



Além disso:

$$E(Y_i) = \beta_1 + \beta_2 X_i$$

$$\text{var}(Y_i) = \sigma^2$$

INFERÊNCIA ESTATÍSTICA

Intervalo de confiança

Considerando que os estimadores dos betas são não-viesados, é possível montar um intervalo no qual o beta estimado, somado e diminuído a um mesmo número arbitrário δ, contenha o verdadeiro beta a uma dada probabilidade. Formalmente:

$$\Pr\left(\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta\right) = 1 - \alpha$$

O intervalo em que o beta está contido se chama intervalo de confiança. O beta pode ou não estar contido no intervalo. Se estiver, a probabilidade de que possamos construir um intervalo desse tipo, que contenha o verdadeiro beta, é 1-α, em que α é o nível de significância do intervalo. Em outras palavras, suponha que α seja 5%. A probabilidade de que o verdadeiro beta esteja num intervalo de confiança de 95% será, portanto, zero ou um.

INFERÊNCIA ESTATÍSTICA

Intervalo de confiança

Estabelecemos anteriormente a hipótese da normalidade sobre os erros. Porém, ela só pode ser aplicada quando a variância populacional dos erros, σ^2 , é conhecida. Como isto raramente acontece, e o que geralmente temos é a estimativa dessa variância, $\hat{\sigma}^2$, devemos utilizar a distribuição t de Student em vez da distribuição normal. Nesse sentido, o intervalo de confiança fica estabelecido como:

$$\Pr\left[\hat{\beta}_2 - t_{(n-2),\alpha/2}\,\mathrm{se}\,(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{(n-2)\alpha/2}\,\mathrm{se}\,(\hat{\beta}_2)\right] = 1 - \alpha$$

Onde o (n-2) significa o grau de liberdade da variável t, e **se** significa o erro padrão, σ.

INFERÊNCIA ESTATÍSTICA

Intervalo de confiança

Vamos estudar alguns exemplos práticos. Lembre-se do modelo econométrico de consumo e renda que ajustamos anteriormente, referente à Tabela I.1 do Rstudio. Digite:

> summary(modelo1)

```
Call:
lm(formula = consumo ~ renda)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.330  -8.601   1.761  14.769  31.306 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.841e+02  4.626e+01  -3.979  0.00157 ** 
renda        7.064e-01  7.827e-03  90.247  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.29 on 13 degrees of freedom
Multiple R-squared:  0.9984,	Adjusted R-squared:  0.9983 
F-statistic:  8145 on 1 and 13 DF,  p-value: < 2.2e-16
```

As variáveis da Tabela I.1 têm 15 observações cada (n=15). Logo, os graus de liberdade são 13 (n-2).

Repare o erro-padrão do coeficiente de renda: 0,007827. Ora, considerando um nível de significância de 5%, para 13 gl, t(0,025) é igual a 2,160. Logo, o intervalo de confiança para beta2 é 0,7064 +– 2,160(0,007827):

$$\Pr(0,6894937 \leq \beta_2 \leq 0,7233063) = 5\%$$

INFERÊNCIA ESTATÍSTICA

Intervalo de confiança

Alguns conceitos importantes:

Erro Tipo I: probabilidade de se rejeitar uma hipótese verdadeira. Está associado ao nível de significância (α). Quanto menor o α, menor a probabilidade de se cometer um Erro Tipo I.

Erro Tipo II: probabilidade de se aceitar uma hipótese falsa. Quanto maior o α, menor a probabilidade de se cometer um Erro Tipo II.

Resta evidente que há um *trade-off* entre os dois tipos de erro. À medida que tentamos evitar o Erro Tipo I, aumentamos a chance de incorrermos no Erro Tipo II, e vice-versa.

Poder do teste: é a probabilidade de não se cometer um Erro Tipo II. Em outras palavras, é a habilidade de rejeitar uma hipótese nula falsa.

INFERÊNCIA ESTATÍSTICA

Intervalo de confiança

Exercício:

Estime o modelo da Tabela 5.5 no Rstudio e construa um intervalo de confiança para beta1 e beta2 com um nível de significância de 5%. Interprete os resultados.

INFERÊNCIA ESTATÍSTICA

Teste de significância

Suponha que na nossa regressão de renda e consumo alguém queira testar a hipótese de que o verdadeiro beta2 é 0,7. Chamemos essa hipótese de h0, ou hipótese nula. Em contrapartida, deseja-se verificar uma hipótese alternativa de que o verdadeiro beta2 é diferente de 0,7. Formalmente:

h0: $\beta_2 = 0,7$ h1: $\beta_2 \neq 0,7$

Vamos relembrar o intervalo de confiança construído anteriormente:

$$\Pr(0,6894937 \leq \beta_2 \leq 0,7233063) = 5\%$$

O intervalo de confiança nos revela que beta2 = 0,7 é um dos valores possíveis do verdadeiro beta a um nível de significância de 5%. Mas existe uma outra abordagem para averiguar a validade de h0, que é a do teste de significância.

INFERÊNCIA ESTATÍSTICA

Teste de significância

O teste t de significância para o parâmetro beta2 pode ser expresso pela fórmula abaixo:

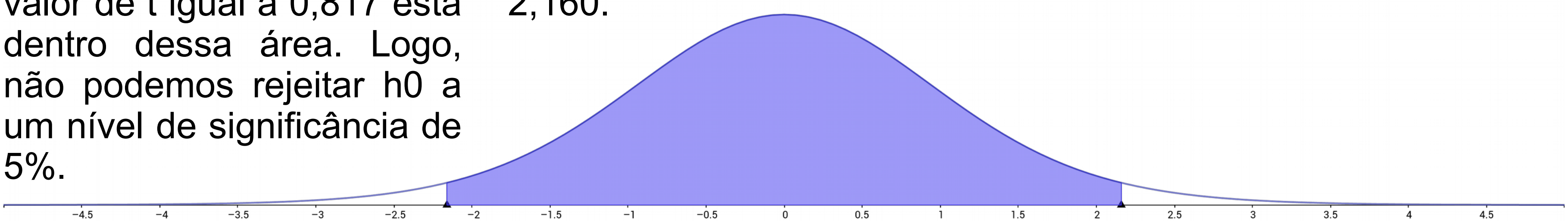
$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{se}\,(\hat{\beta}_2)}$$

Considerando que postulamos que o verdadeiro beta2 seja igual a 0,7, teríamos a seguinte medida de teste:

$$t = \frac{0,7064 - 0,7}{0,007827}$$

O que resulta em t = 0,817. Recordemos que para um nível de significância de 5%, com 13 gl, t(0,025) é igual a 2,160 (por que usar 0,025 em vez de 0,05?). Nosso t calculado, 0,817, é menor do que o t teórico, 2,160.

A área rachurada do gráfico abaixo está entre -2,160 e 2,160. Esta área representa 95% da distribuição. O valor de t igual a 0,817 está dentro dessa área. Logo, não podemos rejeitar h0 a um nível de significância de 5%.



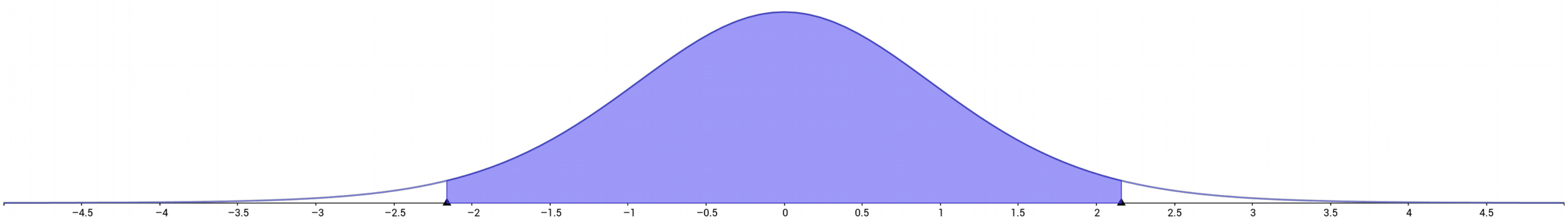
INFERÊNCIA ESTATÍSTICA

Teste de significância

Podemos dar um passo adiante e testar a hipótese nula de beta2 igual a zero. No fundo, queremos testar se nosso beta2 é significativo, ou seja, se a renda(X) exerce alguma influência sobre o consumo (Y). Nossa estatística t nesse caso seria:

$$t = \frac{0,7064 - 0}{0,007827} = 90,25$$

Ora, 90,25 é muito maior do que 2,160. Em termos gráficos, t=90,25 está muito além da região crítica da direita; é um valor tão alto que nem o gráfico o consegue representar. Logo, a hipótese nula de beta2=0 deve ser rejeitada.



INFERÊNCIA ESTATÍSTICA

Teste de significância

O valor de t=90,25 não deveria ser tão estranho para nós, afinal ele já apareceu antes. Ele é reportado pelo R como o "t-value" da variável renda.

```
Call:
lm(formula = consumo ~ renda)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.330  -8.601   1.761  14.769  31.306 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.841e+02  4.626e+01  -3.979  0.00157 ** 
renda        7.064e-01  7.827e-03  90.247  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.29 on 13 degrees of freedom
Multiple R-squared:  0.9984,	Adjusted R-squared:  0.9983 
F-statistic:  8145 on 1 and 13 DF,  p-value: < 2.2e-16
```

Perceba que existe uma coluna ao lado de "t-value" chamada "Pr(>|t|)". Esta coluna expressa o p-valor, que é a probabilidade exata da estatística t quando fazemos beta igual a zero em h0. No caso de beta2, a renda, o p-valor é menor do que $2x10^{-16}$, um número extremamente pequeno. Esta é a probabilidade de cometermos um Erro Tipo I, ou seja, a probabilidade de se rejeitar uma hipótese nula verdadeira é menor que $2x10^{-16}$.

INFERÊNCIA ESTATÍSTICA

ANOVA

No contexto da regressão simples (não é válido para regressão múltipla!), a estatística de teste F é simplesmente o quadrado da estatística de teste t. Note que $90{,}247^2 = 8145$. No entanto, a estatística F representa muito mais do que isso. Lembre-se da seguinte relação: TSS = ESS+RSS. Quais os graus de liberdade associados a cada termo da equação para regressões simples? Em TSS perdemos 1 gl em função da média, logo, temos n-1. Em RSS, n-2, devido à variância do termo de erro. ESS tem somente 1 gl , pois tratamos de somente um estimador para Xi, o beta2. Assim sendo, F é definido como:

$$F = \frac{\text{MSS of ESS}}{\text{MSS of RSS}} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum \hat{u}_i^2/(n-2)} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}$$

INFERÊNCIA ESTATÍSTICA

ANOVA

Voltemos à regressão do consumo e da renda. RSS pode ser obtida pela variância dos resíduos multiplicada por n-1. Já ESS pode ser obtida pelo beta2 ao quadrado vezes a variância da renda multiplicada por n-1. Logo,

```
> var(modelo1$residuals)*(length(modelo1$residuals)-1)
[1] 5349.39

> (modelo1$coefficients[2]^2)*var(tabela_I.1[,2])*(length(tabela_I.1[,2])-1)
  renda
3351407
```

Temos que RSS=5.349,39 e ESS=3.351.407. O F é igual a (3.351.407 / 5.349,39) * 13, onde o 13 é igual a n-2. Logo, F = 8144,53. O comando anova (analysis of variance) do Rstudio resume os resultados:

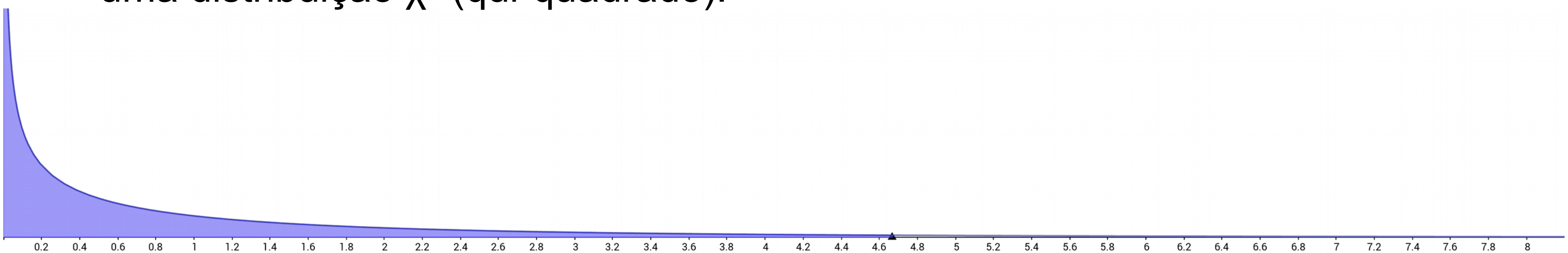
```
Analysis of Variance Table

Response: consumo
          Df  Sum Sq Mean Sq F value    Pr(>F)
renda      1 3351407 3351407  8144.5 < 2.2e-16 ***
Residuals 13    5349     411
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


esesp

INFERÊNCIA ESTATÍSTICA

ANOVA

O valor de F pode ser buscado na tabela estatística do teste F, que segue uma distribuição χ² (qui-quadrado).



Observe no gráfico acima (e olhe na Tabela F) que, para α = 5%, 95% do intervalo de confiança está em valores menores do que F = 4,67. Como F = 8144,53, devemos rejeitar a hipótese nula de que beta2 seja igual a zero. Como veremos adiante, o teste F é útil para regressões múltiplas porque testa a hipótese conjunta de que todos os betas que contenham variáveis explicativas sejam significativamente iguais a zero.

INFERÊNCIA ESTATÍSTICA

Exercícios

Carregue a tabela 5.9 no Rstudio. Ela apresenta o chamado índice do Big Mac, uma medida para aferir Paridade do Poder de Compra (PPP).

1) Ajuste um modelo econométrico de regressão linear simples onde X=PPP e Y=taxa de câmbio.

2) Interprete os resultados.

3) Se a PPP é uma metodologia válida, que resultados você esperaria para os betas a priori?

4) Faça um teste de hipótese com os resultados esperados em 3).

5) Analise os resíduos da regressão, inclusive quanto à normalidade (testes de Shapiro-Wilk e Jarque-Bera).

LOGARITMOS

Elasticidade

Os logaritmos são ferramentas muito úteis na econometria. Uma aplicação importante diz respeito à elasticidade. Ela é expressa pela fórmula $\frac{\partial Y}{\partial X}\frac{X}{Y}$.

Aplique o logaritmo (natural) às séries de consumo e renda da Tabela I.1 e reestime o modelo econométrico. À esquerda, o original. À direita, o log-log. O que mudou?

```
> summary(modelo1)

Call:
lm(formula = consumo ~ renda)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.330  -8.601   1.761  14.769  31.306 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.841e+02  4.626e+01  -3.979  0.00157 ** 
renda        7.064e-01  7.827e-03  90.247  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.29 on 13 degrees of freedom
Multiple R-squared:  0.9984,    Adjusted R-squared:  0.9983 
F-statistic:  8145 on 1 and 13 DF,  p-value: < 2.2e-16
```

```
> summary(modelo2)

Call:
lm(formula = log(consumo) ~ log(renda))

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0115992 -0.0020808  0.0000713  0.0036528  0.0089752 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.79563    0.10387   -7.66 3.59e-06 ***
log(renda)   1.04636    0.01198   87.36  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.005438 on 13 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9982 
F-statistic:  7632 on 1 and 13 DF,  p-value: < 2.2e-16
```

LOGARITMOS

Elasticidade

Em relação a beta2, o modelo original pode ser interpretado da seguinte forma: um aumento de R$ 1,00 na renda leva, em média, a um aumento de R$ 0,71 no consumo. No modelo log-log, um aumento de 1% na renda leva a um aumento, em média, de 1,05% no consumo. Isto se chama elasticidade e designa o percentual de variação da variável dependente em decorrência da variação de 1% na independente.

```
> summary(modelo1)

Call:
lm(formula = consumo ~ renda)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.330  -8.601   1.761  14.769  31.306 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.841e+02  4.626e+01  -3.979  0.00157 ** 
renda        7.064e-01  7.827e-03  90.247  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.29 on 13 degrees of freedom
Multiple R-squared:  0.9984,	Adjusted R-squared:  0.9983 
F-statistic:  8145 on 1 and 13 DF,  p-value: < 2.2e-16
```

```
> summary(modelo2)

Call:
lm(formula = log(consumo) ~ log(renda))

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0115992 -0.0020808  0.0000713  0.0036528  0.0089752 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.79563    0.10387   -7.66 3.59e-06 ***
log(renda)   1.04636    0.01198   87.36  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.005438 on 13 degrees of freedom
Multiple R-squared:  0.9983,	Adjusted R-squared:  0.9982 
F-statistic:  7632 on 1 and 13 DF,  p-value: < 2.2e-16
```

LOGARITMOS

Linearização

Seja $F = AL^{\alpha}K^{1-\alpha}$ uma função Cobb-Douglas com retornos constantes de escala. "F" representa a produção da economia, "A" representa a tecnologia, "L" representa o fator trabalho, "K" representa o fator capital. O parâmetro α representa as elasticidades do trabalho e do capital. A função Cobb-Douglas é uma função não linear. É difícil estimar um modelo para uma função desse tipo, mas o log nos permite especificar um modelo linear, mesmo que a função original não o seja. Aplicando o log na função Cobb-Douglas, temos: $\log F = \log A + \alpha \log L + (1-\alpha)\log K$. Substituindo os termos logF por Y, logA por beta1, logL por X1, logK por X2, α por beta2 e (1-α) por beta3, temos: $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2$

Conforme veremos adiante, esta última equação expressa uma regressão múltipla, porém os termos estão agora muito mais familiares para nós.

LOGARITMOS

Suavização da variância

Aplicar o logaritmo pode contribuir para suavizar a variância de uma série. Formalmente, o teste de Box e Cox retorna uma estatística lambda que pode exigir necessidade de logaritmização caso lambda seja igual a zero. Suavizar a variância é importante, uma vez que o erro-padrão influencia diretamente nas estatísticas de teste.

LOGARITMOS

Exercícios

1) aplique o teste de Box e Cox nos modelos de renda e consumo (original e log). Como você interpreta os resultados? (dica: utilizar comando boxcox).

2) Log-linearize o modelo abaixo. A que conclusão você chega?

$$Y_i = \frac{e^{\beta_1 + \beta_2 X_i}}{1 + e^{\beta_1 + \beta_2 X_i}}$$

3) Estime o modelo CAPM com os dados da Tabela 6.1 no Rstudio. O

modelo CAPM tem a forma: $(\mathrm{ER}_i - r_f) = \beta_i(\mathrm{ER}_m - r_f)$
Onde: ERi é a taxa de retorno esperada do ativo i (exemplo: PETR4).
ERm é a taxa de retorno esperada do mercado (exemplo: IBOVESPA).
rf é a taxa livre de risco (exemplo: SELIC).
O beta é uma medida de risco. Se beta>1, o ativo é de risco; se beta<1, o contrário.
Não confundir este beta com o beta da regressão.

REGRESSÃO MÚLTIPLA

Conceituação

A Tabela 6.4 no Rstudio apresenta as variáveis de mortalidade infantil (CM), PNB per capita (PGNP) e taxa de alfabetização feminina (FLR). Veja os resultados de duas regressões: uma de CM sobre FLR e outra de PGNP sobre FLR.

```
> CM <- tabela_6.4[,1]
> PGNP <- tabela_6.4[,3]
> FLR <- tabela_6.4[,2]
```

```
> summary(modelo_CM)

Call:
lm(formula = CM ~ FLR)

Residuals:
    Min      1Q  Median      3Q     Max 
-86.262 -25.453   0.357  22.591  98.337 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 263.8635    12.2250   21.58   <2e-16 ***
FLR          -2.3905     0.2133  -11.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.02 on 62 degrees of freedom
Multiple R-squared:  0.6696,	Adjusted R-squared:  0.6643 
F-statistic: 125.6 on 1 and 62 DF,  p-value: < 2.2e-16
```

```
> summary(modelo_PGNP)

Call:
lm(formula = PGNP ~ FLR)

Residuals:
    Min      1Q  Median      3Q     Max 
-2026.5  -948.5  -348.5   -70.9 18096.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   -39.30     734.95  -0.053   0.9575  
FLR            28.14      12.82   2.195   0.0319 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2647 on 62 degrees of freedom
Multiple R-squared:  0.07211,	Adjusted R-squared:  0.05714 
F-statistic: 4.818 on 1 and 62 DF,  p-value: 0.03191
```

REGRESSÃO MÚLTIPLA

Conceituação

Qual a finalidade dessas regressões simples? Queremos analisar a influência do PNB per capita sobre a mortalidade infantil mantendo constante a influência da taxa de alfabetização feminina (FLS) sobre ambas as variáveis. Eis as equações dos resíduos de cada regressão:

$$\hat{u}_{1i} = (\mathrm{CM}_i - 263.8635 + 2.3905\ \mathrm{FLR}_i) \qquad \hat{u}_{2i} = (\mathrm{PGNP}_i + 39.3033 - 28.1427\ \mathrm{FLR}_i)$$

Observamos nas equações acima que os resíduos representam, respectivamente, a influência de CM e PGNP líquida de FLR. Portanto, se quisermos a influência líquida de PGNP sobre CM, basta ajustar um modelo sobre os resíduos, com û1i como variável dependente e û2i como variável independente.

```
> summary(modelo_res)

Call:
lm(formula = resid(modelo_CM) ~ resid(modelo_PGNP))

Residuals:
    Min      1Q  Median      3Q     Max 
-84.267 -24.363   0.709  19.455  96.803 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)   
(Intercept)       -4.815e-16  5.176e+00   0.000  1.00000   
resid(modelo_PGNP) -5.647e-03  1.987e-03  -2.842  0.00607 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.41 on 62 degrees of freedom
Multiple R-squared:  0.1152,	Adjusted R-squared:  0.101 
F-statistic: 8.075 on 1 and 62 DF,  p-value: 0.006066
```

O beta2 dos resíduos û2i é -0,0056. Como exercício, faça procedimento análogo para obter a influência líquida de FLR sobre CM. O beta2 dos resíduos de FLR como variável independente deverá ser -2.231586.

esesp

REGRESSÃO MÚLTIPLA

Conceituação

Quanto mais variáveis quisermos analisar, maior o número de regressões simples que deveremos ajustar. Apenas para visualizar a influência líquida de PGNP sobre CM, tivemos que ajustar 2 modelos, observar os resíduos e ajustar um terceiro modelo. Mais 3 modelos seriam necessários para visualizarmos a influência líquida de FLR sobre CM, totalizando 6 modelos de regressão simples. A questão que se coloca é a seguinte: não poderíamos evitar essa trabalheira toda e ajustar apenas um modelo econométrico? Felizmente, a resposta para esta pergunta é positiva. Aí que entra a regressão múltipla. Para sintetizar o raciocínio, vamos ajustar um único modelo utilizando CM como variável dependente e PGNP e FLR como variáveis independentes:

```
> summary(modelo_multiplo)

Call:
lm(formula = CM ~ PGNP + FLR)

Residuals:
    Min      1Q  Median      3Q     Max 
-84.267 -24.363   0.709  19.455  96.803 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 263.641586  11.593179  22.741  < 2e-16 ***
PGNP         -0.005647   0.002003  -2.819  0.00649 ** 
FLR          -2.231586   0.209947 -10.629 1.64e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.75 on 61 degrees of freedom
Multiple R-squared:  0.7077,	Adjusted R-squared:  0.6981 
F-statistic: 73.83 on 2 and 61 DF,  p-value: < 2.2e-16
```

O que podemos dizer sobre os betas desta regressão?

REGRESSÃO MÚLTIPLA

Inclusão de variáveis

Até o momento, vimos como podem ser estabelecidas relações entre uma variável dependente e uma variável independente. Entretanto, é muito comum incluirmos outras variáveis independentes na equação de regressão. Por exemplo, poderíamos supor que a renda de determinada pessoa dependa de variáveis como grau de escolaridade, idade e região geográfica. Isto nos daria a seguinte equação de regressão:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

REGRESSÃO MÚLTIPLA

Inclusão de variáveis

Os pressupostos básicos de uma regressão múltipla são bastante similares aos de uma regressão simples. O fato de adicionarmos mais variáveis tende a aumentar nosso R^2 porque incluímos mais fatores na parte explicativa da equação, deixando menor influência para o termo de erro. Porém, a inclusão de variáveis diminui os graus de liberdade da equação, o que afeta o desempenho dos estimadores. Uma medida mais adequada do R^2, o R^2 ajustado, que considera os graus de liberdade da equação, pode expressar o grau de ajuste do modelo com maior fidedignidade. Vimos um exemplo prático disso na regressão múltipla da mortalidade infantil. O R^2 é 0,7077, ao passo que o R^2 ajustado é 0,6981. Portanto, a mera inclusão de variáveis não necessariamente melhora nossos modelos. A fórmula do R^2 ajustado é $R^2_{\text{adj}} = 1 - \frac{n-1}{n-(p+1)} \times (1-R^2)$ onde n é o número de observações e p é o número de variáveis da regressão.

REGRESSÃO MÚLTIPLA

Exercícios

1) Observe os dois modelos abaixo referentes à Tabela 6.4 do Rstudio. A regressão à direita incluiu a variável TFR, que representa a taxa total de fertilidade.

a) Como você interpreta o beta de TFR?
b) A que se deve a mudança dos coeficientes de PGNP e FLR ao adicionar mais uma variável? A diferença é significativa?
c) Qual dos modelos você escolheria? Que teste balizaria sua decisão?

```
> summary(modelo_multiplo)

Call:
lm(formula = CM ~ PGNP + FLR)

Residuals:
    Min      1Q  Median      3Q     Max 
-84.267 -24.363   0.709  19.455  96.803 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 263.641586  11.593179  22.741  < 2e-16 ***
PGNP         -0.005647   0.002003  -2.819  0.00649 ** 
FLR          -2.231586   0.209947 -10.629 1.64e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.75 on 61 degrees of freedom
Multiple R-squared:  0.7077,	Adjusted R-squared:  0.6981 
F-statistic: 73.83 on 2 and 61 DF,  p-value: < 2.2e-16
```

```
> summary(modelo_multiplo2)

Call:
lm(formula = CM ~ PGNP + FLR + TFR)

Residuals:
   Min     1Q Median     3Q    Max 
-98.17 -18.56   3.32  17.12  98.72 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 168.306690  32.891655   5.117 3.44e-06 ***
PGNP         -0.005511   0.001878  -2.934  0.00473 ** 
FLR          -1.768029   0.248017  -7.129 1.51e-09 ***
TFR          12.868636   4.190533   3.071  0.00320 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.13 on 60 degrees of freedom
Multiple R-squared:  0.7474,	Adjusted R-squared:  0.7347 
F-statistic: 59.17 on 3 and 60 DF,  p-value: < 2.2e-16
```

REGRESSÃO MÚLTIPLA

Exercícios

2) Visualize os dados da Tabela 1.5 do Rstudio. Ela apresenta dados de 21 firmas americanas relacionados ao número de visualizações (em milhões) de suas páginas web (variável impress) e gastos (em milhões de US$) com publicidade (variável adexp).

a) plote uma variável contra outra;
b) ajuste um modelo de regressão linear simples com impress como variável dependente e adexp como variável independente. Chame este modelo de modelo1 e interprete os resultados.
c) crie uma nova variável com o quadrado de adexp. Chame-a de adexp2. Ajuste um modelo de regressão múltipla com impress como variável dependente e adexp e adexp2 como variáveis independentes. Interprete os resultados.
d) qual o melhor modelo? Por quê?
e) há retornos decrescentes de gastos com publicidade? Qual seria o nível ótimo desses gastos?

REGRESSÃO MÚLTIPLA

Exercícios

3) Construa um modelo econométrico de regressão múltipla para explicar a arrecadação do ICMS do Estado do ES. Faça previsão de 1 ano e analise as seguintes medidas de erro de previsão: Erro Quadrático Médio (EQM) e Erro Percentual (EP).

VARIÁVEIS QUALITATIVAS

O que são variáveis dummy?

Variáveis dummy são variáveis que denotam algum estado. Geralmente são utilizadas para representar variáveis qualitativas. Por exemplo:

0 – Não, 1 – Sim
0 – Masculino, 1 – Feminino
0 – Região X, 1 – fora da Região X
0 – maior de idade, 1 – menor de idade

Os valores de 0 ou 1 dependerão da conveniência de como for estruturado o modelo.

Vamos a um exemplo prático!

VARIÁVEIS QUALITATIVAS

Modelos qualitativos

Carregue a Tabela 9.1 no Rstudio. A primeira coluna mostra o salário médio dos professores de escolas públicas em 51 Estados americanos. As duas últimas colunas são variáveis dummy:

D1 = 1 se o Estado é do Norte ou do Centro-Norte (CN)

= 0, caso contrário

D2 = 1 se o Estado é do Sul

= 0, caso contrário

Vamos ajustar o seguinte modelo: $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$

O que este modelo nos diz? O que esperar de cada dummy?

1° caso) Salário médio dos professores dos Estados do Norte ou CN:

$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2$

2° caso) Salário médio dos professores do Sul: $E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3$

3° caso) Salário médio dos professores do Oeste:

$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1$

VARIÁVEIS QUALITATIVAS

Interpretação de resultados

Como podemos interpretar os resultados do modelo qualitativo?

```
> summary(modelo_qualitativo)

Call:
lm(formula = salario ~ D1 + D2)

Residuals:
    Min      1Q  Median      3Q     Max 
-6329.1 -2592.1  -370.6  2143.4 15321.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    26159       1128  23.180   <2e-16 ***
D1             -1734       1436  -1.208   0.2330    
D2             -3265       1499  -2.178   0.0344 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4069 on 48 degrees of freedom
Multiple R-squared:  0.09008,	Adjusted R-squared:  0.05217 
F-statistic: 2.376 on 2 and 48 DF,  p-value: 0.1038
```

O intercepto (beta1) é significativo. Ele nos diz que o salário médio dos professores do Oeste é de $26.159. Os salários médios dos professores do Norte e CN são menores que os do Oeste em $1734. Os salários médios dos professores do Sul, por sua vez, são menores que os do Oeste em $3265.

Resumindo: salários médios dos
Professores do Oeste: $26.159
Professores do Norte e CN: $24.435
Professores do Sul: $22.894

Pergunta: as diferenças salariais são estatisticamente significantes? O que dizer do modelo como um todo?

VARIÁVEIS QUALITATIVAS

Exercícios

1) Refaça a dessazonalização do IPVA utilizando variáveis dummy.

VARIÁVEIS QUALITATIVAS

Exercícios

2) Considere os resultados da seguinte regressão:

$$\begin{aligned}
\hat{Y}_i = 1286 \quad & + 104.97X_{2i} - 0.026X_{3i} + 1.20X_{4i} + 0.69X_{5i} \\
t = \quad (4.67) \quad & \quad (3.70) \quad\quad (-3.80) \quad\quad (0.24) \quad\quad (0.08) \\
& -19.47X_{6i} + 266.06X_{7i} - 118.64X_{8i} - 110.61X_{9i} \\
& \quad (-0.40) \quad\quad (6.94) \quad\quad (-3.04) \quad\quad (-6.14) \\
& R^2 = 0.383 \qquad n = 1543
\end{aligned}$$

Onde: **Yi** = horas de trabalho anuais desejadas pela esposa, calculadas como a soma das horas de trabalho usuais por ano e das semanas procurando por trabalho.
X2 = renda média disp. da esposa. **X3** = renda média disp. do marido no ano passado. **X4** = idade da esposa (em anos). **X5** = escolaridade da esposa (em anos).
X6 (dummy) = 1 se a entrevistada concorda que mulheres trabalhem se elas desejarem; 0, caso contrário.
X7 (dummy) = 1 se o marido da entrevistada concorda com seu trabalho; 0, caso contrário. **X8** = número de crianças menores do que 6 anos de idade.
X9 = número de crianças entre 6 e 13 anos.

VARIÁVEIS QUALITATIVAS

Exercícios

Continuação do exercício 2)

a) os sinais das variáveis quantitativas fazem sentido econômico? Justifique sua resposta.
b) como você interpreta as variáveis dummy? São estatisticamente significantes?
c) as variáveis de idade e escolaridade são significativas? Por quê? Interprete.

FIM

Finalizamos nossos estudos por ora. Espero que tenham aproveitado. A econometria é um campo científico muito vasto, cada vez mais interdisciplinar, e este foi só o começo da jornada.

Caso queira entrar em contato, estarei disponível no e-mail:
mfsalomao@sefaz.es.gov.br

Obrigado e até a próxima.

Martinho de Freitas Salomão

SOLUÇÕES EDUCACIONAIS

Presenciais

A Distância

Customizadas

Lato e Stricto Sensu

FaceEsesp

esesp.es.gov.br

esesp