

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
1/16

Fúlvio Nedel
SPB/UFSC

Introdução


Objetivos da análise

As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

Introdução ao uso do



em Ciências da Saúde

3. Leitura, limpeza e manejo de dados
- a. Leitura: importação de dados e seleção de variáveis

Fúlvio Borges Nedel

Departamento de Saúde Pública – SPB

Centro de Ciências da Saúde – CCS

Universidade Federal de Santa Catarina – UFSC

Grups de Recerca d'Amèrica i Àfrica Llatines – GRAAL

<http://graal.uab.cat>

19 de dezembro de 2017

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
2/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise

As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

1 Introdução

- Objetivos da análise
- As variáveis de análise

2 Leitura do arquivo

3 Seleção das variáveis de interesse

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
3/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise

As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

- Em 2011 o Serviço de Saúde Comunitária do Grupo Hospitalar Conceição (SSC/GHC), em Porto Alegre, RS, iniciou um estudo sobre o processo saúde-doença-atenção de pessoas com Hipertensão (HAS) ou Diabetes Mellitus (DM) usuárias do Serviço
- No estudo, foi realizado um inquérito sobre uma amostra dos usuários, em que se perguntou o peso e altura do indivíduo.
- Usaremos um extrato dessa base de dados, que pode ser baixado nesse [link](#).¹ Baixe o arquivo de dados e salve-o em um diretório para este curso, no seu computador
- Faremos uma análise exploratória do estado nutricional.

¹ Usaremos 300 registros, selecionados por conveniência. Assim, os resultados aqui encontrados não se aplicam nem à amostra nem à população de onde ela foi coletada. Entretanto, nestes exercícios, nossa base de dados será tratada como uma amostra aleatória da população-alvo. O nº de registro do usuário foi modificado, por questões éticas.

Variável dependente

- 1 Descrever o Índice de Massa Corporal (IMC) e analisar fatores associados à sua média.
- 2 Descrever a frequência de categorias do estado nutricional e analisar fatores possivelmente associados à obesidade.

Variáveis independentes

- 1 sexo
- 2 idade
- 3 faixa etária
- 4 condição socioeconômica
- 5 participação em grupos de promoção da saúde

- O IMC é calculado como a razão entre o peso em quilos e o quadrado da altura em metros: $IMC = \frac{Kg}{m^2}$
- Sobrepeso é definido como $25 \leq IMC < 30Kg/m^2$
- Obesidade é definida como $IMC \geq 30Kg/m^2$

As perguntas da entrevista

peso

"u47. Qual o seu peso?"

altura

"u48. Qual a sua altura?"

Tabela: Variáveis independentes, nome e rótulo.

Nome	Rótulo
sexo	u8. Sexo:
dataentr	u5. Data da entrevista:
datanasc	u7. Qual é a sua data de nascimento?
abepcls	Classificação socioeconômica ABEP modificada
grupohas	u53. Desde <6 MESES ATRÁS> o(a) Sr.(a) participou de algum grupo de hipertensos no <UNIDADE DE SAÚDE DE REFERÊNCIA>?
grupodm	u63. Desde <6 MESES ATRÁS> o(a) Sr.(a) participou de algum grupo de diabéticos no <UNIDADE DE SAÚDE DE REFERÊNCIA>?

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
7/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise
As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

- 1 Introdução
 - Objetivos da análise
 - As variáveis de análise

- 2 Leitura do arquivo

- 3 Seleção das variáveis de interesse

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
8/16

Fúlvio Nedel
SPB/UFSC

Introdução
Objetivos da análise
As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

Abra o arquivo de sintaxe criado anteriormente. (exercício do módulo 2)

Crie uma nova linha com um comentário explicando os passos que se seguirão e execute a partir dali os comandos. Algo como:

```
# Leitura de dados  
# -----  
...
```


Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
9/16

Fúlvio Nedel
SPB/UFSC

Introdução
Objetivos da análise
As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

- O arquivo de dados foi gerado no `$P$$`, e está em formato `.sav`
- O pacote `Hmisc` tem funções para facilitar a leitura de arquivos em diferentes formatos, inclusive `SAV`. Ative o pacote (com `library(Hmisc)`) e leia o arquivo `usuariosCursoR.sav` com a função `spss.get`
- Indique as variáveis de data: `datevars = "..."`
- Lembre-se de destinar a ação a um objeto:
`nome do objeto <- spss.get(...)`

```
library(Hmisc)
cursoR <- spss.get( file = "usuariosCursoR.sav",
                    datevars = c("dataentr", "datanasc") )
```

Ignore os avisos. Eles poderiam ser evitados com o argumento `use.value.labels=FALSE`, mas aí... (?`spss.get` para ver o que aconteceria)

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
9/16

Fúlvio Nadel
SPB/UFSC

Introdução
Objetivos da análise
As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

- O arquivo de dados foi gerado no SPSS, e está em formato `.sav`
- O pacote `Hmisc` tem funções para facilitar a leitura de arquivos em diferentes formatos, inclusive SAV. Ative o pacote (com `library(Hmisc)`) e leia o arquivo `usuariosCursoR.sav` com a função `spss.get`
- Indique as variáveis de data: `datevars = "..."`
- Lembre-se de destinar a ação a um objeto:
`nome do objeto <- spss.get(...)`

```
library(Hmisc)
cursoR <- spss.get( file = "usuariosCursoR.sav",
                    datevars = c("dataentr", "datanasc") )
```

Ignore os avisos. Eles poderiam ser evitados com o argumento `use.value.labels=FALSE`, mas aí... (?`spss.get` para ver o que aconteceria)

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
10/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise

As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

Confirme que o objeto (**cursoR**) está presente no espaço de trabalho:

```
ls()
```

```
[1] "cursoR"
```

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
10/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise
As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

Confirme que o objeto (**cursoR**) está presente no espaço de trabalho:

```
ls()  
[1] "cursoR"
```

Verifique a classe do objeto:

```
class(cursoR)  
[1] "data.frame"
```

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
10/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise
As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

Confirme que o objeto (**cursoR**) está presente no espaço de trabalho:

```
ls()  
[1] "cursoR"
```

Verifique a classe do objeto:

```
class(cursoR)  
[1] "data.frame"
```

Verifique o nº de registros no banco de dados:

```
nrow(cursoR) # nº de linhas numa matriz ou banco de dados  
[1] 300
```

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
10/16

Fúlvio Nedel
SPB/UFSC

Introdução
Objetivos da análise
As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

Confirme que o objeto (**cursoR**) está presente no espaço de trabalho:

```
ls()  
[1] "cursoR"
```

Verifique a classe do objeto:

```
class(cursoR)  
[1] "data.frame"
```

Verifique o nº de registros no banco de dados:

```
nrow(cursoR) # nº de linhas numa matriz ou banco de dados  
[1] 300
```

Verifique o nº de variáveis no banco de dados:

```
ncol(cursoR) # nº de colunas numa matriz ou banco de dados  
[1] 169
```

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
11/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise

As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

- 1 Introdução
 - Objetivos da análise
 - As variáveis de análise

- 2 Leitura do arquivo

- 3 Seleção das variáveis de interesse

Use a função **names()** para listar os nomes das variáveis:

names()(cursoR)

[1]	"nquest"	"dataentr"	"u6"	"datanasc"	"sexo"	"u9"
[7]	"u10"	"u10.1"	"u11"	"u12"	"u13"	"u14"
[13]	"u15"	"u16"	"u17"	"u18"	"u19"	"u20"
[19]	"u21"	"u22"	"u23"	"u24.1"	"u24.2"	"u24.3"
[25]	"u24.4"	"u25"	"u26"	"u27"	"u28"	"u29"
[31]	"u30"	"u31"	"u32"	"u33"	"u34.1"	"u34.2"
[37]	"u34.3"	"u34.4"	"u34.5"	"u34.6"	"u35"	"u36"
[43]	"u36.1"	"u37"	"u38"	"u39"	"u40"	"u41"
[49]	"u42"	"u43"	"u44"	"u45"	"u46"	"peso"
[55]	"altura"	"u49"	"u50"	"u51"	"u52"	"grupohas"
[61]	"u54"	"u55.1"	"u55.2"	"u55.3"	"u55.4"	"u55.5"
[67]	"u56.1"	"u56.2"	"u56.3"	"u56.4"	"u56.5"	"u57.1"
[73]	"u57.2"	"u57.3"	"u57.4"	"u57.5"	"u58"	"u59"
[79]	"u60"	"u61"	"u62"	"grupodm"	"u64"	"u65.1"
[85]	"u65.2"	"u65.3"	"u65.4"	"u65.5"	"u66.1"	"u66.2"
[91]	"u66.3"	"u66.4"	"u66.5"	"u67.1"	"u67.2"	"u67.3"
[97]	"u67.4"	"u67.5"	"u68"	"u69"	"u70"	"u71"
[103]	"u72.1"	"u72.2"	"u72.3"	"u72.4"	"u72.5"	"u73.1"
[109]	"u73.2"	"u73.3"	"u73.4"	"u73.5"	"u74"	"u75"
[115]	"u76"	"u77"	"u78"	"u79"	"u80"	"u81"
[121]	"u82"	"u83"	"u84"	"u85"	"u86"	"u87"
[127]	"u88"	"u89"	"u90.1"	"u90.2"	"u90.3"	"u90.4"
[133]	"u90.5"	"u90.6"	"u90.7"	"u90.8"	"u90.9"	"u91"
[139]	"u92"	"u93"	"uidade"	"ufxetar"	"tempentr"	"problema"
[145]	"remedio"	"taf"	"sedent"	"naf"	"ptcage"	"cage"
[151]	"imc"	"dieta3"	"dieta2"	"dietapts"	"ncmhas"	"ncmmhas"
[157]	"tpconshas"	"ncmdm"	"ncmmdm"	"tpconsdm"	"ncodt"	"tpconsodt"
[163]	"abep"	"abep2"	"abepcls"	"abepX2"	"escola"	"morisky"

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
12/16

Fúlvio Nedel
SPB/UFSC

Introdução

Objetivos da análise

As variáveis de análise

Leitura do arquivo

Seleção das variáveis
de interesse

Use a função `names()` para listar os nomes das variáveis:

`names(cursorR)`

```
[1] "nquest"      "dataentr"    "u6"          "datanasc"    "sexo"        "u9"
[7] "u10"         "u10.1"       "u11"         "u12"         "u13"         "u14"
[13] "u15"         "u16"         "u17"         "u18"         "u19"         "u20"
[19] "u21"         "u22"         "u23"         "u24.1"       "u24.2"       "u24.3"
[25] "u24.4"       "u25"         "u26"         "u27"         "u28"         "u29"
[31] "u30"         "u31"         "u32"         "u33"         "u34.1"       "u34.2"
[37] "u34.3"       "u34.4"       "u34.5"       "u34.6"       "u35"         "u36"
[43] "u36.1"       "u37"         "u38"         "u39"         "u40"         "u41"
```

Há muito mais variáveis que as que podem nos interessar.

Vamos manter apenas as necessárias para alcançar os objetivos enunciados.

```
[85] "u56.2"       "u56.3"       "u56.4"       "u56.5"       "u56.6"       "u56.7"
[91] "u66.3"       "u66.4"       "u66.5"       "u67.1"       "u67.2"       "u67.3"
[97] "u67.4"       "u67.5"       "u68"         "u69"         "u70"         "u71"
[103] "u72.1"       "u72.2"       "u72.3"       "u72.4"       "u72.5"       "u73.1"
[109] "u73.2"       "u73.3"       "u73.4"       "u73.5"       "u74"         "u75"
[115] "u76"         "u77"         "u78"         "u79"         "u80"         "u81"
[121] "u82"         "u83"         "u84"         "u85"         "u86"         "u87"
[127] "u88"         "u89"         "u90.1"       "u90.2"       "u90.3"       "u90.4"
[133] "u90.5"       "u90.6"       "u90.7"       "u90.8"       "u90.9"       "u91"
[139] "u92"         "u93"         "uidade"      "ufxetar"     "tempentr"    "problema"
[145] "remedio"     "taf"         "sedent"     "naf"         "ptcage"      "cage"
[151] "imc"         "dieta3"      "dieta2"     "dietapts"    "ncmhas"      "ncmhas"
[157] "tpconshas"   "ncmdm"       "ncmmdm"     "tpconsdm"    "ncodt"       "tpconsodt"
[163] "abep"        "abep2"       "abepcls"    "abepX2"      "escola"      "morisky"
```

- O banco de dados é organizado com cada registro nas *filas* e cada *variável* nas *colunas*, que são entendidas pelo R como `data.frame[filas,colunas]`
- As filas e colunas podem ser chamadas pelo seu número ou nome (como vimos)
- As variáveis de interesse eram: `sexo`, `dataentr`, `datanasc`, `peso`, `altura`, `abepcls`, `grupohas`, `grupodm`
- Diremos ao R então que nos faça uma cópia do banco `cursoR` apenas com essas variáveis
- O comando pode ser escrito em um só passo, mas pode ser mais fácil de entender a sintaxe se primeiro criamos um *vetor com os nomes das variáveis*

- O banco de dados é organizado com cada registro nas *filas* e cada variável nas *colunas*, que são entendidas pelo R como `data.frame[filas,colunas]`
- As filas e colunas podem ser chamadas pelo seu número ou nome (como vimos)
- As variáveis de interesse eram: `sexo`, `dataentr`, `datanasc`, `peso`, `altura`, `abepcls`, `grupohas`, `grupodm`
- Diremos ao R então que nos faça uma cópia do banco `cursoR` apenas com essas variáveis
- O comando pode ser escrito em um só passo, mas pode ser mais fácil de entender a sintaxe se primeiro criamos um *vetor com os nomes das variáveis*

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
13/16

Fúlvio Nedel
SPB/UFSC

Introdução
Objetivos da análise
As variáveis de análise
Leitura do arquivo

Seleção das variáveis
de interesse

- O banco de dados é organizado com cada registro nas *filas* e cada variável nas *colunas*, que são entendidas pelo R como `data.frame[filas,colunas]`
- As filas e colunas podem ser chamadas pelo seu número ou nome (como vimos)
- **As variáveis de interesse eram:** `sexo`, `dataentr`, `datanasc`, `peso`, `altura`, `abepcls`, `grupohas`, `grupodm`
- Diremos ao R então que nos faça uma cópia do banco `cursoR` apenas com essas variáveis
- O comando pode ser escrito em um só passo, mas pode ser mais fácil de entender a sintaxe se primeiro criamos um *vetor com os nomes das variáveis*

Introdução ao R
3.a Importação de
um banco e
seleção de variáveis
13/16

Fúlvio Nedel
SPB/UFSC

Introdução
Objetivos da análise
As variáveis de análise
Leitura do arquivo

Seleção das variáveis
de interesse

- O banco de dados é organizado com cada registro nas *filas* e cada variável nas *colunas*, que são entendidas pelo R como `data.frame[filas,colunas]`
- As filas e colunas podem ser chamadas pelo seu número ou nome (como vimos)
- As variáveis de interesse eram: `sexo`, `dataentr`, `datanasc`, `peso`, `altura`, `abepcls`, `grupohas`, `grupodm`
- Diremos ao R então que nos faça uma cópia do banco `cursoR` apenas com essas variáveis
- O comando pode ser escrito em um só passo, mas pode ser mais fácil de entender a sintaxe se primeiro criamos um *vetor com os nomes das variáveis*

- O banco de dados é organizado com cada registro nas *filas* e cada variável nas *colunas*, que são entendidas pelo R como `data.frame[filas,colunas]`
- As filas e colunas podem ser chamadas pelo seu número ou nome (como vimos)
- As variáveis de interesse eram: `sexo`, `dataentr`, `datanasc`, `peso`, `altura`, `abepcls`, `grupohas`, `grupodm`
- Diremos ao R então que nos faça uma cópia do banco `cursoR` apenas com essas variáveis
- O comando pode ser escrito em um só passo, mas pode ser mais fácil de entender a sintaxe se primeiro criamos um *vetor* com os nomes das variáveis

Criar um banco com variáveis selecionadas:

```
vars <- c('peso', 'altura', 'sexo', 'dataentr', 'datanasc',  
          'abepcls', 'grupohas', 'grupodm')  
x <- cursoR[vars]
```

Temos então um data frame com todos os registros de “cursoR” e apenas as oito variáveis selecionadas:

```
class(x) ; nrow(x) ; ncol(x)  
[1] "data.frame"  
[1] 300  
[1] 8  
names(x)  
[1] "peso"      "altura"    "sexo"      "dataentr"  "datanasc"  "abepcls"  
[7] "grupohas"  "grupodm"
```

Veja também a função `subset`

?subset

Tudo funcionou e o objeto 'x' é o banco de dados de interesse!!

Podemos então:

- 1 chamá-lo 'cursoR', sobrescrevendo o antigo, que não nos interessa mais;
- 2 salvá-lo no computador como um arquivo de dados do R (extensão '.rdata');
- 3 remover os outros objetos da sessão de trabalho; e
- 4 carregar o arquivo de dados criado,

para continuar o trabalho com uma sessão “limpa”.

```
cursoR <- x  
save(cursoR, file="cursoR.RData")
```

```
ls() # verificar os objetos no espaço de trabalho  
rm(list=ls()) # apagar os objetos do espaço de trabalho
```


Apêndice

- Selecionar as variáveis pelo seu nome facilita a leitura humana da sintaxe, mas pode ser mais difícil de digitar e, eventualmente, algum comando necessitará a referência numérica²
- Se possível, dê nomes significativos às variáveis, é mais fácil trabalhar com uma variável chamada “sexo” que com uma variável chamada “u8” (por exemplo)
- Em bases com muitas variáveis pode ser difícil encontrar o nº de ordem das variáveis de interesse. Veja abaixo um exemplo de uso da função `%in%`:

Quais as variáveis de `cursoR` estão citadas em `vars`?

```
(nvar <- which(colnames(cursoR) %in% vars))  
[1] 1 2 3 4 5 6 7 8
```

E poderíamos então criar o mesmo banco, com as variáveis na ordem do banco original:

```
x2 <- cursoR[nvar]  
names(x2)  
[1] "peso"      "altura"    "sexo"      "dataentr"  "datanasc"  "abepcls"  
[7] "grupohas"  "grupodm"
```

²Veja, por exemplo, o comando `names(...)` <- ...