

Análise Estatística no Ambiente Computacional *R* / *R-Commander*

Material suplementar de:

BARBETTA, P. A., REIS, M. M., BORNIA, A. C. - Estatística para cursos de Engenharia e Informática. 4 ed. Rio de Janeiro: LTC, 2024.

Prefácio

O “R” é um ambiente de software livre voltado para a Estatística, podendo ser usado em várias plataformas, como Windows, Linux e MacOS. Ele tem uma estrutura central própria, complementada com milhares de pacotes voltados para situações específicas.

O *R Commander* (pacote “Rcmdr”) é uma interface gráfica que permite fazer procedimentos básicos de manipulação de dados, análise estatística e gráficos através de *menus*. Esse pacote realiza as análises solicitadas pelo *menu* e gera a linguagem computacional *R* dessas análises.

Usuários do *R*, em geral, preferem usar o ambiente proporcionado pelo *R-Studio* (<https://www.rstudio.com>), que mostra várias janelas na tela e facilita a organização das análises, mas você precisa conhecer as funções do *R* para realizar as análises. Pode-se, também, incluir o pacote *Rcmdr* no ambiente do *R-Studio*.

Estas notas sobre a utilização do *R Commander* não pretendem ser completas no sentido de explorar tudo o que contempla esse software, mas oferecer ao estudante os procedimentos comumente usados numa disciplina de Estatística.

Sumário

1 – Instalação	4
1.1 – Instalação do programa básico R.....	4
1.2 – Instalação do <i>R Commander</i>	5
2 – Os dados	9
2.1 – Carregar arquivo de dados no formato <i>RData</i>	9
2.2 – Importar arquivo de dados	11
2.3 – Salvar ou exportar arquivo de dados	14
2.4 – Definir um subconjunto do arquivo de dados	14
2.5 – Recodificar variáveis do arquivo de dados	16
2.6 – Reordenar níveis dos fatores	19
2.7 – Calcular nova variável	20
2.8 – Converter variável numérica para fator	21
3 – Análise exploratória de dados	25
3.1 – Distribuição de frequências.....	25
3.2 – Medidas descritivas	27
3.3 – Histograma e diagrama em caixas.....	30
3.4 – Diagrama de dispersão.....	32
3.6 – Coeficientes de correlação	35
3.7 – Tabela de contingência	37
4 – Distribuições de probabilidade	39
4.1 – Distribuição binomial	40
4.2 – Distribuição normal	44
4.3 – Gráfico de probabilidade normal.....	48
5 – Intervalos de confiança e testes de hipóteses para médias	50
5.1 – Intervalo de confiança e teste para uma média.....	50
5.2 – Intervalo de confiança e teste para duas amostras pareadas.....	51
5.3 – Intervalo de confiança e teste para duas amostras independentes	53
5.4 – Análise de variância com um fator.....	55
5.5 – Análise de variância para projeto fatorial	58
5.6 – Abordagem não paramétrica	59
6 – Modelos de regressão	60
6.1 – Regressão linear simples: gráfico.....	60
6.2 – Regressão linear simples: resumo analítico	62
6.3 – Regressão linear múltipla.....	65

6.4 – Análise de resíduos.....	69
Bibliografia.....	71

1 – Instalação

1.1 – Instalação do programa básico R

A página principal do projeto R está no endereço: <http://www.r-project.org/>, mas para baixar o sistema você precisa usar um *CRAN Mirror*. Se entrar no endereço principal, você tem no lado esquerdo da tela a opção de ser direcionado para um *CRAN* de preferência. No Brasil temos os seguintes *CRAN Mirror*:

<http://cran-r.c3sl.ufpr.br/> Universidade Federal do Paraná

<http://cran.fiocruz.br/> Fundação Oswaldo Cruz, Rio de Janeiro

<http://www.vps.fmvz.usp.br/CRAN/> Universidade de São Paulo, São Paulo

<http://brieger.esalq.usp.br/CRAN/> Universidade de São Paulo, Piracicaba

Ao entrar ou ser direcionado para um *CRAN*, você deve escolher a versão compatível com o sistema operacional de sua máquina (Windows, Linux ou MacOS). Na sequência, vamos considerar o sistema Windows.

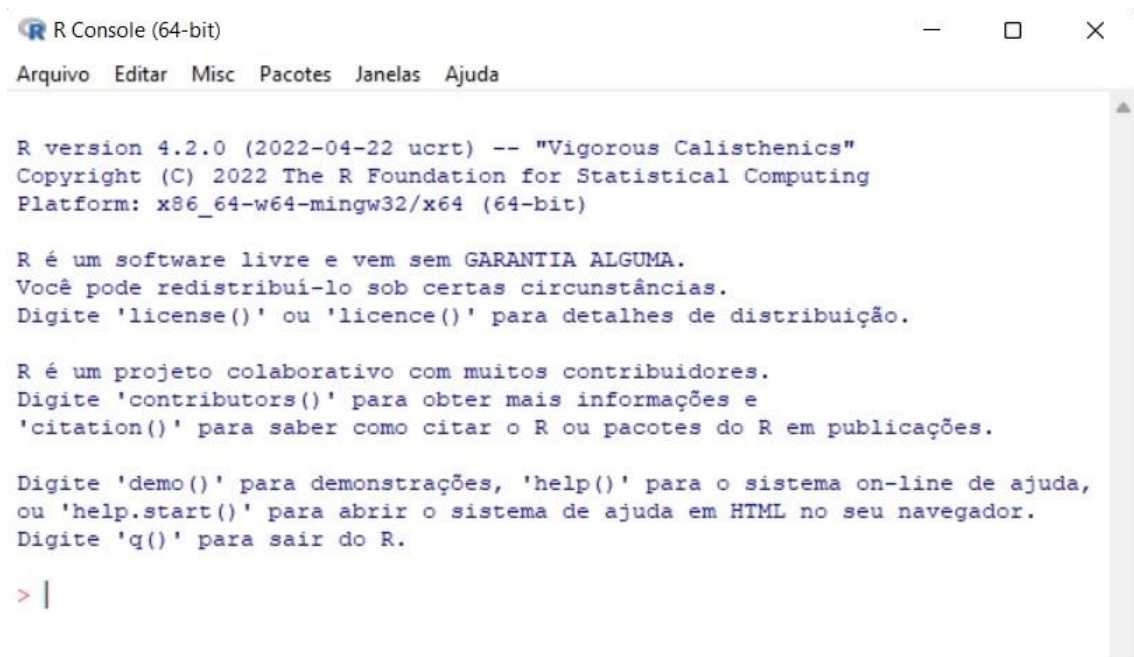
Clicar em *Download R for Windows*. Na página seguinte você deve baixar o programa básico clicando em *base*, onde vai aparecer a opção de baixar a última versão do *R para Windows*.¹

Depois de baixada a última versão, você deve executar o programa de instalação. O software pergunta se quer inicialização padrão ou

¹ Estas notas foram preparadas com a versão 4.2.0, sistema *Windows*.

personalizada. Para usar o *Rcmdr* é recomendável responder inicialização personalizada e optar pelo modo de exibição SDI; quanto as outras opções pode usar o padrão sugerido. Ao final da instalação, o *R* abre uma janela como mostrado na Figura 1.1.

Figura 1.1 – Janela inicial do *R* no sistema Windows.



```
R version 4.2.0 (2022-04-22 ucrt) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distribuição.

R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R em publicações.

Digite 'demo()' para demonstrações, 'help()' para o sistema on-line de ajuda,
ou 'help.start()' para abrir o sistema de ajuda em HTML no seu navegador.
Digite 'q()' para sair do R.

> |
```

1.2 – Instalação do *R Commander*

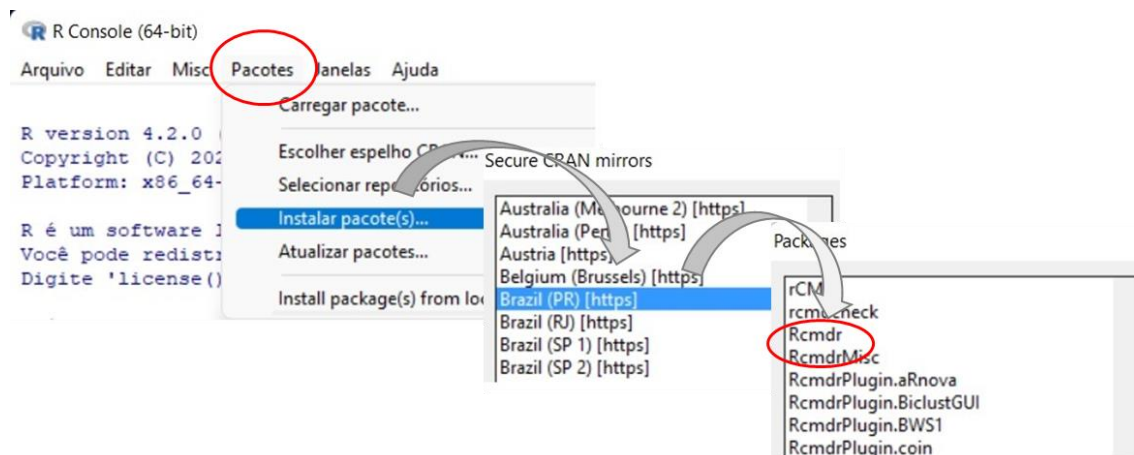
O *R Commander* é um pacote do *R* de nome *Rcmdr*. Para instalá-lo, você deve usar o procedimento geral de instalação de pacotes no *R*, para isto, inicie o software *R*.²

Uma maneira de instalar o *Rcmdr* é através do *menu* principal do *R*, clicando em *Pacotes*, depois em *Instalar pacote(s)*, como mostra a Figura 1.2. O

² A apresentação que segue baseia-se na versão 2.7.2 do pacote *Rcmdr*, sistema Windows.

sistema vai solicitar para você escolher o *CRAN Mirror* de onde buscar o(s) pacote(s). Depois disto, o sistema te apresentará uma longa lista de pacotes em ordem alfabética. Selecione *Rcmdr*.

Figura 1.2 – Ilustração de como instalar o pacote *Rcmdr* via *menu*.



Alternativamente, você pode instalar o pacote *Rcmdr* usando o comando de instalação de pacotes. Quando aparecer a linha de comando com sinal “>”, você digita:

```
>install.packages("Rcmdr")
```

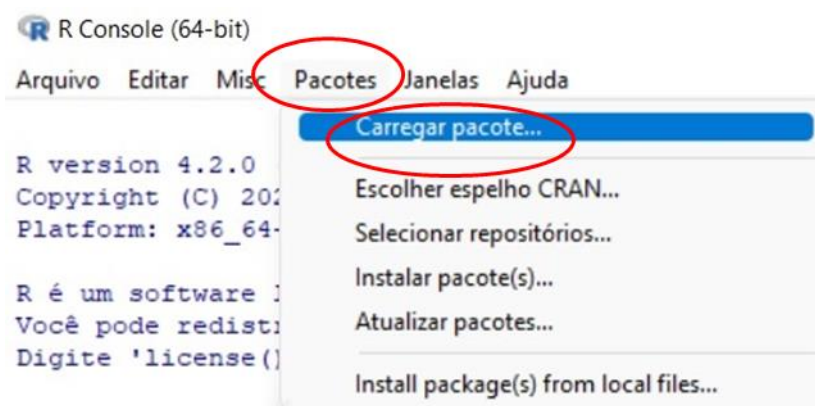
O sistema vai solicitar para você escolher o *CRAN Mirror* de onde buscar o(s) pacote(s). Escolhendo o *CRAN Mirror*, a instalação deve ter início.

A instalação do pacote *Rcmdr* costuma ser um pouco demorada, porque o sistema também instala uma série de pacotes que são necessários para executar o *Rcmdr*, chamados de *pacotes dependentes*. Fique atento nas mensagens, porque pode acontecer de o sistema dar uma mensagem de erro ou de alerta pelo fato de algum pacote não poder ser instalado

automaticamente. Neste caso, antes de iniciar o *Rcmdr*, você deve instalar esse pacote manualmente, seguindo o mesmo procedimento descrito no início deste Capítulo.

Em toda seção de uso do *R* você deve carregar os pacotes que vai usar. Em especial, ilustraremos o carregamento do pacote *Rcmdr*. Isto pode ser feito pelo *menu* principal, como mostra a Figura 1.3. Ao aparecer a lista de pacotes instalados, escolher *Rcmdr*.

Figura 1.3 – Carregar o pacote *Rcmdr* via *menu*.



Alternativamente, o carregamento de pacotes também pode ser feito pela linha de comandos, digitando após o sinal “>”:

```
>library(Rcmdr)
```

Na primeira vez que carregar o *Rcmdr*, o sistema vai dar a seguinte mensagem:

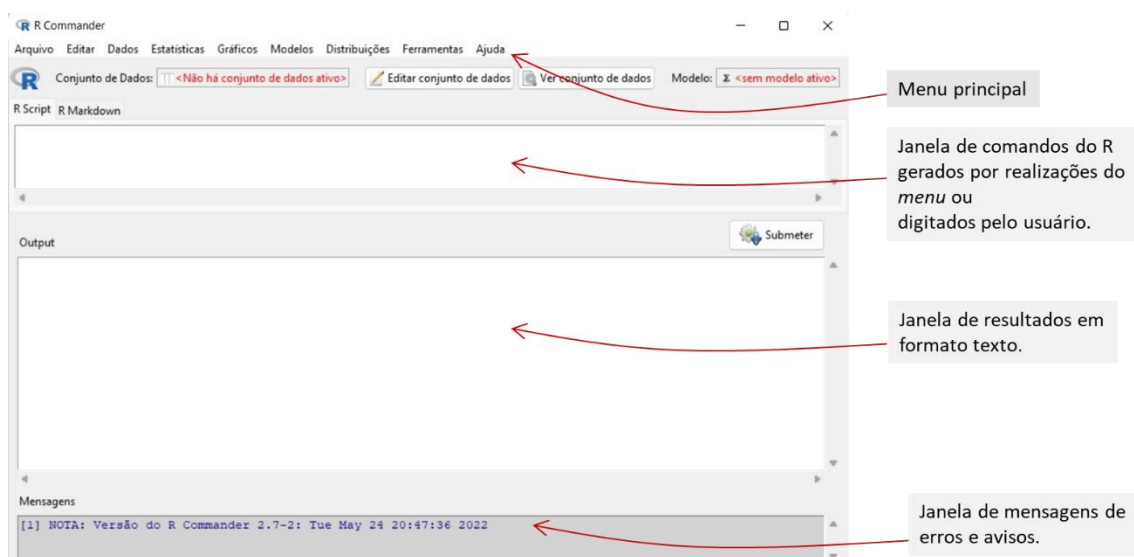
Os seguintes pacotes usados no *Rcmdr* não foram instalados...

Instalar esses pacotes?

Você deve responder que sim e aceitar as sugestões do sistema. Novamente, você precisa ter um pouco de paciência até que o *R* instale todos os pacotes necessários para o perfeito funcionamento do *Rcmdr*. Ficar atento às mensagens para verificar se algum pacote não pode ser instalado automaticamente, exigindo que você o faça separadamente.

Ao iniciar o pacote *Rcmdr*, você vai deparar com uma tela como apresentada na Figura 1.4. Ao longo dessas notas serão apresentados os procedimentos mais comuns no *R Commander*.

Figura 1.4 – A tela inicial do *Rcmdr*.



2 – Os dados

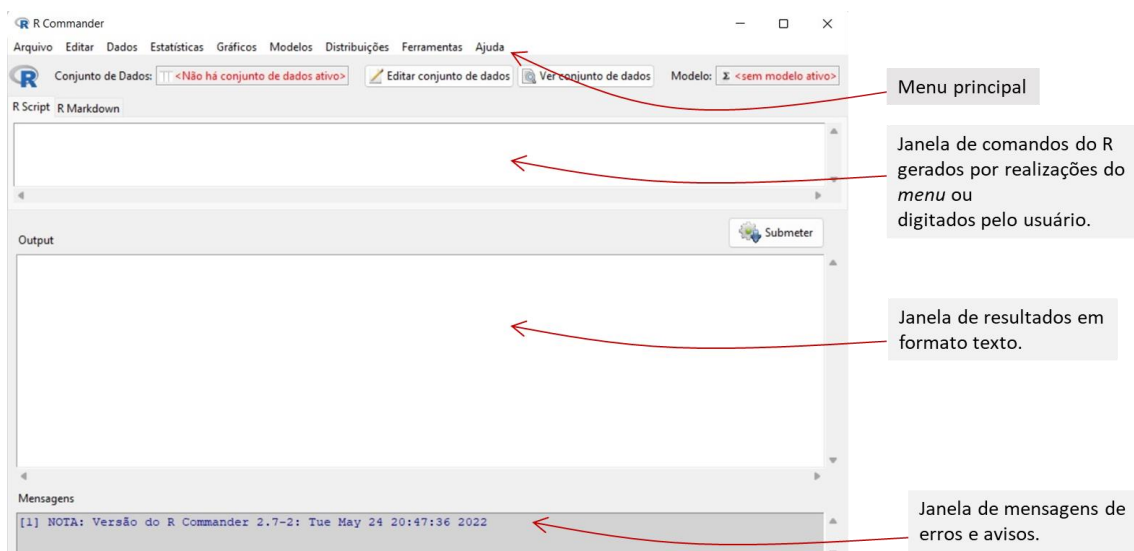
O primeiro desafio ao trabalhar com um software estatístico é como lidar com os dados:

- como carregar os dados;
- como importar dados de outro software;
- como salvar seus dados;
- como exportar seus dados;
- como fazer modificações nos dados.

2.1 – Carregar arquivo de dados no formato *RData*

O formato usual de um arquivo de dados no *R* é com a extensão *RData*. Se você já tem um arquivo nesse formato em seu computador, então você precisa carregá-lo no *R Commander*. Vamos ilustrar o processo com o arquivo *amostraEnem2019.RData*, disponível no material suplementar do livro. A Figura 2.1 ilustra o processo.

Figura 2.1 – Carregando um arquivo *RData* no *R Commander*.



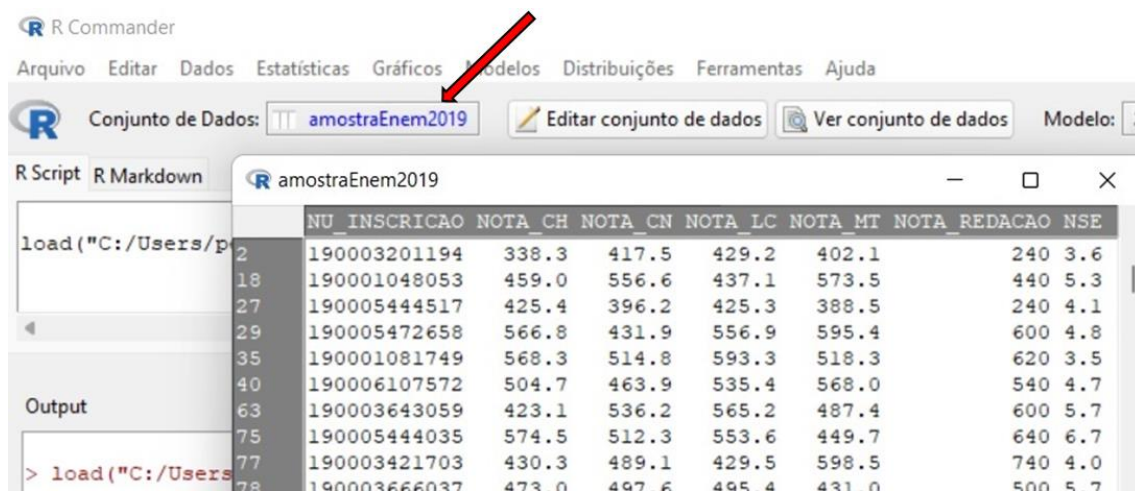
Ao selecionar *Dados >> Carregar conjunto de dados...*, como mostra a Figura 2.1, o *R Commander* abre o explorador de arquivos de seu computador para você escolher o diretório e o arquivo. Feito isto, o *R Commander* apresenta o nome do arquivo em azul (abaixo do Menu principal) e o código computacional de carregamento do arquivo (na janela *R Script*), como mostra a Figura 2.2. Além disto, na janela da base, o software fornece a informação de que o arquivo tem 2.000 linhas e 11 colunas.

Figura 2.2 – Arquivo de dados carregado no *R Commander*.



Se você quiser ver o arquivo no *R Commander*, basta clicar na caixa Ver Conjunto de Dados. O software mostrará uma janela com o arquivo de dados ativo (Figura 2.3).

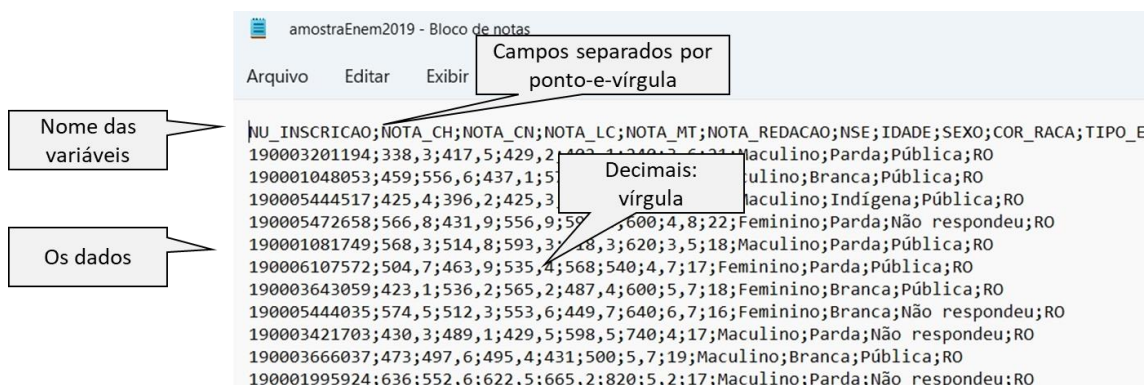
Figura 2.3 – Arquivo de dados carregado no *R Commander*.



2.2 – Importar arquivo de dados

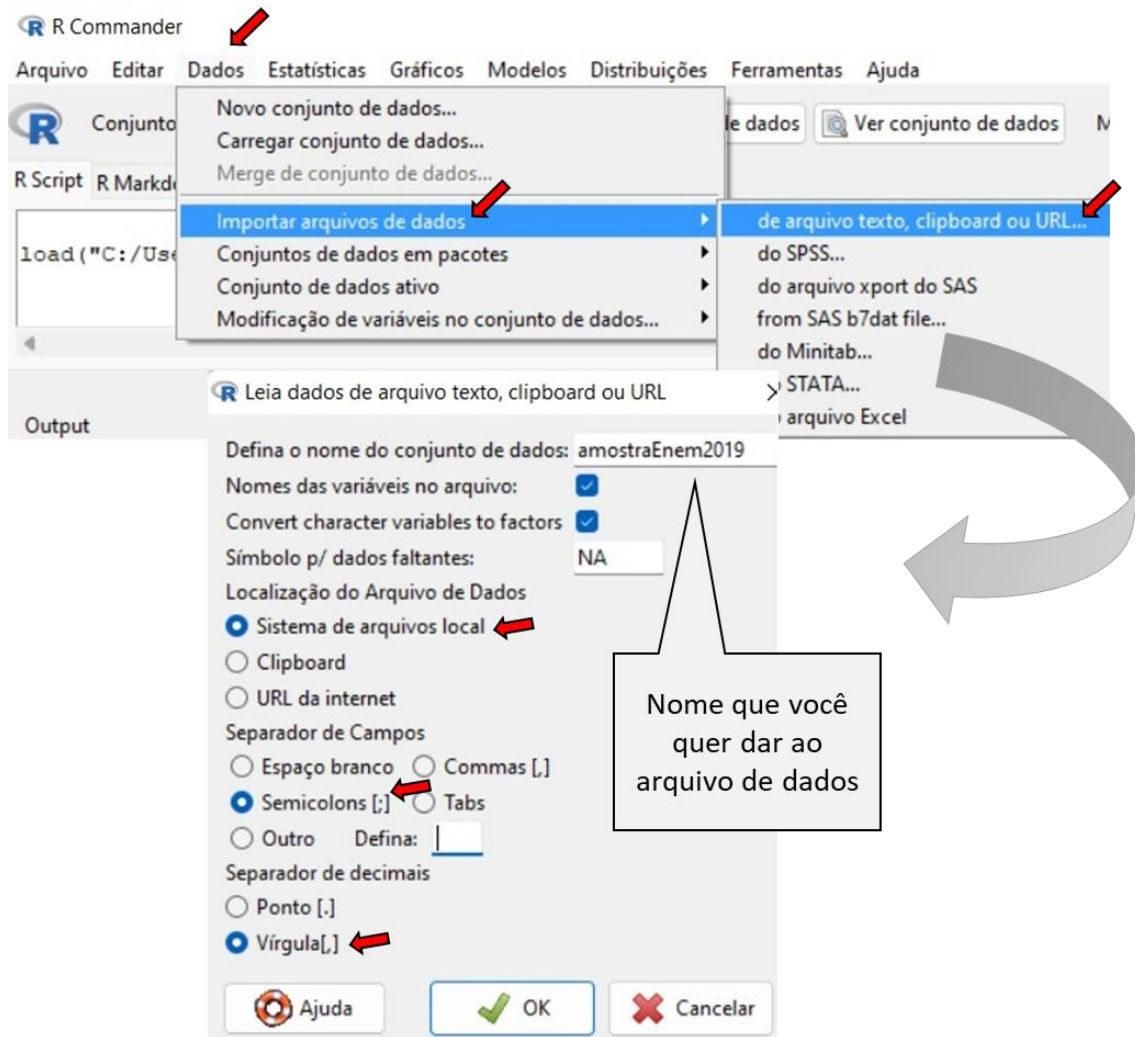
É comum termos dados em arquivos de diferentes formatos. O caso mais comum são arquivos do tipo CSV ou Texto, porque quase todo software tem opção de exportar para esses formatos. Tomemos, como exemplo, o arquivo *amostraEnem2019*, que também está disponível no material suplementar do livro no formato *texto*. A Figura 2.4 mostra as primeiras linhas e colunas desse arquivo.

Figura 2.4 – Parte de um arquivo em formato texto.



A Figura 2.5 mostra o esquema para importar esse tipo de arquivo. Depois de selecionar *Dados >> Importar arquivo de dados >> de arquivo texto, clipboard ou URL ...*, o software abre uma janela para você especificar alguns detalhes do arquivo que se quer importar. Depois disto, o *R Commander* abre o explorador de arquivos de seu computador para você escolher o diretório e o arquivo que deseja importar.

Figura 2.5 – Importando um arquivo em formato texto.



Se você tem um arquivo numa planilha do *Excel* ou do *Calc* que não seja exageradamente grande, você pode copiar o arquivo (*CTRL-C*), incluindo a primeira linha com o nome das variáveis, e seguir os procedimentos da Figura 2.5, marcando *Clipboard* em *localização do arquivo de dados*. Células em branco (sem valor ou atributo) serão repassadas como *missing* para o *R*.³

³ O símbolo de *missing* no *R* para variáveis numéricas é “NA”.

Atenção: Se seu arquivo de dados tiver apóstolo (‘) ou aspas (“) nos rótulos de alguma variável, o arquivo importado apresentará erro. Recomenda-se evitar caracteres especiais.

2.3 – Salvar ou exportar arquivo de dados

Durante uma seção do *R Commander* você pode querer guardar seu arquivo de dados *RData* para ser usado posteriormente. Para isto, você pode fazer a seguinte sequência através do Menu:

Dados >> Conjunto de dados ativo >> Salvar conjunto de dados ativo

O software abrirá o explorador de arquivos para você escolher o local que você quer guardar seu arquivo de dados. De forma similar, você pode exportar seu arquivo de dados para o formato CSV ou Texto, por:

Dados >> Conjunto de dados ativo >> Exportar conjunto de dados ativo

O *R Commander* abrirá uma janela para você definir o nome do arquivo e especificar o separador de campos e o de decimais. Terminado isto, você terá o explorador de arquivos para finalizar.

2.4 – Definir um subconjunto do arquivo de dados

Muitas vezes a análise que você quer fazer corresponde a um subgrupo dos dados do arquivo que você tem a disposição, seja em termos das variáveis

e/ou das observações. Por exemplo, no arquivo *amostraEnem2019* você poderia querer apenas as observações relativas ao Estado de São Paulo e somente as variáveis *NOTA_MT* (nota em Matemática) e *NSE* (Nível Socioeconômico). Conforme ilustrado na Figura 2.6, você deve fazer a sequência através dos *menus*:

Dados >> Conjunto de dados ativo >> Definir subconjunto de dados

O software abrirá uma janela. Se você não quiser todas as variáveis, deve desmarcar a caixa de *Incluir todas as variáveis* e marcar as variáveis que você quer no novo arquivo.⁴

Se você não quer todas as observações, você deve colocar a condição apropriada. No exemplo, queremos incluir apenas as observações cujo campo *UF_RESIDENCIA* corresponda a *SP*, então escrevemos:

$$UF_RESIDENCIA == "SP"$$

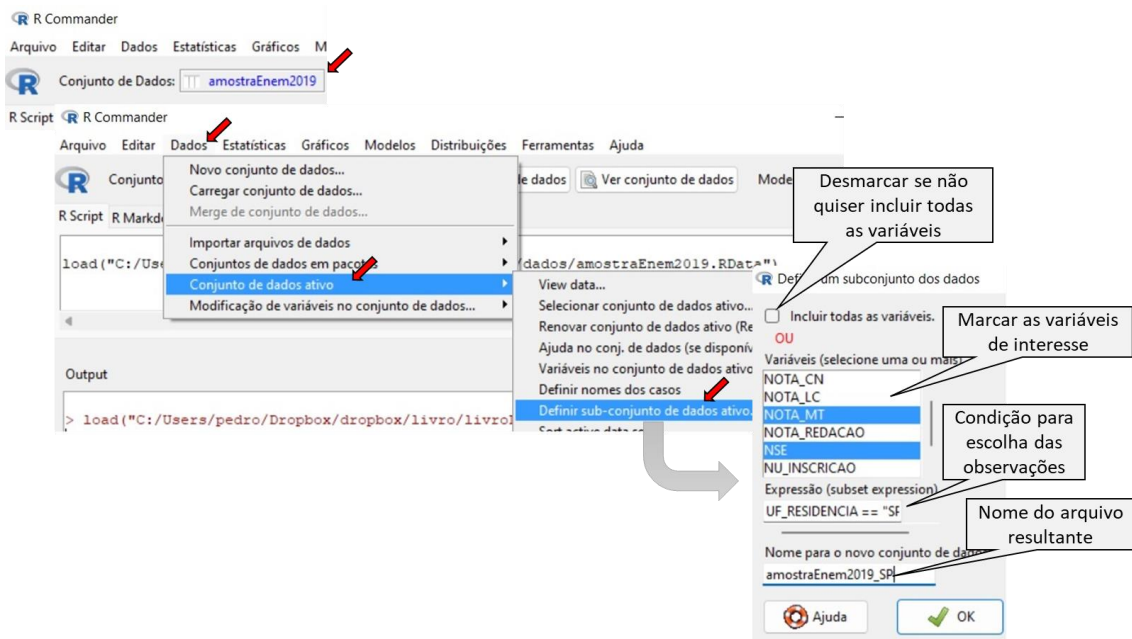
Alguns cuidados nessa fase:

- O *R* diferencia letras maiúsculas de minúsculas.
- O nome do campo (ou da variável) deve ser exatamente igual ao definido no arquivo de dados.
- Observar que usamos o símbolo “==” para relação de igualdade, já que na computação “=” é usado exclusivamente para atribuir valor a uma variável.
- O atributo de *UF_RESIDENCIA* (*SP*), foi colocado entre aspas (“SP”); isto deve ser feito se a variável não for numérica.

⁴ Para selecionar mais de uma variável, mantenha a tecla *Ctrl* apertada e escolha as variáveis.

- Se você quiser manter o arquivo original durante a sessão (recomendado), você deve atribuir um nome diferente para o arquivo formado por subgrupo de variáveis e de casos.

Figura 2.6 – Selecionando um subgrupo do arquivo ativo.



Realizando os procedimentos descritos na Figura 2.6, o novo arquivo de dados (*amostraEnem2019_SP*) ficará como arquivo ativo no *R Commander*, mas o arquivo que tinha sido carregado anteriormente (*amostraEnem2019*) poderá ser ativado a qualquer momento, basta clicar no nome do arquivo ativo que o *R Commander* abrirá uma janela para você escolher um arquivo dentre aqueles carregados (ou construídos) durante a sessão.

2.5 – Recodificar variáveis do arquivo de dados

Considere que se queira recodificar a variável *NOTA_MT* em três categorias, assim definidas:

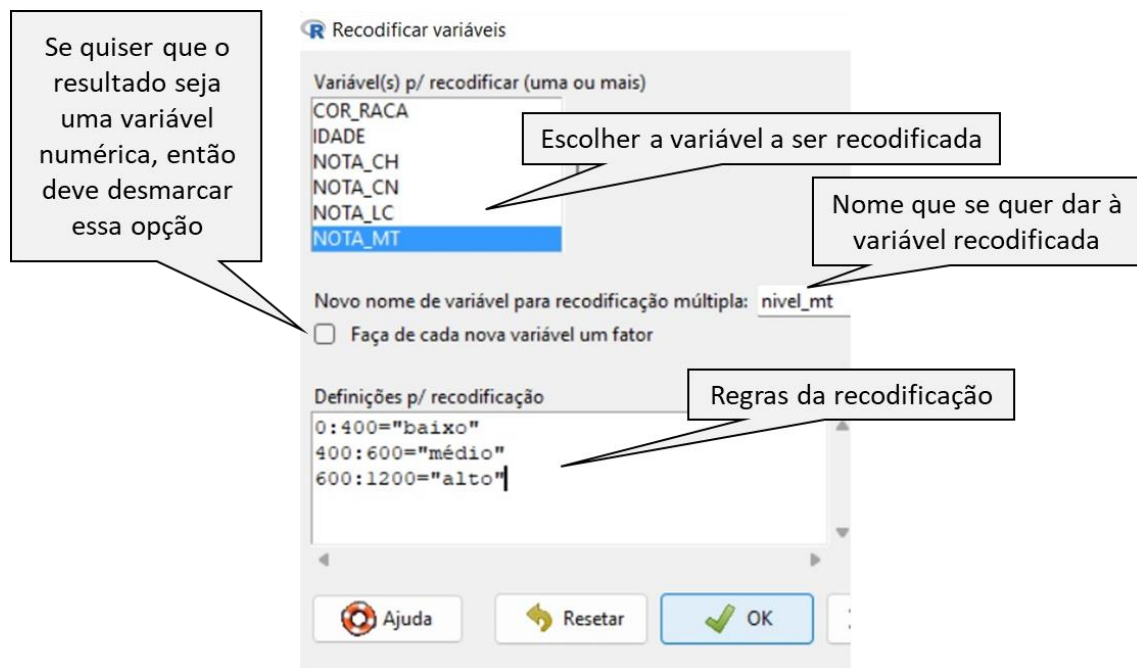
- *baixo* se $NOTA_MT \leq 400$
- *médio* se $400 < NOTA_MT \leq 600$
- *alto* se $NOTA_MT > 600$

A nova variável que contempla esses três níveis será chamada de *niveis_mt*. Para fazer isto, devemos executar nos *menus* do *R Commander* a seguinte sequência:

Dados >> Modificação de variáveis no arquivo de dados >> Recodificar variáveis ...

O *R Commander* abrirá uma janela para preenchimento, conforme ilustra a Figura 2.7.

Figura 2.7 – Recodificando a variável *NOTA_MT* do arquivo *amostraEnem2019*.




Inicialmente, devemos escolher a variável que queremos fazer a recodificação, que aparece no topo da Figura 2.7, em azul. Em seguida, o nome que queremos atribuir a essa variável recodificada, onde escolhemos *nivel_mt*.

A janela já vem com marcação na caixa de “Faça de cada variável um fator”. O termo *fator* refere ao formato usual de uma variável não numérica. Sendo a variável recodificada do tipo *fator*, devemos definir as suas categorias entre aspas, por isto do lado direito das igualdades dessa janela tem-se: “baixo”, “médio” e “alto” (entre aspas). Se você estiver criando uma variável numérica, então não deve colocar os números entre aspas e a caixa “Faça de cada variável um fator” deve ser desmarcada.

Do lado esquerdo das igualdades, tem-se números separados por “:” (dois pontos). Assim, 0:400 significa o intervalo de 0 a 400. Se um número coincidir com um dos limites do intervalo, vale a declaração que vier primeiro. Nesses intervalos valem também “lo” e “hi”. Assim, o primeiro e o terceiro intervalos poderiam ser assim escritos: *lo:400* e *600:hi*, respectivamente.

Depois de feita essa recodificação, se você clicar em “Ver conjunto de dados”, verá que o conjunto de dados contempla a variável *nivel_mt* em seu lado direito, conforme ilustra a Figura 2.8.

Figura 2.8 – Parte do arquivo de dados com o resultado da recodificação.



	NOTA_MT	NOTA_REDACAO	NSE	IDADE	SEXO	COR_RACA	TIPO_ESCOLA	UF_RESIDENCIA	nivel_mt
...	402.1	240	3.6	21	Maculino	Parda	Pública	RO	médio
	573.5	440	5.3	18	Maculino	Branca	Pública	RO	médio
	388.5	240	4.1	18	Maculino	Indígena	Pública	RO	baixo
	595.4	600	4.8	22	Feminino	Parda	Não respondeu	RO	médio
	518.3	620	3.5	18	Maculino	Parda	Pública	RO	médio
	568.0	540	4.7	17	Feminino	Parda	Pública	RO	médio
	487.4	600	5.7	18	Feminino	Branca	Pública	RO	médio
	449.7	640	6.7	16	Feminino	Branca	Não respondeu	RO	médio
	598.5	740	4.0	17	Maculino	Parda	Não respondeu	RO	médio
	431.0	500	5.7	19	Maculino	Branca	Pública	RO	médio
	665.2	820	5.2	17	Maculino	Parda	Não respondeu	RO	alto
...	696.9	880	6.1	18	Maculino	Parda	Pública	RO	alto

Uma mesma recodificação pode ser realizada para várias variáveis, bastando marcar as variáveis em que a recodificação será realizada. O que você escrever na caixa de “Novo nome de variável para recodificação múltipla” será o prefixo das variáveis decorrentes da recodificação. Se deixar essa caixa em branco, então o software irá substituir as variáveis marcadas pelos resultados da recodificação.

Mais exemplos de recodificação você pode encontrar ao clicar em “Ajuda” na janela de recodificação.

2.6 – Reordenar níveis dos fatores

Ao realizar algum procedimento estatístico com a variável *nivel_mt* criada conforme descrito na seção anterior, o *R* apresentará as categorias ordinais dessa variável na seguinte ordem: *alto*, *baixo* e *médio*, ou seja, em ordem alfabética. Contudo, é natural querermos que a ordem de apresentação seja na ordem natural do significado desses atributos, ou seja: *baixo*, *médio* e *alto*. Para isto, fazer:

Dados >> Modificação de variáveis no arquivo de dados >> Reordenar níveis de fatores

O software apresentará uma janela em que você poderá ajustar a ordem desejada de apresentação dos atributos ordinais: 1 para o primeiro; 2 para o segundo; e assim por diante.

2.7 – Calcular nova variável

Outro procedimento muito comum em um arquivo de dados é a necessidade de se fazer algum cálculo com as variáveis quantitativas do arquivo de dados. No arquivo *amostraEnem2019* podemos estar interessados em obter a média das cinco notas do Enem. Neste caso, fazemos a seguinte sequência:

Dados >> Modificação de variáveis no arquivo de dados >> Computar nova variável

Na janela que se abre você deve definir o nome da nova variável e a expressão matemática. Em nosso exemplo, definimos:

- Novo nome de variável: MEDIA
- Expressão p/ computar: $(\text{NOTA_CH} + \text{NOTA_CN} + \text{NOTA_LC} + \text{NOTA_MT} + \text{NOTA_REDACAO})/5$

Realizando esse procedimento, uma nova variável, de nome MEDIA, será criada no arquivo de dados. Se para algum indivíduo não houver alguma das cinco notas, o valor da variável MEDIA para esse indivíduo será *missing*, simbolizado por NA.

2.8 – Converter variável numérica para fator

O *R Commander* tem procedimentos próprios para variáveis quantitativas e para variáveis qualitativas, então se seu arquivo tem alguma variável qualitativa com códigos numéricos é conveniente você transformá-la para *fator*.

No arquivo mostrado na Figura 2.9, a variável *sexo* está codificada com 1 e 2, sendo 1 para o sexo masculino e 2 para o sexo feminino. Ao importar esse arquivo, todas as variáveis vão ser do tipo *numérico* (ou tipo *inteiro*, já que não há casas decimais).

Figura 2.9 – Parte de um arquivo com alturas de indivíduos, altura de seus pais e sexo.

Indivíduo	altura	alt_pai	alt_mae	sexo
1	173	185	170	2
2	164	168	156	1
3	172	175	167	1
4	165	177	163	2
5	167	179	169	2
6	189	190	174	1
...

Uma função importante do *R*, mas não contemplada no *menu* do *R Commander*, é `str()`, que mostra a estrutura do arquivo colocado dentro do parêntesis. Considerando que você importou o arquivo indicado na Figura 2.9 e deu o nome de *dados_de_alturas*, você pode escrever na janela de Script:

```
> str(dados_de_alturas)
```

e, com o cursor na linha do comando, clicar em *Submeter* (lado direito da janela de Script). O software mostrará alguns detalhes do arquivo, como mostra a Figura 2.10.

Figura 2.10 – Estrutura do arquivo *dados_de_alturas*.

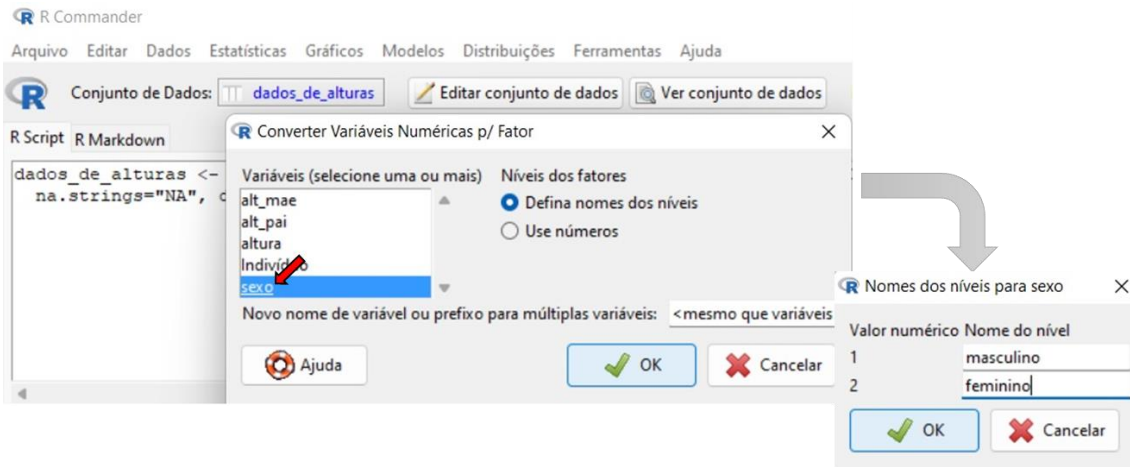
```
Output
> str(dados_de_alturas)
'data.frame': 22 obs. of 5 variables:
 $ Indivíduo: int  1 2 3 4 5 6 7 8 9 10 ...
 $ altura   : int  173 164 172 165 167 189 172 158 192 175 ...
 $ alt_pai  : int  185 168 175 177 179 190 187 169 195 172 ...
 $ alt_mae  : int  170 156 167 163 169 174 175 150 172 168 ...
 $ sexo     : int  2 1 1 2 2 1 2 2 1 1 ...
```

Repare que a variável *sexo* está no formato *int* (inteiro). Para converter *sexo* para *fator*, faça:

*Dados >> Modificação de variáveis no arquivo de dados >> Converter
variável numérica para fator*

O *R Commander* abrirá uma janela como indicado na Figura 2.10. Você deve marcar a variável que quer converter; decidir se quer colocar nomes nos níveis ou manter os números; e se quer criar nova variável ou substituir a existente. Se optar por definir nomes aos níveis, após o software pedir confirmação para substituir a variável, ele abrirá uma janela para você escrever os nomes (lado direito da Figura 2.11).

Figura 2.11 – Convertendo a variável *sexo* para fator.



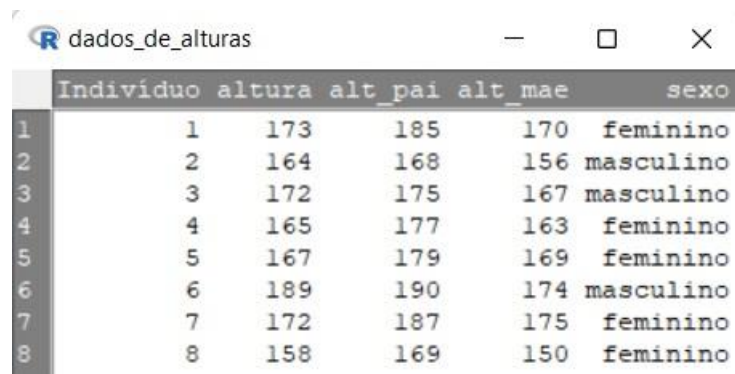
Depois que você executar os procedimentos da Figura 2.11, a estrutura do arquivo de dados mostra os rótulos da variável *sexo* (Figura 2.12).

Figura 2.12 – Estrutura do arquivo depois da mudança na variável *sexo*.

```
> str(dados_de_alturas)
'data.frame': 22 obs. of 5 variables:
 $ Indivíduo: int  1 2 3 4 5 6 7 8 9 10 ...
 $ altura   : int  173 164 172 165 167 189 172 158 192 175 ...
 $ alt_pai  : int  185 168 175 177 179 190 187 169 195 172 ...
 $ alt_mae  : int  170 156 167 163 169 174 175 150 172 168 ...
 $ sexo     : Factor w/ 2 levels "masculino","feminino": 2 1 1 2 2 1 2 2 1 1 ...
```

Note que agora a variável *sexo* está identificada como *fator*. Se você clicar na caixa “Ver conjunto de dados” o software mostra os dados como na Figura 2.13.

Figura 2.13 – Arquivo de dados após a transformação da variável *sexo* para *fator*.



```
R dados_de_alturas
```

	Individuo	altura	alt_pai	alt_mae	sexo
1	1	173	185	170	feminino
2	2	164	168	156	masculino
3	3	172	175	167	masculino
4	4	165	177	163	feminino
5	5	167	179	169	feminino
6	6	189	190	174	masculino
7	7	172	187	175	feminino
8	8	158	169	150	feminino

3 – Análise exploratória de dados

Neste Capítulo vamos usar o *R Commander* para realizar análises estatísticas. Iniciaremos com tabelas e medidas descritivas, depois veremos como fazer gráficos.

3.1 – Distribuição de frequências

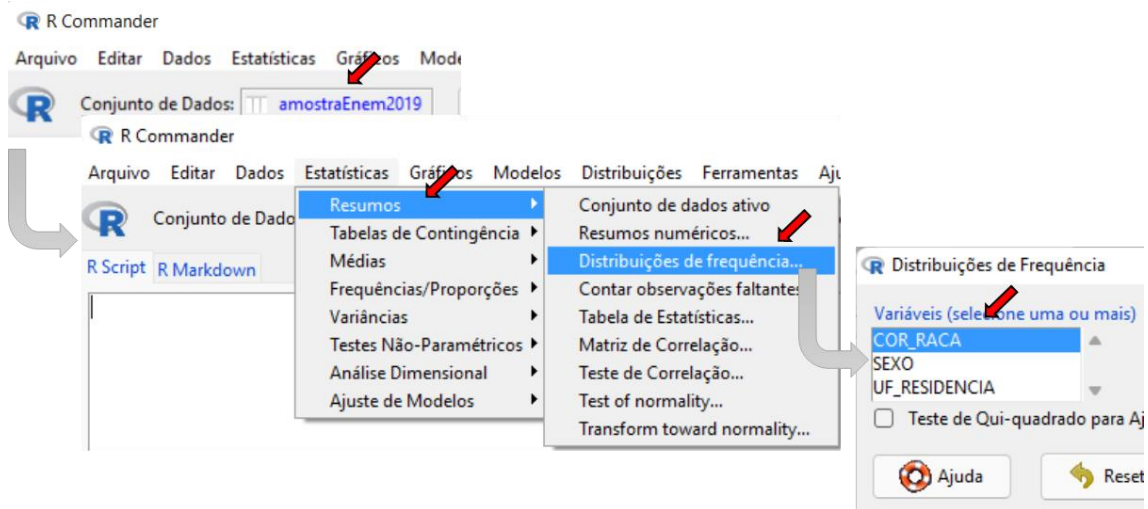
Em se tratando de variáveis qualitativas, que no *R* geralmente é definida como *fator*, você pode obter facilmente uma distribuição de frequências, fazendo a seguinte sequência nos *menus*:

Estatísticas >> Resumos >> Distribuições de frequências

Na janela que se abre, escolher a(s) variável(eis), como mostra a Figura 3.1, que ilustra a realização de uma distribuição de frequências para a variável *COR-RACA* do arquivo *amostraEnem2019*.⁵

⁵ O carregamento e a importação de um arquivo para o *R Commander* são discutidos na seção 2, seções 2.1 e 2.2.

Figura 3.1 – Esquema para fazer uma distribuição de frequências de variável qualitativa.



Ao realizar os procedimentos descritos na Figura 3.1, o *R Commander* apresenta os resultados em formato texto (Figura 3.2).

Figura 3.2 - Distribuição de frequências da cor ou raça declarada pelo candidato.

```
Output
counts:
COR_RACA
  Branca      Parda      Preta      Amarela      Indígena Não declarado
      754        904        239        50          11          42

percentages:
COR_RACA
  Branca      Parda      Preta      Amarela      Indígena Não declarado
  37.70      45.20      11.95      2.50        0.55        2.10
```

Você deve ter notado que ao pedir uma distribuição de frequências aparecem apenas as de variáveis qualitativas (tipo *fator*) do arquivo de dados. Se você tem uma variável quantitativa discreta que não precise de agrupamentos, você pode transformá-la em fator (ver seção 2.8) e usar os procedimentos descritos na Figura 3.1 para obter a distribuição de frequências.

No caso de ser uma variável contínua, ou mesmo uma variável discreta que precise de agrupamentos, você pode efetuar uma recodificação (ver seção 2.5) ou, ainda, de forma automática, fazer:

Dados >> Modificação de variáveis no arquivo de dados >> Agrupar em classes uma variável numérica

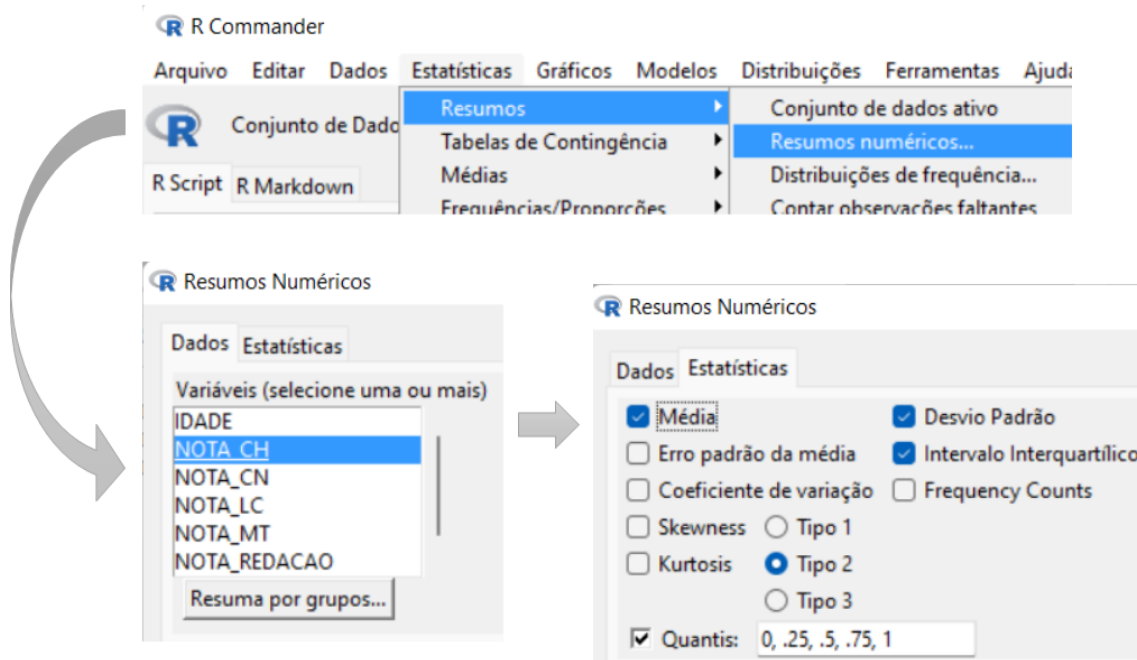
3.2 – Medidas descritivas

Em geral, ao analisarmos observações de uma variável quantitativa, desejamos obter medidas descritivas, como média, desvio padrão etc. Com o arquivo de interesse ativo, fazer a seguinte sequência nos *menus*:

Estatísticas >> Resumos >> Resumos numéricos

A Figura 3.3 mostra os procedimentos, considerando o arquivo *amostraEnem2019* e a variável *NOTAS_CH*. Na parte de cima é mostrada a sequência através dos *menus*; abaixo e à esquerda como escolher a(s) variável(eis) e à direita as opções de medidas disponíveis que aparece ao clicar na aba *Estatísticas*.

Figura 3.3 – Esquema para obter medidas descritivas



Ao clicar em *OK*, o software apresenta os seguintes resultados como mostrado na Figura 3.4⁶

Figura 3.4 – Medidas descritivas das notas em Ciências Humanas

mean	sd	IQR	0%	25%	50%	75%	100%	n
513,4	80,44	119,4	325,9	452,2	517,7	571,5	741,9	2000

Sendo:

- mean média aritmética simples;
- sd desvio padrão;
- IQR desvio interquartilico;
- 0% valor mínimo;
- 25% primeiro quartil;

⁶ Antes de pedir ao software as medidas descritivas optamos em fazer com que as decimais sejam separadas por vírgula e os números apresentados com quatro dígitos significativos. Isto pode ser feito submetendo o seguinte comando antes de solicitar as medidas descritivas: `options(OutDec= ",", digits=4)`.

- 50% mediana;
- 75% terceiro quartil;
- 100% máximo;
- n número de observações que foram usadas nos cálculos.⁷

Muitas vezes queremos medidas descritivas separadas por categorias de alguma variável qualitativa, como medidas descritivas das notas de Ciências Humanas por cor ou raça do candidato.

Observar, no lado esquerdo e de baixo da Figura 3.3, a caixa *Resuma por grupos*. Clicando nesta caixa, você pode inserir uma variável tipo fator que defina os grupos de interesse. Exemplificaremos com medidas descritivas de *NOTAS_CH* por *COR_RACA* do candidato. Considerando, porém, que algumas categorias de *COR_RACA* têm poucas observações, decidimos fazer agrupamentos de categorias, efetuando as seguintes regras de recodificação:⁸

- "Branca"="Branca ou Amarela"
- "Amarela"="Branca ou Amarela"
- "Preta"="Preta ou Indígena"
- "Indígena"="Preta ou Indígena"

Após a recodificação, fizemos os procedimentos para obter medidas descritivas, como indicado na Figura 3.3, mas antes de apertar OK, entramos na caixa *Resuma por grupos* e colocamos a variável (fator) *COR_RACA*

⁷ Se houvesse valores faltantes, haveria nova coluna, indicada com NA, dando a quantidade de valores faltantes.

⁸ Ver seção 2.5 para saber como fazer recodificações.

recodificada, que foi definida como *COR_RACA_recod*. Os resultados fornecidos pelo software são apresentados na Figura 3.5.

Figura 3.5 – Medidas descritivas das notas em Ciências Humanas, por cor ou raça do candidato.

	mean	sd	IQR	0%	25%	50%	75%	100%	NOTA_CH:n
Branca ou Amarela	534,8	79,54	115,9	326,2	478,9	540,0	594,8	741,9	804
Parda	498,6	79,30	118,2	325,9	439,2	500,2	557,4	699,4	904
Preta ou Indígena	501,4	72,95	103,6	347,0	450,7	507,4	554,3	682,9	250
Não declarado	492,3	74,39	101,6	336,8	443,8	494,9	545,4	624,8	42

Para o objetivo de apresentar alguma medida descritiva específica e os dados segregados por uma ou mais variáveis qualitativas, pode-se usar os procedimentos:

Estatísticas >> Resumos >> Tabelas estatísticas

3.3 – Histograma e diagrama em caixas

Um histograma de frequências de uma variável quantitativa pode ser facilmente realizado com a seguinte sequência:

Gráficos >> Histograma

Na janela aberta pelo software, aba *Dados*, você escolhe a variável. Na aba *Opções* você pode escolher o número de classes do histograma. Se deixar *<auto>*, o software aplica o método de Sturges para determinar o número de classes. Você também pode optar pelo tipo de frequências (contagens, porcentagens ou densidades de frequência), além de poder colocar título no gráfico e nos eixos. Com o esquema da Figura 3.6, o software abre uma janela gráfica e mostra o histograma da Figura 3.7.

Figura 3.6 – Esquema para o Rcmdr fazer um histograma

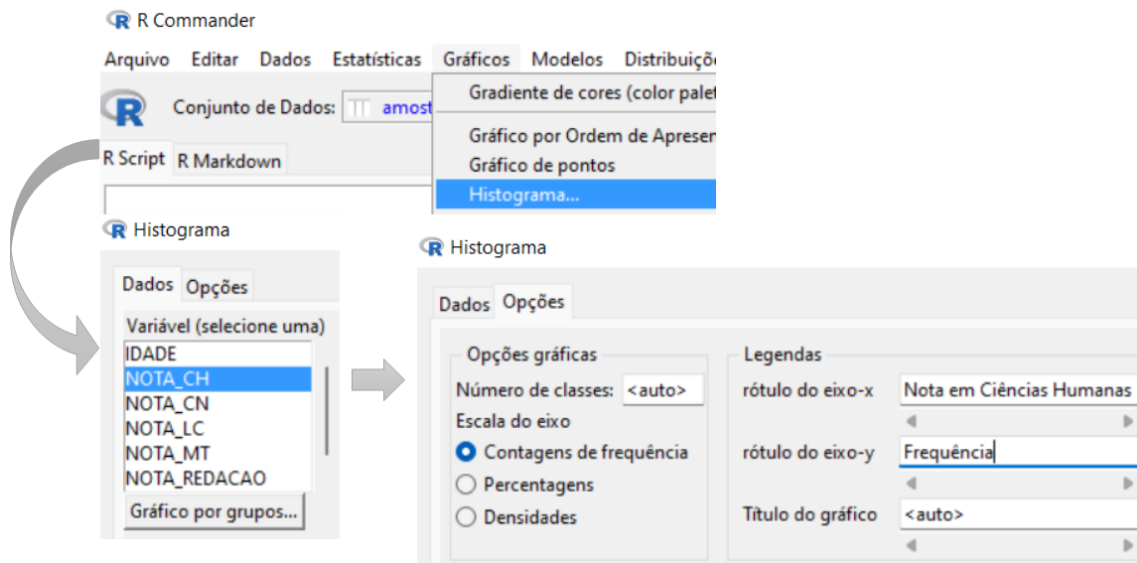
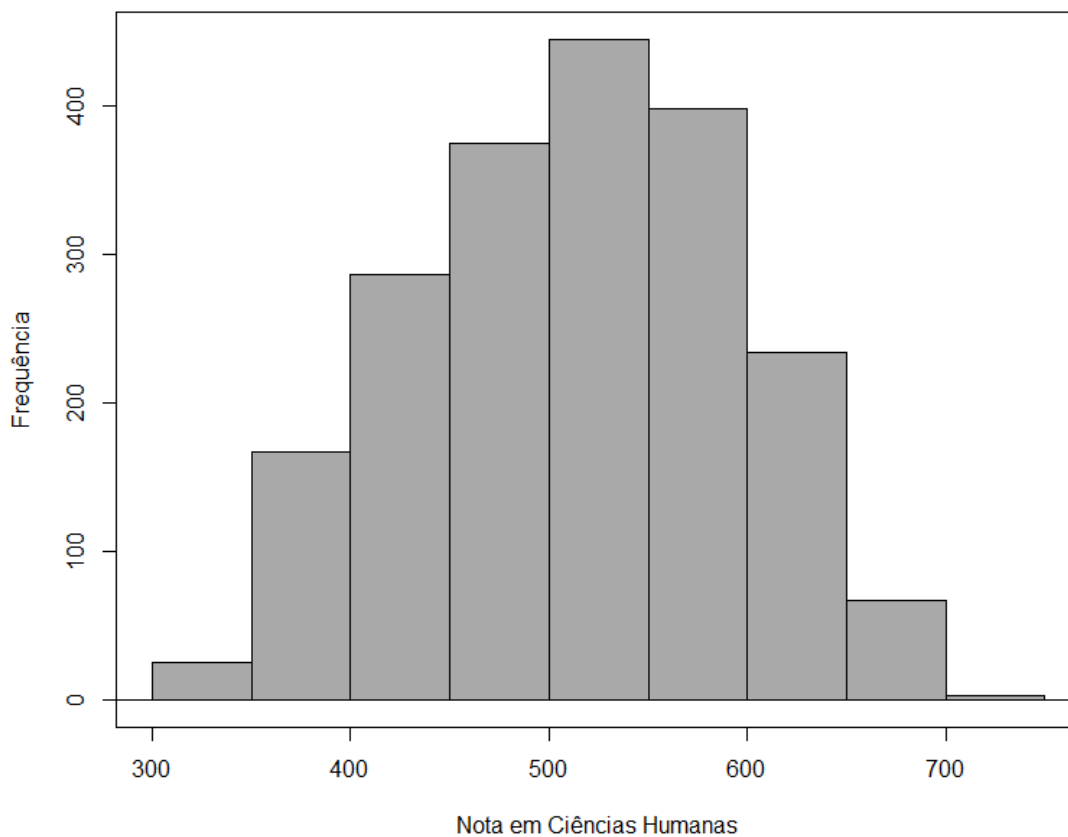


Figura 3.7 – Histograma de frequências das notas de Ciências Humanas.

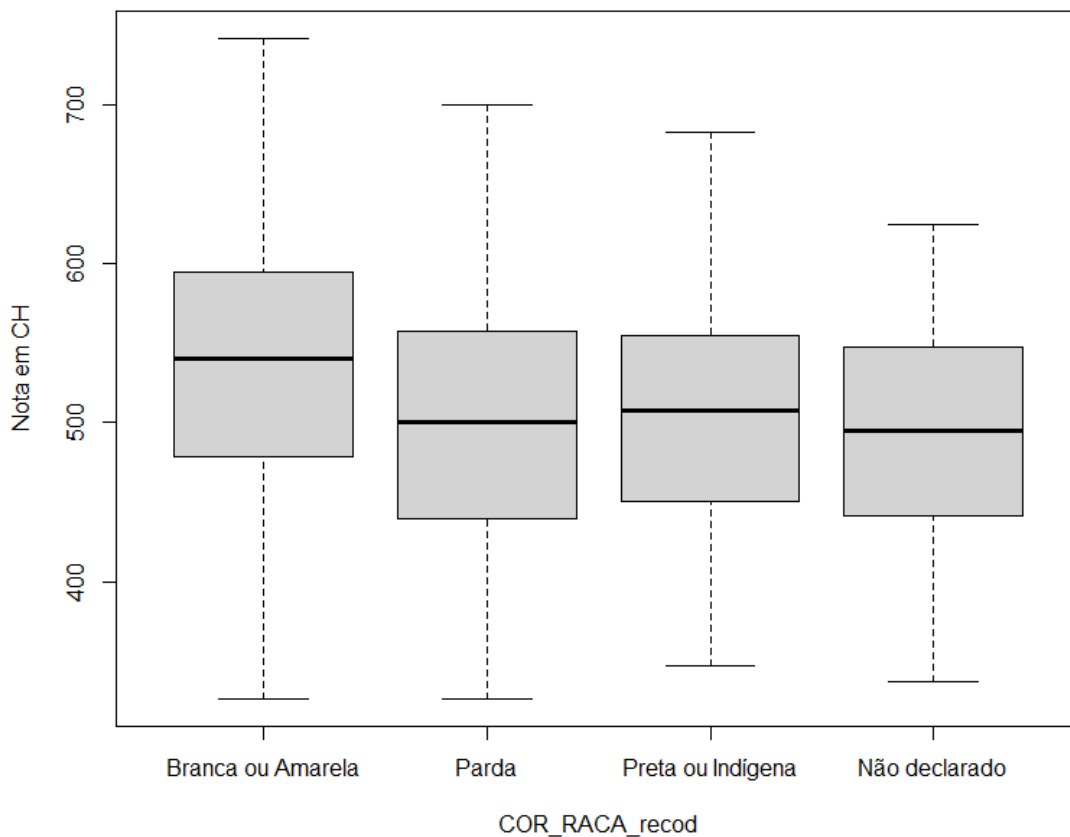


O diagrama em caixas (*boxplot*) pode ser feito pela sequência:

Gráficos >> Boxplot

O software abre uma janela em que na aba *Dados* você marca a variável que quer o gráfico e, na aba *Opções*, você escolhe como deseja que o software identifique os valores discrepantes e quais os títulos que você quer que apareçam. Além disto, você pode abrir a caixa *Gráfico por grupos* e escolher uma variável, tipo fator, que define os grupos. A Figura 3.8 mostra diagramas em caixas em que se escolheu a variável *NOTA_CH* e, em *Gráfico por grupos*, o fator *COR_RACA_recod*.

Figura 3.8 – Diagrama em caixas das notas de Ciências Humanas, por cor ou raça do candidato.



3.4 – Diagrama de dispersão

Diagrama de dispersão entre um par de variáveis quantitativas pode ser obtido através da sequência:

Gráficos >> Diagrama de dispersão

Conforme mostra a Figura 3.9, o software abre uma janela. Na aba *Dados* você escolhe as duas variáveis; na aba *Opções* você define várias características do gráfico. No presente exemplo, nós apenas definimos os rótulos dos eixos. Como resultado, o software apresenta o gráfico mostrado na Figura 3.10.

Figura 3.9 – Esquema para fazer diagrama de dispersão.

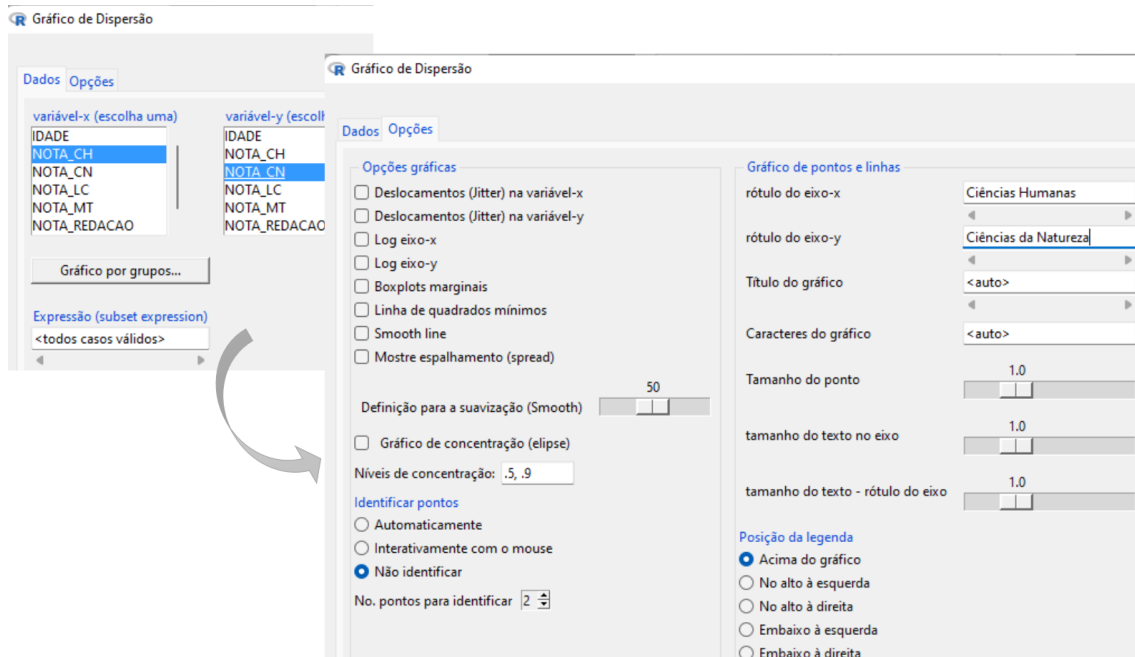
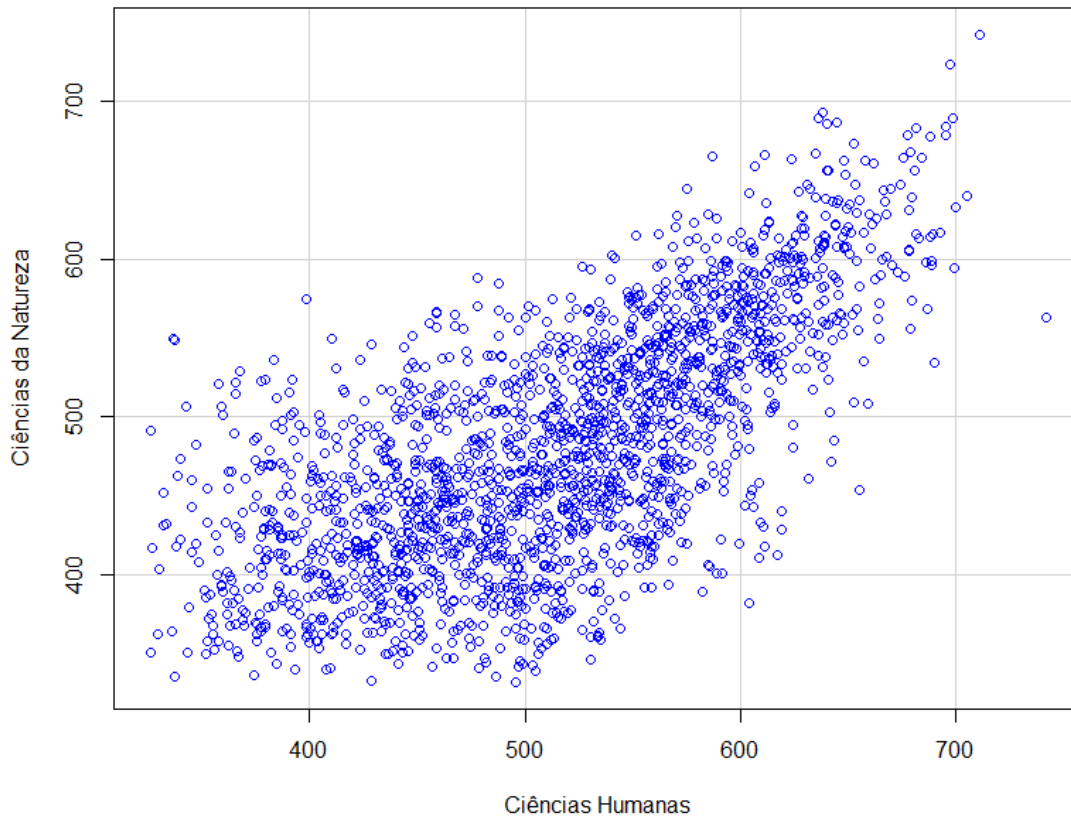


Figura 3.10 – Diagrama de dispersão entre as notas de Ciências Humanas e Ciências da Natureza.

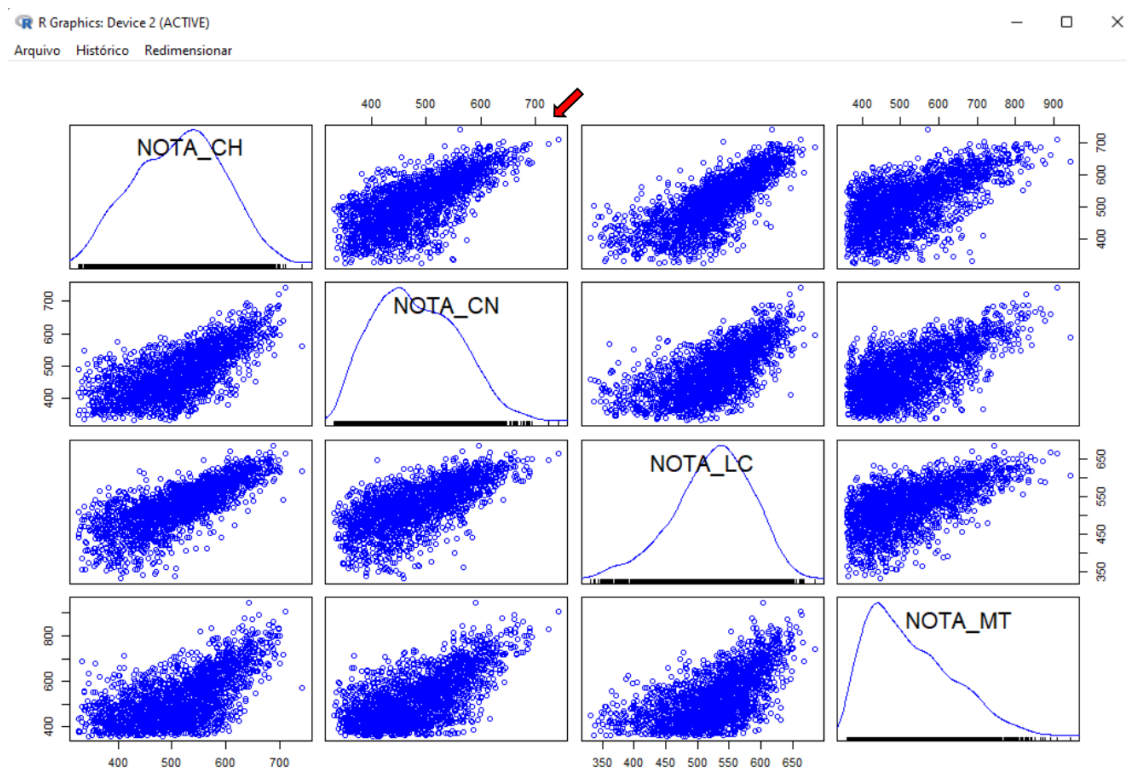


Você também pode construir um painel com diagramas de dispersão entre vários pares de variáveis, fazendo-se:

Gráficos >> Matriz de dispersão

Na janela que se abre, aba *Dados*, você escolhe as variáveis quantitativas que pretende colocar no painel; na aba *Opções* você pode optar pelo conteúdo da análise univariada que aparece na diagonal principal do painel e outras características gráficas. A Figura 3.11 mostra um painel (matriz) de diagramas de dispersão com as quatro notas das provas objetivas do Enem, sendo que nas *opções* optamos por manter o padrão sugerido pelo software.

Figura 3.11 – Painel de diagramas de dispersão entre notas do Enem.



Na diagonal principal tem-se um histograma suavizado de cada variável. Fora da diagonal principal tem-se os diagramas de dispersão de cada par de variáveis. O diagrama marcado por uma seta, por exemplo, tem no eixo horizontal a *NOTA_CH* e no eixo vertical a *NOTA_CN*, sendo as escalas das notas indicadas acima e à direita do painel.

3.6 – Coeficientes de correlação

Coeficientes de correlação entre variáveis quantitativas podem ser obtidos fazendo-se a sequência:

Estatísticas >> Resumos >> Matriz de Correlação

Conforme mostra a Figura 3.12, o software abre uma janela para você escolher as variáveis. Nessa janela você também pode escolher o tipo de correlação, dentre as opções:

- *Produto momento de Pearson*: é o coeficiente de correlação tradicional, conforme apresentado em nosso livro;
- *Sperman (rank order)*: calcula o coeficiente de correlação considerando a ordenação (1, 2, 3, ..., n) conforme a ordem

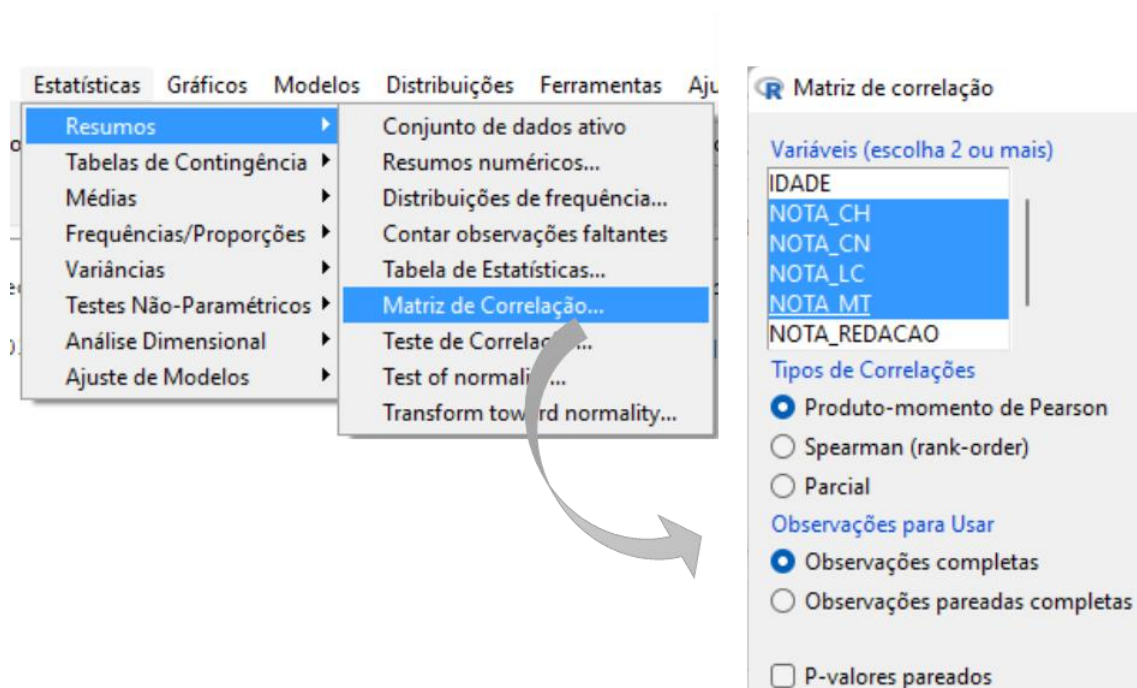
decrecente dos valores, particularmente importante quando as variáveis têm distribuições muito assimétricas e/ou presença de valores discrepantes;

- *Parcial*: avalia a correlação de um par de variáveis, controlada pelo mesmo padrão das outras variáveis; por exemplo, você pode querer a correlação entre as notas nas provas, mas considerando estudantes com o mesmo nível socioeconômico.

No cálculo do coeficiente de correlação entre um par de variáveis é necessário que se tenha pares de observações completas, sem dado faltante. Na janela, você deve escolher entre *Observações completas* e *Observações pareadas completas*. No primeiro caso, todo indivíduo que tiver observação faltante em uma das variáveis escolhidas será eliminado do cálculo. Se escolher *Observações pareadas completas*, o indivíduo será excluído da análise só nos cálculos que envolvem a variável com dado faltante.

A janela também apresenta a opção de incluir o *valor-p*, associado a um teste estatístico bilateral, em que a hipótese nula é ausência de correlação linear na população. O teste é compatível com o tipo de correlação escolhido.⁹

Figura 3.12 – Esquema para obter uma matriz de correlações.



⁹ Para testes mais específicos você pode fazer: *Estatísticas >> Resumos >> Teste de correlação*.

Com esses procedimentos, o software apresenta os resultados mostrados na Figura 3.13.¹⁰

Figura 3.13 – Matriz de correlações das notas das provas das três áreas do Enem.

```
Output
> options(OutDec= ",", digits=3)
> cor(amostraEnem2019[,c("NOTA_CH", "NOTA_CN",
"NOTA_LC", "NOTA_MT")], use="complete")
      NOTA_CH  NOTA_CN  NOTA_LC  NOTA_MT
NOTA_CH  1,000    0,688    0,757    0,616
NOTA_CN  0,688    1,000    0,662    0,646
NOTA_LC  0,757    0,662    1,000    0,607
NOTA_MT  0,616    0,646    0,607    1,000
```

3.7 – Tabela de contingência

Para obter uma tabela de contingência entre duas variáveis qualitativas, fazer:

Estatísticas >> Tabelas de contingência >> Tabela de dupla entrada

Na aba *Dados* você escolhe a variável (*fator*) cujas categorias devem ficar nas linhas da tabela e a variável (*fator*) cujas categorias devem ficar nas colunas. Na aba *Opções* você pode optar por querer também tabela com porcentagens, além de poder pedir um teste estatístico de associação.

Como exemplo, solicitamos uma tabela de contingência entre *COR_RACA_recod* (linha) e *TIPO_DE_ESCOLA* (coluna), incluindo

¹⁰ Antes de proceder os procedimentos para realizar a matriz de correlação, submetemos o comando `options(OutDec= ",", digits=3)`, por isto os resultados foram apresentados com vírgula para a separação das decimais e com três dígitos significativos.

porcentagens nas colunas e teste de associação qui-quadrado.¹¹ A Figura 3.14 mostra os resultados.

Figura 3.14 – Tabelas de frequências cruzadas entre cor ou raça declarada e tipo de escola.

```
Output
Frequency table:
COR_RACA_recod      TIPO_ESCOLA
                    Não respondeu Pública Privada
Branca ou Amarela   505      207     92
Parda                599      274     31
Preta ou Indígena   181       63      6
Não declarado        26       14      2

Column percentages:
COR_RACA_recod      TIPO_ESCOLA
                    Não respondeu Pública Privada
Branca ou Amarela   38,5     37,1    70,2
Parda                45,7     49,1    23,7
Preta ou Indígena   13,8     11,3     4,6
Não declarado        2,0      2,5     1,5
Total                100,0    100,0   100,0
Count               1311,0   558,0   131,0
```

Neste exemplo, na janela de mensagens do *Rcmdr* aparece um alerta de que o teste pode não ser válido, porque há frequência esperada pequena (menor que cinco).

¹¹ A variável *COR_RACA_recod* foi criada através de recodificação da variável *COR_RACA* do arquivo *amostraEnem2019.RData*, conforme descrito na seção 3.1. Fizemos, também, uma reordenação das categorias da variável *TIPO_ESCOLA*. Lembramos que no Capítulo 2 tem seções específicas descrevendo sobre recodificação e ordenação de níveis de fatores.

4 – Distribuições de probabilidade

No *menu* principal, aba *Distribuições*, há vários recursos disponíveis para as principais distribuições de probabilidades. De maneira geral, para as distribuições discretas, tem-se as seguintes possibilidades:

- Quantis (valor mínimo de x , tal que $P(X \leq x) \geq p$, sendo $0 \leq p \leq 1$ um argumento dado)¹² ;
- Probabilidades das caudas ($P(X \leq x)$ ou $P(X > x)$, dependendo da escolha entre cauda inferior ou superior, sendo x um argumento dado);
- Função de probabilidade (os pares $(x, p(x))$);
- Gráfico da distribuição, opcionalmente da função de probabilidade ou da função de distribuição acumulada;
- Amostras de observações geradas segundo a distribuição (pode-se, também, pedir para calcular alguma estatística para cada amostra).

Para funções contínuas, tem-se as opções:

- Quantis (valor de x , tal que $P(X \leq x) = p$, sendo $0 \leq p \leq 1$ um argumento dado);
- Probabilidades das caudas ($P(X \leq x)$ ou $P(X > x)$ dependendo da escolha entre cauda inferior ou superior, sendo x um argumento dado);
- Gráfico da distribuição, podendo ser da função de densidade ou da função de distribuição acumulada;

¹² Quando fornecer algum valor fracionário, o símbolo de decimal deve ser “.” (ponto).

- Amostras de observações geradas segundo a distribuição (opcionalmente, pode-se pedir para calcular alguma estatística em cada amostra).

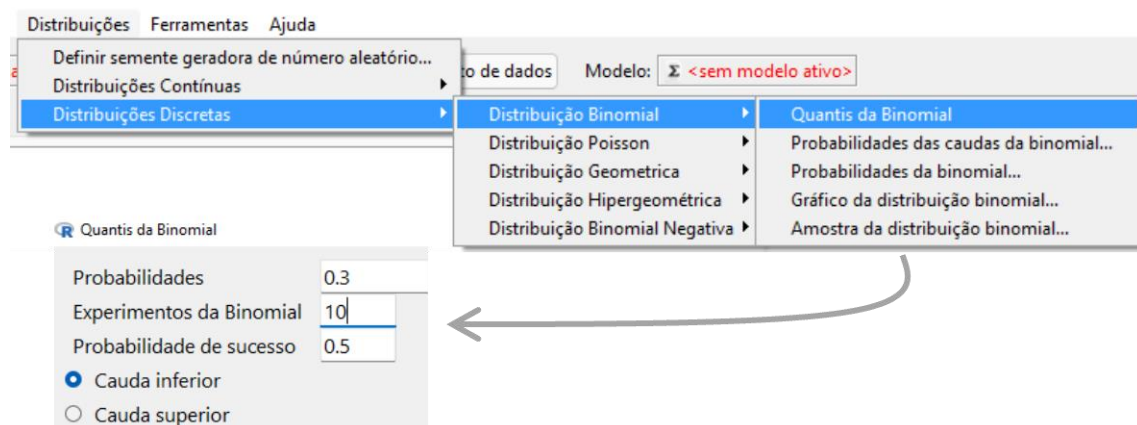
Exemplificaremos, no caso discreto, com exemplos da distribuição binomial e, no caso contínuo, com exemplos usando a distribuição normal.

4.1 – Distribuição binomial

➤ *Quantis*

Ex. Sendo X tendo distribuição binomial com $n = 10$ e $p = \frac{1}{2}$, qual é o valor mínimo de x , tal que $P(X \leq x) \geq 0,3$? A Figura 4.1 mostra como obter esse valor no *Rcmdr*.

Figura 4.1 – Esquema para obter o quantil 0,3 na distribuição binomial de $n = 10$ e $p = 0,5$.



Nota: Repare que usamos *ponto* para representar os números com decimais.

Como resposta o *Rcmdr* fornece o valor 4 (quatro).

➤ *Probabilidades nas caudas*

Como exemplo, vamos considerar X tendo distribuição binomial com $n = 10$ e $p = 0,5$ e vamos calcular $P(X \leq 4)$, ou seja, a probabilidade de se obter um número menor ou igual a quatro. Para resolver esse problema no *Rcmdr*, fazer:

*Distribuições >> Distribuições discretas >> Distribuição binomial >>
Probabilidades das caudas da binomial*

Na janela que se abre, entrar com os valores:

- Valores da variável: 4
- Experimentos da binomial: 10
- Probabilidade de sucessos: 0.5 e
- Marcar *Cauda inferior*.

Como resultado, o *Rcmdr* fornece a probabilidade: 0.3769531.

➤ *Função de probabilidade*

Como exemplo, vamos apresentar a função de probabilidade da binomial com $n = 10$ e $p = 0,5$.

Usando os *menus* do *Rcmdr*, fazer:

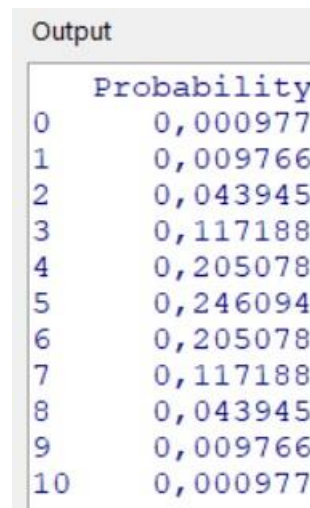
*Distribuições >> Distribuições discretas >> Distribuição binomial >>
Probabilidades da binomial*

Na janela que se abre, preencher:

- Experimentos da binomial: 10
- Probabilidade de sucessos: 0.5

O *Rcmdr* fornece os resultados apresentados na Figura 4.2.

Figura 4.2 – Distribuição de probabilidades de uma binomial com $n = 10$ e $p = 0,5$.



	Probability
0	0,000977
1	0,009766
2	0,043945
3	0,117188
4	0,205078
5	0,246094
6	0,205078
7	0,117188
8	0,043945
9	0,009766
10	0,000977

➤ *Gráfico da distribuição binomial*

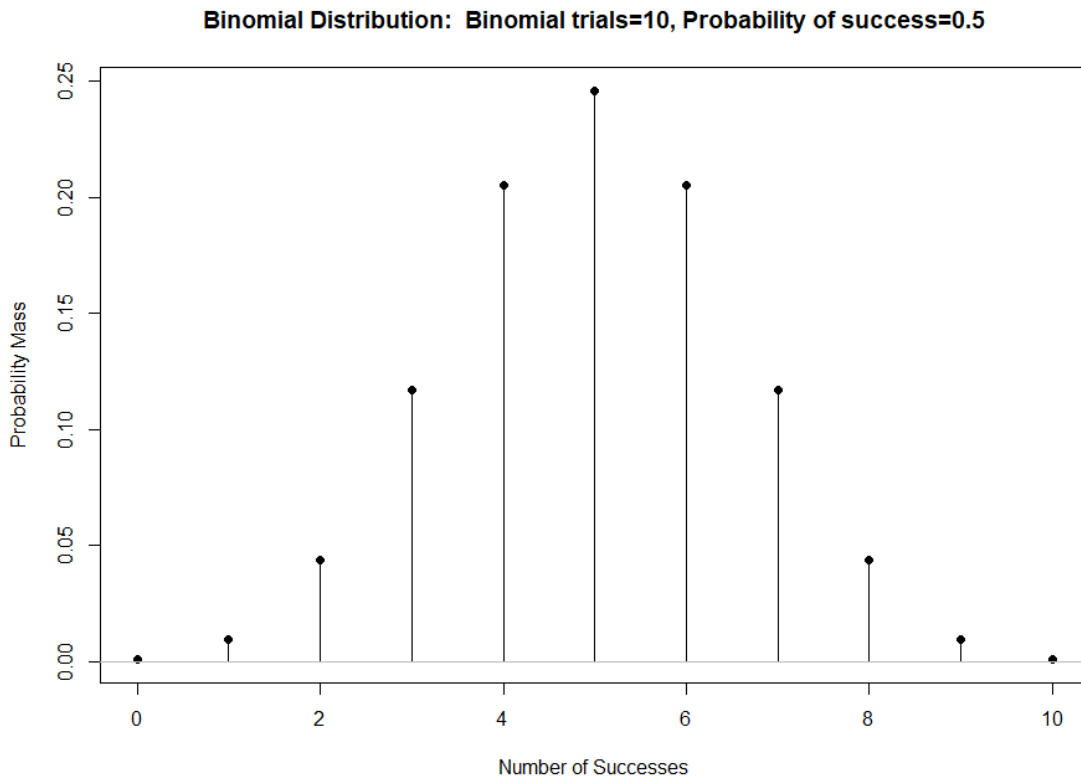
Vamos apresentar a função de probabilidade da binomial com $n = 10$ e $p = 0,5$ em forma gráfica. Podemos obter esse gráfico pelo *Rcmdr*, fazendo:

*Distribuições >> Distribuições discretas >> Distribuição binomial >>
Gráfico da distribuição binomial*

Na janela aberta pelo software, fornecer os parâmetros da distribuição e escolher *Gráfico da função de massa* ou *Gráfico da função cumulativa*. Se o

objetivo é a função de probabilidade, devemos marcar a primeira opção. O *Rcmdr* apresentará o gráfico mostrado na Figura 4.3.

Figura 4.3 – Função de probabilidades da distribuição binomial com $n = 10$ e $p = 0,5$.



➤ *Amostra da distribuição binomial*

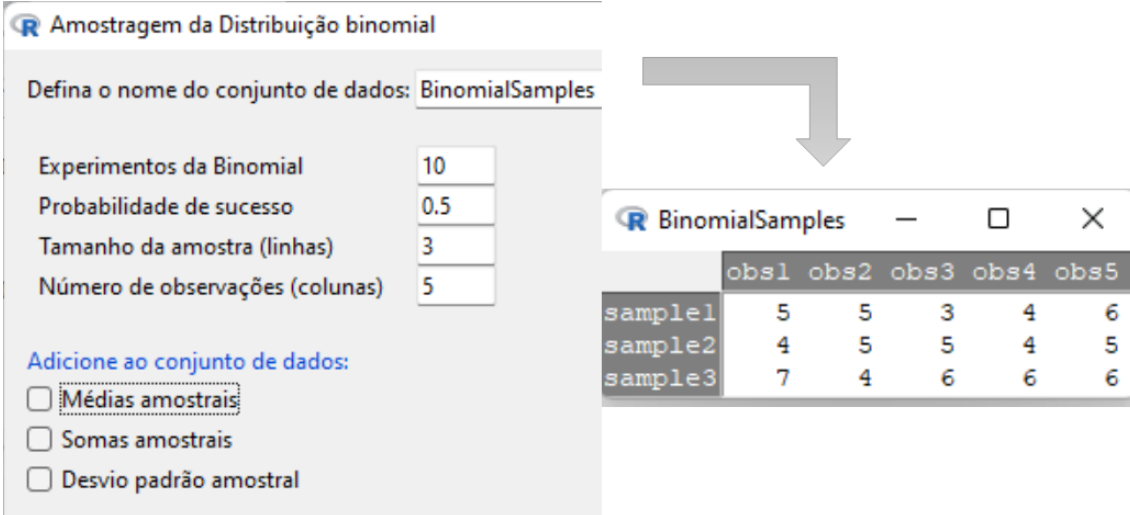
A geração de uma ou mais amostras de uma distribuição binomial pode ser feita com a sequência:

Distribuições >> Distribuições discretas >> Distribuição binomial >> Amostra da distribuição binomial

O *Rcmdr* abre uma janela, conforme mostrado do lado esquerdo da Figura 4.4. No esquema dessa figura, optamos por fazer um arquivo contendo três amostras de tamanho cinco. Observar que colocamos os parâmetros da

distribuição binomial ($n = 10$ e $p = 0,5$), a quantidade de amostras (linhas) e o tamanho dessas amostras (colunas), podendo, ainda, incluir uma coluna com alguma medida descritiva (p. ex., a média) dessas amostras. O lado direito da Figura 4.4 mostra o arquivo criado pelo software (*BinomialSamples*) contendo as amostras.

Figura 4.4 – Geração de três amostras de tamanho cinco de uma binomial com $n = 10$ e $p = 0,5$.



The screenshot shows the R software interface for generating binomial samples. The window title is "Amostragem da Distribuição binomial". The "Defina o nome do conjunto de dados:" field is set to "BinomialSamples". The parameters are: "Experimentos da Binomial" (10), "Probabilidade de sucesso" (0.5), "Tamanho da amostra (linhas)" (3), and "Número de observações (colunas)" (5). Under "Adicione ao conjunto de dados:", the "Médias amostrais" checkbox is checked. An arrow points from the software window to a preview of the resulting data table, titled "BinomialSamples".

	obs1	obs2	obs3	obs4	obs5
sample1	5	5	3	4	6
sample2	4	5	5	4	5
sample3	7	4	6	6	6

Observar que pela notação do software, cada linha corresponde a uma amostra, mas como a geração dos dados é feita de forma independente, você também pode considerar ao contrário, ou seja, cinco amostras de tamanho três.

4.2 – Distribuição normal

➤ *Quantis*

Fazendo a sequência:

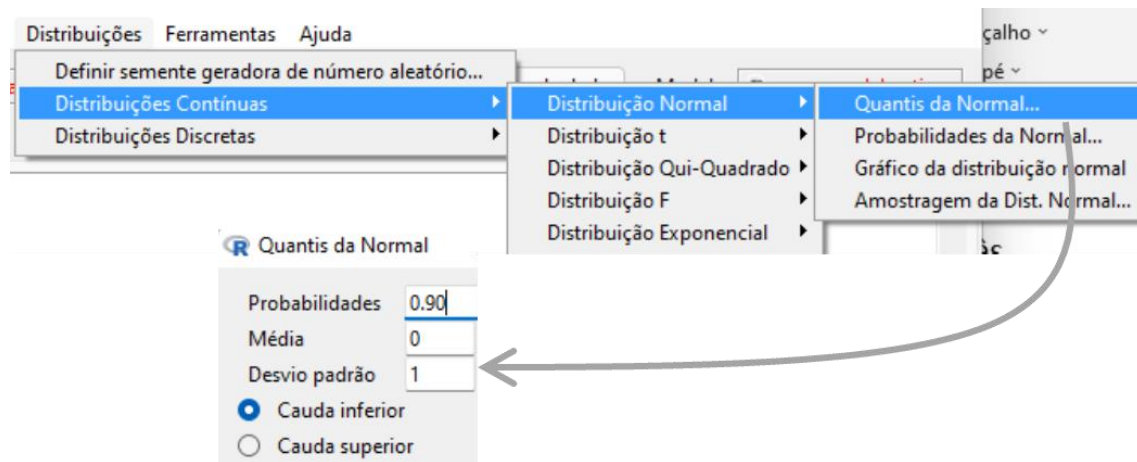
Distribuições >> Distribuições contínuas >> Distribuição normal >>

Quantis da normal

podemos obter o valor x (quantil x) que delimita uma dada área igual a p na cauda inferior (ou uma área igual a $1 - p$ na cauda superior) de uma distribuição normal, ou seja, a inversa da função de distribuição acumulada de uma normal.

A Figura 4.5 mostra o esquema para se obter o valor z da distribuição normal padrão, que deixa uma área de 0,90 na cauda inferior, ou seja, $z = F_z^{-1}(0,90)$.

Figura 4.5 – Esquema para obter o quantil 0,90 na distribuição normal de $\mu = 0$ e $\sigma = 1$.



Como resposta, o *Rcmdr* fornece o quantil 1,281552.

➤ *Probabilidades nas caudas*

Podemos obter a área (ou a probabilidade) numa das caudas de uma distribuição normal, fazendo:

*Distribuições >> Distribuições contínuas >> Distribuição normal >>
Probabilidades da normal*

Para fazer a operação inversa da seção anterior, isto é, $p = F_z(1,281552)$, sendo F_z a função de distribuição acumulada normal padrão, entramos com os seguintes argumentos na janela aberta pelo *Rcmdr*:

- Valores da variável: 1,281552
- Média: 0
- Desvio padrão: 1
- Marcar *Cauda Inferior*.

Com essas entradas, o software fornece a resposta: $p = 0,90$

➤ *Gráfico da distribuição normal*

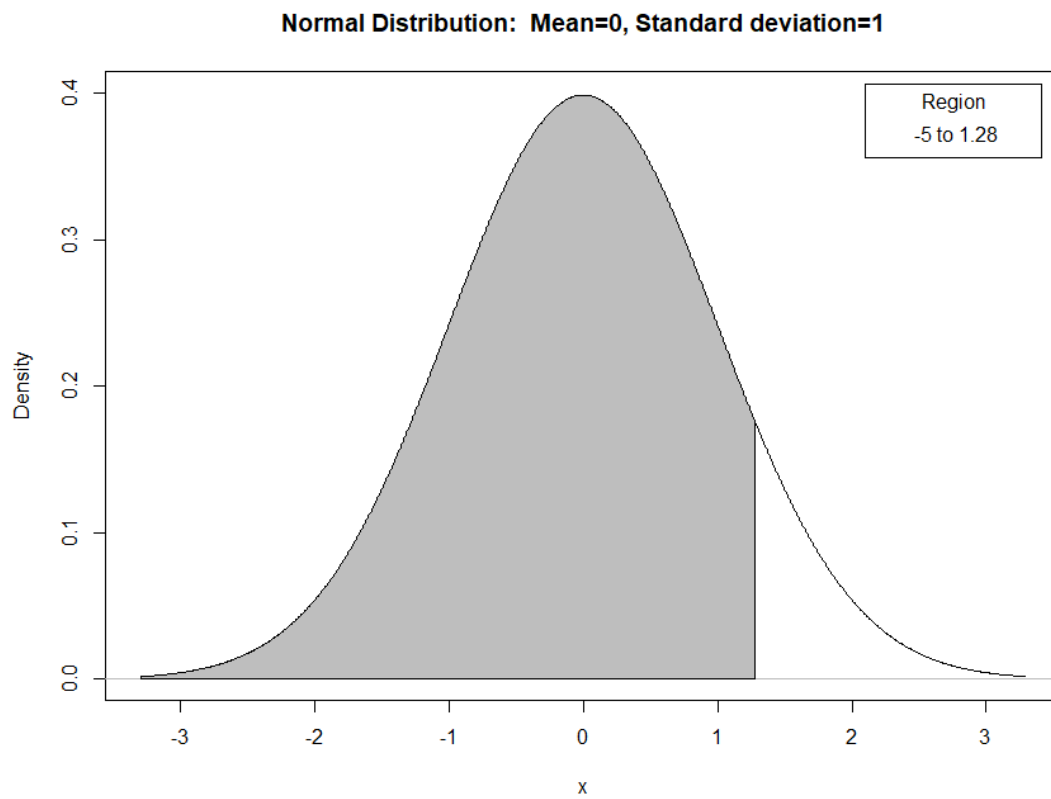
O Gráfico da função de densidade ou da função de distribuição acumulada de uma normal pode ser obtido com a sequência:

*Distribuições >> Distribuições contínuas >> Distribuição normal >>
Gráfico da distribuição normal*

A Figura 4.6 mostra o resultado após as seguintes entradas:

- parâmetros da distribuição normal padrão (média = 0 e desvio padrão = 1);
- escolha da função de densidade; e
- destaque para a região de -5 a 1,28.

Figura 4.6 – Gráfico da distribuição normal padrão e destaque para $P(Z < 1,28)$.



➤ *Amostra da distribuição normal*

A geração de uma ou mais amostras de uma distribuição normal pode ser feita no *Rcmdr* com a sequência:

Distribuições >> Distribuições contínuas >> Distribuição normal >> Amostragem da distribuição normal

O *Rcmdr* abre uma janela em que você deve especificar os parâmetros da distribuição normal (média e desvio padrão), o número de linhas e o número de colunas do arquivo com valores gerados pela distribuição normal. Pelas denominações das linhas e colunas desse novo arquivo de dados, o *Rcmdr*

considera cada linha como uma amostra. Opcionalmente, você pode pedir uma coluna com alguma medida descritiva (p. ex., média, desvio padrão, ...) dessas amostras.

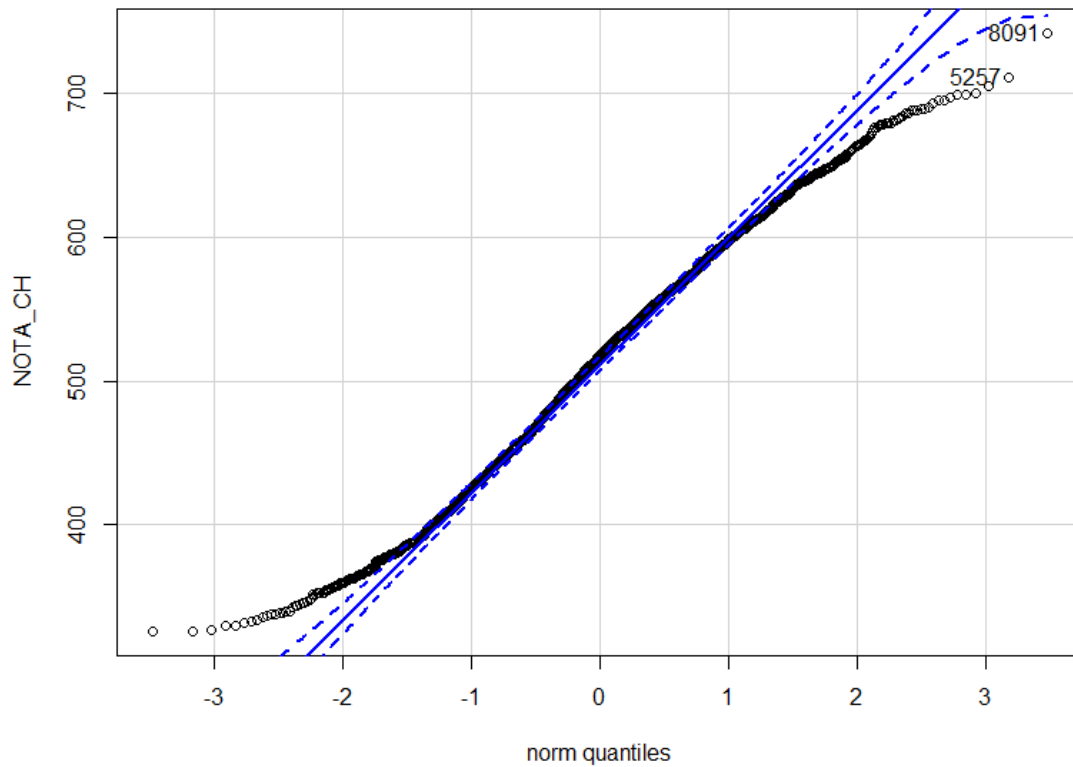
4.3 – Gráfico de probabilidade normal

Uma forma de avaliar se é razoável supor que uma variável aleatória tem distribuição normal, com base numa amostra de observações, é através do gráfico de probabilidade normal. O *Rcmdr* apresenta a opção de *gráfico de comparação de quantis*, que é uma abordagem mais geral que o gráfico de probabilidade normal. Esse gráfico pode ser realizado pela sequência:

Gráficos >> Gráfico de comparação de quantis

Na janela que o software abre, aba *Dados*, você seleciona a variável; na aba *Opções* você pode escolher outros modelos além da normal, a forma de identificação de valores discrepantes e os rótulos nos eixos. Realizando esses procedimentos com a variável *NOTAS_CH* e não alterando as opções sugeridas, obtêm-se o gráfico mostrado na Figura 4.7.

Figura 4.7 – Gráfico de comparação de quantis de *NOTA_CH* com valores teóricos da normal padrão.



O gráfico da Figura 4.7 mostra os pontos na faixa central bem alinhados, mas os pontos das caudas distanciam desse comportamento linear, cuja forma sugere uma distribuição mais achatada que a normal.

5 – Intervalos de confiança e testes de hipóteses para médias

Para análise de uma média ou de diferença de médias, o *Rcmdr* oferece funções que geram resultados em termos de intervalos de confiança e de testes de hipóteses.

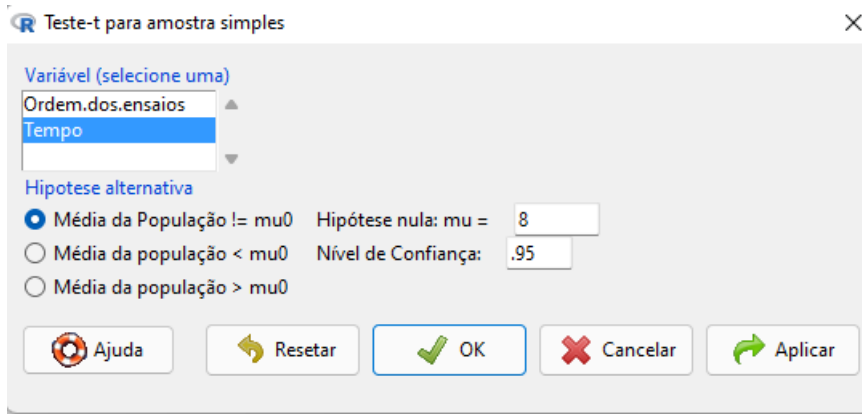
5.1 – Intervalo de confiança e teste para uma média

Tomemos como base o arquivo de dados do Exemplo 9.6, em que se avalia o tempo de resposta em função do tipo de rede de computadores. Inicialmente, consideraremos somente o tipo de rede C1, considerando o problema de se verificar, com base na amostra aleatória de oito observações, se o tempo de resposta se mantém no padrão estabelecido de oito segundos, em média. Depois de importar os dados, fazer através dos *menus* do *Rcmdr*:

Estatísticas >> Médias >> Teste t para uma média

Na janela que se abre, escolhemos a variável *Tempo* e especificamos $H_0: \mu = 8$ e $H_1: \mu \neq 8$ (teste bilateral) e nível de confiança de 0,95 (nível de significância = 0,05), conforme mostra a Figura 5.1.

Figura 5.1 – Janela para especificação de teste de uma média



Apertando *OK*, o software apresenta os resultados como mostrado na Figura 5.2.

Figura 5.2 – Resultados de um teste t para uma média

```
Output
One Sample t-test

data: Tempo
t = 0,73, df = 7, p-value = 0,5
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7,527 8,898
sample estimates:
mean of x
 8,213
```

5.2 – Intervalo de confiança e teste para duas amostras pareadas

Nesta seção usaremos o mesmo arquivo de dados do livro (Exemplo 9.2), cujos dados estão disponíveis em Materiais Suplementares. Para este teste os dados devem estar organizados em duas colunas, associadas às duas variáveis, sendo que nas linhas temos os pares de observações. Usar o *menu* do *Rcmdr*:

Estatísticas >> Médias >> Teste t (dados pareados)

Você deve escolher as duas variáveis pareadas. O software considera a diferença $D = X_1 - X_2$, sendo X_1 a primeira variável escolhida e X_2 a segunda. Seguindo o problema descrito no livro, devemos fazer um teste unilateral à direita ($H_0: \mu_D = 0$ e $H_1: \mu_D > 0$), conforme especificado na Figura 5.3.

Figura 5.3 – Formato dos dados e janelas para especificação de um teste t para dados pareados

Formato dos dados →

Ensaio	X_1	X_2
1	22	25
2	21	28
3	28	26
...

The figure shows two screenshots of the 'Teste-t pareado' software interface. The left screenshot shows the 'Dados' tab with 'Ensaio' selected for both the first and second variables. The right screenshot shows the 'Opções' tab with 'Diferença > 0' selected for the alternative hypothesis and a confidence level of .95.

Vom essas especificações, o software apresenta os resultados mostrados na Figura 5.4.

Figura 5.4 – Resultados de um teste t para dados pareados

```
Output Submeter  
  
Paired t-test  
  
data: X2 and X1  
t = 2,8, df = 9, p-value = 0,01  
alternative hypothesis: true mean difference is greater than 0  
95 percent confidence interval:  
 1,193 Inf  
sample estimates:  
mean difference  
      3,4
```

As três primeiras linhas referem ao teste estatístico, que em função de valor- $p < 0,05$, o teste rejeita H_0 , em favor de H_1 , ao nível de significância de 0,05. A seguir, tem-se o intervalo de confiança para μ_D , que devido a escolha unilateral da hipótese nula, o intervalo só apresenta valor finito do lado esquerdo. Finalmente, tem-se a diferença das médias das amostras, igual a 3,4.

5.3 – Intervalo de confiança e teste para duas amostras independentes

Para duas amostras independentes, usar a sequência de *menus*:

Estatísticas >> Médias >> Teste t para amostras independentes

Voltaremos aos dados do Exemplo 9.6 e vamos testar se há evidência de que os tipos de rede C1 e C2 apresentam tempos esperados de resposta diferentes, considerando o nível de significância de 0,05. Sendo μ_{C1} o tempo de resposta esperado pela rede C1 e μ_{C2} o tempo de resposta esperado pela rede C2, as hipóteses são:

$$H_0: \mu_{C1} = \mu_{C2}$$

$$H_1: \mu_{C1} \neq \mu_{C2}$$

No caso de amostras independentes, as observações da variável resposta dos dois grupos deve estar numa mesma coluna, sendo que a identificação do grupo deve estar em outra coluna, como mostra o lado esquerdo da Figura

5.5. Observe que marcamos a opção de assumir variâncias iguais para ficar compatível com o teste discutido no livro. A Figura 5.6 mostra os resultados

Figura 5.5 – Formato dos dados e janelas para especificação de um teste t para amostras independentes.

Ordem dos ensaios	Tipo de rede	Tempo
1	C1	9,3
3	C1	8,8
4	C1	8,9
7	C1	7,2
9	C1	8,7
10	C1	7,6
15	C1	8,0
19	C1	7,2
2	C2	8,2
5	C2	7,1
8	C2	8,6
12	C2	7,8
16	C2	7,8
17	C2	7,1
20	C2	8,2
23	C2	8,7

Figura 5.6 – Resultados de um teste t para duas amostras independentes.

```

Output
Two Sample t-test

data:  Tempo by Tipo.de.rede
t = 0,76, df = 14, p-value = 0,5
alternative hypothesis: true difference in means between
group C1 and group C2 is not equal to 0
95 percent confidence interval:
-0,4998  1,0498
sample estimates:
mean in group C1 mean in group C2
      8,213          7,938
  
```

Observem que nas primeiras linhas são apresentados os resultados do teste t, os quais levam à aceitação da hipótese nula; e, depois, são apresentados o

intervalo de confiança para a diferença das médias populacionais e os valores das médias amostrais.

5.4 – Análise de variância com um fator

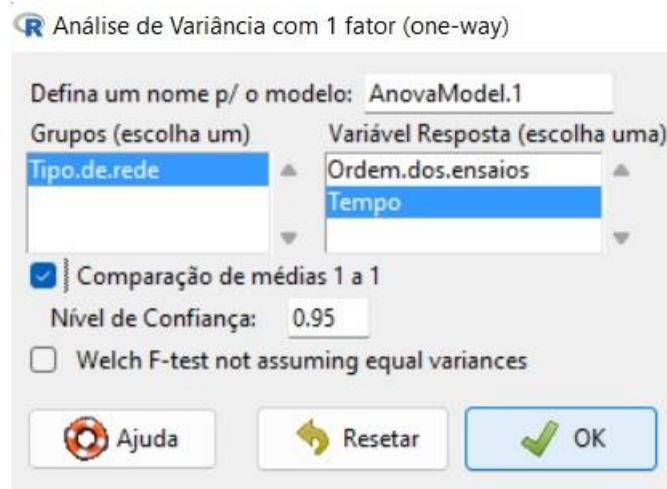
A análise de variância com um fator pode ser feita com a sequência nos *menus* do *Rcmdr*:

Estatísticas >> Médias >> ANOVA para um fator (one way)

Voltando aos dados do Exemplo 9.6, em que se avalia o tempo de resposta em função de três tipos de rede de computadores, a realização da análise de variância pode ser feita para testar a hipótese de que o tempo esperado de resposta é igual para os três tipos de rede de computadores. Tendo o arquivo Exemplo_9.6 ativo no *Rcmdr* e fazendo a sequência nos menus como indicada anteriormente, o *R Commander* apresenta a janela mostrada na Figura 5.7, onde você escolhe a variável resposta e a variável (ou fator) que define os grupos. Nessa janela você também pode pedir *Comparação de médias 1 a 1*, fazendo com que o software faça um teste de igualdade de médias entre todos os pares, procedimento conhecido como *comparações múltiplas*.¹³

¹³ As comparações múltiplas só devem ser consideradas se o teste F rejeita a hipótese nula de igualdade entre todas as médias.

Figura 5.7 – Janela da análise de variância com um fator.



Como resultado desses procedimentos vem, primeiramente, o teste F de igualdade entre as médias populacionais (Figura 5.8).

Figura 5.8 –Resultado do teste F de comparação de várias médias.

```
Output
```

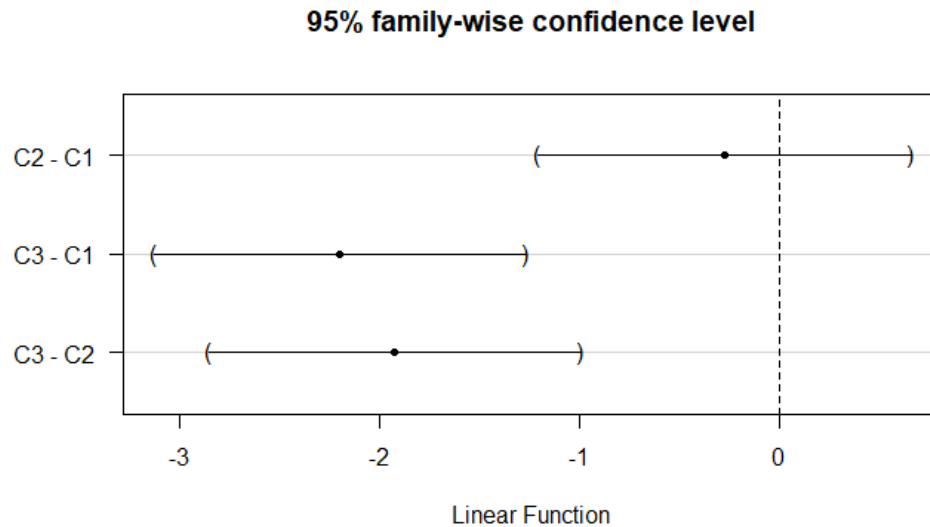
```
> AnovaModel.1 <- aov(Tempo ~ Tipo.de.rede, data=d)
> summary(AnovaModel.1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tipo.de.rede	2	23,0	11,50	21,1	0,0000095 ***
Residuals	21	11,5	0,55		

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1

A tabela da ANOVA mostra que o teste rejeita H_0 , já que o *valor-p* ($Pr(>F)$) é menor que 0,05. Como foi marcada a opção de comparações múltiplas (Comparação de médias 1 a 1), o software mostra resultados analíticos e gráficos dessas comparações. A Figura 5.9 mostra que os intervalos de 95% de confiança envolvendo a rede tipo C3 não contém o zero, ou seja, a rede tipo C3 é diferente, significativamente (nível de significância de 0,05), dos tipos C1 e C2.

Figura 5.9 – Comparações múltiplas com os dados do Exemplo_9.6

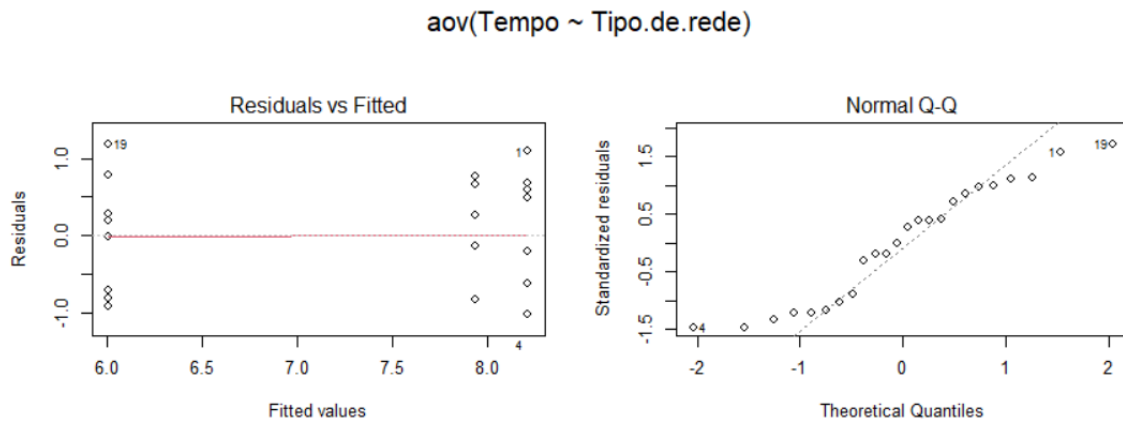


A análise de variância tem algumas suposições para a validade do teste F, que podem ser avaliadas com a análise de resíduos. Ao realizar a ANOVA, o *Rcmdr* ativa a caixa *Modelo*: (à esquerda e logo abaixo do menu principal), com o nome do modelo em azul: *AnovaModel.1* (ou o nome que você deu na janela da ANOVA). Iniciando-se pelo *menu* principal e fazendo a sequência:

Modelos >> Gráficos >> Diagnósticos Gráficos Básicos

O software apresenta um painel com quatro gráficos. A Figura 5.10 mostra um recorte dos dois primeiros: gráfico de *preditos x resíduos* e gráfico de *probabilidade normal*, que são discutidos no livro.

Figura 5.10 – Gráficos de diagnósticos do modelo com base nos resíduos da ANOVA.



5.5 – Análise de variância para projeto fatorial

Nesta seção usaremos o arquivo de dados: *Exemplo_9.9*, cujo problema consiste na análise do tempo de resposta em função da topologia e do protocolo usada na rede de computadores. A análise de variância pode ser realizada pela sequência:¹⁴

Estatísticas >> Médias >> ANOVA multifator (Multi-way)

Com esses procedimentos, o *Rcmdr* apresenta os resultados, como mostrado na Figura 5.11.

¹⁴ Antes de realizar esse procedimento é necessário converter as variáveis Protocolo e Topologia para *fator*, já que essas variáveis estão codificadas com números e o software as considera como variáveis numéricas na importação dos dados. A seção 2.8 descreve como converter variável numérica para fator.

Figura 5.11 – Resultados de testes F em um projeto fatorial.

```
Output
Response: Tempo.de.resposta
          Sum Sq Df F value Pr(>F)
Protocolo      7,15  1    7,51  0,013 *
Topologia     10,36  2    5,44  0,014 *
Protocolo:Topologia  0,26  2    0,14  0,872
Residuals     17,14 18
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05
```

Ao realizar essa análise, o *Rcmdr* ativa o modelo: *Anova.Model.2*, escrito em azul, parte de cima e à direita do painel do software. Tendo o modelo ativo, você pode fazer uma análise de resíduos para diagnóstico do modelo, conforme visto em ANOVA de um fator (seção anterior).

5.6 – Abordagem não paramétrica

Os testes das seções anteriores estão associados às suposições de distribuição normal e variância constante. Quando a análise de resíduos sugerir que essas suposições não podem ser aceitas, uma alternativa são os testes não paramétricos baseados na ordenação das observações. Com o conjunto de dados ativo, inicia-se pelo menu principal, fazendo:

Estatísticas >> Testes não paramétricos

As janelas que se abrem são parecidas com a abordagem paramétrica vista anteriormente, donde acreditamos que o leitor não terá dificuldade em realizar esses testes.

6 – Modelos de regressão

A análise de regressão linear (simples ou múltipla) através do *R Commander* é realizada através dos menus por:

Estatísticas >> Ajuste de modelos >> Regressão linear ou

Estatísticas >> Ajuste de modelos >> Modelo linear

A visualização gráfica de uma regressão linear simples pode ser feita por:

Gráficos >> Diagrama de dispersão

incluindo, na aba *Opções*, marcação na caixa de *Linha de mínimos quadrados*.

6.1 – Regressão linear simples: gráfico

Neste Capítulo, vamos considerar uma parte do arquivo *amostraEnem2019* com os 689 candidatos que têm informação sobre o tipo de escola que estudaram o ensino médio (pública ou privada).¹⁵ Vamos fazer uma regressão com *NOTA_CN* como variável dependente e *NOTA_MT* como variável independente, considerando que o aprendizado de Ciências Naturais depende, em parte, da proficiência em Matemática.

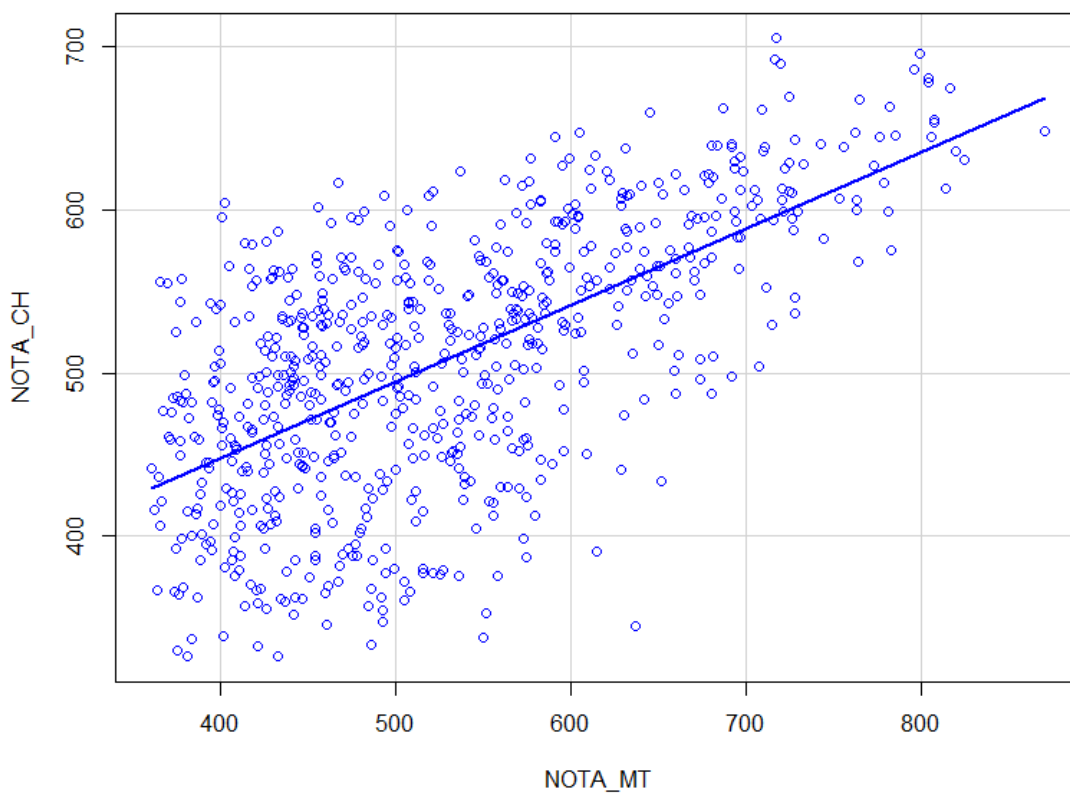
Para visualizarmos graficamente os dados e o ajuste de uma relação linear, podemos fazer:

¹⁵ Na seção 2.4 foi mostrado como obter um subconjunto de dados.

Gráficos >> Diagrama de dispersão

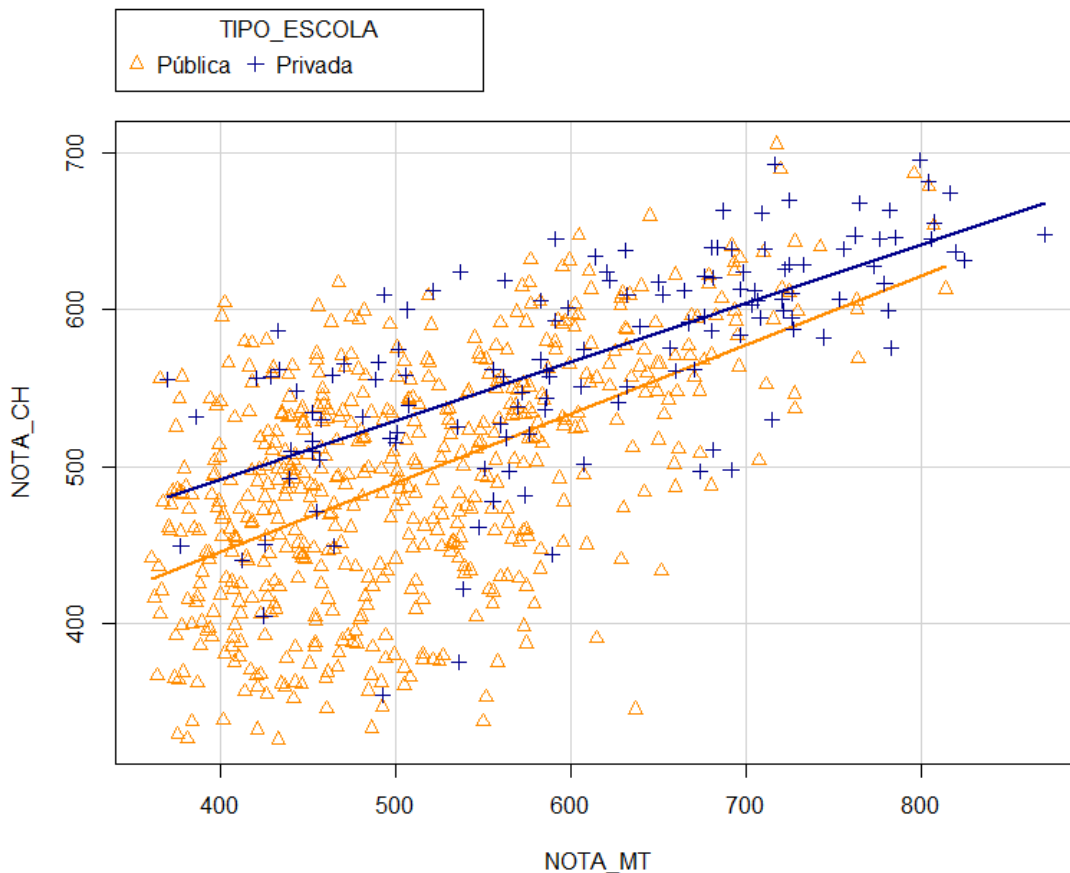
Na aba Dados, marcamos *NOTA_MT* como variável-x e *NOTA_CN* como variável-y. Na aba *Opções*, marcamos *Linha de mínimos quadrados*. Ao clicar *OK*, o *R Commander* apresenta o gráfico mostrado na Figura 6.1.

Figura 6.1 –Ajuste de uma relação linear da *NOTA_CN* em função da *NOTA_MT*.



É possível colocar num mesmo gráfico duas ou mais regressões em termos de algum fator do arquivo de dados. Como exemplo, vamos considerar as mesmas variáveis, mas separando as regressões por tipo de escola, ou seja, uma regressão para cada um dos subconjuntos de dados acerca do tipo de escola (pública ou privada).

Figura 6.2 – Ajuste de uma relação linear da *NOTA_CN* em função da *NOTA_MT*, por tipo de escola (pública ou privada).



6.2 – Regressão linear simples: resumo analítico

Um resumo de uma análise de regressão pode ser obtido por:

Estatísticas >> Ajuste de modelos >> Regressão linear

No topo da janela aberta pelo software, você pode dar um nome ao modelo (por exemplo, *reg1*); na lista de variáveis numéricas apresentada do lado esquerdo da janela, marcar como variável resposta a *NOTA_CN*. Na lista de variáveis numéricas do lado direito da janela, marcar como variável explicativa a *NOTA_MT*. Ao apertar *OK*, o *R Commander* apresenta os principais resultados da regressão, como mostrado na Figura 6.3.

Figura 6.3 – Resultados de uma regressão linear da *NOTA_CN* em função da *NOTA_MT*.

```
Output
> reg1 <- lm(NOTA_CN~NOTA_MT, data=parte_enem)
> summary(reg1)

Call:
lm(formula = NOTA_CN ~ NOTA_MT, data = parte_enem)

Residuals:
    Min       1Q   Median       3Q      Max
-153,16  -42,75    0,72   42,17  180,36

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 239,8059    11,0416    21,7  <2e-16 ***
NOTA_MT      0,4523     0,0204    22,1  <2e-16 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1

Residual standard error: 58 on 687 degrees of freedom
Multiple R-squared:  0,417, Adjusted R-squared:  0,416
F-statistic: 491 on 1 and 687 DF,  p-value: <2e-16
```

Conforme mostra a Figura 6.3, o *R Commander* mostra, inicialmente, uma análise descritiva básica dos resíduos provenientes do ajuste do modelo de regressão. A seguir, os coeficientes da equação, que definem a equação da reta:

$$NOTA_CN = 239,806 + 0,452 \times NOTA_MT$$

Nas mesmas linhas dessa saída computacional, também são apresentadas estimativas dos erros padrões e os dois testes t bilaterais referentes às hipóteses nulas $H_0: \beta_0 = 0$ para o intercepto; e $H_0: \beta_1 = 0$ para o coeficiente angular.

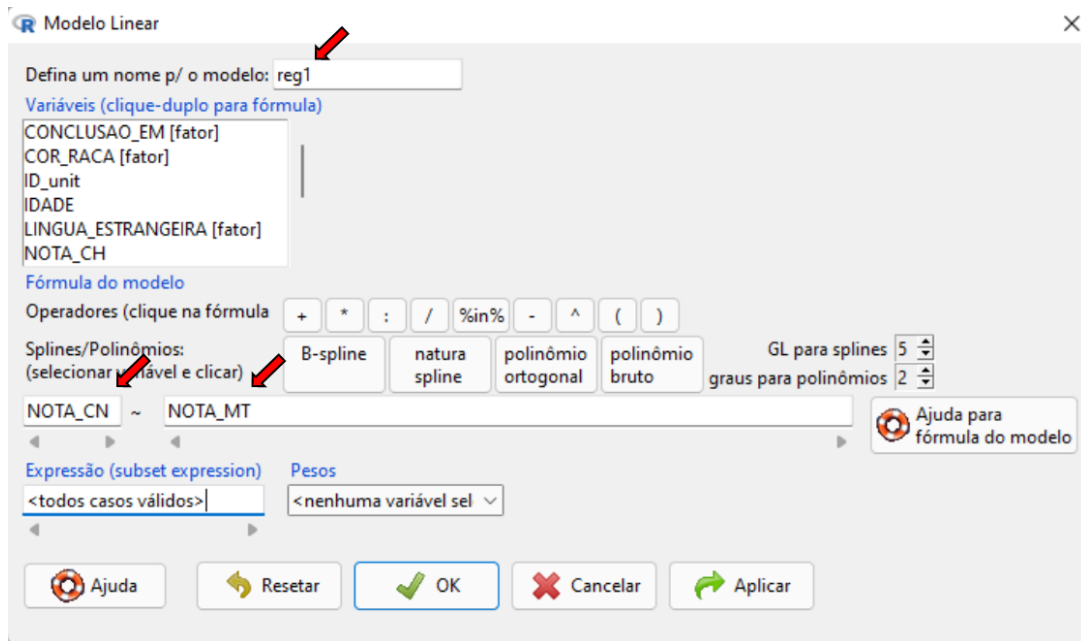
Nas três últimas linhas são apresentadas estimativa do erro padrão do erro, acompanhada dos graus de liberdade; R^2 e R_{aj}^2 ; e o teste F de significância do modelo.

Outra alternativa para fazer a mesma regressão é usando a seguinte sequência:

Estatísticas >> Ajuste de modelos >> Modelo linear

Na janela aberta pelo *Rcmdr*, preencher como é mostrado na Figura 6.4. Esse processo gera código computacional igual ao processo realizado pelo *menu* de *Regressão linear* e, portanto, a mesma saída de resultados. É um pouco mais complexo, mas é bastante útil para modelos mais gerais, como aqueles que incluem variável independente tipo fator, como veremos na próxima seção.

Figura 6.4 – Preenchimento da janela de *Modelo linear* para uma regressão linear simples.



6.3 – Regressão linear múltipla

Para uma regressão usual em que todas as variáveis são quantitativas, você pode usar o processo mais simples, ou seja:¹⁶

Estatísticas >> Ajuste de modelos >> Regressão linear

Por exemplo, se quiser uma regressão para *explicar* a variação das notas de Ciências da Natureza (*NOTAS_CN*) em função das notas em Matemática (*NOTAS_MT*) e nível socioeconômico (*NSE*), que está numa escala numérica de zero a dez, você pode marcar as duas variáveis independentes segurando apertada a tecla *Ctrl* (ver janela à esquerda da Figura 6.4)

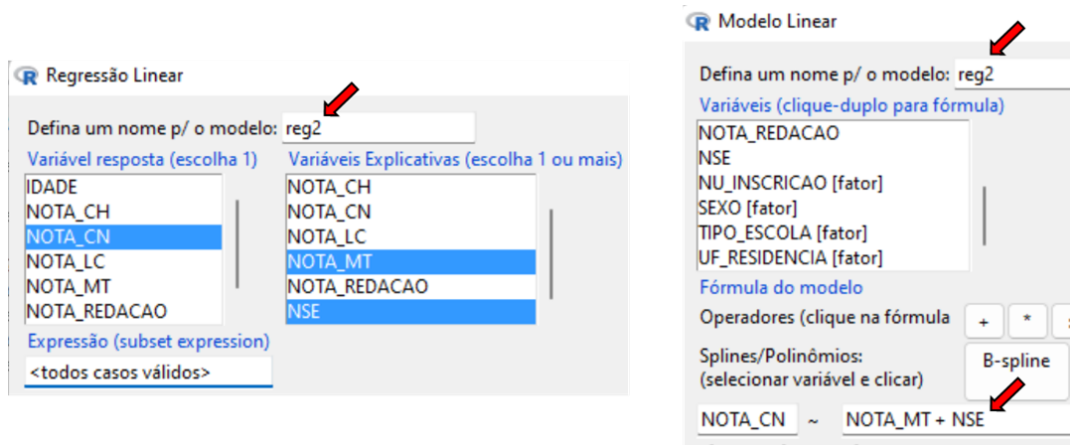
A mesma análise pode ser feita por:

Estatísticas >> Ajuste de modelos >> Modelo linear

Neste caso, as variáveis podem ser inseridas dando dois *clicks* sobre elas ou, ainda, escrevendo-as nas caixas, sendo que as variáveis independentes devem estar separadas com o sinal de adição (+), como mostra o lado direito da Figura 6.5.

¹⁶ Também pode usar essa forma se tiver variáveis independentes do tipo indicadoras (*dummies*), que só assumem os valores 0 e 1.

Figura 6.5 – Duas formas equivalentes para regressão múltipla com variáveis numéricas.



Fazendo um desses procedimentos, o *R Commander* oferece os resultados conforme a Figura 6.6

Figura 6.6 – Resultados de uma regressão múltipla.

```

Output
> reg2 <- lm(NOTA_CN~NOTA_MT+NSE, data=parte_enem)
> summary(reg2)
Call:
lm(formula = NOTA_CN ~ NOTA_MT + NSE, data = parte_enem)

Residuals:
    Min       1Q   Median       3Q      Max
-156,92  -42,44    0,62   39,77  167,02

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  222,207    11,036   20,14 < 2e-16 ***
NOTA_MT       0,375     0,023   16,35 < 2e-16 ***
NSE          11,899     1,795    6,63 6,8e-11 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1

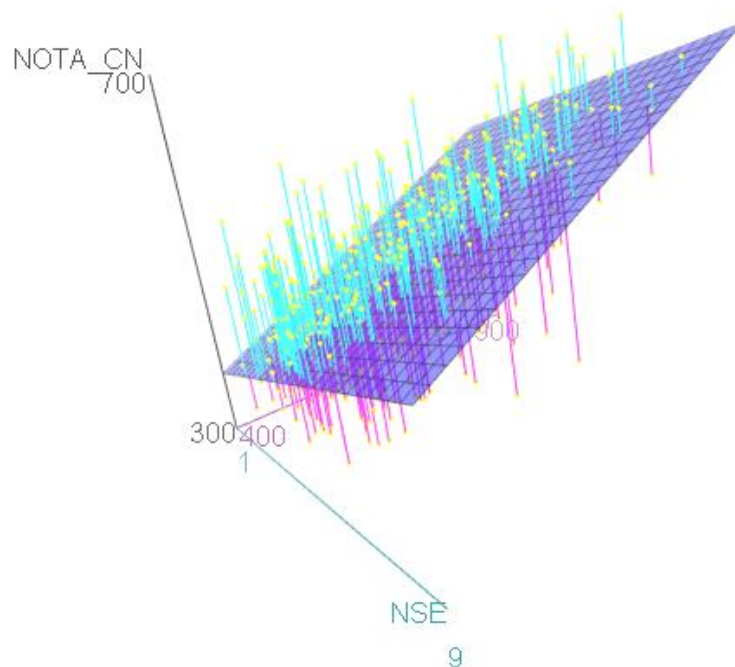
Residual standard error: 56,3 on 686 degrees of freedom
Multiple R-squared:  0,452, Adjusted R-squared:  0,45
F-statistic: 283 on 2 and 686 DF, p-value: <2e-16
    
```

Uma visualização dos pontos do vetor ($NOTA_MT$, NSE , $NOTA_CN$) e do plano formado pelo ajuste de mínimos quadrados pode ser obtido por:

Gráficos >> Gráfico 3D >> Diagrama de dispersão 3D

A Figura 6.7 mostra o resultado gráfico gerado por esse procedimento.

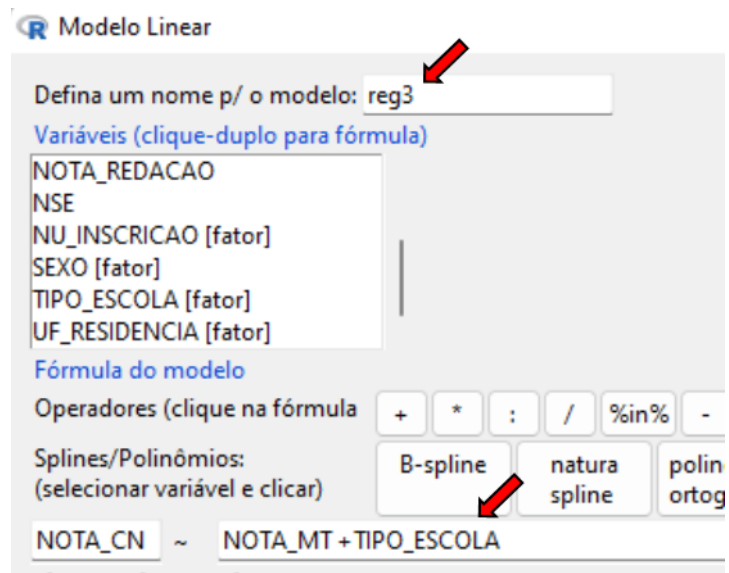
Figura 6.7 – Visualização do ajuste de uma regressão múltipla com duas variáveis independentes quantitativas.



Usando a opção de *modelo linear*, podemos incluir uma variável tipo *fator*, isto porque o próprio software cria as variáveis indicadoras para a regressão. A Figura 6.8 ilustra o processo, considerando a variável dependente $NOTA_CN$ e as variáveis independentes NSE e a indicadora de escola privada. Observar que estamos introduzindo o fator $TIPO_ESCOLA$, mas o software vai criar para a regressão uma variável com valor *um* para escola privada e valor *zero* para escola pública. A ordem da codificação segue a

ordem alfabética, mas fizemos alteração na ordem dos níveis do fator, conforme visto na seção 2.6.

Figura 6.8 – Regressão contendo variável tipo *fator* na lista de variáveis independentes



Executando esse procedimento, obtém-se os seguintes resultados apresentados na tela de saída da Figura 6.9.

Figura 6.9 – Resultados de uma regressão múltipla com variável independente tipo fator.

```
Output
Residuals:
  Min      1Q  Median      3Q      Max
-182,1  -46,4   -3,1    50,7   207,3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      369,47     10,01   36,91  < 2e-16 ***
NSE                20,94      2,12    9,89  < 2e-16 ***
TIPO_ESCOLA[T.Privada]  38,30      7,47    5,13  0,00000038 ***
---
Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Residual standard error: 65,1 on 686 degrees of freedom
Multiple R-squared:  0,266, Adjusted R-squared:  0,264
F-statistic: 124 on 2 and 686 DF, p-value: <2e-16
```

Esse modelo, *reg3*, é chamado de aditivo, porque foi suposto que as diferenças entre escolas públicas e privadas são iguais, independentemente do nível de proficiência do candidato, fazendo com que essa regressão possa ser representada por duas retas paralelas, representando a relação entre *NOTA_MT* e *NOTA_CN* para escolas públicas e para escolas privadas. Em muitos casos, porém, não dá para assumir essa aditividade e se quer um modelo que permita que as retas de escolas públicas e escolas privadas possam ter inclinações diferentes, ou seja, um modelo com *interação*. Para isto, basta substituir o símbolo de adição (+) por um asterisco (*) entre as duas variáveis independentes para as quais se supõe interação. No exemplo, substituir *NOTA_MT + TIPO_ESCOLA* por *NOTA_MT * TIPO_ESCOLA*. Esse modelo é equivalente em fazer regressões separadas: uma para escolas públicas e outra para escolas particulares, cuja representação gráfica foi apresentada na Figura 6.2.

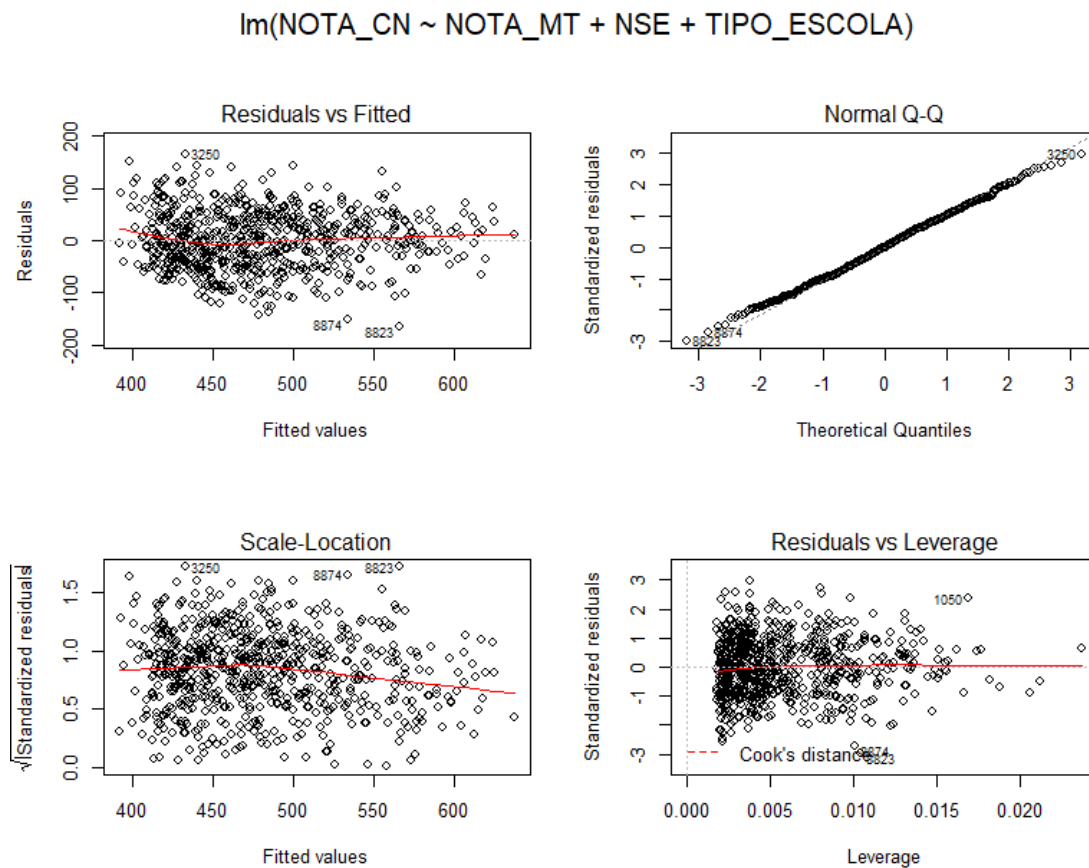
6.4 – Análise de resíduos

A análise de regressão depende de algumas suposições, que podem ser avaliadas por gráficos de resíduos. Após ajustar um modelo de regressão no R Commander, o nome que você deu ao modelo vai aparecer em azul no canto direito superior. Um painel de quatro gráficos de resíduos surge quando você faz:

Modelos >> Gráficos >> Diagnósticos gráficos básicos

Os gráficos de resíduos para a regressão *reg3* são mostrados na Figura 6.10.

Figura 6.10 – Painel de gráficos para diagnóstico das suposições de um modelo de regressão ajustado aos dados.



No *menu Modelos* há vários outros procedimentos que podem ser realizados para completar uma análise de regressão.

Bibliografia

Fox, J. - *Package 'Rcmdr' Reference Manual*. Version 2.7-2, 2022.

<<https://cran.r-project.org/web/packages/Rcmdr/Rcmdr.pdf>> Acessado em 15/04/2022.

Fox, J. - *Getting Started With the R Commander*, 2020.

<<https://cran.r-project.org/web/packages/Rcmdr/vignettes/Getting-Started-with-the-Rcmdr.pdf>> Acessado em 15/04/2022.

Fox, J. - *Using the R Commander: A Point-and-Click Interface for R*. Chapman & Hall/CRC Press (2017).

Fox, J. - *R Commander Installation Notes*.

<<https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/installation-notes.html>> Acessado em 18/04/2022.

Laureto, M. - *Introdução ao R Commander*.

<<http://www.each.usp.br/laureto/cursoR2017/02-RcommanderParte1.pdf>> Acessado em 18/04/2022.

Melo, G.A.; Pinto Jr., J.A. - *R Commander: Facilitando o aprendizado da Estatística*

<<http://www.estadisticacomr.uff.br/wp-content/uploads/2018/10/ApostilaGeorge2016.pdf>> Acessado em 15/04/2022.

Xavier, A; Narimatsu, G.; Carolina, L.; Gonzaga, M. e Luna, A. - *Manual Rcommander*. Departamento de Estatística da UFMG.

<<http://www.each.usp.br/laureto/cursoR2017/02-RcommanderParte1.pdf>>

Acessado em 07/06/2021.