Software R: curso avançado

Felipe Micail da Silva Smolski
Iara Denise Endruweit Battisti

2018-08-04

Sumário

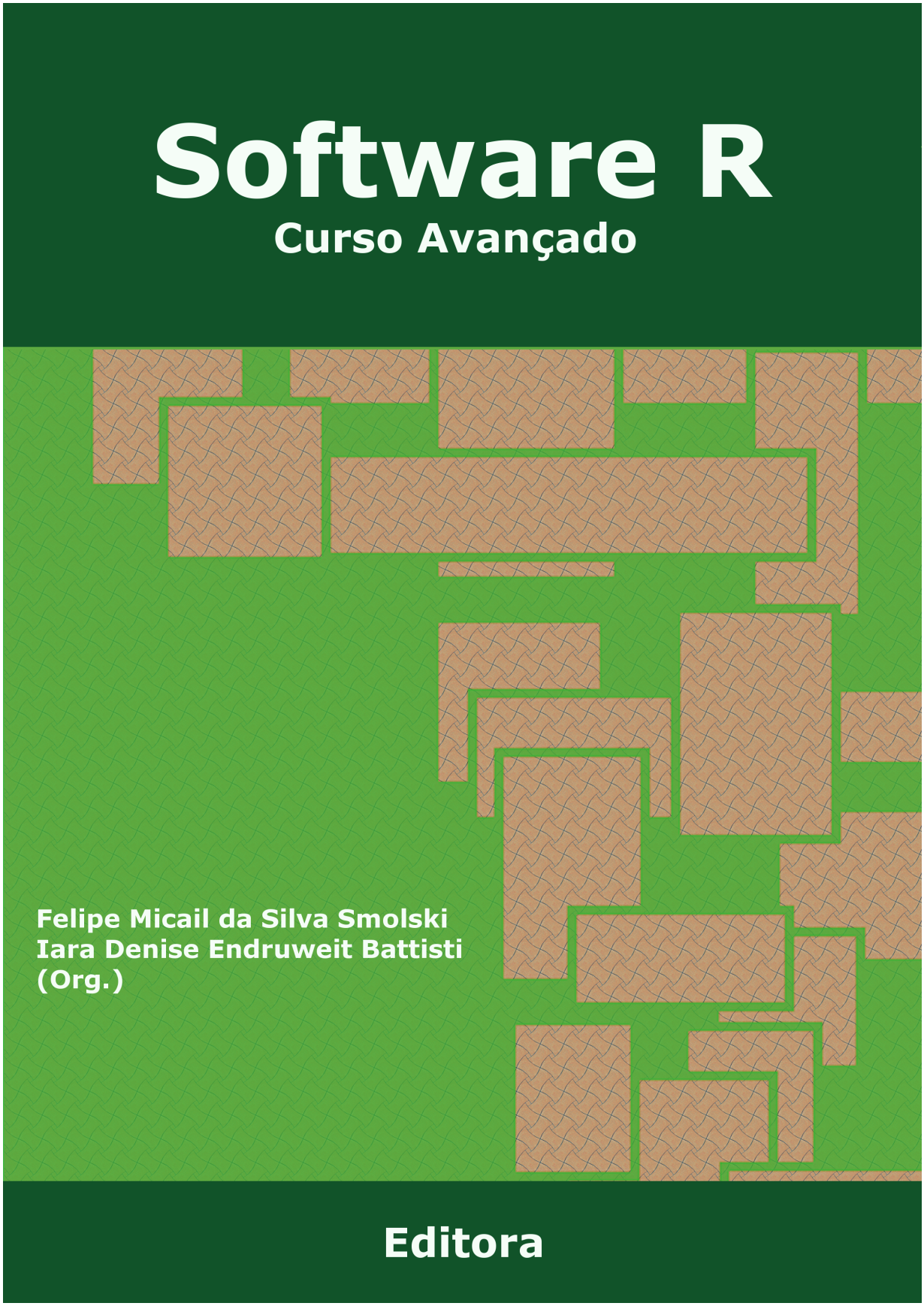
Prefácio

Esta é a estrutura provisória de capítulos do **Curso Avançado em Estatística com R da UFFS**:

- Delineamentos Experimentais
- Análise Fatorial
- Regressão Múltipla
- Regressão Logística

Algumas sugestões a incluir:

- Produção de Mapas
- Análise de Clusters
- Manipulação de bases de dados
- Análise de redes
- O que mais?


Software R
Curso Avançado
Felipe Micail da Silva Smolski
Iara Denise Endruweit Battisti
(Org.)
Editora

Introdução

Capítulo 1

Delineamentos Experimentais

A experimentação é uma parte da estatística probabilística que estuda o planejamento, execução, coleta de dados, análise de dados e interpretação dos resultados provenientes de um experimento.

Um experimento é um procedimento planejado com base em uma hipótese, que tem por objetivo provocar fenômenos (tratamentos) de forma controlada, analisando e interpretando os resultados obtidos.

O tratamento é o método, elemento ou material cujo efeito desejamos avaliar em um experimento. Por exemplo: formas de preparo de solo, diferentes cultivares, doses de adubação, controle de insetos e outras pragas, controle de uma doença. Num experimento, somente o tratamento variade uma unidade experimental para outra, as demais condições são mantidas constantes, salvo erros não controláveis.

E alguns experimentos, utiliza-se a testemunha (nas ciências agrárias e ambientais) ou placebo (na saúde), que são as unidades experimentais que não recebem tratamento.

A unidade experimental é a unidade que recebe o tratamento uma vez e, normalmente são chamadas de parcelas. A escolha da unidade experimental depende dos tipos de tratamentos que serão avaliados. Podem ser: uma área de campo, um vaso com solo, um animal, uma placa de Petri, uma planta. Em áreas de campo, normalmente utiliza-se a bordadura. Num experimento, recomenda-se, no mínimo, a utilização de 20 UEs.

Em um experimento, a variável a ser avaliada chamamos de variável resposta. Por exemplo, núumero de grãos por planta, número de folhas por planta, altura das plantas.

1.1 Princípios básicos da Experimentação

1.1.1 Repetição

A repetição consiste na aplicação do mesmo tratamento sobre duas ou mais unidades experimentais. Permite estimar o erro experimental e avaliar de forma mais precisa o efeito de cada tratamento.

O erro experimental é caracterizado pela variância entre as unidades experimentais que receberam o mesmo tratamento.

1.1.2 Casualização

A casualização consiste na aplicação dos tratamentos aleatoriamente (sorteio) sobre as unidades experimentais. A casualização é usada para obter a independência dos erros, ou seja, evitar que determinados tratamentos sejam favorecidos.

1.1.3 Controle local

Quando tiver heterogeneidade no material experimental: plantas de diferentes alturas, animais de diferentes idades, solo com declividade, deve-se separar o material em grupos homogêneos e aplicar o tratamento uma vez dentro de cada grupo (blocos). A homogeneidade ou não do material dá origem aos tipos de delineamentos:

- Delineamento Inteiramente Casualizado (DIC): material experimental homogêneo;
- Delineamento Blocos Casualizados (DBC): material experimental com uma fonte de heterogeneidade;
- Delineamento Quadrado Latino (DQL): material experimental com duas fontes de heterogeneidade.

1.2 Análise de Variância

Para saber se existe diferença significativa entre as médias resultados dos efeitos de tratamentos, realiza-se a Análise de Variância (ANOVA).

Tabela 1.1: Nome da Tabela

Fonte de Variação	Graus de Liberdade (GL)	Soma de Quadrados (SQ)	Quadrado Médio (QM)	Falc	P
Tratamento	I-1	SQtrat	QMat	QMatr/QMerro	P
Erro	GLerro	SQerro	QMerro		
Total	IJ-1	SQtotal			

1.3 Hipóteses estatísticas

- H0: Não existe diferença entre as médias dos tratamentos
- H1: Existe, pelo menos, uma diferença entre as médias dos tratamentos

1.4 Delineamento Inteiramente Causalizado (DIC)

É utilizado quando as unidades experimentais são homogêneas. É o mais simples dos delineamentos e os tratamentos são designados às unidades experimentais de forma casualizada, por meio de um único sorteio. Usado principalmente em pequenos animais, casas de vegetação e em laboratórios.

Exemplo: Um produtor deseja avaliar 4 variedades de pera (A, B, C e D). Para tanto, instalou um experimento no delineamento inteiramente casualizado, utilizando 5 repetições por variedade. Os resultados, peso médio do fruto, estão apresentados a seguir:

Existe diferença significativa entre as variedades de pera, considerando o peso médio dos frutos de cada variedade?

Para responder esta pergunta, utilizamos a Análise de Variância (ANOVA).

No software RStudio:

Criar o arquivo acima em planilha eletrônica. Nomear como DIC e salvar em formato .xls.

Importar no RStudio:

```
require(readxl)
url <- "https://github.com/Smolski/softwarelivrer/raw/master/avancado/dic.xls"
destfile <- "dic.xls"
curl::curl_download(url, destfile)
```

	A	B	C
1	Variedade	Repeticao	Peso
2	A	1	78
3	A	2	88
4	A	3	72
5	A	4	74
6	A	5	98
7	B	1	79
8	B	2	56
9	B	3	71
10	B	4	96
11	B	5	55
12	C	1	63
13	C	2	68
14	C	3	58
15	C	4	79
16	C	5	59
17	D	1	60
18	D	2	65
19	D	3	59
20	D	4	54
21	D	5	58

Figura 1.1: Variedades de pera separadas por grupos em faixas de peso e repetição

```
DIC <- read_excel(destfile)
attach(DIC)
```

O comando que gera a análise de variância é o `aov()` e o comando que exibe o quadro da ANOVA é o `anova`. Então, podemos gerar o quadro da análise de uma são vez associando os dois comandos.

```
anova=aov(Peso~Variedade)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Variedade    3   1414   471.3   3.775 0.0319 *
## Residuals   16   1997   124.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hipóteses estatísticas:

- H0: ti = 0 (as médias dos tratamentos nãao diferem entre si)
- H1: ti $\neq$ 0 (existe, no mínimo, uma diferença entre as médias dos tratamentos)

Como p = 0,0319 (0,01 $\leq$ p "menor ou igual a" 0,05), rejeita-se H0 com nível de significância de 5% e conclui-se que existe diferença significativa entre as médias dos tratamentos.

Para saber quais as médias que diferem, utilizamos o teste de Tukey.

```
attach(DIC)
```

```
## The following objects are masked from DIC (pos = 3):
##
##     Peso, Repeticao, Variedade
```

```
TukeyHSD(anova,as.factor("Variedade"),ordered=TRUE)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = Peso ~ Variedade)
##
## $Variedade
##     diff        lwr     upr     p adj
## C-D  6.2 -14.016299 26.4163 0.8163995
## B-D 12.2  -8.016299 32.4163 0.3429223
## A-D 22.8   2.583701 43.0163 0.0244592
## B-C  6.0 -14.216299 26.2163 0.8303280
```

```
## A-C 16.6  -3.616299 36.8163 0.1281553
## A-B 10.6  -9.616299 30.8163 0.4602137
```

Para que o RStudio apresente uma tabela com as médias e letras indicando quais as médias que diferiram, devemos instalar o pacote `agricolae`.

```
library(agricolae)
HSD.test(anova,as.factor("Variedade"),console=TRUE)
```

```
##
## Study: anova ~ as.factor("Variedade")
##
## HSD Test for Peso
##
## Mean Square Error:  124.825
##
## Variedade,  means
##
##   Peso       std r Min Max
## A 82.0 10.862780 5  72  98
## B 71.4 17.096783 5  55  96
## C 65.4  8.561542 5  58  79
## D 59.2  3.962323 5  54  65
##
## Alpha: 0.05 ; DF Error: 16
## Critical Value of Studentized Range: 4.046093
##
## Minimun Significant Difference: 20.2163
##
## Treatments with the same letter are not significantly different.
##
##   Peso groups
## A 82.0      a
## B 71.4     ab
## C 65.4     ab
## D 59.2      b
```

*Médias dos tratamentos não seguidas por mesma letra diferem pelo teste de Tukey, ao nível de 5% de significância.

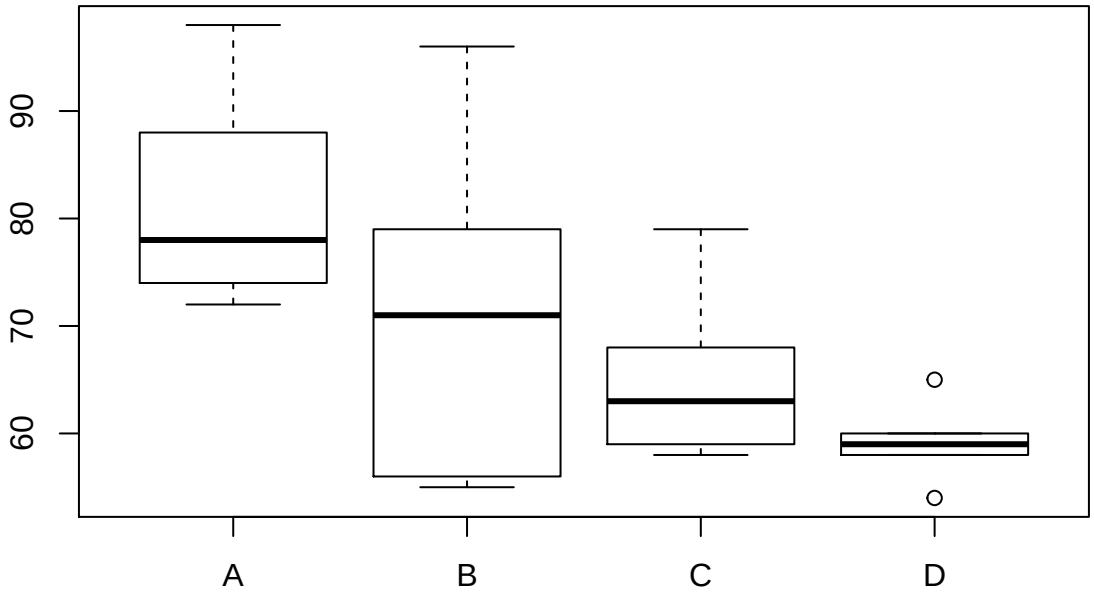
Conclusão: A variedade de pera A apresentou o maior peso médio dos frutos, que não diferiu significativamente do peso médio das variedades B e C. A variedade de pera D apresentou o menor peso médio dos frutos, que não diferiu significativamente do peso médio das variedades B e C. As variedades B e C apresentaram peso médio dos frutos intermediário.

```
attach(DIC)
```

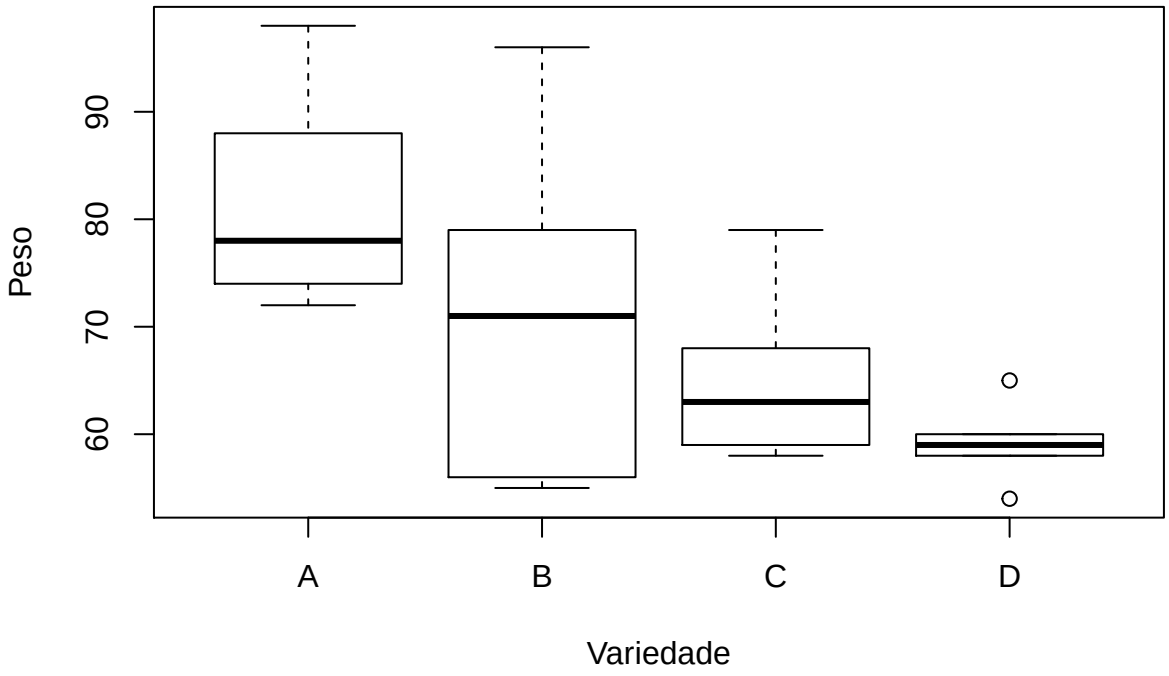
```
## The following objects are masked from DIC (pos = 4):
##
##     Peso, Repeticao, Variedade

## The following objects are masked from DIC (pos = 5):
##
##     Peso, Repeticao, Variedade
```

```
boxplot(Peso~Variedade)
```




```
boxplot(Peso~Variedade,xlab="Variedade",ylab="Peso")
```



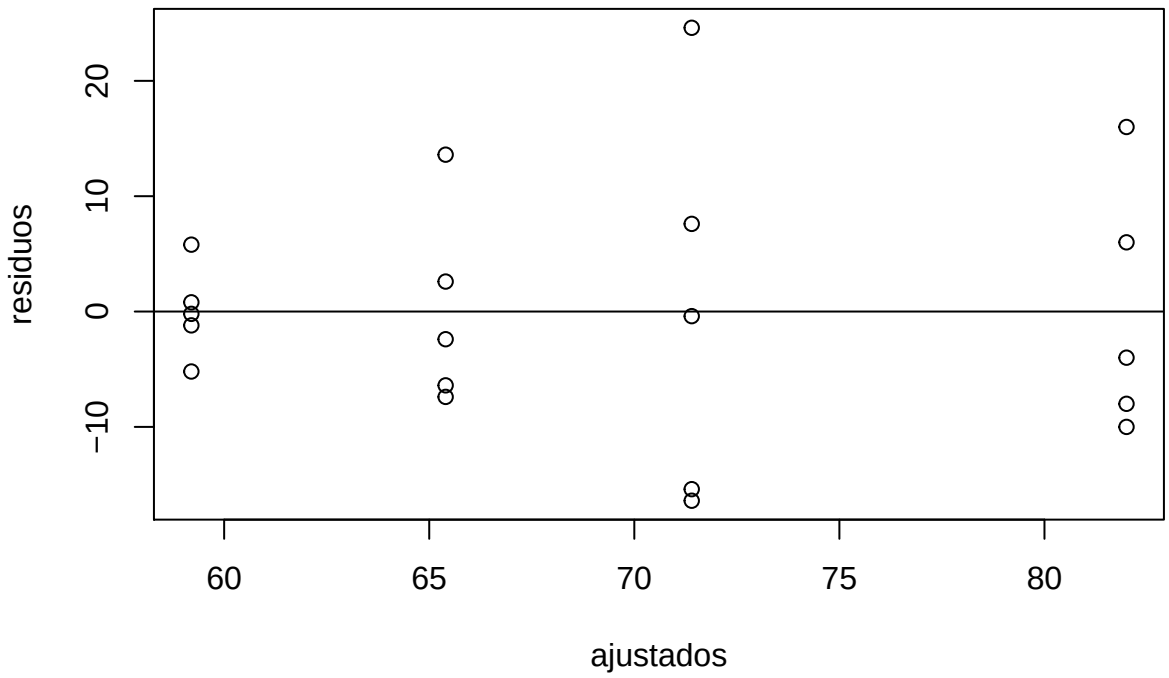

```
tapply(Peso,Variedade,mean)
```

```
##    A    B    C    D
## 82.0 71.4 65.4 59.2
```

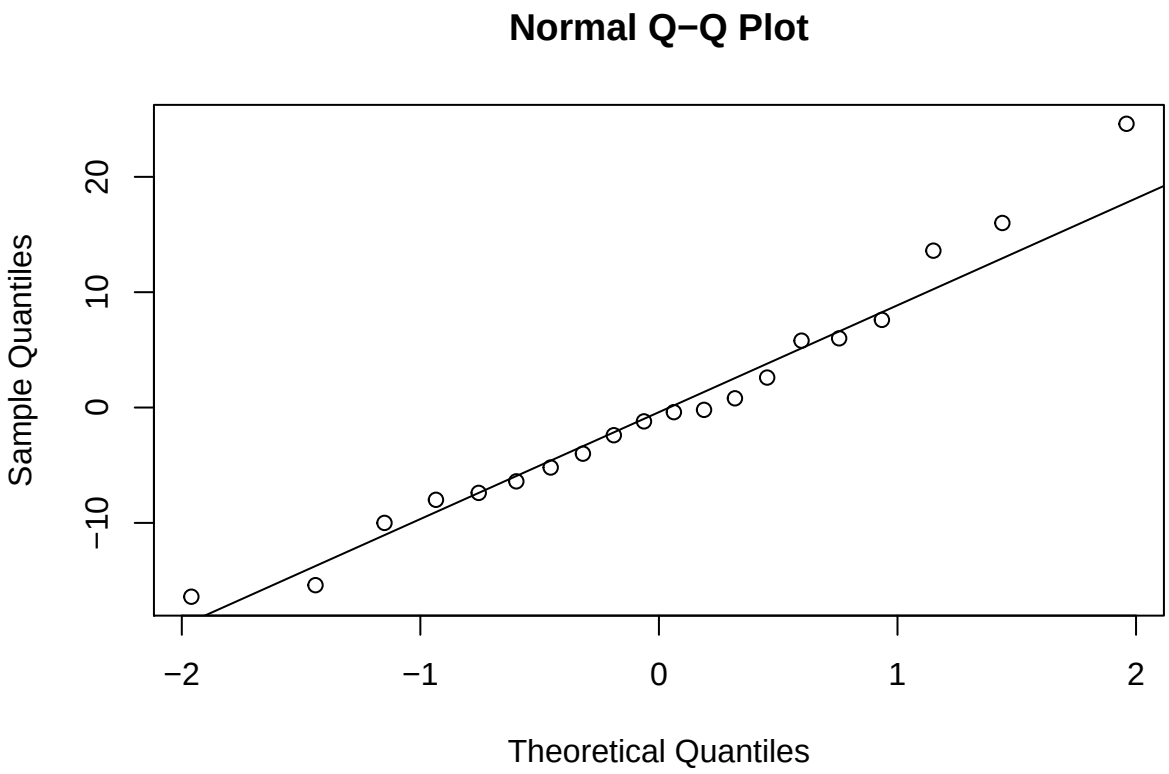
```
tapply(Peso,Variedade,sd)
```

```
##         A         B         C         D
## 10.862780 17.096783  8.561542  3.962323
```

```
residuos=residuals(anova)
ajustados=fitted(anova)
plot(ajustados,residuos)
abline(h=0)
```




```
qqnorm(residuos)
qqline(residuos)
```



1.5 Delineamento Blocos Casualizados (DBC)

É utilizado quando as unidades experimentais são heterogêneas. Os tratamentos são designados às unidades experimentais de forma casualizada, por meio de sorteio por blocos. Na área agrícola, é usado principalmente em áreas de campo e grandes animais.

Exemplo: Uma Nutricionista elaborou 4 dietas e quer aplicá-las em 20 pessoas a fim detestar suas eficiências quanto à perda de peso. Porém ela notou que entre essas 20 pessoas existem 5 grupos de faixas iniciais de peso. Então, para aumentar a eficácia do teste ela separou os 20 indivíduos em 5 grupos de faixas de peso.

Criar o arquivo acima em planilha eletrônica. Nomear como DBC e salvar em formato .xls.

Importar no RStudio:

```
require(readxl)
url <- "https://github.com/Smolski/softwarelivrer/raw/master/avancado/dbc.xls"
destfile <- "dbc.xls"
curl::curl_download(url, destfile)
```

	A	B	C
1	Tratamentos	Blocos	Perda
2	Dieta 1	peso A	2
3	Dieta 2	peso A	5
4	Dieta 3	peso A	2
5	Dieta 4	peso A	5
6	Dieta 1	peso B	3
7	Dieta 2	peso B	7
8	Dieta 3	peso B	4
9	Dieta 4	peso B	3
10	Dieta 1	peso C	2
11	Dieta 2	peso C	6
12	Dieta 3	peso C	5
13	Dieta 4	peso C	4
14	Dieta 1	peso D	4
15	Dieta 2	peso D	5
16	Dieta 3	peso D	1
17	Dieta 4	peso D	3
18	Dieta 1	peso E	2
19	Dieta 2	peso E	5
20	Dieta 3	peso E	4
21	Dieta 4	peso E	4

Figura 1.2: Indivíduos separados por grupos em faixas de peso

```
DBC  <- read_excel(destfile)

attach(DBC)
anova=aov(Perda~Tratamentos+Blocos)
summary(anova)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Tratamentos  3   25.2     8.4   6.000 0.00973 **
## Blocos       4    3.2     0.8   0.571 0.68854
## Residuals   12   16.8     1.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hipóteses estatísticas:

- H0: ti = 0 (as médias dos tratamentos não diferem entre si)
- H1: ti $\neq$ 0 (existe, no mínimo, uma diferença entre as médias dos tratamentos)

Como p = 0,00973 (p $\leq$ 0,01), rejeita-se H0 com nível de significância de 1% e conclui-se que existe diferença significativa entre as médias dos tratamentos.

- H0: σ^2 blocos = 0
- H1: σ^2 blocos $\leq$ 0

Como p = 0,68854 (p $\leq$ 0,05), não rejeita-se H0 e conclui-se que a variância entre os blocos não é significativa.

```
attach(DBC)
HSD.test(anova,as.factor("Tratamentos"),console=TRUE)
```

```
##
## Study: anova ~ as.factor("Tratamentos")
##
## HSD Test for Perda
##
## Mean Square Error:  1.4
##
## Tratamentos,  means
##
##          Perda       std r Min Max
## Dieta 1   2.6 0.8944272 5   2   4
## Dieta 2   5.6 0.8944272 5   5   7
## Dieta 3   3.2 1.6431677 5   1   5
## Dieta 4   3.8 0.8366600 5   3   5
##
```

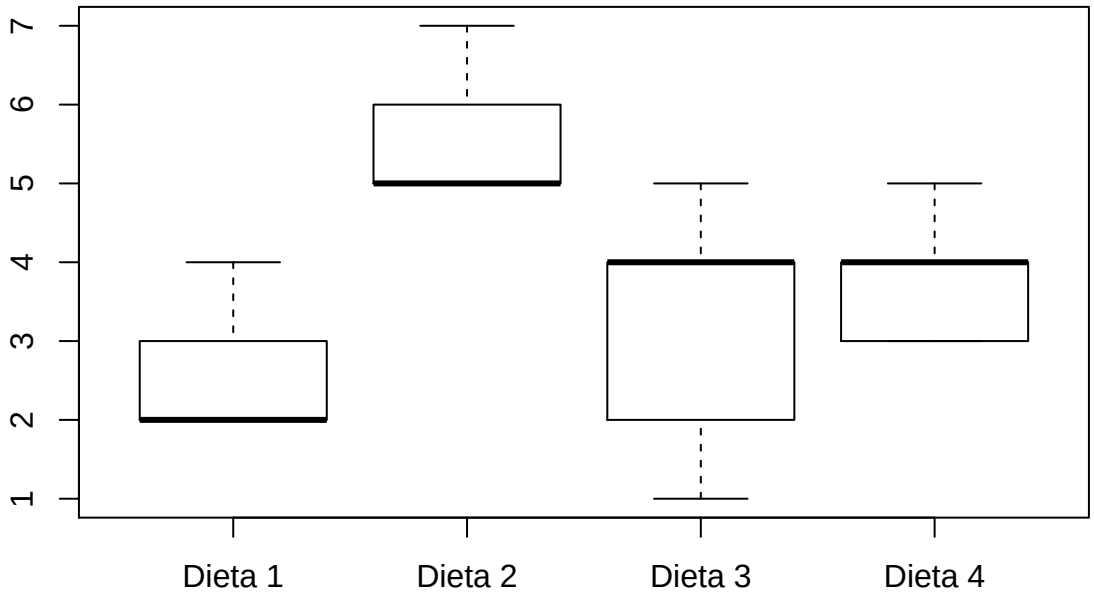
```
## Alpha: 0.05 ; DF Error: 12
## Critical Value of Studentized Range: 4.19866
##
## Minimun Significant Difference: 2.221722
##
## Treatments with the same letter are not significantly different.
##
##         Perda groups
## Dieta 2   5.6      a
## Dieta 4   3.8     ab
## Dieta 3   3.2      b
## Dieta 1   2.6      b
```

Médias dos tratamentos não seguidas por mesma letra diferem pelo teste de Tukey, ao nível de 5% de significância.

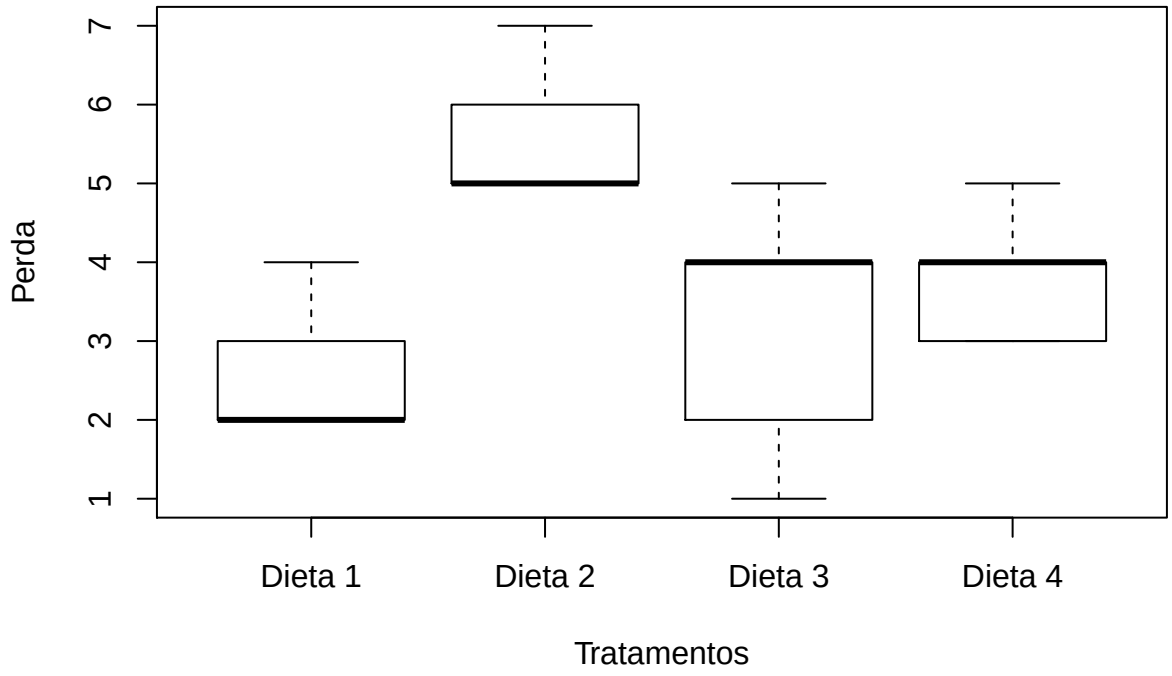
Conclusão: A dieta que resultou na maior perda de peso foi a dieta 2, que não diferiu da dieta 4. A dieta que resultou na menor perda de peso foi a dieta 1, que não diferiu das dietas 3 e 4.

Medidas descritivas com a variável resposta:

```
boxplot(Perda~Tratamentos)
```




```
boxplot(Perda~Tratamentos,xlab="Tratamentos",ylab="Perda")
```



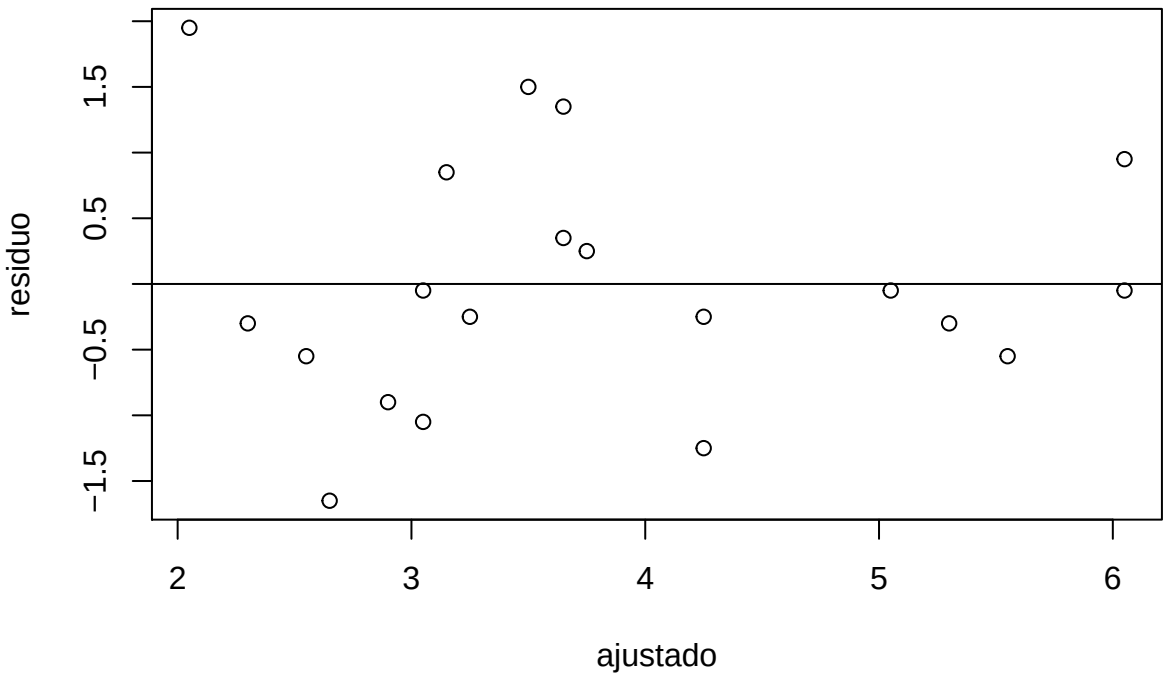

```
tapply(Perda,Tratamentos,mean)
```

```
## Dieta 1 Dieta 2 Dieta 3 Dieta 4
##     2.6     5.6     3.2     3.8
```

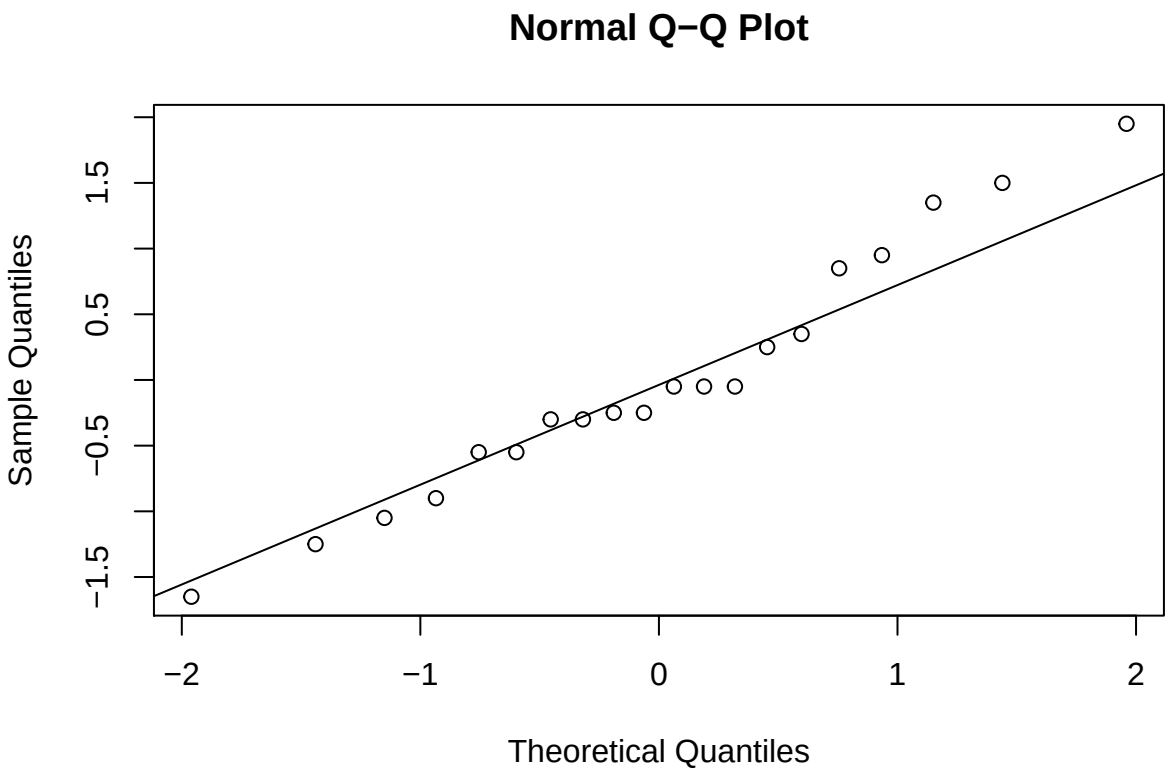
```
tapply(Perda,Tratamentos,sd)
```

```
##   Dieta 1   Dieta 2   Dieta 3   Dieta 4
## 0.8944272 0.8944272 1.6431677 0.8366600
```

```
residuo=residuals(anova)
ajustado=fitted(anova)
plot(ajustado,residuo)
abline(h=0)
```




```
qqnorm(residuo)
qqline(residuo)
```


Normal Q−Q Plot
Sample Quantiles
−1.5
−0.5
0.5
1.5
−2
−1
0
1
2
Theoretical Quantiles

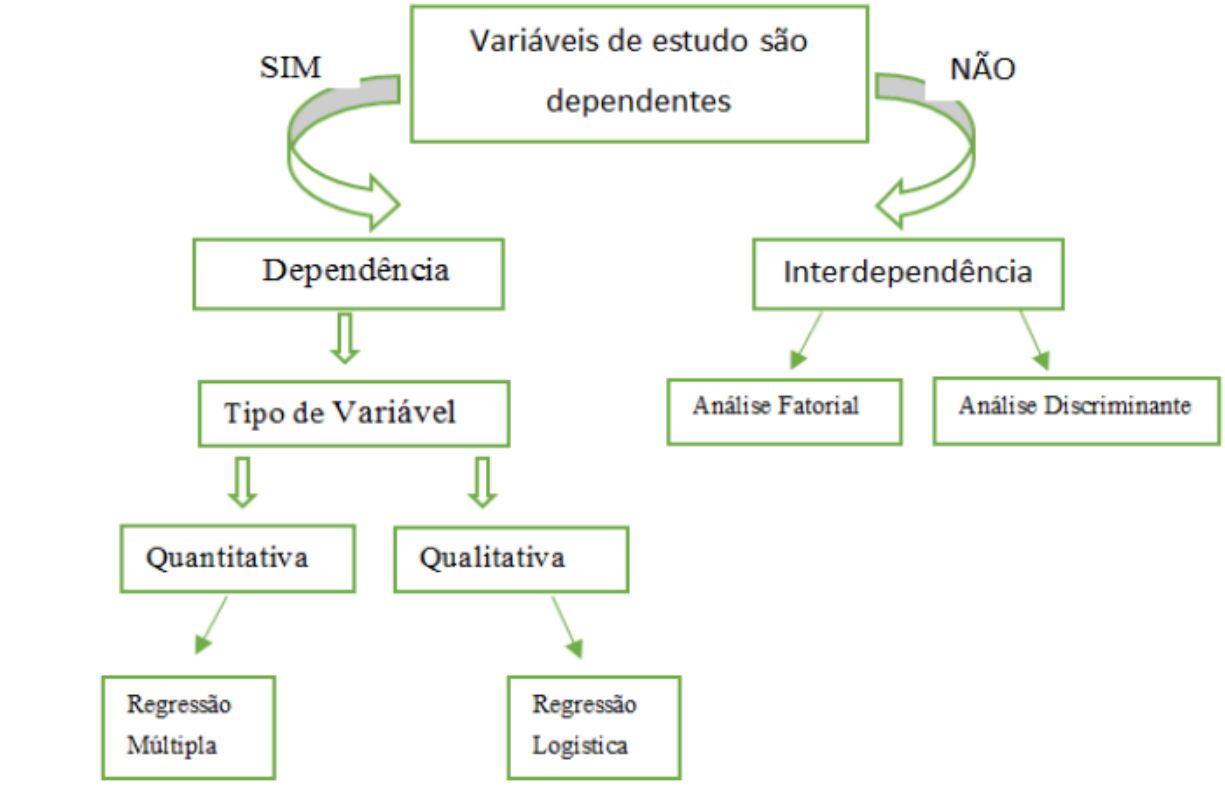
Capítulo 2

Análise Fatorial

A análise fatorial é um método estatístico utilizado para descrever a variabilidade entre variáveis observadas e possivelmente correlacionadas em termos de um número potencialmente menor de variáveis não observadas chamadas fatores.

Assim, é possível que as variaçõess de três ou quatro variáveis observadas possam ser explicadas por somente um fator, o que evidencia a utilidade da análise fatorial para descrever um conjunto de dados utilizando para isso apenas alguns fatores.

Diferentemente da análise de variância, regressão e análise discriminante, onde uma das variáveis é identificada como a variável dependente, examina-se todo o conjunto de relações interdependentes entre variáveis.



Fonte: Azevedo, R. Métodos Quantitativos: Análise multivariada.

A análise fatorial aborda o problema de analisar a estrutura das inter-relações (correlações) entre um grande número de variáveis (escores de testes, itens de testes, respostas de questionários), definindo um conjunto de dimensões latentes comuns, chamados fatores. Então, a análise fatorial, permite primeiro identificar as dimensões separadas da estrutura e então determinar o grau em que cada variável é explicada por cada dimensão. Uma vez que essas dimensões e a explicação da cada variável estejam determinadas, os dois principais usos da análise fatorial podem ser conseguidos:

- **Resumo**: ao resumir os dados, a análise fatorial obtém dimensões latentes que, quando interpretadas e compreendidas, descrevem os dados em um núumero muito menor de conceitos do que as variáveis individuais originais.

- **Redução de dados**: pode ser obtida calculando escores para cada dimensão latente e substituindo as variáveis originais pelos mesmos.

As técnicas analíticas fatoriais podem ser classificadas quanto aos seus objetivos como **exploratória** ou **confirmatória**. Exploratória, útil na busca da estrutura em um conjunto de variáveis ou como um método de redução de dados. Sob esta perspectiva, as técnicas analíticas fatoriais "consideram o que os dados oferecem"e não estabelecem restriçoes *a priori* sobre o número de componentes a serem extraídos. O uso da análise fatorial em situações, que se deseja testar hipóteses envolvendo questões sobre, quais variáveis deveriam ser agrupadasem fator ou número exato de fatores, por exemplo, a análise fatorial desempenha um

papel confirmatório, ou seja, avalia o grau em que os dados satisfazem a estrutura esperada.

Exemplo: em maketing, fatores associados às características do produto, clientes e até mesmo da organização.

Em estudos visando analisar o inter-relacionamento e o agrupamento de indivíduos, cidade ou regiões em grupos homogêneos em relação à mobilidade, preferências pessoais, condições de desenvolvimento, entre outras variáveis.

2.1 Pressupostos

A análise fatorial clássica exige que alguns pressupostos sejam satisfeitos, quais sejam (MALHOTRA, 2001):

a. Normalidade dos dados: apesar deste pressuposto não ser crítico quando a estimação é realizada por mínimos quadrados ordinários, a exigência de normalidade auxilia na análise, evitando possíveis assimetrias e a presença de *outliers*.

b. Variáveis quantitativas medidas em escala Intervalar ou de Razão. Esse pressuposto é crítico, pois a análise deve ser realizada com variáveis quantitatias e, frequentemente, alguns estudos são realizados utilizando variáveis ordinais (as quaiss são qualitativas) na análise fatorial clássica (o que é errado de muitas maneiras).

c. Como diretriz inicial deve haver ao menos quatro a cinco vezes mais observações do que variáveis.

2.2 Estatísticas Associadas a Análise Fatorial

Em geral, as estatísticas utilizadas no processo de análise fatorial são (AAKER-KUMARDAY, 2001):

- Teste de esfericidade de Bartlett: estatística de teste usada para examinar a hipótese de que as variáveis não sejam correlacionadas na população, ou seja, a matriz de correlação da população é uma matriz identidade onde cada variável se correlaciona perfeitamente com ela própria (r=1), mas não apresenta correlação com as outras variáveis (r=0).

- Matriz de correlação: o triângulo inferior da matriz exibe as correlações simples, r, entre todos os pares possíveis de variáveis incluídas na análise, enquanto os elementos da diagonal, que são todos iguais a 1, em geral são omitidos.

- Comunalidade: porção da variância que uma variável compartilha com todas as outras variáveis consideradas, sendo também a proporção de variância explicada pelos fatores

comuns.

- Autovalor: representa a variância total explicada por cada fator.
- Cargas fatoriais: correlação simples entre as variáveis e os fatores.
- Gráfico das cargas dos fatores: gráfico das variáveis originais utilizando as cargas fatoriais como ordenadas.
- Matriz de fatores ou matriz principal: contém as cargas fatoriais de todos as variáveis em todos os fatores extraídos.
- Escores fatoriais: escores compostos estimados para cada entrevistado nos fatores derivados.
- Medida de adequacidade da amostra de Kaiser-Meyer-Olkin (KMO): é o índice usado para avaliar a adequacidade da análise fatorial. Valores altos (entre 0,5 e 1,0) indicam que a análise fatorial é apropriada. Valores abaixo de 0,5 indicam que a análise fatorial pode ser inadequada.
- Percentagem de variância: percentagem da variância total atribuída a cada fator.
- Resíduos: diferenças entre as correlações observadas, dadas na matriz de correlação de entrada (input) e as correlações reproduzidas, conforme estimadas pela matriz de fatores.
- Scree plot: gráfico dos autovalores versus número de fatores por ordem de extração.

Exemplo 1:

(MALHOTRA, 2001) Suponhamos que um pesquisador queira avaliar os benefícios que os consumidores esperam de um dentifrício. Foi entrevistada uma amostra de 30 pessoas em um supermercado, para que indicassem seu grau de concordância com as seguintes afirmações, utilizando uma escala de 7 pontos (1= discordância total, 7 =concordância total).

- V1: É importante comprar um creme dental que evite cáries.
- V2: Gosto de um creme dental que clareie os dentes.
- V3: Um creme dental deve fortificar as gengivas.
- V4: Prefiro um creme dental que refresque o hálito.
- V5: Manter os dentes sadios não é uma vantagem importante de um creme dental.
- V6: O aspecto mais importante na compra de um creme dental é tornar os dentes atraentes.

Inicialmente podemos explorar algumas estatísticas descritivas relacionadas às variáveis pesquisadas, utilizando a função `summary`:

```
require(readxl)
```

```
## Carregando pacotes exigidos: readxl
```

```
url <- "https://github.com/Smolski/softwarelivrer/raw/master/avancado/creme_dental_exemp
destfile <- "creme_dental_exemplo1.xlsx"
curl::curl_download(url, destfile)
creme_dental_exemplo1 <- read_excel(destfile)

attach(creme_dental_exemplo1)
summary(creme_dental_exemplo1)
```

```
##        v1              v2            v3            v4            v5
##  Min.   :1.000   Min.   :2.0   Min.   :1.0   Min.   :2.0   Min.   :1.0
##  1st Qu.:2.000   1st Qu.:3.0   1st Qu.:2.0   1st Qu.:3.0   1st Qu.:2.0
##  Median :4.000   Median :4.0   Median :4.0   Median :4.0   Median :3.5
##  Mean   :3.933   Mean   :3.9   Mean   :4.1   Mean   :4.1   Mean   :3.5
##  3rd Qu.:6.000   3rd Qu.:5.0   3rd Qu.:6.0   3rd Qu.:5.0   3rd Qu.:5.0
##  Max.   :7.000   Max.   :7.0   Max.   :7.0   Max.   :7.0   Max.   :7.0
##        v6
##  Min.   :2.000
##  1st Qu.:3.000
##  Median :4.000
##  Mean   :4.167
##  3rd Qu.:4.750
##  Max.   :7.000
```

2.3 Passos da Análise Fatorial

Basicamente, os seguintes passos conduzem a análise fatorial: entrada de dados, cálculo das correlações entre as variáveis, extração inicial dos fatores e a rotação da matriz.

2.3.1 Construção da Matriz de Correlação

Entrada de Dados (BASE): os dados de entrada da análise fatorial geralmente tomam a forma de um conjunto de valores de variáveis para cada objeto ou indivíduo na amostra. Toda matriz, cujos componentes ofereçam uma medida de similaridade entre variáveis, pode ser passível de análise fatorial. A medida de similaridade não precisa ser uma correlação, embora, geralmente, ou seja:

Para que a análise fatorial seja adequada, as variáveis devem ser correlacionadas.

Espera-se também que as variáveis altamente correlacionadas umas com as outras se correlacionem também com o(s) mesmo(s) fatore(s).

Note que existem correlações amostrais positivas e negativas relativamente elevadas entre V1 (prevenção de cáries), V3 (gengivas fortes) e V5 (dentes sadios). Espera-se que essas variáveis se relacionem com o mesmo conjunto de fatores. Verificam-se também correlações relativamente elevadas entre V2 (clareie os dentes), V4 (hálito puro) e V6 (dentes atraentes). Essas variáveis também devem correlacionar-se com os mesmos fatores.

```
matcor <- cor(creme_dental_exemplo1)
print(matcor, digits = 2)
```

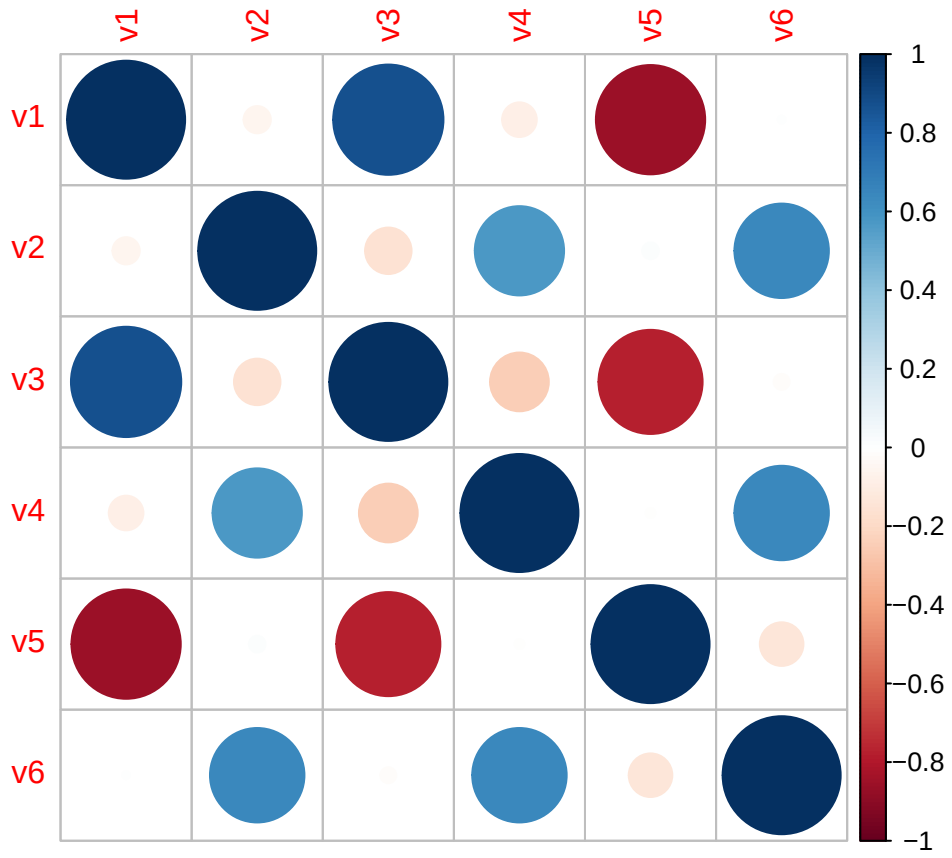
```
##         v1     v2     v3      v4      v5      v6
## v1  1.0000 -0.053  0.873 -0.0862 -0.8576  0.0042
## v2 -0.0532  1.000 -0.155  0.5722  0.0197  0.6405
## v3  0.8731 -0.155  1.000 -0.2478 -0.7778 -0.0181
## v4 -0.0862  0.572 -0.248  1.0000 -0.0066  0.6405
## v5 -0.8576  0.020 -0.778 -0.0066  1.0000 -0.1364
## v6  0.0042  0.640 -0.018  0.6405 -0.1364  1.0000
```

```
require(corrplot)
```

```
## Carregando pacotes exigidos: corrplot
```

```
## corrplot 0.84 loaded
```

```
corrplot(matcor, method="circle")
```



Na figura acima, as correlações estão em cor azul porque são positivas, com tons mais fortes para as correlações mais altas.

Para testar a conveniência do modelo fatorial pode-se aplicar o teste de esfericidade de Bartlett para testar a hipótese nula, de que as variáveis não sejam correlacionadas na população. Um valor elevado da estatística de teste favorece a rejeição da hipótese nula.

Também, a medida de adequacidade da amostra de Kaiser-Meyer-Olkin (KMO) compara as magnitudes dos coeficientes de correlação observados com as magnitudes dos coeficientes de correlação parcial. Pequenos valores de KMO indicam que as correlações entre os pares de variáveis não podem ser explicadas por outras variáveis, indicando que a análise fatorial não é adequada.

Hipóteses:

Ho: A matriz de correlação da população é uma matriz identidade, ou seja as variáveis não são correlacionadas na população.

H1: A matriz de correlação da população não é uma matriz identidade, ou seja as variáveis são correlacionadas na população.

```
#install.packages("psych")
require(psych)
```

```
## Carregando pacotes exigidos: psych
```

```
cortest.bartlett(creme_dental_exemplo1)
```

```
## R was not square, finding R from data
```

```
## $chisq
## [1] 111.3138
##
## $p.value
## [1] 9.017094e-17
##
## $df
## [1] 15
```

Veja que a hipótese nula de que a matriz de correlação da população seja uma matriz identidade é rejeitada pelo teste de esfericidade de Bartlett. A estatística qui-quadrado aproximada é 111,314, com 15 graus de liberdade, significativa ao nível de 0,05.

```
KMO(creme_dental_exemplo1)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = creme_dental_exemplo1)
## Overall MSA =  0.66
## MSA for each item =
##   v1   v2   v3   v4   v5   v6
## 0.62 0.70 0.68 0.64 0.77 0.56
```

A estatística KMO maior que 0,5 também concorda quanto ao fato de que a análise fatorial pode ser considerada uma técnica apropriada para analisar a matriz de correlação.

2.3.2 Método de Análise Fatorial

As duas abordagens básicas são a análise de componentes principais (ACP) e a análise fatorial (AFC) comum ou análise fatorial exploratória (AFE), embora existam diferentes métodos de extração de fatores da matriz de correlações, que de forma geral, são métodos numericamente complexos. Na análise de componentes principais, o objetivo da extração de fatores é encontrar um conjunto de fatores que formem uma combinação linear das variáveis originais ou da matriz de correlações. Assim, se as variáveis X1 , X2 , X3 , . . . , Xn são altamente correlacionadas entre si, elas serão combinadas para formar um fator, e assim, sucessivamente, com todas as demais variáveis da matriz de correlação.

A análise fatorial exploratória pode trazer informações importantes sobre a estrutura multivariada de um instrumento de mensuração, identificando os construtos teóricos.

O segundo objetivo da analise fatorial exploratória está relacionado à redução de dados e descoberta de ponderações ótimas para as variáveis mensuradas, de forma que um grande conjunto de variáveis possa ser reduzido a um conjunto menor de índices sumários que tenham máxima variabilidade e fidedignidade. A redução de dados é especialmente possível pela aplicação da Análise dos Componentes Principais (ACP) e não pelo uso da analise fatorial comum (AFC), havendo uma diferença fundamental entre os dois métodos: a ACP trabalha com a variância total observada, enquanto a AFC trabalha somente com a variância partilhada dos itens (variância erro e variância única são excluídas) (LAROS, 2012).

Na AFC, os fatores são estimados para explicar as covariâncias entre as variáveis observadas, portanto os fatores são considerados como as causas das variáveis observadas. Já na ACP, os componentes são estimados para representar a variância das variáveis observadas de uma maneira tão econômica quanto possível. Os componentes principais são somas otimamente ponderadas das variáveis observadas, neste sentido, as variáveis observadas são consideradas as causas dos componentes principais (LAROS, 2012).

Assim, recomenda-se a ACP, quando o objetivo é determinar o número mínimo de fatores que respondem pela máxima variância nos dados, sendo os fatores chamados componentes principais (MALHOTRA, 2001).

Obs.:

cor = TRUE: as componentes principais serão geradas a partir da matriz de correlação.

cor = FALSE: as componentes principais serão geradas a partir da matriz de covariância.

```
fit<-princomp(creme_dental_exemplo1,cor=TRUE)
fit
```

```
## Call:
## princomp(x = creme_dental_exemplo1, cor = TRUE)
##
## Standard deviations:
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
## 1.6526307 1.4893352 0.6645283 0.5841726 0.4273502 0.2919051
##
##  6  variables and  30 observations.
```

```
summary(fit)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
```

```
## Standard deviation     1.6526307 1.4893352 0.66452834 0.58417262
## Proportion of Variance 0.4551981 0.3696865 0.07359965 0.05687627
## Cumulative Proportion  0.4551981 0.8248846 0.89848425 0.95536053
##                            Comp.5     Comp.6
## Standard deviation     0.42735024 0.29190514
## Proportion of Variance 0.03043804 0.01420144
## Cumulative Proportion  0.98579856 1.00000000
```

A função `summary(fit)` mostra a aplicação da análise de componentes principais. O fator 1 responde por 45,52% da variância total. Da mesma forma, o segundo fator responde por 36,97% da variância total, sendo que os dois primeiros fatores respondem por 82,49% da variância total. Várias considerações devem integrar a análise do núumero de fatores que devem ser usados na análise.

2.3.3 Determinação do núumero de fatores

A fim de reduzir as informações presentes nas variáveis originais, deve-se reduzir o número de fatores. Na literatura, diversos processos são sugeridos: determinação a priori, observação dos autovalores, representação gráfica (**scree plot**), testes de significância entre outros.

2.3.3.1 Determinação a priori

Quando o pesquisador, com base na experiência que apresentação em relação ao assunto, decide quantos fatores deseja utilizar.

2.3.3.2 Autovalores

Como o autovalor representa a quantidade de variância associada ao fator, incluem-se apenas os fatores com variância maior que 1.

2.3.3.3 Gráfico de declive (scree plot)

Trata-se de uma representação gráfica dos autovalores associada ao número de fatores na ordem de extração. O ponto em que a inclinação suaviza indica o número de fatores a ser usados, que em geral é superior ao revelado pelos autovalores.

2.3.3.4 Percentagem da variância

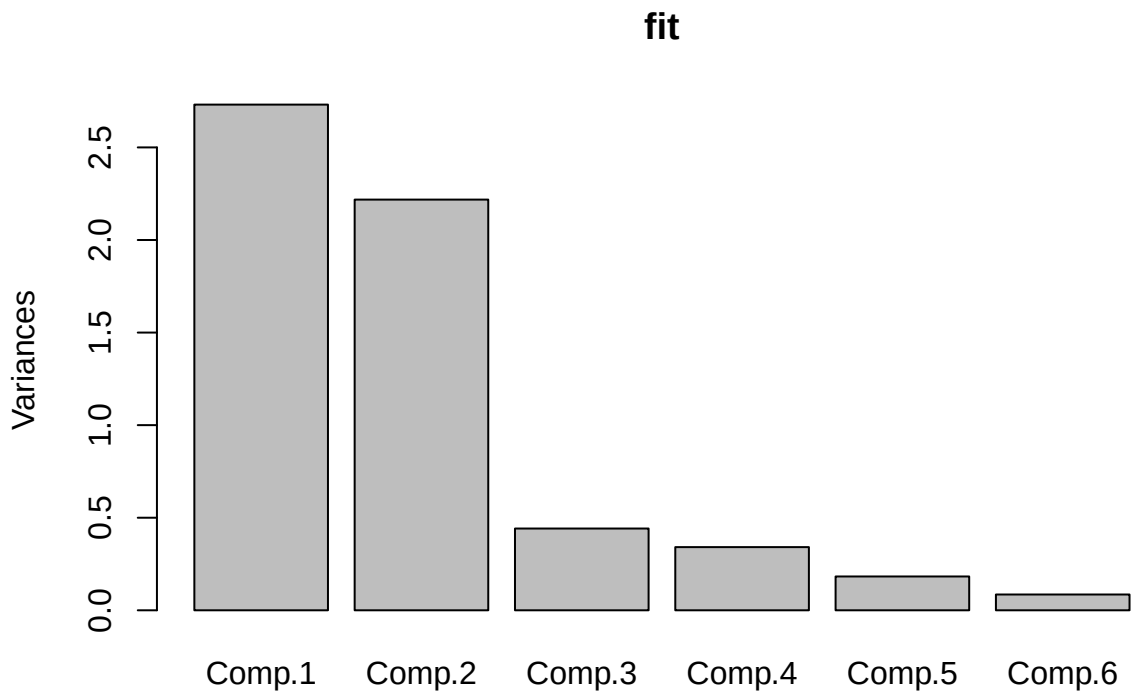
Determina que o núumero de fatores extraídos seja de no mínimo 60% da variância.

2.3.3.5 Teste de significância

É possível reter apenas os fatores estatisticamente significativos com base na significância estatística dos autovalores separados.

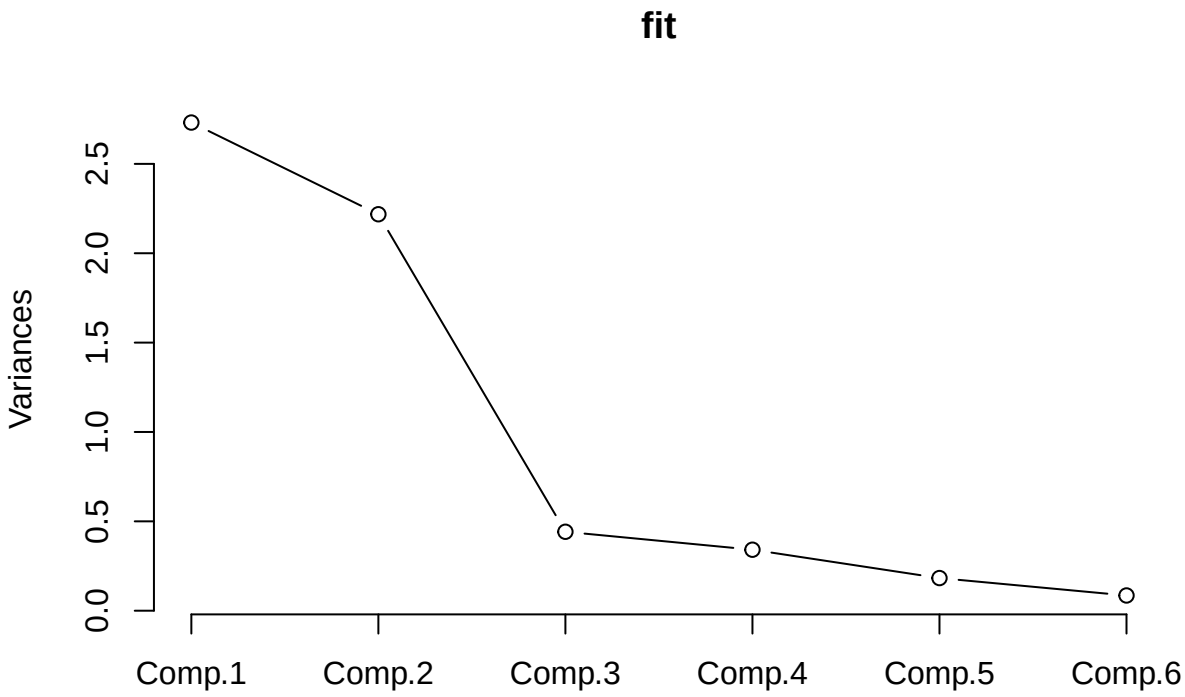
Abaixo vamos apresentar o `scree-plot`, em formato do gráfico de barras para o nosso exemplo

```
screeplot(fit)
```



Note que as duas primeiras componentes, aparecem em destaque, ocorrendo uma ligeira suavização das alturas nas demais colunas.

```
plot(fit,type="lines")
```



2.3.4 Análise de Componentes Principais

Rodando a Análise de Componentes Principais no R, temos:

```
PCAdente<-principal(creme_dental_exemplo1, nfactors=2,
                n.obs=30,rotate="none", scores=TRUE)
PCAdente
```

```
## Principal Components Analysis
## Call: principal(r = creme_dental_exemplo1, nfactors = 2, rotate = "none",
##     n.obs = 30, scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      PC1   PC2   h2    u2 com
## v1  0.93  0.25 0.93 0.074 1.1
## v2 -0.30  0.80 0.72 0.277 1.3
## v3  0.94  0.13 0.89 0.106 1.0
## v4 -0.34  0.79 0.74 0.261 1.4
## v5 -0.87 -0.35 0.88 0.122 1.3
```

```
## v6 -0.18  0.87 0.79 0.210 1.1
##
##                         PC1  PC2
## SS loadings            2.73 2.22
## Proportion Var         0.46 0.37
## Cumulative Var         0.46 0.82
## Proportion Explained   0.55 0.45
## Cumulative Proportion 0.55 1.00
##
## Mean item complexity =  1.2
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  3.94  with prob <  0.41
##
## Fit based upon off diagonal values = 0.98
```

A matriz de fatores acima, resultante da análise de componentes principais, é composta pelos coeficientes (cargas fatoriais) que expressam as variáveis padronizadas em termos dos fatores. Valores altos das cargas fatoriais, representam boa relação entre a variável e o fator. Essa matriz não rotada, apresenta dificuldades para ser interpretada pelo fato de que, em geral os fatores são correlacionados com muitas variáveis.

Com o processo da rotação, a matriz de fatores resulta numa matriz mais simples, sendo que a rotação não afeta as comunalidades e a porcentagem da variância explicada. No entanto, a percentagem da variância explicada por cada fator varia, sendo redistribuída por rotação (MALHOTRA, 2001).

Obs. comunalidades (*communalities*) são quantidades das variâncias (correlações) de cada variável explicada pelos fatores.

2.3.5 Matriz Rotada do Fator

Com o objetivo de possibilitar uma melhor interpretação dos fatores, é prática comum fazer uma rotação ou uma transformação dos fatores.

O conjunto de cargas fatoriais, obtidas por qualquer método de solução fatorial, quando o número de fatores comuns é maior do que um, não é único, pois outros conjuntos equivalentes podem ser encontrados, por transformações ortogonais de cargas.

Na rotação ortogonal, os eixos são mantidos em ângulo reto, sendo o método mais utilizado o processo varimax. Esse método ortogonal de rotação minimiza o número de variáveis com altas cargas sobre um fator afim de permitir a interpretaçã dos fatores. A rotação ortogonal resulta em fatores nãocorrelacionados ao passo que a rotação oblíqua não

mantém os eixos em ângulo reto e os fatores são correlacionados (MALHOTRA, 2001).

```
PCAdentevarimax<-principal(creme_dental_exemplo1, nfactors=2,
            n.obs=30,rotate="varimax",scores=TRUE)
PCAdentevarimax
```

```
## Principal Components Analysis
## Call: principal(r = creme_dental_exemplo1, nfactors = 2, rotate = "varimax",
##     n.obs = 30, scores = TRUE)
## Standardized loadings (pattern matrix) based upon correlation matrix
##      RC1   RC2   h2    u2 com
## v1  0.96 -0.03 0.93 0.074 1.0
## v2 -0.05  0.85 0.72 0.277 1.0
## v3  0.93 -0.15 0.89 0.106 1.1
## v4 -0.09  0.85 0.74 0.261 1.0
## v5 -0.93 -0.08 0.88 0.122 1.0
## v6  0.09  0.88 0.79 0.210 1.0
##
##                        RC1  RC2
## SS loadings           2.69 2.26
## Proportion Var        0.45 0.38
## Cumulative Var        0.45 0.82
## Proportion Explained  0.54 0.46
## Cumulative Proportion 0.54 1.00
##
## Mean item complexity =  1
## Test of the hypothesis that 2 components are sufficient.
##
## The root mean square of the residuals (RMSR) is  0.07
##  with the empirical chi square  3.94  with prob <  0.41
##
## Fit based upon off diagonal values = 0.98
```

Veja que na matriz rotada, o Fator 1 apresenta altos coeficientes para as variáveis V1 (prevenção de cáries), V3 (gengivas fortes) e coeficiente negativo para V5 (dentes sadios não é importante). O Fator 2 apresenta forte relação com V2 (clareie os dentes), V4 (hálito puro) e V6 (dentes atraentes).

Rotulando:

Nesta fase é usual tentar dar nomes aos fatores. Em muitos casos, isto requer um certo grau de imaginação:

Fator 1: Fator de benefício para a saúde.

Fator 2: Fator de benefício social.

Com os dois fatores acima, podemos concluir sobre o que o consumidor espera de um creme dental.

2.3.6 Autovalores

Para acessar os eingenvalues (autovalores):

```
PCAdentevarimax$values
```

```
## [1] 2.73118833 2.21811927 0.44159791 0.34125765 0.18262823 0.08520861
```

Confirmando, temos autovalores acima de 1, nos dois primeiros casos.

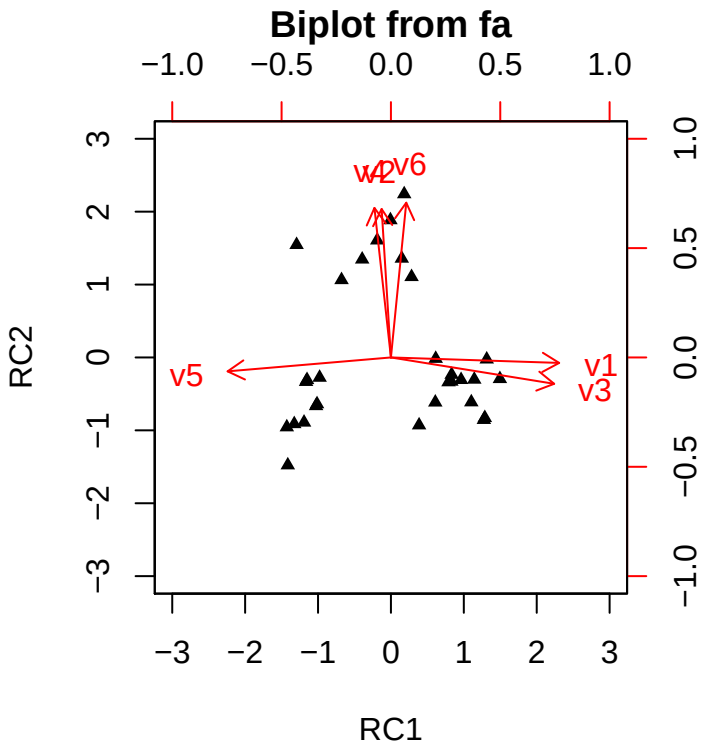
Para visualizar melhor a contribuição de cada variável (peso):

```
PCAdentevarimax$loadings
```

```
##
## Loadings:
##    RC1    RC2
## v1  0.962
## v2         0.848
## v3  0.933 -0.151
## v4         0.855
## v5 -0.934
## v6         0.885
##
##                  RC1   RC2
## SS loadings    2.687 2.263
## Proportion Var 0.448 0.377
## Cumulative Var 0.448 0.825
```

Recurso importante na interpretação dos fatores, o gráfico das variáveis, apresenta ao final do eixo, as variáveis que com cargas mais altas sobre aquele fator. Quanto mais próximas da origem menores as cargas destas variáveis sobre aquele fator. Variáveis distantes dos dois eixos, estão relacionadas a ambos os fatores.

```
biplot(PCAdentevarimax)
```



Os valores dos fatores obtidos para os 30 entrevistados encontram-se na matriz de coeficiente de escore do componente mostrada abaixo. Esta ajuda a entender como cada variável se relaciona aos escores dos componentes calculados para cada participante. Para melhor compreensão da análise dos escores dos entrevistados é importante especificar e comentar o significado de cada fator:

Fator 1: Fator de benefício para a saúde.

Fator 2: Fator de benefício social.

Analisando os escores fatoriais dos entrevistados, destacamos a seguir alguns entrevistados e seus respectivos resultados:

Entrevistado 18: 1.494934982

Este entrevistado se destacou como o primeiro colocado no ranqueamento, obtendo o maior escore ponderado, demonstrando ser bastante atento à prevenção de cáries, gengivas fortes e dentes sadios.

Entrevistado 29: 2.24121650

Este entrevistado se destacou em primeiro no segundo fator, apresentando preocupa-

ção quanto ao beneffício social da dentição: boa aparência dos dentes, hálito puro e boa aparência dos dentes.

Destacando-se os entrevistados de interesse, verifica-se:

```
 [8,]  1.314326185 -0.02535258
 [9,] -1.013720900 -0.63921247
[10,] -1.294148852  1.54533311
[11,]  1.102641284 -0.61319753
[12,] -1.150922200 -0.30734835
[13,]  1.288271583 -0.82866966
[14,]  0.148988842  1.35740692
[15,] -1.326348572 -0.91233215
[16,]  0.789822075 -0.33831055
[17,]  0.608638252 -0.61593673
[18,]  1.494934982 -0.29386303
[19,] -1.026914833 -0.66541567
[20,] -0.394466651  1.34560335
[21,] -1.192010691 -0.89125450
[22,]  0.614234778 -0.01676177
[23,] -0.978198322 -0.27595537
[24,] -0.006906942  1.88496785
[25,]  0.833064467 -0.23598682
[26,] -0.188090764  1.60734167
[27,]  0.828974634 -0.32986524
[28,] -0.677614532  1.06323436
[29,]  0.182192413  2.24121650
[30,] -1.428147525 -0.95604502
```

Capítulo 3

Regressão Múltipla

3.1 Modelo geral

Um modelo de regressão múltipla é expresso como:

$$y_i = \beta_0 + \ \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

em que:

- y_i: valores da variável resposta, $i = 1, 2, ..., n$ observações;
- x: valores das variáveis explicativas, $k = 1, 2, ..., K$ variáveis;
- β_k: parâmetros do modelo;
- ε_i: erro aleatório.

A equação estimada para este modelo é definida como:

$$y_i = b_0 + \ b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

em que:

- b_k: coeficientes estimados.

3.2 Variável dummy

Em algumas situações é necessário introduzir, como variável preditora (independente), uma variável categórica no modelo de regressão linear simples ou múltiplo, como por exemplo, local (urbano ou rural), área (preservada ou degradada), etc, podendo ter mais que duas categorias. Essa variável terá que ser codificada, utilizando somente códigos 0 e 1, assim chamada variável dummy.

O número de variáveis dummy no modelo será sempre igual ao número de categorias da variável preditora original menos 1. Por exemplo:

- Para a variável preditora "local" que assume valores - urbano ou rural, então têm-se a variável dummy local_dummy assumindo 0 para rural e 1 para urbano; também, poderia ser utilizado 1 para rural e 0 para urbano. Uma indicação é que a categoria que assume o valor 0 seja a categoria de referência.

- Para a variável preditora "grau de escolaridade"" que assume valores – ensino fundamental, ensino médio, ensino superior, então têm-se as variáveis dummy: escola1 e escola2, assim definido:

a. escola1=0 e escola2=0 para ensino fundamental;
b. escola1=1 e escola2=0 para ensino médio;
c. escola1=0 e escola2=1 para ensino superior.

Exercício:

1) Utilizando o banco de dados `ARVORE2`, ajuste um modelo de regressão linear simples para predizer a altura das árvores em função do diâmetro. Veja essa relação no diagrama de dispersão. Interprete os resultados.

Relembrando Modelos de Regressão Linear Simples – Curso Básico do Software R:

- 1.1 Ajustar a equação de regressão. Interpretá-la.
- 1.2 Encontrar e interpretar a significância da equação.
- 1.3 Encontrar e interpretar o coeficiente de determinação.
- 1.4 Analisar graficamente os resíduos.
- 1.5 Testar a normalidade dos resíduos.

Adicionalmente - Curso Avançado do Software R:

- 1.6 Analisar pontos *outliers* nos resíduos.

Para análise dos valores *outliers* nos resíduos (*residuals standard* e *residuals studentized*), utilizam-se os seguintes comandos:

```
rstudent(regressao)
```

`rstandard(regressao)`

E o gráfico para verificar valores outliers nos resíduos:

`plot(rstudent(regressao))`

`plot(rstandard(regressao))`

Aqueles valores maiores que |2| são possíveis outliers. Incluir uma linha y =2 e y=-2, para facilitar a visualização de outliers.

- 1.7 Analisar pontos influentes nos resíduos.

Para análise dos valores influentes, utiliza-se:

`dffits(regressao)`

Aqueles valores maiores que `2*(p/n)^(1/2)` são possíveis pontos influentes. Em que, p = número de parâmetros do modelo e n = tamanho da amostra. O gráfico para detectar pontos influentes pode ser elaborado pelo comando:

`plot(dffits(regressao))`

Aqueles valores maiores, em módulo, são possíveis influentes. Incluir linhas para facilitar a visualização de pontos influentes.

Ainda, pode-se utilizar o comando `plot(regressao)` elabora diferentes gráficos para o diagnóstico do modelo.

2) Ajuste um segundo modelo de regressão linear simples para predizer a altura das árvores em função da espécie. Veja essa relação no diagrama de dispersão. Interprete os resultados.

3) Ajuste um terceiro modelo de regressão múltipla para predizer a altura das árvores em função do diâmetro e da espécie. Interprete os resultados.

```
library(readxl)
url <- "https://github.com/Smolski/softwarelivrer/raw/master/avancado/arvore2.xlsx"
destfile <- "arvore2.xlsx"
curl::curl_download(url, destfile)
arvore2 <- read_excel(destfile)
attach(arvore2)
head(arvore2)
```

```
## # A tibble: 6 x 4
##   Nomecientifico            diametro_cm altura_m especie
##   <chr>                           <dbl>    <dbl>   <dbl>
## 1 Sebastiania commersoniana        52.2     15.2       0
```

```
## 2 Sebastiania commersoniana        95       17.3       0
## 3 Sebastiania commersoniana        67.3     16.3       0
## 4 Sebastiania commersoniana        46.3     14         0
## 5 Sebastiania commersoniana        64.1     15         0
## 6 Sebastiania commersoniana       122       22         0
```

```
modelom=lm(altura_m~diametro_cm+especie)
modelom
```

```
##
## Call:
## lm(formula = altura_m ~ diametro_cm + especie)
##
## Coefficients:
## (Intercept)  diametro_cm      especie
##    12.69592      0.05713     -1.62517
```

Modelo:

$$Y = 12,328 + 0,0576x_1 1,423x_2$$

Ou

$$\text{Altura} = 12,328 + 0,0576\text{diâmetro} 1,423\text{espécie}$$

Verificando a significância de cada coeficiente do modelo de regressão múltipla:

```
summary(modelom)
```

```
##
## Call:
## lm(formula = altura_m ~ diametro_cm + especie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2688 -0.7663 -0.1236  0.8132  2.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.69592    0.38639  32.857  < 2e-16 ***
## diametro_cm  0.05713    0.00445  12.837  < 2e-16 ***
## especie     -1.62517    0.24459  -6.644 1.52e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.185 on 102 degrees of freedom
## Multiple R-squared:  0.6995, Adjusted R-squared:  0.6937
```

```
## F-statistic: 118.7 on 2 and 102 DF,  p-value: < 2.2e-16
```

Verificar a significância do modelo completo.

Verificar o coeficiente de determinação do modelo.

Realizar análise dos resíduos.

- gráfico dos resíduos com cada variável preditora
- resíduos padronizados para verificar outlier
- verificar pontos infuentes

A interpretação dos termos de regressão é um pouco mais complicada. Em geral, um modelo com múltiplos preditores indica a diferença média na variável desfecho quando mudamos o valor de uma variável e mantemos a outra constante.

Nesse caso, entre árvores de mesmo diâmetro (x_1), a diferença média esperada da altura (y) para a espécie *Syphoneugena reitzii* em relação a espécie *Sebastiania* commersoniana é de cerca de 1,42m a menos (pois b_3=-1,42).

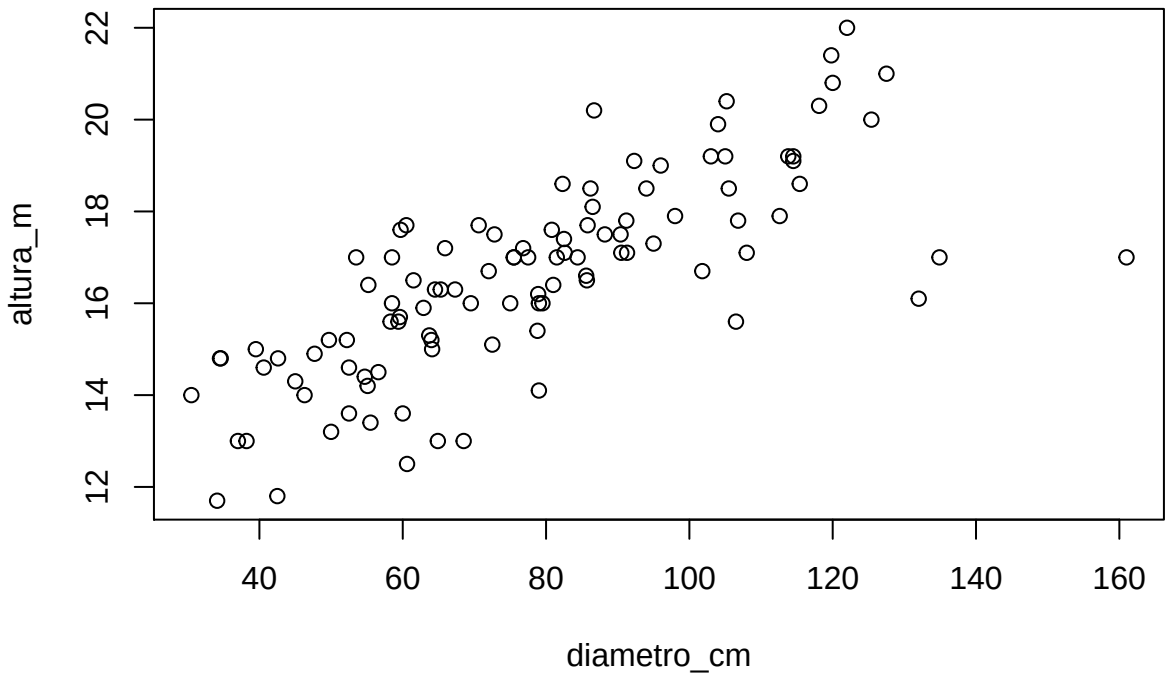
Da mesma forma, árvores da mesma espécie têm, em média, 0,05758m (pois b_2=0,05758) a mais a cada 1 cm de diâmetro.

Como envolvem mais variáveis, não é possível resolver o modelo inteiro num único gráfico. Como alternativa, pode-se plotar a reta para cada espécie (variável categórica).

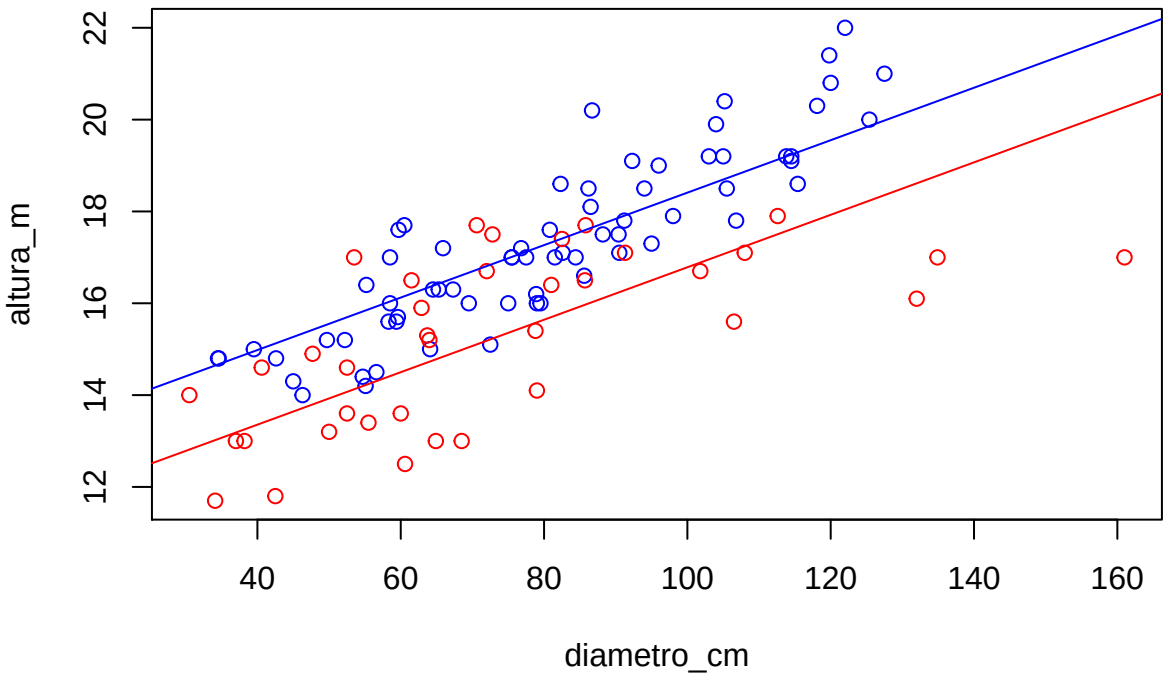
Primeiro, os pontos são plotados. O argumento `type='n'` indica que não é para acrescentar nenhum ponto ao gráfico. Em seguida, os pontos são acrescentados separadamente, com a função points, a qual acrescenta pontos ao gráfico, sendo que o colchetes `[espécie==0]` seleciona somente os casos desejados.

Por fim, acrescentamos as retas de regressão para cada resposta a variável independente espécie. Usamos a função `coef` para extrair os coeficientes de interesse.

```
plot(diametro_cm,altura_m)
```




```
# Gera o gráfico sem pontos
plot(diametro_cm,altura_m,type='n')
# Acrescenta os pontos
points(diametro_cm[especie==0],altura_m[especie==0],col='blue')
points(diametro_cm[especie==1],altura_m[especie==1],col='red')
# Acrescenta as linhas
abline(coef(modelom)[1], coef(modelom)[2], col='blue')
abline(coef(modelom)[1]+coef(modelom)[3], coef(modelom)[2], col='red')
```



3.3 Métodos seleção de variáveis na regressão múltipla

3.3.1 Full model – Modelo completo

Sintaxe no software R para um modelo de regressão múltipla com três variáveis preditivas:

```
regressao=lm(y~x1+x2+x3)

summary(regressao)
```

Existem três métodos de seleção de variáveis para modelos de regressão múltipla: *backward*, *forward* e *stepwise*.

```
regressao=step(lm(y~x1+x2+x3),direction = 'método')
```

3.3.2 Procedimento backward

Considera todas as variáveis inicialmente, testando posteriormente, a permanência de cada uma no modelo. Se p $\leq$ 15%, permanece no modelo (saiu do modelo não entra mais) (Riboldi, 2005).

Passo 1) Ajustar o modelo completo de m variáveis e obter SQR^c_{eg} e σ^2;

Passo 2) Para cada uma das m variáveis do modelo completo do passo 1, considerar o modelo reduzido – retirando esta variável – e calcular SQR^r_{eg} para obter o valor da estatística (slide 24);

Passo 3) Achar o mínimo dos m valores da estatística obtidos no passo 2, denotado por F_{min};

Passo 4) Seja F_{out} o valor da distribuição F com 1 e (n-m-1) gl;

- Se $F_{min} > F_{out}$: interromper o processo e optar pelo modelo completo desta etapa;
- Se $F_{min} < F_{out}$: voltar ao passo 1, iniciando nova etapa em que o modelo completo tem (m-1) variáveis – dada a eliminação da variável cuja estatística é igual a F_{min}.

3.3.3 Procedimento forward

Inclui uma variável de cada vez, se p $\leq$ 20%, entra no modelo. Este método não testa a permanência da variável (entrou no modelo não sai mais) (Riboldi, 2005).

Passo 1) Ajustar o modelo reduzido de m variáveis e obter SQR^c_{eg};

Passo 2) Para cada variável não pertencente ao modelo do passo 1, considerar o modelo completo com adição desta variável extra e calcular SQR^r_{eg} e σ^2 para obter o valor da estatística (slide 26);

Passo 3) Achar o máximo dos valores da estatística obtidos no passo 2, denotado por F_{max};

Passo 4) Seja F_{in} o valor da distribuição F com 1 e (n-m) gl;

- Se $F_{max} > F_{in}$: voltar ao passo 1, iniciando nova etapa em que o modelo reduzido tem (m+1) variáveis – dada a inclusão da variável cuja estatística é igual a F_{max}.
- Se $F_{max} < F_{in}$: interromper o processo e optar pelo modelo reduzido desta etapa;

3.3.4 Procedimento stepwise

Inclui as variáveis passo-a-passo e testa a permanência (as variáveis podem entrar e sair do modelo) (Riboldi, 2005).

Passo 1) Ajustar o modelo reduzido de m variáveis e obter SQR^r_{eg};

Passo 2) Para cada variável não pertencente ao modelo do passo 1, considerar o modelo completo - com adição desta variável extra - e calcular SQR^c_{eg} e σ^2 para obter o valor da estatística (slide 26);

Passo 3) Achar o máximo dos valores da estatística obtidos no passo 2, denotado por F_{max};

Passo 4) Seja Fin o valor da distribuição F com 1 e (n-m) gl;

- Se Fmax > Fin -> passar ao passo 5, com modelo completo composto por (m+1) variáveis – as m variáveis do modelo do passo 1 e a variável cuja estatística é igual a Fmax.

- Se Fmax < Fin -> passar ao passo 5, com modelo completo igual ao modelo do passo 1 ou encerrar o processo se no passo 8 da etapa anterior, nenhuma variável tiver sido eliminada;

Passo 5) Ajustar o modelo completo de k variáveis – sendo k igual a m ou (m+1), e obter SQR^c_{eg} e σ^2;

Passo 6) Para cada uma das k variáveis do modelo completo do passo 5, considerar o modelo reduzido – retirando esta variável – e calcular SQR^r_{eg} para obter o valor da estatística;

Passo 7) Achar o mínimo dos k valores da estatística obtidos no passo 6, denotado por F_{min};

Passo 8) Seja F_{out} o valor da distribuição F com 1 e (n-k-1) gl;

- Se $F_{min} > F_{out}$: não eliminar nenhuma variável e voltar ao passo 1, iniciando nova etapa com modelo reduzido com k variáveis ou encerrar o processo de no passo 4 nenhuma variável tiver sido anexada;

- Se Fmin < Fout: eliminar a variável cuja estatística é igual a Fmin e voltar ao passo 1 iniciando nova etapa com modelo reduzido com (k-1) variáveis.

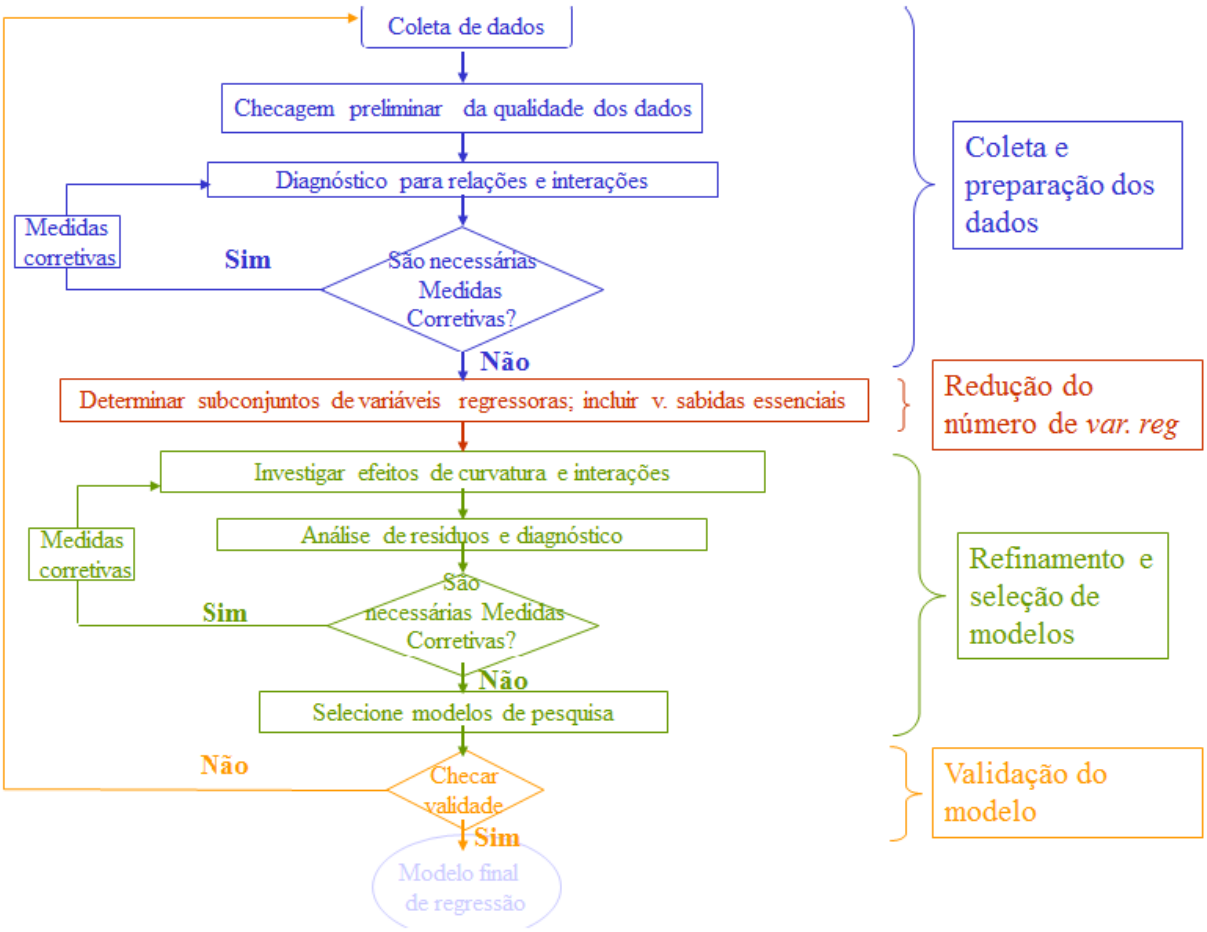


Figura 3.1: Modelagem estatística

Capítulo 4

Regressão Logística

A técnica de regressão logística é uma ferramenta estatística utilizada nas análises preditivas. O interesse em mensurar a probabilidade de um evento ocorrer é extremamente relevante em diversas áreas, como por exemplo em Marketing, Propaganda e Internet, na Aplicação da Lei e Detecção de Fraude, na Assistência Médica, com relação aos Riscos Financeiros e Seguros ou mesmo estudando a Força de Trabalho. É imprescindível elevar o conhecimento sobre quais clientes possuem maior propensão à responder o contato de marketing, quais transações serão fraudulentas, quais e-mails são *spam*, quem efetivamente fará o pagamento de uma obrigação ou mesmo qual criminoso reincidirá.[1]

4.1 O modelo

O modelo de regressão logística é utilizado quando a variável dependente é binária, categórica ordenada ou mesmo categórica desordenada (quando não há relação hierárquica entre elas). Abaixo exemplificam-se algumas perguntas que podem levar a estes três tipos de variáveis.

Tabela 4.1: Tipos de variáveis

Variável dependente binária:	Você votou na última eleição?	0 - Não; 1 - Sim
Variável dependente categórica ordenada:	Você concorda ou desconcorda com o presidente?	1 - Disconcordo; 2 - Neutro; 3 - Concordo
Variável dependente categórica não ordenada:	Se as eleições fossem hoje, em que partido você votaria?	1 - Democratas; 2 - Qualquer um; 3 - Republicanos

[1]Para mais exemplos como estes sobre análises preditivas, ver SIEGEL (2017).

Fonte: Adaptado de TORRES-REYNA (2014).

Nota-se primeiramente que em sendo somente a variável dependente **binária** (0 e 1), é detectada a presença ou não de determinada característica da variável a ser estudada pelo pesquisador. Outros exemplos abrangem a qualificação dos indivíduos estaudados em sendo do sexo feminino (1) ou do sexo masculino (0), se a empresa analisada está inadimplente (1) ou não (0) no mês de referência, etc. Por outro lado, quando a variável dependente é **categórica ordenada**, há uma hierarquia determinada entre as variáveis resposta (neste caso entre Disconcordo, Neutro e Concordo). No terceiro exemplo, a variável resposta é **categórica não ordenada** não possuindo nenhuma relação de ordem entre elas (Democratas, Qualquer um, Republicanos).

A regressão logística a ser estudada neste capítulo será com a variável resposta dependente binária, portanto, tratando os grupos de interesse (variável dependente) com valores de 0 e 1. Sua funcionalidade se ocupa de prever a probabilidade de uma observação estar no grupo igual a 1 ("eventos"), em relação ao grupo igual a zero ("não eventos").

Para a estimação dos coeficientes das variáveis independentes, são utilizados o valor logit ou a razão de desigualdades (HAIR et al., 2009):

$$Logit_i = ln\left(\frac{prob_{eventos}}{1 - prob_{eventos}}\right) = b_0 + b_1X_1 + \ldots + b_nX_n$$

ou

$$Logit_i = \left(\frac{prob_{eventos}}{1 - prob_{eventos}}\right) = e^{b_0 + b_1X_1 + \ldots + b_nX_n}$$

Algumas características importantes da regressão logística: a análise é semelhante à regressão linear simples/múltipla (possui a relação entre a variável dependente e a(s) variável(is) independente(s)); possui testes estatísticos diretos, incorporando variáveis métricas e não-métricas, com efeitos não-lineares; é menos afetada pela não satisfação de normalidade dos dados (pois o termo de erro da variável discreta segue a distribuição binomial) e; foi elaborada para que seja prevista a probabilidade de determinado evento ocorrer (HAIR et al., 2009).

Para otimizar o tempo do estudante, é recomendada a instalação prévia dos pacotes no RStudio a serem utilizados neste capítulo. Segue abaixo o comando a ser efetuado no console do RStudio:

```
install.packages(c("readr","mfx","caret","pRoc","ResourceSelection","modEvA","fore
```

4.2 Regressão Logística Simples

Este primeiro exemplo tratará da regressão logística simples, portanto, utilizando somente uma variável independente, neste caso numérica. Os dados são originados do livro de HOSMER; LEMESCHOW (2000), tratando-se de uma amostra com 100 pessoas. A variável dependente é a ocorrência ou não (1 ou 0) de doença coronária cardíaca (CHD), associando-se com a idade (AGE) dos indivíduos, criando assim um modelo de regressão logística.

```
require(readr)
```

```
## Carregando pacotes exigidos: readr
```

```
chd <- read_delim("https://goo.gl/uDAAHv",
    ";", escape_double = FALSE, col_types = cols(CHD = col_factor(levels = c())),
    trim_ws = TRUE)

summary(chd)
```

```
##       AGE             AGRP      CHD
##  Min.   :20.00   Min.   :1.00   0:57
##  1st Qu.:34.75   1st Qu.:2.75   1:43
##  Median :44.00   Median :4.00
##  Mean   :44.38   Mean   :4.48
##  3rd Qu.:55.00   3rd Qu.:7.00
##  Max.   :69.00   Max.   :8.00
```

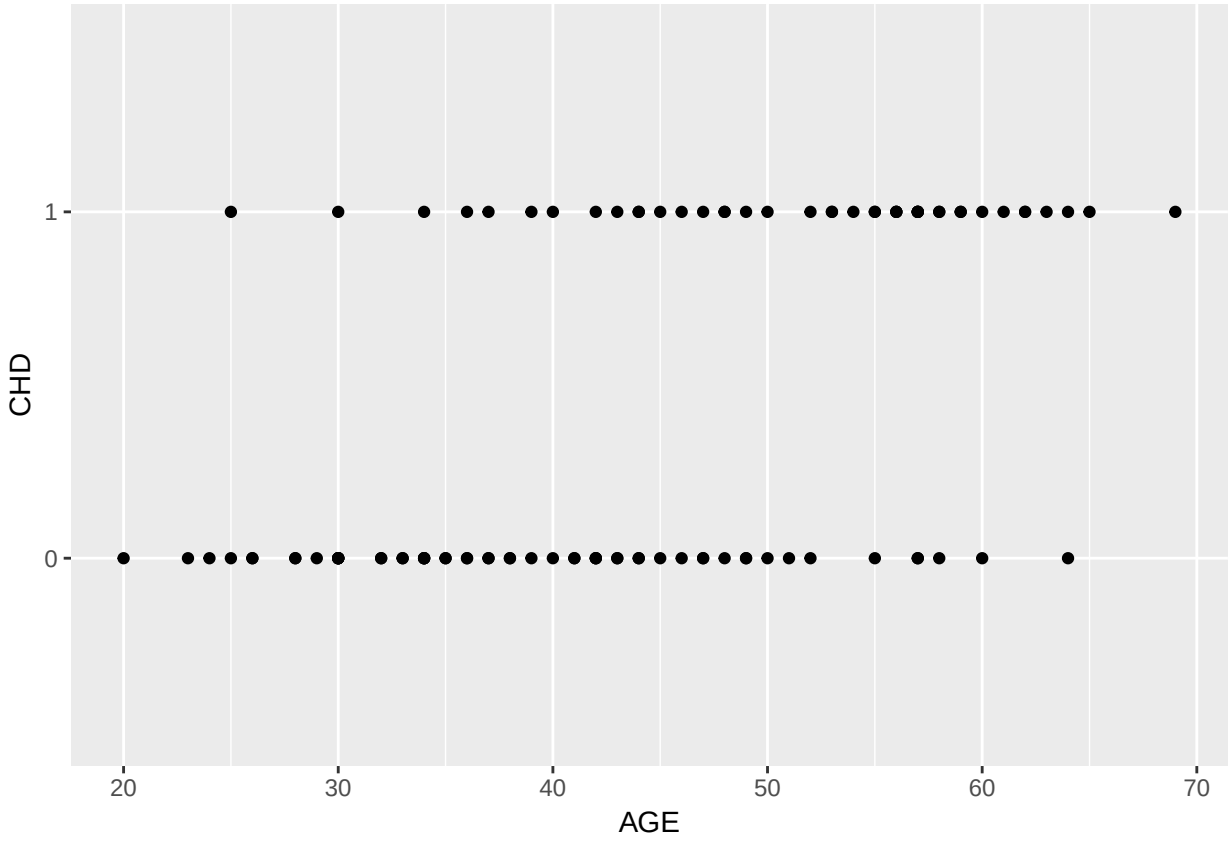
Observa-se na figura abaixo a dispersão dos "eventos" e dos "nao-eventos" da CHD relacionando-se com a variável idade (AGE).

```
require(ggplot2)
```

```
## Carregando pacotes exigidos: ggplot2
```

```
ggplot(chd, aes(x=AGE, y=CHD)) +
  geom_point() +
  stat_smooth(method="glm", method.args=list(family="binomial"), se=FALSE)
```

```
## Warning: Computation failed in `stat_smooth()`:
## y values must be 0 <= y <= 1
```



Monta-se então o modelo de regressão logística com a variável dependente CHD e a variável independente AGE. Abaixo é demonstrada a descrição da equação utilizando o comando `summary()` para o modelo m1 com a sintaxe básica:

```
glm(Y~modelo, family=binomial(link="logit"))
```

Assim é obtida a função de ligação estimada do modelo:

$$\hat{g}(CHD) = -5,309 + 0,1109AGE$$

```
m1=glm(CHD~AGE, family = binomial(link="logit"), data = chd)
summary(m1)
```

```
##
## Call:
## glm(formula = CHD ~ AGE, family = binomial(link = "logit"), data = chd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9718  -0.8456  -0.4576   0.8253   2.2859
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
## AGE          0.11092    0.02406   4.610 4.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 136.66  on 99  degrees of freedom
## Residual deviance: 107.35  on 98  degrees of freedom
## AIC: 111.35
##
## Number of Fisher Scoring iterations: 4
```

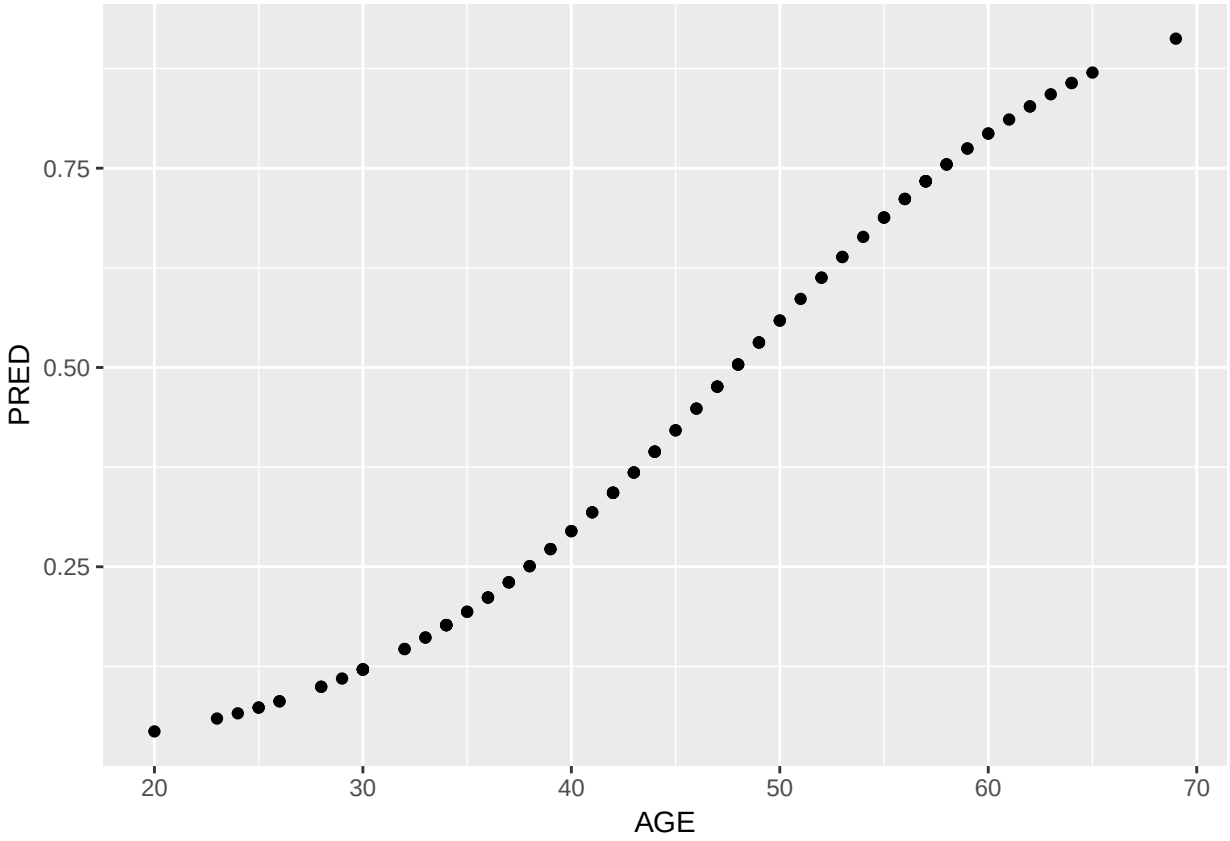
Se observa o intercepto com o valor de -5,309, sendo que para a análise aqui proposta da relação entre CHD e AGE não obtém-se um significado prático para este resultado. No entanto, a variável de interesse é idade, que no modelo de regressão obteve o coeficiente de 0,1109. Pelo fato de ser positivo informa que quando a idade (AGE) se eleva, elevam-se as chances de ocorrência de CHD. De igual forma, nota-se que há significância estatística a $p = 0,001$ na utilização da variável AGE para o modelo, mostrando que possui importância ao modelo de regressão proposto.

Por fim, o modelo é utilizado para construção da predição de todos os valores das idades de todos os indivíduos desta amostra. Para isto, será criada um novo objeto contendo somente a variável dependente do modelo (AGE) e em sequida, é criada nova coluna constando os valores preditos. Assim, pode ser plotado um gráfico completo com todas as probabilidades desta base de dados:

```r
# Filtrando a idade dos indivíduos
IDADE<-chd[,1]

# Criando campo de predição para cada idade dos indivíduos
IDADE$PRED=predict(m1, newdata=IDADE, type="response")

# Plotando a probabilidade predita pelo modelo
require(ggplot2)
ggplot(IDADE, aes(x=AGE, y=PRED)) +
  geom_point()
```



4.2.1 Estimando a Razão de Chances

O modelo de regressão logística, porém, traz os resultados dos estimadores na forma logarítma, ou seja, o log das chances da variável idade no modelo é 0,1109. No entanto, para uma interpretação mais enriquecida da relação da idade com o CHD é necessária a transformação deste coeficiente, ou seja, que seja efetuada a exponenciação da(s) variavel(eis) da regressão. Assim, obtém-se a razão das chances (OR - Odds Ratio em inglês) para as variáveis independentes.

Uma maneira prática de se obter a razão de chances no RStudio é utilizando o pacote mfx. Novamente o intercepto não nos interessa nesta análise mas sim a variável AGE. Como demonstrado abaixo, o resultado da razão de chances da variável AGE foi de 1,1173, o que pode assim ser interpretado: para cada variação unitária na idade (AGE), as chances de ocorrência de CHD aumentam 1,1173 vezes. Dito de outra forma, para cada variação unitária em AGE, aumentam-se 11,73% ((1,1173-1)*100) as chances da ocorrência de CHD.

```
require(mfx)
logitor(CHD~AGE,data = chd)
```

```
## Call:
```

```
## logitor(formula = CHD ~ AGE, data = chd)
##
## Odds Ratio:
##     OddsRatio Std. Err.      z     P>|z|
## AGE  1.117307  0.026882 4.6102 4.022e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.2.2 Determinando o Intervalo de Confiança

A determinação do intervalo de confiança do modelo proposto é relevante para que seja analizada a estimativa do intervalo de predição do coeficiente da variável independente, a um nível de confiança de 95%. Desta forma, em 95% dos casos, o parâmetro dos coeficientes estará dentro deste intervalo.

De forma prática é possível determinar os intervalos de confiança com o comando `confint()` commo observado abaixo, sendo que o coeficiente AGE toma o valor de 1,1173, podendo variar de 1,0692 a 1,1758.

```
exp(cbind(OR=coef(m1), confint(m1)))
```

```
##                       OR        2.5 %    97.5 %
## (Intercept) 0.004944629 0.0004412621 0.0389236
## AGE         1.117306795 1.0692223156 1.1758681
```

4.2.3 Predição de Probabilidades

A partir dos coneficientes do modelo de regressão logística é possível, portanto, efetuar a predição da variável categórica CHD, ou seja, saber a chance de ocorrer CHD com relação à uma determinada idade (AGE). No exemplo abaixo, primeiramente utilizamos a idade média das observações (44,38 anos), criando assim um novo data.frame chamadio media. Para utilizar o valor da idade média na função de regressão obtida ($m1$), utiliza-se a função `predict()`, de acordo com valor da média encontrada (data.frame media). O resultado mostra que para a idade média da amostra, 44,38 anos, há uma probabilidade de 40,44% na ocorrência da doença CHD. Esta ferramenta permite também a comparação pelo pesquisador das diferentes probabilidades entre as diversas idades (variável AGE).

```
media = data.frame(AGE=mean(chd$AGE))
media
```

```
##     AGE
## 1 44.38
```

```
media$pred.prob = predict(m1, newdata=media, type="response")
media
```

```
##     AGE pred.prob
## 1 44.38 0.4044944
```

4.2.4 Matriz de Confusão

Uma maneira prática de qualificar o ajuste do modelo de regressão logística é pela projeção do modelo na tabela de classificação (ou Matriz de Confusão). Para isto, precisa-se criar uma tabela com o resultado da classificação cruzada da variável resposta, de acordo com uma variável dicotômica em que os valores se derivam das probabilidades logísticas estimadas na regressão (HOSMER; LEMESCHOW, 2000). No entanto, é preciso definir uma regra de predição, que dirá se houve acerto ou não da probabilidade estimada com os valores reais, pois as probabilidades variam de 0 a 1 enquanto os valores reais binários possuem valores fixos de 0 "ou" 1.

É intuitivo supor que se as probabilidades aproximam-se de 1 o índivíduo estimado pode ser classificado como $\hat{Y}_i = 1$, bem como de forma contrária, se o modelo estimar probabilidades perto de 0, classificá-la como $\hat{Y}_i = 0$. Mas qual nível utilizar? Para resolver este problema, é preciso em primeiro lugar determinar um ponto de corte para classificar a estimação como 0 ou 1. Usualmente na literatura se utiliza o valor de 0,5 mas dependendo do estudo proposto pode não ser limitado a este nível (HOSMER; LEMESCHOW, 2000).

Após determinado o ponto de corte, é importante avaliar o poder de discriminação do modelo, pelo seu desempenho portanto em classificar os "eventos" dos "não eventos". Cria-se a Matriz de Confusão (vide Tabela xxx) com as observações de Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN) e em seguida determinam-se alguns parâmetros numéricos, a serem descritos abaixo:

Precisão: representa a proporção das predições corretas do modelo sobre o total:

$$ACC = \frac{VP + VN}{P + N}$$

onde P representa o total de "eventos" positivos (Y=1) e N é o total de "não eventos" (Y=0, ou negativo).

Sensibilidade: representa a proportação de verdadeiros positivos, ou seja, a capacidade do modelo em avaliar o evento como $\hat{Y} = 1$ (estimado) dado que ele é evento real $Y = 1$:

		Valor Observado	
		Y=1	Y=0
Valor Estimado	$\hat{Y}_i = 1$	VP	FP
	$\hat{Y}_i = 0$	FN	VN

Figura 4.1: Matriz de Confusão

$$SENS = \frac{VP}{FN}$$

Especificidade: a proporção apresentada dos verdadeiros negativos, ou seja, o poder de predição do modelo em avaliar como "não evento" $\hat{Y} = 0$ sendo que ele não é evento $Y = 0$:

$$SENS = \frac{VN}{VN + FP}$$

Verdadeiro Preditivo Positivo: se caracteriza como proporção de verdadeiros positivos com relação ao total de predições positivas, ou seja, se o evento é real $Y = 1$ dada a classificação do modelo $\hat{Y} = 1$:

$$VPP = \frac{VPP}{VN + FP}$$

Verdadeiro Preditivo Negativo: se caracteriza pela proporção de verdadeiros negativos comparando-se com o total de predições negativas, ou seja, o indivíduo não ser evento $Y = 0$ dada classificação do modelo como "não evento" $\hat{Y} = 0$:

$$VPN = \frac{VN}{VN + FN}$$

Fonte: Adaptado de FAWCETT (2006).

```
require(caret)
```

```
## Carregando pacotes exigidos: caret
```

```
## Carregando pacotes exigidos: lattice
```

```
  pdata <- as.factor(
    ifelse(
      predict(m1, newdata = chd, type = "response")
      >0.5,"1","0"))

confusionMatrix(pdata, chd$CHD, positive="1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 45 14
##          1 12 29
##
##                Accuracy : 0.74
##                  95% CI : (0.6427, 0.8226)
##     No Information Rate : 0.57
##     P-Value [Acc > NIR] : 0.0003187
##
##                   Kappa : 0.4666
##  Mcnemar's Test P-Value : 0.8445193
##
##             Sensitivity : 0.6744
##             Specificity : 0.7895
##          Pos Pred Value : 0.7073
##          Neg Pred Value : 0.7627
##              Prevalence : 0.4300
##          Detection Rate : 0.2900
##    Detection Prevalence : 0.4100
##       Balanced Accuracy : 0.7319
##
##        'Positive' Class : 1
##
```

4.2.5 Curva ROC

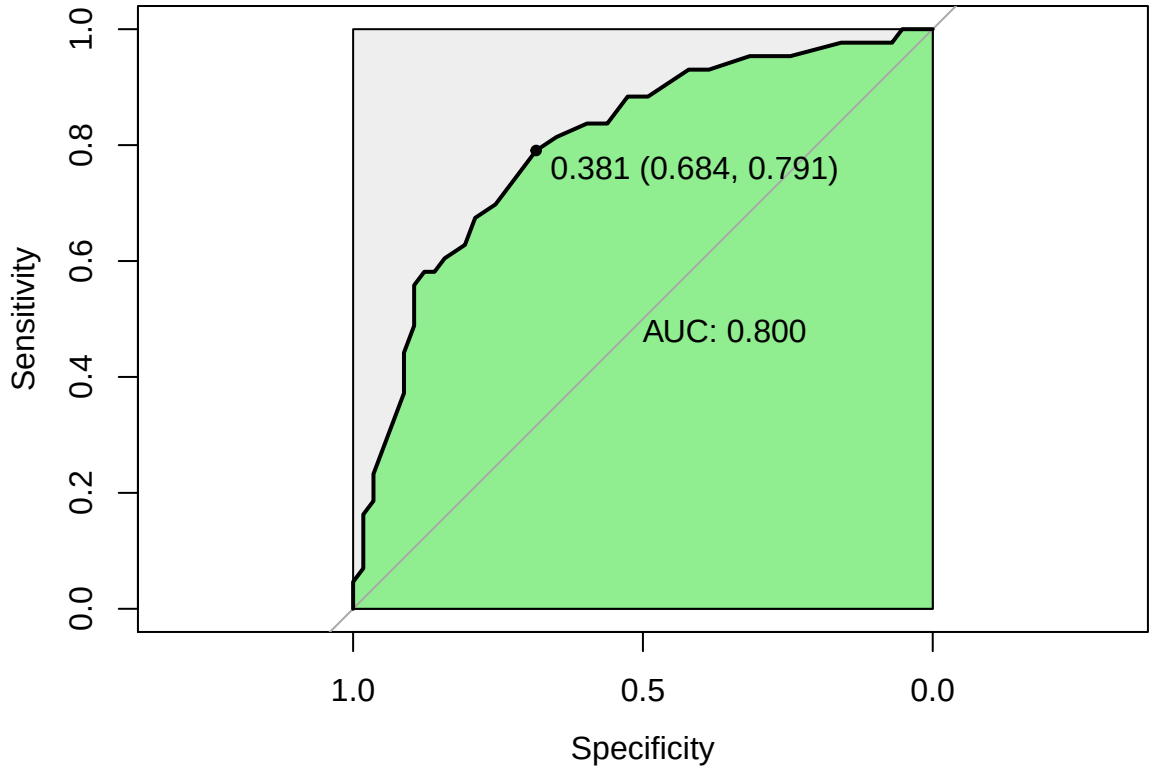
A Curva ROC (Receiver Operating Characteristic Curve) associada ao modelo logístico mensura a capacidade de predição do modelo proposto, através das predições da sensibilidade e da especificidade.

- Passo 1:

```
require(pROC) roc1=plot.roc(chd$CHD,fitted(m1))
```

- Passo 2:

```
plot(roc1,
     print.auc=TRUE,
     auc.polygon=TRUE,
     grud=c(0.1,0.2),
     grid.col=c("green","red"),
     max.auc.polygon=TRUE,
     auc.polygon.col="lightgreen",
     print.thres=TRUE)
```



4.2.6 O teste Hosmer e Lemeshow

```
require(ResourceSelection)
hl=hoslem.test(chd$CHD,fitted(m1),g=10)
hl
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
```

```
## data:  chd$CHD, fitted(m1)
## X-squared = 100, df = 8, p-value < 2.2e-16
```

4.2.7 Pseudo R^{2}

```
require(modEvA)
```

```
## Carregando pacotes exigidos: modEvA
```

```
RsqGLM(m1)
```

```
## $CoxSnell
## [1] 0.2540516
##
## $Nagelkerke
## [1] 0.3409928
##
## $McFadden
## [1] 0.2144684
##
## $Tjur
## [1] 0.2705749
##
## $sqPearson
## [1] 0.2725518
```

4.3 Regressão Logística Múltipla

O exemplo abaixo abordado foi extraído de TORRES-REYNA (2014), onde observa-se o banco de dados criado chamado `mydata`, possuindo as variáveis`country`,`year`,`y`,`y_bin`,`x1`,`x2`,`x3` e`opinion`. A variável dependente é `y_bin`, da qual foi categorizada entre 0 e 1 conforme a ocorrência de valores negativos em`y`. As variáveis independentes do modelo serão`x1`,`x2`e`x3`.

```
require(foreign)
```

```
## Carregando pacotes exigidos: foreign
```

```
mydata <- read.dta("http://dss.princeton.edu/training/Panel101.dta")
summary(mydata)
```

```
##  country      year            y                  y_bin
##  A:10    Min.   :1990   Min.   :-7.863e+09   Min.   :0.0
```

```
##  B:10    1st Qu.:1992   1st Qu.: 2.466e+08   1st Qu.:1.0
##  C:10    Median :1994   Median : 1.898e+09   Median :1.0
##  D:10    Mean   :1994   Mean   : 1.845e+09   Mean   :0.8
##  E:10    3rd Qu.:1997   3rd Qu.: 3.372e+09   3rd Qu.:1.0
##  F:10    Max.   :1999   Max.   : 8.941e+09   Max.   :1.0
##  G:10
##        x1                x2                x3                opinion
##  Min.   :-0.5676   Min.   :-1.6218   Min.   :-1.16539   Str agree:20
##  1st Qu.: 0.3290   1st Qu.:-1.2156   1st Qu.:-0.07931   Agree    :15
##  Median : 0.6413   Median :-0.4621   Median : 0.51419   Disag    :19
##  Mean   : 0.6480   Mean   : 0.1339   Mean   : 0.76185   Str disag:16
##  3rd Qu.: 1.0958   3rd Qu.: 1.6078   3rd Qu.: 1.15486
##  Max.   : 1.4464   Max.   : 2.5303   Max.   : 7.16892
##
```

Utiliza-se uma função para Modelos Lineares Generalizados (glm - em inglês Generalized Linear Models), determinando a variável dependente (y_bin), as variáveis independentes (`x1+x2+x3`), a base de dados a ser utilizada (`data=mydata`) e a família dos modelos (`family = binomial(link="logit")`).

Abaixo os resultados da estimação do modelo utilizando o comando `summary`. Observa-se que os valores `estimados` mostram os coeficientes em formato logarítmo de chances. Assim, quando x3 eleva-se em 1 (uma) unidade, o log das chances esperado para x3 altera-se em 0,7512. Neste ponto, observa-se que as três variáveis independentes possuem efeitos positivos para determinação das chances do preditor ser igual a 1, caso contrário constariam com sinal negativo. A coluna $Pr(>|z|)$ traz os p-valores das variáveis indicando o teste da hipótese nula. Como resultado a variável x3 revelou significância estatística a 10% ($<$0,10), no entanto o valor usual para considerá-la estatísticamente significante é 5% (0,05). Para fins de explanação do modelo, neste trabalho, serão efetuadas as demais análises do modelo de forma explicativa.

```
logit=glm(y_bin~x1+x2+x3, data=mydata, family = binomial(link="logit"))
summary(logit)
```

```
##
## Call:
## glm(formula = y_bin ~ x1 + x2 + x3, family = binomial(link = "logit"),
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0277   0.2347   0.5542   0.7016   1.0839
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)   0.4262     0.6390   0.667   0.5048
## x1            0.8618     0.7840   1.099   0.2717
## x2            0.3665     0.3082   1.189   0.2343
## x3            0.7512     0.4548   1.652   0.0986 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.056  on 69  degrees of freedom
## Residual deviance: 65.512  on 66  degrees of freedom
## AIC: 73.512
##
## Number of Fisher Scoring iterations: 5
```

```
require(stargazer)
stargazer(logit, title="Resultados",type = "text")
```

```
##
## Resultados
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                y_bin
## ------------------------------------------------
## x1                              0.862
##                                (0.784)
##
## x2                              0.367
##                                (0.308)
##
## x3                             0.751*
##                                (0.455)
##
## Constant                        0.426
##                                (0.639)
##
## ------------------------------------------------
## Observations                     70
## Log Likelihood                -32.756
## Akaike Inf. Crit.             73.512
## ================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

A razão de chances (OR - odds ratio em inglês) estimada no modelo terá de ser

transformada por estar apresentada na forma logarítma conforme o modelo de regressão logística o estima. Assim, utiliza-se o pacote `mfx` para efetuar esta transformação para todo o modelo de forma automatizada `(logitor(y_bin~x1+x2+x3,data=mydata))`:

```
require(mfx)
logitor(y_bin~x1+x2+x3,data=mydata)
```

```
## Call:
## logitor(formula = y_bin ~ x1 + x2 + x3, data = mydata)
##
## Odds Ratio:
##    OddsRatio Std. Err.      z   P>|z|
## x1   2.36735   1.85600 1.0992 0.27168
## x2   1.44273   0.44459 1.1894 0.23427
## x3   2.11957   0.96405 1.6516 0.09861 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O resultado acima evidencia que para uma alteração em 1 (uma) unidade em x3, a chance de que y seja igual a 1 aumenta em 112% ((2,12-1)*100). Dito de outra forma, a chance de y=1 é 2,12 vezes maior quando x3 aumenta em uma unidade (sendo que aqui mantêm-se as demais variáveis independentes constantes).

Como visto, para cada variação unitária em x3 o log das chances varia 0,7512. É possível estimar, portanto, a alteração das chances em função das médias dos valores de cada variável x1 e x2, e utilizar como exemplo os valores de 1, 2 e 3 para x3, para assim alcançar os preditores do log das chances nesta simulação, como segue abaixo:

Para facilitar a interpretação do modelo, se torna mais fácil depois de transformado a sua exponenciação dos coeficientes logísticos utilizando o comando `exp(coef(logit))`. Desta forma, para cada incremento unitário em x2 e mantendo as demais variáveis constantes, conclui-se que é 1,443 vezes provável que y seja igual a 1 em oposição a não ser (igual a zero), ou seja, as chances aumentam em 44,30%.

```
exp(coef(logit))
```

```
## (Intercept)          x1          x2          x3
##    1.531417    2.367352    1.442727    2.119566
```

O **intervalo de confiança** do modelo pode ser exposto utilizando o comando `confint` para os coeficientes estimados, como segue abaixo:

```
exp(cbind(OR=coef(logit), confint(logit)))
```

```
##                    OR     2.5 %    97.5 %
## (Intercept) 1.531417 0.4387468  5.625299
## x1          2.367352 0.5129380 11.674641
```

```
## x2          1.442727 0.8041221  2.737965
## x3          2.119566 1.0038973  5.718637
```

A partir do modelo logístico, podemos realizar **predições das probabilidades** de se encontrar o resultado y=1 conforme visto acima. Para isto, como exercício utilizaremos as médias das observações de cada variável independente do modelo. Em primeiro lugar deve ser criado um data.frame com os valores médios, como segue:

```
allmean = data.frame(x1=mean(mydata$x1),
                     x2=mean(mydata$x2),
                     x3=mean(mydata$x3))
allmean
```

```
##          x1        x2       x3
## 1 0.6480006 0.1338694 0.761851
```

Utiliza-se o comando `predict()` para predição do modelo, como segue abaixo, informando o objeto criado com a equação do modelo (logit), a base de dados com as condições dos valores médios (allmean) e o tipo de teste requerido ("response") para predizer as probabilidades. Como resultado, o modelo informa que constando os valores médios das variáveis independentes, obtêm-se a probabilidade de 83% em y se constituir igual a 1.

```
allmean$pred.prob = predict(logit, newdata=allmean, type="response")
allmean
```

```
##          x1        x2       x3 pred.prob
## 1 0.6480006 0.1338694 0.761851 0.8328555
```

4.3.1 Método Stepwise

O método Stepwise auxilia o pesquisador em selecionar as variáveis importantes ao modelo:

```
step(logit, direction = 'both')
```

```
## Start:  AIC=73.51
## y_bin ~ x1 + x2 + x3
##
##        Df Deviance    AIC
## - x1    1   66.736 72.736
## - x2    1   66.996 72.996
## <none>      65.512 73.512
## - x3    1   69.402 75.402
##
## Step:  AIC=72.74
```

```
## y_bin ~ x2 + x3
##
##        Df Deviance    AIC
## - x2    1   67.330 71.330
## <none>      66.736 72.736
## + x1    1   65.512 73.512
## - x3    1   70.032 74.032
##
## Step:  AIC=71.33
## y_bin ~ x3
##
##        Df Deviance    AIC
## <none>      67.330 71.330
## - x3    1   70.056 72.056
## + x2    1   66.736 72.736
## + x1    1   66.996 72.996

##
## Call:  glm(formula = y_bin ~ x3, family = binomial(link = "logit"),
##     data = mydata)
##
## Coefficients:
## (Intercept)           x3
##      1.1339       0.4866
##
## Degrees of Freedom: 69 Total (i.e. Null);  68 Residual
## Null Deviance:       70.06
## Residual Deviance: 67.33     AIC: 71.33
```

4.4 Regressão Logística Múltipla com variável categórica

Abaixo segue um exemplo com uma variável dependente categórica:

- **admin**: Variável dependente = 0 (não admitido) e 1 (admitido)
- **Rank**: Variável independente = ranking da escola de proveniência do candidato
- **Gre**: Variável independente = exames prévios do candidato.
- **Gpa**: Variável independente = exames prévios do candidato.

```
require(readr)
binary <- read_csv("http://www.karlin.mff.cuni.cz/~pesta/prednasky/NMFM404/Data/binary.c

binary$rank <- factor(binary$rank)
```

```
mylogit <- glm(admit ~ gre + gpa + rank, data = binary, family = binomial(link="logit"))
```

```
exp(cbind(OR = coef(mylogit), confint(mylogit)))
```

```
##                    OR       2.5 %    97.5 %
## (Intercept) 0.0185001 0.001889165 0.1665354
## gre         1.0022670 1.000137602 1.0044457
## gpa         2.2345448 1.173858216 4.3238349
## rank2       0.5089310 0.272289674 0.9448343
## rank3       0.2617923 0.131641717 0.5115181
## rank4       0.2119375 0.090715546 0.4706961
```

FAWCETT, T. **An introduction to ROC analysis**. *Pattern Recognition Letters*, [s.l.], 2006.

HAIR, J. F. et al. **Análise Multivariada de Dados**. São Paulo: Bookman, 2009.

HOSMER, D. W.; LEMESCHOW, S. **Applied Logistic Regression**. New York: Wiley, 2000.

SIEGEL, E. **Análise Preditiva: O poder de prever quem vai clicar, comprar, mentir ou morrer**. Rio de Janeiro: Alta Books, 2017.