Efeitos do álcool nos estudos

Grupo 04:
Maria Eduarda Garcia
Mirela Mei

Profª: Ana Amélia

DataSet

Acompanha o desempenho estudantil na educação secundária em duas escolas portuguesas. Consideram-se matérias de matemática e português. Na presente análise, consideramos apenas o primeiro caso.

Os atributos incluem as notas dos alunos, dispositivos demográficos, sociais e relacionados à escola e foram coletados com o uso de relatórios e questionários escolares.

- 395 linhas e 33 colunas
- Sem valores nulos iniciais, sem necessidade de remover registros por missing datas

DataSet – Variáveis

- Escola em que estuda
- Gênero
- Idade
- Área em que reside
- Tamanho da família
- Estado de coabitação dos pais
- Nível de escolaridade da mãe
- Nível de escolaridade do pai
- Trabalho do pai
- Trabalho da mãe
- Razão pela qual escola foi escolhida
- Quem possui a guarda
- Tempo da escola para casa
- Tempo de estudo semanal
- Repetências

DataSet – Variáveis

- Suporte escolar extra
- Suporte familiar
- Classes extra pagas
- Atividades extracurriculares
- Frequentouescola de enfermaria
- Pretende cursar ensino superior
- Acesso à internet
- Participa de relação romântica
- Qualidade da relação familiar
- Tempo livre após a escola
- Frequência com que sai com amigos
- Consumo de álcool diário
- Consumo de álcool aos finais de semana
- Estado de saúde
- Número de faltas
- Nota do primeiro período
- Nota do segundo período
- Nota final

DataSet - Variáveis

As variáveis escolhidas para análise foram: idade, nota final, número de faltas e consumo de álcool aos finais de semana (especificamente nível 4, alto)

Pergunta dessa fase da pesquisa: <u>De que forma a idade, o número de faltas e o alto consumo de álcool em finais de semana impactam na nota final G3)?</u>

Regressão Linear Múltipla

A regressão linear múltipla é uma técnica de dependência, que visa calcular a <u>dependência estatística de uma variável dependente quantitativa em relação a duas ou mais variáveis</u>

A fim de:

- encontrar relação causal entre as variáveis;
- estimar os valores da variável dependente a partir dos valores conhecidos das variáveis independentes.

Modelo

$$E(y) = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta kXk$$

Em que o coeficiente **β**k representa a variação esperada de Y para cada unidade de variação em Xk (k = 1, 2, ..., k), considerando as outras variáveis independentes fixas (no caso, Y)

Regressão Linear Múltipla

Entendem-se as equações da seguinte forma:

Equação de regressão ajustada aos dados: $\hat{y} = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$

Valores preditos: $\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + ... + b_kx_{ik}$

Resíduos: $\hat{e}_i = y_i - \hat{y}_i$

Regressão Linear Múltipla

Para avaliar a qualidade do ajuste, tem-se o coeficiente de determinação, que apresenta uma medida da proporção da variação total que é explicada pelo modelo de regressão: r^2

É o quociente entre a variação explicada e a variação total do modelo.

Pode-se mostrar que:

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i - \overline{y})^2 & = & \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 & + & \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \\ \downarrow & & \downarrow & & \downarrow \\ SST & = & SSE & + & SSR \end{array}$$

$SST \longmapsto$ Soma dos quadrados totais - Variação total

$SSE \longmapsto$ Soma dos quadrados dos resíduos - Variação não explicada

$SSR \longmapsto$ Soma dos quadrados da regressão - Variação explicada

Regressão Linear Múltipla

Temos, então, duas hipóteses:

- h0: $\beta 1 = \beta 2 = \beta 3 = \ldots = \beta k = 0$
  - as variáveis selecionadas não interferem na variável independente
- h1: nem todos os βi são iguais à 0
  - os coeficientes que não são nulos correspondem às variáveis que interferem na variável independente

Regressão Linear Múltipla

As variáveis escolhidas no presente trabalho são:

- VARIÁVEL DEPENDENTE
  - G3: nota final escolar - ao sair do ensino médio
- VARIÁVEIS INDEPENDENTES
  - idade
  - número de faltas
  - alto consumo de álcool aos finais de semana (variável dummy - considerado apenas nível 4)

Estatística Descritiva

tabela de frequências

```
> table(dados$age)

 15  16  17  18  19  20  21  22 
 82 104  98  82  24   3   1   1 
> table(dados$Walc)

  1   2   3   4   5 
151  85  80  51  28 
> table(dados$absences)

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19 
115   3  65   8  53   5  31   7  22   3  17   3  12   3  12   3   7   1   5   1 
 20  21  22  23  24  25  26  28  30  38  40  54  56  75 
  4   1   3   1   1   1   1   1   1   1   1   1   1   1 
> table(dados$G3)

 0  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 
38  1  7 15  9 32 28 56 47 31 31 27 33 16  6 12  5  1 
```

Estatística Descritiva

```
> # Amplitudes
> range(dados$age)
[1] 15 22
> range(dados$absences)
[1]  0 75
> range(dados$G3)
[1]  0 20
```

```
> # Função summary (média, mediana, quartis e valores mín e máx)
> summary(dados$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   15.0    16.0    17.0    16.7    18.0    22.0 
> summary(dados$absences)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   4.000   5.709   8.000  75.000 
> summary(dados$G3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    8.00   11.00   10.42   14.00   20.00 
```

Estatística Descritiva

mediana

```
> medianAge <- median(dados$age)
> print(medianAge)
[1] 17
> medianAbsences <- median(dados$absences)
> print(medianAbsences)
[1] 4
> medianG3 <- median(dados$G3)
> print(medianG3)
[1] 11
```

Estatística Descritiva

moda

```
> getmode <- function(dados) {
+   uniqv <- unique(dados)
+   uniqv[which.max(tabulate(match(dados, uniqv)))]
+ }
> resultAge <- getmode(dados$age)
> print(resultAge)
[1] 16
> resultAbsences <- getmode(dados$absences)
> print(resultAbsences)
[1] 0
> resultG3 <- getmode(dados$G3)
> print(resultG3)
[1] 10
```

Estatística Descritiva

```
> # Função describe e describeBy (média, desvio padrão[sd], erro, mediana)
> describe(dados$age)
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 395 16.7 1.28     17   16.63 1.48  15  22     7 0.46    -0.03 0.06
> describe(dados$absences)
   vars   n mean sd median trimmed  mad min max range skew kurtosis  se
X1    1 395 5.71  8      4    4.24 5.93   0  75    75 3.64    21.31 0.4
> describe(dados$G3)
   vars   n  mean   sd median trimmed  mad min max range  skew kurtosis   se
X1    1 395 10.42 4.58     11   10.84 4.45   0  20    20 -0.73     0.37 0.23
```

Estatística Descritiva

```
> # Erro
> std_mean <- function(x) sd(x)/sqrt(length(x))
> std_mean(dados$age)
[1] 0.06420468
> std_mean(dados$absences)
[1] 0.4026794
> std_mean(dados$G3)
[1] 0.2305174
```

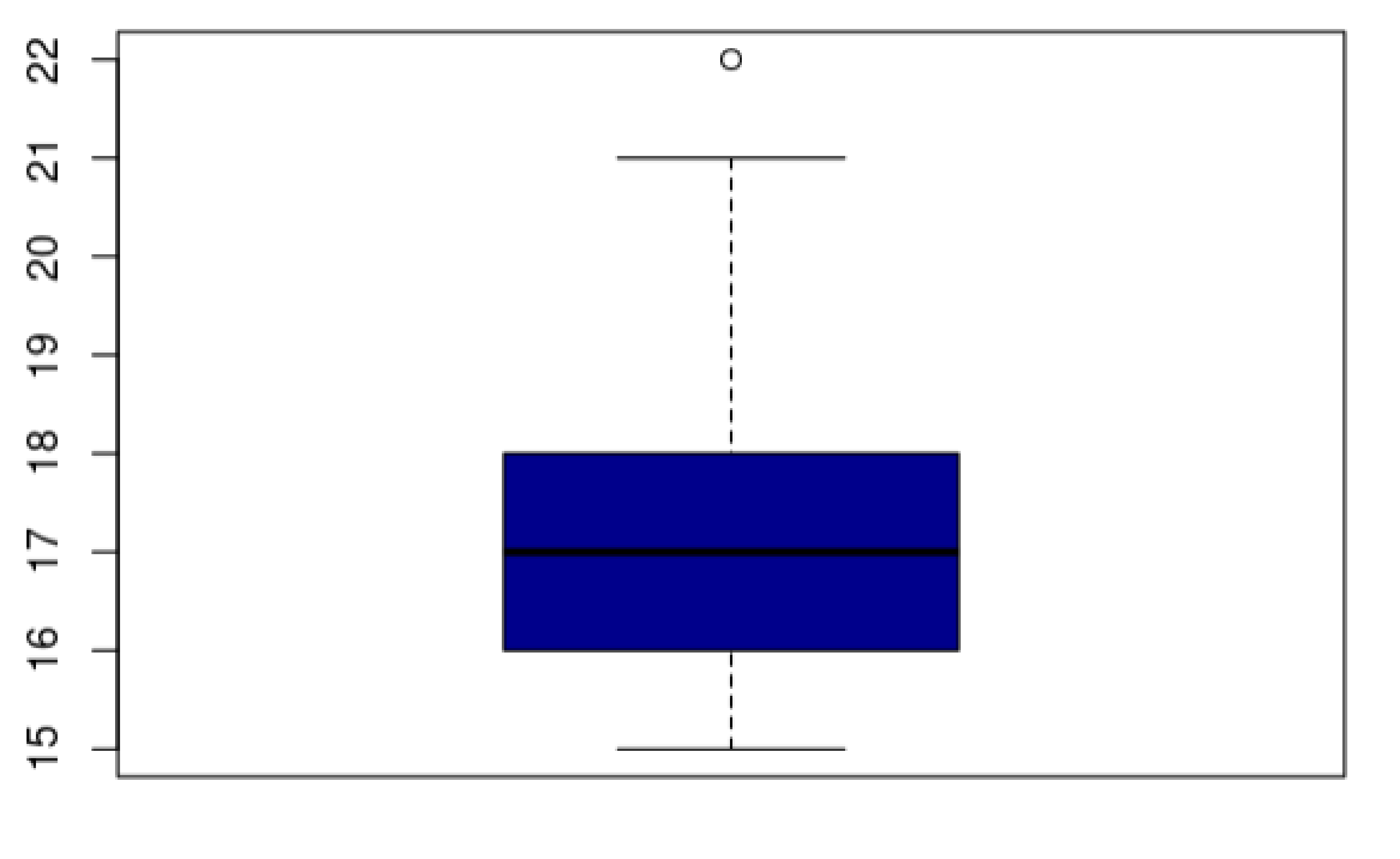
Estatística Descritiva

```
> # Variância
> var(dados$age)
[1] 1.628285
> var(dados$absences)
[1] 64.04954
> var(dados$G3)
[1] 20.98962
```

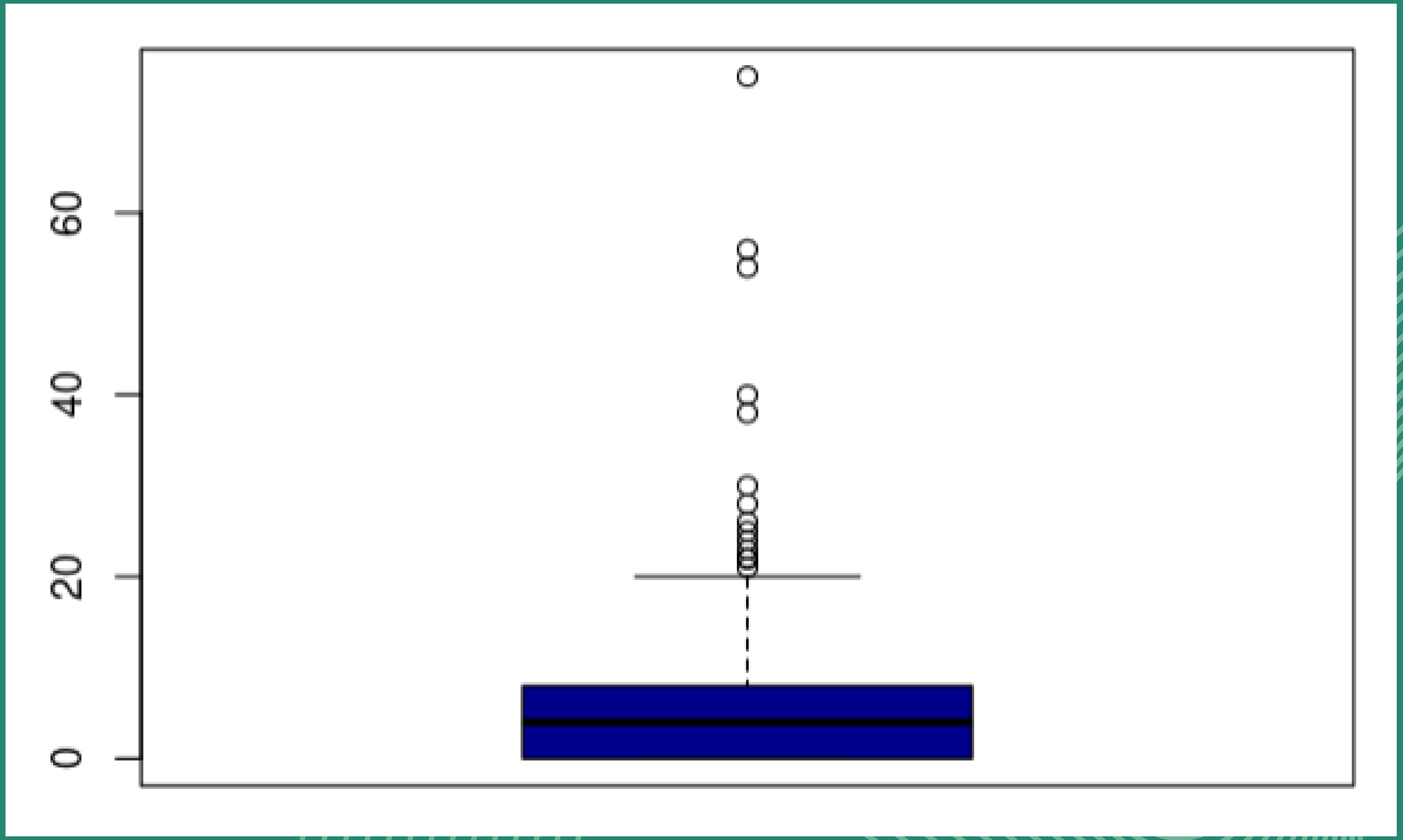
Estatística Descritiva

Boxplot das variáveis idade, faltas e notas finais, respectivamente

Idades



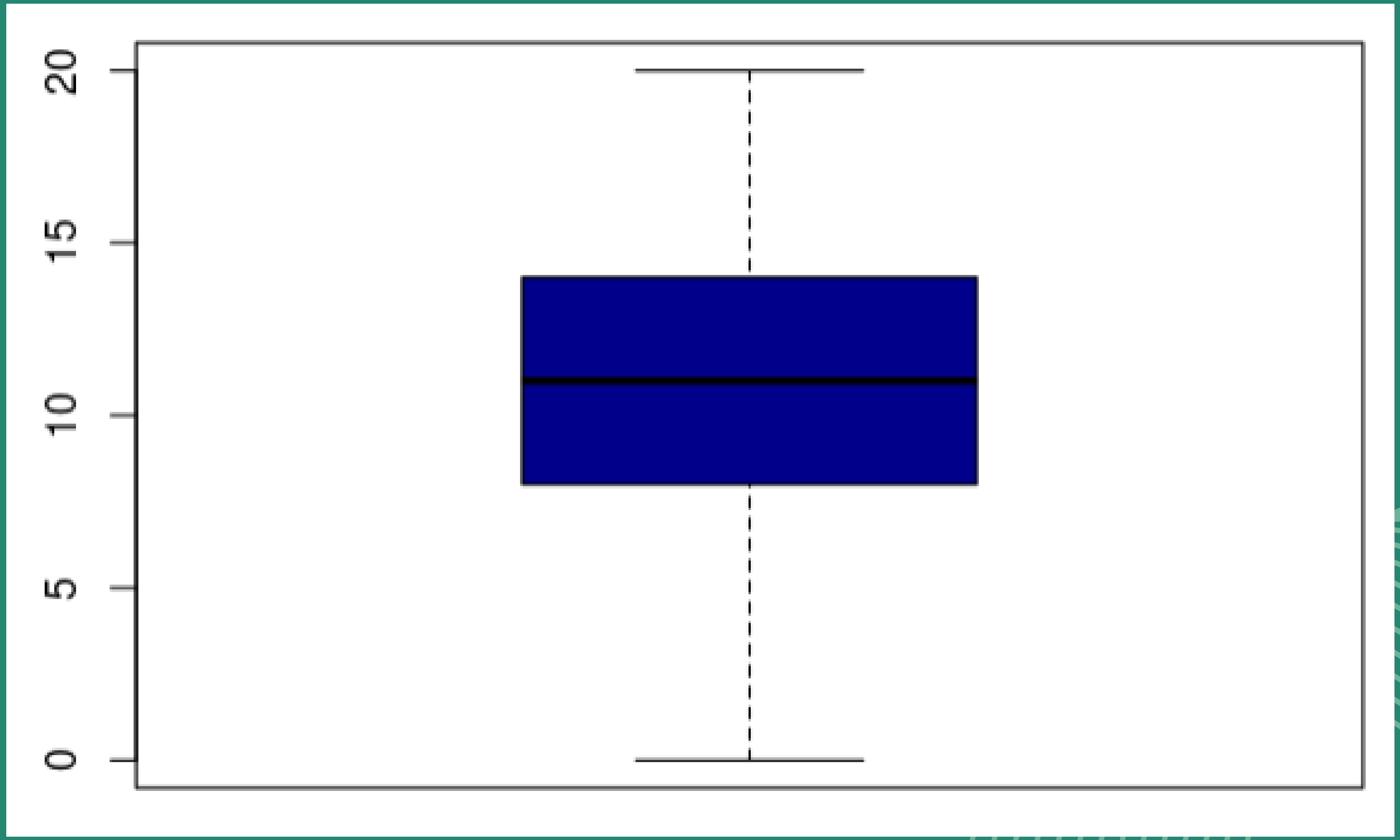
Faltas



Estatística Descritiva

Boxplot das variáveis idade, faltas e notas finais, respectivamente

Notas finais



Estatística Descritiva

Gráfico de setores da variável qualitativa

1 – consumo muito baixo
2 – consumo baixo
3 – consumo médio
4 – consumo alto
5 – consumo muito alto

Estatística Descritiva

A partir do uso de variáveis dummy, a variável qualitativa foi desmembrada, e na presente análise foram utilizadas somente as respostas que indicam o nível 4, que se referem ao alto consumo de álcool em finais de semana

Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x faltas

```
> modAbsence <- lm(dados2$G3 ~ dados2$absences)
> summary(modAbsence)

Call:
lm(formula = dados2$G3 ~ dados2$absences)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.3033  -2.3033   0.5007   3.4811   9.6183 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     10.30327    0.28347  36.347   <2e-16 ***
dados2$absences  0.01961    0.02886   0.679    0.497    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.585 on 393 degrees of freedom
Multiple R-squared:  0.001173,	Adjusted R-squared:  -0.001369 
F-statistic: 0.4615 on 1 and 393 DF,  p-value: 0.4973
```

Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x faltas

```
> summary(rstandard(modAbsence))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.251684 -0.503358  0.109388 -0.000427  0.760816  2.100754
> anova1 <- aov(modAbsence)
> shapiro.test(anova1$residuals)

	Shapiro-Wilk normality test

data:  anova1$residuals
W = 0.93511, p-value = 4.283e-12
```

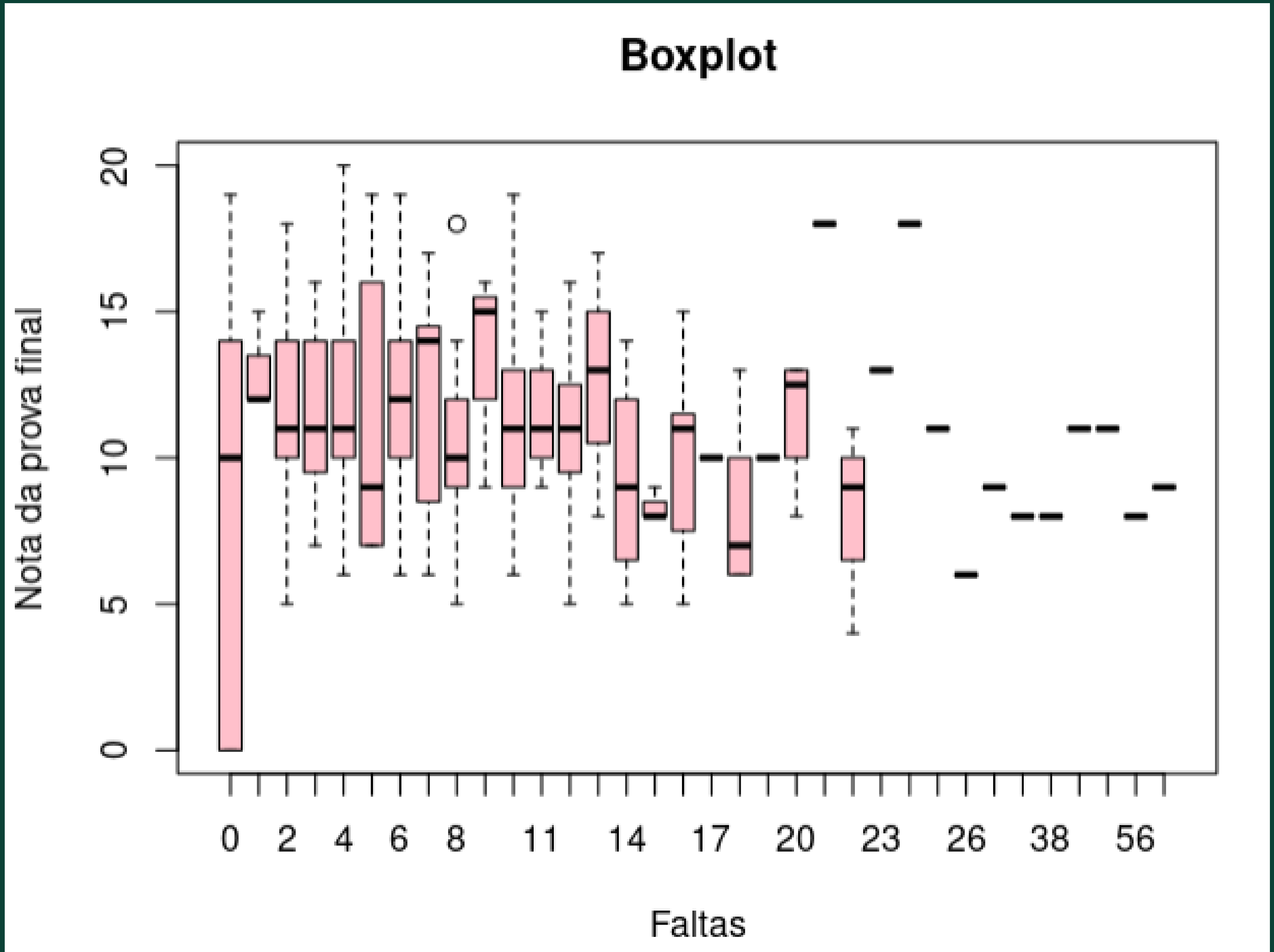
Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x faltas

- Observando o modelo padronizado, tem-se
  - resíduo padronizado deveria estar dentro do intervalo -2 e 2 - modelo pouco ajustado
  - p value muito abaixo de 0,05, - há indícios de normalidade, ou seja, entende-se que uma variável interfere na outra

Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x faltas



Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x idade

```
> modAge <- lm(dados2$G3 ~ dados2$age)
> summary(modAge)

Call:
lm(formula = dados2$G3 ~ dados2$age)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3992  -1.6588   0.3412   3.1809   9.5014

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.1011     2.9928   6.717 6.54e-11 ***
dados2$age   -0.5801     0.1787  -3.246  0.00127 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.527 on 393 degrees of freedom
Multiple R-squared:  0.02611,	Adjusted R-squared:  0.02363
F-statistic: 10.54 on 1 and 393 DF,  p-value: 0.001271
```

Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x idade

```
> summary(rstandard(modAge))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-2.5269341 -0.3673823  0.0755614  0.0000151  0.7038154  2.1196582
> anova2 <- aov(modAge)
> shapiro.test(anova2$residuals)

	Shapiro-Wilk normality test

data:  anova2$residuals
W = 0.94329, p-value = 3.805e-11
```

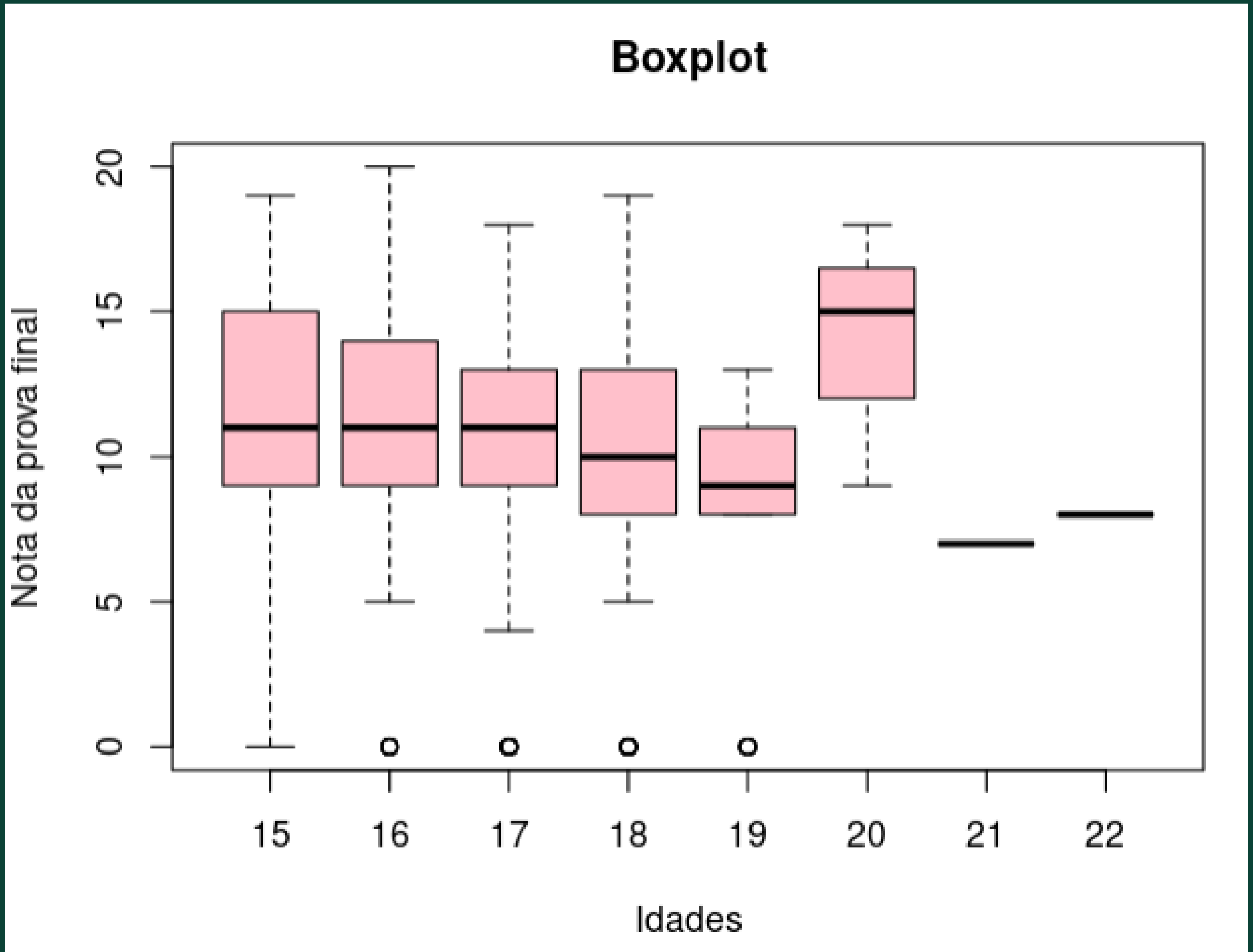
Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x idade

- Observando o modelo padronizado, tem-se
  - resíduo padronizado deveria estar dentro do intervalo -2 e 2 - modelo pouco ajustado
  - p value muito abaixo de 0,05, - há indícios de normalidade, ou seja, entende-se que uma variável interfere na outra

Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x idade



Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x alto consumo de álcool em finais de semana

```
> modAlcohol <- lm(dados2$G3 ~ dados2$Walc_4)
> summary(modAlcohol)

Call:
lm(formula = dados2$G3 ~ dados2$Walc_4)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5233  -1.6863   0.4767   3.4767   9.4767

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.5233     0.2469  42.628   <2e-16 ***
dados2$Walc_4  -0.8370     0.6870  -1.218    0.224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.579 on 393 degrees of freedom
Multiple R-squared:  0.003762,	Adjusted R-squared:  0.001227
F-statistic: 1.484 on 1 and 393 DF,  p-value: 0.2238
```

Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x alto consumo de álcool em finais de semana

```
> summary(rstandard(modAlcohol))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-2.3017 -0.3720  0.1043  0.0000  0.7604  2.0728 
> anova3 <- aov(modAlcohol)
> shapiro.test(anova3$residuals)

	Shapiro-Wilk normality test

data:  anova3$residuals
W = 0.92767, p-value = 6.864e-13
```

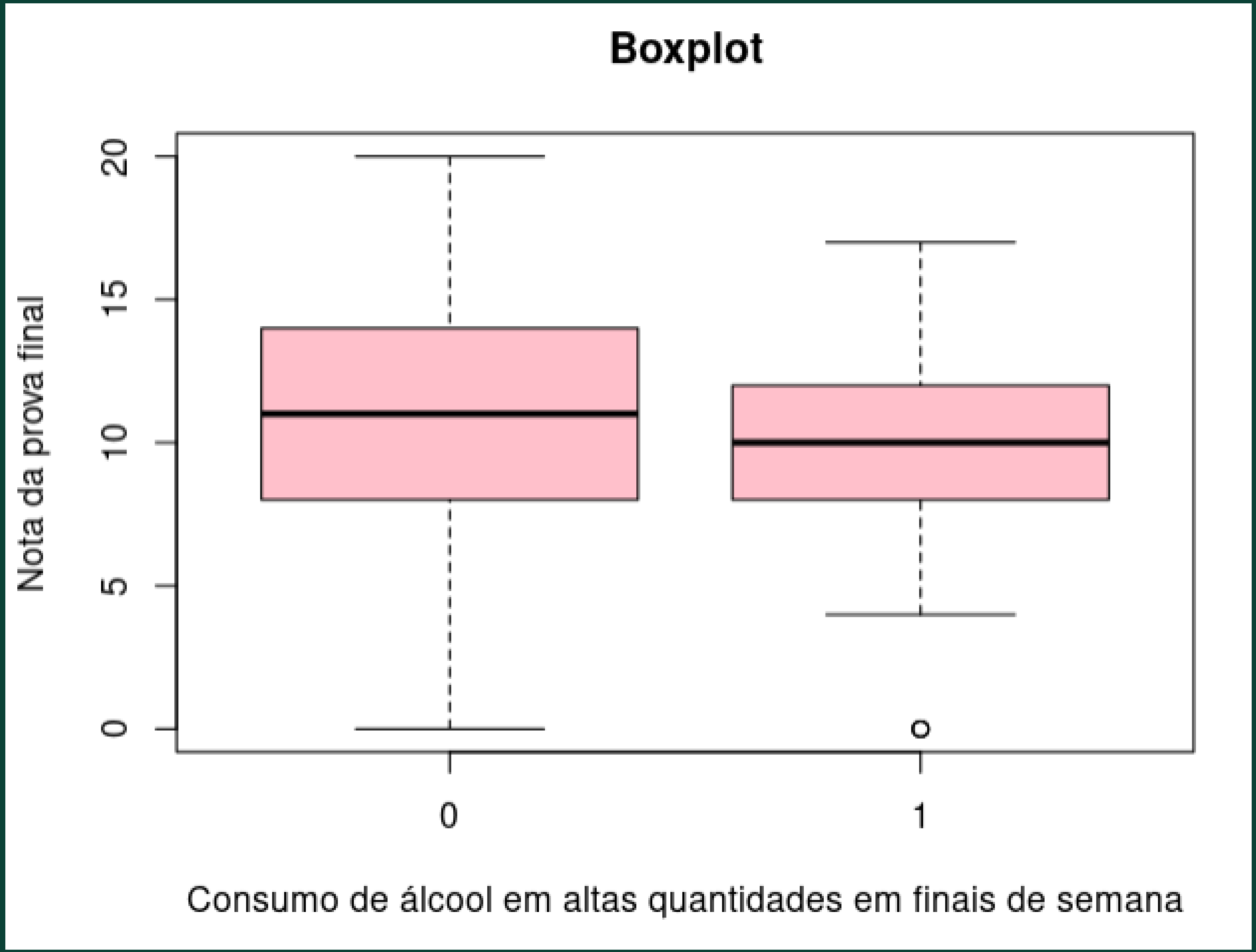
Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x alto consumo de álcool em finais de semana

- Observando o modelo padronizado, tem-se
  - resíduo padronizado deveria estar dentro do intervalo -2 e 2 - modelo pouco ajustado
  - p value muito abaixo de 0,05, - há indícios de normalidade, ou seja, entende-se que uma variável interfere na outra

Apresentação dos resultados

Regressão Linear Simples – Nota final (G3) x alto consumo de álcool em finais de semana



Modelo com todas as variáveis

```
> modelo <- lm(G3 ~ absences + age + Walc_4, dados2)
> summary(modelo)

Call:
lm(formula = G3 ~ absences + age + Walc_4, data = dados2)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.3800  -1.7223   0.3761   3.1264   9.7920 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.89591    3.01564   6.929 1.75e-11 ***
absences     0.04279    0.02913   1.469 0.142680    
age         -0.63440    0.18133  -3.499 0.000521 ***
Walc_4      -1.03031    0.68397  -1.506 0.132779    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Método Backward

```
> step(modelo, direction="backward")
Start:  AIC=1195.01
G3 ~ absences + age + Walc_4

           Df Sum of Sq    RSS    AIC
<none>                  7974.3 1195.0
- absences  1    44.002 8018.3 1195.2
- Walc_4    1    46.279 8020.6 1195.3
- age       1   249.640 8224.0 1205.2

Call:
lm(formula = G3 ~ absences + age + Walc_4, data = dados2)

Coefficients:
(Intercept)     absences          age       Walc_4
   20.89591      0.04279     -0.63440     -1.03031
```

Método Stepwise

```
> step(modelo, direction="both")
Start:  AIC=1195.01
G3 ~ absences + age + Walc_4

           Df Sum of Sq    RSS    AIC
<none>                  7974.3 1195.0
- absences  1    44.002 8018.3 1195.2
- Walc_4    1    46.279 8020.6 1195.3
- age       1   249.640 8224.0 1205.2

Call:
lm(formula = G3 ~ absences + age + Walc_4, data = dados2)

Coefficients:
(Intercept)     absences          age       Walc_4
   20.89591      0.04279     -0.63440     -1.03031
```

Método Forward

```
> modelo <- lm(G3 ~ absences + age + Walc_4, dados2)
> step(modelo, direction="forward")
Start:  AIC=1195.01
G3 ~ absences + age + Walc_4


Call:
lm(formula = G3 ~ absences + age + Walc_4, data = dados2)

Coefficients:
(Intercept)     absences          age       Walc_4
   20.89591      0.04279     -0.63440     -1.03031
```

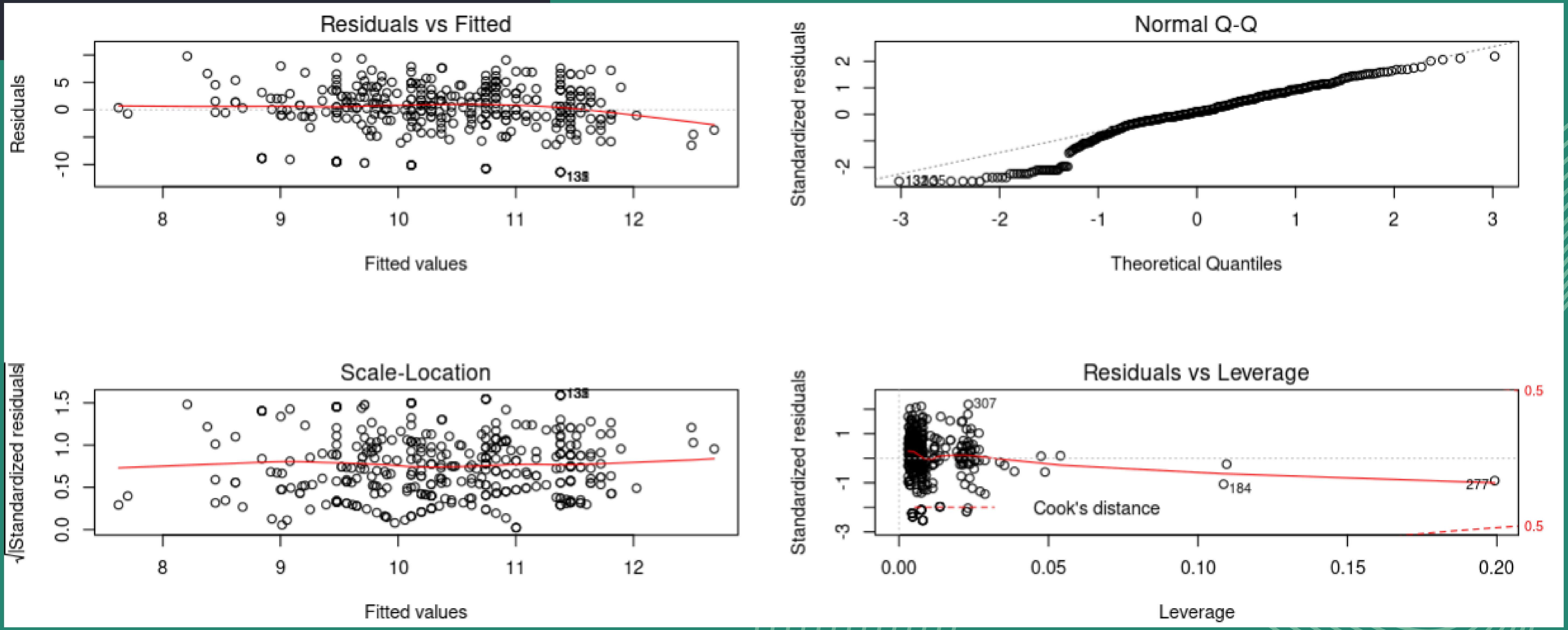
Métodos de checagem

Os três métodos de checagem de modelo (backward, forward e stepwise) indicam que, para construção do modelo definitivo, não se deve tirar nenhuma variável, a fim de não comprometer a qualidade. Assim, mantém-se as variáveis avaliadas durante a regressão linear simples.

Regressão Linear Múltipla

Construção do modelo e análise gráfica – estudo dos resíduos

```
> modelo <- lm(G3 ~ absences + age + Walc_4, dados2)
> # Análise gráfica
> par(mfrow=c(2,2))
> plot(modelo)
> par(mfrow=c(1,1))
```

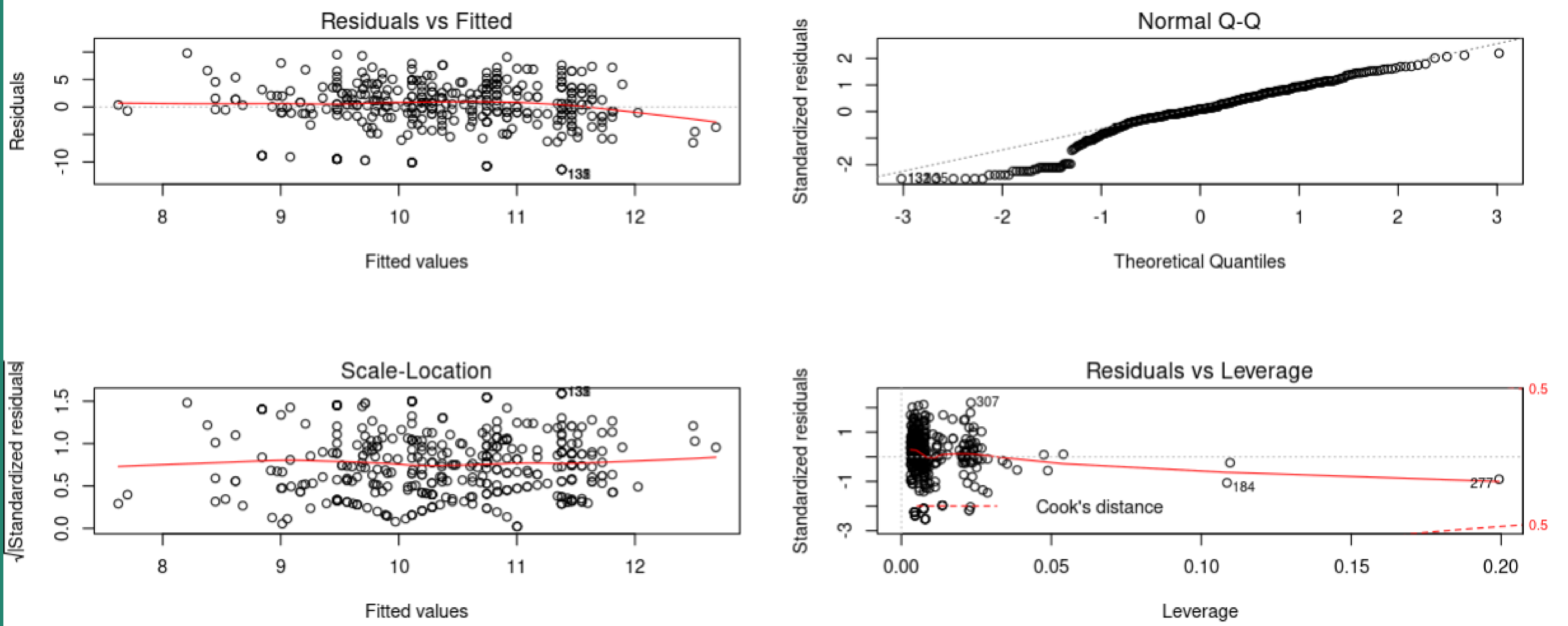


Regressão Linear Múltipla

Construção do modelo e análise gráfica – estudo dos resíduos

Linearidade

Análise de normalidade



Homoscedasticidade

Outliers e pontos influentes

Regressão Linear Múltipla

Construção do modelo e análise gráfica – estudo dos resíduos

Normalidade dos resíduos

```
> shapiro.test(modelo$residuals)

	Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.94979, p-value = 2.502e-10
```

Outliers nos resíduos

```
> summary(rstandard(modelo))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max. 
-2.5299682 -0.3829807  0.0853290 -0.0005498  0.6952566  2.1938392 
```

Regressão Linear Múltipla

Construção do modelo e análise do coeficiente de regressão, t value, p value e erro padrão

```
> summary(modelo)

Call:
lm(formula = G3 ~ absences + age + Walc_4, data = dados2)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.3800  -1.7223   0.3761   3.1264   9.7920 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 20.89591    3.01564   6.929 1.75e-11 ***
absences     0.04279    0.02913   1.469 0.142680    
age         -0.63440    0.18133  -3.499 0.000521 ***
Walc_4      -1.03031    0.68397  -1.506 0.132779    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.516 on 391 degrees of freedom
Multiple R-squared:  0.03574,	Adjusted R-squared:  0.02834 
F-statistic: 4.831 on 3 and 391 DF,  p-value: 0.002585
```

Conclusão

p value = 0,002585, f = 4,831, r^2 ajustado = 0,02834

entende-se que coeficientes angulares da reta são diferentes de 0, o que faz com que rejeitemos h0

entende-se, portanto, que as variáveis selecionadas interferem na variável independente

em outras palavras, a idade, as faltas e o alto consumo de álcool em finais de semana interferem na nota final dos adolescentes entrevistados

Bibliotecas utilizadas

- dplyr - facilita a manipulação de dataframes
- psych - facilita análise multivariada
- ISwR - facilita análises estatísticas iniciais
- pacman - facilita uso de funções nativas do R
- nycflights13 - facilita análise multivariada
- fastDummies - auxílio ao criar variáveis dummy

Referências

- Raciocínio para o R2 (artigo). Disponível em: <https://pt.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/a/r-squared-intuition>. Acesso em 28/10/2022.
- Análise de variância (ANOVA) | Estatística e probabilidade. Disponível em: <https://pt.khanacademy.org/math/statistics-probability/analysis-of-variance-anova-library>. Acesso em 28/10/2022.
- SANT’ANA, R. Disponível em: <http://lite.acad.univali.br/rcurso/anova/>. Acesso em 28/10/2022.
- O Significado e a Interpretação dos P-values (o que os dados dizem). Disponível em: <http://www.bertolo.pro.br/FinEst/Estatistica/EstatisticaNosNegocios/p-

OBRIGADA!