Efeitos do álcool nos estudos

Grupo 04:
Maria Eduarda Garcia
Mirela Mei

Prof.ª Ana Amélia

Dataset

DESCRIÇÃO

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

VARIÁVEIS

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.

DADOS COLETADOS

- UCI Machine Learning Repository
- 395 linhas e 33 colunas
- Não há missing data

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
- Suporte escolar extra
- Suporte familiar
- Classes extra pagas
- Atividades extracurriculares
- Frequentou escola 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

Regressão Logística

DESCRIÇÃO

É uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias.
Então, a partir desse modelo gerado é possível calcular ou prever a probabilidade de um evento ocorrer, dado uma observação aleatória.

Em comparação às técnicas conhecidas em regressão, em especial a regressão linear, a regressão logística distingue-se essencialmente pelo fato de a variável resposta ser categórica. Em geral, a ocorrência do evento de interesse é codificada como “1” e a ausência como “0”

Regressão Logística

DESCRIÇÃO

O modelo de regressão logística permite:

- Modelar a probabilidade de um evento ocorrer dependendo dos valores das variáveis independentes, que podem ser categóricas ou contínuas.
- Estimar a probabilidade de um evento ocorrer para uma observação selecionada aleatoriamente contra a probabilidade do evento não ocorrer
- Prever o efeito do conjunto de variáveis sobre a variável dependente binária.
- Classificar observações, estimando a probabilidade de uma observação estar em uma categoria determinada.

Regressão Logística

NOTAÇÃO CLÁSSICA DO MODELO LINEAR

$$Y = \alpha + \beta X + \varepsilon$$

- Y representa a variável dependente, ou seja, aquilo que queremos entender/explicar/predizer.
- X representa a variável independente.
- O intercepto, (α), representa o valor de Y quando X assume valor zero.
- O coeficiente de regressão, (β), representa a variação observada em Y associada ao aumento de uma unidade em X.
- O termo estocástico, (ε), representa o erro do modelo.
- Tecnicamente, é possível estimar se existe relação linear entre uma variável dependente (Y) e diferentes variáveis independentes.

Regressão Logística

PREMISSAS

- Na regressão logística binária, é necessário que a variável resposta seja binária.
- O resultado desejado deve ser representado pelo fator nível 1 da variável resposta, o indesejado é 0.
- Apenas as variáveis que são significativas devem ser incluídas.
- Variáveis independentes devem ser essencialmente independentes umas das outras. Deve haver pouca ou nenhuma multi-colinearidade.
- As probabilidades de log e as variáveis independentes devem estar linearmente relacionadas.
- A regressão logística deve ser aplicada apenas a tamanhos de amostra massivos.

Regressão Logística

EXEMPLO

Suponha que queira-se analisar a ocorrência da apneia do sono, que é um distúrbio do sono potencialmente grave. Existem vários fatores que podem influenciar nesse distúrbio, mas para este exemplo, iremos considerar apenas dois: idade e peso. Digamos que para esta análise, tenhamos uma amostra de cem indivíduos, contendo a idade, o peso e se ele tem apneia ou não, este é o nosso conjunto de observações.
A variável dependente é a ocorrência ou não da apneia do sono, ter apneia é igual a 1, não ter apneia é igual a 0. As variáveis independentes são a idade e o peso. Para este exemplo, o que a regressão logística propõe é que, a partir dessas informações, é possível gerar um modelo logístico que possa prever a probabilidade de uma pessoa ter apneia do sono, baseando-se no peso e idade desta pessoa.

Regressão Logística

ESCOLHA DAS VARIÁVEIS

As variáveis escolhidas para a analise foram:
'Está em uma relação amorosa' (*romantic*), G3 (nota final escolar) e número de faltas (*absences*).

PERGUNTA DA ANÁLISE

> Alunos que não faltaram e tiraram notas acima da média possuem maior probabilidade de não estar em uma relação amorosa?

Estatística Descritiva

ESTÁ EM UMA RELAÇÃO AMOROSA

Tabelas de frequência:

```
> table(dados$romantic)

 no yes
263 132
> prop.table(table(dados$romantic))

       no       yes
0.6658228 0.3341772
>
```

Estatística Descritiva

ESTÁ EM UMA RELAÇÃO AMOROSA

Estatística Descritiva

G3 (NOTA FINAL ESCOLAR)

Tabelas de frequência:

```
> 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
> prop.table(table(dados$G3))

          0           4           5           6           7           8
0.096202532 0.002531646 0.017721519 0.037974684 0.022784810 0.081012658
          9          10          11          12          13          14
0.070886076 0.141772152 0.118987342 0.078481013 0.078481013 0.068354430
         15          16          17          18          19          20
0.083544304 0.040506329 0.015189873 0.030379747 0.012658228 0.002531646
>
```

Estatística Descritiva

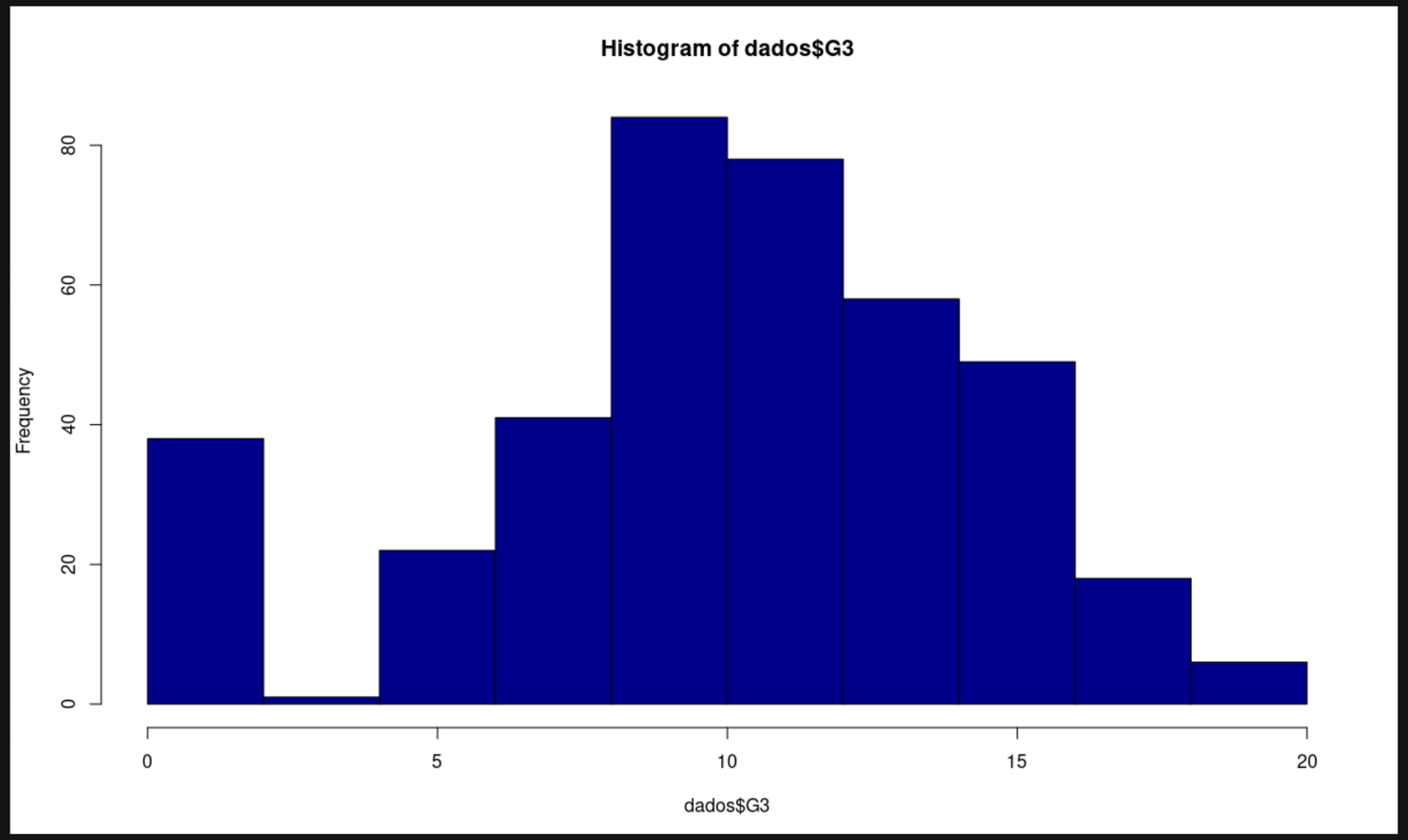
G3 (NOTA FINAL ESCOLAR)

Medidas de tendência central e de dispersão:

```
> summary(dados$G3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   11.00   10.42   14.00   20.00
> medianG3 <- median(dados$G3)
> print(medianG3)
[1] 11
> getmode <- function(dados) {
+   uniqv <- unique(dados)
+   uniqv[which.max(tabulate(match(dados, uniqv)))]
+ }
> resultG3 <- getmode(dados$G3)
> print(resultG3)
[1] 10
> describe(dados$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
> 
```

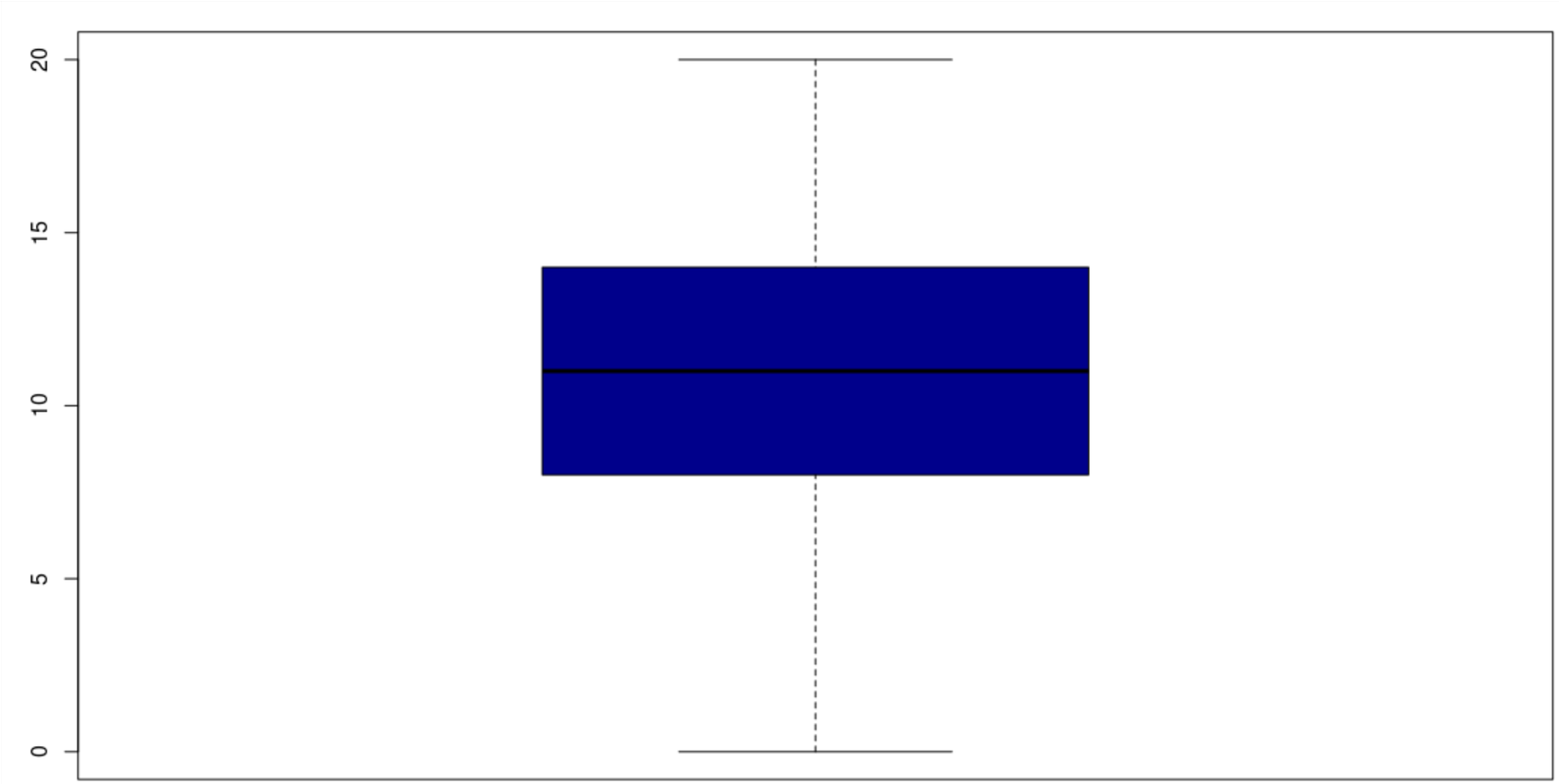
Estatística Descritiva

G3 (NOTA FINAL ESCOLAR)



Estatística Descritiva

G3 (NOTA FINAL ESCOLAR) - BOXPLOT



Estatística Descritiva

NÚMERO DE FALTAS

Tabelas de frequência:

```
> 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
> prop.table(table(dados$absences))

          0           1           2           3           4           5
0.291139241 0.007594937 0.164556962 0.020253165 0.134177215 0.012658228
          6           7           8           9          10          11
0.078481013 0.017721519 0.055696203 0.007594937 0.043037975 0.007594937
         12          13          14          15          16          17
0.030379747 0.007594937 0.030379747 0.007594937 0.017721519 0.002531646
         18          19          20          21          22          23
0.012658228 0.002531646 0.010126582 0.002531646 0.007594937 0.002531646
         24          25          26          28          30          38
0.002531646 0.002531646 0.002531646 0.002531646 0.002531646 0.002531646
         40          54          56          75
0.002531646 0.002531646 0.002531646 0.002531646
>
```

Estatística Descritiva

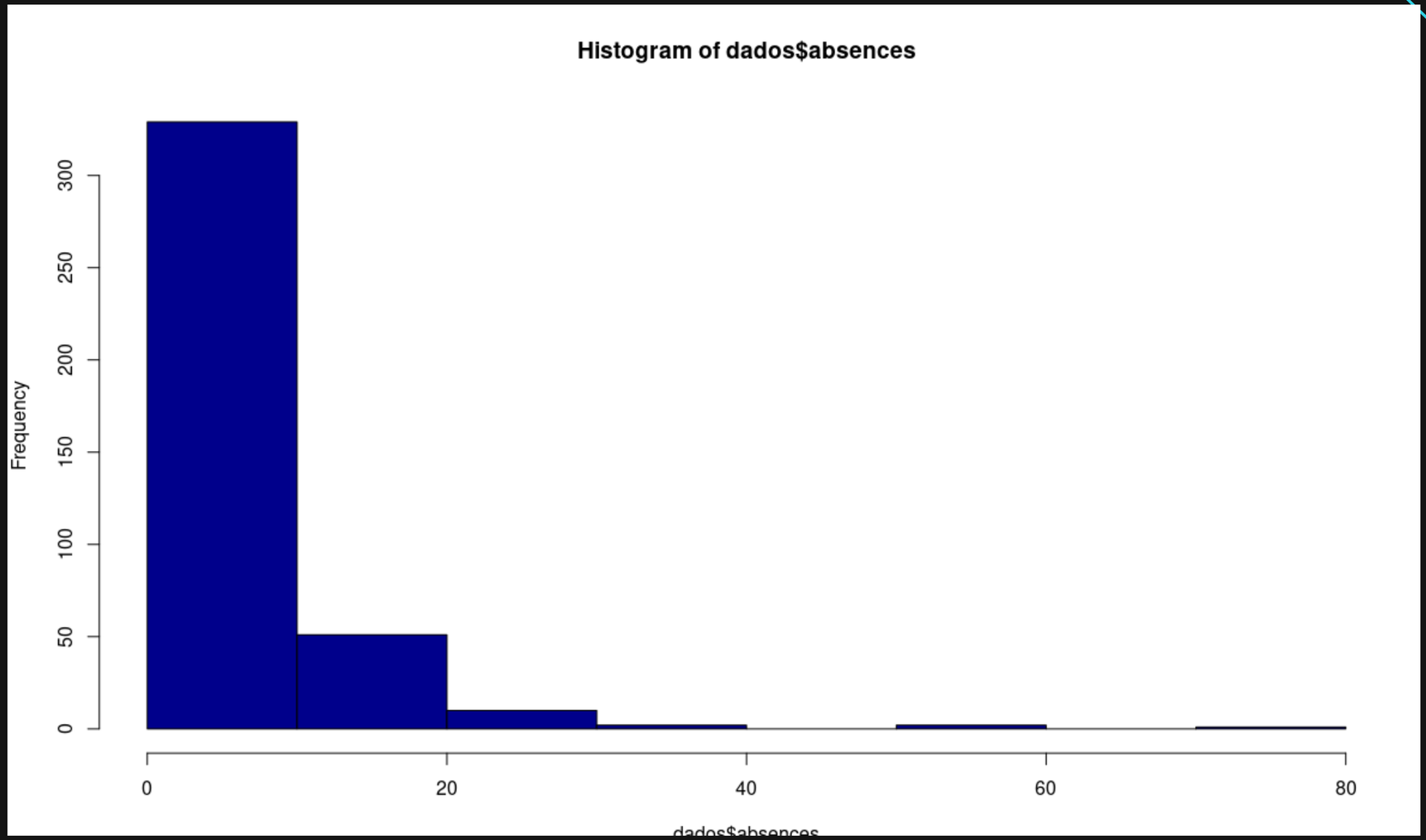
NÚMERO DE FALTAS

Medidas de tendência central e de dispersão:

```
> summary(dados$absences)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   4.000   5.709   8.000  75.000
> medianAbsences <- median(dados$absences)
> print(medianAbsences)
[1] 4
> getmode <- function(dados) {
+   uniqv <- unique(dados)
+   uniqv[which.max(tabulate(match(dados, uniqv)))]
+ }
> resultAbsences <- getmode(dados$absences)
> print(resultAbsences)
[1] 0
> describe(dados$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
> |
```

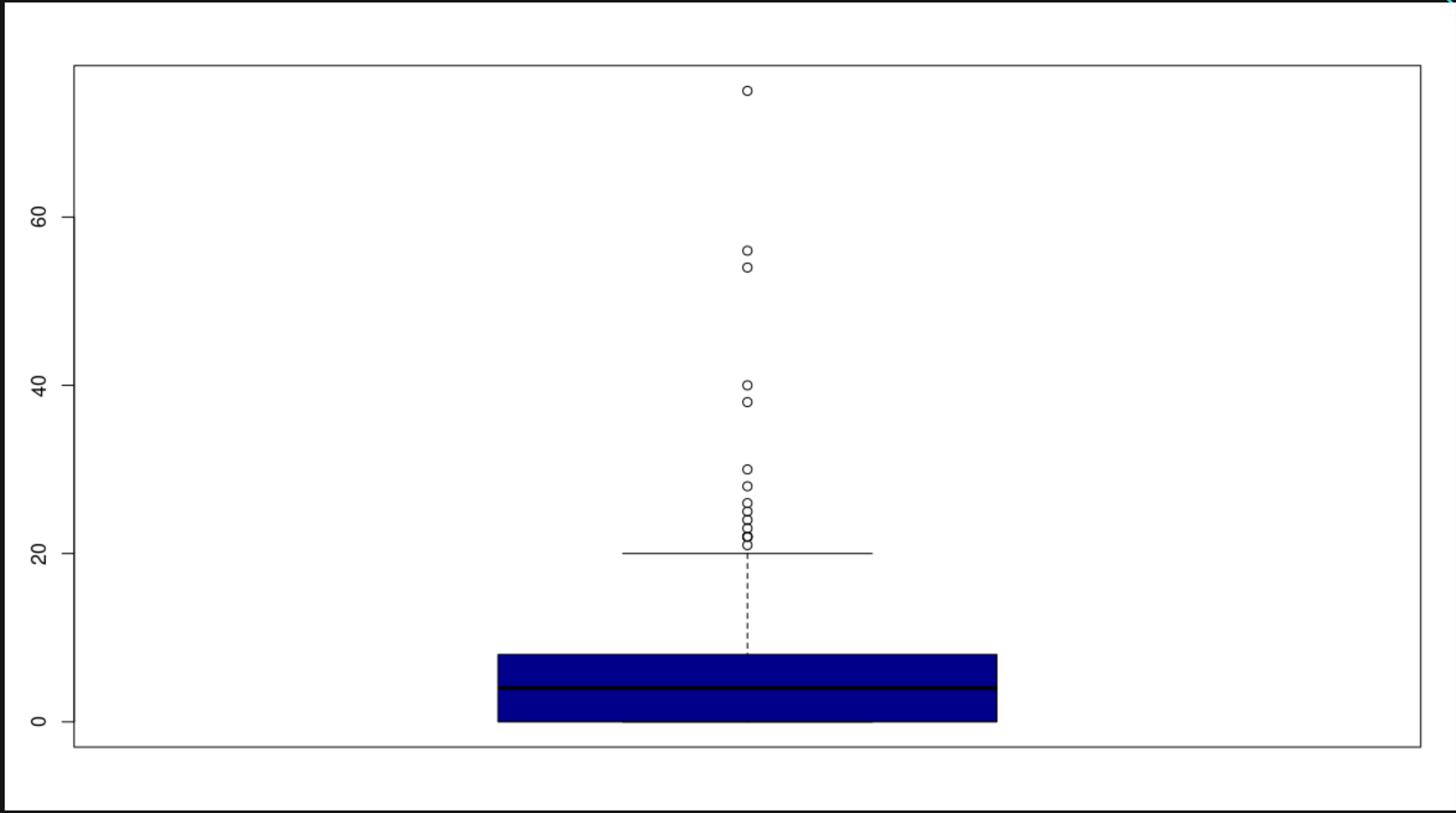
Estatística Descritiva

NÚMERO DE FALTAS



Estatística Descritiva

NÚMERO DE FALTAS - BOXPLOT



Regressão Logística

VARIÁVEIS

As variáveis escolhidas para realizar a regressão logística foram:

- Variáveis independentes
  - absences - número de faltas
  - G3 - nota final
- Variável dependente
  - romantic - se se encontra ou não em relação amorosa (yes/no)

> Entende-se por *dados2* o dataset com apenas essas três colunas de interesse, sendo removidas as iniciais do dataset que não serão utilizadas.

Regressão Logística

VARIÁVEIS

Avaliando o balanceamento das variáveis, tem-se que estão distribuídas, não se concentrando em valores que não se alteram entre os registros:

```
> summary(dados2)
 romantic     absences           G3       
 no :263   Min.   : 0.000   Min.   : 0.00  
 yes:132   1st Qu.: 0.000   1st Qu.: 8.00  
           Median : 4.000   Median :11.00  
           Mean   : 5.709   Mean   :10.42  
           3rd Qu.: 8.000   3rd Qu.:14.00  
           Max.   :75.000   Max.   :20.00  
```

Regressão Logística

VARIÁVEL CATEGÓRICA - CATEGORIA DE REFERÊNCIA

A categoria de referência da variável *romantic* foi concluída pela função *levels*, que retorna que a categoria "não" é a primeira categoria considerada.

```
> # Checagem das categorias de referência
> levels(dados2$romantic) # categoria de referência: "não"
[1] "no"  "yes"
```

Regressão Logística

CHECAGEM DE PRESSUPOSTOS

1) Variável dependente dicotômica: categorias mutuamente exclusivas
   - Verdadeiro, uma vez que ou o aluno está ou não está em relação amorosa
2) Observações independentes
   - Verdadeiro, não há medidas repetidas uma vez que cada registro se refere a um aluno diferente

Regressão Logística

CHECAGEM DE PRESSUPOSTOS

Para a verificação dos próximos pressupostos, será necessária a criação de um modelo com as variáveis escolhidas, explicitando na função a família de distribuição binomial, sendo a função de ligação *logit,* configurando assim uma regressão logística binária.

```
> # construção do modelo
> mod <- glm(romantic ~ absences + G3,
+            family = binomial(link = 'logit'), data = dados2)
```

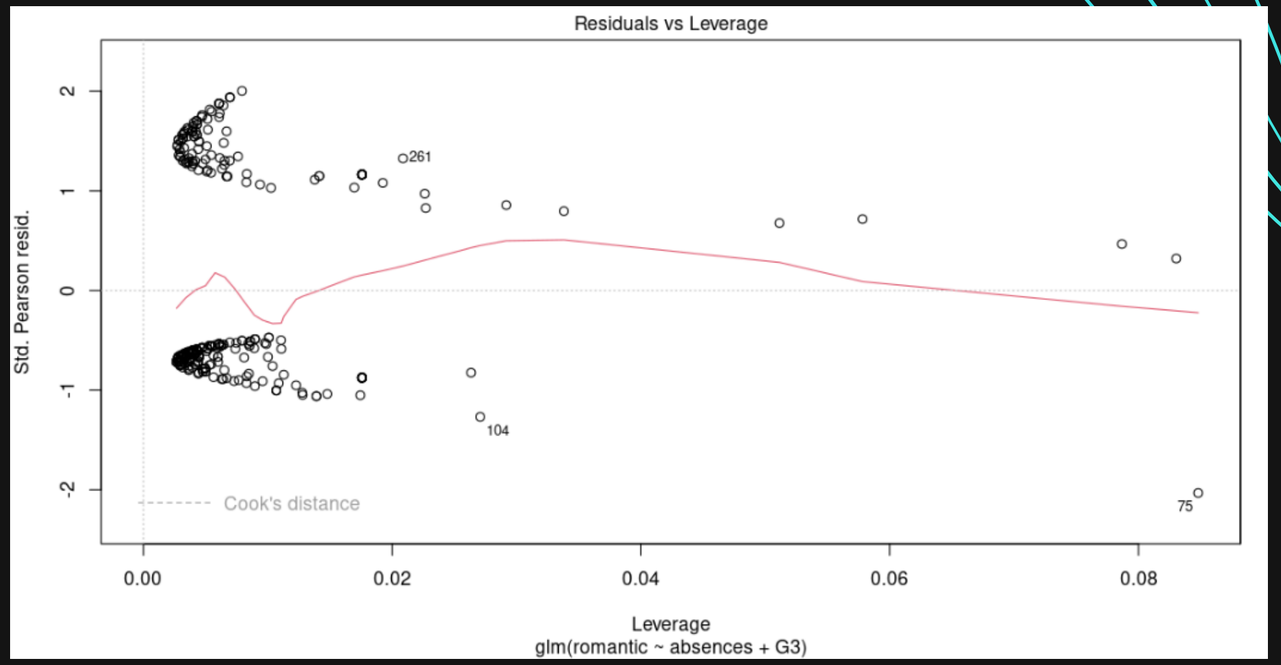
Regressão Logística

CHECAGEM DE PRESSUPOSTOS

3) Ausência de outliers ou pontos de alavancagem

- Para avaliar a existência de outliers, foi realizado um gráfico que considera os registros fora da média ou atípicos aqueles que se encontram abaixo da linha vermelha pontilhada. Esta não aparece em nosso diagrama, o que mostra que esse pressuposto foi atendido.

```
plot(mod, which = 5)
```



Regressão Logística

CHECAGEM DE PRESSUPOSTOS

3) Ausência de outliers ou pontos de avalancagem

- Avalia-se, também, a análise do resíduo padronizado, que tem seus valores mínimo e máximo entre -3 e +3, apresentando novamente que os valores estão dentro de um intervalo aceitável.

```
> summary(stdres(mod))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-2.0245388 -0.6984311 -0.5945901 -0.0002112  1.1595731  1.9956683
```

Regressão Logística

CHECAGEM DE PRESSUPOSTOS

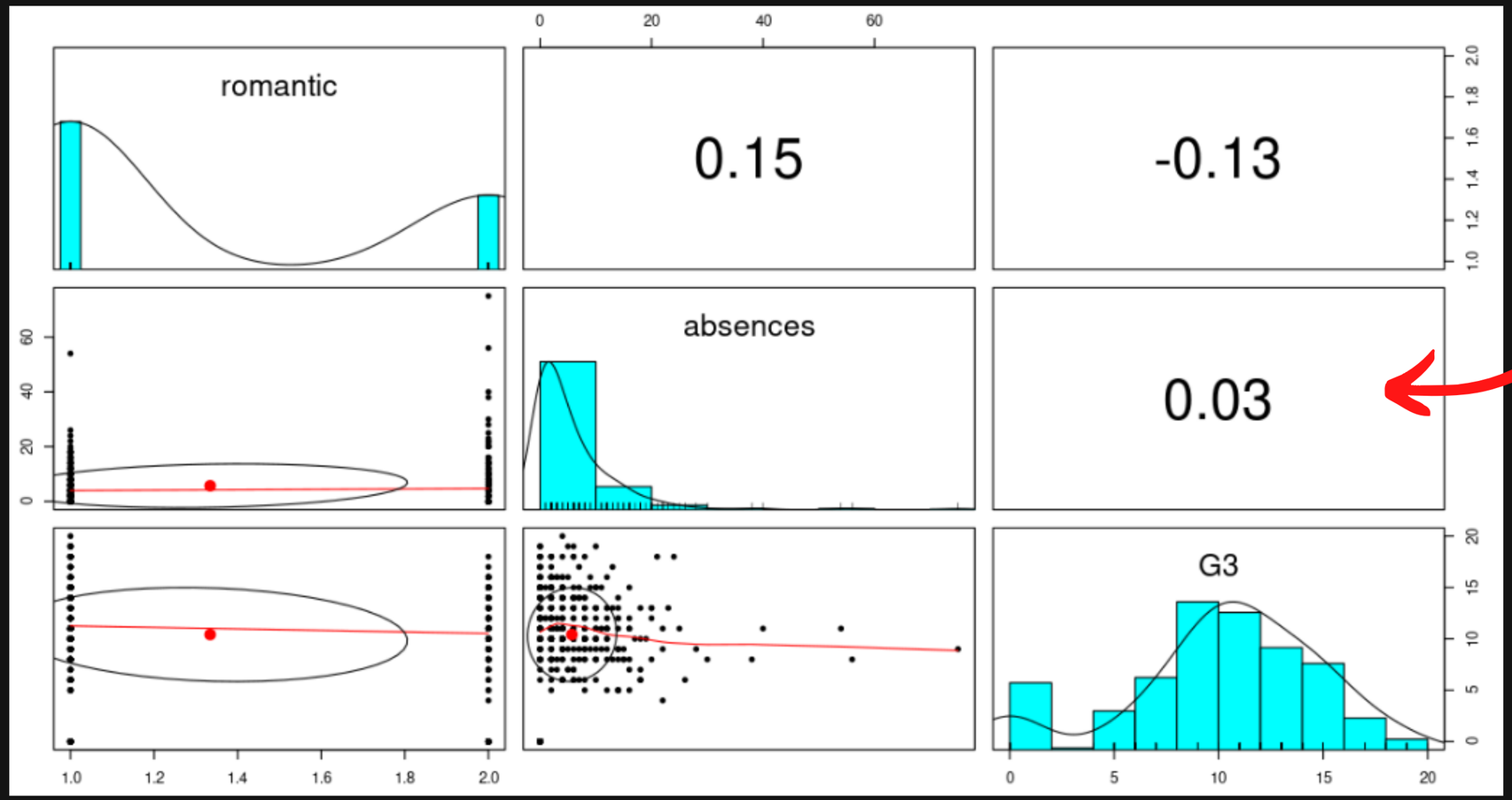
4) Ausência de multicolinearidade

- Nesse passo, pretende-se avaliar se existe alguma correlação alta entre as variáveis independentes por si só. Utilizando a função *pairs.panels*, pensando a partir da correlação de Pearson, que o valor aceitável dessa multicolinearidade é até 0,8. Entre as variáveis G3 e absences, tem-se o valor 0,03, evidenciando que esta não existe de maneira significativa.

```
pairs.panels(dados2)
```

Regressão Logística

CHECAGEM DE PRESSUPOSTOS



Regressão Logística

CHECAGEM DE PRESSUPOSTOS

4) Ausência de multicolinearidade

- Podemos avaliar a ausência de multicolinearidade também através do fator de inflação, que indica um problema de multicolinearidade quando está acima de 10, o que também não foi indicado aqui.

```
> vif(mod)
absences       G3
1.008133 1.008133
```

Regressão Logística

CHECAGEM DE PRESSUPOSTOS

5) Relação linear entre cada variável independente contínua e o logito da dependente

- Realização do teste de Box-Tidell - inclusão no modelo a interação de cada variáel contínua e seu próprio log.

```
> intlog <- dados2$absences * log(dados2$absences)
> dados2$intlog <- intlog
> intlog2 <- dados2$G3 * log(dados2$G3)
> dados2$intlog2 <- intlog2
> modint <- glm(romantic ~ G3 + absences + intlog + intlog2,
+               family = binomial(link='logit'), data = dados2)
```

Regressão Logística

CHECAGEM DE PRESSUPOSTOS

Se interação não significativa (p value acima de 0,05 e coeficientes positivos - ou próximos disto), teste obteve sucesso. Esse resultado só foi atingido com a interação de absences e si mesma, mas por pouco não com G3 e si mesma.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8290  -0.8355  -0.6754   1.1971   2.2189

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.32145    2.77685  -2.637  0.00837 **
G3           1.69670    0.84142   2.016  0.04375 *
absences     0.26693    0.10729   2.488  0.01285 *
intlog      -0.05196    0.02769  -1.876  0.06060 .
intlog2     -0.49662    0.24631  -2.016  0.04377 *
```

Regressão Logística

CHECAGEM DE PRESSUPOSTOS

5) Relação linear entre cada variável independente contínua e o logito da dependente

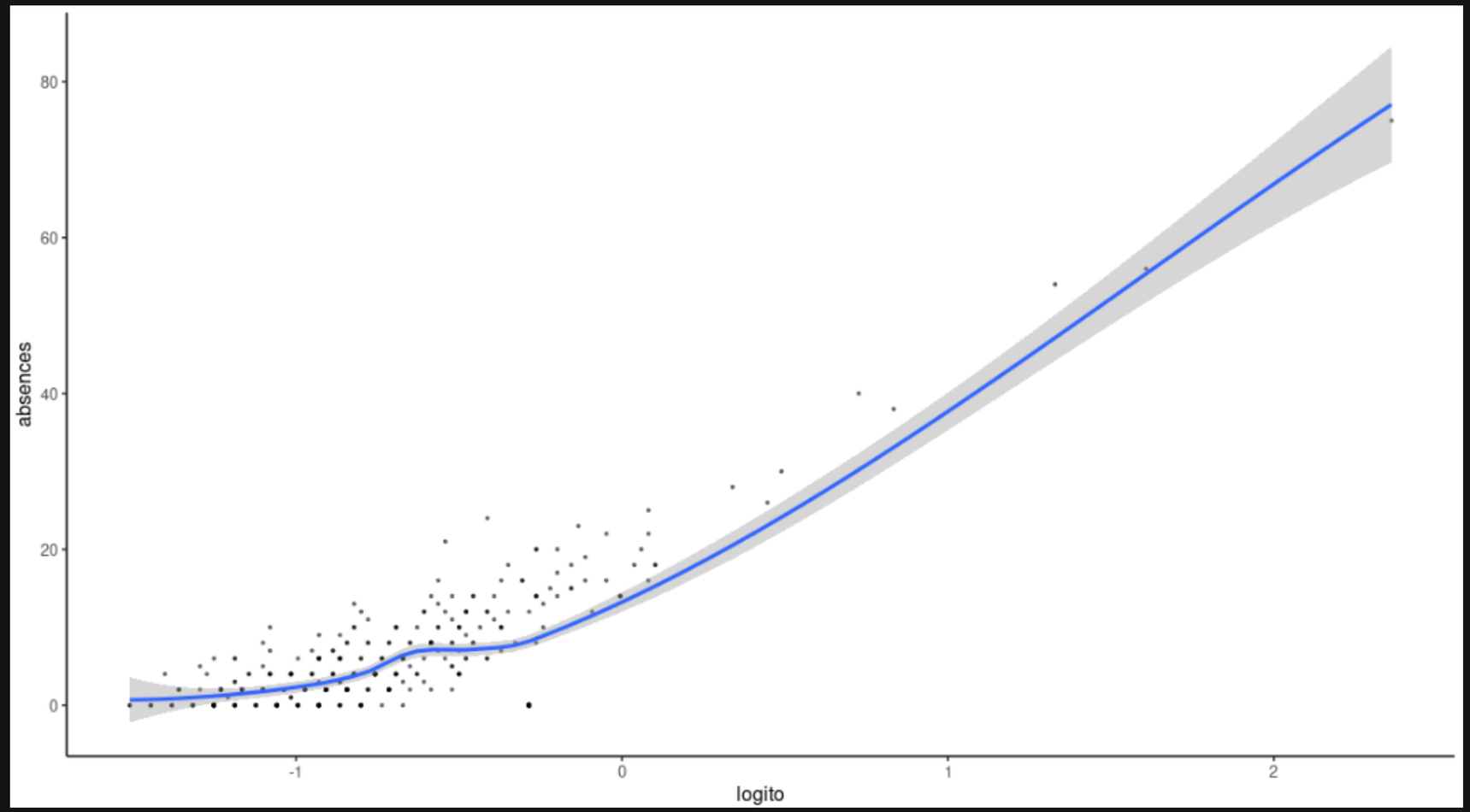
- Realização do logito da variável dependente, armazenado no dataset como nova coluna de dados. Para ver a relação entre o logito e cada uma das variáveis independentes, plota-se um gráfico, e quanto mais o plot se aproximar a uma reta, melhor atende-se ao pressuposto de linearidade.

```
> logito <- mod$linear.predictors
> dados2$logito <- logito
> ggplot(dados2, aes(logito, absences)) +
+   geom_point(size = 0.5, alpha = 0.5) +
+   geom_smooth(method = 'loess') +
+   theme_classic()
`geom_smooth()` using formula = 'y ~ x'
> ggplot(dados2, aes(logito, G3)) +
+   geom_point(size = 0.5, alpha = 0.5) +
+   geom_smooth(method = 'loess') +
+   theme_classic()
`geom_smooth()` using formula = 'y ~ x'
```

Regressão Logística

CHECAGEM DE PRESSUPOSTOS

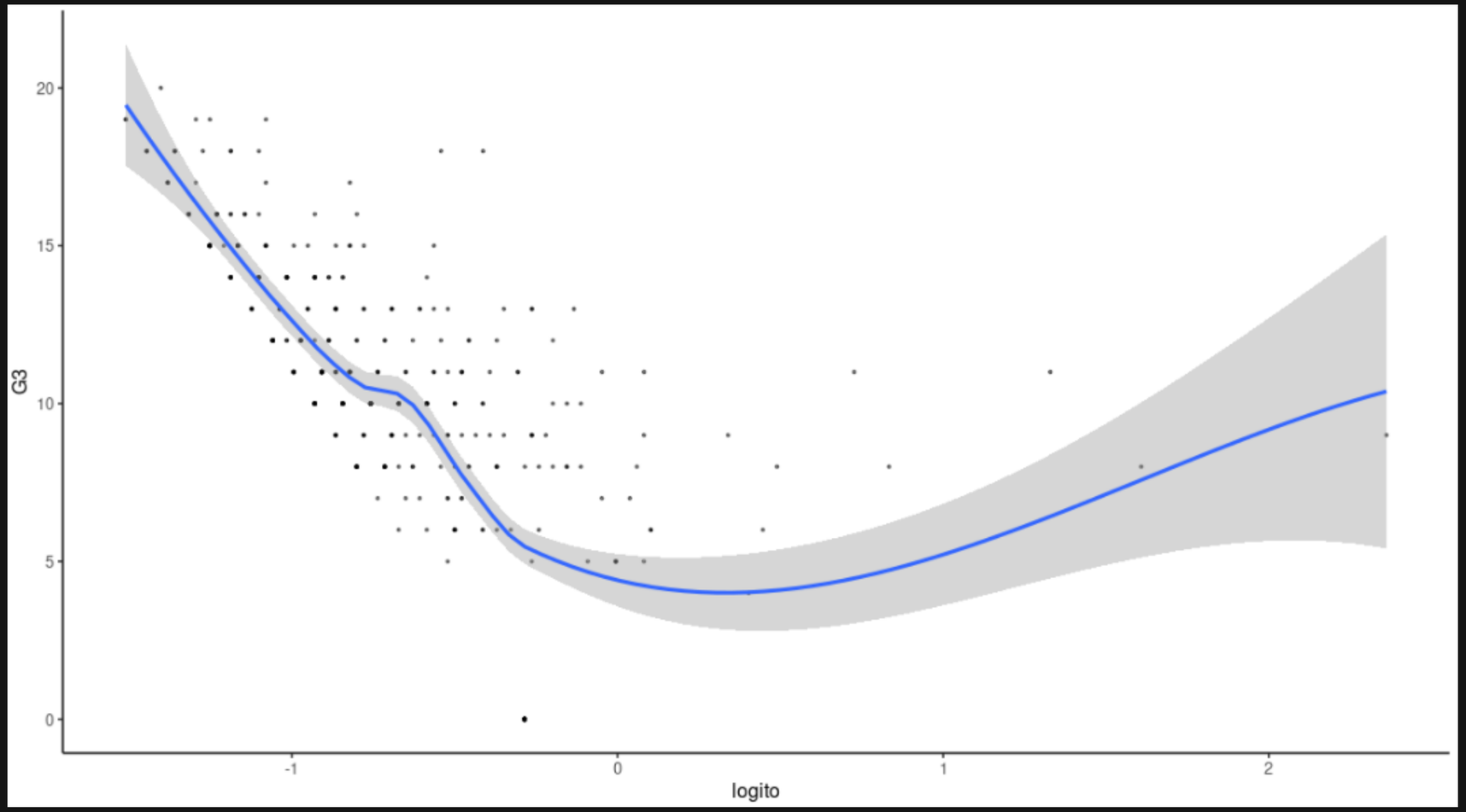
absences x logito



Regressão Logística

CHECAGEM DE PRESSUPOSTOS

G3 x logito



Regressão Logística

ANÁLISE DO MODELO E INTERPRETAÇÃO DE RESULTADOS

1) Efeitos gerais

- Valores de p menores que 0,05, sendo previsores estatisticamente significativos na previsão de estar ou não em uma relação amorosa.

```
> Anova(mod, type="II", test="Wald")
Analysis of Deviance Table (Type II tests)

Response: romantic
          Df  Chisq Pr(>Chisq)
absences   1 8.5212   0.003510 **
G3         1 7.4169   0.006461 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regressão Logística

ANÁLISE DO MODELO E INTERPRETAÇÃO DE RESULTADOS

2) Efeitos específicos

- Resíduos dentro do intervalo -3 e +3 esperado
- A coluna *Estimate* apresenta os coeficientes das variáveis, que sendo valores positivos, indicam que tanto a variável *G3* quando a *absence* apresentam influência sobre a variável *romantic*. O resultado alcançado reforça as hipóteses levantadas na checagem de pressupostos.

```
> summary(mod)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7685  -0.8925  -0.7790   1.3007   1.7917

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.28520    0.26782  -1.065  0.28692
absences     0.04302    0.01474   2.919  0.00351 **
G3          -0.06445    0.02366  -2.723  0.00646 **
```

Regressão Logística

ANÁLISE DO MODELO E INTERPRETAÇÃO DE RESULTADOS

3) Odds Ratio - interpretação dos coeficientes: razão das chances

- Uso do erro padrão no cálculo SPSS e intervalo de confiança de 95% - obtido através da função *confint.default*
- O coeficiente é colocado como expoente do valor *e* (Euler) - obtido através da função *coef*.
- Resultado demonstra o quanto as variáveis interferem na dependente - *absences* mais que *G3*, mas de intensidade muito reduzida.

```
> exp(cbind(OR = coef(mod), confint.default(mod)))
                    OR     2.5 %    97.5 %
(Intercept) 0.7518632 0.4448074 1.2708831
absences    1.0439577 1.0142353 1.0745511
G3          0.9375846 0.8950909 0.9820957
```

Regressão Logística

ANÁLISE DO MODELO E INTERPRETAÇÃO DE RESULTADOS

4) Odds Ratio - utilizando log-likelihood

- O coeficiente é colocado como expoente do valor e (Euler) - obtido através da função *coef*.
- Resultado demonstra o quanto as variáveis interferem na dependente - *absences* mais que *G3*, mas de intensidade muito reduzida.

```
> exp(cbind(OR = coef(mod), confint(mod)))
Waiting for profiling to be done...
                    OR     2.5 %    97.5 %
(Intercept) 0.7518632 0.4418365 1.2678007
absences    1.0439577 1.0156989 1.0762769
G3          0.9375846 0.8947899 0.9820136
```

Regressão Logística

ANÁLISE DO MODELO E INTERPRETAÇÃO DE RESULTADOS

5) Pseudo r2

- Adaptação do r2 da regressão linear, numa tentativa de demonstrar o quanto a variável independente varia de acordo com o modelo. Aqui, realizado através do método *Nagelkerke*, tendo um valor ajustado entre 0 e 1 e, ainda assim, apresenta um valor muito baixo, retratando que a variável não varia significativamente conforme o modelo, demonstrando que o resultado não foi favorável à nossa pergunta.

```
> PseudoR2(mod, which = "Nagelkerke")
Nagelkerke
0.05652436
```

Regressão Logística

CONCLUSÃO

Tendo em vista a pergunta inicial "Alunos que não faltaram e tiraram notas acima da média possuem maior probabilidade de não estar em uma relação amorosa?", entende-se que a resposta seja negativa, ou seja, as variáveis número de faltas e nota final não interferem na presença ou não de relacionamentos amorosos.

Ainda que a variável *absences* apresente resultados mais favoráveis que a *G3*, entende-se que o modelo gerado falhou em tentar explicar a variação da variável dependente *romantic.*

Bibliotecas utilizadas

- pacman: uma ferramenta de gerenciamento de pacotes R que combina a funcionalidade das funções relacionadas à biblioteca base em funções nomeadas intuitivamente
- car: funções para Acompanhar J. Fox e S. Weisberg, An R Companion to Applied Regression, Terceira Edição, Sage, 2019.
- descTools: é uma extensa coleção de diversas funções estatísticas básicas e wrappers não disponíveis no sistema básico R para descrição eficiente de dados.
- biocManager: permite aos usuários instalar e gerenciar pacotes do projeto Bioconductor, que se concentra na análise estatística e na compreensão de dados genômicos de alto rendimento.

Referências

Alcohol Effects On Study. Disponível em: <https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study?resource=download>. Acesso em: 17 nov. 2022.

DE AZEVEDO GONZALEZ, L. [s.l: s.n.]. Disponível em: <https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf>.

FERNANDES, A. A. T. et al. Read this paper if you want to learn logistic regression. Revista de Sociologia e Política, v. 28, n. 74, 2020.

BATTISTI, I. D. E.; SMOLSKI, F. M. DA S. Capítulo 7 Regressão Logística | Software R: curso avançado. [s.l: s.n.].

O que é regressão logística? Disponível em: <https://www.tibco.com/pt-br/reference-center/what-is-logistic-regression>.

DOS, C. modelo estatístico. Disponível em: <https://pt.wikipedia.org/wiki/Regress%C3%A3o_log%C3%ADstica>. Acesso em: 17 nov. 2022.

Field, A. P., Miles, J., & Field, Z. (2012). Discovering statistics using R/Andy Field, Jeremy Miles, Zoë Field.

Obrigada!