FGV EMAp
ESCOLA DE MATEMÁTICA APLICADA

# Trabalho A2 - Modelagem Estatística

Edilton Brandão de Sousa

Rio de Janeiro
2023

Edilton Brandão de Sousa

# Trabalho A2 - Modelagem Estatística

Este trabalho visa empregar os princípios de modelagem estatística, adquiridos ao longo do curso, para analisar e interpretar um conjunto de dados específico, com o objetivo de responder a uma pergunta determinada, ressaltando a importância dos resultados alcançados e a relevância do problema em estudo.

Professor: Luiz Max F. Carvalho

Rio de Janeiro
2023

# Sumário

# 1 Introdução

## 1.1 Problema e sua relevância

O ambiente escolar é um cenário complexo, onde diversos fatores podem influenciar o desempenho acadêmico dos alunos. Aspectos como renda, escolaridade dos pais, idade, horas de estudo, têm sido amplamente estudados como determinantes do sucesso educacional. Por outro lado, compreender como o envolvimento romântico dos alunos afeta seu desempenho acadêmico é uma questão bastante interessante e de grande relevância, uma vez que pode ter implicações tanto para o bem-estar dos alunos quanto para as estratégias de ensino e apoio educacional.

Nesse contexto, esse trabalho tem como objetivo investigar o impacto dos relacionamentos amorosos no desempenho acadêmico dos alunos. Pretendo analisar se a presença de um relacionamento amoroso está associada a variações nas notas dos alunos, controlando outros fatores relevantes que também podem influenciar o desempenho escolar.

Para alcançar esses objetivos, utilizarei as principais técnicas de modelagem estatística estudadas durante do curso e análise de dados para explorar um conjunto de dados abrangente sobre alunos de duas escolas. Faremos uso de variáveis como gênero, idade, nível de educação dos pais, qualidade das relações familiares, tempo de estudo, consumo de álcool, entre outros, para controlar possíveis fatores de confusão e obter estimativas do efeito dos relacionamentos amorosos no desempenho acadêmico.

## 1.2 Dados

Os dados utilizados neste estudo se referem ao desempenho dos alunos no ensino secundário de duas escolas públicas localizadas na região do Alentejo, em Portugal, durante o ano letivo de 2005-2006. Os conjuntos de dados foram coletados por meio de relatórios escolares e questionários e incluem notas dos alunos, características socioeconômicas, demográficas e comportamentais.

Os dados originais estavam divididos em dois conjuntos, um para a disciplina de Matemática (student-mat) e outro para a disciplina de Língua Portuguesa (student-por). Os conjuntos de dados foram utilizados em estudos anteriores, como mencionado em [1][Cortez e Silva, 2008], que abordaram tarefas de classificação binária/cinco níveis e regressão.

Uma observação importante é que o atributo alvo G3, que representa a nota final dos alunos, possui uma forte correlação com os atributos G2 e G1. Essa correlação existe porque o G3 é a nota final obtida no 3º período, enquanto G1 e G2 correspondem às notas obtidas nos 1º e 2º períodos, respectivamente. Portanto, prever o G3 sem levar em conta G2 e G1 é mais difícil, mas é uma previsão mais útil para análises posteriores.

Tabela 1: Descrição dos atributos

| **Atributo** | **Descrição (Domínio)** |
|---|---|
| sex | Sexo do estudante (binário: feminino ou masculino) |
| age | Idade do estudante (15 a 22) |
| school | Escola do estudante (binário: Gabriel Pereira ou Mousinho da Silveira) |
| address | Tipo de endereço residencial do estudante (binário: urbano ou rural) |
| Pstatus | Situação de convivência dos pais (binário: morando juntos ou separados) |
| Medu | Educação da mãe do estudante (numérico: de 0 a 4)[a] |
| Mjob | Profissão da mãe do estudante (nominal)[b] |
| Fedu | Educação do pai do estudante (numérico: de 0 a 4)[a] |
| Fjob | Profissão do pai do estudante (nominal)[b] |
| guardian | Responsável pelo estudante (nominal: mãe, pai ou outro) |
| famsize | Tamanho da família (binário: $\leq 3$ ou $> 3$) |
| famrel | Qualidade dos relacionamentos familiares (numérico: de 1 - 5) |
| reason | Motivo para escolher esta escola (nominal: distância, curso,etc) |
| traveltime | Temp.casa p/escola (1 - <15 min., 2 - 15 a 30 min., 3 - 30 min. a 1 hr ou 4 - >1 hr) |
| studytime | Temp.estudo semanal (num: 1 - <2 hrs, 2 - 2 a 5 hrs, 3 - 5 a 10 hrs ou 4 - >10 hrs) |
| failures | Número de reprovações em períodos anteriores (numérico: n se $1 \leq$ n <3, senão 4) |
| schoolsup | Suporte educacional extra da escola (binário: sim ou não) |
| famsup | Suporte educacional familiar (binário: sim ou não) |
| activities | Atividades extracurriculares (binário: sim ou não) |
| paidclass | Aulas extras pagas (binário: sim ou não) |
| internet | Acesso à internet em casa (binário: sim ou não) |
| nursery | Frequência à escola maternal (binário: sim ou não) |
| higher | Deseja prosseguir com ensino superior (binário: sim ou não) |
| romantic | Relacionamento romântico (binário: sim ou não) |
| freetime | Tempo livre após a escola (num: de 1 - muito baixo a 5 - muito alto) |
| goout | Sair com amigos (numérico: de 1 - muito baixo a 5 - muito alto) |
| Walc | Consumo de álcool nos finais de semana (num: de 1 - muito baixo a 5 - muito alto) |
| Dalc | Consumo de álcool nos dias úteis (num: de 1 - muito baixo a 5 - muito alto) |
| health | Estado de saúde atual (numérico: de 1 - muito ruim a 5 - muito bom) |
| absences | Número de faltas na escola (numérico: de 0 a 93) |
| G1 | Nota do primeiro período (numérico: de 0 a 20) |
| G2 | Nota do segundo período (numérico: de 0 a 20) |
| G3 | Nota final (numérico: de 0 a 20) |

*a*: 0 – nenhum, 1 – educação primária (4ª série), 2 – 5ª a 9ª série, 3 – educação secundária ou 4 – educação superior.
*b*: professor, relacionado à área de saúde, serviços públicos (por exemplo, administrativo ou policial), em casa ou outro.

# 2 Metodologia

## 2.1 Modelo

O modelo utilizado para esse estudo foi a regressão linear multivariada, uma escolha fundamentada na natureza contínua da variável dependente - a média das notas G1, G2 e G3, aqui denominada como GF - e na suposição de existência de uma relação

linear entre as variáveis envolvidas. Outro fator relevante que reforça a escolha deste modelo é que a variável resposta GF apresenta uma distribuição normal (resultado observado durante o processo de análise exploratória dos dados), premissa fundamental para a eficácia da regressão linear.

No entanto, foi considerado importante excluir as notas G1 e G2 da análise devido à sua alta correlação com G3, o que resultou num coeficiente de determinação ($R^2$) do modelo de regressão linear ligeiramente inferior. Apesar disso, essa diminuição no $R^2$ não é particularmente preocupante no contexto do nosso estudo, já que o objetivo principal é avaliar o impacto da variável 'romantic', em conjunto com outras variáveis pertinentes.

## 2.2 Ajuste

O ajuste do modelo de regressão linear foi realizado utilizando o método dos mínimos quadrados. A função *lm* do pacote estatístico R foi empregada para realizar esse ajuste.

```
# importando os dados
data = read_delim("data/student-mat.csv", delim = ";")

# realizando a regressao linear usando a funcao lm
modelo <- lm(GF ~ romantic + failures + ... , data)

# resumo do modelo para ver os coeficientes e outras informacoes
summary(modelo)
```

Incluímos a variável *romantic* como preditora principal do desempenho acadêmico (GF), juntamente com outras variáveis independentes relevantes identificadas durante a análise exploratória.

## 2.3 Avaliação do modelo

Iniciamos a avaliação do modelo através da análise dos resíduos. Esta análise é feita tanto visualmente, pela inspeção da distribuição dos resíduos em gráficos, quanto por testes estatísticos, nos quais verificamos se os resíduos seguem uma distribuição normal - uma pressuposição fundamental para a validade de nossas inferências.

Posteriormente, partimos para o cálculo de métricas específicas que avaliam a eficácia do modelo. Adotamos o coeficiente de determinação ($R^2$) como uma dessas métricas, que nos dá uma ideia da porcentagem da variação do desempenho acadêmico que é explicada pelas variáveis independentes incorporadas no modelo.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \tag{1}$$

onde:

- $SS_{\text{res}}$ é a soma dos quadrados dos resíduos, dado por $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,
- $SS_{\text{tot}}$ é a soma total dos quadrados, dado por $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Outra métrica a ser utilizada é o Erro Quadrático Médio da Raiz (Root Mean Squared Error - RMSE). Esta métrica nos fornece uma medida da magnitude dos erros do nosso modelo, sendo expressa na mesma unidade que a variável resposta, o que facilita a sua interpretação.

O RMSE é calculado utilizando a seguinte fórmula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

onde:

- $y_i$ são os valores observados,
- $\hat{y}_i$ são os valores previstos pelo modelo,
- $n$ é o número total de observações.

Um RMSE menor indica um ajuste do modelo mais preciso aos dados. No entanto, é importante notar que o RMSE não pode ser analisado de forma isolada, mas deve ser interpretado no contexto das variáveis dependentes específicas e da escala dos dados.

Durante o processo de seleção do modelo, consideramos várias especificações possíveis e nos apoiamos no Critério de Informação de Akaike (AIC) como uma métrica para comparar o ajuste e a complexidade desses modelos. O AIC é calculado como:

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{3}$$

onde:

- $k$ é o número de parâmetros estimados no modelo,
- $\hat{L}$ é o valor máximo da função de verossimilhança para o modelo estimado.

O AIC tem o objetivo de penalizar a complexidade do modelo (representada pelo número de parâmetros) e recompensar o bom ajuste do modelo (representado pela verossimilhança). Em geral, o modelo com o menor valor de AIC é preferido, indicando um bom equilíbrio entre ajuste e complexidade.

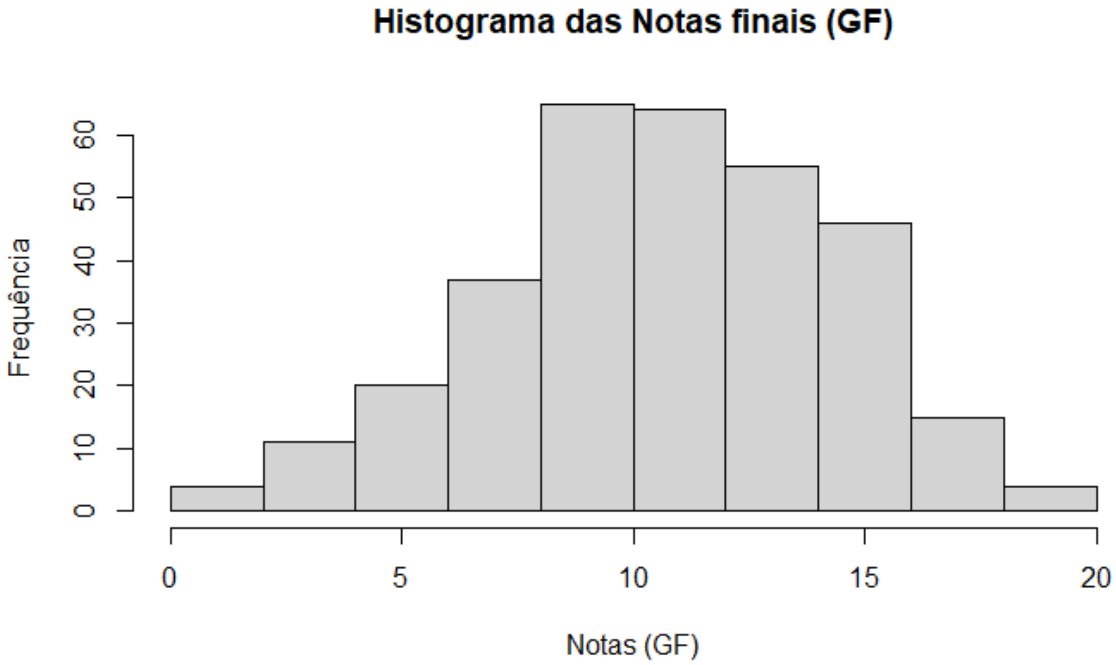
Além disso, utilizamos os valores p associados a cada coeficiente para determinar se estes são estatisticamente significativos, ou seja, se diferem de zero de uma maneira que não pode ser atribuída apenas à variação aleatória. Damos uma atenção especial ao p-valor correspondente à variável 'romantic', considerando o foco principal de nossa análise.

# 3 Resultados

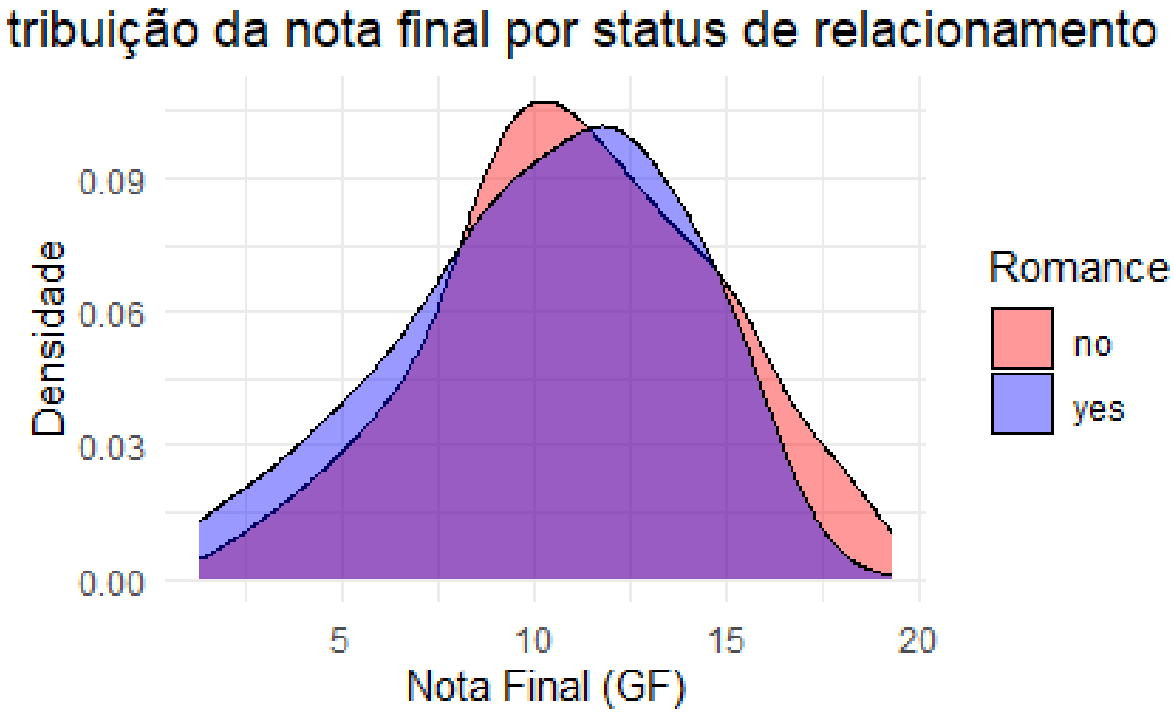
## 3.1 Análise exploratória

**Dados ausentes e modificações**

O conjunto de dados encontra-se limpo, sem valores faltantes. Foi criada uma nova coluna chamada GF nos dados com as médias das notas G1, G2 e G3. Essa será a variável resposta principal.

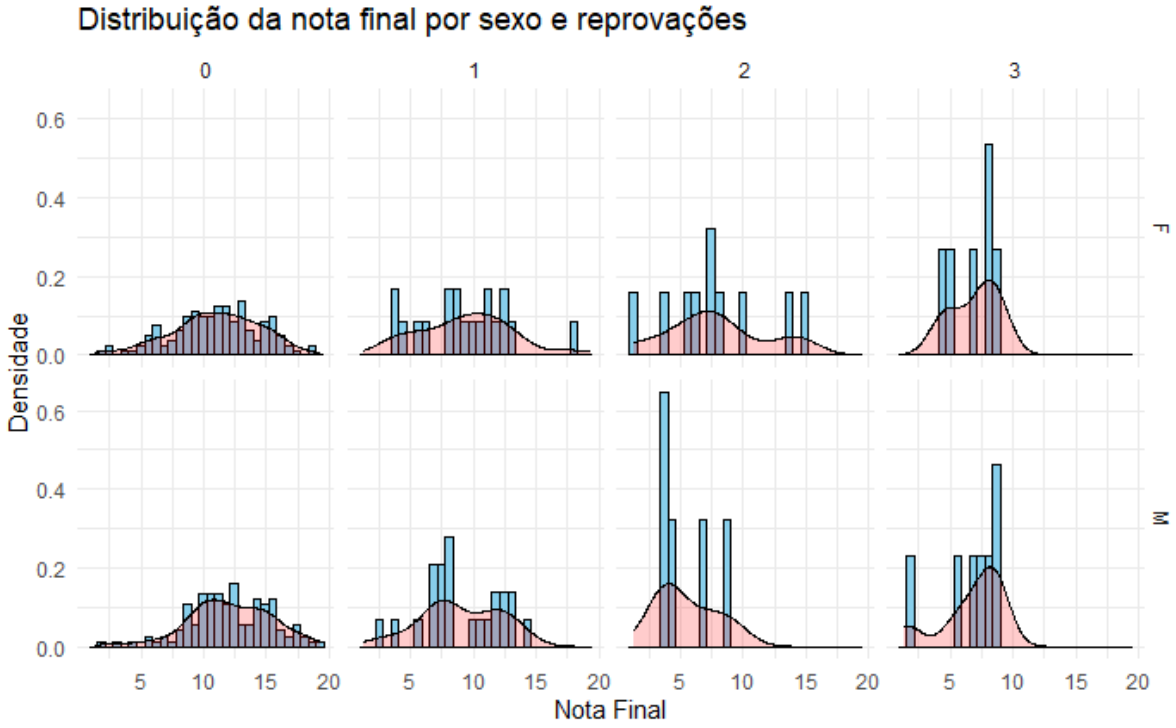



Essa foi uma parte bastante importante que reforçou a escolha do modelo como uma regressão linear, pois as notas aparentemente seguiam uma distribuição normal.

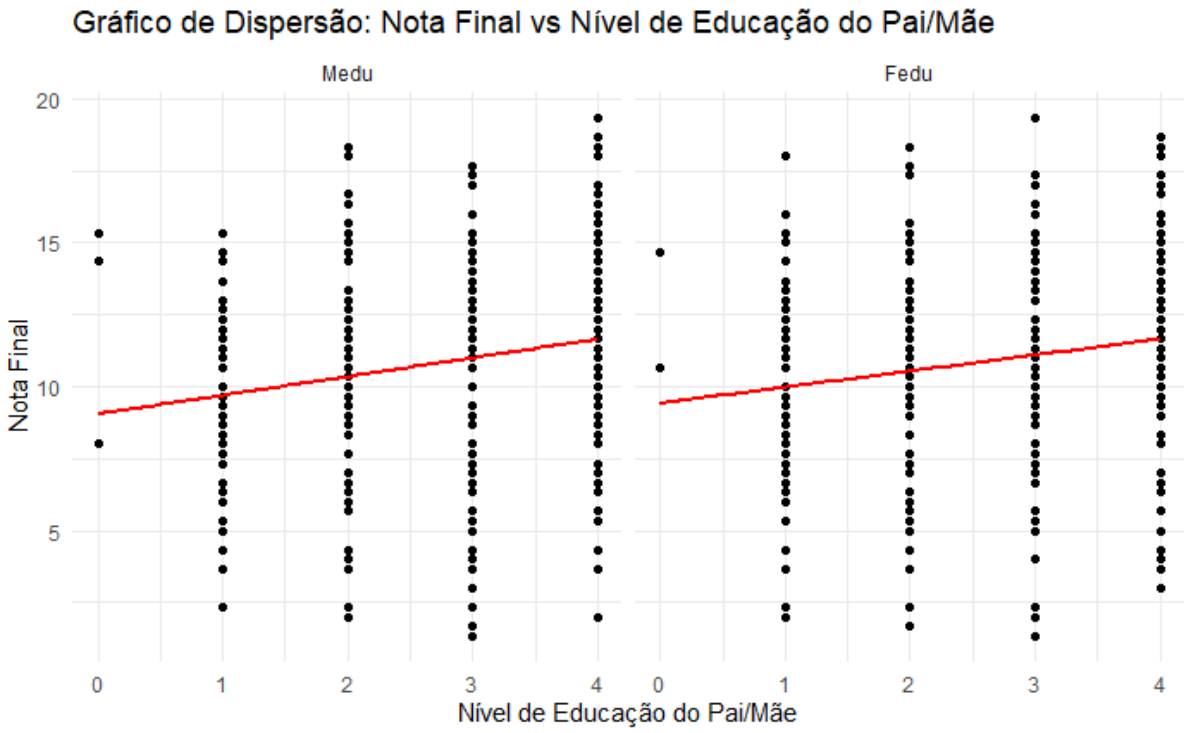



Resolvi analisar esse gráfico, pois a quantidade de estudantes que não namoram era mais que o dobro dos que namoram, e esse gráfico não leva em conta esse pro-

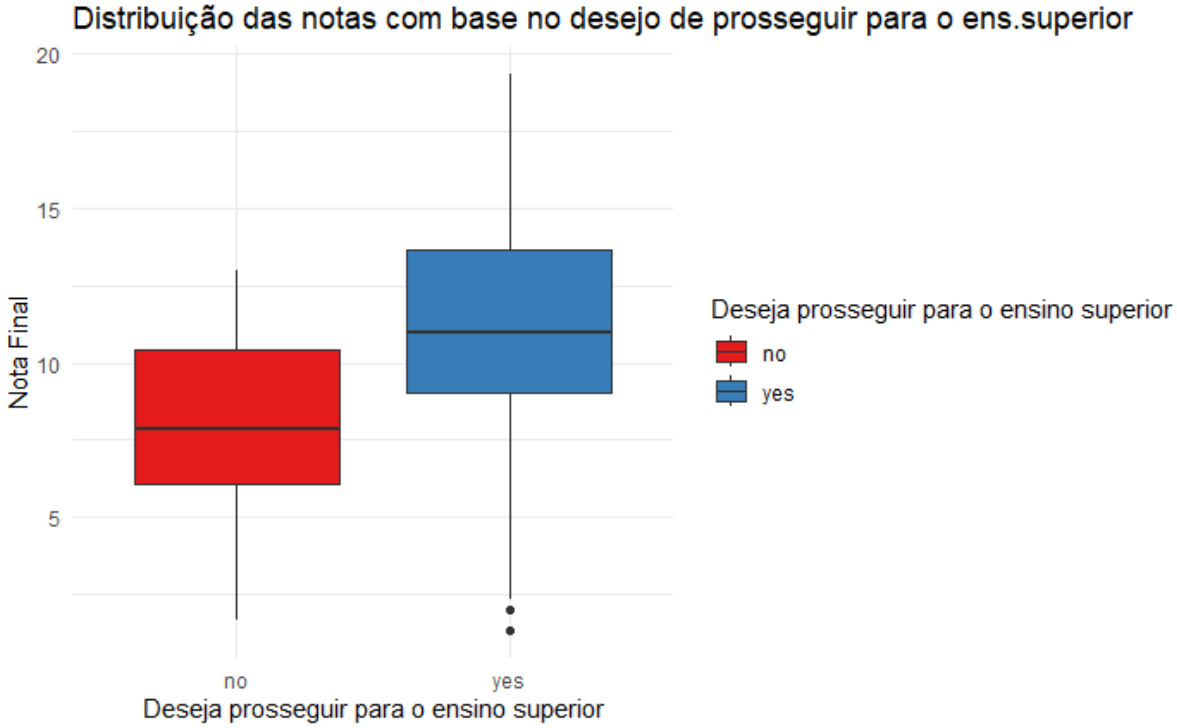
blema. Nota-se que a média das notas dos estudantes que estão em um relacionamento é maior do que daqueles que não estão. No entanto, podemos observar que a distribuição do gráfico azul (estudantes que estão em um relacionamento) tem uma "cauda"mais pesada à esquerda, enquanto no gráfico vermelho ocorre exatamente o contrário. Logo, não conseguimos ainda tirar muitas conclusões sobre esses dados, pois as médias podem ser influenciadas por valores extremos que não conseguimos observar aqui.




Em ambos os sexos, observa-se que as notas dos alunos que mais reprovaram em anos anteriores estão distribuídas nas notas mais baixas.




Observa-se que o nível de educação dos pais está positivamente relacionado ao desempenho escolar dos filhos. Pais com níveis mais altos de educação podem influenciar positivamente o aprendizado de seus filhos.




Realmente, é de se esperar que estudantes que desejam cursar um ensino superior estejam mais motivados a estudar e, consequentemente, tendam a tirar as melhores notas.

Referência que serviu de inspiração para essa análise exploratória: [2]:

## 3.2 Modelo ajustado e suas previsões

Segue abaixo um resumo da comparação de todos os modelos testados:

Tabela 2: Resultados dos modelos

| Model | R-squared | AIC | RMSE |
|---|---|---|---|
| Model 1 | 0.1716043 | 1693.146 | 3.319019 |
| Model 2 | 0.1759601 | 1691.454 | 3.310281 |
| Model 3 | 0.2095538 | 1688.093 | 3.242104 |
| Model 4 | 0.1908876 | 1691.586 | 3.280161 |
| Model 5 | 0.2066570 | 1687.268 | 3.248039 |
| Model 6 | 0.2262052 | 1681.259 | 3.207773 |

O modelo escolhido foi o 6:

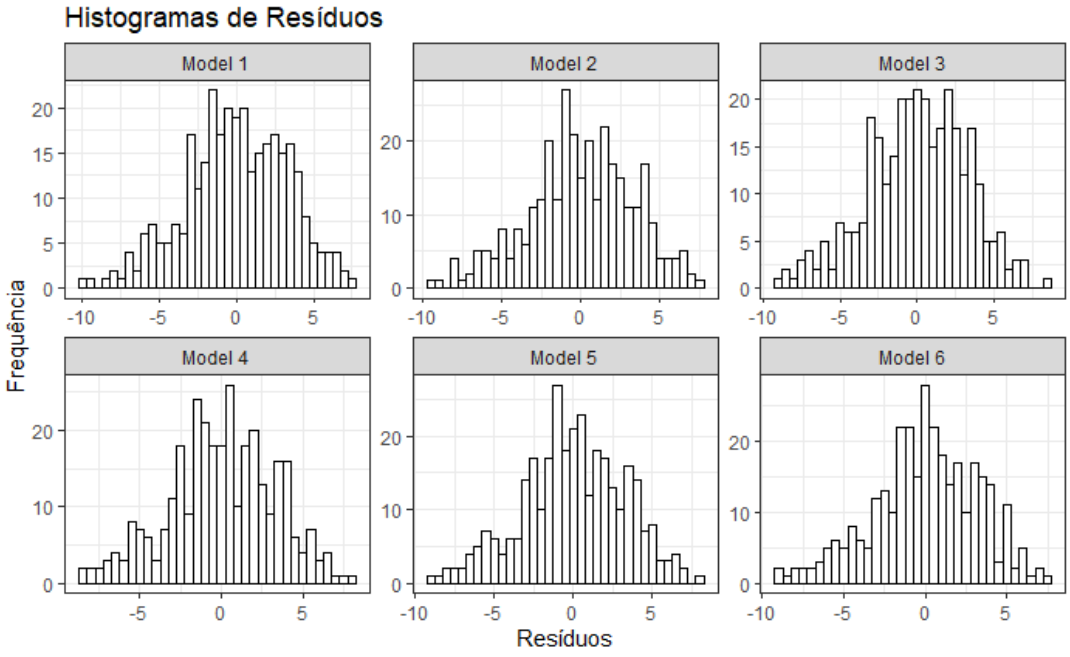
```
Call:
lm(formula = GF ~ failures + Medu * higher + schoolsup + romantic +
    Walc * goout, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9237 -1.7833  0.1145  2.3515  7.6011

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     16.5622     2.1182   7.819 8.33e-14 ***
failures        -1.6563     0.2604  -6.360 7.20e-10 ***
Medu            -1.6589     0.8277  -2.004  0.04592 *
higheryes       -2.5607     1.8387  -1.393  0.16473
schoolsupyes    -1.5439     0.5508  -2.803  0.00538 **
romanticyes     -0.6755     0.3908  -1.728  0.08489 .
Walc            -1.0234     0.4749  -2.155  0.03191 *
goout           -1.0558     0.3496  -3.020  0.00274 **
Medu:higheryes   2.0861     0.8416   2.479  0.01371 *
Walc:goout       0.3032     0.1282   2.366  0.01861 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 311 degrees of freedom
Multiple R-squared:  0.2262,  Adjusted R-squared:  0.2038
F-statistic:  10.1 on 9 and 311 DF,  p-value: 1.211e-13
```

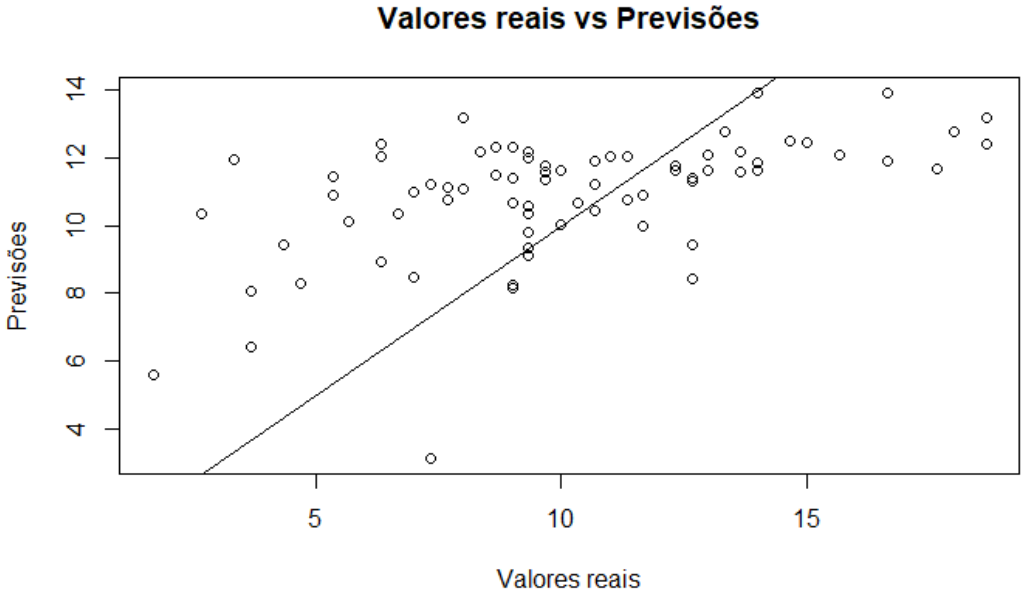
Resíduos de todos os modelos:




Aparentemente, os resíduos estão seguindo uma distribuição normal, o que é ótimo para a nossa suposição do modelo.

## 3.3 Previsões

Fizemos previsões sobre o conjunto de testes que separamos dos dados completos durante o processo de AED:



# 4 Discussão

## 4.1 O que foi aprendido

Após todo o processo de análise exploratória, concluímos que a variável *romantic* tinha um efeito negativo nas notas, embora bastante pequeno, devido ao alto valor-p observado em todos os modelos testados.

Aqui está o coeficiente estimado para a variável *romantic* no modelo final, juntamente com o intervalo de confiança:

O coeficiente estimado de *romantic* é -0.6754887, com um intervalo de confiança de 95% entre -1.444433 e 0.09345548.

Comentários sobre cada coeficiente, erros padrão e valores-p das variáveis do nosso modelo final:

- O coeficiente estimado para *failures* é -1.6563, com um erro padrão de 0.2604 e um valor-p extremamente baixo ($p < 0.001$), indicando uma associação significativa e negativa entre o número de reprovações anteriores e as notas finais.
- O coeficiente estimado para *Medu* é -1.6589, com um erro padrão de 0.8277 e um valor-p de 0.04592, indicando uma associação negativa entre o nível de educação da mãe e as notas finais, embora com um nível de significância marginal.
- O coeficiente estimado para *higheryes* isoladamente é -2.5607, com um erro padrão de 1.8387 e um valor-p de 0.16473, sugerindo uma associação negativa entre a aspiração de ensino superior e as notas finais, mas sem um nível significativo de evidência estatística.
- O coeficiente estimado para *schoolsupyes* é -1.5439, com um erro padrão de 0.5508 e um valor-p de 0.00538, indicando uma associação significativa e negativa entre o suporte educacional extraescolar e as notas finais.
- O coeficiente estimado para *romanticyes* é -0.6755, com um erro padrão de 0.3908 e um valor-p de 0.08489, sugerindo uma possível associação negativa entre estar em um relacionamento e as notas finais, embora com um nível de significância marginal.
- O coeficiente estimado para *Walc* é -1.0234, com um erro padrão de 0.4749 e um valor-p de 0.03191, indicando uma associação significativa e negativa entre o consumo de álcool nos fins de semana e as notas finais.
- O coeficiente estimado para *goout* é -1.0558, com um erro padrão de 0.3496 e um valor-p de 0.00274, indicando uma associação significativa e negativa entre sair com os amigos e as notas finais.
- O coeficiente estimado para a interação entre *Medu* e *higheryes* é 2.0861, com um erro padrão de 0.8416 e um valor-p de 0.01371, sugerindo que a influência da educação da mãe nas notas finais pode variar dependendo da aspiração de ensino superior.
- O coeficiente estimado para a interação entre *Walc* e *goout* é 0.3032, com um erro padrão de 0.1282 e um valor-p de 0.01861, sugerindo que o efeito do consumo de

álcool nos fins de semana nas notas finais pode depender da frequência de sair com os amigos.

Uma vez que o intervalo de confiança para o coeficiente de *romantic* inclui o valor zero, não podemos rejeitar a hipótese nula de que o coeficiente é igual a zero. Isso significa que não há evidências de que a variável *romantic* tenha um efeito significativo nas notas.

### 4.2 Limitações e direções futuras

Nosso modelo está focado exclusivamente nas notas de matemática. É importante ressaltar que o comportamento que observamos pode ter um efeito diferente nas notas de português. Existem outros fatores e variáveis específicas que podem influenciar o desempenho dos alunos nessa disciplina.

Além disso, é importante destacar que nosso modelo atual apresenta algumas limitações em termos de capacidade preditiva. Como podemos observar no gráfico de previsões, as predições do modelo têm uma dispersão considerável em relação aos valores reais das notas.

Dessa forma, para direções futuras deste trabalho, é recomendado realizar análises separadas para as notas de português, a fim de obter insights mais abrangentes e precisos sobre essa disciplina específica. Além disso, é importante buscar maneiras de aprimorar o modelo existente, como a inclusão de novas variáveis relevantes ou a utilização de técnicas mais avançadas de modelagem. Explorar outras abordagens de modelagem também pode ser uma opção interessante para melhorar a capacidade preditiva do modelo.

## Referências


[1] P. Cortez e A. M. G. Silva. "Using Data Mining to Predict Secondary School Student Performance". Em: *Proceedings of 5th Annual Future Business Technology Conference*. Ed. por A. Brito e J. Teixeira. Porto, 2008.

[2] Hindelya. *Students Grade Prediction*. `https://www.kaggle.com/code/hindelya/students-grade-prediction`. Acesso em: 27 de junho de 2023.


Os dados utilizados e a análise exploratória completa pode ser encontrada no repositório do trabalho:

https://github.com/edilton-bs/a2stats-model.