**UNIVERSIDADE ESTADUAL DE MATO GROSSO DO SUL**
PRÓ REITORIA DE PESQUISA E PÓS-GRADUAÇÃO *STRICTU SENSU*
Mestrado Profissional em Matemática/PROFMAT
UNIDADE UNIVERSITÁRIA DE DOURADOS

# Evolução do rendimento escolar no Ensino Médio público do Brasil: Uma análise utilizando regressão linear

LEANDRO RODRIGO MORAIS
Mestrado Profissional em Matemática: PROFMAT/SBM

Orientadora: Prof.ª Dra. MARINA RODRIGUES MAESTRE

DOURADOS - 2021

**UNIVERSIDADE ESTADUAL DE MATO GROSSO DO SUL**
PRÓ REITORIA DE PESQUISA E PÓS-GRADUAÇÃO *STRICTU SENSU*
Mestrado Profissional em Matemática/PROFMAT
UNIDADE UNIVERSITÁRIA DE DOURADOS

# Evolução do rendimento escolar no Ensino Médio público do Brasil: Uma análise utilizando regressão linear

LEANDRO RODRIGO MORAIS
Mestrado Profissional em Matemática: PROFMAT/SBM

Dissertação apresentada ao curso de Mestrado Profissional em Matemática/PROFMAT da Universidade Estadual de Mato Grosso do Sul, como requisito parcial para obtenção do título de Mestre em Matemática.

BANCA EXAMINADORA

Prof.ª Dra. MARINA RODRIGUES MAESTRE
UEMS - Universidade Estadual de Mato Grosso do Sul

Prof. Dr. VANDO NARCISO
UEMS - Universidade Estadual de Mato Grosso do Sul

Prof. Dr. ALEXANDRE PITANGUI CALIXTO
UFGD - Universidade Federal da Grande Dourados

DOURADOS - 2021

M825 Morais, Leandro Rodrigo
Evolução do rendimento escolar no ensino médio público do Brasil: uma análise utilizando regressão linear / Leandro Rodrigo Morais. – Dourados, MS: UEMS, 2021.
69 p.

Dissertação (Mestrado) – Matemática – Universidade Estadual de Mato Grosso do Sul, 2021.
Orientadora: Prof. Drª Marina Rodrigues Maestre

1. Indicadores educacionais 2. Análise de regressão 3. Software R. I. Maestre, Marina Rodrigues II. Título

CDD 23. ed. – 510.07

| | | |
|---|---|---|
| UEMS Universidade Estadual de Mato Grosso do Sul | **UNIVERSIDADE ESTADUAL DE MATO GROSSO DO SUL**<br>**PROGRAMA DE MESTRADO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL - PROFMAT** | PROFMAT |

**LEANDRO RODRIGO MORAIS**

***EVOLUÇÃO DO RENDIMENTO ESCOLAR NO ENSINO MÉDIO PÚBLICO DO BRASIL: UMA ANÁLISE UTILIZANDO REGRESSÃO LINEAR***

Produto Final do Curso de Mestrado Profissional apresentado ao Programa de Pós-Graduação *Stricto Sensu* em Matemática em Rede Nacional, da Universidade Estadual de Mato Grosso do Sul, como requisito final para a obtenção do Título de Mestre em Matemática.

**Aprovado em: 14 de dezembro de 2021.**

**BANCA EXAMINADORA:**

Marina Rodrigues Maestre.
Profa. Dra. Marina Rodrigues Maestre (UEMS)
Universidade Estadual de Mato Grosso do Sul

Prof. Dr. Vando Narciso (UEMS)
Universidade Estadual de Mato Grosso do Sul
(participação realizada à distância por videoconferência)

Prof. Dr. Alexandre Pitangui Calixto (UFGD)
Universidade Federal da Grande Dourados
(participação realizada à distância por videoconferência)

*Dedico este trabalho primeiramente aquele que creio ser o senhor de todos nós, "Deus" e a Nossa Senhora Aparecida, por nos momentos difíceis terem atendido minhas preces. Aos ex-colegas de trabalho, familiares e colegas de longa data que infelizmente vieram a perder suas vidas devido ao vírus da Covid-19. Aos amigos e familiares, em especial a minha mãe Maria Ferreira do Nascimento, nobre senhora que nunca mediu esforços para que eu pudesse continuar meus estudos. E por fim aquela a quem escolhi a dividir minhas alegrias e angústias, minha amada esposa Mayara Freitas da Silva Morais, por não me abandonar e estar sempre me motivando a continuar e não permitir que eu distanciasse da realização de um dos meus sonhos.*

**Agradecimentos**

A Deus por ter me dado forças nas horas que pensei em desistir. A UEMS - Universidade Estadual de Mato Grosso do Sul, por ofertar o programa de mestrado que nos conduz a um aprendizado diferente e especial. A minha orientadora Prof.ª Dra. Marina Rodrigues Maestre que me auxiliou ao longo destes semestres para que eu pudesse concluir esta dissertação. Também agradeço a todos os professores que me acompanharam nesta jornada e aos meus amigos de turma pelos momentos inesquecíveis que passamos juntos, serão sempre lembrados.

*Ninguém vai bater tão forte como a vida, mas a questão não é o quão forte você consegue bater. É o quão forte você consegue apanhar e continuar seguindo em frente. A vitória é feita assim. A vida já é difícil normalmente, mas, em alguns momentos, ela pode até te derrubar. No entanto, é preciso levantar sempre. (Rocky Balboa)*

**Resumo:** De 2010 à 2019, segundo dados divulgados pelo INEP, houve uma redução no número de estudantes que constituem as turmas do Ensino Médio público do Brasil e ao mesmo tempo ocorreu um leve aumento nas horas aulas diárias estudadas. Também pode-se notar que a taxa de aprovação, nesse mesmo período, passou de 76% para 85%. Diante dessas informações surgem as seguintes questões: a quantidade de alunos em sala influencia na aprendizagem? Ou, aumentando a carga horária em sala irá resultar em uma melhor aprendizagem? Ou, ainda de uma forma mais enfática, diminuindo a quantidade de alunos em sala e aumentando a carga horária o resultado será satisfatório ao ponto de elevar a taxa de aprovação? Tais indagações foram analisadas e respondidas por meio de modelos de análise de regressão utilizando o *software* R. Os resultados obtidos por tais modelos afirmam que reduzindo o número de estudantes em sala e/ou aumentando a carga horária de estudo terá como resposta um aumento na taxa de aprovação, mostrando que as variáveis relacionadas aos questionamentos iniciais influenciam na aprendizagem dos estudantes do Ensino Médio público do Brasil.

**Palavras-chave:** Indicadores Educacionais, Análise de Regressão, *Software* R.

**Abstract:** From 2010 to 2019, according to data released by INEP, there was a reduction in the number of students that make up public High School classes in Brazil and, at the same time, there was a slight increase in the number of daily classes studied. It can also be noted that the pass rate, in the same period, went from 76% to 85%. Given this information, the following questions arise: does the number of students in the classroom influence learning? Or, will increasing classroom hours result in better learning? Or, even more emphatic, will reducing the number of students in the classroom and increasing the study hours will the result be satisfactory to the point of raising the pass rate? Such questions were analyzed and answered through regression analysis models using the R statistical software. The results obtained by such models state that reducing the number of students in the classroom and/or increasing the study hours will have as an answer an increase in the pass rate, showing that the variables related to the initial questions influence the learning of public High School students in Brazil.

**Key words:** Educational Indicators, Regression Analysis, R Software

# Lista de Figuras

# Sumário

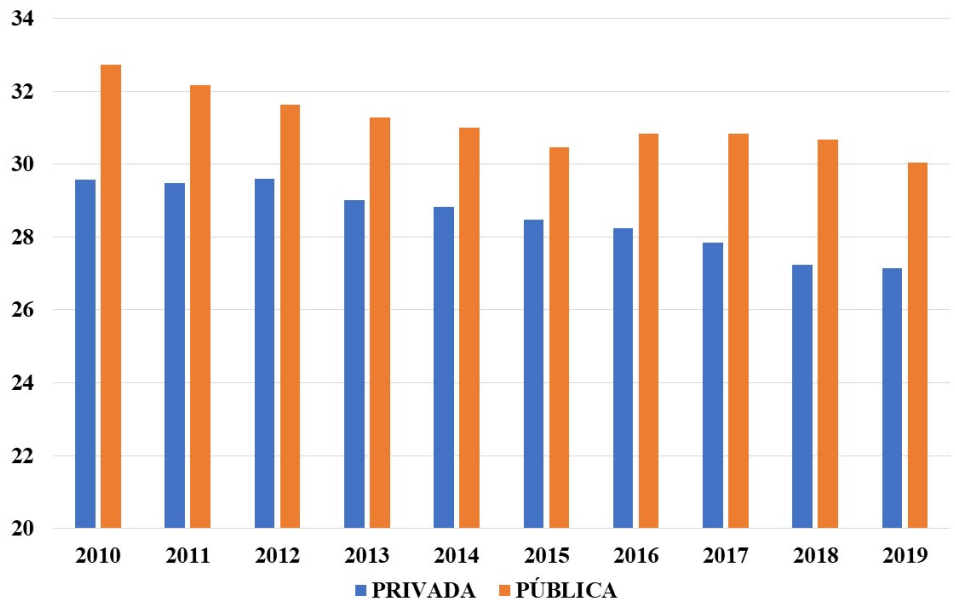
# 1 Introdução

O número de estudantes em sala de aula, a proporção de professores por grupos de alunos, o tamanho dos espaços físicos disponíveis são questões que afetam o desempenho e a aprendizagem, e estão ligadas ao aproveitamento do ensino. Caso a sala esteja superlotada, será muito mais difícil para os professores darem uma devida atenção a cada aluno individualmente. Apesar da socialização ser um fator importante para incentivar e aumentar o interesse pelo aprendizado, há um limite para que o número de estudantes por turma não passe a prejudicar o ensino. Mais do que a quantidade, é necessário prezar pela qualidade.

Segundo o artigo 6° da Constituição Federal de 1988, a Educação é estabelecida como um direito social e que o Ensino deve possuir um padrão de qualidade (BRASIL, 1988). Tais direitos também estão presentes na Lei de Diretrizes e Bases da Educação Nacional de 1996, em seu artigo 25 diz que as autoridades responsáveis devem estabelecer uma relação adequada entre o número de alunos, professor e carga horária (BRASIL, 1996).

Em 2012 a comissão de Educação, Cultura e Esporte do Senado aprovou o projeto de lei n. 504, de 2011 alterando o parágrafo único do art. 25 da Lei n. 9.394, de 20 de dezembro de 1996, para estabelecer o número máximo de alunos por turma na pré-escola e nos ensinos fundamental e médio (BRASIL, 2012).

Figura 1: Média de alunos por turma no Ensino Médio do Brasil

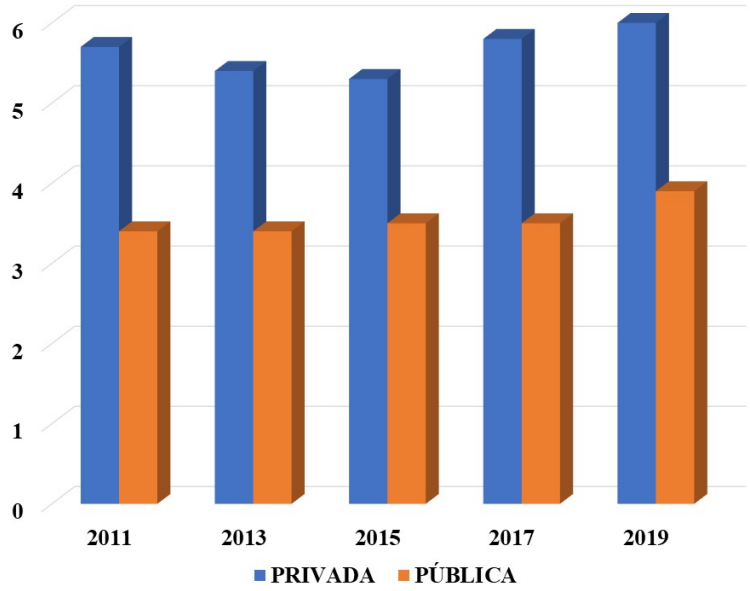



**Fonte:** Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Mas a realidade nas escolas públicas é outra. De acordo com Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP de 2010 a 2019, como mostra a Figura 1, apesar de ter ocorrido uma redução, o ensino médio público do país neste período sempre teve mais alunos por turma do que o ensino particular (INEP, 2010 - 2019).

Também com base nos dados do INEP sobre o Índice de Desenvolvimento da Educação Básica – IDEB, utilizado para medir a qualidade do aprendizado nacional e deles estabelecer metas para a melhoria do ensino, pela Figura 2 pode-se observar que nesse mesmo período o ensino particular obteve um melhor desempenho em relação ao público.

Figura 2: IDEB do Ensino Médio




**Fonte:** Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

A avaliação do rendimento escolar possui dois objetivos principais: identificar as dificuldades de aprendizagem dos alunos para ajudá-los a superá-las; e avaliar a eficácia do Ensino, sendo considerado um parâmetro de análise para o trabalho desenvolvido em sala de aula e na escola, refletindo assim a qualidade do Ensino (HAYDT, 1997).

Assim, com base nas Figuras 1 e 2 é possível destacar que o ensino público em relação ao ensino privado tende a desejar, pois em turmas reduzidas, para o professor, é mais fácil acompanhar o ritmo de aprendizagem de cada estudante, já turmas acima do limite tendem a dispersão. Assim os professores acabam perdendo tempo desnecessário em tentar manter o controle e a organização quando na verdade deveriam estar se dedicando ao ensino.

Por isso essa pesquisa pretende verificar se a quantidade de alunos por turma do Ensino Médio público do Brasil se reflete na taxa de aprovação, ou seja, quer-se responder a seguinte pergunta: Caso a quantidade de estudantes por turma do ensino médio público diminua, isso pode refletir em um aumento no desenvolvimento de seus conhecimentos, resultando em um aumento do número médio da taxa de aprovação nacional? Uma vez que com menos estudantes em sala o professor passa a ter mais tempo ao atendimento individual ao educando podendo assim, sanar melhor as possíveis dúvidas.

Nesse mesmo sentido, se aumentar a quantidade de horas estudadas em sala isso também resultará em um aumento na taxa de aprovação nacional do Ensino Médio Público do Brasil? Ou ainda, com este mesmo público, se aumentar a quantidade de horas diárias estudadas e ao mesmo tempo diminuir a quantidade de alunos por turma isso resultará em um aumento na taxa de aprovação nacional?

# 2 Referencial teórico

O desenvolvimento desse trabalho se dará utilizando cálculos estatísticos com o *software* R, tais detalhes se darão nas próximas seções.

## 2.1 O *software* R

Em 1996 os professores do departamento de estatística da Universidade de Auckland, Nova Zelândia, Ross Ihaka e Robert Gentleman desenvolveram o *software* R. O R é um *software* livre e colaborativo com uma linguagem de programação, por isso está em constante atualização (R CORE TEAM, 2020).

A utilização deste *software* se concentra na área de estatística por conta de sua capacidade gráfica e análise de dados, mas também é utilizado por desenvolvedores de *software* de estatística. Por ter inúmeros pesquisadores utilizando o *software* R, este está em constante atualização. Em sua versão padrão possui um número limitado de pacotes que são bibliotecas para funções específicas, mas também é possível adicionar pacotes desenvolvidos por outros pesquisadores.

> Em geral, as licenças para softwares estatísticos comerciais apresentam um custo elevado, o que gera grande dificuldade de manipulação e aprendizagem aos alunos que necessitam desta ferramenta para produzir suas análises. Outro aspecto importante de se ressaltar é que os softwares comerciais possuem seu código fechado, não permitindo ao usuário qualquer intervenção, ficando o mesmo sujeito à tarefa de operacionalizar os menus (ALCOFORADO; LEVY, 2017).

O R está disponível para download em seu site https://r-project.org/ sendo compatível com diferentes sistemas operacionais (Windows, MAC ou Linux) e sua instalação é bem simples.

## 2.2 Modelo de regressão

A teoria de Regressão teve origem no século XIX com Francis Galton. Em um de seus trabalhos ele estudou a relação entre a altura dos pais e dos filhos, procurando saber como a altura do pai influenciava a altura do filho. Galton notou que pais com baixa estatura tendem a ter filhos também com baixa estatura, porém os filhos têm altura média maior do que a altura média de seus pais. O mesmo acontecendo em sentido contrário, com pais de estatura alta. Essa observação Galton chamou de regressão, ou seja, existe uma tendência de os dados regredirem à média (DEMÉTRIO; ZOCCHI, 2006).

Quando são analisados dados que sugerem a existência de uma relação funcional entre duas variáveis, surge então o problema de se determinar uma função matemática que exprima esse relacionamento. A análise de regressão linear é uma relação entre a variável dependente (Y) e uma ou várias variáveis independentes $(X_1, X_2, ..., X_p)$. A regressão linear é responsável por determinar a equação que melhor representa a dispersão gráfica entre a variável dependente e a(s) variável(is) independente(s).

Segundo Bussab e Morettin (2010), o modelo de regressão linear pode ser escrito como:

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, ..., n \tag{1}$$

Ainda segundo os autores esses modelos podem ser classificados em três diferentes formas que são: Modelos lineares, modelos linearizáveis e modelos não lineares.

i. Modelos lineares: são aqueles que se referem ao modo como os parâmetros entram no modelo de forma linear. Seja (2) uma equação que representa o modelo linear.

$$y = \alpha + \beta x \tag{2}$$

ii. Modelos linearizáveis: são aqueles que não são lineares em sua forma inicial, mas podem ser transformados em lineares, por meio de transformações das variáveis.

$$y = \alpha.\beta^x \tag{3}$$

Assim, aplicando logaritmo em ambos os lados de (3), tem-se:

$$\begin{aligned} \log(y) &= \log(\alpha.\beta^x) \\ \log(y) &= \log(\alpha) + \log(\beta^x) \\ \log(y) &= \log(\alpha) + x.\log(\beta) \end{aligned} \tag{4}$$

Fazendo $y_1 = \log(y), \alpha_1 = \log(\alpha)$ e $\beta_1 = \log(\beta)$ em (4) resulta em:

$$y_1 = \alpha_1 + x.\beta_1 \tag{5}$$

Que é um modelo linear.

iii. Modelos não lineares: são aqueles não lineares em relação aos parâmetros e não possuem transformação conhecida capaz de torná-los lineares. Como exemplo o

Modelo de Gompertz (THOMAS, 2016):

$$y = e^{\alpha+\beta.\rho.x} + \epsilon \tag{6}$$

Em que $\alpha, \beta \in \mathbb{R}$ e $0 < \rho < 1$ e a função de Gompertz $e^{\alpha+\beta.\rho.x}$ é monotonicamente crescente.

## 2.3 Regressão Linear Simples

Uma regressão linear simples é definida como uma relação entre a variável dependente $(Y)$ e uma variável independente $(X)$ e é representada pela equação:

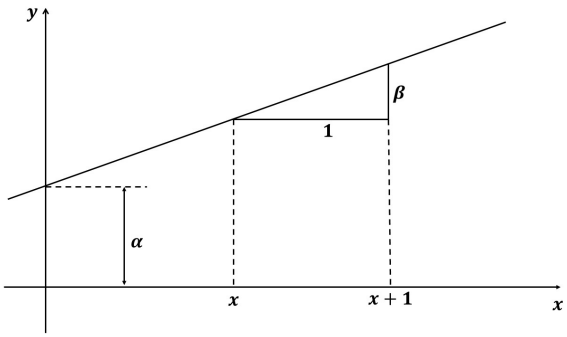
$$y_i = \alpha + \beta.x_i + \epsilon_i, \quad i = 1, 2, ..., n \tag{7}$$

Onde:

- $y_i$ representa o valor da variável dependente, $Y$, na observação $i = 1, 2, ..., n$;
- $x_i$ representa o valor da variável independente, $X$, na observação $i = 1, 2, ..., n$;
- $\alpha$ é o intercepto e representa o ponto onde a reta corta o eixo das ordenadas;
- $\beta$ é o coeficiente angular que representa o quanto varia a média de $Y$ para um aumento de uma unidade da variável $X$;
- $\epsilon_i, i = 1, 2, ..., n$ são variáveis aleatórias que correspondem ao erro.

A Figura 3 representa a interpretação geométrica dos parâmetros $\alpha$ e $\beta$.

Figura 3: Interpretação geométrica dos parâmetros $\alpha$ e $\beta$




**Fonte:** Adaptado de Bussab e Morettin (2010, p.450).

## 2.4 Pressupostos de uma Regressão Linear Simples

Segundo Lewis-Beck (1980), os seguintes pressupostos precisam ser satisfeitos para que uma regressão linear seja adequada:

i. a relação entre as variáveis dependente e independente deve ser linear;

ii. as variáveis foram medidas adequadamente, ou seja, assume-se que não há erro sistemático de mensuração;

iii. a expectativa da média do termo de erro é igual a zero;
Em (7) foi definido o modelo de regressão linear simples, pressupondo então que $E(\epsilon_i) = 0$ para $i = 1, ..., n$, tem-se:

$$\begin{aligned} E(y_i) &= E(\alpha + \beta x_i + \epsilon_i) \\ E(y_i) &= \alpha + \beta x_i + E(\epsilon_i) \\ E(y_i) &= \alpha + \beta x_i \end{aligned} \tag{8}$$

iv. homocedasticidade, ou seja, a variância do termo de erro é constante para os diferentes valores da variável independente. Lembrando que $E(\epsilon_i) = 0$, para $i = 1, ..., n$ e também que $E(\epsilon_i^2) = \sigma^2$ Assim:

$$\begin{aligned} var(\epsilon_i) &= E(\epsilon_i^2) - [E(\epsilon_i)^2] \\ var(\epsilon_i) &= \sigma^2 - 0^2 \\ var(\epsilon_i) &= \sigma^2 \end{aligned} \tag{9}$$

E consequentemente,

$$\begin{aligned} var(y_i) &= var(\alpha + \beta x_i + \epsilon_i) \\ var(y_i) &= var(\alpha + \beta x_i) + var(\epsilon_i) \end{aligned}$$

Mas $\alpha + \beta x_i$ é um termo constante, logo $var(\alpha + \beta x_i) = 0$, como $var(\epsilon_i) = \sigma^2$ então:

$$var(y_i) = \sigma^2$$

v. ausência de autocorrelação, ou seja, os termos de erros são independentes entre si,

o que significa:

$$\begin{aligned} cov(\epsilon_i, \epsilon_j) &= E(\epsilon_i \epsilon_j) - E(\epsilon_i)E(\epsilon_j) \\ cov(\epsilon_i, \epsilon_j) &= E(\epsilon_i \epsilon_j) \\ cov(\epsilon_i, \epsilon_j) &= 0 \end{aligned}$$

Para $i \neq j$, para $i, j = 1, 2, ..., n$.

vi. a variável independente não deve ser correlacionada com o termo de erro;

vii. nenhuma variável teoricamente relevante para explicar Y foi deixada de fora do modelo e nenhuma variável irrelevante para explicar Y foi incluída no modelo;

viii. assume-se que o termo de erro tem uma distribuição normal;
De iii. e v. tem-se:

$$\epsilon_i \sim N(0, \sigma^2), i = 1, 2, ..., n \tag{10}$$

e, portanto,

$$y_i \sim N(\alpha + \beta x_i, \sigma^2), i = 1, 2, ..., n \tag{11}$$

ix. há uma adequada proporção entre o número de casos e o número de parâmetros estimados.

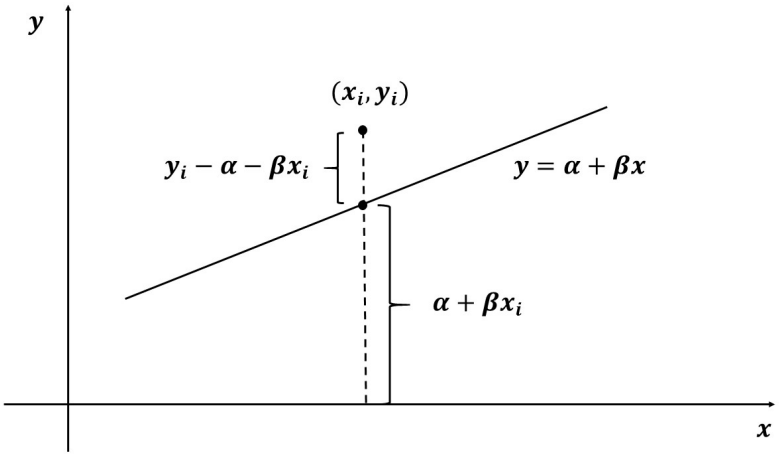
## 2.5 Estimação dos parâmetros do modelo (Método dos Mínimos Quadrados)

Supondo existir uma relação linear entre as variáveis $X$ e $Y$, então resta determinar os parâmetros $\alpha$ e $\beta$. Uma maneira de determinar esses parâmetros foi proposta por Gauss (1795) chamada de Método dos Mínimos Quadrados.

O objetivo desse método é minimizar a distância de cada ponto $(x_i, y_i)$ do gráfico a cada ponto $(x_i, \alpha + \beta x_i)$ da reta. A distância entre esses dois pontos é $|y_i - \alpha - \beta x_i|$ e a soma dos quadrados dessas distâncias é:

$$q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{12}$$

Figura 4: Distância de um ponto $(x_i, y_i)$ à reta $y = \alpha + \beta x$




**Fonte:** Souza (2003, p.1)

Os candidatos a ponto de mínimo da função (12) são aqueles para os quais são nulas as derivadas parciais de $q$ em relação a cada um de seus parâmetros, isto é:

$$\frac{\partial q}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \tag{13}$$

$$\frac{\partial q}{\partial \beta} = -2\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = 0 \tag{14}$$

Tendo em vista que:

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\alpha - \sum_{i=1}^{n}\beta x_i = \sum_{i=1}^{n} y_i - n\alpha - \beta\sum_{i=1}^{n} x_i$$

E que:

$$\sum_{i=1}^{n} x_i(y_i - \alpha - \beta x_i) = \sum_{i=1}^{n} x_i y_i - \alpha\left(\sum_{i=1}^{n} x_i\right) - \beta\left(\sum_{i=1}^{n} x_i^2\right)$$

Resulta no seguinte sistema de equações, denominadas "equações normais" do problema, cujas incógnitas são os parâmetros $\alpha$ e $\beta$ da equação $y = \alpha + \beta x$:

$$\begin{cases} n\alpha + \beta\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} y_i \\ \alpha\left(\sum_{i=1}^{n} x_i\right) + \beta\left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{15}$$

Sabendo que $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, representam as médias de $X$ e $Y$, respectivamente. Assim, isolando $\alpha$ na primeira equação do sistema (15), obtém-se:

$$\begin{aligned} n\alpha = \sum_{i=1}^{n} y_i - \beta\sum_{i=1}^{n} x_i \quad &\Rightarrow \quad \alpha = \frac{\sum_{i=1}^{n} y_i}{n} - \beta\frac{\sum_{i=1}^{n} x_i}{n} \\ \alpha \quad &= \quad \overline{y} - \beta\overline{x} \end{aligned} \tag{16}$$

Agora substituindo (16) na segunda equação de (15).

$$\alpha\left(\sum_{i=1}^{n} x_i\right) + \beta\left(\sum_{i=1}^{n} x_i^2\right) = \sum_{i=1}^{n} x_i y_i \quad \Rightarrow \quad (\overline{y} - \beta\overline{x})\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \Rightarrow$$

$$\beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - (\overline{y} - \beta\overline{x})\sum_{i=1}^{n} x_i \quad \Rightarrow \quad \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i + \beta\overline{x}\sum_{i=1}^{n} x_i \Rightarrow$$

$$\beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \overline{y}n\overline{x} + \beta\overline{x}n\overline{x} \quad \Rightarrow \quad \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \overline{y}n\overline{x} + \beta n\overline{x}^2 \Rightarrow$$

$$\beta\sum_{i=1}^{n} x_i^2 - \beta n\overline{x}^2 = \sum_{i=1}^{n} x_i y_i - \overline{y}n\overline{x} \quad \Rightarrow \quad \beta\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \sum_{i=1}^{n} x_i y_i - \overline{y}n\overline{x} \Rightarrow$$

$$\beta \quad = \quad \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}n\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{17}$$

Portanto, os valores de $\alpha$ e $\beta$ de (15) são:

$$\alpha = \overline{y} - \beta\overline{x} \text{ e } \beta = \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}n\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}$$

## 2.6 Regressão Linear Múltipla

A diferença entre a regressão linear múltipla e a regressão linear simples é que na múltipla são consideradas duas ou mais variáveis independentes enquanto que na regressão linear simples é considerada apenas uma. Assim, tem-se o seguinte modelo teórico:

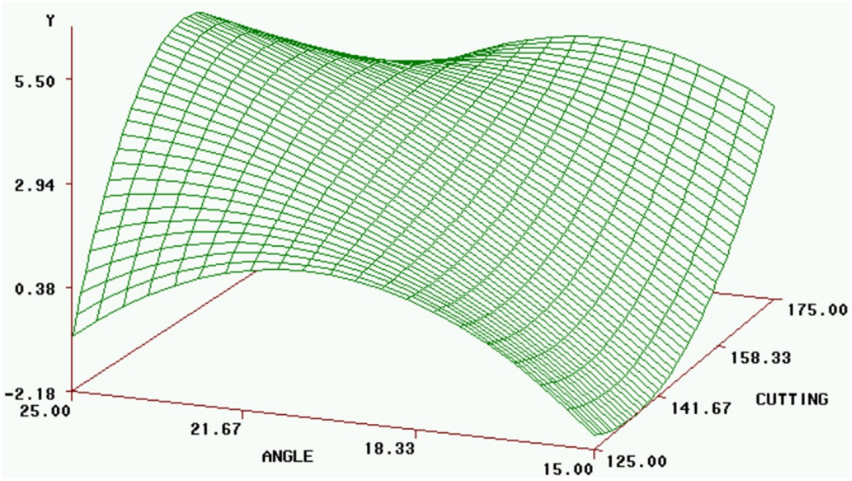
$$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + ... + \pi x_{ip} + \epsilon_i, i = 1, ..., n \tag{18}$$

Onde:

- $y_i$ representa o valor da variável dependente na observação $i, i = 1, 2, ..., n$;
- $x_{i1}, x_{i2}, ..., x_{ip}, i = 1, 2, ..., n$ são os valores da $i$-ésima observação das $p$ variáveis independentes;
- $\alpha, \beta, \gamma, ..., \pi$ são os parâmetros ou coeficientes de regressão;
- $\epsilon_i, i = 1, 2, ..., n$ correspondem aos erros aleatórios.

A equação (18) descreve um hiperplano $p$-dimensional referente às variáveis explicativas como mostra a Figura 5.

Figura 5: Hiperplano $p$-dimensional referente às variáveis independentes




**Fonte:** Rodriguês (2012, p.24)

## 2.7 Pressupostos de uma Regressão Linear Múltipla

Os pressupostos de uma regressão linear múltipla são:

i. a relação entre as variáveis dependente e independentes deve ser linear;

ii. as variáveis foram medidas adequadamente, ou seja, assume-se que não há erro sistemático de mensuração;

iii. a expectativa da média do termo de erro é igual a zero;

iv. homoscedasticidade, ou seja, a variância do termo de erro é constante para os diferentes valores da variável independente;

v. ausência de autocorrelação, ou seja, os termos de erros são independentes entre si;

vi. a variável independente não deve ser correlacionada com o termo de erro;

vii. nenhuma variável teoricamente relevante para explicar Y foi deixada de fora do modelo e nenhuma variável irrelevante para explicar Y foi incluída no modelo;

viii. assume-se que o termo de erro tem uma distribuição normal.
Logo $e_i \sim N(0, \sigma^2), i = 1, ..., n$ assim $y$ tem uma distribuição normal com variância $\sigma^2$ e para o caso da equação (18) tem-se:

$$E(Y) = \alpha + \beta X_1 + \gamma X_2 + ... + \pi X_p$$

ix. há uma adequada proporção entre o número de casos e o número de parâmetros estimados;

x. as variáveis independentes não apresentam alta correlação, o chamado pressuposto de não multicolinearidade.

## 2.8 Transformação de dados

Segundo Allaman (2019), a transformação de dados é uma forma possível de contornar o problema de pelo menos um dos pressupostos da regressão linear não ser atendido, assim uma transformação com os dados originais se faz necessário. É importante ressaltar que a transformação não garante que todos os pressupostos sejam atendidos e, portanto, deve-se fazer uma nova análise dos resíduos para checagem. A transformação só serve para realizar os testes de hipóteses, desse modo, a apresentação dos resultados deve ser feita com a variável original.

Caso o modelo de regressão testado falhe em algum dos pressupostos é possível aplicar uma transformação nas variáveis do modelo. Dentre as diversas transformações existentes, aqui será apresentada a transformação logarítmica e a de Box-Cox, ambas exemplificadas por Allaman (2019).

### 2.8.1 Transformação logarítmica

Geralmente quando os dados são contínuos e mesmo assim os resíduos não apresentam uma distribuição normal, então a transformação logarítmica pode ser útil para contornar tal violação dos pressupostos, não importando neste caso, se a base é 10 ou $e$ (logaritmo neperiano). Portanto, seja X uma variável aleatória mensurada na unidade de observação, tem-se a seguinte transformação:

$$Y_i = log(X_i) \text{ ou } Y_i = ln(X_i)$$

Caso alguns dos elementos for igual a zero, a transformação pode ser realizada da seguinte forma:

$$Y_i = log(X_i + k) \text{ ou } Y_i = ln(X_i + k)$$

Onde:

- $X_i$ são os elementos da variável original do modelo, com $i = 1, 2, ..., n$;
- $Y_i$ são os elementos da variável transformada, com $i = 1, 2, ..., n$;
- $k$ uma constante maior que zero.

### 2.8.2 Transformação de Box-Cox

Esse método de transformação das variáveis foi proposto em 1964 pelos estatísticos britânicos George Edward Pelham Box e Sir David Roxbee Cox, ficando esse processo conhecido como transformação Box-Cox (BOX; COX, 1964).

Tomando $X_1, X_2, ..., X_n$ como os dados originais do modelo, sobre eles é aplicada a transformação Box-Cox que permite estabelecer o parâmetro $\lambda$ tal que:

$$Y_i(\lambda) = \begin{cases} \dfrac{X_i^{\lambda} - 1}{\lambda}, & \text{se } \lambda \neq 0 \\ log(X_i), & \text{se } \lambda = 0 \end{cases} \tag{19}$$

Estabelecido o valor de $\lambda$ é encontrado os valores dos dados transformados $Y_1, Y_2, ..., Y_n$.

### 2.8.3 Transformação utilizada neste trabalho

Nas seções 4.2 e 4.3 poderá ser observado que os modelos iniciais a serem testados falham em alguns de seus pressupostos, assim será realizada a transformação sobre uma das variáveis do modelo, no caso a variável em questão é "*alunos por turma*". Chegou-se a confirmação dessa variável pois essa se correlaciona com o termo de erro do modelo em 4.2 e possui alta correlação com a variável "*horas estudadas*" em 4.3.

Inicialmente, formam realizadas as transformações elencadas no item 2.8, mas os novos modelos de regressão que continham a variável transformada continuavam falhando no mesmo pressuposto. Até que foi testada a seguinte transformação:

$$Y_i = \frac{\sqrt{X_i}}{e^{X_i}}$$

Onde:

- $X_i$ a variável original, e
- $Y_i$ é a variável transformada.

Neste caso, todos os pressupostos do modelo de regressão linear utilizando a variável $Y_i$ foram atendidos. Mas, foi pensado em simplificar essa transformação. Assim foi realizada uma nova transformação:

$$Y_i = \frac{X_i}{e^{X_i}}$$

Que novamente verificou-se que todos os pressupostos são atendidos, e por fim chegou-se a transformação:

$$Y_i = \frac{1}{e^{X_i}} \tag{20}$$

Esta, mais simples que as anteriores e contemplando todos os pressupostos de modelo de regressão linear utilizando a variável transformada.

Onde, $X_i$ e $Y_i$ são respectivamente, os elementos das variáveis *"alunos por turma"* e *"AlunosT"*. Esta última sendo a variável transformada, ficando seus elementos, como mostra a Tabela 1

Tabela 1: Variáveis *"Média de alunos por turma"* e *"AlunosT"*

| **Ano** | **Média de alunos por turma** | **AlunosT** |
|---|---|---|
| **2010** | 32,87 | $5{,}305610\times10^{-15}$ |
| **2011** | 32,23 | $1{,}300621\times10^{-14}$ |
| **2012** | 31,70 | $1{,}709480\times10^{-14}$ |
| **2013** | 31,33 | $2{,}748800\times10^{-14}$ |
| **2014** | 31,07 | $3{,}209740\times10^{-14}$ |
| **2015** | 30,57 | $5{,}291970\times10^{-14}$ |
| **2016** | 30,93 | $3{,}692080\times10^{-14}$ |
| **2017** | 30,90 | $3{,}804530\times10^{-14}$ |
| **2018** | 30,73 | $4{,}509520\times10^{-14}$ |
| **2019** | 30,03 | $9{,}081060\times10^{-14}$ |

# 3 Material e Métodos

A seguir, será apresentada a metodologia proposta nessa dissertação. Primeiramente serão apresentados os dados utilizados na pesquisa, após, os métodos empregados para desenvolver o modelo de regressão linear.

## 3.1 Material

Os dados utilizados nessa dissertação foram obtidos no site do (INEP, 2010 - 2019). Os indicadores analisados são: média de alunos por turma, média de horas-aula diária e taxa de rendimento. Vale ressaltar que os indicadores de média de alunos por turma e taxa de rendimento são divulgados desde 2007, mas o indicador média de horas-aula diária começou a ser divulgado apenas em 2010. Assim, para manter o mesmo parâmetro, as análises realizadas se compreendem de 2010 a 2019. Uma outra observação é que alguns desses indicadores já possuem informações sobre o ano de 2020, mas como este foi um ano atípico por conta da pandemia de Covid-19, então decidiu-se permanecer com o primeiro recorte temporal.

Todos os indicadores educacionais utilizados possuem informações referentes ao Brasil, das cinco regiões brasileiras, dos estados da federação, municípios e também os dados individuais de cada escola. Esses indicadores são divulgados em planilhas disponíveis para download no site do INEP, sendo possível verificar a dependência administrativa da escola (privada, municipal, estadual ou federal) e também se sua localização é urbana ou rural. As informações estão distribuídas entre Ensino Fundamental (inicial e final) e Ensino Médio.

Os dados utilizados nessa dissertação são referentes aos indicadores citados acima de 2010 a 2019 do Ensino Médio. Assim, foi necessária uma organização das informações as quais serão apresentadas em forma de tabela na seção 4 deste trabalho.

## 3.2 Métodos

Nesta seção estão apresentados os pressupostos para desenvolver uma regressão linear. Lembrando que neste trabalho é utilizado o *Software* R versão 4.0.3.

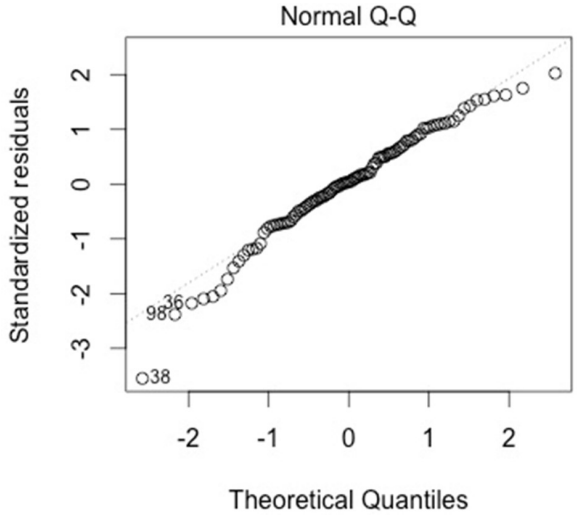
### 3.2.1 Normalidade dos resíduos

A normalidade dos resíduos pode ser analisada através de gráficos ou por testes de hipóteses. Aqui será analisado o gráfico QQ mais conhecido como *QQ plot*, que é

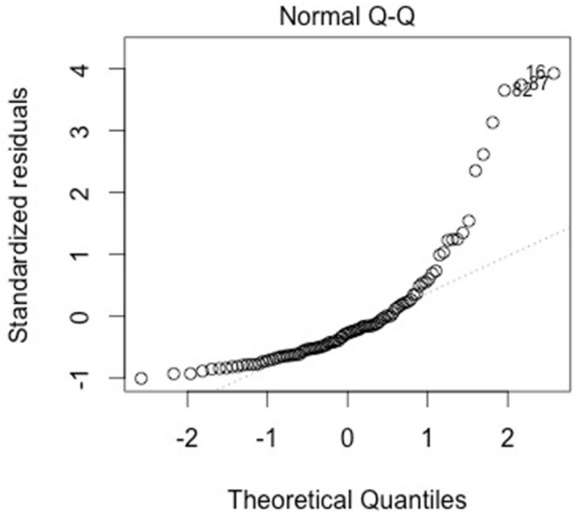
um gráfico de dispersão utilizado para avaliar se um conjunto de dados possui alguma distribuição teórica, como normal, polinomial, exponencial, logarítmica dentre outras.

Figura 6: *QQ plot*

(a) Resíduos apresentam distribuição normal




(b) Resíduos não apresentam distribuição normal




**Fonte:** https://data.library.virginia.edu/diagnostic-plots/

Para que os resíduos sejam considerados normais, eles têm de estar linearmente distribuídos sobre a linha tracejada da Figura 6. Os pontos na Figura 6a não estão perfeitamente distribuídos, mas estão aproximados da linha tracejada, então pode-se considerar que os dados da Figura 6a seguem uma distribuição normal. Já na Figura 6b os pontos estão distribuídos em formato exponencial, logo os resíduos de gráficos com esse formato não são considerados com distribuição normal.

Outra forma de avaliar a normalidade dos resíduos quando sua amostra é menor que 30 elementos foi proposto por Shapiro e Wilk (1965). Essa técnica é conhecida hoje como teste de Shapiro-Wilk, representado pela letra $W$. As hipóteses testadas são:

$$\begin{cases} H_0 : \text{os resíduos seguem uma distribuição normal} \\ H_1 : \text{os resíduos não seguem uma distribuição normal} \end{cases}$$

A estatística do teste Shapiro-Wilk é:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Onde $x_i$ são os valores das amostras ordenadas, ou seja, $x_1 < x_2 < ... < x_n, \overline{x}$ é a média dos valores da amostra e $b$ é determinada da seguinte forma:

$$b = \begin{cases} \sum_{i=1}^{\frac{n}{2}} a_{n-i+1}(x_{n-i+1} - x_i), \text{ se n é par} \\ \\ \sum_{i=1}^{\frac{n+1}{2}} a_{n-i+1}(x_{n-i+1} - x_i), \text{ se n é ímpar} \end{cases} \tag{21}$$

Em que $a_{n-i+1}$ são constantes geradas pelas médias, variâncias e covariâncias de uma amostra de tamanho $n$.

Para realizar esse mesmo teste utilizando o *software* R basta inserir o comando `shapiro.test`. Esse teste já está disponível na biblioteca do *software* em sua versão padrão, não sendo necessário instalar pacote adicional. As informações obtidas são:

Figura 7: *Output* teste Shapiro-Wilk

```
        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.94724, p-value = 0.6359
```

Neste caso, são obtidos $W_{cal}$ que representa o valor calculado do teste de Shapiro-Wilk e *p-value* ou em português p-valor, mede o quanto os dados do modelo são compatíveis com a hipótese nula.

### 3.2.2 *Outliers* dos resíduos

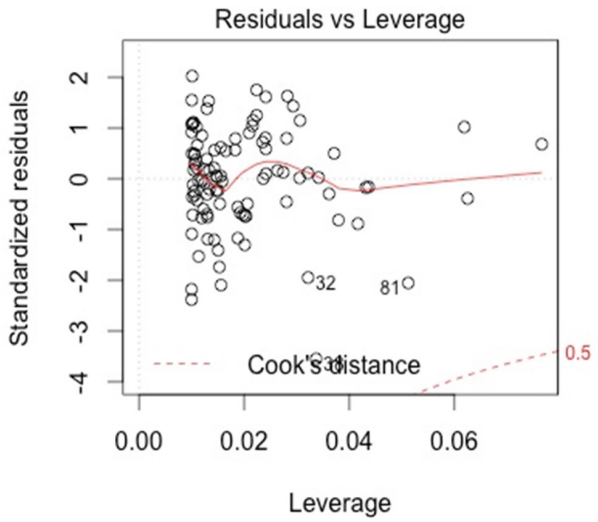
A análise gráfica para verificar a existência de *outlier*s em um modelo, consiste em observar pontos afastados da maioria dos demais. A existência desses pontos faz surgirem dúvidas se esses pontos afastados estão interferindo de algum modo no modelo testado.

O gráfico *Residuals vs Leverage* (em tradução livre resíduos vs pontos de alavancagem) é considerado o mais indicado para essa análise gráfica. Nele os *outliers* ou pontos de alavancagem estão além da linha tracejada, como mostra a Figura 8b. Neste caso, o R já indica com um rótulo o elemento do modelo que está mais afastado.

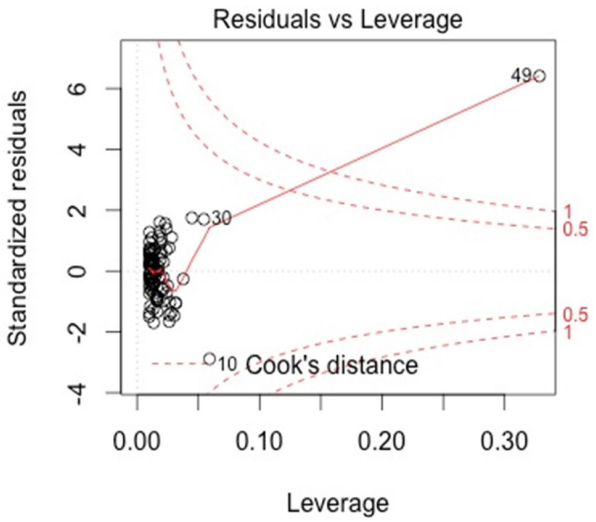
A Figura 8a não possui nenhum ponto além da linha tracejada, então pode-se considerar que os dados da Figura 8a não apresenta nenhum ponto de alavancagem. Ou seja, o modelo testado não possui *outliers*.

Figura 8: Resíduos em relação aos pontos de alavancagem

(a) Resíduos não possuem *outliers*




(b) Resíduos possuem *outliers*




**Fonte:** https://data.library.virginia.edu/diagnostic-plots/

Para realizar a análise da existência de *outliers* no R através de testes é utilizada a função `summary(rstandard(modelo))` do pacote `stats`. Essa função já está disponível na biblioteca do *software* em sua versão padrão. As informações obtidas são:

Figura 9: *Output* da função summary

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.46050 -0.69928 -0.08365  0.03896  0.89794  1.53472
```

Em sequência da esquerda para a direita na Figura 9 são apresentados o valor mínimo, primeiro quartil, mediana (ou segundo quartil), média, terceiro quartil e valor máximo. Aqui todos os valores estão padronizados, e caso haja algum ponto de alavancagem este terá um valor padronizado superior a 3, ou inferior a $-3$ e além disso a mediana tem de ficar próxima a zero.

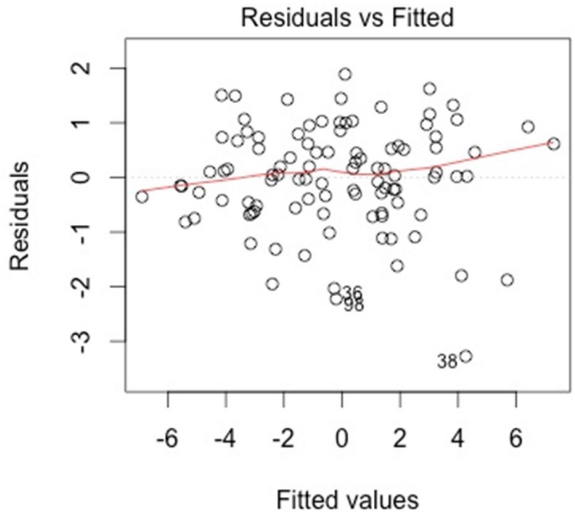
### 3.2.3 Independência dos resíduos

A análise gráfica para a independência dos resíduos pode ser realizada em dois diferentes gráficos.

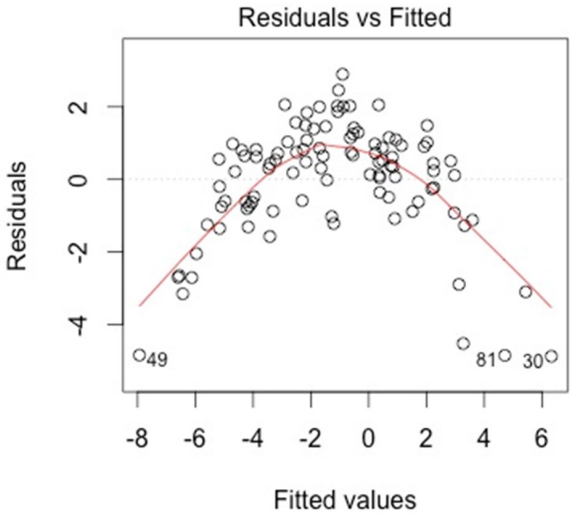
Para que haja independência dos resíduos a distribuição dos pontos deve acontecer com a mesma amplitude em torno do zero, ou seja, para que os resíduos sejam considerados independentes eles têm de estar igualmente espalhados em torno de uma linha horizontal, de preferência se essa linha for sobre o ponto zero do eixo dos resíduos.

Figura 10: Resíduos em relação aos valores ajustados

(a) Resíduos são independentes




(b) Resíduos não são independentes

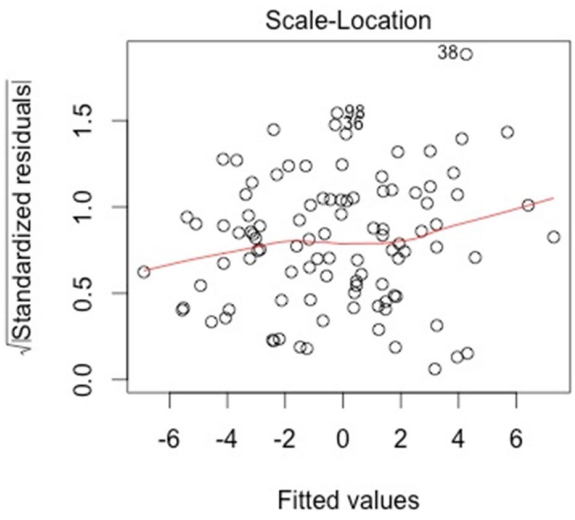



**Fonte:** https://data.library.virginia.edu/diagnostic-plots/

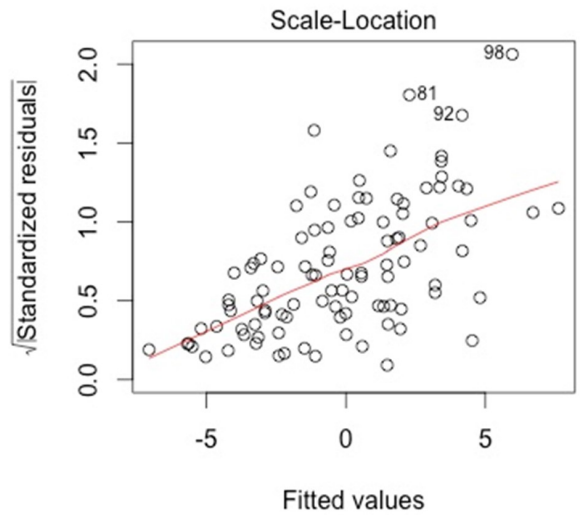
Assim, pode-se afirmar que os resíduos da Figura 10a são independentes, enquanto que os resíduos da Figura 10b não são independentes.

Figura 11: Raiz quadrada dos resíduos padronizados em relação aos valores ajustados

(a) Resíduos são independentes




(b) Resíduos não são independentes




**Fonte:** https://data.library.virginia.edu/diagnostic-plots/

Na Figura 11 tem-se a raiz quadrada dos resíduos padronizados em relação aos valores ajustados, e neste caso a análise a ser feita é praticamente a mesma que realizada nos gráficos da Figura 10. A Figura 11a apresenta uma linha próxima da horizontal com os resíduos distribuídos aleatoriamente. Já a Figura 11b possui uma linha quase que linear,

porém com uma inclinação e com os pontos apresentando uma maior dispersão depois de 0. Assim, a melhor figura que representa a independência dos resíduos é a Figura 11a.

Outra forma de analisar a independência dos resíduos foi proposta por Durbin e Watson (1950). Hoje essa técnica é conhecida como teste Durbin-Watson, representada por DW e as hipóteses testadas são:

$$\begin{cases} H_0 : \text{não existe autocorrelação entre os resíduos} \\ H_1 : \text{existe autocorrelação entre os resíduos} \end{cases}$$

A estatística do teste é:

$$DW = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2}$$

Onde $\epsilon_i$ é o termo do erro do modelo na $i$-ésima observação. Para tomar a decisão de aceitar ou rejeitar $H_0$ é necessário verificar o valor de DW e compará-lo aos valores críticos de $d_L$ e $d_U$ do Quadro 1 de Durbin-Watson. O Quadro 1 informa as decisões a serem tomadas em função dos valores críticos.

Quadro 1: Quadro de decisão em função de $d_L$ e $d_U$

| | Zona de rejeição ou aceitação de $H_0$ | | | | |
|---|---|---|---|---|---|
| DW | $[0, d_L)$ | $[d_L; d_U)$ | $[d_U; 4-d_U)$ | $[4-d_U; 4-d_L)$ | $[4-d_L; 4]$ |
| Decisão | Rejeitar $H_0$ Autocorrelação positiva | Nada se pode concluir | Não rejeitar $H_0$ Os resíduos são independentes | Nada se pode concluir | Rejeitar $H_0$ Autocorrelação negativa |

No R é possível realizar o teste de Durbin-Watson através do pacote "`car`" em sua versão atual, 3.0.11 (FOX; WEISBERG, 2019). Para utilizar o pacote "`car`" primeiro é necessário instalá-lo e após carregá-lo executando, respectivamente, as seguintes linhas de comando:

```
> install.packages("car")
> library("car")
```

Com o pacote "`car`" instalado e carregado agora é possível realizar a estatística de Durbin-Watson no R seguindo o comando:

```
> durbinWatsonTest(model, ...)
```

As informações obtidas desse teste são:

Figura 12: *Output* do teste de Durbin-Watson

```
lag Autocorrelation D-W Statistic p-value
  1      -0.5453145      3.016875   0.116
Alternative hypothesis: rho != 0
```

Onde, D-W Statistic é a estatística de Durbin-Watson, ou seja, o valor calculado da estatística. *P-value* é o p-valor da hipótese a ser testada.

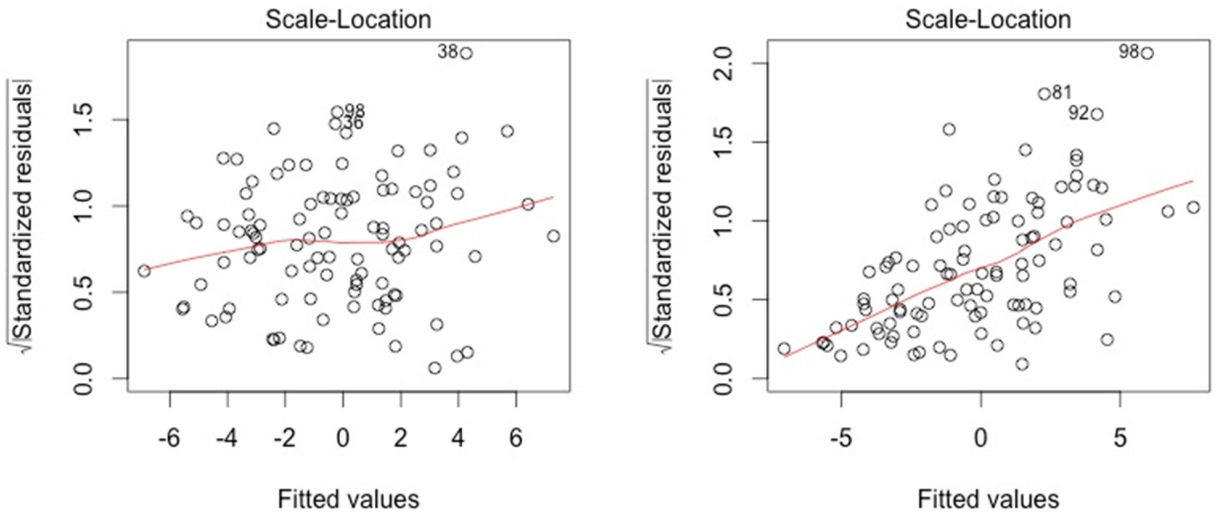
### 3.2.4 Homocedasticidade dos resíduos

Pode-se analisar a homocedasticidade dos resíduos através de gráficos ou por testes de hipóteses. Para esse pressuposto é necessário analisar o gráfico da raiz quadrada dos resíduos padronizados pelos valores previstos dos resíduos que é o gráfico *Scale-Location*. Este é um gráfico de dispersão utilizado para verificar a suposição da igualdade da variância entre os resíduos, caso essa igualdade exista, haverá uma linha no sentido horizontal.

A linha da Figura 13a não está perfeitamente horizontal, mas está aproximada. Já na Figura 13b a linha está com uma inclinação ferindo a condição inicial de homocedasticidade dos elementos. Assim, a Figura 13a é a que melhor representa a homocedasticidade entre os resíduos.

Figura 13: Raiz quadrada dos resíduos padronizados em relação aos valores ajustados

(a) Há homocedasticidade entre os resíduos (b) Não há homocedasticidade entre os resíduos




**Fonte:** https://data.library.virginia.edu/diagnostic-plots/

Dentre os testes existentes para avaliar a homocedasticidade dos resíduos o teste

utilizado neste trabalho foi desenvolvido por Breusch e Pagan (1979). As hipóteses testadas são:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 \\ H_1 : \text{pelo menos um dos } \sigma_i^2 \text{ é diferente, } i = 1, 2, ..., n \end{cases}$$

Para realizar o teste Breusch-Pagan utilizando o R é necessário o pacote "`lmtest`" desenvolvido por Zeileis e Hothorn (2002). Primeiro é necessário instalar o pacote e após carregar, executando respectivamente as seguintes linhas de comando:

```
> install.packages("lmtest")
> library("lmtest")
```

Com o pacote "`lmtest`" instalado e carregado agora é possível realizar a estatística de Breusch-Pagan no R seguindo o comando:

```
> bptest(formula, varformula = NULL, data = list())
```

Onde:

- formula: o modelo a ser testado;
- varformula: uma equação que descreve apenas as variáveis explicativas potenciais para a variância;
- data: são os dados do modelo.

As informações obtidas desse teste são:

Figura 14: *Output* do teste Breusch-Pagan

```
        studentized Breusch-Pagan test

data:  mod
BP = 0.34589, df = 1, p-value = 0.5565
```

Onde:

- BP é a estatística de Breusch-Pagan, ou seja, o valor calculado da estatística;
- df os graus de liberdade, pois esse teste segue a distribuição qui-quadrado;
- *P-value* é o p-valor da hipótese a ser testada.

# 4 Resultados e discussões

Os dados analisados aqui foram obtidos no site do INEP e organizados na Tabela 2. Vale ressaltar que todas as informações da Tabela 2 são referentes ao Ensino Médio público do Brasil.

Tabela 2: Ensino Médio público do Brasil

| **Ano** | **Horas estudadas** | **Taxa de aprovação em (%)** | **Taxa de reprovação em (%)** | **Taxa de abandono em (%)** | **Média de alunos por turma** |
|---|---|---|---|---|---|
| 2010 | 4,5 | 76,17 | 12,6 | 11,23 | 32,87 |
| 2011 | 4,5 | 76,17 | 13,33 | 10,5 | 32,23 |
| 2012 | 4,53 | 77,47 | 12,43 | 10,1 | 31,7 |
| 2013 | 4,67 | 79,07 | 12,03 | 8,9 | 31,33 |
| 2014 | 4,77 | 79,2 | 12,37 | 8,43 | 31,07 |
| 2015 | 4,8 | 80,7 | 11,77 | 7,53 | 30,57 |
| 2016 | 4,8 | 80,37 | 12,33 | 7,3 | 30,93 |
| 2017 | 4,9 | 82,07 | 11,2 | 6,73 | 30,9 |
| 2018 | 4,97 | 82,2 | 10,93 | 6,87 | 30,73 |
| 2019 | 5,07 | 85,27 | 9,43 | 5,3 | 30,03 |

**Fonte:** INEP

## 4.1 Regressão linear simples entre horas estudadas e taxa de aprovação

O modelo de regressão linear em questão pretende verificar se a quantidade de horas estudadas pelos estudantes do Ensino Médio público do Brasil se reflete na taxa de aprovação, ou seja, quer-se responder a seguinte pergunta: Caso os estudantes que cursam o ensino médio público do país permaneçam mais horas diárias em aula desenvolvendo suas habilidades, com o decorrer dos anos isso pode realmente refletir em um maior desenvolvimento de seu conhecimento resultando em um aumento do número médio da taxa de aprovação nacional?

Mas antes disso, é necessário verificar os pressupostos da regressão linear simples. Inicialmente é preciso carregar os dados a serem analisados no R, a versão utilizada aqui é a 4.0.3. Vale ressaltar que neste trabalho é utilizado o RStudio em sua versão 1.4.1717, este é uma interface mais intuitiva que a do R, mas em ambos os resultados são os mesmos.

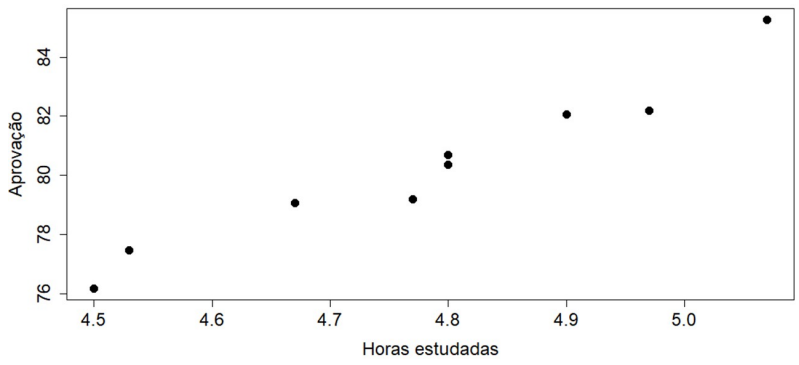
No RStudio os dados são carregados seguindo o caminho: *File* > *Import Dataset* > *From Excel....* Lembrando que os dados desta análise, que aparecem na Tabela 1, foram organizados numa planilha antes de serem importados. Após, as informações carregadas são nomeadas como *"dados"*.

Neste modelo, a variável dependente é taxa de aprovação e a variável independente é horas estudadas. Agora com as variáveis já conhecidas é preciso verificar se há alguma relação linear entre elas, no R a função "`plot`" é utilizada para a construção de gráficos. Assim:

```
> plot(dados$Horas,dados$Aprovação, xlab="Horas estudadas", ylab="Aprova-
ção", cex=1.5, cex.lab=1.5, cex.axis=1.5, pch=19)
```

Para essas variáveis tem-se o seguinte gráfico de dispersão:

Figura 15: Relação linear entre taxa de aprovação e horas estudadas



Observando a Figura 15 é possível notar que os pontos distribuídos não formam perfeitamente uma relação linear, mas se aproximam dela, não tendo nenhuma semelhança com o gráfico de uma relação quadrática ou exponencial. Mas realizando o teste de coeficiente de correlação linear de Pearson entre essas variáveis no R através da função:

```
> cor.test(dados$Horas,dados$Aprovação)
```

Obtém-se:

Figura 16: *Output* teste de coeficiente de correlação

```
        Pearson's product-moment correlation

data:  dados$Horas and dados$Aprovação
t = 13.782, df = 8, p-value = 7.415e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9131868 0.9953232
sample estimates:
      cor
0.9795848
```

Do teste de coeficiente de correlação linear de Pearson realizado no R a correlação entre as variáveis é de aproximadamente $0,98$. Como essa é uma correlação forte e além disso o p-valor desse teste é de $7,415 \times 10^{-7}$ indicando a rejeição da hipótese de que o coeficiente de correlação seja igual a zero. Assim pode-se afirmar que a distribuição gráfica dessas variáveis possui uma relação linear.

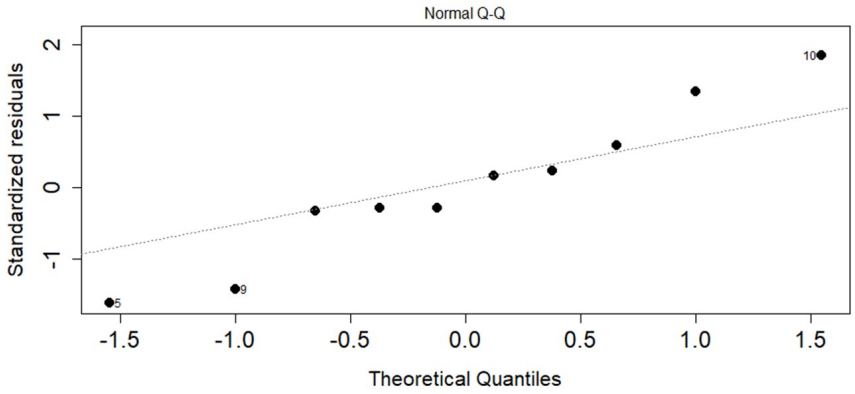
A seguir serão analisados os pressupostos do modelo de regressão para informar com mais precisão se realmente essas variáveis formam uma regressão linear. Mas, antes disso, é preciso construir tal modelo no R. O modelo a ser analisado será chamado de "`mod`", e para criá-lo é utilizada a função "`lm`" do inglês *linear model*. Assim:

```
> mod <- lm(Aprovação ~ Horas, dados)
```

Os gráficos utilizados para a análise dos pressupostos de normalidade, *outliers*, independência e homocedasticidade dos resíduos são obtidos pelo comando `plot(mod)` no R. O primeiro pressuposto a ser analisado nos resíduos é a sua normalidade. Para realizar a análise gráfica deste pressuposto é utilizado o gráfico *QQ plot*.

Na Figura 17 para que os resíduos apresentem uma distribuição normal, os pontos do gráfico têm de estar sobre a linha pontilhada. Neste caso, não estão perfeitamente alinhados, mas estão aproximados da linha fugindo um pouco em alguns deles.

Figura 17: Normalidade dos resíduos




Disto, é possível considerar que há uma distribuição normal dos resíduos, mas aplicando o teste de Shapiro-Wilk no modelo é possível ter uma afirmação mais precisa sobre tal.

Figura 18: *Output* teste Shapiro-Wilk

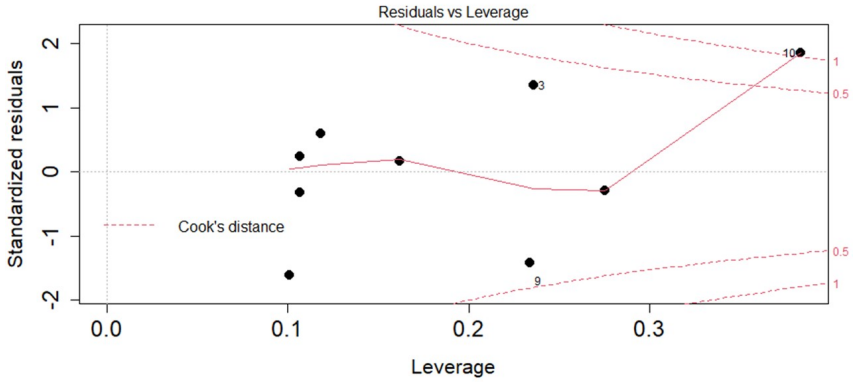
```
        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.95912, p-value = 0.7758
```

Ressaltando que o teste de Shapiro-Wilk tem as seguintes hipóteses:

$$\begin{cases} H_0 : \text{os resíduos do modelo seguem uma distribuição normal} \\ H_1 : \text{os resíduos do modelo não seguem uma distribuição normal} \end{cases}$$

Considerando o teste para 5% de significância e observando que o p-valor obtido nesse teste foi de $0,7758$ então, neste caso, a hipótese nula é considerada, podendo afirmar que os resíduos do modelo em questão seguem uma distribuição normal.

Figura 19: *Outliers* nos resíduos



O próximo pressuposto a ser observado é a existência de *outliers*. Realizando uma análise da Figura 19 é possível observar que existe um ponto além da linha vermelha tracejada no gráfico, mas utilizando a comando `summary(rstandard(mod))` é possível afirmar se o ponto em questão pode ser considerado como ponto de alavancagem.

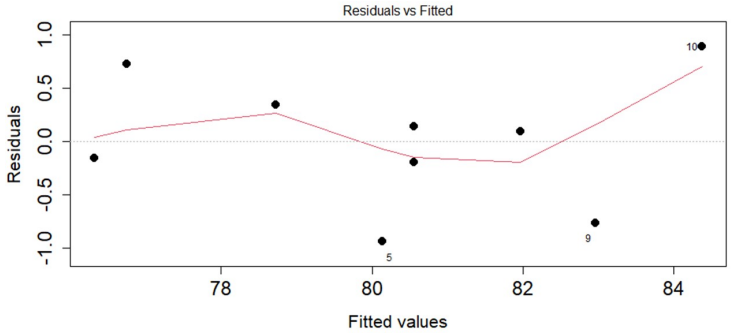
Da análise, como mostra a Figura 20, os valores padronizados estão entre $-3$ e 3, indicando que não há a existência de pontos de alavancagem, além da média estar bem próximo de zero. Assim, é possível afirmar que não existe outlier nos resíduos do modelo em questão.

Figura 20: *Output* função summary

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.61067 -0.32025 -0.06067  0.02673  0.50937  1.85389
```

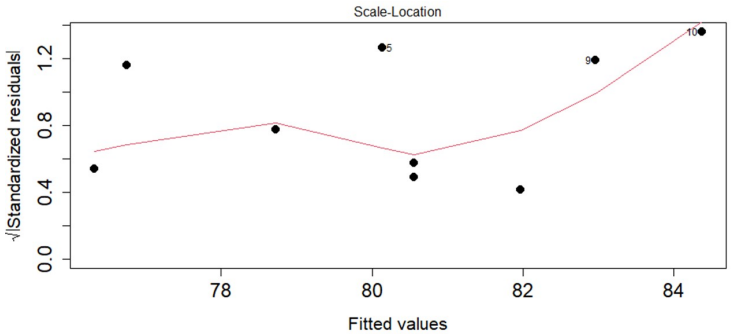
O terceiro pressuposto a ser analisado é a independência dos resíduos. Realizando a análise gráfica.

Figura 21: Independência dos resíduos




Na Figura 21, não é possível afirmar se a distribuição dos pontos em torno do zero acontece com a mesma amplitude. Já na Figura 22, para que exista a independência, a linha vermelha tem de ser aproximadamente horizontal, o que não acontece.

Figura 22: Independência dos resíduos



Desse modo, pela análise gráfica não é possível afirmar se os resíduos do modelo em questão são independentes. Disto, aplicando o teste de Durbin-Watson no modelo é possível afirmar se há a independência dos resíduos.

Figura 23: *Output* teste Durbin-Watson

```
lag Autocorrelation D-W Statistic p-value
  1      -0.3617799      2.451174    0.78
Alternative hypothesis: rho != 0
```

Ressaltando que o teste de Durbin-Watson tem as seguintes hipóteses:

$$\begin{cases} H_0 : \text{não existe autocorrelação nos resíduos do modelo} \\ H_1 : \text{existe autocorrelação nos resíduos do modelo} \end{cases}$$

Considerando um nível de significância 5% e observando que o p-valor obtido nesse teste foi de $0,78$ então a hipótese nula é aceita, assim pode-se afirmar que no modelo em questão os resíduos são independentes.

O último pressuposto a ser analisado é sobre a homocedasticidade dos resíduos. Pela análise gráfica, para que exista a homocedasticidade no modelo, no gráfico da Figura 21 teria de haver uma linha vermelha no sentido horizontal, o que não acontece. Mas aplicando o teste Breusch-Pagan no modelo tem-se:

Figura 24: *Output* teste Breusch-Pagan

```
        studentized Breusch-Pagan test

data:  mod
BP = 1.7428, df = 1, p-value = 0.1868
```

Ressaltando que as hipóteses do teste de Breusch-Pagan são:

$$\begin{cases} H_0 : \text{há homocedasticidade dos resíduos do modelo} \\ H_1 : \text{não há homocedasticidade dos resíduos do modelo} \end{cases}$$

Para um nível de 5% de significância e pelo fato do p-valor do teste de Breusch-Pagan ser de $0,1868$, então é considerada a hipótese nula, logo pode-se afirmar que há homoscedasticidade entre os resíduos das variáveis.

Os testes de hipóteses para os resíduos do modelo confirmam que todos os pressupostos são atendidos. Assim, resta realizar uma análise do modelo utilizando a função `summary(mod)`.

Figura 25: *Output* função summary

```
Call:
lm(formula = Aprovação ~ Horas, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-0.93748 -0.18160 -0.02833  0.29384  0.89331

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.734      4.875   2.612    0.031 *
Horas         14.131      1.025  13.782 7.42e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6139 on 8 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9545
F-statistic:   190 on 1 and 8 DF,  p-value: 7.415e-07
```

A Figura 25 fornece um resumo das informações do modelo que está sendo analisado. Onde:

- *"call"* é a descrição do modelo de regressão linear analisado;
- *"residuals"* são os valores dos resíduos que não estão padronizados;
- *"intercept"* é o ponto em que a reta corta o eixo $y$, neste caso é a taxa de aprovação (variável dependente) esperada quando horas estudadas (variável independente) é zero.

A variável independente *"horas estudadas"* tem um p-valor de $7,42 \times 10^{-7}$. O p-valor para a variável independente é baseado no teste $t$ de *Student* que possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{o coeficiente é igual a zero} \\ H_1 : \text{o coeficiente é diferente de zero} \end{cases}$$

Para um teste ao nível de 5% de significância e com um p-valor obtido de $7,42 \times 10^{-7}$, então rejeita-se a hipótese nula, portanto o coeficiente é diferente de zero podendo ser interpretado, mostrando que a variável independente tem um impacto sobre a variável dependente.

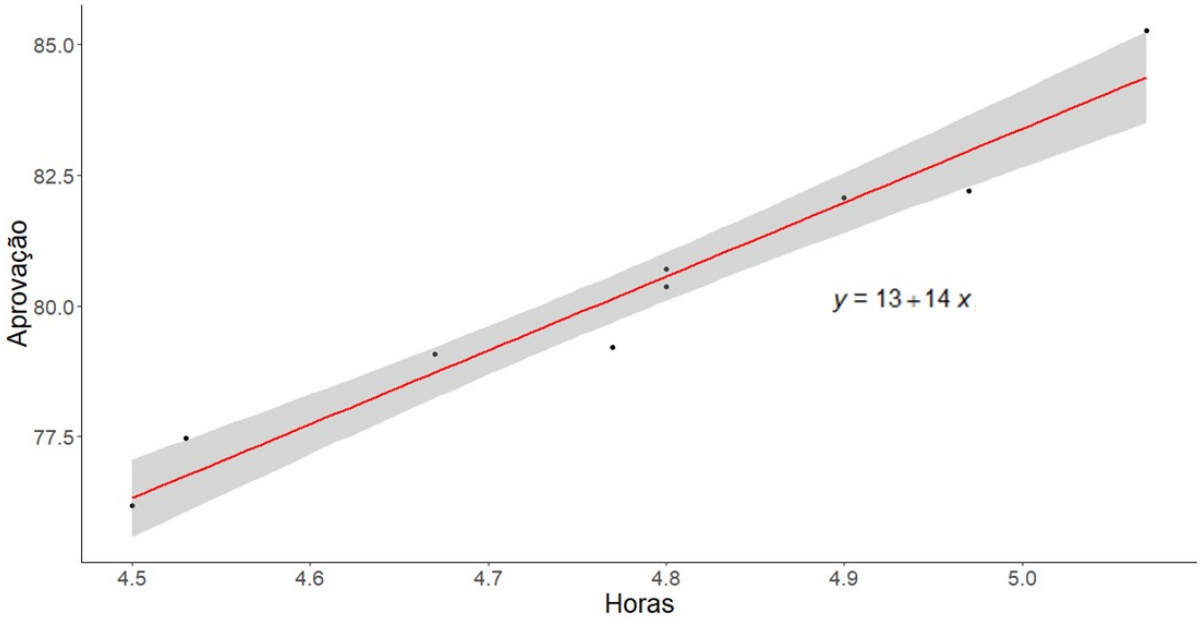
Então, interpretando o modelo, a cada hora estudada a mais em sala de aula a taxa média de aprovação dos estudantes do Ensino Médio público no Brasil aumenta 14,1. Outra análise que também pode ser realizada envolve o *"Multiple R-squared"* que é o $R^2$. Neste caso, seu valor visto em porcentagem é de $95,96\%$ significando que a quantidade de horas estudadas tem um impacto de $95,96\%$ na taxa média de aprovação.

Após realizado todas as análises gráficas, testes do modelo e verificado que ele pode ser interpretado, agora será construído um gráfico de dispersão das variáveis do modelo e a equação da reta que melhor lhes representa. Isso é feito utilizando a função:

```
> ggplot(data = dados, mapping = aes(x = Horas , y = Aprovação))
+ geom_point() + geom_smooth(method = "lm", col = "red")
+ stat_regline_equation(aes(label = paste(..eq.label..,..adj.rr.label..,
+ sep ="*plain(\",\")~~")),+ label.x = 4.9,label.y = 80) + theme_classic()
+ theme(text = element_text(family = "Times New Roman", size = 20))
```

E obtém-se o seguinte gráfico:

Figura 26: Taxa de aprovação × Horas estudadas




Na Figura 26, a reta de regressão linear tem a equação $y = 13 + 14x$. A faixa na cor cinza é o intervalo de confiança dos pontos que se distanciam da reta com 95% de certeza.

## 4.2 Regressão linear simples entre alunos por turma e taxa de aprovação

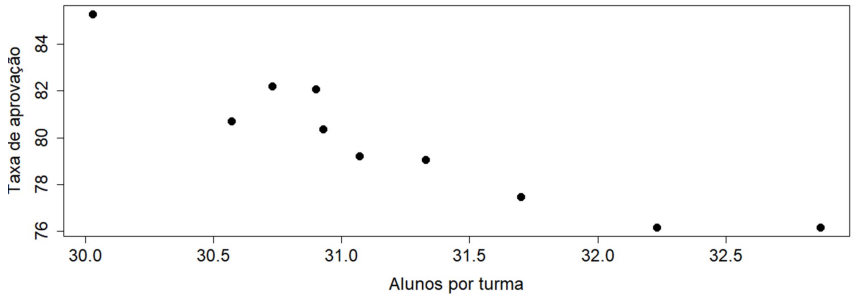
O modelo de regressão linear em questão pretende verificar se a quantidade de alunos por turma do Ensino Médio público do Brasil se reflete na taxa de aprovação nacional. Em outras palavras, através dessa regressão pretende responder a seguinte

pergunta: Caso as turmas do Ensino Médio que compõem o ensino público do Brasil diminuírem a quantidade de alunos com o passar dos anos isso pode realmente refletir em um maior desenvolvimento de seu conhecimento resultando em um aumento do número médio da taxa de aprovação nacional?

Mas antes disso, é necessário verificar os pressupostos de uma regressão linear simples. Neste modelo, a variável dependente é taxa de aprovação e a variável independente é média de alunos por turma. Com as variáveis estabelecidas e com os dados carregados no R é preciso verificar se há alguma relação linear entre elas. No R a função `"plot"` é utilizada para a construção de gráficos. Assim:

```
> plot(dados$Alunos,dados$Aprovação, xlab="Alunos por turma", ylab="Taxa
de aprovação", cex=1.5, cex.lab=1.5, cex.axis=1.5, pch=19)
```

Figura 27: Relação linear entre taxa de aprovação e alunos por turma




Observando a Figura 27 é possível notar que os pontos não formam perfeitamente uma relação linear, mas se aproxima dela. Desse modo, realiza-se o teste de coeficiente de correlação entre essas duas variáveis no R através da função:

```
cor.test(dados$Alunos,dados$Aprovação)
```

Obtendo:

Figura 28: *Output* teste de coeficiente de correlação

```
        Pearson's product-moment correlation

data:  dados$Alunos and dados$Aprovação
t = -6.7289, df = 8, p-value = 0.0001482
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9816903 -0.6965376
sample estimates:
       cor
-0.9218701
```

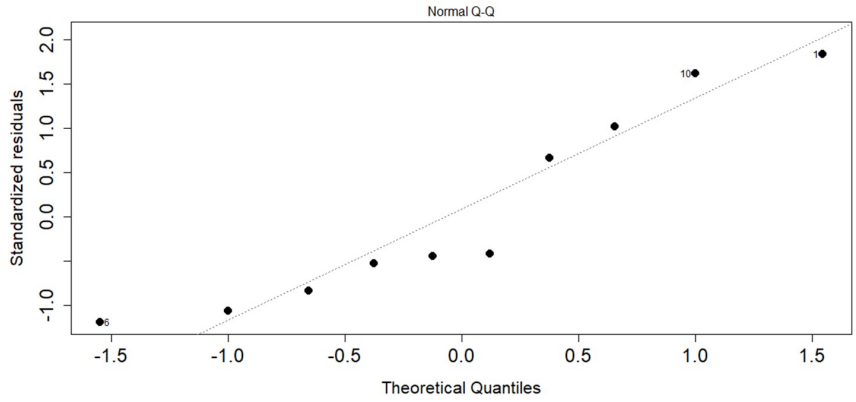
Do teste de coeficiente de correlação linear de Pearson realizado no R a correlação entre as variáveis é de aproximadamente $-0,92$. Como essa é uma correlação forte e além disso o p-valor desse teste é de $0,0001482$ indicando a rejeição da hipótese de que o coeficiente de correlação seja igual a zero. Assim pode-se afirmar que a distribuição gráfica dessas variáveis possui uma relação linear.

Assim, serão analisados os pressupostos do modelo de regressão para informar com mais precisão se realmente essas variáveis formam uma regressão linear. Mas antes disso, no R, é preciso construir tal modelo. O modelo a ser analisado será chamado de "`mod`", e para criá-lo é utilizada a função "`lm`" do inglês *linear model*. Assim:

```
> mod <- lm(Aprovação ~ Alunos, dados)
```

Os gráficos utilizados para a análise dos pressupostos de normalidade, *outliers*, independência e homocedasticidade dos resíduos são obtidos pelo comando `plot(mod)` no R. O primeiro pressuposto a ser analisado nos resíduos é a sua normalidade. Para realizar a análise gráfica deste pressuposto é utilizado o gráfico *QQ plot*.

Figura 29: Normalidade dos resíduos




Para que os resíduos apresentem uma distribuição normal, os pontos da Figura 29 têm de estar sobre a linha pontilhada. Neste caso, não estão perfeitamente alinhados, mas estão aproximados dela. Disto é possível considerar que há uma distribuição normal dos resíduos, mas aplicando o teste de Shapiro-Wilk no modelo é possível ter uma afirmação mais precisa sobre tal.

Figura 30: *Output* teste Shapiro-Wilk

```
        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.86878, p-value = 0.09674
```

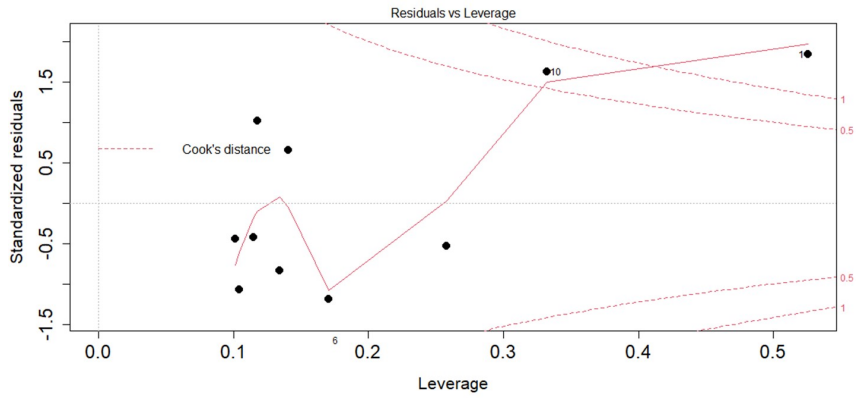
Lembrando que o teste de Shapiro-Wilk possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{os resíduos do modelo seguem uma distribuição normal} \\ H_1 : \text{os resíduos do modelo não seguem uma distribuição normal} \end{cases}$$

Considerando o teste para 5% de significância e observando que o p-valor obtido nesse teste foi de $0,09674$ então, neste caso, a hipótese nula é considerada podendo afirmar que os resíduos do modelo em questão seguem uma distribuição normal.

O próximo pressuposto é a existência de *outliers*. Realizando uma análise gráfica da Figura 31 pode ser observado que existe um ponto além da linha vermelha pontilhada, então utilizando a função `summary(rstandard(mod))` é possível afirmar se o ponto em questão pode ser considerado um ponto de alavancagem.

Figura 31: *Outliers* nos resíduos



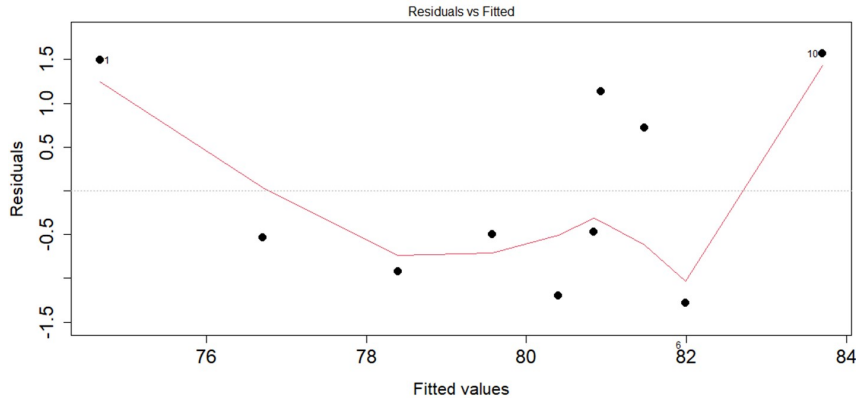

Das informações, como mostra a Figura 32, os valores padronizados estão entre $-1,19$ e $1,84$ indicando que não há a existência de pontos de alavancagem. Assim, é possível afirmar que não há a existência de outlier nos resíduos do modelo em questão.

Figura 32: *Output* função summary

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.19410 -0.76122 -0.43493  0.06341  0.92900  1.83679
```

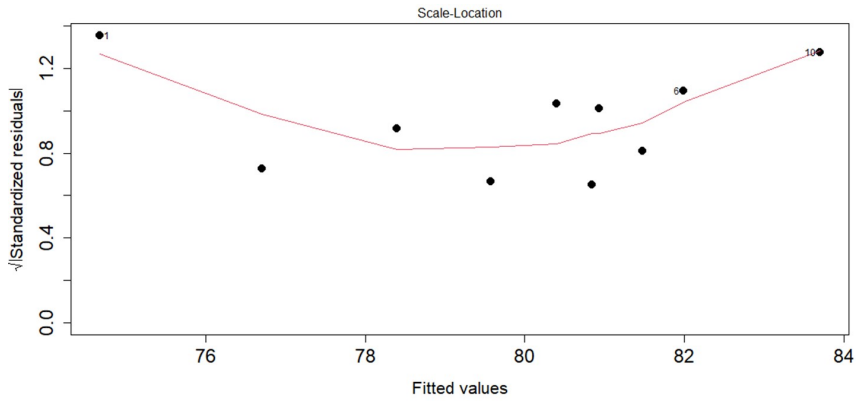
A independência dos resíduos, é o terceiro pressuposto a ser analisado podendo ser observado pelas Figura 33 e Figura 34.

Figura 33: Independência dos resíduos




Na Figura 33 não é possível afirmar se a distribuição dos pontos em torno do zero acontece com a mesma amplitude, já na Figura 34 para que exista a independência a linha vermelha tem de ser aproximadamente horizontal, o que não acontece.

Figura 34: Independência dos resíduos




Disto não é possível afirmar pela análise gráfica se os resíduos do modelo em questão são independentes. Logo, aplicando o teste de Durbin-Watson obtém-se:

Figura 35: *Output* teste Durbin-Watson

```
lag Autocorrelation D-W Statistic p-value
  1       0.3850863     0.8108377   0.006
Alternative hypothesis: rho != 0
```

Ressaltando que o teste de Durbin-Watson possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{não existe autocorrelação nos resíduos do modelo} \\ H_1 : \text{existe autocorrelação nos resíduos do modelo} \end{cases}$$

Tomando um nível de 5% de significância para o teste e observando que o p-valor obtido foi de $0,006$ então a hipótese nula é rejeitada, pois o p-valor obtido é menor que o nível de significância do teste. Assim, é possível afirmar que os resíduos não são independentes, ou ainda que existe autocorrelação entre eles.

Como o modelo de regressão linear testado falhou no pressuposto de independência dos resíduos, ou seja, a variável *"alunos por turma"* e o termo de erro são correlacionáveis, então esse modelo não pode ser interpretado.

Assim, para que todos os pressupostos de um modelo de regressão linear sejam atendidos, a variável *"alunos por turma"* irá sofrer a seguinte transformação:

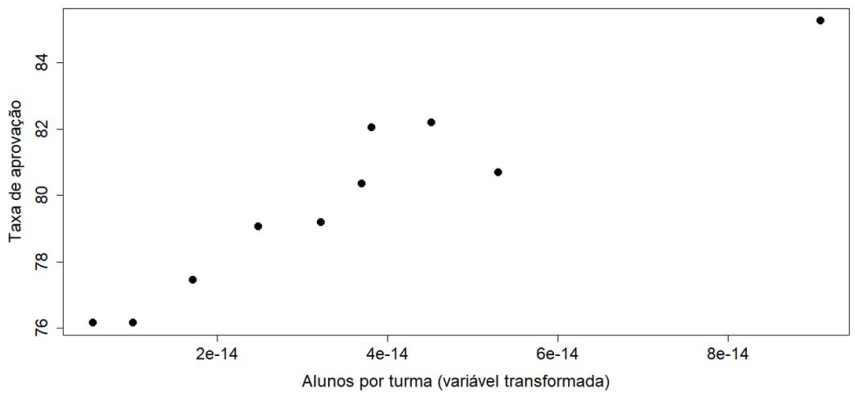
$$Y_i = \frac{1}{e^{X_i}}$$

Onde, $X_i$ e $Y_i$ são respectivamente, os elementos das variáveis *"alunos por turma"* e *"AlunosT"*.

Para esse novo modelo que será analisado, a variável dependente é *"taxa de aprovação"* e a variável independente é *"AlunosT"* (variável transformada). Com as variáveis estabelecidas é preciso verificar se há relação de linearidade entre elas.

```
> plot(dados$AlunosT, dados$Aprovação, xlab="Alunos por turma (variável
transformada)", ylab="Taxa de aprovação", cex=1.5, cex.lab=1.5,
cex.axis=1.5, pch=19)
```

Figura 36: Relação linear entre taxa de aprovação e AlunosT




Observando a Figura 36 é possível notar que os pontos não formam perfeitamente uma relação linear, mas se aproxima dela. Desse modo, realiza-se o teste de coeficiente de correlação entre essas duas variáveis no R através da função:

```
> cor.test(dados$Aprovação, dados$AlunosT)
```

Obtendo:

Figura 37: *Output* teste de coeficiente de correlação

```
        Pearson's product-moment correlation

data:  dados$Aprovação and dados$AlunosT
t = 7.9378, df = 8, p-value = 4.619e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7676558 0.9865126
sample estimates:
      cor
0.9419861
```

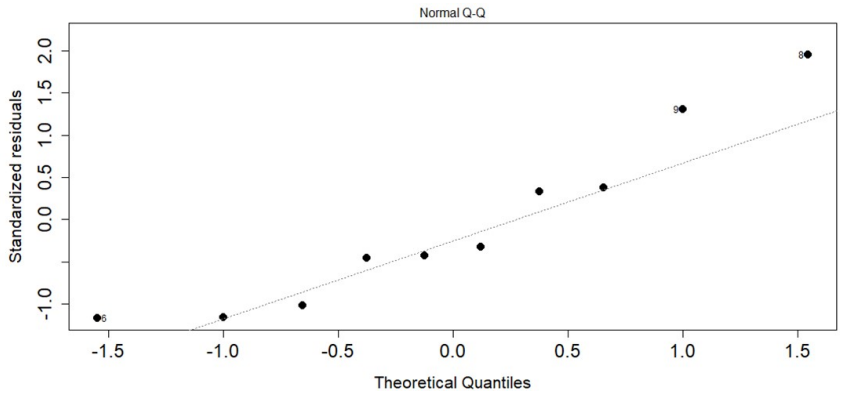
Do teste de coeficiente de correlação linear de Pearson realizado no R a correlação entre as variáveis é de aproximadamente $0,94$. Como essa é uma correlação forte e além disso o p-valor desse teste é de $4,619 \times 10^{-5}$ indicando a rejeição da hipótese de que o coeficiente de correlação seja igual a zero. Assim pode-se afirmar que a distribuição gráfica dessas variáveis possui uma relação linear.

Assim, serão analisados os pressupostos do modelo de regressão para informar com mais precisão se realmente essas variáveis formam uma regressão linear. Mas antes disso, no R, é preciso construir tal modelo. O modelo a ser analisado será chamado de "`mod2`", e para criá-lo é utilizada a função "`lm`" do inglês *linear model*. Assim:

```
> mod2 <- lm(Aprovação ~ AlunosT, dados)
```

Os gráficos utilizados para a análise dos pressupostos de normalidade, *outliers*, independência e homocedasticidade dos resíduos são obtidos pelo comando `plot(mod2)` no R. O primeiro pressuposto a ser analisado nos resíduos é a sua normalidade. Para realizar a análise gráfica deste pressuposto é utilizado o gráfico *QQ plot*.

Figura 38: Normalidade dos resíduos



Para que os resíduos apresentem uma distribuição normal, os pontos da Figura 38 têm de estar sobre a linha pontilhada. Neste caso, não estão perfeitamente alinhados, mas estão aproximados dela. Disto é possível considerar que há uma distribuição normal dos resíduos, mas aplicando o teste de Shapiro-Wilk no modelo é possível ter uma afirmação mais precisa sobre tal.

Figura 39: *Output* teste Shapiro-Wilk

```
        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.90644, p-value = 0.2574
```

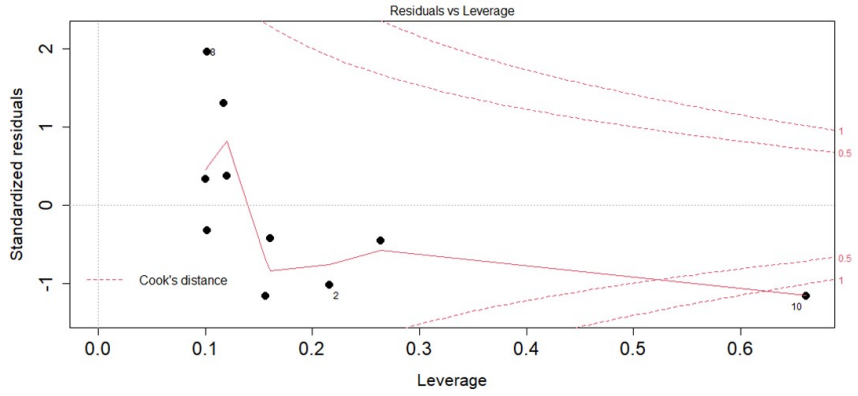
Lembrando que o teste de Shapiro-Wilk possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{os resíduos do modelo seguem uma distribuição normal} \\ H_1 : \text{os resíduos do modelo não seguem uma distribuição normal} \end{cases}$$

Considerando o teste para 5% de significância e observando que o p-valor obtido nesse teste foi de $0,2574$ então, a hipótese nula é considerada podendo afirmar que os resíduos do modelo em questão seguem uma distribuição normal.

O próximo pressuposto é a existência de *outliers*. Realizando uma análise gráfica da Figura 40 pode ser observado que existe um ponto além da linha vermelha pontilhada, então utilizando a função `summary(rstandard(mod2))` é possível afirmar se o ponto em questão pode ser considerado um ponto de alavancagem.

Figura 40: *Outliers* nos resíduos



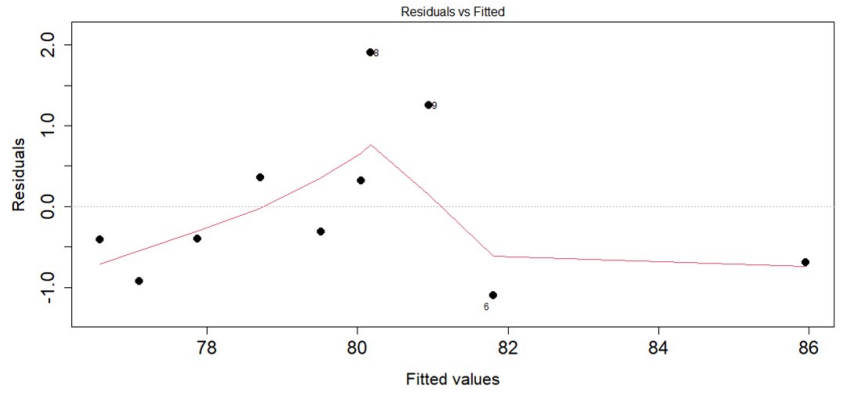

Das informações, como mostra a Figura 41, os valores padronizados estão entre $-1,17$ e $1,96$ indicando que não há a existência de pontos de alavancagem. Assim, é possível afirmar que não há a existência de outlier nos resíduos do modelo em questão.

Figura 41: *Output* função summary

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.17178 -0.88018 -0.37489 -0.05958  0.36493  1.95611
```

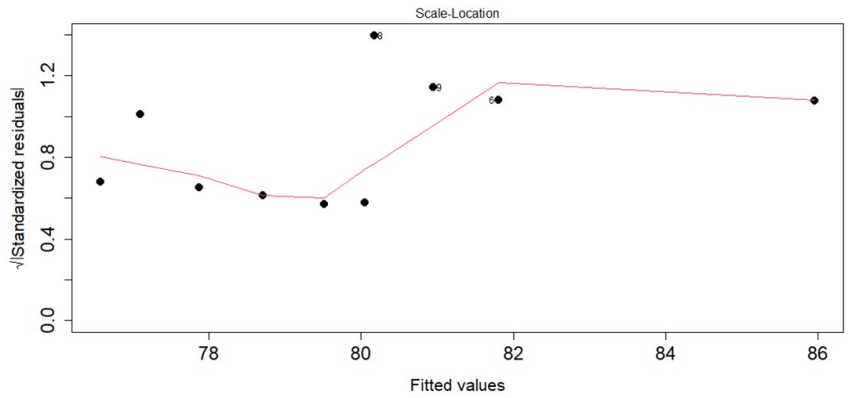
A independência dos resíduos, é o terceiro pressuposto a ser analisado podendo ser observado pelas Figura 42 e Figura 43.

Figura 42: Independência dos resíduos




Na Figura 42 não é possível afirmar se a distribuição dos pontos em torno do zero acontece com a mesma amplitude, já na Figura 43 para que exista a independência, a linha vermelha tem de ser aproximadamente horizontal, o que não acontece.

Figura 43: Independência dos resíduos




Disto não é possível afirmar pela análise gráfica se os resíduos do modelo em questão são independentes. Logo, aplicando o teste de Durbin-Watson obtém-se:

Figura 44: *Output* teste Durbin-Watson

```
lag Autocorrelation D-W Statistic p-value
  1       0.310246      1.302725   0.124
Alternative hypothesis: rho != 0
```

Ressaltando que o teste de Durbin-Watson possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{não existe autocorrelação nos resíduos do modelo} \\ H_1 : \text{existe autocorrelação nos resíduos do modelo} \end{cases}$$

Considerando o teste com 5% de significância e observando que o p-valor obtido neste teste foi de $0,124$ então a hipótese nula é aceita. Assim, pode-se afirmar que no modelo em questão os resíduos são independentes.

O último pressuposto a ser analisado é sobre a homocedasticidade dos resíduos. Pela análise gráfica, para que exista a homocedasticidade no modelo, no gráfico da Figura 42 teria de haver uma linha vermelha no sentido horizontal, o que não acontece. Mas aplicando o teste Breusch-Pagan no modelo tem-se:

Figura 45: *Output* teste Breusch-Pagan

```
        studentized Breusch-Pagan test

data:  mod
BP = 0.29964, df = 1, p-value = 0.5841
```

Lembrando que as hipóteses do teste de Breusch-Pagan são:

$$\begin{cases} H_0 : \text{há homocedasticidade dos resíduos do modelo} \\ H_1 : \text{não há homocedasticidade dos resíduos do modelo} \end{cases}$$

Para um nível de 5% de significância e pelo fato do p-valor do teste de Breusch-Pagan ser de $0,5841$, então é considerada a hipótese nula, logo pode-se afirmar que há homoscedasticidade entre os resíduos do modelo.

Os testes de hipótese para os resíduos do modelo confirmam que todos os pressupostos são atendidos. Assim, resta realizar uma análise do modelo utilizando a função `summary(mod2)`.

A Figura 46 fornece um resumo das informações do modelo que está sendo analisado. Onde:

- *"call"* é a descrição do modelo de regressão linear analisado;
- *"residuals"* são os valores dos resíduos que não estão padronizados;
- *"intercept"* é o ponto em que a reta corta o eixo $y$, neste caso é a variável *"Aprovação"* (variável dependente) quando a variável *"AlunosT"* (variável independente) é zero.

Figura 46: *Output* função summary

```
Call:
lm(formula = dados$Aprovação ~ dados$AlunosT)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1030 -0.6218 -0.3573  0.3517  1.9006

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.599e+01  5.863e-01 129.613 1.40e-14 ***
dados$AlunosT 1.098e-14  1.384e+13   7.938 4.62e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.025 on 8 degrees of freedom
Multiple R-squared:  0.8873,    Adjusted R-squared:  0.8733
F-statistic: 63.01 on 1 and 8 DF,  p-value: 4.619e-05
```

A variável independente *"AlunosT"* tem um *p-valor* de $4,62 \times 10^{-5}$. O p-valor para a variável independente é baseado no teste t de $t$ de *Student* que possui as seguintes hipóteses:

$$\begin{cases} H_0 : \text{o coeficiente é igual a zero} \\ H_1 : \text{o coeficiente é diferente de zero} \end{cases}$$

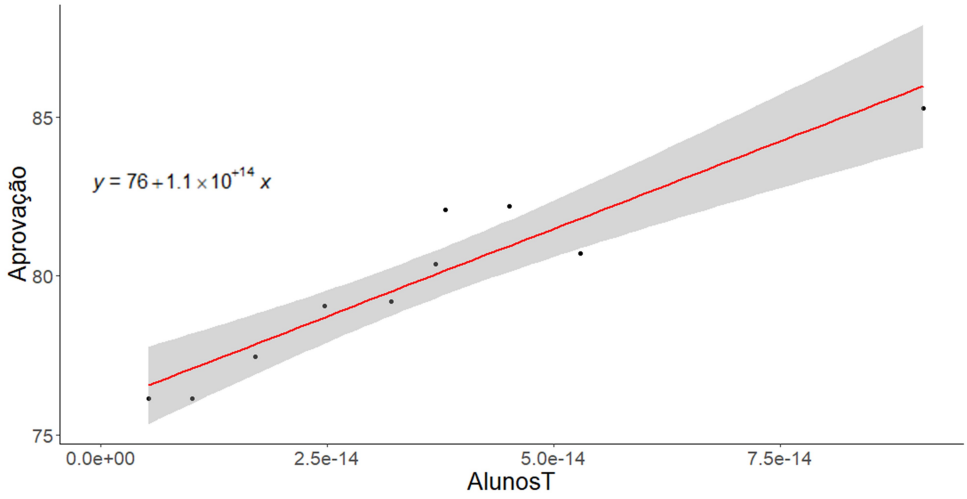
Para um teste ao nível de 5% de significância e com um p-valor obtido de $0,0000462$, então rejeita-se a hipótese nula, portanto o coeficiente é diferente de zero podendo ser interpretado, mostrando que a variável independente tem um impacto sobre a variável dependente.

Após realizado todas as análises gráficas, testes do modelo e verificado que ele pode ser interpretado, agora será construído um gráfico de dispersão das variáveis e a equação da reta que melhor lhes representa. Isso é feito utilizando a função:

```
> ggplot(data = dados, mapping = aes(x = Aprovação, y = AlunosT))
+ geom_point() + geom_smooth(method = "lm", col = "red")
+ stat_regline_equation(aes(label = paste(..eq.label..,
sep = "*plain(\",\")~~")) + label.x = 77, label.y = 1) + theme_classic()
```

E obtém-se o seguinte gráfico:

Figura 47: Taxa de aprovação × AlunosT




Na Figura 47, a reta de regressão linear que tem a equação $y = 76 + 1,1 \times 10^{-14}x$. A faixa na cor cinza é o intervalo de confiança dos pontos que se distanciam da reta com 95% de certeza.

Este modelo envolve uma variável transformada e sobre ela foi realizada a checagem dos pressupostos para uma regressão linear simples, onde todos foram atendidos. Então nesse caso toda a apresentação dos resultados tem de ser realizada com a variável original.

Lembrando que a variável original é "*alunos por turma*" e nomeando seus elementos como $a_1, a_2, \ldots, a_{10}$ e a variável transformada é "*AlunosT*" e seus elementos chamando-os de $b_1, b_2, \ldots, b_{10}$. Como $a_1 > a_2 > \cdots > a_{10}$ e pelo fato da variável original ser inversamente proporcional a variável transformada então $b_1 < b_2 < \cdots < b_{10}$.

Na Figura 47 a reta que melhor representa a regressão do `mod2` é crescente, com a variável independente "*AlunosT*" no eixo das abscissas e a variável dependente "*taxa de aprovação*" no eixo das ordenadas. Assim, quanto mais se distancia o valor de *"AlunosT"* para a direita da origem, maior o valor da *"aprovação"*. Em suma, quanto maior o valor de *"AlunosT"* maior a *"taxa de aprovação"*.

Como as variáveis *"alunos por turma"* e *"AlunosT"* são inversamente proporcionais, então o gráfico de regressão no qual a variável *"alunos por turma"* apareça no eixo das abscissas e *"taxa de aprovação"* no eixo das ordenadas teríamos uma reta decrescente indicando que quanto maior a quantidade de alunos menor a taxa de aprovação.

Disto, é possível afirmar que a quantidade de alunos influencia na aprendizagem da turma.

## 4.3 Regressão linear múltipla entre horas estudadas, alunos por turma e taxa de aprovação

A regressão linear simples é composta por uma variável dependente e uma variável independente, já a regressão linear múltipla que é uma extensão da anterior é possível adicionar mais de uma variável independente. Aqui será trabalhado com a variável dependente taxa de aprovação e com variáveis independentes horas estudadas e média de alunos por turma.

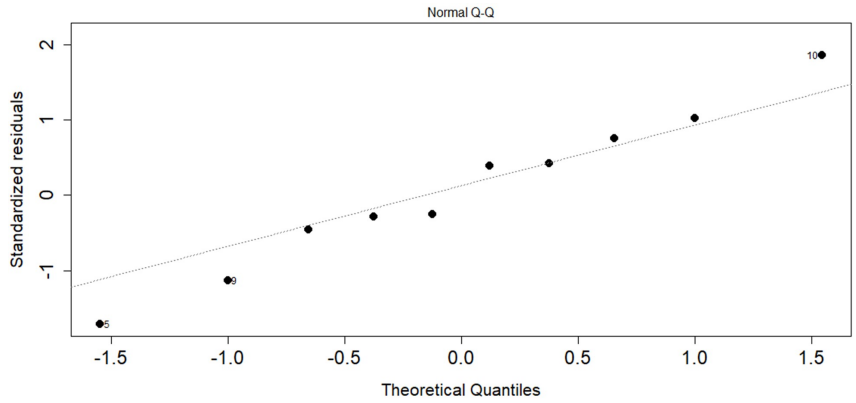
Pretende-se responder a seguinte pergunta: Caso a quantidade de horas estudadas pelos estudantes do Ensino Médio público do Brasil aumente e concomitantemente a quantidade de alunos por turma diminuir, essa condição irá refletir em um melhor aprendizado por parte dos educandos? Pois o professor terá mais tempo para trabalhar com seus discentes, podendo realizar atividades diferenciadas para potencializar sua aprendizagem e também realizar um atendimento individual para assim sanar as possíveis dúvidas.

Antes de partir para a análise de regressão é preciso construir um modelo e avaliar seus pressupostos. Conhecendo as variáveis e chamando o modelo de "`mod`", tem-se:

```
> mod <- lm(Aprovação ~ Horas + Alunos, dados)
```

Os gráficos utilizados para a análise dos pressupostos de normalidade, *outliers*, independência e homocedasticidade dos resíduos são obtidos pelo comando `plot(mod)` no R. O primeiro pressuposto a ser analisado nos resíduos é a sua normalidade. Para realizar a análise gráfica deste pressuposto é utilizado o gráfico *QQ plot*.

Figura 48: Normalidade dos resíduos




Para que os resíduos apresentem uma distribuição normal, os pontos da Figura 48 têm de estar sobre a linha pontilhada. Neste caso, isso não acontece, mas está bem próximo disso. Assim, pode-se considerar que há uma distribuição normal entre os resíduos. Mas aplicando o teste de Shapiro-Wilk, obtém-se:

Figura 49: *Output* teste Shapiro-Wilk

```
        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.97904, p-value = 0.9598
```

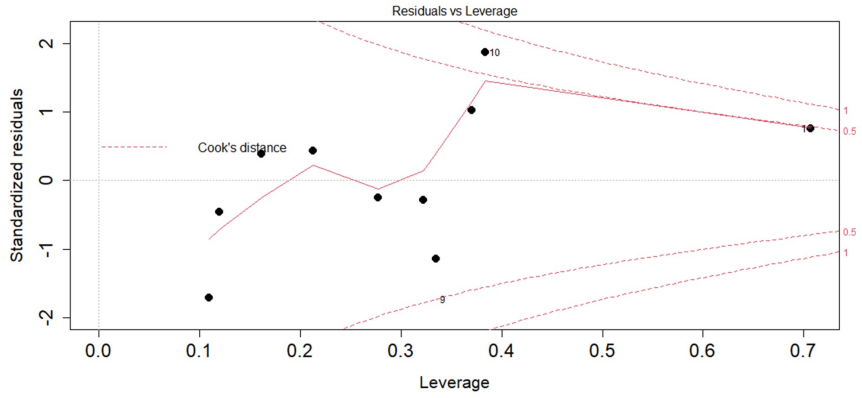
Lembrando que as hipóteses desse teste são:

$$\begin{cases} H_0 : \text{os resíduos do modelo seguem uma distribuição normal} \\ H_1 : \text{os resíduos do modelo não seguem uma distribuição normal} \end{cases}$$

Considerando o teste para 5% de significância e observando que o p-valor obtido nesse teste foi de $0,9598$ então, neste caso, a hipótese nula é considerada, podendo afirmar que os resíduos do modelo em questão seguem uma distribuição normal.

O segundo pressuposto é a existência de *outliers*. Pela análise gráfica da Figura 50, nota-se que não há pontos além da linha vermelha tracejada, mas pela função `summary(rstandard(mod))` pode-se afirmar se há ou não pontos de alavancagem que possam influenciar na existência de *outliers*.

Figura 50: *Outliers* nos resíduos



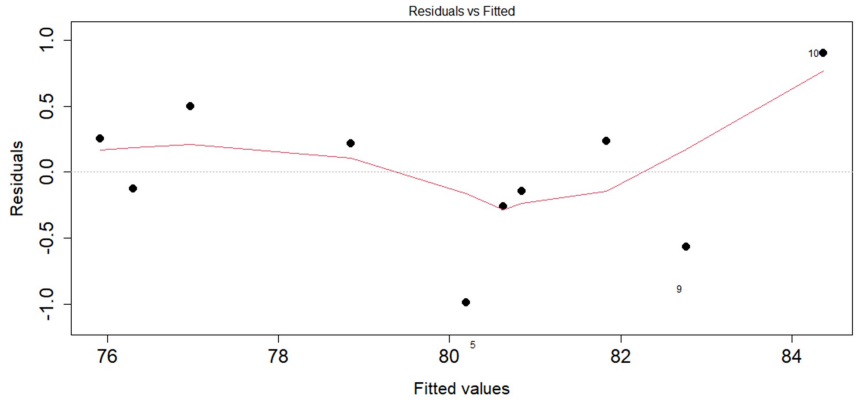

Da Figura 51 os valores padronizados de mínimo e máximo estão dentro do intervalo $[-3, 3]$ indicando que não há a existência de pontos de alavancagem. Assim, pode-se afirmar que não existe outlier nos resíduos do modelo em questão.

Figura 51: *Output* função summary

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.71399 -0.41409  0.06939  0.06175  0.67300  1.86577
```

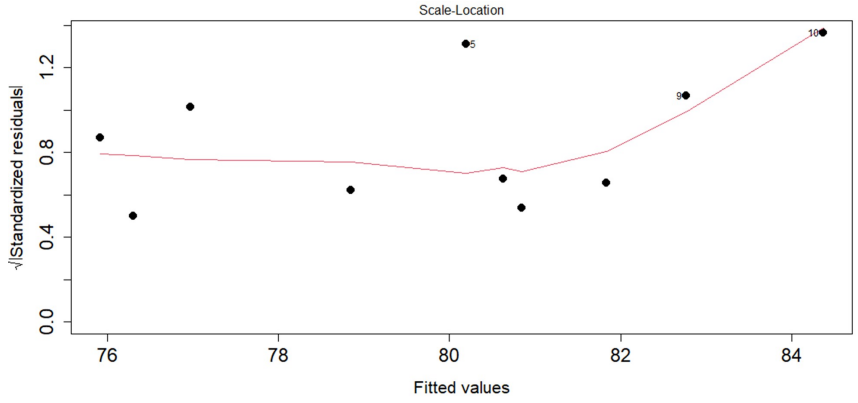
O terceiro pressuposto a ser analisado é a independência dos resíduos. Pela análise gráfica tem-se:

Figura 52: Independência dos resíduos




Da Figura 52, não é possível afirmar que a distribuição dos pontos em torno do zero acontece com a mesma amplitude.

Figura 53: Independência dos resíduos




Na Figura 53, para que haja a independência, a linha vermelha deveria ser aproximadamente horizontal. Assim, não é possível afirmar pela análise gráfica se os resíduos do modelo são independentes. Então, aplicando o teste de Durbin-Watson no modelo é possível afirmar se há ou não a independência dos resíduos.

Figura 54: *Output* teste Burbin-Watson

```
lag Autocorrelation D-W Statistic p-value
  1      -0.2751418        2.2201   0.768
Alternative hypothesis: rho != 0
```

Lembrando que o teste de Durbin-Watson tem as seguintes hipóteses:

$$\begin{cases} H_0 : \text{não existe autocorrelação nos resíduos do modelo} \\ H_1 : \text{existe autocorrelação nos resíduos do modelo} \end{cases}$$

Considerando um nível de 5% de significância, nota-se que o p-valor obtido no teste é de $0,768$ então a hipótese nula é aceita. Assim, no modelo em questão os resíduos são independentes.

O quarto pressuposto a ser analisado é sobre a homocedasticidade dos resíduos. Pela análise gráfica, para que exista a homocedasticidade no modelo, na Figura 53 teria de haver uma linha vermelha no sentido horizontal, o que não acontece. Mas aplicando o teste Breusch-Pagan no modelo tem-se:

Figura 55: *Output* teste Breusch-Pagan

```
        studentized Breusch-Pagan test

data:  mod
BP = 1.9994, df = 2, p-value = 0.368
```

Lembrando que as hipóteses do teste de Breusch-Pagan são:

$$\begin{cases} H_0 : \text{há homocedasticidade dos resíduos do modelo} \\ H_1 : \text{não há homocedasticidade dos resíduos do modelo} \end{cases}$$

Para um nível de 5% de significância e pelo p-valor do teste de Breusch-Pagan ser $0,368$ então é considerada a hipótese nula. Logo, pode-se afirmar que há homocedasticidade entre os resíduos do modelo.

Até o presente momento, os testes de hipótese para os resíduos realizados confirmam que todos os pressupostos são atendidos. Havendo uma relação linear com os resíduos, seguindo uma distribuição normal, não tendo pontos influentes e havendo a existência de homoscedasticidade.

O pressuposto que surge na regressão linear múltipla é que não deve haver multicolinearidade entre as variáveis independentes, ou seja, o coeficiente de correlação linear entre as variáveis *"Alunos por turma"* e *"Horas estudadas"* não deve ser superior a $0,9$ ou inferior a $-0,9$. Pela função: `with(dados, cor.test(Alunos, Horas, alternative = "two.sided", method = "pearson"))`, tem-se:

Figura 56: *Output* correlação de Pearson

```
        Pearson's product-moment correlation

data:  Alunos and Horas
t = -6.2567, df = 8, p-value = 0.0002439
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9791046 -0.6605805
sample estimates:
       cor
-0.9112152
```

Da Figura 56 pode-se observar que o coeficiente de correlação linear de Pearson entre as variáveis independentes do modelo é aproximadamente $-0,911$. Como essa é uma correlação forte e além disso o p-valor desse teste é de $0,0002439$ indicando a rejeição da hipótese de que o coeficiente de correlação é igual a zero. Desse modo, pode-se afirmar que há multicolinearidade entre essas duas variáveis independentes. Assim, o modelo em questão não pode ser interpretado, pois as variáveis em questão são fortemente correlacionadas, desse modo é muito difícil haver variação entre uma sem que haja em outra.

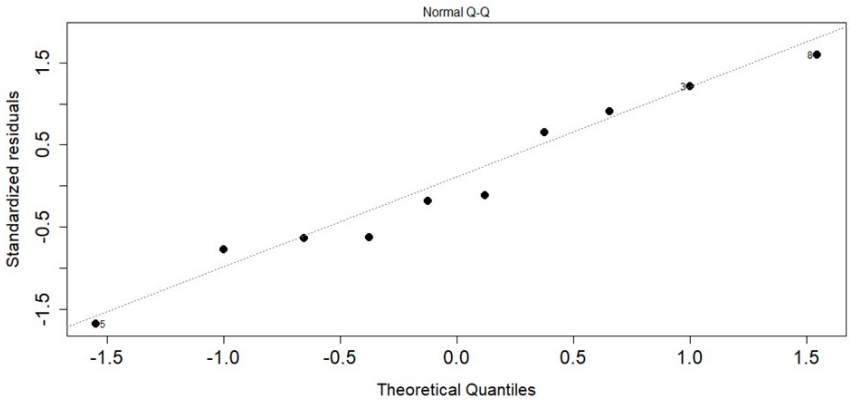
Então será construído um novo modelo de regressão linear múltipla, agora utilizando a variável *"AlunosT"* da Tabela 1 ao invés de *"alunos por turma"*, como já realizado no item 4.2.

Antes de partir para a análise de regressão é preciso construir um modelo e avaliar seus pressupostos. Conhecendo as variáveis e chamando o modelo de "`mod2`", tem-se:

```
> mod2 <- lm(Aprovação ~ Horas + AlunosT, dados)
```

Os gráficos utilizados para a análise dos pressupostos de normalidade, *outliers*, independência e homocedasticidade dos resíduos são obtidos pelo comando `plot(mod2)` no R. O primeiro pressuposto a ser analisado nos resíduos é a sua normalidade. Para realizar a análise gráfica deste pressuposto é utilizado o gráfico *QQ plot*.

Figura 57: Normalidade dos resíduos



Para que os resíduos apresentem uma distribuição normal, os pontos da Figura 57 têm de estar sobre a linha pontilhada. Neste caso, isso não acontece, mas está bem próximo disso. Assim, pode-se considerar que há uma distribuição normal entre os resíduos. Mas aplicando o teste de Shapiro-Wilk, obtém-se:

Figura 58: *Output* teste Shapiro-Wilk

```
        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.96169, p-value = 0.805
```

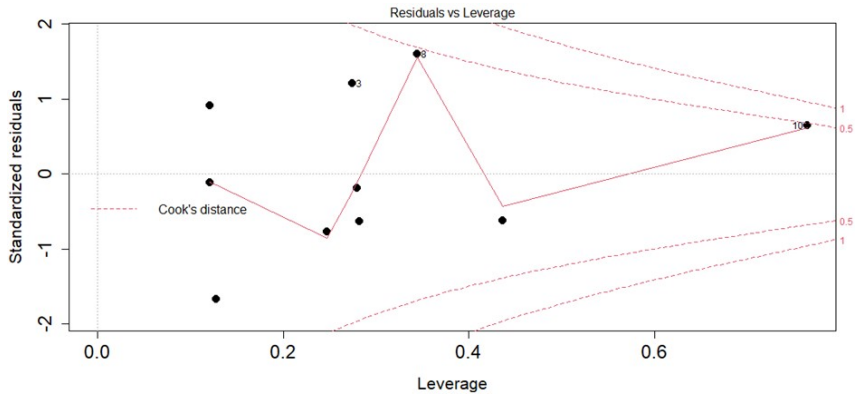
Lembrando que as hipóteses desse teste são:

$$\begin{cases} H_0 : \text{os resíduos do modelo seguem uma distribuição normal} \\ H_1 : \text{os resíduos do modelo não seguem uma distribuição normal} \end{cases}$$

Considerando o teste para 5% de significância e observando que o p-valor obtido nesse teste foi de $0,805$ então, neste caso, a hipótese nula é considerada, podendo afirmar que os resíduos do modelo em questão seguem uma distribuição normal.

O segundo pressuposto é a existência de *outliers*. Pela análise gráfica da Figura 59, nota-se que não há pontos além da linha vermelha tracejada, mas pela função `summary(rstandard(mod2))` pode-se afirmar se há ou não pontos de alavancagem que possam influenciar na existência de *outliers*.

Figura 59: *Outliers* nos resíduos



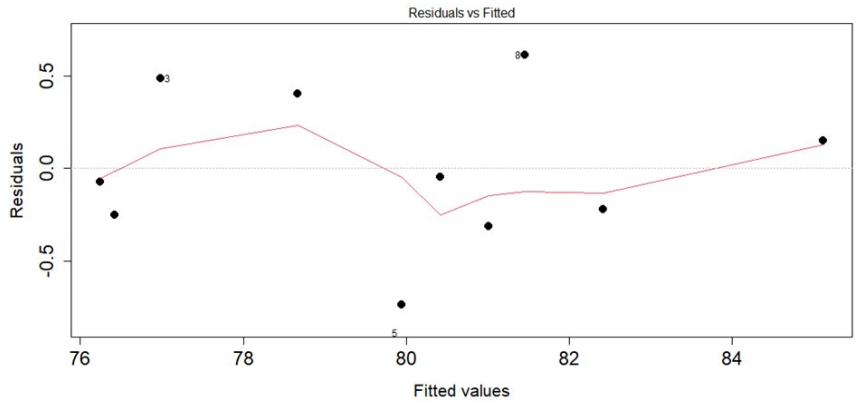

Da Figura 60 os valores padronizados de mínimo e máximo estão dentro do intervalo [-3,3] indicando que não há a existência de pontos de alavancagem. Assim, pode-se afirmar que não existe outlier nos resíduos do modelo em questão.

Figura 60: *Output* função summary

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.67700 -0.62829 -0.14811  0.03758  0.84551  1.59812
```

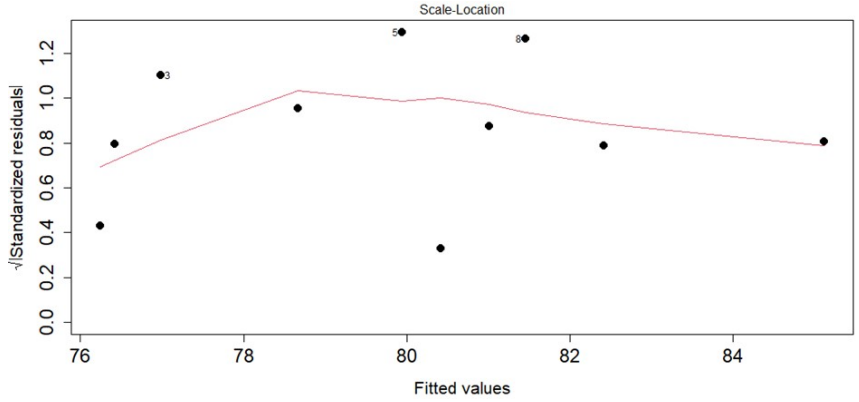
O terceiro pressuposto a ser analisado é a independência dos resíduos. Pela análise gráfica tem-se:

Figura 61: Independência dos resíduos




Da Figura 61, não é possível afirmar que a distribuição dos pontos em torno do zero acontece com a mesma amplitude.

Figura 62: Independência dos resíduos




Na Figura 62, para que haja a independência a linha vermelha deveria ser aproximadamente horizontal. Assim, não é possível afirmar pela análise gráfica se os resíduos do modelo são independentes. Então, aplicando o teste de Durbin-Watson no modelo é possível afirmar se há ou não a independência dos resíduos.

Figura 63: *Output* teste Durbin-Watson

```
lag Autocorrelation D-W Statistic p-value
  1     -0.09973123      2.181736   0.844
Alternative hypothesis: rho != 0
```

Lembrando que o teste de Durbin-Watson tem as seguintes hipóteses:

$$\begin{cases} H_0 : \text{não existe autocorrelação nos resíduos do modelo} \\ H_1 : \text{existe autocorrelação nos resíduos do modelo} \end{cases}$$

Considerando um nível de 5% de significância, nota-se que o p-valor obtido no teste é de $0,844$ então a hipótese nula é aceita. Assim, no modelo em questão os resíduos são independentes.

O quarto pressuposto a ser analisado é sobre a homocedasticidade dos resíduos. Pela análise gráfica, para que exista a homocedasticidade no modelo, na Figura 61 teria de haver uma linha vermelha no sentido horizontal, o que não acontece. Mas aplicando o teste Breusch-Pagan no modelo tem-se:

Figura 64: *Output* teste Breusch-Pagan

```
        studentized Breusch-Pagan test

data:  mod
BP = 1.0716, df = 2, p-value = 0.5852
```

Lembrando que as hipóteses do teste de Breusch-Pagan são:

$$\begin{cases} H_0 : \text{há homocedasticidade dos resíduos do modelo} \\ H_1 : \text{não há homocedasticidade dos resíduos do modelo} \end{cases}$$

Para um nível de 5% de significância e pelo p-valor do teste de Breusch-Pagan ser $0,5852$ então é considerada a hipótese nula. Logo, pode-se afirmar que há homocedasticidade entre os resíduos do modelo.

Até o presente momento, os testes de hipótese para os resíduos realizados confirmam que todos os pressupostos são atendidos. Havendo uma relação linear com os resíduos, seguindo uma distribuição normal, não tendo pontos influentes e havendo a existência de homoscedasticidade.

Disto, realizando o teste de correlação linear de Pearson entre as variáveis independentes do modelo, pela função:

```
> with(dados, cor.test(AlunosT,Horas, alternative = "two.sided", method =
"pearson"))
```

Tem-se:

Figura 65: *Output* correlação de Pearson

```
        Pearson's product-moment correlation

data:  AlunosT and Horas
t = 5.8146, df = 8, p-value = 0.0003985
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6215404 0.9761756
sample estimates:
      cor
0.8992542
```

Da Figura 65 pode-se observar que o coeficiente de correlação linear de Pearson entre as variáveis *"AlunosT"* e *"Horas estudadas"* é de $0,899$. Como essa correlação é aproximadamente $0,9$ e além disso o p-valor desse teste é de $0,0003985$ indicando a rejeição da hipótese de que o coeficiente de correlação é igual a zero. Desse modo, é possível afirmar que não há multicolinearidade entre as variáveis independentes. Disto resta analisar o modelo utilizando a função `summary(mod2)`.

A Figura 66 fornece um resumo das informações do modelo que está sendo analisado. São elas:

- *"call"* a descrição do modelo de regressão linear analisado;
- *"residuals"* são os valores dos resíduos que não estão padronizados.

Figura 66: *Output* função summary

```
Call:
lm(formula = Aprovação ~ Horas + AlunosT, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73920 -0.24413 -0.06163  0.33956  0.61081

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.110e+01  8.105e+00   3.837 0.006401 **
Horas       9.989e+00  1.802e+00   5.542 0.000867 ***
AlunosT     3.722e-13  1.457e+13   2.555 0.037808 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.472 on 7 degrees of freedom
Multiple R-squared:  0.9791,    Adjusted R-squared:  0.9731
F-statistic: 163.9 on 2 and 7 DF,  p-value: 1.322e-06
```

O *estimate* da variável *"horas estudadas"* é $9,989$ com p-valor de $0,000867$. Então tem-se as seguintes hipóteses:

$$\begin{cases} H_0 : \text{o coeficiente é igual a zero} \\ H_1 : \text{o coeficiente é diferente de zero} \end{cases}$$

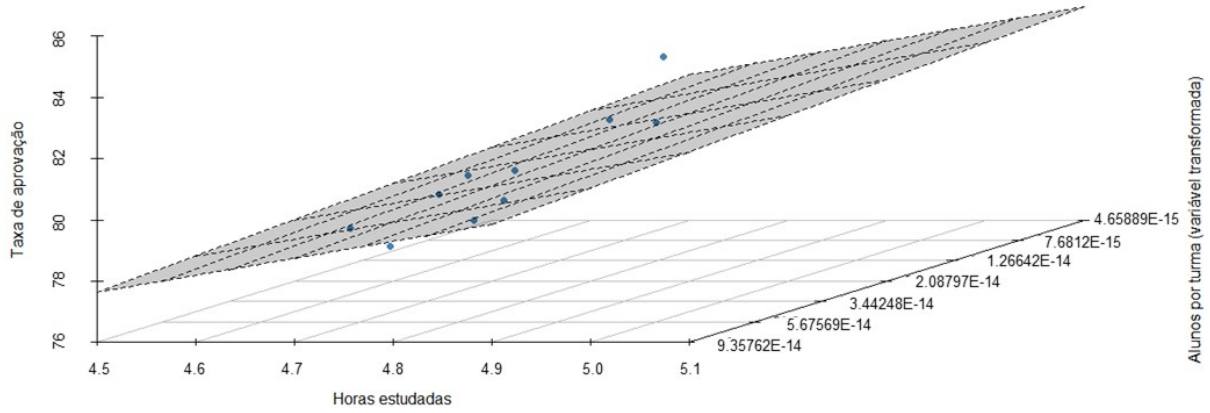
Para um teste de 5% de significância e conhecendo o p-valor da variável, então a hipótese nula é rejeitada pois p-valor é menor que 5%. Assim, a estimativa de horas estudadas indica que a cada hora a mais estudada em sala de aula a taxa de aprovação aumenta aproximadamente 10%.

Já para a variável média de alunos por turma o p-valor obtido é de $0,038$ inferior a 5% então a hipótese nula é rejeitada. Logo, a estimativa de alunos por turma indica que a cada aluno a menos em sala de aula a taxa de aprovação aumenta em $3,722 \times 10^{-13}$.

Após analisar todos os testes dos pressupostos será construído o gráfico de dispersão das variáveis. Como o modelo em questão possui três variáveis então seu gráfico é tridimensional e ao invés de uma reta, um plano irá representar melhor a dispersão dos pontos. Para isso será utilizada a seguinte função:

```
> graph <- scatterplot3d(dados$Aprovação ~ dados$Horas + dados$AlunosT,
+ pch = 19, angle = 30, color = "steelblue", box = FALSE, + xlab="Horas
estudadas", ylab=" Taxa de aprovação ", zlab="Alunos por turma (variável
transformada)")
> graph$plane3d(mod, col="black", draw_polygon = TRUE)
```

Figura 67: Taxa de aprovação × Horas estudadas e Alunos por turma (variável transformada)




O plano que aparece no gráfico é a previsão do modelo. Se o modelo estivesse acertando 100%, todos os pontos do gráfico estariam sobre esse plano. Como há pontos acima e abaixo do plano, o modelo tem erros que são esperados, os chamados resíduos.

Segundo a Equação 18, o plano da Figura 67 pode ser escrito como: $y = 31,1 + 9,989X_1 + 3,722 \times 10^{-13}X_2$, onde $X_1$ é a variável *"Horas estudadas"* e $X_2$ a variável *"AlunosT"*.

Mas como este modelo envolve uma variável transformada e sobre ela foi realizada a checagem dos pressupostos para uma regressão linear múltipla, onde todos foram atendidos. Então nesse caso toda a apresentação dos resultados tem de ser realizada com a variável original.

Da equação $y = 31,1 + 9,989X_1 + 3,722 \times 10^{-13}X_2$ do `mod2`, como o fator $3,722 \times 10^{-13}X_2$ é próximo a zero, então para qualquer valor de $X_2$ os valores de $y$ variam próximos a $y = 31,1 + 9,989X_1$.

Lembrando que $X_1$ é a variável *"Horas estudadas"* e $X_2$ a variável *"AlunosT"*. Mas as variáveis *"alunos por turma"* e *"AlunosT"* são inversamente proporcionais, assim: $a_1 > a_2 > \cdots > a_{10}$ e $b_1 < b_2 < \cdots < b_{10}$, onde $a_i$ e $b_i$, com $i = 1, 2, 3, \ldots, 10$, são respectivamente, os elementos da variável original *"alunos por turma"* e os elementos da variável transformada *"AlunosT"*.

Então pode-se afirmar que aumentando a quantidade de horas estudadas em sala e concomitantemente diminuindo a quantidade de alunos por turma obtém-se um aumento na taxa de aprovação.

# Considerações Finais

Diante das análises realizadas sobre os modelos vistos nas seções anteriores, é possível afirmar que aumentando a carga-horária do ensino em sala para os estudantes do Ensino Médio da rede pública do Brasil tem-se um aumento no rendimento na taxa de aprovação nacional. O mesmo também ocorre quando o número de estudantes em sala diminui. Por fim, a regressão linear múltipla mostrou que as variáveis independentes, horas estudadas e alunos por turma, ambas têm efeito sobre a variável dependente, no caso, taxa de aprovação.

Disto, com os dados analisados, pode-se afirmar que as variáveis horas estudadas e média de alunos por turma no Ensino Médio público do Brasil interferem no aumento da taxa de aprovação nacional. Ou seja, caso os estudantes que cursam o ensino médio público do país permaneçam mais horas diárias em aula e concomitantemente a quantidade de alunos que constituem as turmas diminuírem, com o decorrer dos anos essa condição realmente irá se refletir em um maior desenvolvimento acumulado de seus conhecimentos, resultando em um aumento do número médio da taxa de aprovação nacional.

# Referências Bibliográficas


1 ALCOFORADO, Luciane Ferreira; LEVY, Ariel. Visualização de dados com o *software* R. Niterói: LFA, 2017.

2 ALLAMAN, Ivan Bezerra. Transformação de dados. Ilhéus: Universidade Estadual de Santa Cruz, 2019.

3 BOX, George Edward Pelham; COX, David Roxbee. An analysis of transformations. Journal of the Royal Statistical Society, Edinburgh, Biometrika, Vol. 26, Issue 2, p. 211–252, 1964.

4 BRASIL. Constituição da República Federativa do Brasil, Brasília, DF, 1988.

5 BRASIL. Lei n. 9.394, de 20 de dezembro de 1996. Lei de Diretrizes e Bases da Educação Nacional, LDB 9394/1996, Brasília, 1996.

6 BRASIL. Projeto de Lei do Senado n. 504, de 2011, Brasília, 2012.

7 BREUSCH, Trevor Stanley; PAGAN, Adrian Rodney. A simple test for heteroscedasticity and random coefficient variation, Econometrica, Biometrika, Vol. 47, p. 1287-1294, 1979.

8 BUSSAB, Wilton de Oliveira; MORETTIN, Pedro Alberto. Estatística Básica. 6. ed. São Paulo: Saraiva, 2010.

9 DEMÉTRIO, Clarice Garcia Borges; ZOCCHI, Sílvio Sandoval. Modelos de Regressão. Piracicaba: ESALQ/USP, 2006.

10 FOX, John; WEISBERG, Sanford (2019). An R companion to applied regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

11 GAUSS, Carl Friedrich (1795). Method of Least Squares. Unpublished.

12 HAYDT, Regina Cazaux. Avaliação do processo ensino-aprendizagem. 6ª. ed. São Paulo: Ática, 1997.

13 INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). Indicadores Educacionais, 2010-2019. Brasília, 02 Fevereiro 2020. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/complexidade-de-gestao-da-escola>

14 LEWIS-BECK, Michael Steven. Applied Regression: an introduction. Series quantitative applications in the social sciences, SAGE University Paper, 1980.

15 R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

16 RODRIGUÊS, Sandra Cristina Antunes. Modelo de Regressão Linear e suas Aplicações. Universidade da Beira Interior. Covilhã. 2012.

17 SHAPIRO, Samuel Sanford; WILK, Martin Breadbury. An analysis of variance test for normality (complete samples), Biometrika, Vol. 52, p.591-611, 1965.

18 SOUZA, Marcone Jamilson Freitas. Ajuste de curvas pelo Método dos Quadrados Mínimos. Universidade Federal de Ouro Preto. Ouro Preto, p. 5. 2003. Disponível em: http://www.decom.ufop.br/prof/marcone/Disciplinas/MetodosNumericoseEstatisticos/QuadradosMinimos.pdf

19 THOMAS, Gustavo. Regressão não linear. Relatório sobre regressão não linear para a disciplina de Regressão e Covariância. Escola Superior de Agricultura "Luiz de Queiroz". Piracicaba, p. 16. 2016. Disponível em: https://edisciplinas.usp.br/pluginfile.php/2340838/modresource/content/0/GustavoRelatorio.pdf

20 ZEILEIS, Achim; HOTHORN, Torsten. Diagnostic Checking in Regression Relationships. R News 2(3), p. 7-10, 2002.

# Anexos

**Constituição Federal de 1988**

**Art. 6°** São direitos sociais a educação, a saúde, a alimentação, o trabalho, a moradia, o transporte, o lazer, a segurança, a previdência social, a proteção à maternidade e à infância, a assistência aos desamparados, na forma desta Constituição.

**Art. 25.** Será objetivo permanente das autoridades responsáveis alcançar relação adequada entre o número de alunos e o professor, a carga horária e as condições materiais do estabelecimento.

**Parágrafo único**. Cabe ao respectivo sistema de ensino, à vista das condições disponíveis e das características regionais e locais, estabelecer parâmetro para atendimento do disposto neste artigo.

**Art. 206.** O ensino será ministrado com base nos seguintes princípios:

i. igualdade de condições para o acesso e permanência na escola;

ii. liberdade de aprender, ensinar, pesquisar e divulgar o pensamento, a arte e o saber;

iii. pluralismo de idéias e de concepções pedagógicas, e coexistência de instituições públicas e privadas de ensino;

iv. gratuidade do ensino público em estabelecimentos oficiais;

v. valorização dos profissionais da educação escolar, garantidos, na forma da lei, planos de carreira, com ingresso exclusivamente por concurso público de provas e títulos, aos das redes públicas; (Redação dada pela Emenda Constitucional nº 53, de 2006)

vi. gestão democrática do ensino público, na forma da lei;

vii. garantia de padrão de qualidade;

viii. piso salarial profissional nacional para os profissionais da educação escolar pública, nos termos de lei federal. (Incluído pela Emenda Constitucional nº 53, de 2006)

ix. garantia do direito à educação e à aprendizagem ao longo da vida. (Incluído pela Emenda Constitucional nº 108, de 2020)

**Parágrafo único**. A lei disporá sobre as categorias de trabalhadores considerados profissionais da educação básica e sobre a fixação de prazo para a elaboração ou adequação de seus planos de carreira, no âmbito da União, dos Estados, do Distrito Federal e dos Municípios.

**Projeto de Lei do Senado n. 504, de 2011**

Altera o parágrafo único do art. 25 da Lei n. 9.394, de 20 de dezembro de 1996 (Lei de Diretrizes e Bases da Educação), para estabelecer o número máximo de alunos por turma na pré-escola e no ensino fundamental e médio.

Em 2012, a Comissão de Educação, Cultura e Esporte do Senado aprovou o projeto de lei que alteraria a redação do artigo 25 da Lei de Diretrizes e Bases da Educação Nacional passando a ter a seguinte redação:

**Art. 25.** Será objetivo permanente das autoridades responsáveis alcançar relação adequada entre o número de alunos e o professor, a carga horária e as condições materiais do estabelecimento.

**Parágrafo único**. Cabe ao respectivo sistema de ensino, à vista das condições disponíveis e das características regionais e locais, estabelecer parâmetros para atendimento do disposto no caput deste artigo, assegurado que o número máximo de alunos por turma não exceda a:

i. vinte e cinco, na pré-escola e nos dois anos iniciais do ensino fundamental;

ii. trinta e cinco nos anos subsequentes do ensino fundamental e no ensino médio.