ISSN: 2317-0840

Aplicação da análise de regressão linear no rendimento escolar do Ensino Médio público do Brasil

Leandro R. Morais[1], Marina R. Maestre[†2]

[1] *Universidade Federal da Grande Dourados (UFGD).*
[2] *Universidade Estadual de Mato Grosso do Sul (UEMS).*

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 statistical R 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.*

Keywords: *Educational Indicators; Regression Analysis; R Software.*

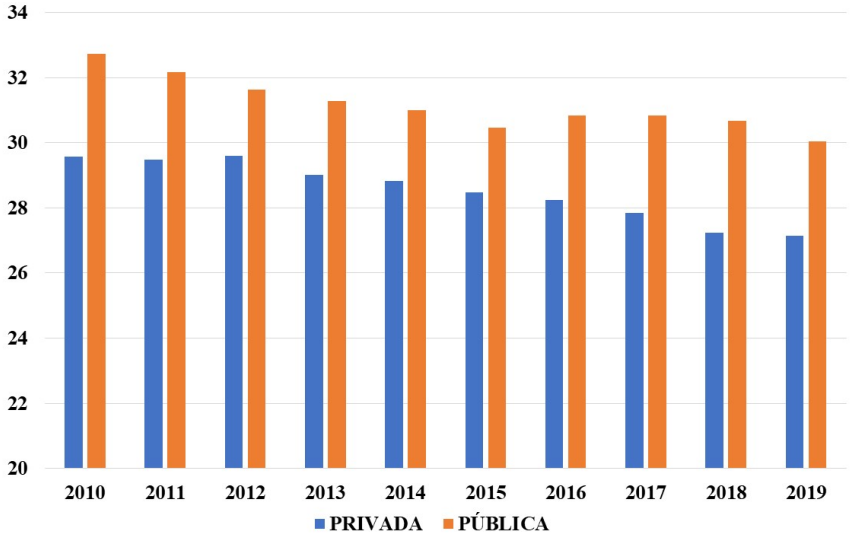
Introdução

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.

[†]Autora correspondente: marina.maestre.estatistica@gmail.com.

Porém, 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).

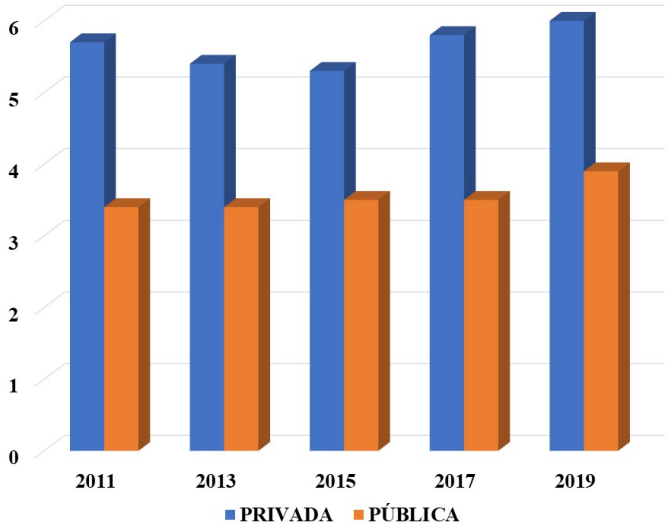
Figura 1: Média de alunos por turma no Ensino Médio do Brasil



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

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.

Assim, com base nas Figuras 1 e 2 é possível destacar que o ensino público em relação ao ensino privado deixa 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, sem necessidade, perdendo tempo 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, deseja-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?

Referencial teórico

Teoria 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 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).

Regressão Linear Simples

Uma regressão linear simples é definida como uma relação entre a variável dependente (Y) e uma variável independente (X). Segundo Bussab e Morettin (2010), o modelo de regressão linear pode ser escrito como:

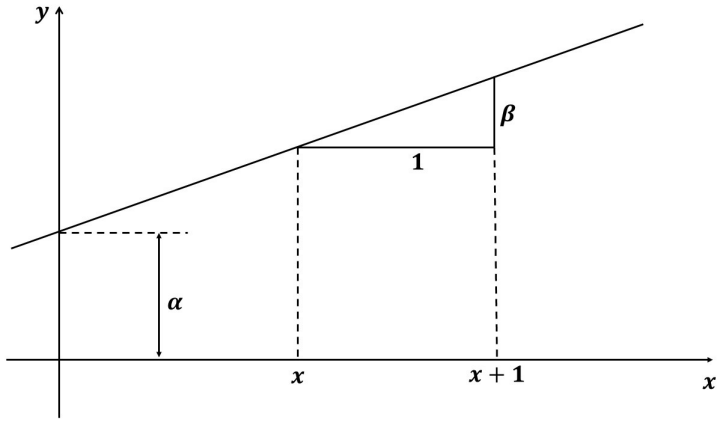
$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, ..., n \tag{1}$$

Onde:

- y_i representa o valor da variável dependente, Y;
- x_i representa o valor da variável independente, X;
- α é o intercepto e representa o ponto onde a reta corta o eixo das ordenadas;
- β é o coeficiente angular que representa o quanto varia a média de Y para um aumento de uma unidade da variável X;
- ϵ_i são variáveis aleatórias que correspondem ao erro.

A Figura 3 representa a interpretação geométrica dos parâmetros α e β.

Figura 3: Interpretação geométrica dos parâmetros α e β



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

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, nós temos o seguinte modelo teórico:

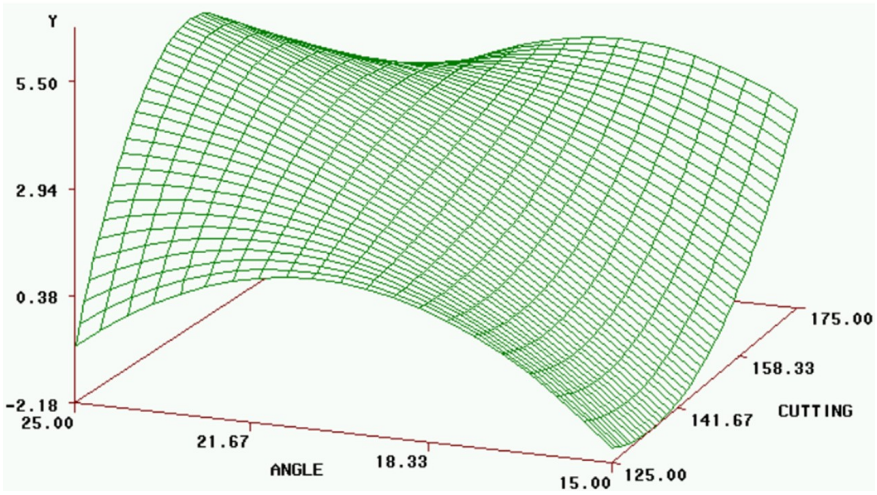
$$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + ... + \pi x_{ip} + \epsilon_i, i = 1, ..., n \tag{2}$$

Onde:

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

A equação (2) descreve um hiperplano p-dimensional referente às variáveis explicativas como mostra a Figura 4.

Figura 4: Hiperplano p-dimensional referente às variáveis independentes



Fonte: Adaptado de Rodriguês (2012, p.24).

Pressupostos de uma regressão linear

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, verificada pelo teste de correlação linear de Pearson;

ii O termo de erro segue uma distribuição normal, analisado pelo teste Shapiro-Wilk (1965);

iii Existência de *outliers* ou pontos de alavancagem;

iv Ausência de autocorrelação, ou seja, os termos de erros são independentes entre si, observado pelo teste Durbin-Watson (1950);

v Homocedasticidade, ou seja, a variância do termo de erro é constante para os diferentes valores da variável independente, analisado pelo teste Breusch-Pagan (1979)

vi As variáveis independentes não apresentam alta correlação entre si, analisado pelo coeficiente de correlação linear de Pearson. Este pressuposto é analisado apenas para regressões lineares múltiplas.

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 é feita para realizar os testes de hipóteses, desse modo, a apresentação dos resultados deve ser realizada com a variável original.

Nas seções que irá discutir sobre regressão linear simples entre alunos por turma e taxa de aprovação e regressão linear múltipla entre taxa de aprovação, alunos por turma e taxa de aprovação 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 na regressão simples e possui alta correlação com a variável "horas estudadas" na regressão múltipla.

Inicialmente, foram realizadas as transformações logarítmica e a de Box-Cox, porém nos novos modelos de regressão, a variável transformada continuou falhando no mesmo pressuposto, até que foi testada a seguinte transformação:

$$Y_i = \frac{1}{e^{X_i}} \tag{3}$$

Onde:

- X_i a variável original "*alunos por turma*", e
- Y_i é a variável transformada "$AlunosT$".

O *Software* R

As regressões lineares serão analisadas por cálculos estatísticos utilizando o *software* R. Por conta de sua capacidade gráfica e análise de dados, o R é muito utilizado para tal finalidade (R CORE TEAM, 2020).

Metodologia

Os dados analisados nesse trabalho são: média de alunos por turma, média de horas-aulas diária e taxa de rendimento. Todos estes foram obtidos no site do INEP. Lembrando que as informações são referentes ao Ensino Médio público do Brasil e foram organizadas conforme a Tabela 1.

Tabela 1: Ensino Médio público do Brasil

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

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

Como o objetivo aqui é verificar se a quantidade de horas estudadas e/ou a média de alunos por turma influenciam na taxa de aprovação, serão construídos e analisados modelos de regressão com tais dados descritos da seguinte forma:

- Regressão linear simples entre horas estudadas e taxa de aprovação.

$$ta_i = \alpha + \beta he_i + \epsilon_i \tag{4}$$

- Regressão linear simples entre alunos por turma e taxa de aprovação.

$$ta_i = \alpha + \beta al_i + \epsilon_i \tag{5}$$

- Regressão linear múltipla entre horas estudadas, alunos por turma e taxa de aprovação.

$$ta_i = \alpha + \beta he_i + \gamma al_i + \epsilon_i \tag{6}$$

Onde:

- ta = Taxa de aprovação;
- he = Horas estudadas;
- al = Alunos por turma;

Com $i = 1, 2, ..., 10$, representando os dados de 2010 à 2019.

Resultados e discussão

Os modelos são analisados por meio de testes de hipótese, em todos os casos o nível de significância considerado é de 5%.

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. Para isso, é necessário verificar os pressupostos de uma regressão linear simples. Neste modelo, a variável dependente é taxa de aprovação e a variável independente é horas estudadas.

Então, no R, será construído um modelo de regressão linear com essas variáveis. Esse modelo será chamado de `mod` e para criá-lo é utilizada a função "`lm`" do inglês *linear model*. Logo:

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

Inicialmente, é verificado se há alguma relação linear entre as variáveis por meio do teste de coeficiente de correlação linear de Pearson, utilizando o comando: `cor.test(dados$Horas,dados$Aprovação)`.

Do teste, a correlação entre as variáveis é de $0,9795848$. Segundo Bussab e Morettin (2010) essa é uma correlação forte, e além disso o p-valor é igual a $7,415 \times 10^{-7}$ indicando a rejeição da hipótese de que o coeficiente de correlação seja igual a zero. Assim, é possível afirmar que a distribuição gráfica dessas variáveis possui uma relação linear.

Pelo comando `shapiro.test` é possível aplicar o teste de Shapiro-Wilk ao modelo. Obtendo o valor $0,95912$ como a estatística do teste, e o p-valor igual a $0,7758$. Neste caso, a hipótese nula é considerada, podendo afirmar que os resíduos do modelo em questão seguem uma distribuição normal.

No R a existência de *outliers* é verificada pela função `summary(rstandard(modelo))`. Caso exista algum ponto de alavancagem este terá valor superior a 3, ou inferior a -3. No modelo, os valores de mínimo e máximo são respectivamente $-1,61067$ e $1,85389$, indicando que não há a existência de *outliers* nos resíduos do modelo.

Agora, a independência dos resíduos é analisado por meio do teste de Durbin-Watson através do comando: `durbinWatsonTest(model, ...)`. Tem-se um p-valor para o teste de $0,78$, não rejeitando a hipótese nula. Assim, é possível afirmar que no modelo analisado os resíduos são independentes.

O último pressuposto do modelo a ser observado é a homocedasticidade dos resíduos. No R, o teste de Breusch-Pagan é realizado por meio do comando: `bptest(formula, varformula = NULL, data = list())`. O p-valor obtido neste teste é de $0,1868$, então é considerada a hipótese nula. Logo, pode-se afirmar que há homoscedasticidade entre os resíduos das variáveis.

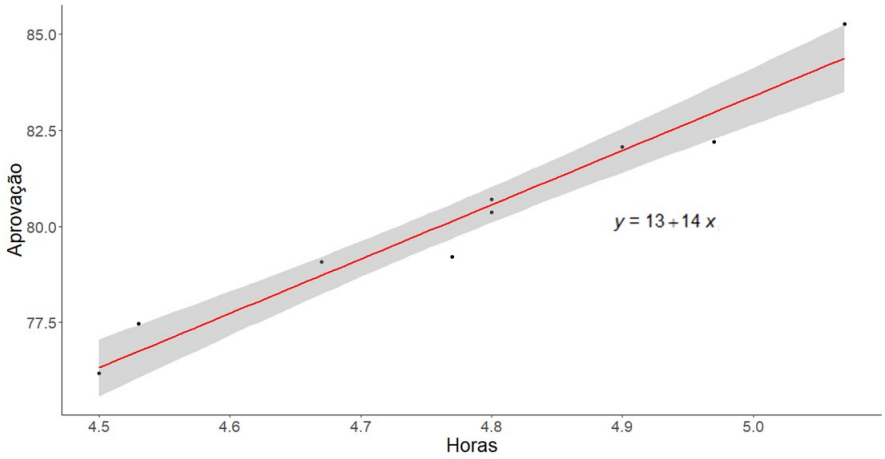
Os testes de hipótese para os resíduos do modelo confirmam que todos os pressupostos são atendidos. Então resta construir o gráfico de dispersão das variáveis e a equação da reta que melhor lhes representa. Isso é feito utilizando a função:

```
> ggplot(data = dados, mapping = aes(x = 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))
```

Obtendo o seguinte gráfico:

Na Figura 5, a reta de regressão linear tem a equação $y = 13 + 14x$. A faixa presente na figura é o intervalo de confiança dos pontos que se distanciam da reta com 95% de certeza.

Figura 5: Taxa de aprovação × Horas estudadas



Fonte: Autores.

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. Nesse modelo, a variável dependente é taxa de aprovação e a variável independente é média de alunos por turma. Tal modelo será construído no R e nomeado de `mod`, ficando:

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

A análise da linearidade entre as variáveis, normalidade, *outliers* e independência dos resíduos encontra-se resumida na Tabela 2.

Tabela 2: Relação linear, normalidade, *outliers* e independência dos resíduos

Correlação linear		teste Shapiro - Wilk	função summary		teste Durbin - Watson
coeficiente	p-valor	p-valor	mínimo	máximo	p-valor
$-0,921870$	$0,0001482$	$0,09674$	$-1,1941$	$1,83679$	$0,006$

Fonte: Autores.

Da Tabela 2, nota-se que o coeficiente de correlação linear entre as variáveis do modelo é de aproximadamente $-0,92$. Como essa correlação é forte e além do p-valor desse teste ser igual a $0,0001482$, indicando a rejeição da hipótese de que o coeficiente de correlação possa ser igual a zero. Desse modo é possível afirmar que as variáveis dependente e independente seguem uma relação linear.

Nesta mesma tabela, é possível observar que o p-valor obtido para o teste de Shapiro-Wilk foi de $0,09674$ indicando que a hipótese nula é considerada. Assim, os resíduos do modelo seguem uma distribuição normal. Agora, com relação a existência de *outliers*, os valores mínimo e máximo, respectivamente, são: $-1,1941$ e $1,83679$ indicando que não há a existência de *outliers* nos resíduos do modelo.

Mas ao analisar a independência dos resíduos, nota-se que o p-valor é de $0,006$ então a hipótese nula é rejeitada. Pois, neste caso, o p-valor obtido no teste é menor que o seu nível de significância. Logo os resíduos não são independentes.

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 transformação elencada no referencial teórico.

Para esse novo modelo que será analisado, a variável dependente é *"taxa de aprovação"* e a variável independente é *"AlunosT"* (variável transformada), ficando da seguinte forma:

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

A Tabela 3, apresenta os resultados dos testes realizados sobre o `mod2`.

Tabela 3: Relação linear, normalidade, *outliers*, independência e homocedasticidade

Correlação linear		teste Shapiro - Wilk	função summary		teste Durbin - Watson	teste Breusch - Pagan
coeficiente	p-valor	p-valor	mínimo	máximo	p-valor	p-valor
$0,9419861$	$4,619 \times 10^{-5}$	$0,2574$	$-1,17178$	$1,95611$	$0,124$	$0,5841$

Fonte: Autores.

O coeficiente de correlação entre as variáveis é de aproximadamente $0,94$, como essa correlação é forte, e além do p-valor ser $4,619 \times 10^{-5}$, indicando a rejeição de que o coeficiente de correlação possa ser igual a zero. Assim, é possível afirmar que as variáveis do modelo seguem uma relação linear.

O p-valor obtido no teste de Shapiro-Wilk é de $0,2574$ indicando que os resíduos do modelo seguem uma distribuição normal. Por meio da função `summary` os valores de máximo e mínimo, respectivamente, são $-1,17178$ e $1,95611$, assim não há a existência de *outliers* nos resíduos do modelo.

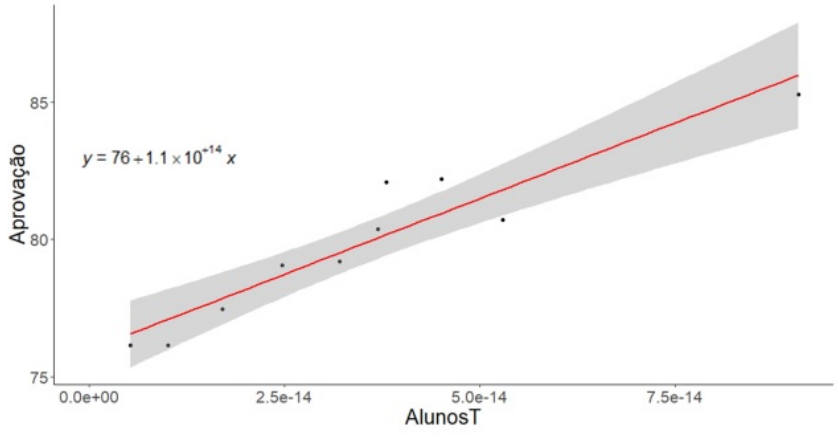
Do teste Durbin-Watson é obtido um p-valor de $0,124$, desse modo os resíduos do modelo são independentes. E por meio do teste de Breusch-Pagan, pode-se afirmar que há homoscedasticidade entre os resíduos do modelo, pois o p-valor obtido nesse teste foi de $0,5841$.

Os testes de hipótese para os resíduos do modelo confirmam que todos os pressupostos são atendidos. Então resta construir o 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()
```

Obtendo o seguinte gráfico:

Figura 6: Taxa de aprovação × AlunosT



Fonte: Autores.

Na Figura 6, a reta de regressão linear tem a equação $y = 76 + 1,1 \times 10^{-14}x$. A faixa presente na figura é o intervalo de confiança dos pontos que se distanciam da reta de regressão com 95% de certeza.

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. Nesta seção, será trabalhado com a variável dependente taxa de aprovação e com as variáveis independentes horas estudadas e média de alunos por turma. O modelo a ser analisado será chamado de `mod` e no R basta entrar com o comando:

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

Na Tabela 4, pode ser observado os resultados dos testes realizados sobre `mod`.

Tabela 4: Normalidade, *outliers*, independência, homocedasticidade e correlação linear

teste Shapiro - Wilk	função summary		teste Durbin - Watson	teste Breusch - Pagan	Correlação linear	
p-valor	mínimo	máximo	p-valor	p-valor	coeficiente	p-valor
0,9598	-1,71399	1,86577	0,768	0,368	-0,9111152	0,0002439

Fonte: Autores.

No teste de Shapiro-Wilk, o p-valor obtido é $0,9598$ indicando que os resíduos do modelo seguem uma distribuição normal. Com relação a existência de *outliers* no modelo, verificada pela função `summary`, é obtido o valor mínimo de $-1,71399$ e máximo de $1,86577$, indicando que não há a existência de *outliers* nos resíduos do modelo.

Os resíduos do modelo são independentes, pois o p-valor do teste de Durbin-Watson é de $0,768$. E por meio do teste de Breusch-Pagan pode-se afirmar que os resíduos do modelo são homocedásticos pois o p-valor para esse teste foi de $0,368$.

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 $-0,9$. Mas como mostra na Tabela 4, o coeficiente de correlação linear de Pearson é aproximadamente $-0,911$, então há multicolinearidade entre as variáveis independentes. Assim, o modelo em questão não pode ser interpretado, pois as variáveis do modelo 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"* ao invés de *"alunos por turma"*, como já realizado no item anterior. Esse novo modelo será chamado de `mod2`, ficando:

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

Na Tabela 5, pode ser observado os resultados dos testes realizados sobre `mod2`.

Tabela 5: Normalidade, *outliers*, independência, homocedasticidade e correlação linear

teste Shapiro - Wilk	função summary		teste Durbin - Watson	teste Breusch - Pagan	Correlação linear	
p-valor	mínimo	máximo	p-valor	p-valor	coeficiente	p-valor
0,805	-1,677	1,59812	0,844	0,5852	0,8992542	0,0003985

Fonte: Autores.

Observando a Tabela 5 é possível afirmar que todos os pressupostos de uma regressão linear múltipla são aceitos, pois o p-valor para o teste de Shapiro-Wilk é de $0,805$ indicando que

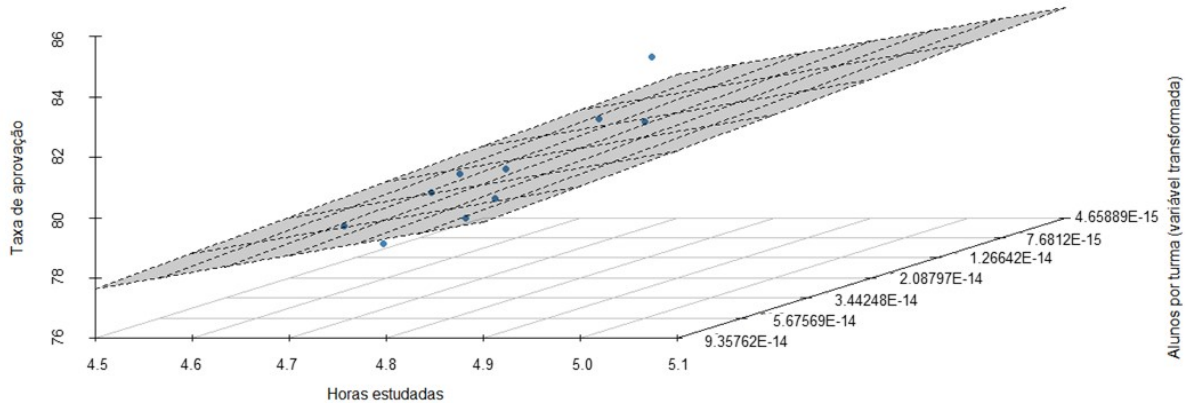
os resíduos do modelo seguem uma distribuição normal. Os valores de máximo e mínimo da função `summary` são, respectivamente, $-1,677$ e $1,59812$, indicando que não existem *outliers* nos resíduos do modelo. Além disso, os testes de Durbin-Watson e Breusch-Pagan apresentam p-valores, respectivamente, $0,844$ e $0,5852$ afirmando que os resíduos do modelo são independentes e homocedásticos.

O coeficiente de correlação entre as variáveis independentes é $0,8992542$ com um p-valor de $0,0003985$, indicando a rejeição de que o coeficiente possa ser igual a zero. Assim, é possível afirmar que as variáveis independentes do novo modelo seguem uma relação linear.

Após verificar que todos os pressupostos para o modelo são atendidos, resta construir o gráfico de dispersão das variáveis.

```
> 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 7: Taxa de aprovação × Horas estudadas e Alunos por turma (variável transformada).



Fonte: Autores.

O plano que aparece na Figura 7 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.

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.

Agradecimentos

Agradecimento à Capes pelo apoio financeiro.

Referência

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

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

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

DURBIN, James; WATSON, Geoffrey Stuart. Test for Serial Correlation in Least Squares Regression, *Biometrika*, 37, p. 409-428, 1950.

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>

LEWIS-BECK, Michael Steven. *Applied Regression: an introduction. Series Quantitative Applications in the Social Sciences*. [S.l.]: SAGE University Paper, 1980.

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

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

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