UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 5: DIAGNÓSTICO DO MODELO DE REGRESSÃO LINEAR SIMPLES E ESTIMAÇÃO PONTUAL DO MODELO DE REGRESSÃO LINEAR MÚLTIPLO

Salvador

2023

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 5: DIAGNÓSTICO DO MODELO DE REGRESSÃO LINEAR SIMPLES E ESTIMAÇÃO PONTUAL DO MODELO DE REGRESSÃO LINEAR MÚLTIPLO

Atividades de laboratório apresentadas ao Instituto de Matemática e Estatística da Universidade Federal da Bahia como parte das exigências da disciplina Análise de Regressão ministrada pela professora Dra. Edleide de Brito.

Salvador

2023

SUMÁRIO

1 Introdução

Neste trabalho, serão realizadas duas atividades: a primeira consiste no ajuste de modelos de regressão linear simples utilizando o conjunto de dados "trees". Esse conjunto de dados contém informações de 31 cerejeiras da espécie Black Cherry, localizadas na Floresta Nacional de Allegheny.

Primeiramente, será ajustado um modelo de regressão linear simples para o volume da madeira em função da altura das árvores. Em seguida, será realizado uma avaliação dos resíduos *Jackknife* para diagnosticar possíveis violações do modelo.

Além disso, serão consideradas diferentes transformações para a variável resposta (volume da madeira). Para cada uma dessas transformações, também será ajustado um modelo de regressão linear simples e avaliado os resíduos de *Jackknife*. Em seguida, será verificado qual transformação seria mais apropriada dentro da família proposta por Box Cox.

Já a segunda atividade, consistirá na análise de regressão linear múltipla utilizando o conjunto de dados de 403 afro-americanos residentes no Estado da Virginia, entrevistados em um estudo sobre a prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares.

Primeiramente, a base de dados será estudada em busca de possíveis inconsistências e dados ausentes. Além disso, será modificada a escala de algumas variáveis para facilitar a interpretação dos resultados.

Será proposta a criação de novas variáveis com base nas variáveis disponíveis. Em seguida, será ajustado um modelo de regressão linear múltipla em que a variável reposta será o Índice de Massa Corporal (IMC). Os parâmetros do modelo ajustado serão interpretados e será verificado a significância de cada um deles.

Assim, através do modelo de regressão linear múltiplo será possível compreender a relação entre as características dos afro-americanos residentes no Estado da Virginia e o IMC.

2 Pesquisa sobre Cerejeiras Black Cherry

2.1 Análise Descritiva

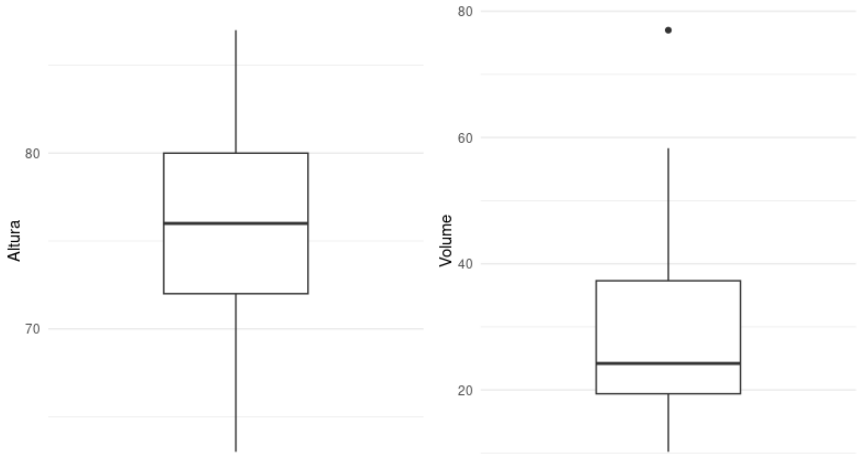
Observando a Tabela 2.1, a altura apresenta média e mediana iguais e a moda é relativamente próxima aos valores da mediana e da média. Logo, isso poderia indicar que a distribuição é simétrica. A moda da variável volume está bem próxima do valor mínimo e bem distante da mediana e média, indicando uma clara assimetria na variável. A altura apresenta um baixíssimo coeficiente de variação quando comparado com o volume, indicando que altura tem uma menor variabilidade do que o volume.

Tabela 2.1: Sumarização das variáveis altura e volume das cerejeiras da Floresta Nacional de Allegheny

	Altura	Volume
Mínimo	63,0	10,2
1º quartil	72,0	19,4
Mediana	76,0	24,2
Média	76,0	30,2
Moda	80,0	10,3
3º quartil	80,0	37,3
Máximo	87,0	77,0
CV	0,08	0,54

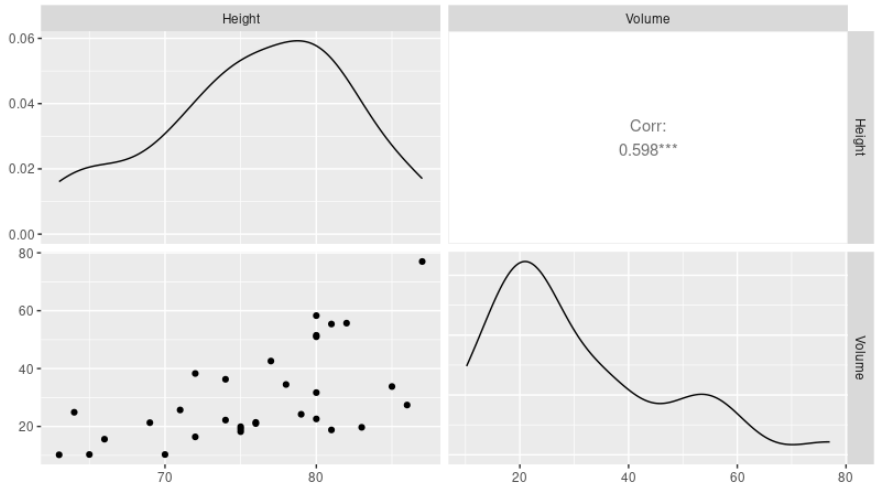
Com os *boxplots* da Figura 2.1, a simetria da variável altura é ainda mais evidente. Da mesma forma, é possível observar que a variável volume tem uma variabilidade consideravelmente maior entre a mediana e o terceiro quartil, indicando uma forte assimetria à direita.

Figura 2.1: *Boxplots* da altura e volume das cerejeiras da Floresta Nacional de Allegheny



Ao analisar o gráfico de dispersão entre as variáveis volume e altura, ilustrado na Figura 2.2, é possível observar que parece haver uma relação linear positiva relativamente fraca entre o volume e a altura, apesar dessa relação se tornar menos notável para valores de altura acima de 80. O coeficiente de correlação de *Pearson* entre essas duas variáveis foi igual a 0,598, indicando que a altura e o volume apresentam correlação moderada positiva entre si.

Figura 2.2: Correlação e gráfico de dispersão e densidade da altura e volume das cerejeiras da Floresta Nacional de Allegheny



2.2 Modelo de regressão linear simples

A Tabela 2.2, apresenta uma sumarização da regressão linear em que a altura das cerejeiras é utilizada para predizer o seu volume. A estimativa do intercepto, nesse caso, não possui interpretação. A estimativa do coeficiente angular aponta que para cada aumento de uma unidade na altura das arvores, espera-se um aumento médio de 1,54 unidades no seu volume.

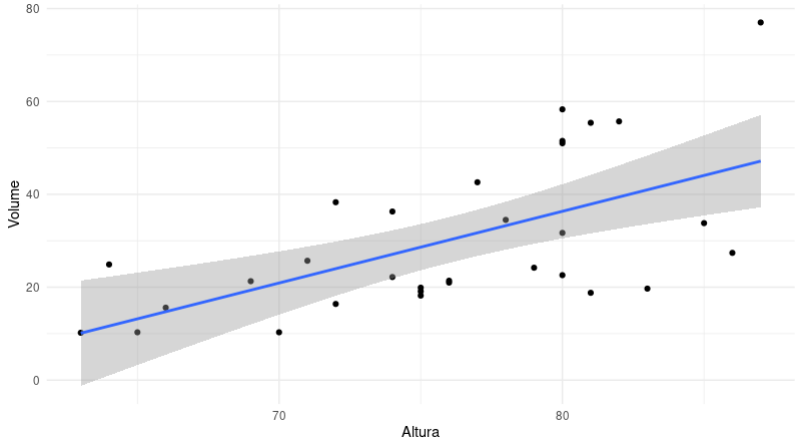
Tabela 2.2: Sumarização da regressão linear em que a altura das cerejeiras é utilizada para predizer o seu volume

Coeficientes	Estimativa	EP	Est. T	Valor-p	R^2
β_0	-87,12	29,27	-2,98	0,006	0,36
β_1	1,54	0,38	4,02	$< 0,001$	

O modelo de regressão linear sugere que a altura das cerejeiras pode ser um fator significativo na predição do seu volume. Uma vez que, o valor-p para o coeficiente angular indica o que esse coeficiente é significativo a um nível de 5%. Mas, apesar do coeficiente ser significativo, o coeficiente de determinação evidencia que apenas 36% da variação no volume das cerejeiras pode ser explicada pela variação na altura.

Observando o gráfico de dispersão da Figura 2.3, é possível notar que o modelo não explica tão bem o volume das árvores.

Figura 2.3: Gráfico de dispersão e curva de regressão linear ajustada aos dados das cerejeiras da Floresta Nacional de Allegheny

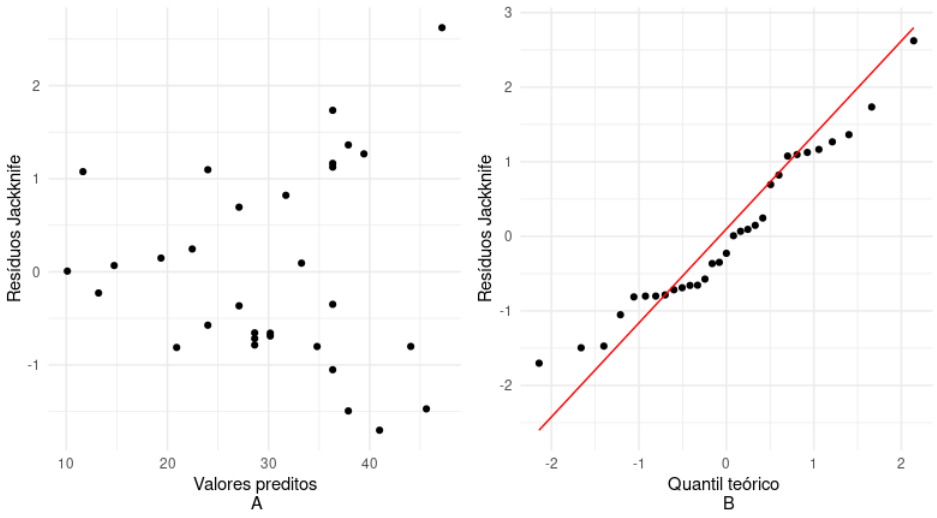


Até o valor de altura igual a 80, a curva de regressão linear acompanha a tendência dos pontos no gráfico. Acima desse valor, parece que a distância entre as observações e a curva de regressão linear é maior, não seguindo tão bem a tendência de pontos no gráfico.

Análise dos resíduos

O gráfico dos resíduos *Jackknife* contra valores preditos na Figura 2.4 (A), evidencia uma maior variabilidade nos resíduos entre os maiores valores preditos, ou seja, a variação dos resíduos está aumentando a medida que os valores preditos crescem. Realizando o teste de *Goldfeld-Quandt*, foi obtido um p-valor de aproximadamente zero, rejeitando a hipótese nula, a um nível de significância de 5%, de que a variância dos resíduos é constante. Portanto, os resíduos violam a suposição de homocedasticidade.

Figura 2.4: Gráfico dos resíduos Jackknife versus valores ajustados e quantil teórico da regressão linear dos dados das cerejeiras da Floresta Nacional de Allegheny



O *QQ-plot* dos resíduos *Jackknife*, Figura 2.4 (B), mostra que os resíduos não estão tão próximos aos quantis teóricos e a normalidade dos resíduos pode estar sendo prejudicada pelas caudas da distribuição. Mas, o teste de *Shapiro-Wilk* indicou um p-valor de 0,16, não rejeitando a hipótese nula, a um nível de significância de 5%, de que os resíduos vem de uma distribuição normal. Portanto, é possível afirmar que a suposição de normalidade dos resíduos não foi violada pelo modelo.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente zero, apontando há autocorrelação serial nos resíduos a um nível de 5%. Ou seja, há violação do pressuposto de independência.

2.3 Modelo de regressão linear simples transformado

A Tabela 2.3, apresenta uma sumarização das regressões lineares em que a altura das cerejeiras é utilizada para predizer o seu volume transformado. As transformações consideradas são $\sqrt{Y}$, $\log(Y)$ e Y^2 e a interpretação para cada um dos modelos é:

- Modelo de regressão com a transformação $\sqrt{Y}$: este modelo indica que para cada unidade de aumento na altura das árvores, a raiz quadrada do volume aumenta em média 0,14 unidades. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 39% da variação no volume pode ser explicada pela variação na altura;

- Modelo de regressão com a transformação $\log(Y)$: este modelo indica que para cada unidade de aumento na altura das árvores, o volume estimado aumenta em $1 - e^{0,5} \times 100\% = 5{,}14\%$. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 42% da variação no volume pode ser explicada pela variação na altura;

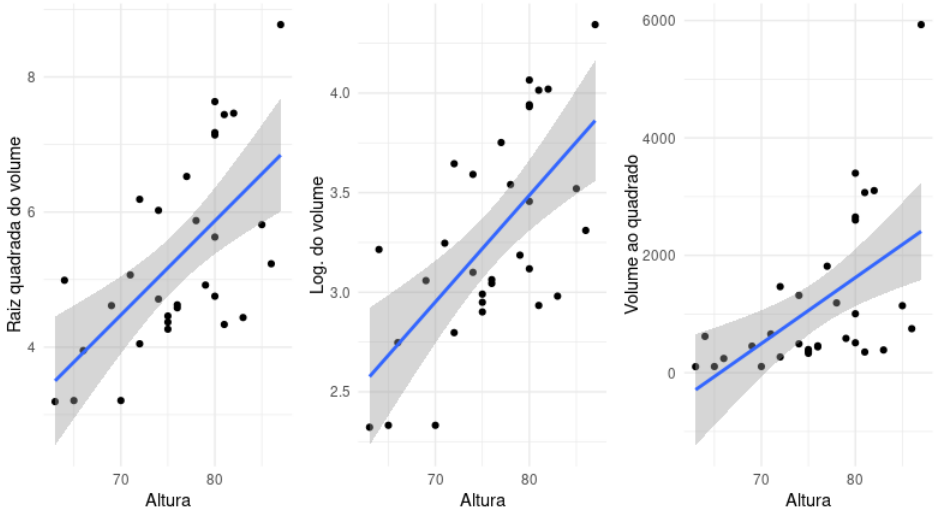
- Modelo de regressão com a transformação Y^2: este modelo indica que para cada unidade de aumento na altura das árvores, o quadrado do volume aumenta em média 112,41 unidades. O valor-p sugere que este coeficiente é significativo e o R^2 indica que cerca de 30% da variação no volume pode ser explicada pela variação na altura.

Tabela 2.3: Sumarização da regressão linear em que a altura das cerejeiras é utilizada para predizer o seu volume para cada modelo transformado

Transf.	Coef.	Estimativa	EP	Est. T	Valor-p	R^2
$\sqrt{Y}$	β_0	-5,28	2,46	-2,14	0,04	0,39
	β_1	0,14	0,03	4,31	$< 0{,}001$	
$\log(Y)$	β_0	-0,80	0,89	-0,90	0,38	0,42
	β_1	0,05	0,01	4,59	$< 0{,}001$	
Y^2	β_0	-7371,17	2440,94	-3,02	0,005	0,30
	β_1	112,41	32,01	3,51	0,002	

O modelo com a transformação logarítmica é o modelo que mais explica o volume das árvores, uma vez que ele possui um coeficiente de determinação mais alto. Mas, os três coeficientes de determinação são próximos e graficamente (Figura 2.5) os modelos com as transformações $\sqrt{Y}$ e $\log(Y)$ parecem explicar o volume das cerejeiras da mesma forma. Enquanto o modelo com a transformação Y^2 apresenta um valor discrepante, que afeta negativamente a precisão do modelo.

Figura 2.5: Gráfico de dispersão e curva de regressão linear ajustada aos dados das cerejeiras da Floresta Nacional de Allegheny para cada modelo transformado

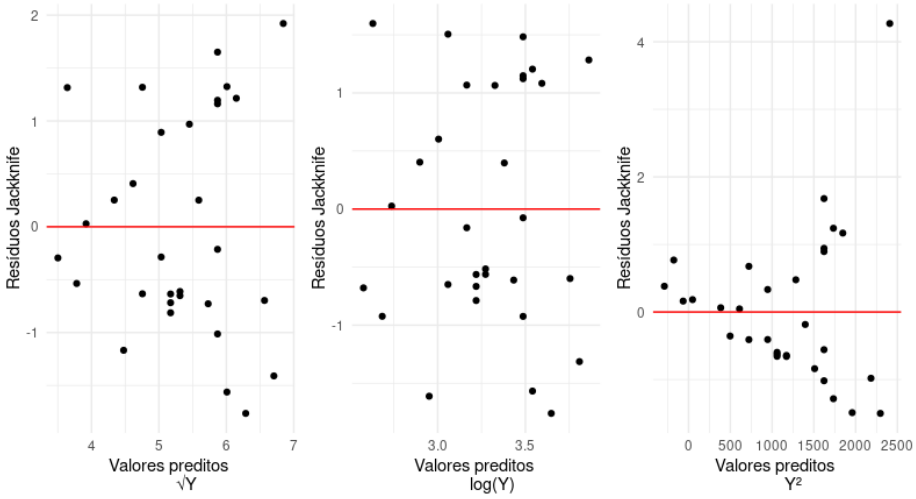


Além disso, a curva de regressão linear acompanha a tendência dos pontos no gráfico, mas grande parte das observações estão fora do intervalo de confiança em todos os modelos.

Análise dos resíduos

Os gráficos dos resíduos em relação aos valores preditos, na Figura 2.6, revelam uma maior variabilidade nos resíduos nos modelos que utilizam as transformações $\sqrt{Y}$ e Y^2 em valores preditos mais altos. Isso significa que a dispersão dos resíduos aumenta à medida que os valores preditos aumentam. Por outro lado, no modelo com a transformação logarítmica, os resíduos apresentam uma variabilidade constante em todos os valores preditos.

Figura 2.6: Gráfico dos resíduos *Jackknife* versus valores ajustados da regressão linear dos dados das cerejeiras da Floresta Nacional de Allegheny para cada modelo transformado



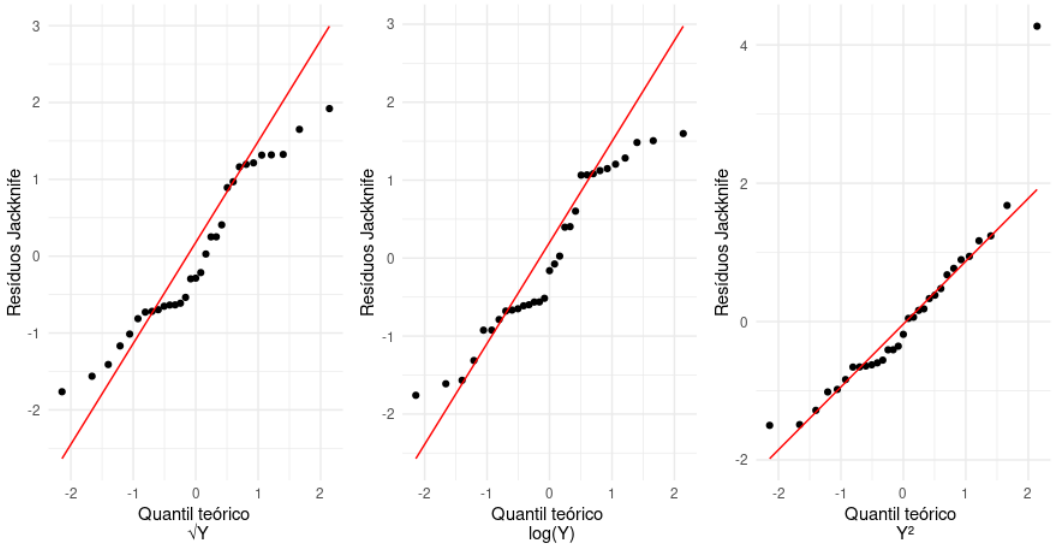
Os resultados dos testes de *Goldfeld-Quandt*, apresentados na Tabela 2.4, confirmam a análise gráfica realizada. Os valores-p dos testes indicam uma rejeição da hipótese de homocedasticidade para os resíduos dos modelos que utilizam as transformações $\sqrt{Y}$ e Y^2, a um nível de significância de 5%. Isso significa que os resíduos desses modelos não obedecem à suposição de homocedasticidade. Por outro lado, os resíduos do modelo com a transformação logarítmica não violam essa suposição.

Tabela 2.4: Valores-p dos testes utilizados para verificar os pressupostos dos modelos em cada transformação

Teste	$\sqrt{Y}$	$\log(Y)$	Y^2
Shapiro-Wilk	0,05	0,02	0,04
Goldfeld-Quandt	0,001	0,06	<0,001
Durbin-Watson	<0,001	<0,001	<0,001

O gráfico *QQ-plot* dos resíduos *Jackknife*, apresentado na Figura 2.7, revela que os resíduos não estão muito próximos dos quantis teóricos, indicando que a normalidade dos resíduos pode ser afetada pelas caudas da distribuição nos modelos com as transformações de raiz quadrada e logarítmica. Por outro lado, os resíduos do modelo com a transformação ao quadrado estão próximos aos quantis teóricos, com exceção de uma única observação discrepante.

Figura 2.7: Gráfico dos resíduos *Jackknife* versus quantil teórico da regressão linear dos dados das cerejeiras da Floresta Nacional de Allegheny para cada modelo transformado



O teste de *Shapiro-Wilk*, realizado a um nível de significância de 5%, rejeitou a hipótese nula de que os resíduos seguem uma distribuição normal em todos os modelos. Mas, para o modelo com a transformação quadrada, o resultado do teste certamente está sendo afetado por uma única observação. Portanto, podemos afirmar que a suposição de normalidade

dos erros foi violada apenas para os modelos com as transformações raiz quadrada e logarítmica.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente zero para todos os modelos, apontando que há autocorrelação serial nos resíduos a um nível de 5%. Ou seja, há violação do pressuposto de independência.

2.4 Modelo de regressão linear simples transformado (Box-Cox)

Na família proposta por Box-Cox, a transformação mais adequada é aquela que utiliza um parâmetro de transformação (lambda) próximo a -0,2, conforme mostrado na Figura 8. No entanto, o intervalo de confiança indica que a transformação logarítmica também é uma opção viável, uma vez que o intervalo de confiança inclui o valor zero.

Figura 2.8: Log-verossimilhanças para parâmetro da transformação de potência Box-Cox

A Tabela 2.5, apresenta uma sumarização da regressão linear em que a altura das cerejeiras é utilizada para predizer o seu volume transformado. A estimativa do coeficiente angular, que é significativo a um nível de 5%, aponta que para cada aumento de uma unidade na altura das arvores, espera-se um aumento médio de 0,03 unidades na função *Box-Cox* do volume.

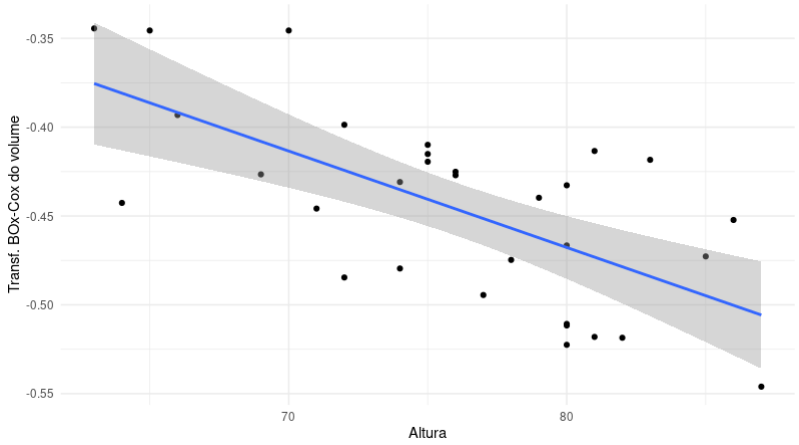
Tabela 2.5: Sumarização da regressão linear, com a transformação Box-Cox, em que a altura das cerejeiras são utilizadas para predizer o seu volume

Coeficientes	Estimativa	EP	Est. T	Valor-p	R^2
β_0	0,18	0,49	0,38	0,71	0,43
β_1	0,03	0,01	4,67	$< 0,001$	

O coeficiente de determinação indica que cerca de 42% da variabilidade observada no volume das cerejeiras pode ser explicada pela variação na altura nesse modelo.

Observando o gráfico de dispersão da Figura 2.9, é possível notar que o modelo não explica tão bem o volume das árvores. A curva de regressão linear acompanha a tendência dos pontos no gráfico, mas há poucas observações e grande parte delas estão fora do intervalo de confiança.

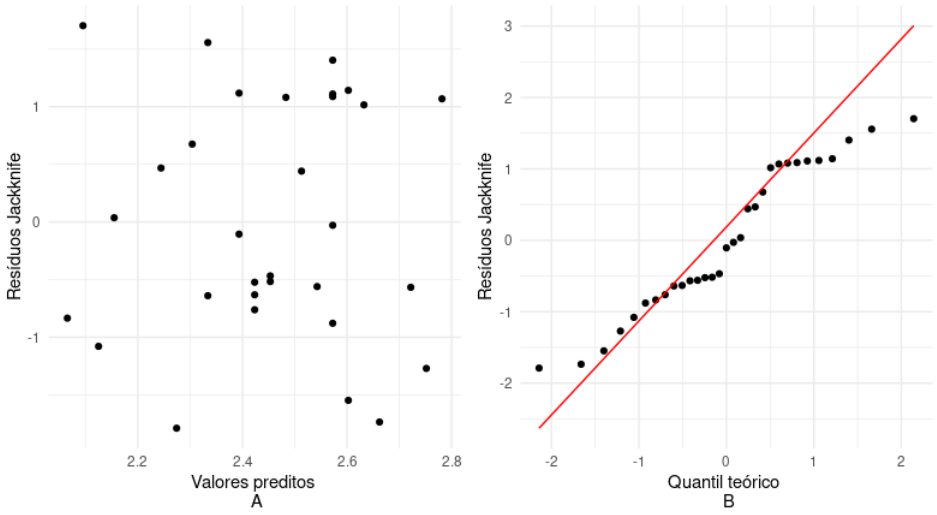
Figura 2.9: Gráfico de dispersão e curva de regressão linear, com a transformação Box-Cox, ajustada aos dados das cerejeiras da Floresta Nacional de Allegheny



Análise dos resíduos

O gráfico dos resíduos versus valores preditos, Figura 2.10, indica que dispersão dos resíduos permanece uniforme em relação aos valores preditos. O teste *Goldfeld-Quandt* corrobora a análise gráfica, apontando a não rejeição da hipótese de homocedasticidade, com um valor-p de aproximadamente 0,18. Logo, os resíduos não violam essa suposição.

Figura 2.10: Gráfico dos resíduos Jackknife versus valores ajustados e quantil teórico da regressão linear, com a transformação Box-Cox, dos dados das cerejeiras da Floresta Nacional de Allegheny



O gráfico *QQ-plot* dos resíduos, Figura 2.10 (B), aponta que os resíduos não estão tão próximos aos quantis teóricos e a normalidade dos resíduos pode estar sendo afetada pelas caudas da distribuição. O teste de *Shapiro-Wilk* indicou um p-valor de 0,04, rejeitando a hipótese nula, a um nível de 5%, de que os resíduos são normalmente distribuídos. Portanto, é possível afirmar que a suposição de normalidade dos resíduos foi violada pelo

modelo.

Já o teste de *Durbin-Watson* resultou num valor-p de aproximadamente zero, apontando há autocorrelação serial nos resíduos a um nível de 5%. Ou seja, há violação do pressuposto de independência.

3 Pesquisa sobre prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares

3.1 Manipulação dos dados

Os dados de 403 afro-americanos residentes no Estado da Virginia, apresentam as variáveis: colesterol total, glicose estabilizada, colesterol bom, razão entre colesterol total e colesterol bom, hemoglobina glicada, município de residência, idade em anos, sexo, altura, peso, pressão sanguínea sistólica e diastólica (1^{o} e 2^{o} medidas), cintura e quadril.

Nessa manipulação dos dados, algumas variáveis tiveram suas unidades convertidas: as variáveis altura, cintura e quadril foram convertidas para metros e a variável peso foi convertida para quilos.

Duas novas variáveis foram calculadas a partir das variáveis existentes. A variável IMC (Índice de Massa Corporal) foi calculada dividindo o peso em quilogramas pelo quadrado da altura em metros. A variável RCQ (Relação Cintura Quadril) foi calculada dividindo a cintura pela medida do quadril.

Por haver muitos valores ausentes, as variáveis que denotam as pressões sanguíneas sistólica e diastólica na 2^{o} medida foram retiradas dos dados. Além disso, as linhas que contêm valores ausentes foram removidas. Após a remoção, restaram 377 observações no conjunto de dados.

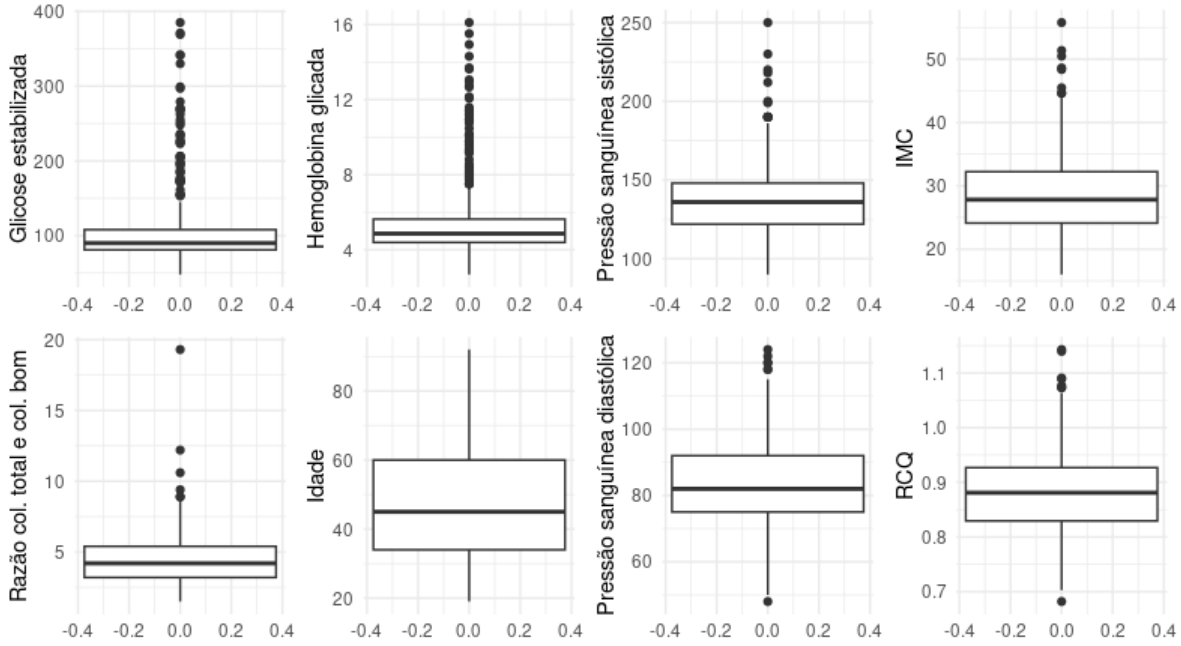
Além disso, como a razão entre o colesterol total e colesterol bom, o IMC e o RCQ são funções de variáveis que estão nos dados, essas variáveis que deram origem as funções serão desconsideradas, assim como as variáveis qualitativas. Portanto, as variáveis que serão utilizadas no modelo de regressão são: glicose estabilizada, razão entre colesterol total e colesterol bom, hemoglobina glicada, idade em anos, pressão sistólica, pressão diastólica e as novas variáveis IMC e RCQ.

3.2 Análise Descritiva

Ao observar os *boxplots* da Figura 3.1, é possível notar que há presença de muitos valores atípicos, principalmente nas variáveis glicose estabilizada e hemoglobina glicada. Esses valores atípicos não devem ser retirados da análise, pois eles fazem parte da natureza das variáveis e parecem indicar alguma informação sobre a presença de diabetes nesses indivíduos, já que ambas variáveis avaliam a glicose no sangue.

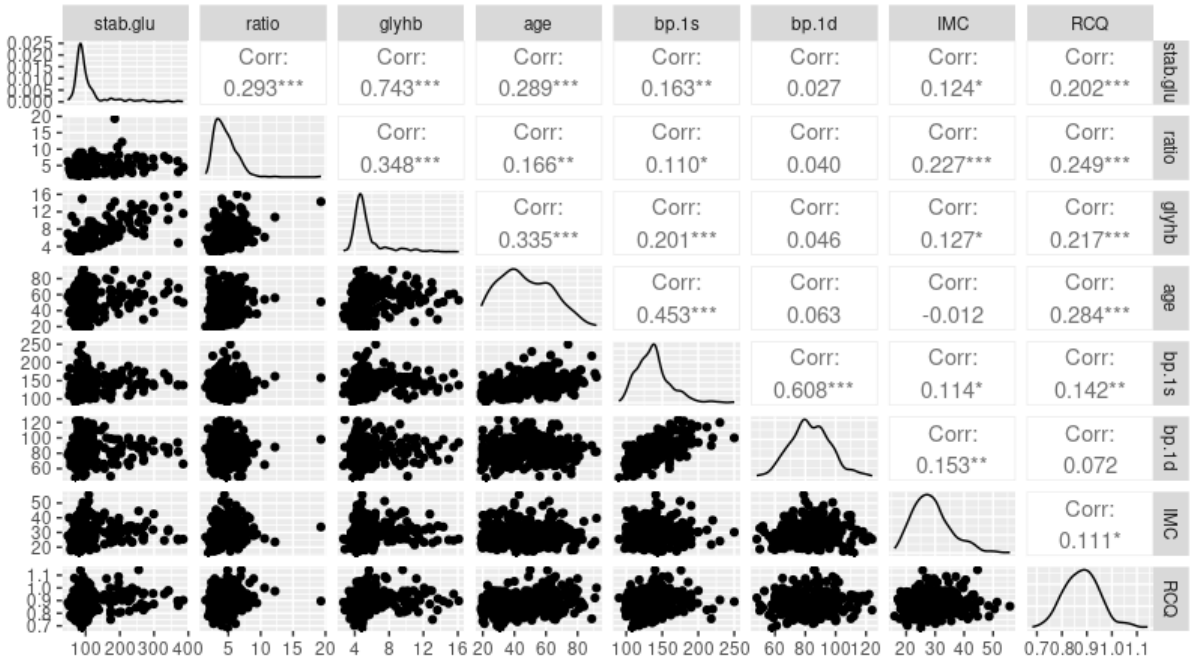
As variáveis pressão sanguínea diastólica, IMC e RCQ aparentam uma leve simetria, isso pode ser observado nos *boxplots* da Figura 3.1 e no gráfico da densidade das variáveis na Figura 3.2.

Figura 3.1: *Boxplots* das características de afro-americanos residentes no Estado da Virginia (EUA)



As variáveis, no geral, não aparentam ter uma relação linear visível entre si, quando se observa os gráficos de dispersão na Figura 3.2, com exceção das relações entre as variáveis glicose estabilizada e hemoglobina glicada, e entre pressão sanguínea diastólica e sistólica, que parecem ser linear. Esta análise é reiterada quando observamos o coeficiente de correlação linear de *Pearson*, apenas entre glicose estabilizada e hemoglobina glicada e entre pressão sanguínea diastólica e sistólica que apresentaram o valor do coeficiente de correlação maior que 0,6.

Figura 3.2: Correlação e gráfico de dispersão e densidade das características de afro-americanos residentes no Estado da Virginia (EUA)



3.3 Modelo de regressão linear múltiplo

A Tabela 3.1, apresenta uma sumarização da regressão linear em que as características dos afro-americanas resisdentes no Estado da Virginia são utilizadas para predizer o seu IMC.

Tabela 3.1: Sumarização da regressão linear em que as características dos afro-americanos residentes no Estado da Virginia (EUA) são utilizadas para predizer o seu IMC

Coeficientes	Estimativa	EP	Est. T	Valor-p
β_0	14,76	4,40	3,35	<0,001
Glicose est.	0,007	0,009	0,76	0,45
Razão colest.	0,74	0,21	3,61	< 0,001
Hemoglobina	0,08	0,23	0,34	0,73
Idade	-0,05	0,03	-1,93	0,06
Pressão sist.	0,02	0,02	0,75	0,45
Pres. dias.	0,06	0,03	1,70	0,09
RCQ	5,69	4,80	1,19	0,23

Na Tabela 3.1, é possível notar que a estimativa de β_0, apesar de ser um valor teoricamente possível do IMC, não faz sentido, pois o indivíduo não pode ter valores de RCQ, glicose estabilizada,pressão sistólica e diastólica iguais a 0, por exemplo.

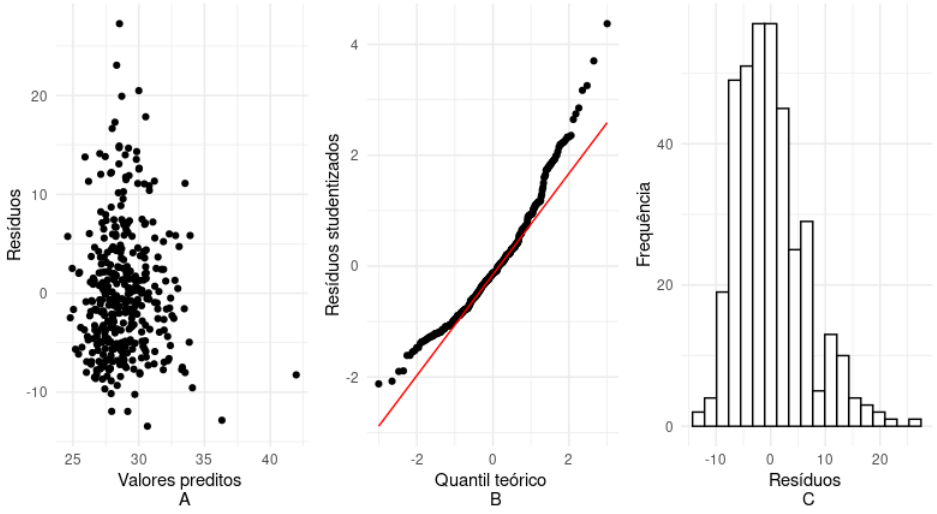
Apenas a variável razão colesterol total e colesterol bom foi significativa no modelo de regressão de múltipla, ao nível de significância de 5%. Além disso, o coeficiente de

determinação foi aproximadamente igual a $R^2 = 0,09$, ou seja, este o modelo de regressão linear múltiplo com estas variáveis independentes explicam apenas 9% da variabilidade do IMC. Portanto, parece que o modelo de regressão linear múltiplo para variável IMC com essas variáveis selecionadas não foi relevante em descrever a natureza da variável IMC.

3.3.1 Análise dos resíduos

No gráfico de dispersão dos resíduos contra os valores preditos na Figura 3.3, conseguimos perceber que não parece haver nenhuma tendência de de crescimento ou diminuição da variação dos resíduos ao longo dos valores preditos, desse modo, a variação dos resíduos parece ser constante. Realizando o teste de *Goldfeld-Quandt* para avaliar a homoscedasticidade dos resíduos, foi obtido um p-valor de 0.73, logo, a hipótese de homocedasticidade não foi rejeitada ao nível de significância de 5%. Portanto, é possível dizer que a suposição da variância constante dos resíduos não foi violada.

Figura 3.3: Gráfico dos resíduos versus valores ajustados e quantil teórico da regressão linear dos dados e afro-americanos residentes no Estado da Virginia (EUA)



No gráfico QQ-plot e no histograma dos resíduos contidos também na Figura 3.3, nota-se um grande problema no que tange à normalidade nas caudas da distribuição do resíduos, indicando uma assimetria à direita incomum a distribuição normal. Realizando o teste de *Shapiro-Wilk* para avaliar a normalidade dos resíduos, foi obtido um p-valor aproximadamente igual a 0, rejeitando a hipótese nula de que os resíduos seguem uma distribuição normal, ao nível de significância de 5%. Portanto, podemos dizer que a suposição de normalidade dos resíduos foi violada.

Para avaliar se os resíduos são independentes, foi realizado o teste de *Durbin-Watson*, que afere se os resíduos não têm autocorrelação serial. O p-valor obtido foi igual a 0.66, logo, não rejeita-se a hipótese nula de não autocorrelação serial dos resíduos, ao nível

de significância de 5%. Desse modo, podemos dizer a suposição de independência dos resíduos não foi violada

4 Conclusão

4.1 Atividade 1

A análise dos resíduos dos modelos de regressão linear simples revelou que eles não atendem a todas as suposições da regressão linear. A suposição de homocedasticidade foi violada para o modelo de regressão linear simples e para os modelos com as transformações raiz quadrada e quadrada. A suposição de normalidade foi violada para o modelo com a transformação raiz quadrada, logarítmica e Box-Cox. Já a suposição de independência foi violada para todos os modelos de regressão considerados.

Os coeficientes de todos os modelos são estatisticamente significativos. Em termos de ajuste aos dados e capacidade de explicar a variação no volume das cerejeiras, os modelos de regressão linear simples com as transformações logarítmica e Box-Cox apresentam o melhor desempenho, com o coeficientes de determinação mais altos.

No entanto, cada uma dessas transformações resultou em diferentes relações entre a altura e o volume das cerejeiras e é importante considerar a interpretabilidade dos coeficientes. Por possuir uma interpretabilidade maior, o modelo com a transformação logarítmica é o mais adequado para quantificar a relação.

Mas, os pressupostos estão sendo violados, sobretudo o de independência: há auto-correlação serial. Diante disso, é importante ter cautela ao interpretar os resultados, eles podem não garantir uma estimativa precisa ou confiável da relação entre a altura e o volume das cerejeiras.

Além disso, é necessário considerar que os modelos de regressão linear simples têm limitações inerentes, e a inclusão de outras variáveis ou técnicas mais avançadas de modelagem estatística poderia melhorar a precisão das estimativas. Portanto, é recomendável realizar análises adicionais e considerar outras abordagens para explorar e modelar adequadamente a relação entre a altura e o volume das cerejeiras na Floresta Nacional de Allegheny.

4.2 Atividade 2

Na segunda atividade, foi realizada a manipulação dos dados, com o intuito de checar inconsistências, observar valores faltantes, mudar a escala das variáveis para o padrão brasileiro, criar novas variáveis a partir de variáveis já existentes no banco de dados e excluir aquelas que não são mais úteis para a nossa análise.

Na análise descritiva, foi observado que há muitos valores discrepantes nas variáveis glicose estabilizada e hemoglobina glicada, contudo esses valores nos trazem informação sobre possíveis diabéticos, já que ambas as variáveis avaliam a glicose no sangue. Além disso, foi notado que tirando as variáveis glicose estabilizada e hemoglobina glicada, e pressão sanguínea sistólica e diastólica, as variáveis não estão muito correlacionadas e não apresentam uma relação linear visível entre si.

Um modelo de regressão linear múltiplo para predizer o IMC foi proposto. Contudo, apenas a variável explicativa razão colesterol total e colesterol bom foi significativa, ao nível significância de 5%, no modelo. Além disso, o coeficiente de determinação para este modelo foi de apenas $R^2 = 0,09$, o que significa que apenas 9% da variabilidade do IMC pode ser explicada pelo modelo. Portanto, o modelo linear múltiplo com estas variáveis não parece ser tão adequado.

Ademais, o modelo não violou a suposição de homoscedasticidade e independência dos resíduos, porém violou a suposição de normalidade dos resíduos. Portanto, além de ser um modelo com baixíssima explicabilidade, é um modelo que viola as suposições requeridas para um modelo de regressão linear. Desse modo, a alternativa seria inserir outras variáveis explicativas no modelo de regressão linear múltiplo ou mudar de modelo de regressão.