LABORATÓRIO 5: Regressão Linear

Fernado Bispo, Jeff Caponero

Sumário

Introdução

O laboratório desta semana está subdividido em duas partes com análises de dois conjuntos de dados distintos que visa a continuidade da aplicação das técnicas de Regressão Linear Simples com a aplicabilidade das técnicas de análise de resíduos e transformação de variáveis inclusive. Para melhor desenvolvimento do processo de analise, este relatório foi dividido em duas partes contendo as análises de cada um dos conjuntos de dados e contando com suas respectivas apresentações sobre o contexto a ser analisado.

Parte 1: Regressão Linear Simples - Diagnóstico do modelo

Metodologia

O primeiro conjunto de dados a ser analisado é denominado *trees*, disponível no pacote *datasets*, contém informações de 31 cerejeiras (*Black cherry*) da Floresta Nacional de Allegheny, relativas a três características numéricas contínuas:

- Volume de madeira útil (em metros cúbicos (m^3));
- Altura (em metros (m));
- Circunferência (em metros(m)) a 1,37 de altura.

Para esta atividade **serão considerados apenas as informações referentes ao volume e altura das árvores**. Com base nestes dados se desenvolverá:

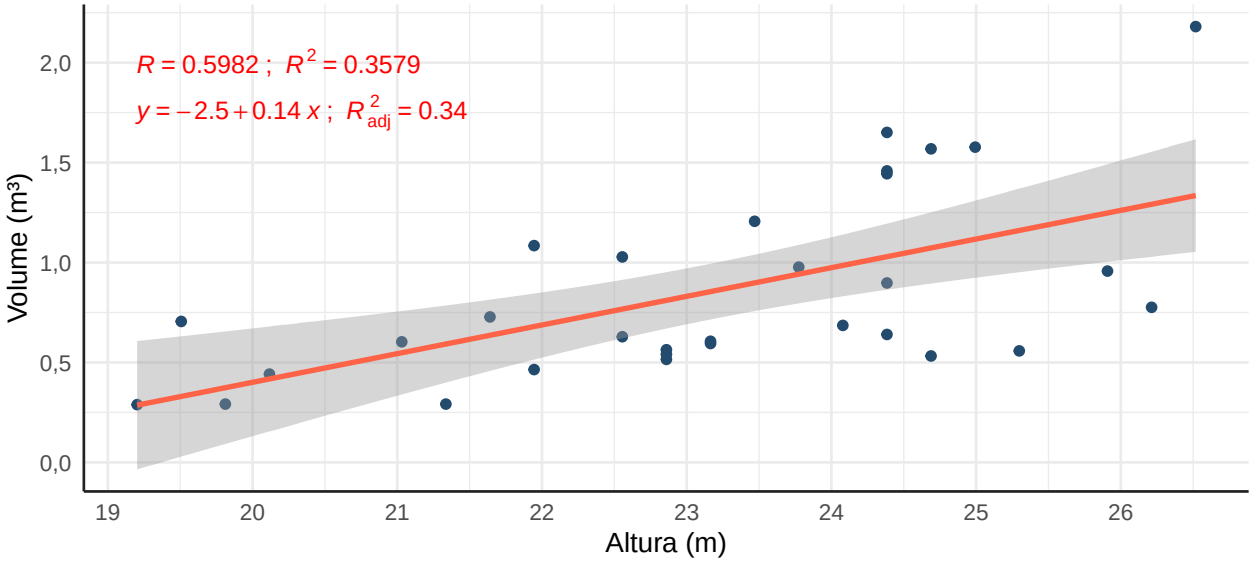
(a) Ajuste de modelo linear simples para volume como função da altura da árvore;
(b) Avalação gráfica dos resíduos Jacknife para diagnóstico do modelo ajustado;
(c) Transformações da característrica utilizada como variável resposta do modelo;
(d) Avaliação da transformação mais apropriada dentro da família proposta por Box e Cox;
(e) Indicação da melhor transformação analisada.

Resultados

Ajuste do Modelo

Figura 1: Modelo Ajustado entre Volume e Altura

Diagrama de dispersão com equação da reta de regressão ajustada, Coeficiente de Correlação de Pearson, Coeficiente de Determinação, Coeficiente de Determinação Ajustado, Reta de Regressão e Intervalo de Confiança



Com base na Figura 1 é possível sugerir uma relação positiva entre as variáveis **Volume** e **Altura**, fato confirmado pelo Coeficiente de Correlação de Pearson ($R = 0{,}598$) que após o teste de hipótese para avaliar a significância da correlação estimada, demonstrou **possuir correlação não nula**. A reta de regressão ajustada segue a seguinte equação:

$$\hat{Y}_i = -2,5 + 0,14X_i$$

A Tabela 1 traz os resultados do teste de hipóteses para correlação e o Intervalo de Confiança para o verdadeiro valor da correlação, podendo concluir, com base no p-valor menor que o nível de significância ($\alpha = 5\%$), que a hipótese nula (H_0) foi rejeitada, assumindo-se a hipótese alternativa (H_1) que afirma que $\rho \neq 0$.

Tabela 1: Teste de Hipótese para Correlação entre Volume e Altura

	t	p-valor	LI	LS
Altura	4,02051	0,00038	0,30952	0,78598

Nota: Teste realizado com 5% de significância

O Coeficiente de Determinação (R^2) apresenta um valor baixo, podendo afirmar que apenas aproximadamente 36% da variabilidade dos dados está sendo explicada pelo modelo de regressão calculado.

Significância do Modelo

Após o ajuste do modelo existe a necessidade de se avaliar a significância do mesmo, o teste de hipótese para tal situação será realizado, contendo as seguintes hipóteses:

$$H_0 : \hat{\beta}_1 = 0$$
$$H_1 : \hat{\beta}_1 \neq 0.$$

As Tabelas 4 e 5 trazem os principais resultados da tabela ANOVA e do Intervalo de Confiança para os parâmetros, possibilitando assim inferir sobre o modelo ajustado.

Tabela 2: Análise de Variância (ANOVA)

	GL^1	Soma de Quadrados	Quadrado Médio	Estatística F-Snedecor	p-valor
Regressão	1	2,326	2,326	16,1645	0,0004
Resíduos	29	4,174	0,144		

Legenda:
[1] GL: Graus de Liberdade

Tabela 3: Intervalos de Confiança para os parâmetros estimados no MRLS.

	LI^1	LS^2
$\hat{\beta}_0$	-4,162	-0,772
$\hat{\beta}_1$	0,070	0,216

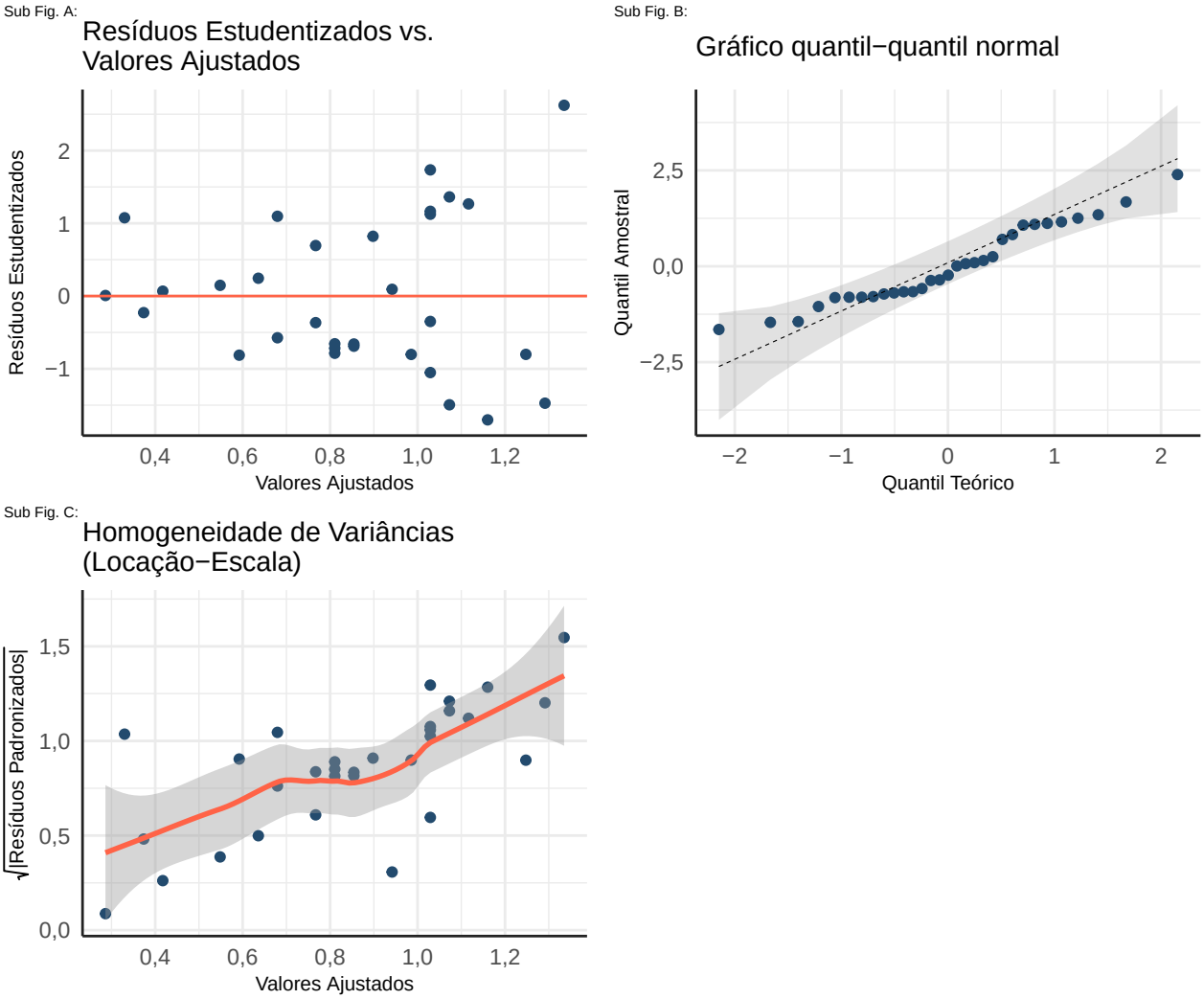
Legenda:
[1] LI: Limite Inferior (2,5%)
[2] LS: Limite Superior (97,5%)
[*] Nível de Significância de 5%.

Com base na Tabela 2, avaliando o p-valor é possível afirmar que o modelo é significante rejeitando assim H_0 que tem como pressuposto $\hat{\beta}_1 = 0$.

Através dos Intervalos de Confiança calculados (Tabela 3) é possível afirmar **com 95% de confiança que o verdadeiro valor de β_0 está entre (-4,1624; -0,7717) e que o verdadeiro valor de β_1 está entre (0,0704; 0,2163)**.

Análise de Resíduos

Figura 2: Análise de resíduos do modelo ajustado



A Figura 2 Sub.Fig A apresenta um comportamento assimétrico dos resíduos, podendo ser constatado uma pequena variabilidade inicial e um aumento desta à medida que os valores ajustados aumentam, caracterizando uma heterocedasticidade. A Sub.Fig C, que trada da Homogeneidade de Variâncias (Locação-Escala) ressalta que há um problema na variabilidade dos dados, corroborando com a interpretação feita na análise da Sub.Fig. A, de que há uma mudança na variabilidade dos dados, caracterizando Heterocedasticidade dos dados. A Sub.Fig. B que traz o gráfico para avaliação da normalidade dos dados, mostra que apesar dos dados não estarem precisamente sobre a reta de referência, os mesmos estão contidos na região pertencente ao Intervalo de Confiança - IC, podendo assumir que há normalidade, contudo tal avaliação será confirmada após os Testes de Diagnóstico.

Testes de Diagnósticos do Modelo

Para avaliar se o modelo atende aos pressupostos, além da análise gráfica podem ser realizados testes de diagnósticos, que são testes de hipóteses para avaliação dos pressupostos que são:

- Normalidade;

  H_0 : Os resíduos possuem normalidade.

  H_1 : Os resíduos **não** possuem normalidade.

- Homoscedasticidade (Homogeneidade de Variância);

  H_0 : Os resíduos possuem variância constante.

  H_1 : Os resíduos **não** possuem variância constante.

- Linearidade;

- Independência.

  H_0: Existe correlação serial entre os resíduos.

  H_1: **Não** existe correlação serial entre os resíduos.

Para tanto serão utilizados os seguintes testes:

- Shapiro-Wilk, para avaliar a Normalidade;
- Breush-Pagan, para avaliar a Homoscedasticidade;
- Durbin-Watson, para avaliar a Independência.

Tabela 4: Testes de Diagnósticos dos Resíduos

	Estatística de teste	p-valor
Shapiro-Wilk	0.9508	0.164
Breush-Pagan	7.4901	0.0062
Durbin-Watson	0.5009	<0,0001

A Tabela 4 traz os testes de diagnósticos realizados para avaliar o modelo de regressão ajustado, conforme análise gráfica dos resíduos há a confirmação da heterocedasticidade de variância conforme o p-valor obtido pelo teste de Breush-Pagan (0,006) bem como a dependência entre as características confirmado pelo p-valor do teste de Durbin-Watson (<0,0001), em ambos os testes a hipótese nula (H_0) foi rejeitada com base p-valor, como tentativa de contornar a quebra dos pressupostos se faz necessária a transformação da variável resposta.

Transformações dos Dados

Para esta análise a característica em estudo (Volume) será muitas vezes representada pela letra Y para melhor representação.

Tendo em vista que o modelo não atendeu aos pressupostos se faz necessário a utilização de técnicas para buscar uma melhora de performace do modelo antes da possibilidade de descarte e para tanto algumas transformações são sugeridas, sendo estas:

- $T_1 = \sqrt{Y}$;
- $T_2 = log(Y)$;
- $T_3 = Y^2$.

Sendo Y a variável resposta do modelo representada pelo Volume.

Tabela 5: Medidas Resumo da característica Volume com e sem transformações.

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV
1. Volume	0,289	0,541	0,685	0,854	1,085	2,180	0,465	0,545
2. √Volume	0,537	0,735	0,828	0,894	1,041	1,477	0,239	0,267
3. log(Volume)	-1,242	-0,615	-0,378	-0,292	0,081	0,780	0,526	-1,805
4. Volume²	0,083	0,293	0,470	0,940	1,176	4,754	1,051	1,119

A Tabela 5 traz a sumarização da característica em analise em sua forma natural juntamente com as formas transformadas para tentar identificar o comportamento destas transformações, podendo constatar um aumento na variabilidade dos dados, com base no valor do Coeficiente de Variação, com exceção da variável sob a transformação Y^2. Seguem as equações das retas ajustadas após a transformação.

$$\sqrt{\hat{Y}_i} = -0,89 + 0,077X_i$$

$$\log(\hat{Y}_i) = -4,4 + 0,18X_i$$

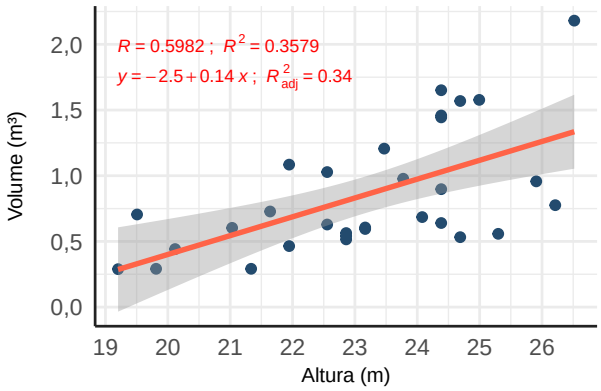
$$\hat{Y}_i^{\,2} = -5,9 + 0,3X_i$$

A Figura 3 traz os diagramas de dispersão com a reta ajustada para o modelo sem transformação e para cada transformação realizada, a fim de possibilitar a identificação das diferenças entre cada modelo.

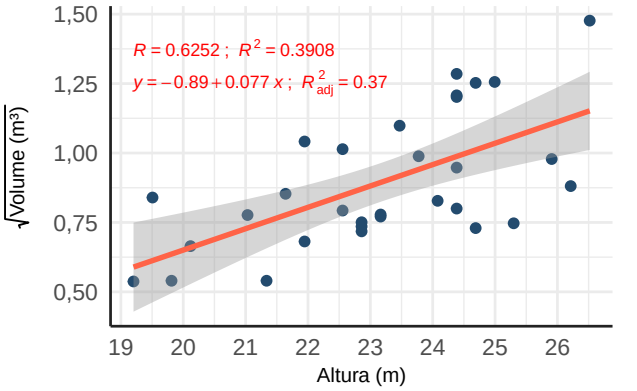
Figura 3: Modelo ajustado e suas transformações

Comparativo entre o modelo ajustado sem transformação com os modelos após a transformações da variável resposta.

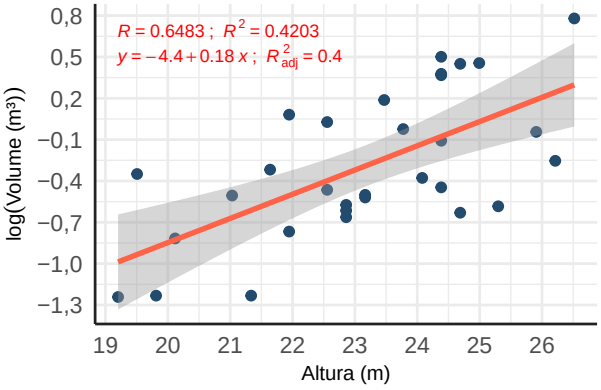
Sub Fig. A:



Sub Fig. B:



Sub Fig. C:



Sub Fig. D:

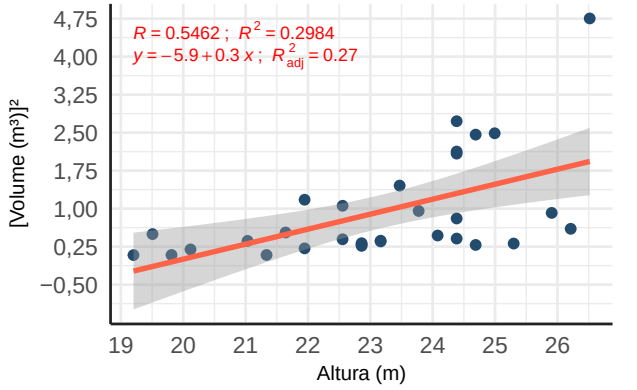
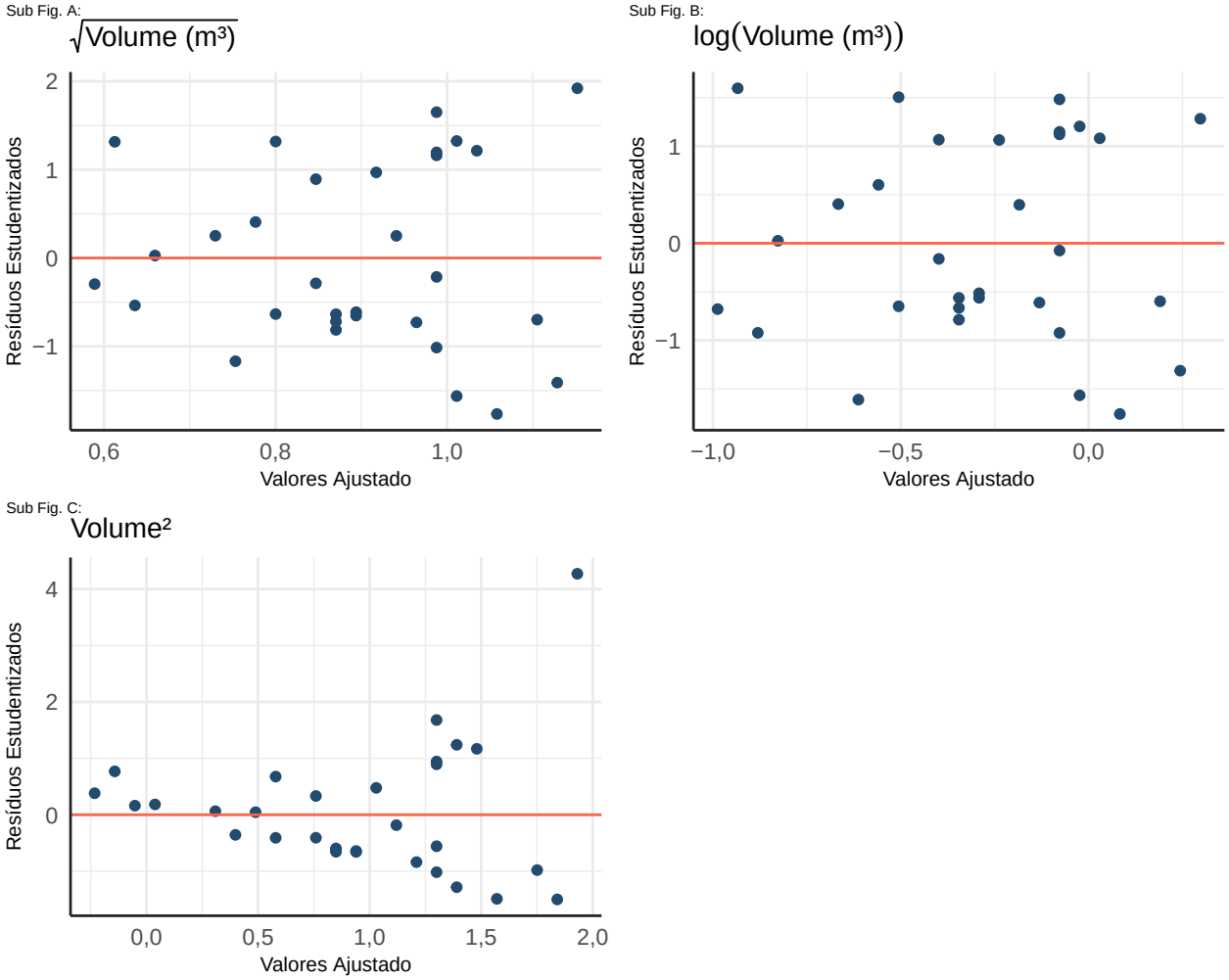


Figura 4: Gráficos dos Resíduos Estudentizados vs. Valores Ajustados das características transformadas.



A Figura 5 traz os gráficos comparativos dos resíduos estudentizados (*Jacknife*) versus os valores ajustados das variáveis transformadas, sendo possível constatar que houve uma aparente melhora na variabilidade nas Sub.Figs. A e B, contudo houve uma maior evidenciação do ponto atípico na Sub.Fig C, sendo este influente o suficiente para interferir a variabilidade total deste modelo, sendo descartado este modelo transformado pela quebra do pressuposto da homogeneidade de variâncias. Dentre os modelos sob a transformação $\sqrt{Y}$ e $log(Y)$, a figura que traz o comparativo dos resíduos sob a transformação $log(Y)$ (Sub.Fig. B) apresenta uma maior homogeneidade de variância, sendo o modelo mais adequado dentre os modelos transformados.

Como forma de confirmar a avaliação feita sobre a analise gráfica, foi construída a Tabela 6 com os testes de diagnósticos dos resíduos do modelo sob a transformação $log(Y)$.

Tabela 6: Testes de Diagnósticos dos Resíduos após transformação log(Y)

	Estatística de teste	p-valor
Shapiro-Wilks	0.9131	0.0155
Breush-Pagan	0.4757	0.4904
Durbin-Watson	0.5066	<0,0001

A Tabela 6 traz os testes de diagnóstico para a o modelo após a transformação da variável resposta, sendo possível constatar que após a transformação o novo modelo não rejeita a hipótese nula (H_0) para o teste de homogeneidade de variância, corroborando com a análise gráfica, contudo, os demais testes não foram bem sucedidos.

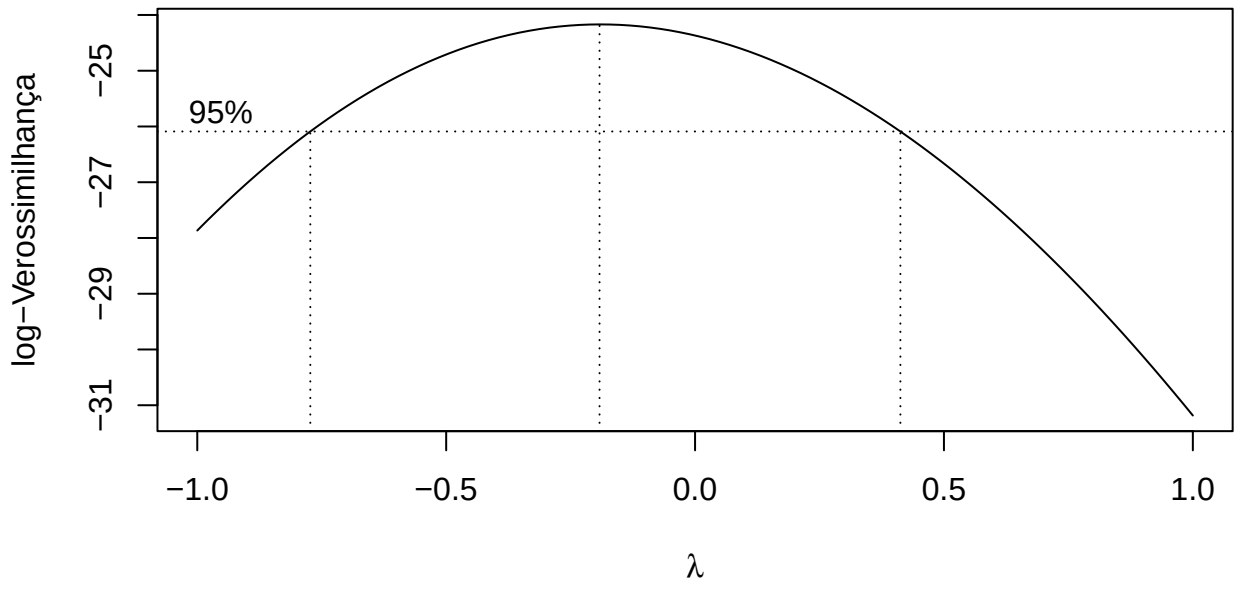
Tendo em vista não ser a tarefa mais simples a aplicação de diversas transformações e os devidos testes para avaliar o melhor possível modelo a ser utilizado, que minimize a variância residual, a opção mais adequada é a escolha do modelo baseado na família de transformações de Box-Cox, definida por:

$$Y(\lambda) = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda}, \text{ se } \lambda \neq 0 \\ log(Y), \text{ se } \lambda = 0 \end{cases}$$

sendo λ o parâmetro da transformação.

Para tanto a Figura 5 foi construída com base na função de λ para a escolha da transformação apropriada.

Figura 5: Transformação Box−Cox



Analisando o gráfico das **famílias de transformações Box-Cox** é possível identificar que $-0,5 < \lambda_{max} < 0$ ($\lambda_{max} \approx$ -0,192), partindo do princípio que o valor zero está incluso no intervalo de valores possíveis de λ que minimizam a variância residual, mesmo o zero não sendo o máximo valor assumido, ainda assim, visando a escolha de uma transformação que possibilite uma interpretação facilitada, **a escolha da transformação** $log(Y)$ **torna-se a escolha mais assertiva**, conforme conclusão anteriormente feita baseado na análise gráfica.

Conclusão

Após as análises realizadas sobre modelo ajustado foi possível constatar uma melhora no desempenho do modelo, quanto a variabilidade dos dados, após a transformação da variável resposta para a transformação $log(Y)$, fato constatado através da Figura 3 por meio do valor do Coeficiente de Determinação R^2 bem como através da análise da Figura 4, por meio dos resíduos.

Apesar de não se ter conseguido um modelo que preenchesse todos os pressupostos, ainda assim, na possibilidade desse modelo ter sido satisfatório, poder-se-ia interpretar a sua utilização da seguinte forma: Para cada aumento de um metro na altura da árvore, há uma redução média de $exp(-4,4)$ m^3 no volume ou aproximadamente 0,012 m^3.

Parte 2: Regressão Linear Múltipla - Estimação pontual

Metodologia

O segundo conjunto de dados a ser analisado, contém informações de 403 afro-americanos residentes no Estado da Virginia (EUA), entrevistados em um estudo referente à prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares. As características apresentadas são:

- Colesterol total;
- Glicose estabilizada;
- Lipoproteína de alta densidade (colesterol bom);
- Razão colesterol total e colesterol bom;
- Hemoglobina glicada;
- Município de residência (Buckingham ou Louisa);
- Idade (em anos);
- Sexo;
- Altura (em polegadas);
- Peso (em libras);
- Pressão sanguínea sistólica (1ª medida);
- Pressão sanguínea diastólica (1ª medida);
- Pressão sanguínea sistólica (2ª medida);
- Pressão sanguínea diastólica (2ª medida);
- Cintura (em polegadas);
- Quadril (em polegadas).

Com base nestes dados se desenvolverá:

(a) Verificação de possíveis inconsistências nos dados, dados ausentes e identificar a escala de cada característica;
(b) Conversão das medidas das características expressas em escalas não usuais ao padrão brasileiro;

(c) Tentar identificar uma variável que poderia ser considerada a variável resposta em uma análise de regressão e na medida do possível, explorar a relação dessa variável com as demais;
(d) Proposição de novas variáveis, baseadas nas variáveis disponíveis;
(e) Ajuste de um Modelo de Regressão Linear Múltiplo - MRLM, interpretação e verificação da significância dos parâmetros do modelo ajustado (apenas para variáveis quantitativas).

Resultados

Análise descritiva dos dados

Numa primeira análise do conjunto de dados se constatou que as características **Pressão sanguínea sistólica (2ª medida)** e **Pressão sanguínea diastólica (2ª medida)** possuem uma quantidade muito grande de dados ausentes, cerca de 65% de ausência de dados, ou seja, das 403 observações coletadas, 262 estão ausentes para estas características, diante desta grande falta de dados essas características serão descartadas.

Se constatou também que as observações das características **altura, peso, cintura** e **quadril** estão representadas em unidades do Sistema Imperial, diferentes das praticadas no Brasil, sendo necessária a conversão para o Sistema Internacional, a fim de facilitar a interoperabilidade para a nossa realidade cotidiana. Para tanto as características que possuem medidas originais em **polegadas (in), libras (lb), polegadas (in) e polegadas (in)** respectivamente, foram convertidas para **metro (m), quilograma (kg), centímetro (cm) e centímetro (cm)** respectivamente.

Sendo parte primordial para qualquer estudo, a fase exploratória dos dados está representada inicialmente na Tabela 7 com a sumarização das características em análise.

Tabela 7: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
Altura	1,32	1,60	1,68	1,68	1,75	1,93	0,10	0,06	0,03	-0,21
Cintura	66,04	83,82	93,98	96,27	104,14	142,24	14,55	0,15	0,47	-0,17
Colesterol total	78,00	179,00	204,00	207,85	230,00	443,00	44,45	0,21	0,92	2,54
Glicose estabilizada	48,00	81,00	89,00	106,67	106,00	385,00	53,08	0,50	2,75	8,10
Hemoglobina glicada	2,68	4,38	4,84	5,59	5,60	16,11	2,24	0,40	2,23	4,98
Idade	19,00	34,00	45,00	46,85	60,00	92,00	16,31	0,35	0,32	-0,67
Lipoproteína de alta densidade	12,00	38,00	46,00	50,45	59,00	120,00	17,26	0,34	1,19	1,93
Peso	44,91	68,49	78,24	80,55	90,72	147,42	18,30	0,23	0,72	0,67
Pressão sanguínea diastólica (1ª medida)	48,00	75,00	82,00	83,32	90,00	124,00	13,59	0,16	0,27	0,04
Pressão sanguínea sistólica (1ª medida)	90,00	121,00	136,00	136,90	147,00	250,00	22,74	0,17	1,10	2,38
Quadril	76,20	99,06	106,68	109,32	116,84	162,56	14,37	0,13	0,80	0,83
Razão colesterol total e colesterol bom	1,50	3,20	4,20	4,52	5,40	19,30	1,73	0,38	2,20	13,17

Regressão Linear Múltipla

Com base na proposta da coleta de dados, a característica que apresenta melhor adequação para ser a variável resposta de um modelo de regressão é o **peso**, pois intuitivamente é a que apresenta melhor correlação. Desta forma, será feita a avaliação dessa característica com as demais a fim de se identificar as possíveis correlações intuídas.

Tabela 8: Análise de Variância (ANOVA).

	GL^1	Soma de Quadrados	Quadrado Médio	Estatística F-Snedecor	p-valor
Colesterol total	1	403,501	403,5011	8,5784	0,0036
Glicose estabilizada	1	3.722,204	3.722,2039	79,1337	<0,0001
Lipoproteína de alta densidade	1	10.435,440	10.435,4396	221,8564	<0,0001
Razão colesterol total e colesterol bom	1	5,385	5,3850	0,1145	0,7353
Hemoglobina glicada	1	4,542	4,5418	0,0966	0,7562
Idade	1	2.012,321	2.012,3210	42,7817	<0,0001
Altura	1	5.511,324	5.511,3239	117,1702	<0,0001
Pressão sanguínea sistólica (1ª medida)	1	1.873,619	1.873,6191	39,8330	<0,0001
Pressão sanguínea diastólica (1ª medida)	1	1.775,081	1.775,0806	37,7380	<0,0001
Cintura	1	76.162,631	76.162,6313	1.619,2098	<0,0001
Quadril	1	7.814,820	7.814,8199	166,1423	<0,0001
Resíduos	365	17.168,474	47,0369		

Legenda:
[1] GL: Graus de Liberdade

A análise da Tabela 8 permite avaliar que apenas para a razão colesterol total e colesterol bom e para a Hemoglobina glicada o peso do paciente está relacionado aos resultados obtidos de forma significativa. Para todos os demais a correlação não é evidente.

Realizando a RLM para a variável IMC

Comummente nas avaliações na área de saúde o peso isoladamente é menos significativo que quando dividido pelo quadrado da altura, constituindo o que se convencionou de Índice de Massa Corpórea (IMC). Refazendo os calculos anteriores para essa nova variável temos o que se pode ver nas tabelas a seguir.

Tabela 9: Análise de Variância (ANOVA).

	GL^1	Soma de Quadrados	Quadrado Médio	Estatística F-Snedecor	p-valor
Colesterol total	1	125,860	125,8604	15,2529	0,0001
Glicose estabilizada	1	201,777	201,7772	24,4532	<0,0001
Lipoproteína de alta densidade	1	1.036,773	1.036,7727	125,6457	<0,0001
Razão colesterol total e colesterol bom	1	87,165	87,1647	10,5634	0,0013
Hemoglobina glicada	1	5,057	5,0570	0,6129	0,4342
Idade	1	56,523	56,5228	6,8500	0,0092
Pressão sanguínea sistólica (1ª medida)	1	248,162	248,1619	30,0746	<0,0001
Pressão sanguínea diastólica (1ª medida)	1	146,950	146,9504	17,8088	<0,0001
Cintura	1	9.570,640	9.570,6402	1.159,8587	<0,0001
Quadril	1	2.049,798	2.049,7982	248,4135	<0,0001
Resíduos	366	3.020,070	8,2516		

Legenda:
[1] GL: Graus de Liberdade

A análise da Tabela 9 indica que o peso precisa ser avaliado em função da altura para desterminar a razão colesterol total e colesterol bom, uma vez que na ponderação do IMC esta condição deixa de estar significativamente correlacionada. Já para a Hemoglobina glicada o peso do paciente ou seu IMC estão relacionados aos resultados obtidos de forma significativa. Para todos os demais a correlação não é evidente.

Conclusões

Embora o banco de dados utilizado tenha diversos dados ausentes e use um sistema métrico diverso, o tratamento dos dados permitiu a avaliação da correlação entre as variáveis. Observou-se que o peso do paciente apenas tem relação significativa com sua razão de colesterol total e colesterol bom e suas medidas de Hemoglobina glicada, entretanto a ponderação do peso por meio do Índice de Massa Corpórea (IMC), deixa claro que apenas a Hemoglobina glicada tem relação com o peso quando se leva em consideração a altura do paciente. Como um trabalho futuro, pode-se pensar em uma análise envolvendo variáveis resposta menos intuitivas que o peso do paciente ou seu IMC, que tem potencial para revelar outras relações entre as condições clínicas do paciente.