# Análise de Modelo de Regressão Linear para o nível de Glicose de Indígenas Norte-Americanas

Vitória Nascimento de Jesus Sesana

## Sumário

# 1 Resumo

O Instituto Nacional de Diabetes e Doenças Digestivas E Renais dos Estados UnidoS realizou, em 1988, uma pesquisa em que foram observadas mulheres descendentes dos Pimas, povo indígena norte-americano, que vivem nos arredores de Phoenix-Arizona, para prever qual dessas mulheres teriam ou não diabetes de acordo com características físicas analisadas. No entanto, o intuito deste relatório é analisar como essas características afetam o nível de glicose dessas mulheres. Os dados originais da pesquisa contém 9 variáveis, sendo 8 relacionadas a características médicas de 768 mulheres descendentes desse grupo e que possuíssem pelo menos 21 anos.

## 2 Sobre o modelo de regressão linear

Entender como um elemento se comporta a partir de um conjunto de dados é um dos objetivos para construir modelos de regressão, além de ter a possibilidade de predize-los. No caso do modelo de regressão linear, essa relação é expressa como uma função linear, como por exemplo $y = a + bx$

Com base em métodos estatísticos e matemáticos, consegue-se elaborar uma função que tenta explicar como uma variável se comporta de acordo com outras variáveis. Entretanto, o processo de modelagem dos dados apresenta diversas etapas que vão desde a interpretação inicial dos dados até averiguar a qualidade do modelo. Para identificar se o modelo construído é adequado ou não ele precisa se adequar às suposições exigidas. O resultado da construção de um modelo de regressão linear múltipla é obter estimativas dos coeficientes para cada covariável que explique a relação entre o conjunto de dado à variável de interesse, a fim de obter valores próximos dos valores observados.

## 3 Análise Descritiva

O base de dados com 9 características coletadas da população de estudo está disponibilizada no site do Kaggle, com o número total de 768 observações.

Tabela 1: Primeiras observações da população de estudo

| | Y | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
|---|---|---|---|---|---|---|---|---|---|
| Observação 1 | 148 | 6 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| Observação 2 | 85 | 1 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| Observação 3 | 183 | 8 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| Observação 4 | 89 | 1 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| Observação 5 | 137 | 0 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

Onde:

- (Y) **Glucose**: Concentração de glicose por meio de teste oral de tolerância à glicose;
- ($X_1$) **Pregnancies**: Quantidade de vezes que a mulher engravidou;
- ($X_2$) **BloodPressure**: Pressão arterial diastólica (mm Hg);
- ($X_3$) **SkinThickness**: Espessura cutânea triciptal (mm);
- ($X_4$) **Insulin**: 2 horas de insulina no soro (mu U/ml);
- ($X_5$) **BMI**: Índice de massa corporal (IMC);
- ($X_6$) **DiabetesPedigreeFunction**: Diabetes em função da ancestralidade, indica a probabilidade de diabetes com base no histórico familiar.
- ($X_7$) **Age**: Idade em anos das mulheres observadas;
- ($X_8$) **Outcome**: Indicador da presença de diabetes (0 = não possui, 1 = possui);

No entanto, a coluna *Outcome* (X8), indicador de diabetes (0=saudavel, 1=diabético), não será utilizada no modelo, pois essa era a coluna de variável resposta proposto pelo desafio inicial, que não é o interesse desse relatório. Desse modo, a base passa a ter 8 variáveis, sendo a *Glucose* (Y) a variável de interesse (ou resposta) e as demais, $X_i \quad i = 1, \ldots, 7$, as variáveis explicativas (ou covariáveis).

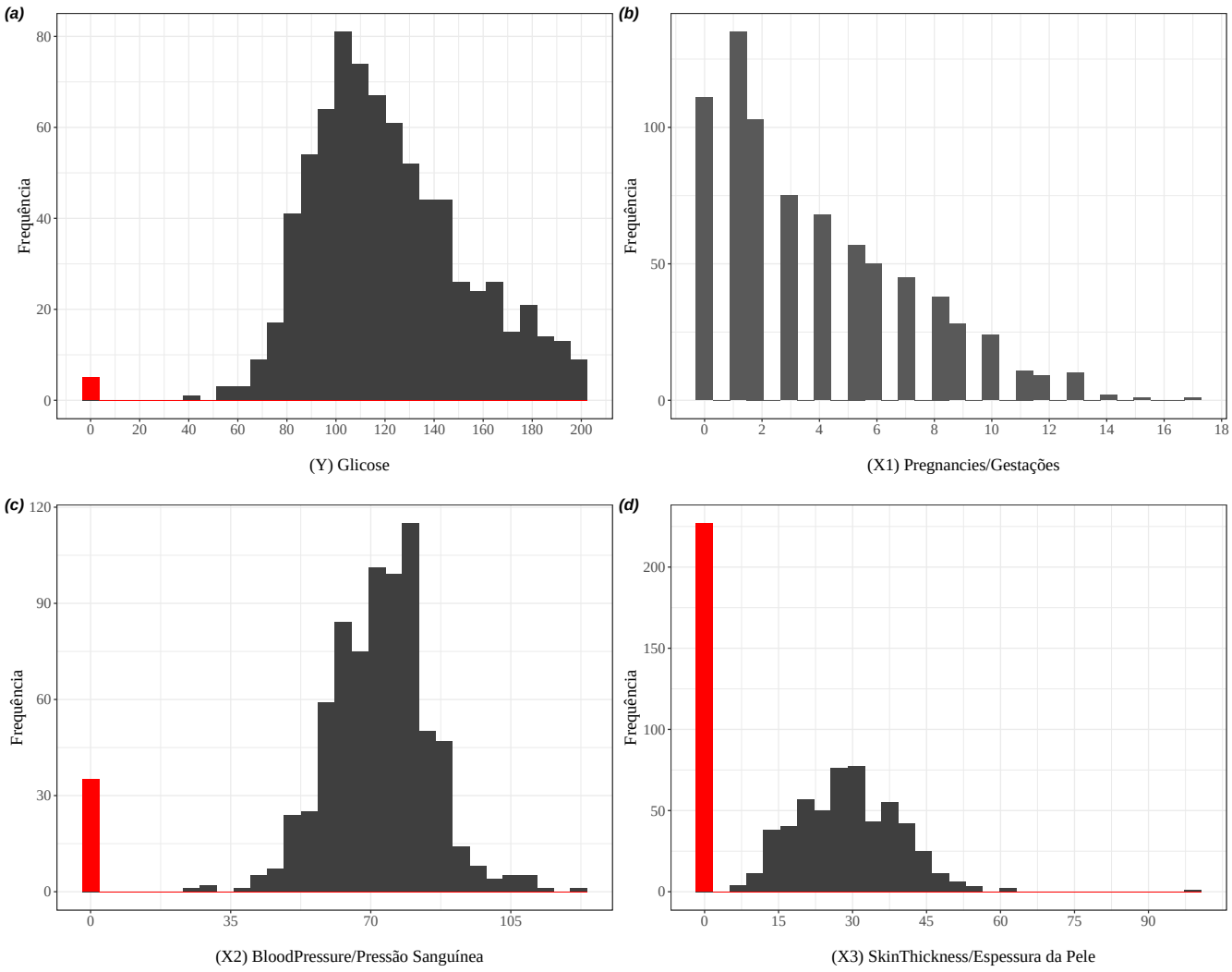


Figura 1: Histograma das colunas Y, X1, X2 e X3

A Figura 1 mostra a frequência dos níveis de glicose (a), quantidade de gestações (b), pressão sanguínea (c) e espessura da pele (d). Os valores destacados mostram resultados impossíveis para essas variáveis, como o indivíduo possuir 0 pressão sanguínea ou 0 nível de glicose no sangue. Esses valores estão iguais a zero, pois indicam valores faltantes na base de dados, ou seja, que por algum motivo não foram coletados. No gráfico (d) é o mais alarmante, pois há muita quantidade de dados que não condiz com o cenário real do estudo.

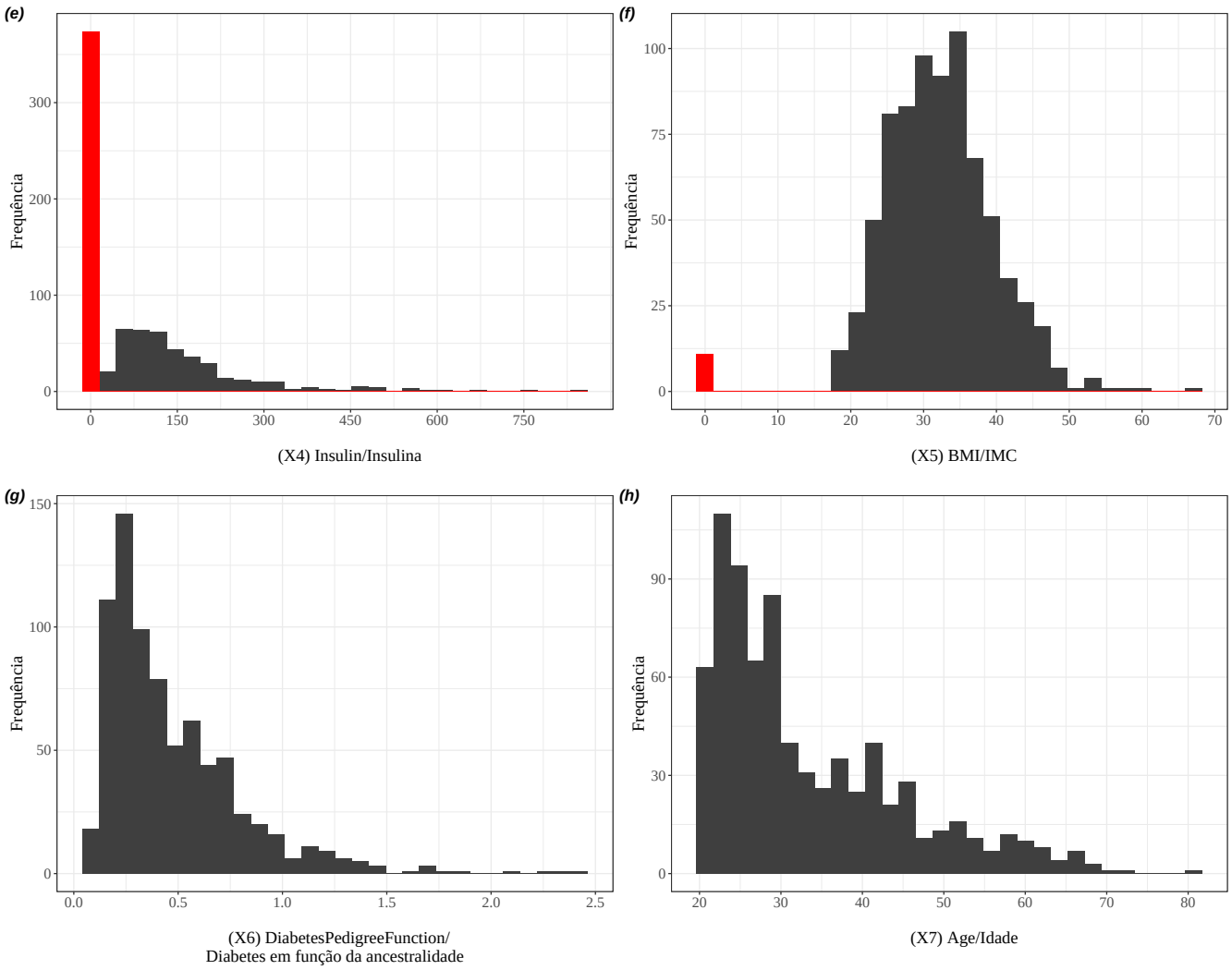


Figura 2: Histograma das colunas X4, X5, X6 e X7

Do mesmo jeito que a Figura 1, a Figura 2 apresenta valores destacados, porém para as variáveis insulina (e), IMC (f), função de diabetes (g) e idade (h). Neste caso, apenas os valores da insulina e IMC apresentam dados que não condizem. O IMC apresenta poucos dados surreais, ao contrário da insulina que possui diversas

Esses valores destacados indicam dados que não foram coletados e eles impactam na construção do modelo. Para compreender melhor o nível do problema, foi montado a seguinte tabela:

Tabela 2: Observações de dados faltantes

| | Quantidade de dados faltantes | Percentual em relação à própria variável (%) | Percentual em relação aos dados faltantes totais (%) |
|---|---|---|---|
| Y | 5 | 0.65 | 0.77 |
| X1 | 0 | 0 | 0 |

| | Quantidade de dados faltantes | Percentual em relação à própria variável (%) | Percentual em relação aos dados faltantes totais (%) |
|---|---|---|---|
| X2 | 35 | 4.56 | 5.37 |
| X3 | 227 | 29.56 | 34.82 |
| X4 | 374 | 48.7 | 57.36 |
| X5 | 11 | 1.43 | 1.69 |
| X6 | 0 | 0 | 0 |
| X7 | 0 | 0 | 0 |
| total | 652 | - | - |

Descritiva Dados Faltantes

Para obter bons valores nas estimações, irá eliminar as observações em que a variável de interesse possui dados faltantes, no caso foram 5 observações. Desse modo, a base fica com 763 observações.

Vale se atentar que a característica nível de insulina (X4), possui muitos NA's (dados faltantes), representando cerca de 48.7% do total de elementos dessa variável. Geralmente quando isso ocorre o ideal é eliminar essa variável da base de dados para elaborar o modelo.

No entanto, a variável continuará na base de dados utilizada para estimar o modelo, tendo em vista que a insulina é relacionada com o nível de glicose.

Há métodos para lidar com os dados faltantes. Como substituir os valores de acordo com a média do conjunto de dados, medianas e até mesmo elaborar um modelo de regressão para esses valores.

No R, há a biblioteca **MICE**, pacote que é voltado para esse problema e que possui diversos. Neste caso, escolhi o método "*cart*" que utiliza a técnica de Árvores de classificação e regressão para imputar os valores faltantes.[1]

Realizada a imputação dos dados faltantes, vamos verificar como é a correlação entre as variáveis com o gráfico abaixo.

A correlação é um valor que varia de -1 a 1 e quanto mais próximo do 0 menor a relação entre as variáveis, quanto os valores se aproximam aos intervalos extremos, mais forte a correlação é. Na Figura 3 percebemos que poucas variáveis possuem forte relação entre si.

Visualiza-se também que a variável de interesse, glicose (Y), possui relação positiva com todas as variáveis explicativas, no entanto elas são relações bem fracas (abaixo de 0.3), com exceção da Insulina (X4) onde há uma relação moderada (0.54). Isso pode significar que as características selecionadas para explicar o nível de glicose podem não ser as únicas características que expliquem esse fator a ser ajustado.

## 4 Modelagem

Após os dados tratados e analisados, segue-se com as etapas para a construção do modelo para a variável glicose.

[1]Para verificar quais as outras opções para substituir dados faltantes acesse a documentação do pacote Mice

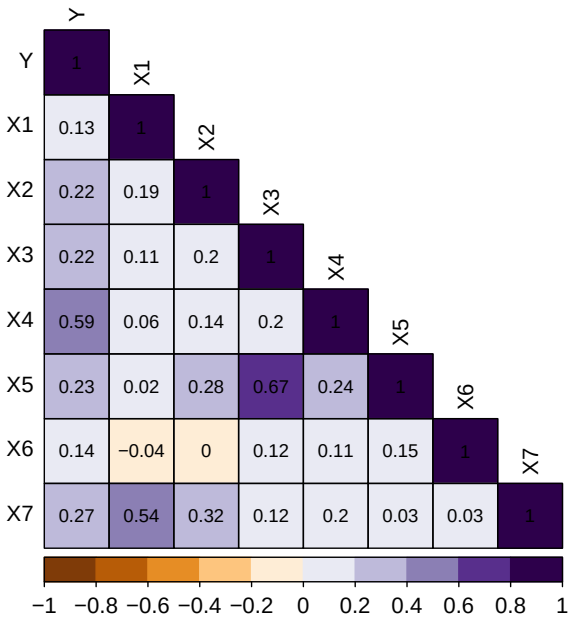


Figura 3: Matriz de Correlação das Variáveis

## Multicolinearidade

Antes de elaborar o primeiro modelo, é necessário verificar se as covariáveis são linearmente independentes entre si. Caso haja dependência, ou seja, combinações lineares entre as covariáveis, ocorre a redundância dessas variáveis no modelo, o que não contribui para a estimação dos coeficientes, já que essas estimativas podem não ser significativas, mesmo que a variável seja importante para explicar a variável resposta (aumento do erro tipo II).

Alguns métodos para averiguar a existência de multicolinearidade são: matriz de correlação, fator inflacionário da variância (VIF) e os autovalores da matriz de correlação.

### Matriz de correlação

A matriz de correlação vista na Figura 3 é útil para analisar se os dados assumem o pressuposto de não colinearidade, porém, é analisado somente as covariáveis, ou seja, desconsidera-se as correlações entre a variável Y. Caso alguma apresente um valor maior que 0.8 então há fortes indícios de multicolinearidade

### Fator Inflacionário da Variância

O VIF analisa o quanto a estimativa do coeficiente de uma variável explicativa é afetada pela combinação linear dessa variável com as demais. Se a variável apresenta pouca dependência linear, mais próximo de 1 o VIF fica, caso contrário o VIF tende ao infinito. Um ponto de corte recorrente para interpretar se a variável possui combinação linear é se o seu valor VIF é maior que 10.

Tabela 3: Fator Inflacionário da Variância

| X1 | X2 | X3 | X4 | X5 | X6 | X7 |
|---|---|---|---|---|---|---|
| 1.435 | 1.224 | 1.873 | 1.116 | 2.016 | 1.037 | 1.603 |

Pela Table 3 percebe-se que todas as variáveis explicativas ficam em torno de 1. Por meio desse resultado e do que foi observado pela matriz de correlação, conclui-se que não há fortes relações entre as covariáveis. Desse modo, o estimador para os coeficientes do modelo não é afetado pelas relações existentes entre as covariáveis, mesmo que existam.

## 1º Modelo

Com a multicolinearidade dos dados analisadas e não identificada, o próximo passo é construir o modelo inical. Este irá considerar interações 2 a 2, por se tratar de um conjunto de dados onde os efeitos das variáveis explicativas podem ser afetadas pelo efeito de covariável. Esta interação pode explicar melhor a nossa variável de interesse.

Tabela 4: Estimativas do 1º Modelo (inicialmente)

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| (Intercepto) | 37.6109 | 28.0712 | 1.3398 | 0.1807 | nao |
| X1 | 1.0709 | 2.3340 | 0.4589 | 0.6465 | nao |
| X2 | 0.3650 | 0.3735 | 0.9774 | 0.3287 | nao |
| X3 | 0.4085 | 0.7959 | 0.5133 | 0.6079 | nao |
| X4 | 0.3724 | 0.0578 | 6.4476 | 0.0000 | sim |
| X5 | 0.5052 | 1.0418 | 0.4849 | 0.6279 | nao |
| X6 | -51.2806 | 18.3811 | -2.7899 | 0.0054 | sim |
| X7 | 0.6073 | 0.6695 | 0.9071 | 0.3646 | nao |
| X1:X2 | -0.0034 | 0.0265 | -0.1284 | 0.8979 | nao |
| X1:X3 | -0.0260 | 0.0375 | -0.6946 | 0.4875 | nao |
| X1:X4 | 0.0032 | 0.0028 | 1.1118 | 0.2666 | nao |
| X1:X5 | 0.0008 | 0.0608 | 0.0125 | 0.9901 | nao |
| X1:X6 | 1.1549 | 0.9753 | 1.1841 | 0.2367 | nao |
| X1:X7 | -0.0274 | 0.0276 | -0.9937 | 0.3207 | nao |
| X2:X3 | -0.0020 | 0.0097 | -0.2009 | 0.8408 | nao |
| X2:X4 | -0.0011 | 0.0007 | -1.5296 | 0.1265 | nao |
| X2:X5 | -0.0033 | 0.0121 | -0.2761 | 0.7826 | nao |
| X2:X6 | 0.2745 | 0.2238 | 1.2263 | 0.2205 | nao |
| X2:X7 | 0.0022 | 0.0083 | 0.2688 | 0.7881 | nao |
| X3:X4 | -0.0002 | 0.0009 | -0.1846 | 0.8536 | nao |
| X3:X5 | -0.0101 | 0.0121 | -0.8330 | 0.4051 | nao |
| X3:X6 | -0.2276 | 0.3837 | -0.5932 | 0.5532 | nao |
| X3:X7 | 0.0109 | 0.0090 | 1.2031 | 0.2293 | nao |
| X4:X5 | -0.0019 | 0.0014 | -1.2998 | 0.1941 | nao |

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| X4:X6 | -0.0100 | 0.0164 | -0.6094 | 0.5424 | nao |
| X4:X7 | -0.0024 | 0.0006 | -3.8418 | 0.0001 | sim |
| X5:X6 | 1.3790 | 0.5319 | 2.5926 | 0.0097 | sim |
| X5:X7 | -0.0057 | 0.0180 | -0.3198 | 0.7492 | nao |
| X6:X7 | -0.1953 | 0.3028 | -0.6450 | 0.5191 | nao |

A tabela acima informa o valor das estimativas dos coeficientes de cada covariável e suas interações, junto com as informações relacionadas ao teste de significância da covariável, onde analisa se essas são estatisticamente influentes no modelo. Caso a hipótese nula seja rejeitada(P-Valor $< 1\%$), isso mostra que a covariável não impacta no modelo, pois o resultado do teste informa que a variável é estatisticamente igual a 0.

Percebe-se que a maioria dos coeficientes estimados desse modelo não são estatisticamente significantes para a variável de interesse (Y), ou seja, essa função possui variáveis que não impactam significativamente no nível da glicose. No entanto, escolher as variáveis que irão fazer parte do modelo, tanto retirar quanto adicionar, pode fazer com que as estimações dos coeficientes de outras variáveis passem a se tornar significantes para o modelo. Essa técnica de obter diferentes modelos a partir da escolha das variáveis é chamado de seleção de variáveis.

Existem esses métodos para selecionar variáveis:*backward*, *foward*, *stepwise* ou *all regression*.

Com a seleção de modelos realizada, basta compara-las e verificar qual apresenta melhor ajuste aos dados. Há métodos de comparação entre os modelos, que são: AIC, BIC ou coeficiente de determinação ($R^2$).

A função *gmulti()*, do pacote **glmulti**, consegue selecionar os modelos com o método de seleção *all regression* e classificá-los de acordo com o método de comparação escolhido. Neste caso foi selecionado o método AIC, ou seja, quanto menor o AIC, melhor o modelo.

Tabela 5: Estimativas do 1º Modelo (após seleção de variáveis)

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| (Intercepto) | 48.0216 | 8.3965 | 5.7193 | 0.0000 | sim |
| X2 | 0.3675 | 0.1185 | 3.1003 | 0.0020 | sim |
| X4 | 0.3695 | 0.0489 | 7.5579 | 0.0000 | sim |
| X6 | -33.4528 | 8.9279 | -3.7470 | 0.0002 | sim |
| X7 | 0.6941 | 0.1242 | 5.5879 | 0.0000 | sim |
| X2:X4 | -0.0009 | 0.0006 | -1.5321 | 0.1259 | nao |
| X4:X5 | -0.0023 | 0.0009 | -2.6078 | 0.0093 | sim |
| X6:X5 | 1.0952 | 0.2410 | 4.5436 | 0.0000 | sim |
| X4:X7 | -0.0022 | 0.0006 | -3.8547 | 0.0001 | sim |

Esse seria o primeiro modelo com ajustes proposto para explicar a variável de interesse. Verifica-se a redução do número de variáveis e a maioria das estiamtivas. Com o modelo gerado, basta verificar se ele se adequa aos pressupostos.

### Valores Extremos

Encontrar possíveis valores extremos que possam impactar nas estimativas dos coeficientes é fundamental para retirá-los ou transformá-los e assim evitar que influenciem o modelo.

Há 3 tipos de valores extremos:

- **Pontos Atípicos**: valores discrepantes dos resíduos estudentizados;
- **Pontos de Alavancagem**: valores discrepantes entre os valores da diagonal principal da matriz hat dos dados.
- **Pontos Influentes**: classificado de acordo com a distância de cook para cada observação.

O que mais impacta o modelo é o ponto influente, já que estes interferem bruscamente nos valores das estimativas.

As 3 medições podem ser resumidas em um único gráfico:

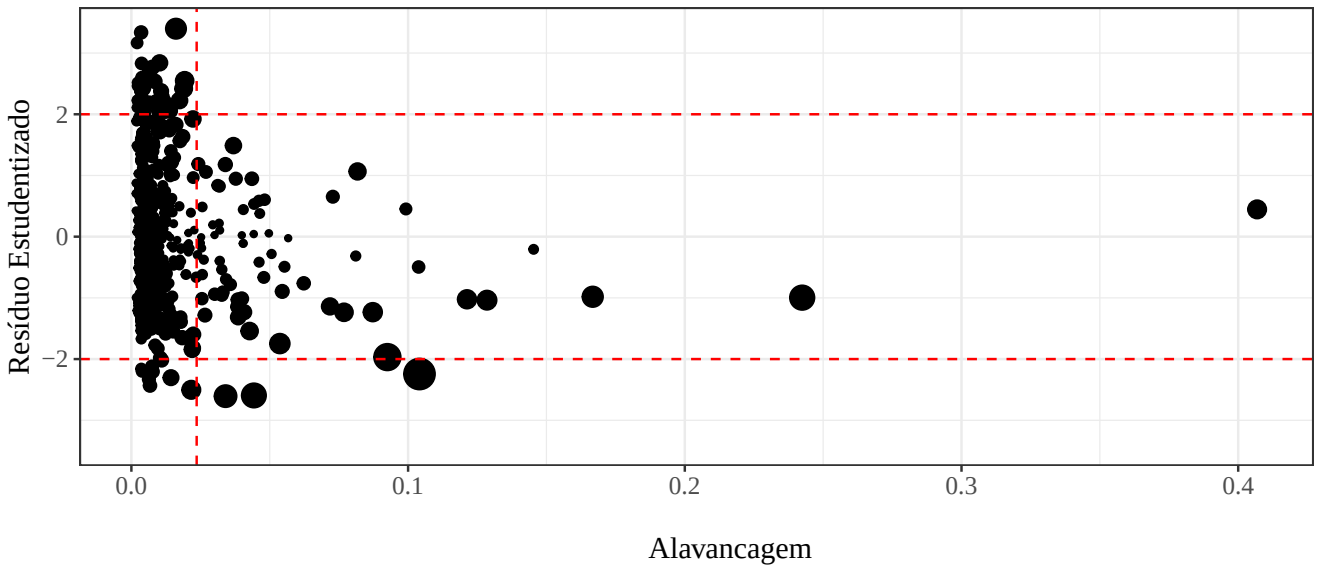


Figura 4: Gráfico Geral de Diagnóstico de Valores Extremos - 1º Modelo

Ele verifica tanto os pontos de alavancagem (eixo x), pontos atípicos (eixo y), e influentes (tamanho dos pontos). Há alguns pontos atípicos (valores maiores que 2 ou menores que -2) e pontos de alavancagem (valores maiores que 0.024)

Como este é um documento estático, não há a possibilidade de verificar de maneira interativa o tamanho dos pontos e se eles são ou não pontos influentes de acordo com a distância de cook. Para isso há a seguinte tabela que informa os pontos que são atípicos, de alavancagem e suas distâncias de cook:

Tabela 6: Pontos Atípicos e de Alavancagem do 1º Modelo

| | Resíduos | Diagonal da Matriz Hat | Distância Cook | São Pontos Influentes? |
|---|---|---|---|---|
| Obs.: 533 | -2.605678 | 0.0340317 | 0.0263753 | nao |

| | Resíduos | Diagonal da Matriz Hat | Distância Cook | São Pontos Influentes? |
|---|---|---|---|---|
| Obs.: 580 | -2.594357 | 0.0442713 | 0.0343808 | nao |
| Obs.: 600 | -2.243971 | 0.1041165 | 0.0646758 | nao |

Observa-se que mesmo tendo 3 pontos que sejam atípicos e de alavancagem, não há nenhuma observação de que esses pontos sejam pontos influentes, dado que o ponto de corte para a distância de cook deste modelo seja de 0.93.

### Normalidade

A fim de analisar se o modelo atende ao pressuposto de normalidade dos erros, foi construído os seguintes gráficos que apresentam informações sobre os resíduos:

No 1º gráfico da Figura 5 é construído o histograma dos valores dos resíduos com o intuito de verificar se a distribuição se assemelha a uma distribuição normal[2], onde, a primeira vista, percebe-se uma certa semelhança.

Já o 2º gráfico trata-se de um Q-Q plot, também conhecido como gráfico Quantil-Quantil, tendo como finalidade comparar duas distribuições por meio dos quantis dos valores observados com os quantis teóricos, neste caso os quantis teóricos se referem à distribuição normal.

A interpretação desse gráfico sugere que quanto mais os valores plotados permanecem sobre a reta, mais próximo da distribuição teórica os valores observados estão. Neste caso, no entanto, à medida que valores observados crescem, mais distante da reta os pontos estão, sendo um indicativo de que os erros/resíduos não seguem uma distribuição normal.

Como os gráficos apenas dão interpretações subjetivas para supor ou não a normalidade, os testes de hipóteses são utilizados para verificar, com um maior rigor, a normalidade do conjunto de dados.

Há diversos testes de hipóteses na literatura, cada um com suas características. Os detalhes das escolhas não serão apresentados, mas fica a título de curiosidade saber quais testes existem e como eles são estruturados. No geral, a hipótese nula supõe normalidade aos dados, enquanto a hipótese alternativa não garante essa característica. Desse modo, se o p-valor apresentado for menor que o nível de significância escolhido (1%, 5% ou 10%), deve-se rejeitar H0, ou seja, não há como supor a normalidade, caso contrário, pode-se supor a normalidade.

Tabela 7: Testes de Normalidade dos Resíduos - 1º Modelo

| | Estatística | P-Valor | Resultado |
|---|---|---|---|
| Asymptotic one-sample Kolmogorov-Smirnov test | 0.0626 | 0.005 | rejeita H0 |
| Lilliefors (Kolmogorov-Smirnov) normality test | 0.0626 | 0.000 | rejeita H0 |
| Cramer-von Mises normality test | 0.8320 | 0.000 | rejeita H0 |
| Shapiro-Wilk normality test | 0.9783 | 0.000 | rejeita H0 |
| Shapiro-Francia normality test | 0.9786 | 0.000 | rejeita H0 |

[2]Simétrica na média, ocasionando na congruência ou aproximação dos valores da média, mediana e moda.

| | Estatística | P-Valor | Resultado |
|---|---|---|---|
| Anderson-Darling normality test | 5.2002 | 0.000 | rejeita H0 |

Todos os resultados apresentados pelos testes de normalidade escolhidos rejeitam H0 a um nível de 1% de significância, ou seja, o indicativo é de que não há normalidade nos erros. Isso impacta na estimação dos parâmetros do 1º Modelo e uma alternativa para solucionar esse problema é aplicar a transformação box-cox na variável de interesse. A transformação box-cox é:

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

Essa transformação afeta……. pretende-se normalizar os erros.

Para obter o valor de $\lambda$, é utilizado o método de estimação por máximo verosimilhança (EMV). Utilizando a função *boxcox()* do pacote *MASS*, verificou-se $\lambda = 0.26$, portanto diferente de zero. Esse valor é o parâmetro que indica o poder da transformação de box-cox. A partir disso, $Y_*$ passará a ser a nossa variável de interesse.

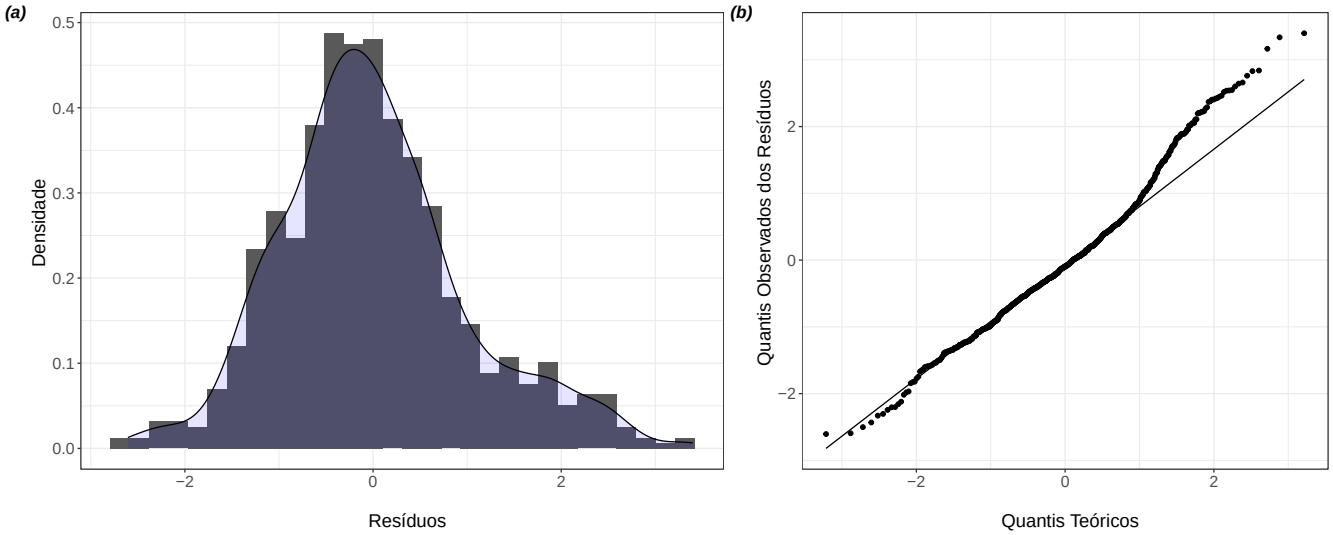


Figura 5: Gráficos para Verificação da Normalidade - 1º Modelo

## 2º Modelo

O próximo passo após aplicar a transformação box-cox é construir outro modelo a partir dos dados transformados. Sendo assim, as estimativas do 2º Modelo, que utiliza o $Y^*$ em vez do $Y$, são dadas por:

Tabela 8: Estimativas do 2º Modelo (inicialmente)

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| (Intercepto) | 7.0184 | 0.8111 | 8.6526 | 0.0000 | sim |
| X1 | 0.0298 | 0.0674 | 0.4422 | 0.6585 | nao |
| X2 | 0.0118 | 0.0108 | 1.0954 | 0.2737 | nao |
| X3 | 0.0100 | 0.0230 | 0.4357 | 0.6632 | nao |
| X4 | 0.0114 | 0.0017 | 6.8039 | 0.0000 | sim |
| X5 | 0.0178 | 0.0301 | 0.5899 | 0.5555 | nao |
| X6 | -1.4040 | 0.5311 | -2.6433 | 0.0084 | sim |
| X7 | 0.0160 | 0.0193 | 0.8287 | 0.4075 | nao |
| X1:X2 | -0.0002 | 0.0008 | -0.2769 | 0.7819 | nao |
| X1:X3 | -0.0006 | 0.0011 | -0.5687 | 0.5697 | nao |
| X1:X4 | 0.0001 | 0.0001 | 1.0819 | 0.2797 | nao |
| X1:X5 | 0.0001 | 0.0018 | 0.0310 | 0.9753 | nao |
| X1:X6 | 0.0336 | 0.0282 | 1.1917 | 0.2338 | nao |
| X1:X7 | -0.0007 | 0.0008 | -0.8301 | 0.4067 | nao |
| X2:X3 | 0.0000 | 0.0003 | -0.1502 | 0.8807 | nao |
| X2:X4 | 0.0000 | 0.0000 | -1.4690 | 0.1423 | nao |
| X2:X5 | -0.0001 | 0.0003 | -0.4169 | 0.6769 | nao |
| X2:X6 | 0.0073 | 0.0065 | 1.1285 | 0.2595 | nao |
| X2:X7 | 0.0001 | 0.0002 | 0.3848 | 0.7005 | nao |

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| X3:X4 | 0.0000 | 0.0000 | -0.4878 | 0.6258 | nao |
| X3:X5 | -0.0002 | 0.0003 | -0.6533 | 0.5138 | nao |
| X3:X6 | -0.0060 | 0.0111 | -0.5430 | 0.5873 | nao |
| X3:X7 | 0.0003 | 0.0003 | 1.1634 | 0.2450 | nao |
| X4:X5 | -0.0001 | 0.0000 | -1.3566 | 0.1753 | nao |
| X4:X6 | -0.0005 | 0.0005 | -0.9679 | 0.3334 | nao |
| X4:X7 | -0.0001 | 0.0000 | -4.3022 | 0.0000 | sim |
| X5:X6 | 0.0373 | 0.0154 | 2.4273 | 0.0155 | nao |
| X5:X7 | -0.0002 | 0.0005 | -0.3726 | 0.7096 | nao |
| X6:X7 | -0.0042 | 0.0087 | -0.4799 | 0.6314 | nao |

Percebe-se que a maioria das variáveis não são estatisticamente significantes, assim como no 1º modelo. Por conta disso deve-se realizar a seleção de variáveis para obter um modelo parcimonioso[3]. Com o mesmo método de seleção de variáveis do 1º Modelo, *all regression*, consegue-se os seguintes resultados:

Tabela 9: Estimativas do 2º Modelo (após seleção de variáveis)

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| (Intercepto) | 8.2108 | 0.1097 | 74.8261 | 0.0000 | sim |
| X4 | 0.0113 | 0.0014 | 8.1611 | 0.0000 | sim |
| X6 | -1.4419 | 0.3833 | -3.7617 | 0.0002 | sim |
| X4:X2 | 0.0000 | 0.0000 | -1.9813 | 0.0479 | nao |
| X4:X5 | -0.0001 | 0.0000 | -2.7802 | 0.0056 | sim |
| X6:X2 | 0.0087 | 0.0048 | 1.7900 | 0.0739 | nao |
| X6:X5 | 0.0267 | 0.0072 | 3.7060 | 0.0002 | sim |
| X2:X7 | 0.0002 | 0.0000 | 4.6626 | 0.0000 | sim |
| X7:X3 | 0.0001 | 0.0001 | 1.4454 | 0.1488 | nao |
| X4:X7 | -0.0001 | 0.0000 | -4.5828 | 0.0000 | sim |

Este modelo foi o melhor modelo após a transformação box-cox. Ele possui o AIC de 1572.58, bem menor se comparado ao modelo anterior.

#### Valores extremos

Verifica-se novamente, por meio dos resíduos, se há valores extremos que influenciam nas estimações do novo modelo.

Com o gráfico, aparentemente aparecem pontos que são atípicos(valores dos resíduos maiores que 2 ou menores que menos 2) e de alavancagem(valores da diagonal da matriz hat maiores que 0.0262). Resta saber se, pelo tamanho dos essa característica, já que seu histograma se assemelha com a densidade da

[3]Modelo com o menor número de variável significantes possíveis

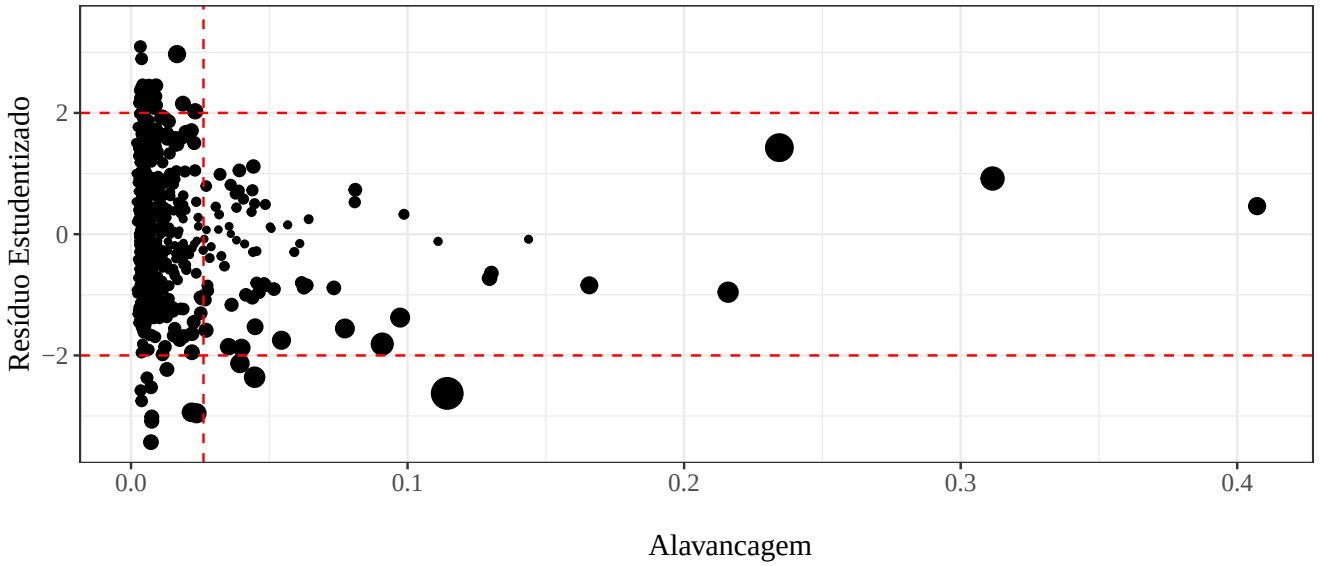


Figura 6: Gráfico Geral de Diagnóstico de Valores Extremos - 2º Modelo

normal e os pontos entre os resíduos e os valores ajustados estão mais alinhados no Q-Q plot se comparado com o 1º Modelo.

Tabela 10: Testes de Normalidade dos Resíduos - 2º Modelo

| | Estatística | P-Valor | Resultado |
|---|---|---|---|
| Asymptotic one-sample Kolmogorov-Smirnov test | 0.0340 | 0.3417 | não rejeita H0 |
| Lilliefors (Kolmogorov-Smirnov) normality test | 0.0343 | 0.0336 | não rejeita H0 |
| Cramer-von Mises normality test | 0.2068 | 0.0044 | rejeita H0 |
| Shapiro-Wilk normality test | 0.9938 | 0.0032 | rejeita H0 |
| Shapiro-Francia normality test | 0.9938 | 0.0039 | rejeita H0 |
| Anderson-Darling normality test | 1.4143 | 0.0012 | rejeita H0 |

Além dos gráficos, os testes de *Asymptotic one-sample Kolmogorov-Smirnov* e *Lilliefors (Kolmogorov-Smirnov)* não rejeitaram H0 a um nível de 1% de significância, ou seja, há evidências de que os resíduos do 2º Modelo apresentam uma distribuição normal. No entanto, é observado que os demais testes de normalidade não possuem sinais para não rejeitar a hipótese nula. Mesmo assim, continua-se com a análise dos pressupostos desse modelo.

### Homocedasticidade

Supor que a variância dos resíduos é constante faz parte das verificações de pressupostos do modelo, pois se o modelo não atende essa característica, pode impactar nas estimativas do modelo.

Para verificar a homocedasticidade dos erros, é utilizado o teste de *Breusch-Pagan*. Esse teste consiste em verificar se a variância dos erros dados os valores das covariáveis são constantes (hipótese nula) ou se eles variam (hipótese alternativa).

Com a função *bptest*, do pacote *lmtest*, consegue-se obter o resultado do teste para o 2º Modelo.

Tabela 11: Teste de Homocedasticidade

| | Estatística | P-Valor |
|---|---|---|
| Breusch-Pagan test | 23.63966 | 0.0049089 |

Como o P-Valor é menor que o nível de significância de 1%, há evidências que permite rejeitar a hipótese nula, indicando que a heterocedasticidade está presente no modelo.

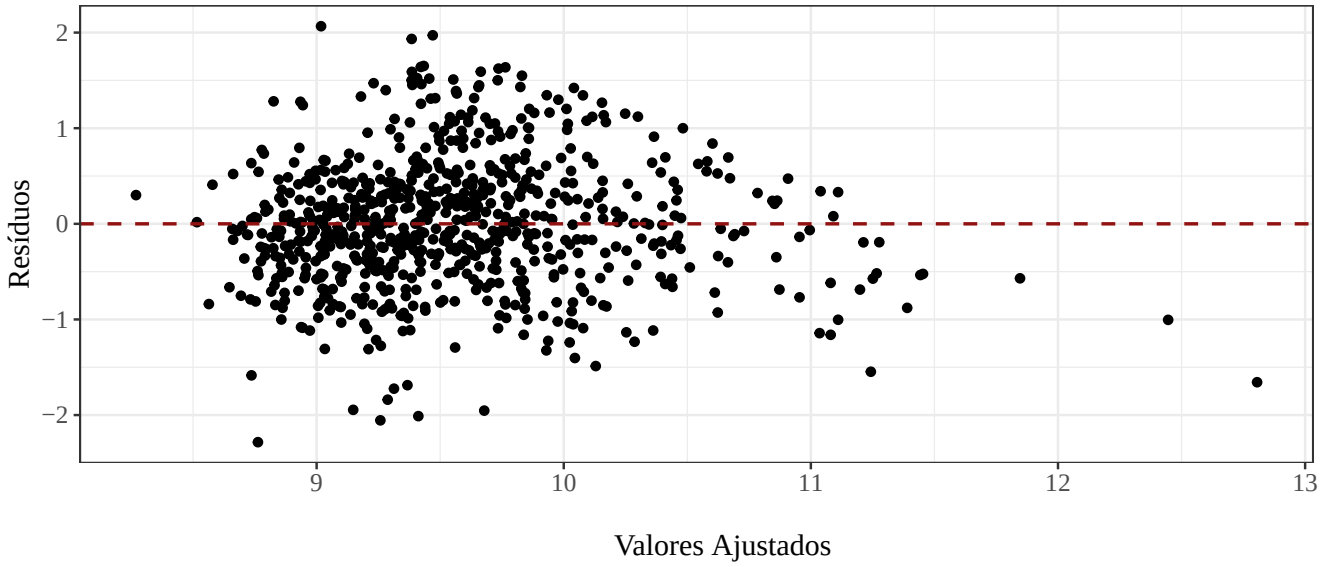


Figura 7: Gráfico da Dispersão dos Resíduos em Relação aos Valores Estimados - 2º Modelo

Esse gráfico corrobora com o resultado do teste, apresentando fortes indícios da presença de heterocedasticidade. Percebe-se que à medida que os valores estimados da variável de interesse aumentam, a variância dos resíduos fica menos dispersa em torno de 0, que é a média teórica dos erros e os valores dos resíduos inclinam-se para valores negativos. Por conta disso os resíduos aparentam não ter variância constante, não homogênea.

## 3º Modelo

Com o objetivo de incorporar a heterocedasticidade presente no 2º modelo, foi utilizado o Modelo Aditivo generalizado para localização, escala e forma (GAMLSS). Esse modelo atribui a variabilidade das variáveis explicativas ao ajuste do modelo, moldando conforme a heterocedasticidade dos dados.

No R, aplica-se a função *gamlss*, do pacote **gamlss**, para obter o novo modelo. Desse modo são obtidos os seguintes resultados:

Tabela 12: Estimativas de $\mu$ para o 3º Modelo

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| (Intercepto) | 8.1889 | 0.1086 | 75.4196 | 0.0000 | sim |
| X4 | 0.0113 | 0.0014 | 8.1649 | 0.0000 | sim |
| X6 | -1.4697 | 0.3804 | -3.8632 | 0.0001 | sim |
| X4.X2 | 0.0000 | 0.0000 | -2.2260 | 0.0263 | nao |
| X4.X5 | -0.0001 | 0.0000 | -2.5818 | 0.0100 | nao |
| X6.X2 | 0.0077 | 0.0049 | 1.5713 | 0.1165 | nao |
| X6.X5 | 0.0296 | 0.0072 | 4.1059 | 0.0000 | sim |
| X2.X7 | 0.0003 | 0.0000 | 6.1690 | 0.0000 | sim |
| X4.X7 | -0.0001 | 0.0000 | -4.0236 | 0.0001 | sim |

Tabela 13: Estimativas de $\sigma^2$ para o 3º Modelo

| | Estimativa | Desvio Padrão | T | P-Valor | Estatísticamente Significante? |
|---|---|---|---|---|---|
| (Intercepto) | -1.0450 | 0.1472 | -7.0990 | 0.0000 | sim |
| X5 | 0.0084 | 0.0039 | 2.1559 | 0.0314 | nao |
| X7 | 0.0106 | 0.0024 | 4.5089 | 0.0000 | sim |

As covariáveis utilizadas para o 3º Modelo de $\mu$ são as mesmas que a do 2º Modelo, porém as estimativas dos coeficientes sofreram alterações por conta do ajuste feito para considerar a heterocedasticidade constatadas no 2º Modelo.

Já para selecionar as variáveis estatisticamentes significativas para obter o modelo de $\sigma^2$, foram retiradas uma por uma as variáveis que apresentaram o P-Valor menor que 0.05. Com isso temos as estimativas dos valores de interesse onde $Y \sim N(\mu_i, \sigma_i^2)$

Após a incorporação da heterocedasticidade, também se faz necessário verificar se o modelo está adequado às suposições dos resíduos.

**Valores Extremos**

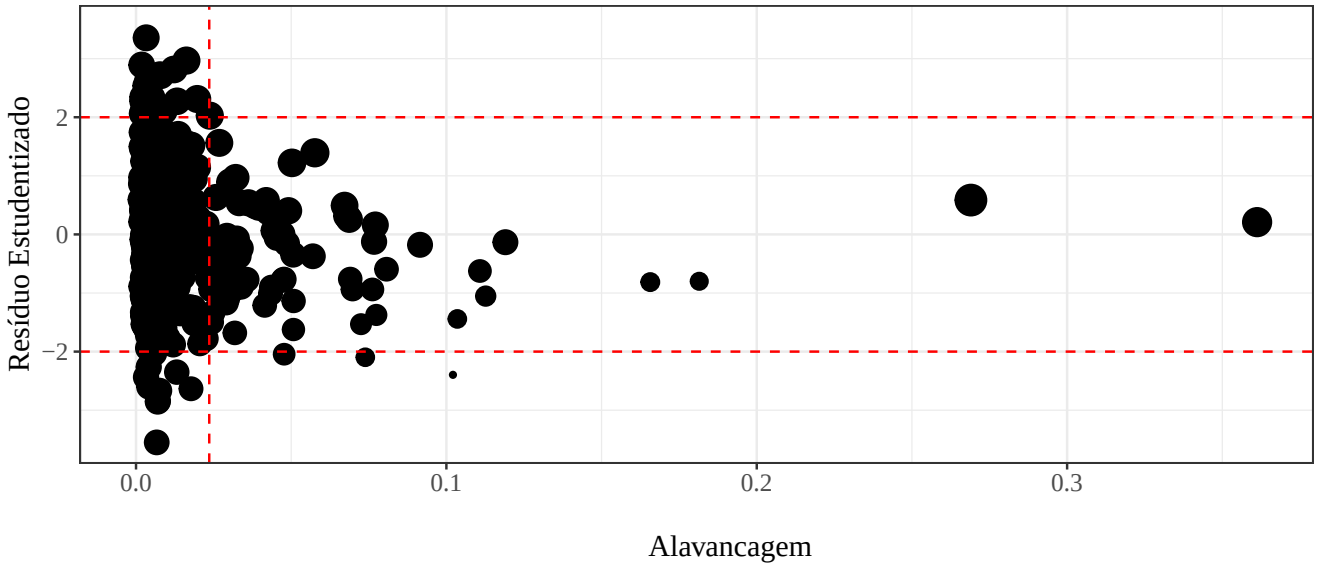


Figura 8: Gráfico Geral de Diagnóstico de Valores Extremos - 3º Modelo

Na Figura 8 é observável a existência de alguns pontos atípicos e de alavancagem.

Tabela 14: Pontos Atípicos e de Alavancagem do 2º Modelo

| | Resíduos | Diagonal da Matriz Hat | Distância Cook | São Pontos Influentes? |
|---|---|---|---|---|
| Obs.: 389 | -2.096620 | 0.0738765 | -0.0185830 | nao |
| Obs.: 545 | 2.026908 | 0.0237694 | 0.0054835 | nao |
| Obs.: 580 | -2.044855 | 0.0477263 | -0.0113872 | nao |
| Obs.: 600 | -2.396675 | 0.1021161 | -0.0302859 | nao |

As observações visualizadas na Figura 8 não são pontos influentes, pois a distância de cook é menor que o ponto de corte de 0.93. Portanto, não há outilers que impactam no 3º Modelo.

## Normalidade

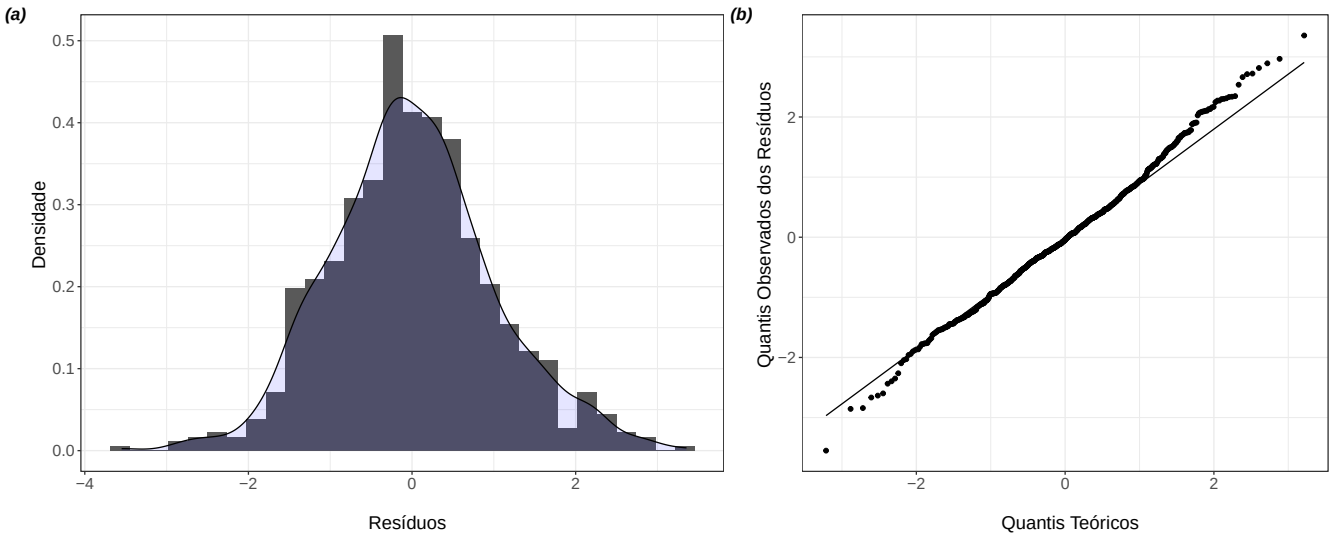


Figura 9: Gráficos para Verificação da Normalidade - 3º Modelo

Analisando o Q-Q plot do 3º Modelo, interpreta-se que os erros se aproximam da distribuição normal, mas ainda há valores que divergem do que seria estipulado pelo gráfico.

Tabela 15: Testes de Normalidade dos Resíduos - 3º Modelo

| | Estatística | P-Valor | Resultado |
|---|---|---|---|
| Asymptotic one-sample Kolmogorov-Smirnov test | 0.0347 | 0.3186 | não rejeita H0 |
| Lilliefors (Kolmogorov-Smirnov) normality test | 0.0351 | 0.0263 | não rejeita H0 |
| Cramer-von Mises normality test | 0.1721 | 0.0122 | não rejeita H0 |
| Shapiro-Wilk normality test | 0.9951 | 0.0146 | não rejeita H0 |
| Shapiro-Francia normality test | 0.9948 | 0.0121 | não rejeita H0 |
| Anderson-Darling normality test | 1.1350 | 0.0057 | rejeita H0 |

Diferente do modelos anterior, onde apenas os testes *Asymptotic one-sample Kolmogorov-Smirnov* e *Lilliefors (Kolmogorov-Smirnov)* não rejeitaram a hipótese nula, neste, os testes *Cramer-von Mises*, *Shapiro-Wilk* e *Shapiro-Franci* também supõe evidências para não rejeitar a hipótese nula, ou seja, a distribuição dos resíduos do 3º Modelo se aproxima da normal.

## Homocedasticidade

Devido ao fato da heterocedasticidade já ter sido incorporada neste modelo, não é necessário verificar se há a presença ou não de homocedasticidade na variabilidade dos resíduos.

### Autocorrelação

A autocorrelação dos erros analisa a existência de dependência entre os erros. Geralmente, esse pressuposto é verificado em modelos com dados temporais ou espaciais, já que estes podem ser influenciados por valores anteriores ou por valores próximos, respectivamente. O cenário estudado não se encaixa em nenhum desses casos, mas será realizado mesmo assim.

Um dos métodos utilizados para identificar a existência de correlação é aplicar o teste *Breusch-Godfrey*. Após a aplicação do teste, verificou que o p-valor (0.694) é maior que o nível de significância de 1%, dando evidências para não rejeitar a hipótese nula, ou seja, não há indícios de autocorrelação entre os erros neste modelo.

## 5 Conclusões

Portanto, o modelo proposto para utilizar que chegamos é o 3º Modelo. Para chegar a esse resultado foi necessário a manipulação da variável resposta (Glicose) por conta da não normalidade dos dados iniciais, devido a isso, a interpretação da variável resposta estimada é prejudicada. Também ocorreu a incorporação da heterocedasticidade no modelo. Ademais, não tivemos indícios de valores extremos, os dados não apresentam multicolinearidade entre as covariáveis e nem autocorrelação entre os erros.

$$Y \sim N(\mu_i, \sigma_i^2)$$

Para cada valor variável aleatória de interesse (Glicose), terá um valor que dependerá da média $\mu_i$ e $\sigma_i^2$.

Onde, pela Table 12 de estimativas do 3º Modelo:

$$\mu_i = B_0 + B_4X_4 + B_6X_6 + +B_{42}X_4X_2 + B_{45}X_4X_5 + B_{62}X_6X_2 + B_{65}X_6X_5 + B_{27}X_2X_7 + B_{47}X_4X_7$$

E pela Table 13 de estimativas das variâncias:

$$\sigma_i^2 = B_0 + B_5X_5 + B_7X_7$$

Pelo modelo de estimação para $\mu_i$, percebe-se que a quantidade de gestações das mulheres ($X_1$) e o nível de espessura da pele ($X_3$) não impactam tanto no nível de glicose ($Y$), já que elas não foram incorporadas ao modelo ajustado.

Já para a variância dos dados $\sigma_i^2$, o nível de glicose é impactado apenas pela variabilidade de do índice corporal ($X_5$) e pela idade ($X_7$) das mulheres.

Apesar disso, o modelo proposto pode não ser o mais adequado para a variável resposta, tendo em vista que o nível de glicose pode ter outros fatores que interferem na sua estimação que não foram adotadas no modelo, como visto nas Figura 3.

# 6 Referências

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press.

Montogomery, D.C., Peck, A.E., & Vining, G.G., (2012). Introduction to Linear Regression Analysis.

Johnson, R.A. & Wichern, D.W., (2007). Applied Multivariate Statistical Analysis.