Universidade de Brasília

DEPARTAMENTO DE ESTATÍSTICA

25 julho 2023

# Trabalho em Grupo - Regressão

Profª. Maria Teresa Costa Leão

Análise de Regressão Linear

Alunos:
Bruno Gondim Toledo | Matrícula: 15/0167636
Arthur Rodrigues | Matrícula: 19/0127376
Lucas Menezes | Matrícula: 18/0105418

# Introdução

O projeto em questão visa realizar análises estatísticas abrangentes sobre o mercado imobiliário em uma localizada não especificada, com foco nos preços de venda de casas. O objetivo principal é desenvolver um modelo preditivo capaz de estimar o preço de venda de residências com base em várias características relevantes da casa e do seu entorno.

As características que serão poderão utilizadas para realizar essas predições incluem o tamanho da casa em pés quadrados, o número de quartos e banheiros, a presença de ar condicionado, o tamanho da garagem, a existência ou ausência de piscina, o ano de construção, a qualidade da construção, o tamanho do terreno e a proximidade com a rodovia. Esses atributos foram selecionados com base na suposição de que eles podem exercer influência significativa sobre os preços de venda das casas.

O conjunto de dados utilizado no projeto compreende informações detalhadas sobre essas onze características de 522 casas.

Por meio da combinação de técnicas estatísticas avançadas e visualizações gráficas, espera-se obter insights valiosos sobre os principais fatores que impactam o preço de venda de casas na cidade em questão. Esses resultados podem ser úteis para compradores, vendedores e profissionais do mercado imobiliário, fornecendo informações sólidas para auxiliar em decisões de investimento e estratégias de negociação.

Além disso, a modelagem regressiva é interessante para determinar com quais características mínimas é possível obter com boa margem de confiança a resposta a que se quer chegar, no caso, o preço das casas; sem ter que coletar tantas variáveis, o que encarece o processo de levantamento de dados.

Para realizar as análises e visualizações dos dados, o software R na versão 4.1.3 foi escolhido como a ferramenta principal, além de ferramentas auxiliares como o SAS, o Overleaf e o rmarkdown, utilizado junto ao R no RStudio.

# Metodologia

## Boxplot

O boxplot é uma representação gráfica na qual se pode perceber de forma mais clara como os dados estão distribuídos.

A porção inferior do retângulo diz respeito ao primeiro quartil, enquanto a superior indica o terceiro quartil. Já o traço no interior do retângulo representa a mediana do conjunto de dados, ou seja, o valor em que o conjunto de dados é dividido em dois subconjuntos de mesmo tamanho. A média é representada pelo losango branco e os pontos são *outliers*. Os *outliers* são valores discrepantes da série de dados, ou seja, valores que não demonstram a realidade de um conjunto de dados.

## Teste de Normalidade de Shapiro-Wilk

O Teste de Shapiro-Wilk é utilizado para verificar a aderência de uma variável quantitativa ao modelo da Distribuição Normal, sendo mais recomendado para amostras pequenas. A suposição de normalidade é importante para a determinação do teste a ser utilizado. As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \text{A variável segue uma distribuição Normal} \\ H_1 : \text{A variável segue outro modelo} \end{cases}$$

# Modelo de regressão linear

Modelos de regressão linear são utilizados para fazer predições de uma variável explicativa quantitativa Y com base em valores de variáveis quantitativas ou qualitativas X. O modelo se diz linear pois os parâmetros interagem entre sí de maneira linear. Um modelo de regressão linear múltiplo generalizado é dado pela expressão abaixo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$$

Onde: $Y_i$ é o valor da variável resposta para uma dada observação; $\beta_0$ é o valor esperado de Y caso os valores de todas as variáveis explicativas sejam 0, também chamado de intercepto; $\beta_k$ é o aumento na variável resposta causado pelo aumento de uma unidade da variável $X_k$; $X_{ki}$ é o valor observado da k-ésima variável explicativa na i-ésima observação, essas variáveis podendo ser transformações de outras variáveis, contanto que $X_k$ sejam independentes; $\epsilon_i$ é o componente do erro aleatório.

Para o modelo de regressão linear, se assume que $\epsilon_i$ são independentes e identicamente distribuidos, seguindo distribuição normal com média 0 e variância constante $\sigma^2$.

A partir do modelo acima, um modelo pode ser ajustado para fazer predições da variável explicativa, dado pela expressão abaixo:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$$

Onde: $\hat{Y}_i$ é o valor predito da variável resposta para uma dada observação; $\hat{\beta}_0$ é o valor predito de Y caso os valores de todas as variáveis explicativas sejam 0, também chamado de intercepto; $\hat{\beta}_k$ é a estimativa do aumento na variável resposta causado pelo aumento de uma unidade da variável $X_k$; $X_{ki}$ é o valor observado da k-ésima variável explicativa na i-ésima observação, essas variáveis podendo ser transformações de outras variáveis, contanto que $X_k$ sejam independentes.

O modelo também pode ser escrito de forma matricial, dado pela expressão abaixo:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Onde: $\mathbf{Y}$ é o vetor das observações da variável resposta; $\mathbf{X}$ é a matriz das observações das variáveis explicativas; $\boldsymbol{\beta}$ é o vetor dos parâmetros; $\boldsymbol{\epsilon}$ é o vetor dos erros das observações.

O modelo para predições em forma matricial é dado por:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Onde: $\hat{\mathbf{Y}}$ é o vetor das predições da variável resposta; $\mathbf{X}$ é a matriz das observações das variáveis explicativas; $\hat{\boldsymbol{\beta}}$ é o vetor dos estimadores dos parâmetros.

# Estimativa dos Parâmetros

Os parâmetros dos modelos de regressão podem ser facilmente estimados em forma matricial. O vetor de estimadores de mínimos quadrados é dado pela seguinte formula:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Sendo:

- $\mathbf{Y}$ = vetor de observações da variável resposta
- $\mathbf{X}$ = matriz de observações do conjunto de variáveis explicativas

# Definição de termos relacionados a modelos de regressão linear

## Soma de Quadrados dos Resíduos (SSE)

Mede a variação total dos resíduos depois de ajustar um modelo de regressão para os dados. O cálculo de um resíduo é dado pela formula abaixo:

$$e_i = y_i - \hat{y}_i$$

Onde:

- $n$ = tamanho da amostra
- $y_i$ = i-ésimo valor da variável $Y$
- $\hat{y}_i$ = predição da variável $Y$ para a i-ésima observação

O SSE é dado então pela formula:

$$SSE = \sum_{i=1}^{n} e_i^2$$

Sendo:

- $n$ = tamanho da amostra
- $e_i$ = resídou para a i-ésimo observação

## Soma de Quadrados da Regressão (SSReg)

Mede a variação total explicada pelo conjunto de variáveis $\mathbf{X}$. O seu calculo é dado pela seguinte formula:

$$SSReg = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2$$

Onde:

- $n$ = tamanho da amostra
- $\bar{y}$ = média dos valores da variável $Y$
- $\hat{y}_i$ = predição da variável $Y$ para a i-ésima observação

## Soma de Quadrados Total (SSTO)

Mede a variação total da variável explicativa Y. É calculada segundo a seguinte formula:

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Sendo:

- $n$ = tamanho da amostra
- $y_i$ = i-ésimo valor da variável $Y$
- $\bar{y}$ = média dos valores da variável $Y$

Também é possível encontrar SSTO a partir de SSReg e SSE, seguindo a seguinte equação:

$$SSTO = SSReg + SSE$$

## Quadrado Médio dos Resíduos (MSE)

É uma estimativa não viesada da variância $\sigma^2$ dos erros $\epsilon_i$ nos modelos de regressão linear. É um componente utilizado no cálculo de várias estatísticas para testes relacionados à regressão. Seu calculo é dado pela seguinte formula:

$$MSE = \frac{SSE}{n - p}$$

Onde:

- $n$ = tamanho da amostra
- $p$ = número de parâmetros no modelo
- $SSE$ = soma de quadrados dos resíduos

## Quadrado Médio da Regressão (MSReg)

É um componente utilizado no cálculo de várias estatísticas para testes relacionados à regressão. Seu calculo é dado pela seguinte formula:

$$MSReg = \frac{SSReg}{p - 1}$$

Onde:

- $p$ = número de parâmetros no modelo
- $SSReg$ = soma de quadrados da regressão

## Resíduo Studentizado Externamente

Mede o afastamento de uma dado resíduo em relação aos demais resíduos. Serve para avaliar se uma observação é discrepante em termos da variável resposta. Seu cálculo é dado pela seguinte expressão:

$$t_i = e_i \left[ \frac{n - p - 1}{\sqrt{SSE\,(1 - h_{ii}) - e_i^2}} \right]^{1/2}$$

Sendo:

- $e_i$ = resídou para a i-ésimo observação

- $n$ = tamanho da amostra
- $p$ = número de parâmetros no modelo
- $SSE$ = soma de quadrados dos resíduos
- $h_{ii}$ = alavanca para a i-ésimo observação

## DFFITS

Mede a influência que uma observação tem na predição de seu própio valor. Serve para avaliar se uma observação é influente no calculo de seu valor predito. Seu calculo é dado pela formula:

$$DFFITS_i = t_i \left( \frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

Sendo:

- $t_i$ = resíduo studentizado externamente para a i-ésimo observação
- $h_{ii}$ = alavanca para a i-ésimo observação

## Distância de Cook (Cook's D)

Mede a influência que uma observação tem na predição de todos valores da variável resposta. Serve para avaliar se uma observação é influente no calculo de todos valores preditos. Seu calculo é dado pela formula:

$$D_i = \frac{e_i^2}{pMSE} \left( \frac{h_{ii}}{(1 - h_{ii})^2} \right)$$

Sendo:

- $e_i$ = resídou para a i-ésimo observação
- $p$ = número de parâmetros no modelo
- $MSE$ = quadrado médio dos resíduos
- $h_{ii}$ = alavanca para a i-ésimo observação

## DFBETAS

Mede a influência que uma observação tem na estimação dos parâmetros do modelo. Serve para avaliar se uma observação é influente no calculo dos estimadores do modelo. Seu calculo para um dado parâmetro k é dado pela formula:

$$\left(DFBETAS_k\right)_i = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

Sendo:

- $\hat{\beta}_k$ = estimativa do k-ésimo parâmetro
- $\hat{\beta}_{k(i)}$ = estimativa do k-ésimo parâmetro sem a i-ésima observação
- $MSE_{(i)}$ = quadrado médio dos resíduos sem a i-ésima observação
- $c_{kk}$ = k-ésimo valor na diagonal da matriz $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$, onde:
  - $\mathbf{X}$ = matriz de observações do conjunto de variáveis explicativas

# Métodos de Seleção de Modelos

## Coeficiente de Determinação ($R^2$)

Caclula quanto que a informação de uma variável ou um conjunto de variáveis explicativas **X** reduz a variância de uma variável resposta Y. É utilizado para medir quão bem as variáveis explicativas **X** explicam a variável resposta Y. Procura-se um modelo com um alto coeficiente de determinação. Em estudos de regressão, seu caclulo é dado pela seguinte expressão:

$$R^2 = 1 - \frac{SSE}{SSTO}$$

Sendo:

- $SSE$ = soma de quadrados dos resíduos
- $SSTO$ = soma de quadrados total

## Coeficiente de Determinação Ajustado ($R_a^2$)

Estimativa para o grau de explicação do modelo. Similar ao coeficiente de determinação, mas penalizado pela inclusão de novas variáveis no modelo. Procura-se um modelo com um alto coeficiente de determinação ajustado. Seu caclulo é dado pela seguinte expressão:

$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}$$

Sendo:

- $n$ = tamanho da amostra
- $p$ = número de parâmetros no modelo
- $SSE$ = soma de quadrados dos resíduos
- $SSTO$ = soma de quadrados total

## $C_p$ de Mallows

Mede grau de ajuste de um modelo, levando em consideração o número de parâmetros no modelo. Procura-se um modelo com um $C_p \approx p$, sendo p o número de parâmetros no modelo. O seu cálculo é dado pela expressão:

$$C_p = \frac{SSE_p}{MSE} - (n - 2p)$$

Sendo:

- $n$ = tamanho da amostra
- $p$ = número de parâmetros no modelo testado
- $SSE_p$ = soma de quadrados dos resíduos do modelo com p variáveis
- $MSE$ = quadrado médio dos resíduos do modelo com todas variáveis

## Métodos de seleção automática

São algoritmos iterativos que buscam o melhor modelo com base em uma serie de testes F de ausência de regressão e testes T de ausência de um parâmetro. Os métodos visam um modelo em que todas as variáveis inclusas foram significantes, ou todas as retiradas foram não significantes em ao menos uma iteração. Para todos os métodos, não há garantia que o melhor modelo será encontrado.

#### Método Backward

Inicia-se com todas as variáveis no modelo. Depois, faz-se um teste F de ausência de regressão e se significativo, fazem-se testes T de ausência de um parâmetro para todas as variáveis incluídas, e se algum teste for não significativo, retira-se a variável associada ao teste com o maior p-valor, e se repetem os testes com o modelo novo. Se repete o procedimento até que todos p-valores sejam significativos.

#### Método Forward

Inicia-se com nenhuma variáveis no modelo. Depois, faz-se um teste F de ausência de regressão e se significativo, fazem-se testes T de ausência de um parâmetro para todas as variáveis não incluídas, e se algum teste for significativo, inclui-se a variável associada ao teste com o menor p-valor, e se repetem os testes com o modelo novo. Se repete o procedimento até que todos p-valores sejam não significativos.

#### Método Stepwise

Inicia-se com nenhuma variáveis no modelo. Depois, faz-se um teste F de ausência de regressão e se significativo, fazem-se testes T de ausência de um parâmetro para todas as variáveis não incluídas, e se algum teste for significativo, inclui-se a variável associada ao teste com o menor p-valor. Em seguida, faz-se um teste F de ausência de regressão e se significativo, fazem-se testes T de ausência de um parâmetro para todas as variáveis incluídas, e se algum teste for não significativo, retira-se a variável associada ao teste com o maior p-valor. Repetem-se os testes com o modelo novo até que nenhuma variável nova entre e nenhuma antiga saia. Esse método funciona como uma junção do método *Forward* e o método *Backward*.

# Padronização de Variáveis

Para o modelo de regrassão seguir todos os seus pressupostos, as vezes é necessário reparametrizar ou padronizar algumas variáveis, para evitar problemas de não linearidade entre as variáveis explicativas e a variável resposta ou até mesmo de multicolinearidade entre as variáveis explicativas.

## One Hot Encoding

Esse procedimento é utilizado para transformação de variáveis qualitativas. Supondo que a variável qualitativa possui j categorias, são criadas j novas variáveis indicadoras para serem utilizadas no modelo, essas assumindo valores de 0 ou 1. Essas variáveis indicam se a observação pertence à uma categoria por receber um valor de 1 na variável indicativa associada à categoria, e um valor de 0 nas demais categorias. Podem ser criadas também j-1 novas variáveis, e designar uma categoria como a categoria padrão, representada quando todas as variáveis indicativas recebem um valor de 0.

# Testes de hipóteses

## Teste de Ausência de Regressão

Testa se as variáveis explicativas $X_k$ tem influência na variável resposta Y.

Hipóteses:

$$\begin{cases} H_0 : \beta_k = 0, k = 1, 2, ..., p-1 \\ H_1 : \exists \beta_k \neq 0, k = 1, 2, ..., p-1 \end{cases}$$

Estatística do Teste:

$$F = \frac{MSReg}{MSE} \sim F_{p-1,n-p}$$

Sendo:

- $MSReg$ = quadrado médio da regressão
- $MSE$ = quadrado médio dos resíduos

### Teste T de Ausência de um Parâmetro

Testa se uma variável explicativa $X_k$ tem influência na variável resposta Y.

Hipóteses:

$$\begin{cases} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{cases}$$

Estatística do Teste:

$$T = \frac{\hat{\beta}_k}{S(\hat{\beta}_k)} \sim t_{n-p}$$

Sendo:

- $\hat{\beta}_k$ = estimativa do k-ésimo parâmetro
- $S(\hat{\beta}_k)$ = raiz quadrada da variação da estimativa do k-ésimo parâmetro

# Análises exploratórias

## Preço de venda da casa

A fim de analisar o comportamento da variável $X_1$ (tamanho da casa) dispõe-se o gráfico e o quadro a seguir:

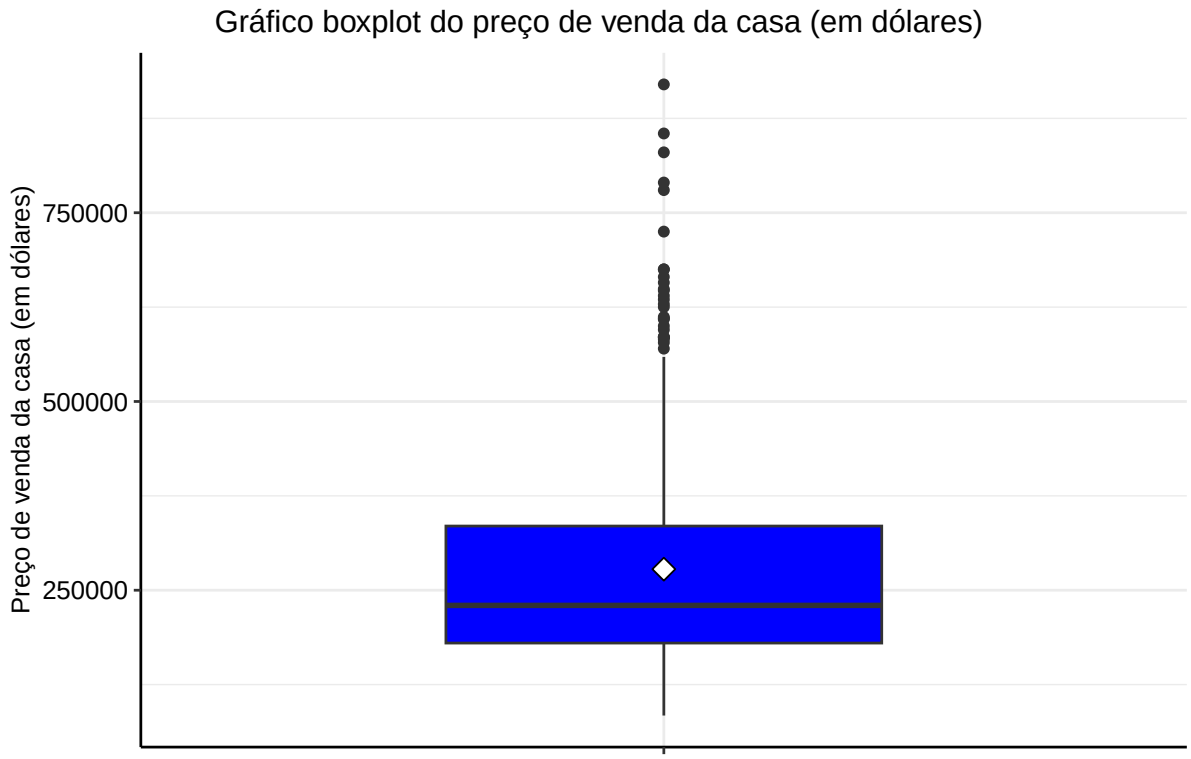



Table 1: Medidas Resumo do preço de venda da casa (em dólares)

| Estatística | Valor |
|---|---|
| **Minimo** | 84000 |
| **Primeiro Quartil.** | 180000 |
| **Mediana** | 229900 |
| **Média** | 227894 |
| **Terceiro Quartil** | 335000 |
| **Maximo** | 92000 |
| **Desvio Padrão** | 137923 |

Diante da Figura 1 e do Quadro 1, conclui-se que o menor preço de venda foi de 84 mil dólares, enquanto que a casa mais cara saiu no valor de 920 mil dólares. Além disso, o preço médio das casas foi $277.894,00 dólares.É importante notar que a média pode ser influenciada pelos valores extremos, ou seja, pelos outliers presentes no limite superior na distribuição dos preços. Dito isso, para analisar o preço médio, pode ser utilizado a mediana, tendo em vista que essa medida é resistente à outliers e separa os dados da distribuição ao meio. Dessa forma, a mediana foi de $229.900,00 dólares.

Vale ressaltar que 75% das casas foram vendidas por um preço igual ou inferior a $335.000,00 dólares, que pode ser visto atráves do terceiro quartil, Já o desvio padrão, apresentou um valor de $137.923, mostrando que há uma grande variabilidade nos preços das casas, que pode ser notada pela amplitude do boxplot.

## Tamanho da casa

Para analisar o comportamento da variável $X_2$(tamanho da casa) dispõe-se o gráfico e o quadro a seguir:

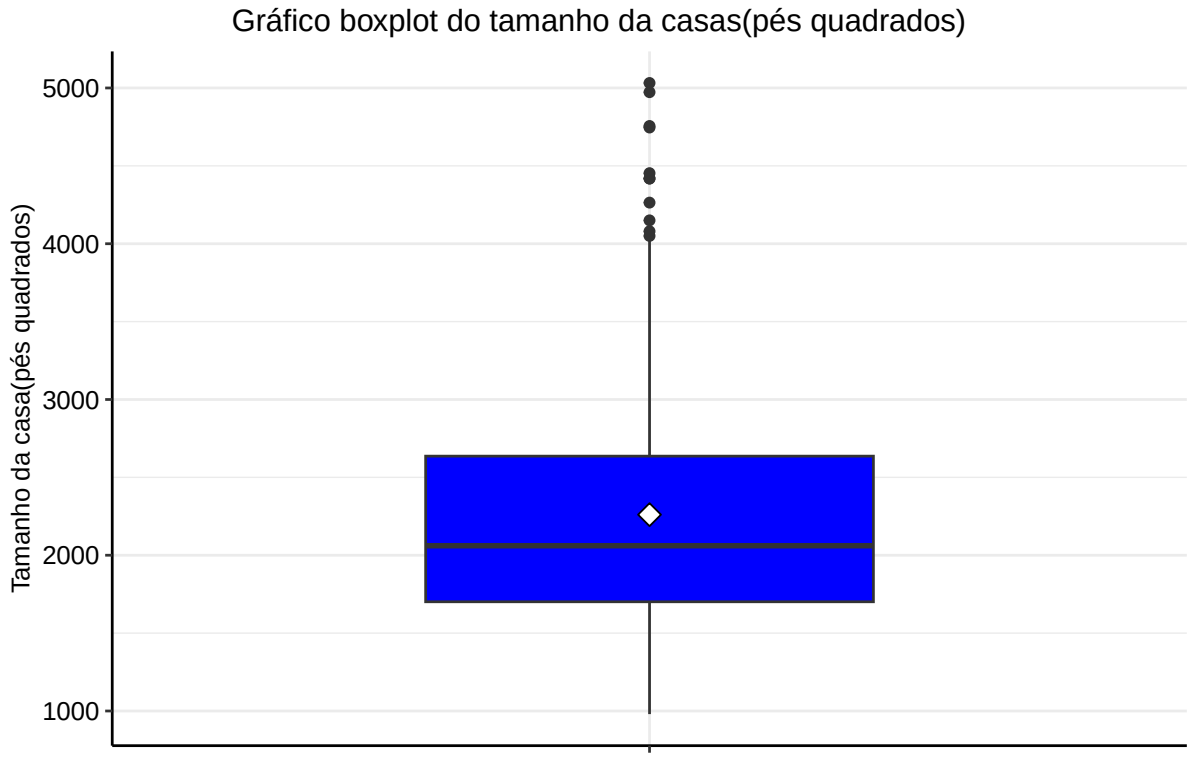



Table 2: Medidas Resumo tamanho das casas (pés quadrados)

| Estatística | Valor |
|---|---|
| **Minimo** | 980.00 |
| **Primeiro Quartil.** | 1701.00 |
| **Mediana** | 2061.00 |
| **Média** | 2261.00 |
| **Terceiro Quartil** | 2636.00 |
| **Maximo** | 5032.00 |
| **Desvio Padrão** | 711.07 |

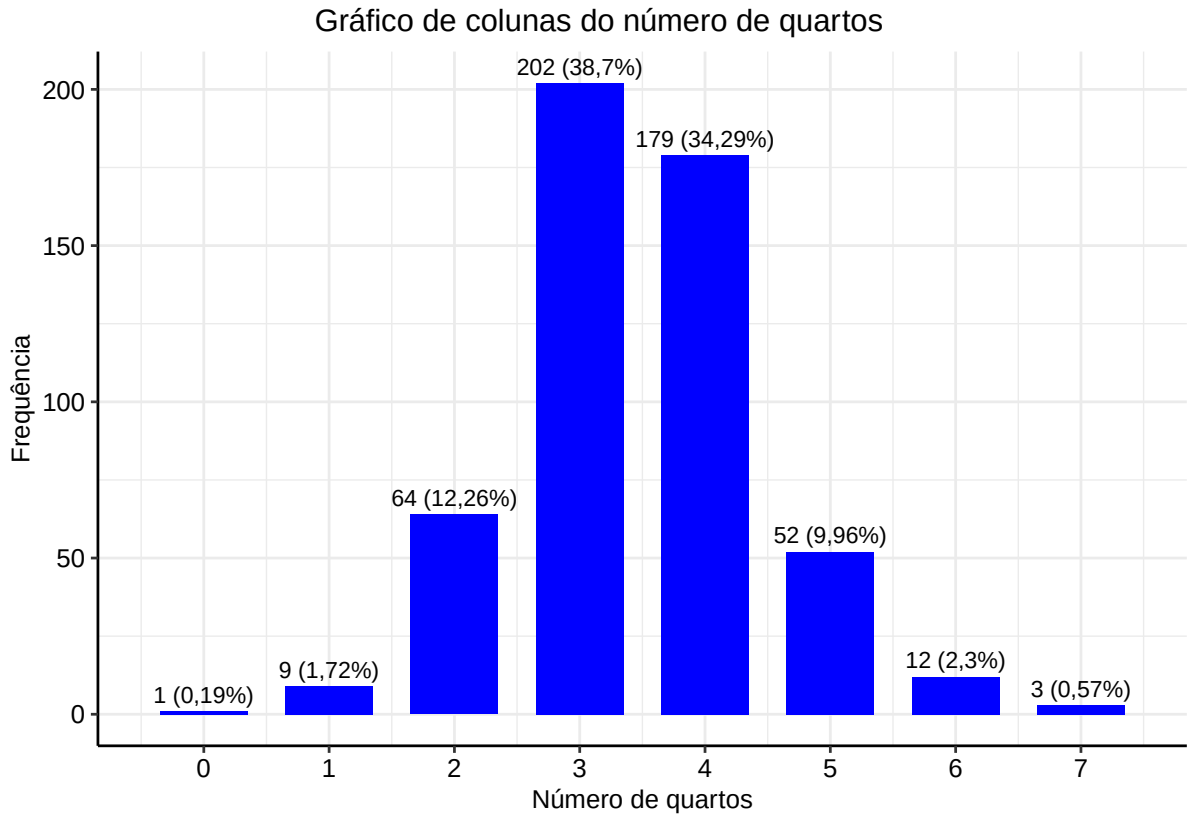
Com base na Figura 1 e no Quadro 1, conclui-se que a média de tamanho das casas é de aproximadamente 2261 pés quadrados.Ademais, o valor mínimo é de 980 pés quadrados e o valor máximo é de 5032 pés quadrados, mostrando que a amostra abrange desde casas pequenas até propriedades bastante extensas.

Nesse sentido, destaca-se a amplitude do tamanho das casas, que é a diferença entre o maior e o menor valor observado. Neste caso, a amplitude é de 4052 pés quadrados, mostrando uma grande variação dos no tamanho das casas que pode ser explicada pelo desvio padrão, que é de aproximadamente 711,07 pés quadrados.

Além disso, a mediana, por sua vez, é de 2061 pés quadrados, indicando uma assimetria à direita na distribuição dos dados,visto que a média é maior que a mediana. Essa assimetria pode ser explicada pela presença de valores maiores (outliers) que puxam a média para cima.

## Número de quartos

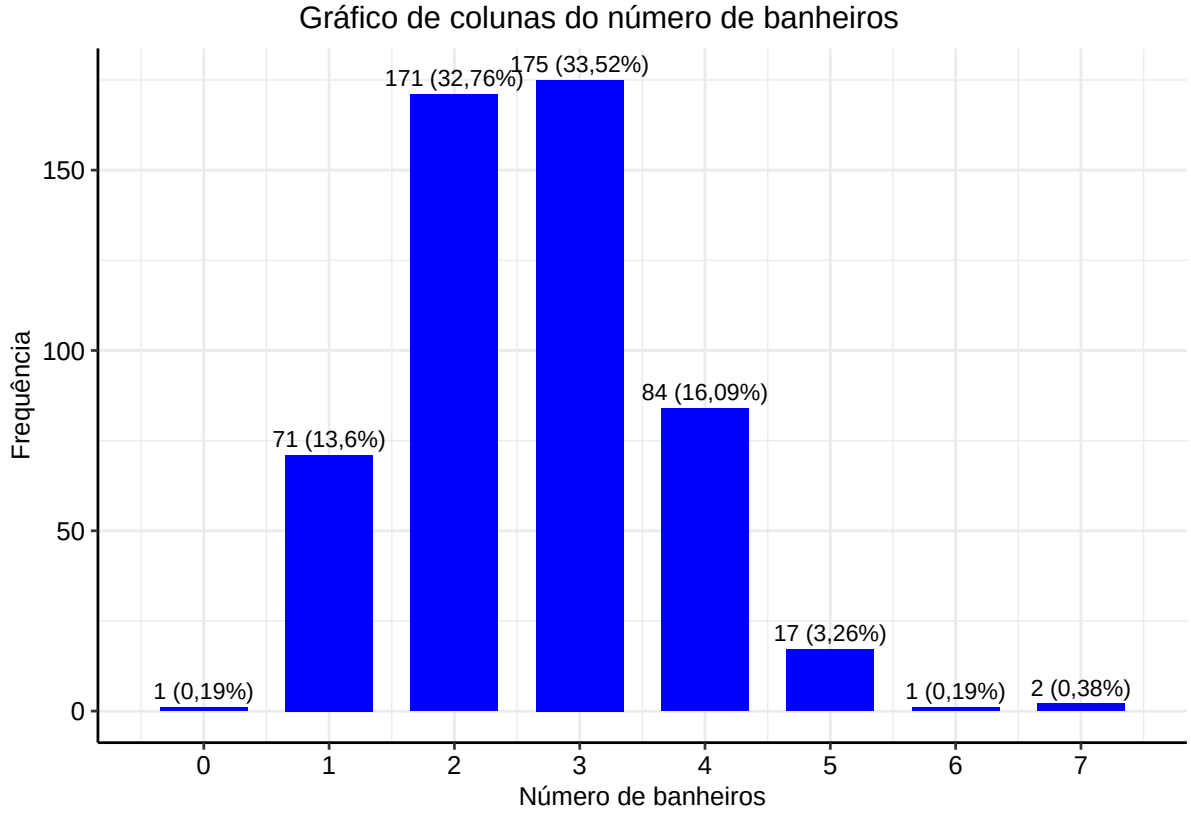
A análise descritiva da variável $X_3$ (Número de quartos) é ilustrada na figura abaixo, apresentando um gráfico de colunas. Esse gráfico oferece uma representação visual dos dados relacionados ao número de quartos nas residências estudadas.




A partir da Figura 2, observa-se que a maioria das casas construídas entre os anos de 1885 e 1998 possuía entre 3 e 4 quartos, representando quase 73% do total. Além disso, foi identificado que apenas 1 casa (0,19%) não possuía nenhum quarto, enquanto em apenas 3 casas (0,57%) foram encontrados sete quartos

## Número de banheiros

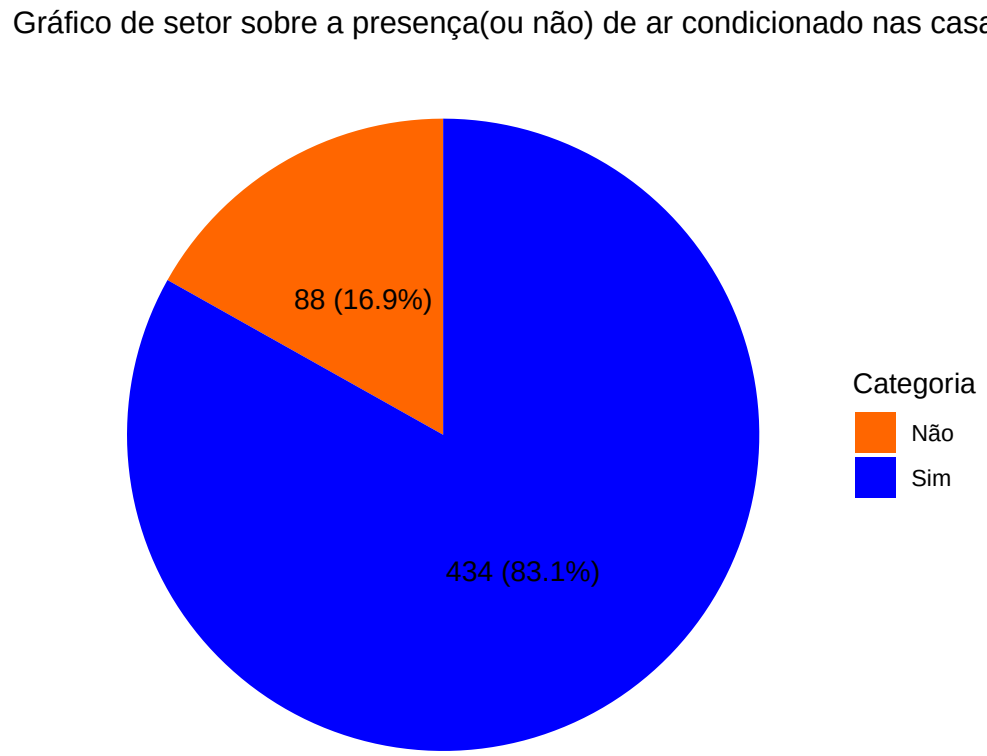
Este estudo visa analisar a variável $X_4$, que representa o número de banheiros em uma casa. Para facilitar a compreensão dos dados, a figura a seguir apresenta uma visualização gráfica da distribuição dessa variável.




Com base na Figura 3, podemos observar que a maioria das casas construídas entre os anos de 1885 e 1998 possui entre 2 e 3 banheiros, representando 66,28% do total. Em contrapartida, 16% das casas possuem 4 banheiros, enquanto 13,6% contam com apenas 1 banheiro. É importante destacar que uma parcela inferior a 0,5% possui entre 6 e 7 banheiros, e somente 1 casa (0,19%) não possui banheiro.

## Presença de ar condicionado

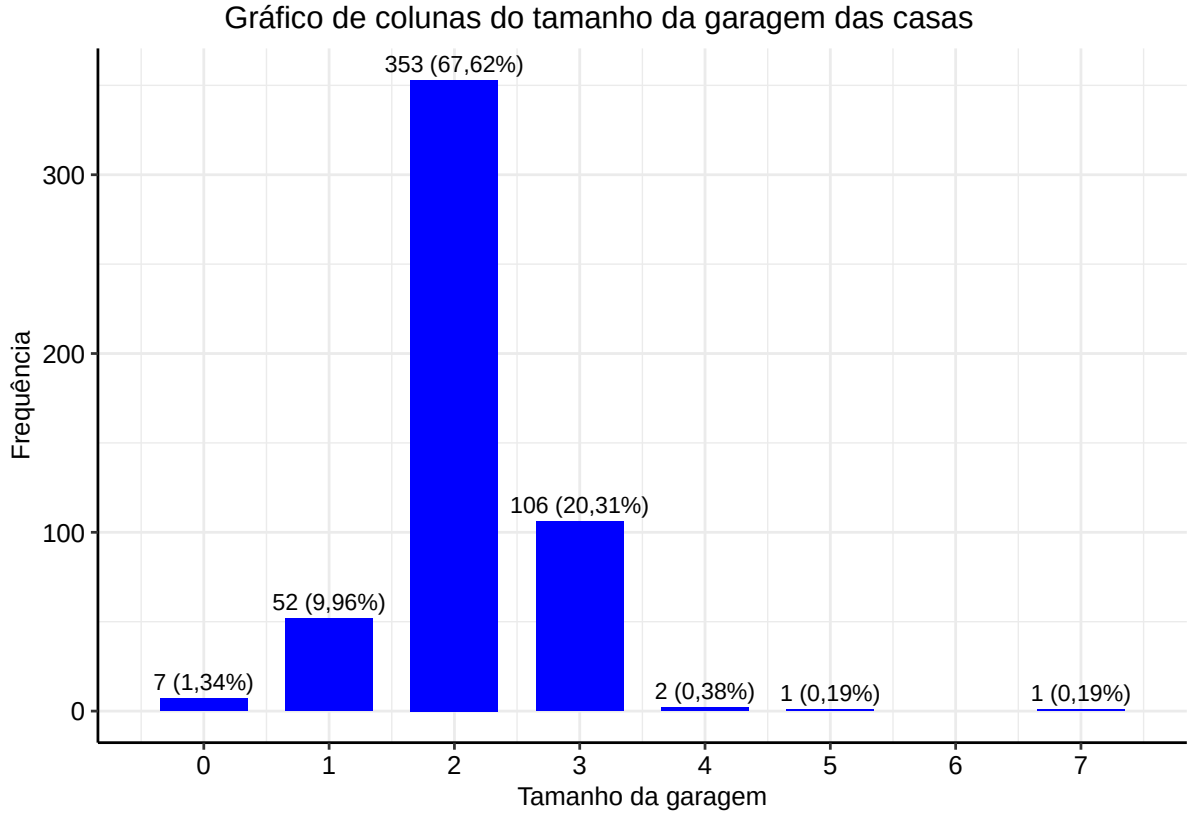
A análise descritiva da variável $X_5$, que representa a presença(ou não) de ar condicionado nas casas, é ilustrada através da figura apresentada abaixo, a qual consiste em um gráfico de setor.




Diante da figura 4, constata-se que 83% das casas construídas nos anos de 1885 a 1998 não tinham ar condicionado.

## Tamanho da garagem

Nesta análise descritiva, vamos examinar a variável $X_6$, que representa o tamanho da garagem, ou seja, o número de carros que podem ser guardados na garagem. A seguir, apresenta-se a figura que ilustra essa variável.

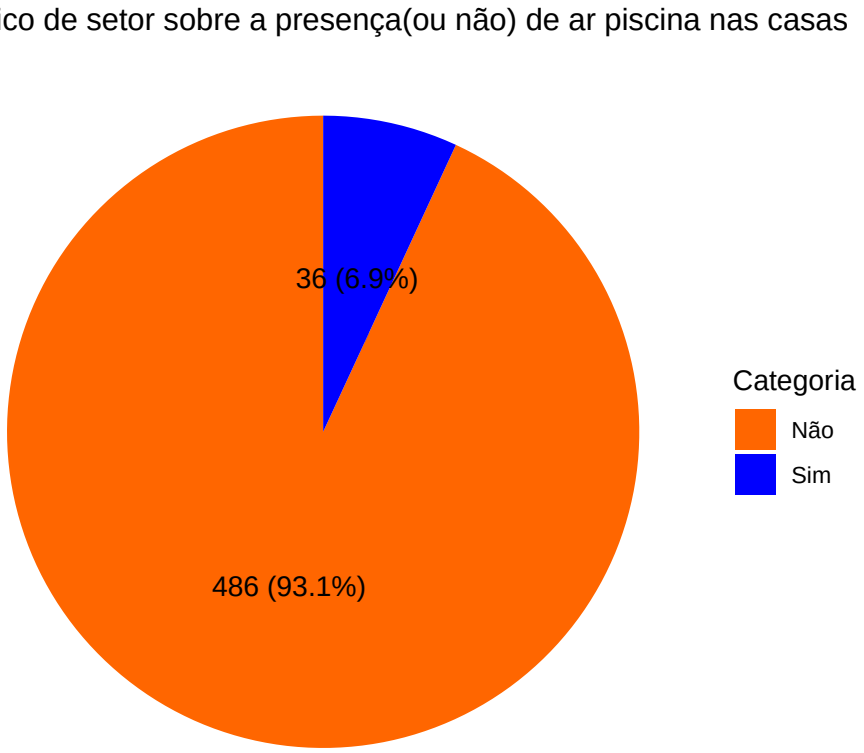



De acordo com a Figura 5, conclui-se que a maioria das casas construídas (cerca de 67,62%) tem espaço para aproximadamente 2 carros em suas garagens. Além disso, cerca de 20% das casas têm capacidade para acomodar 3 carros.Por outro lado, casas com espaço para apenas 1 carro representam menos de 10% do total. Destaca-se também as casas com espaço para 4, 5 ou 7 carros que são ainda mais raras, compreendendo menos de 1% do total cada.

Vale a pena ressaltar que existem mais casas sem garagem (cerca de 1,34% do total) do que casas com capacidade para comportar entre 4 a 7 carros juntas.

## Presença de piscina

Com o objetivo de examinar a variável $X_7$, que indica se as casas possuem piscina ou não, foi realizada uma análise descritiva, cujos resultados são apresentados na figura abaixo.




Atráves da Figura 6, constata-se que mais de 90% das casas construídas entre os anos de 1885 e 1998 possuem piscinas, ao mesmo tempo que apenas 6,9% não usufruem disso.

## Idade das casas

No banco de dados original, a coluna $X_8$ constava o ano de construção da casa. Esta é uma informação potencialmente útil para a análise, porém carecendo de uma transformação para fazer sentido numéricamente. Para tal, transformamos a informação contida na variável para 'Idade da casa', considerando a amplitude abrangida pelo banco de dados. Considerando a casa mais nova presente no banco de dados (Ano 1998) como idade 0, e a casa mais antiga da casa (Ano 1885) como idade 113; e todas as demais seguindo o mesmo padrão.

A fim de examinar a distribuição da variável $X_8$, que agora representa a idade das casas em anos, foram elaborados um boxplot e um quadro com medidas resumo.

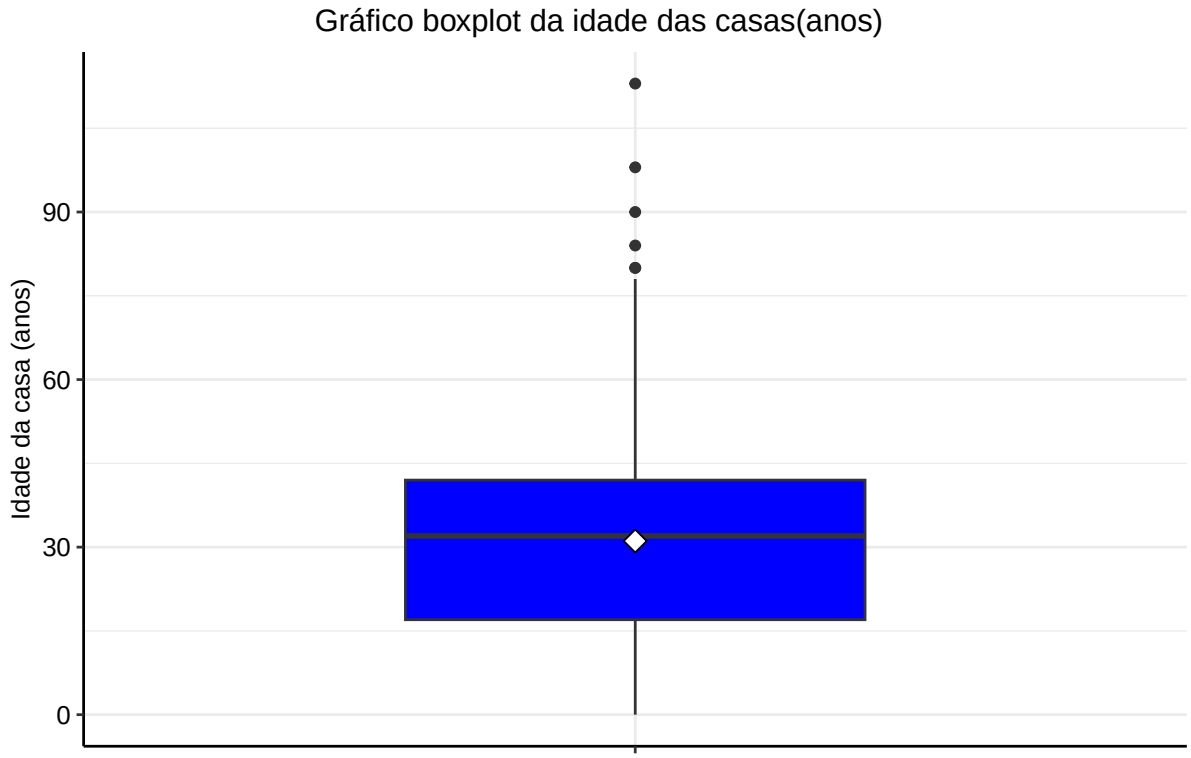



Table 3: Medidas Resumo da idade das casas(anos)

| Estatística | Valor |
|---|---|
| **Minimo** | 0.00 |
| **Primeiro Quartil.** | 17.00 |
| **Mediana** | 32.00 |
| **Média** | 31.10 |
| **Terceiro Quartil** | 42.00 |
| **Maximo** | 113.00 |
| **Desvio Padrão** | 17.64 |

Analisando a Figura 7 e no Quadro 2 constata-se que a idade média das casas e de 31,1 anos. Além disso, a casa mais nova tem 0 anos, ou seja, foi construída em 1998, enquanto que a casa mais velha tem 113 anos (1885). No entanto, a presença de outliers no limite inferior, pode estar puxando a média para baixo, tornando-a menor que a mediana. Nesse caso, a média não é a melhor medida de tendência central, pois é influenciada por valores extremos. Dessa forma, a mediana é a melhor medida para representar a idade média das casas, já que não é afetada por outliers.

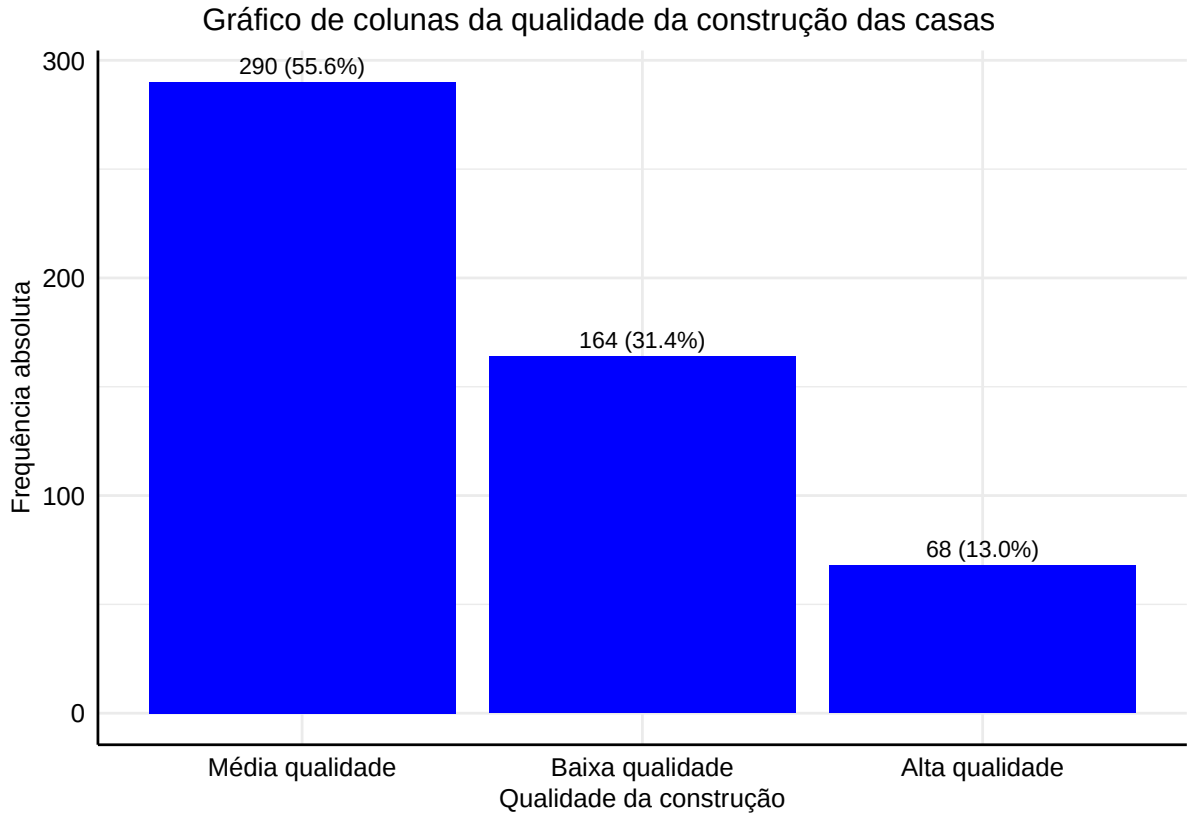
Nesse contexto, a mediana está em torno de 32 anos, ou seja, metade das casas tem uma idade menor ou igual a 32 anos, enquanto a outra metade tem uma idade maior ou igual a esse valor. Vale ressaltar

também que o valor do terceiro quartil é de 42 anos, indicando que 75% das casas têm uma idade menor ou igual a 42 anos.

A amplitude dos da idade das casas é de 113 anos, mostrando uma grande variação da idade que é explicada pelo O desvio padrão de 38.99 anos, indicando uma dispersão das idades em relação à média.

## Qualidade da construção

Com o intuito de analisar a variável $X_9$, que representa a qualidade da construção,dispõe-se a figura abaixo:




Baseando-se na Figura 8, comprova-se que mais da metade das casas (55,6%) foram construídas com uma qualidade média. Enquanto isso, 31% das casas têm uma baixa qualidade e somente 13% delas são consideradas de alta qualidade.

## Tamanho do terreno

Com o propósito de realizar uma análise descritiva da variável $X_{10}$ que representa o tamanho do terreno (pés quadrados), foram apresentadas a figura e o quadro abaixo:

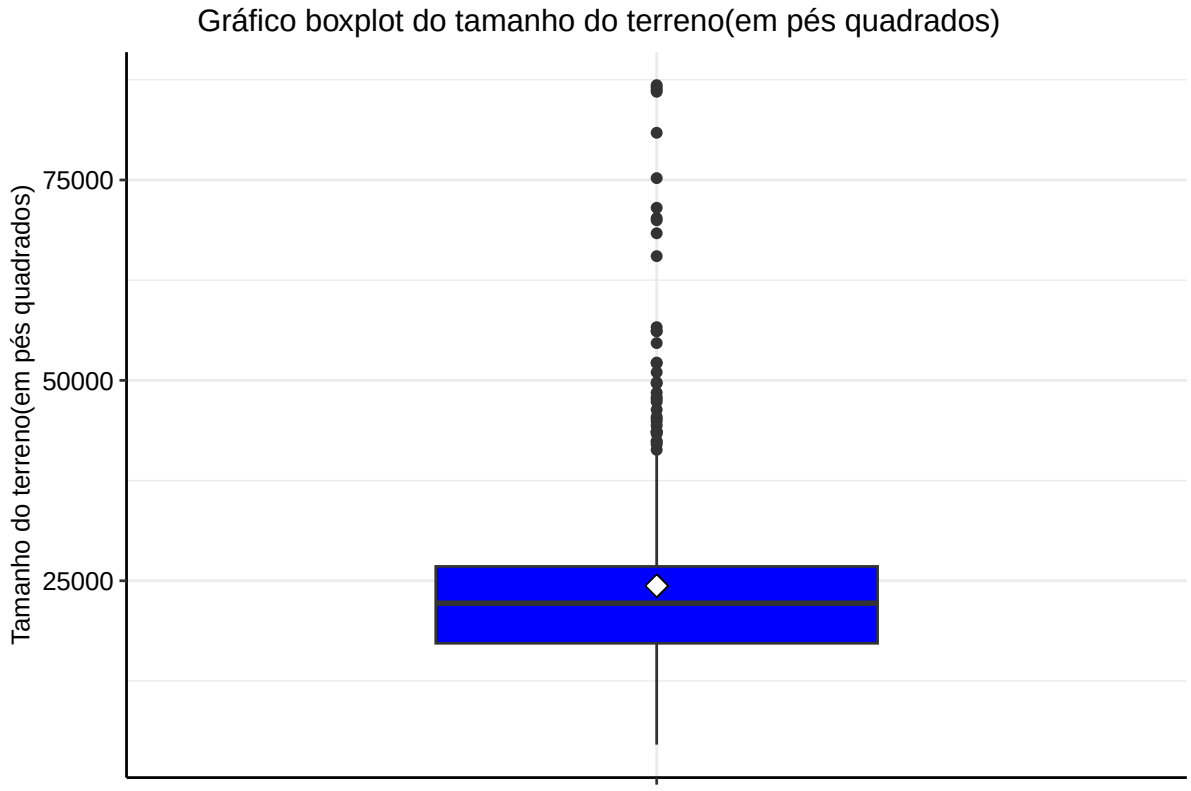



Table 4: Medidas Resumo do tamanho do terreno(em pés quadrados)

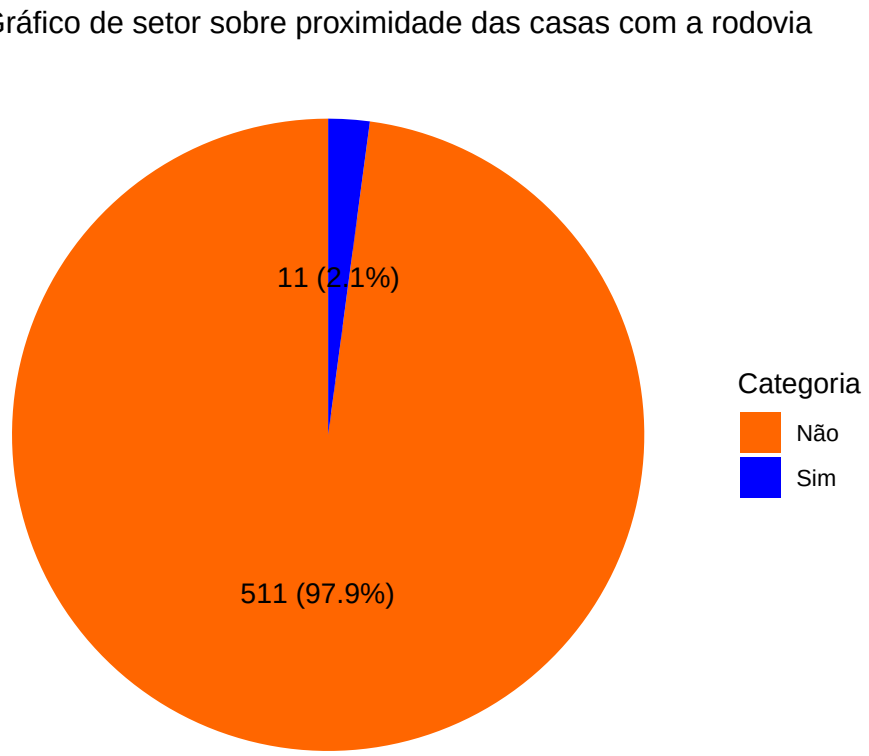
| Estatística | Valor |
|---|---|
| **Minimo** | 4560.00 |
| **Primeiro Quartil.** | 17205.00 |
| **Mediana** | 22200.00 |
| **Média** | 24370.00 |
| **Terceiro Quartil** | 26787.00 |
| **Maximo** | 86830.00 |
| **Desvio Padrão** | 11684.08 |

Com base a Figura 9 e no Quadro 3 comprova - se que o tamanho médio do terreno das casas 24.370 pés quadrados. Vale a pena lembrar que essa medida pode ser distorcida pela presença de outliers no limite superior do boxplot.Nesse contexto, destaca-se a mediana que é de 22.200 pés quadrados, o que significa que metade dos terrenos tem um tamanho menor ou igual a 22.200, enquanto a outra metade possui um tamanho maior ou igual. Desse modo, a mediana é um medida que pode substituir a média, tendo em vista que é resistente a outliers

Além disso, o menor terreno em que a casa construída possui 4560 pés quadrados, ao mesmo tempo que a casa o maior terreno foi de 86.830 pés quadrados. Essa grande amplitude entre o valor mínimo e máximo sugere que os terrenos variam consideravelmente em tamanho, abrangendo desde terrenos menores até propriedades bastante espaçosas. Essa amplitude pode ser analisada através do desvio padrão que é de 29.784,69 pés quadrados, trazendo à tona uma dispersão do tamanho dos terrenos em relação à média

## Proximidade da rodovia

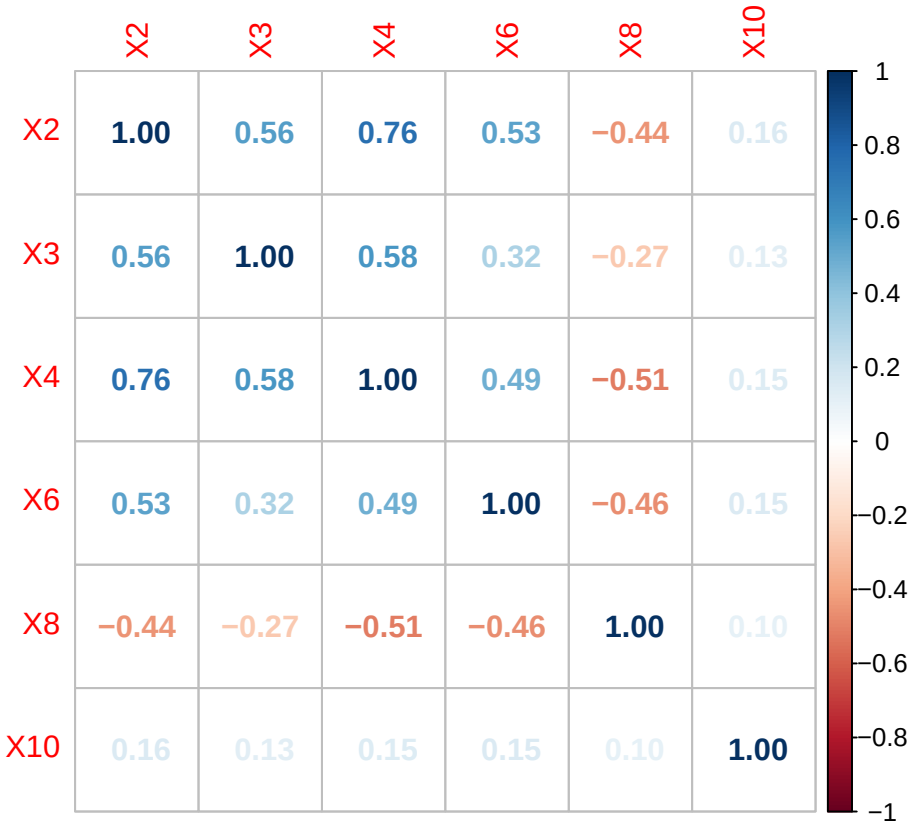
A análise descritiva da figura em questão tem como foco a variável $X_{11}$, que corresponde à proximidade das casas em relação à rodovia.




Analisando a Figura 10, comprova-se que 97,9% das casas que foram construídas entre 1885 e 1998 não são próximas da rodovia. Apenas 11(2,1%) casas ficam perto das rodovias.

## Correlograma das variáveis

Com o objetivo de análisar a existência ou não das variáveis,dispõe-se o correlograma a seguir:



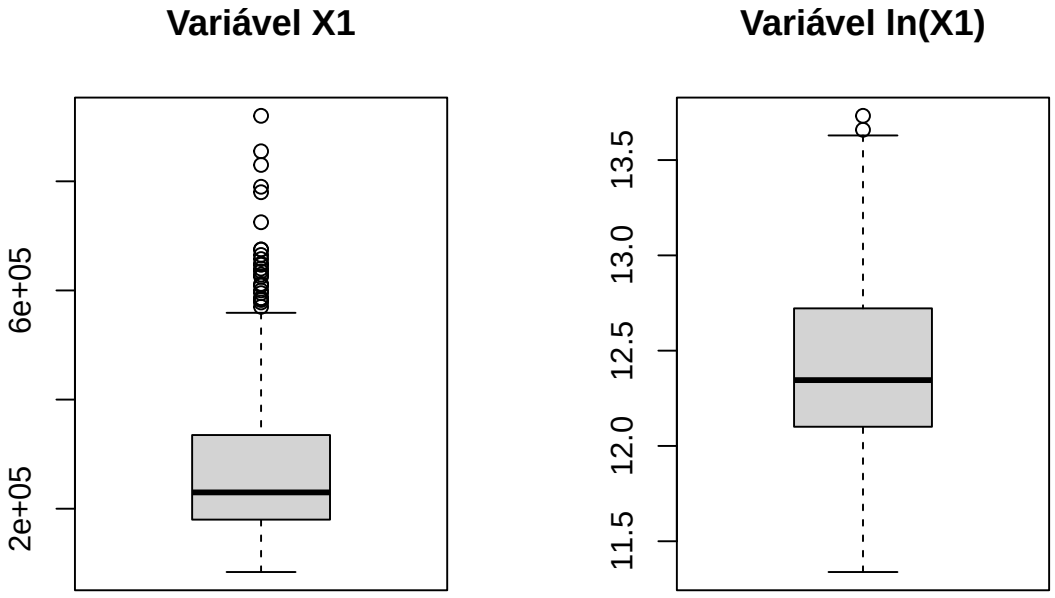

De acordo com o correlograma,conclui-se que há uma correlação:

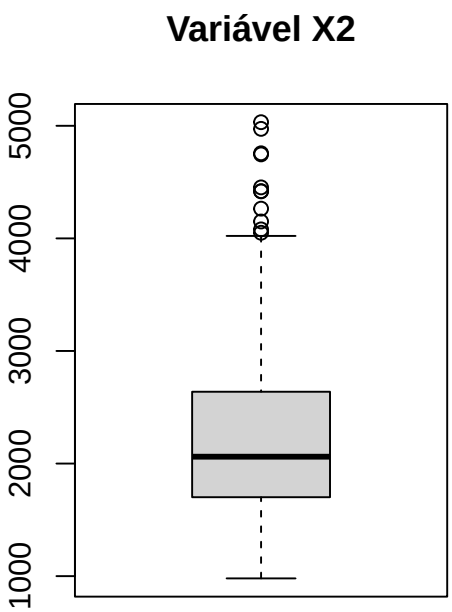
- Moderada(Positiva) entre $X_2$ e $X_3$, $X_2$ e $X_4$, $X_2$ e $X_6$, $X_3$ e $X_4$,$X_4$ e $X_6$
- Moderada(Negativa) entre $X_4$ e $X_8$
- Forte entre $X_2$ e $X_4$

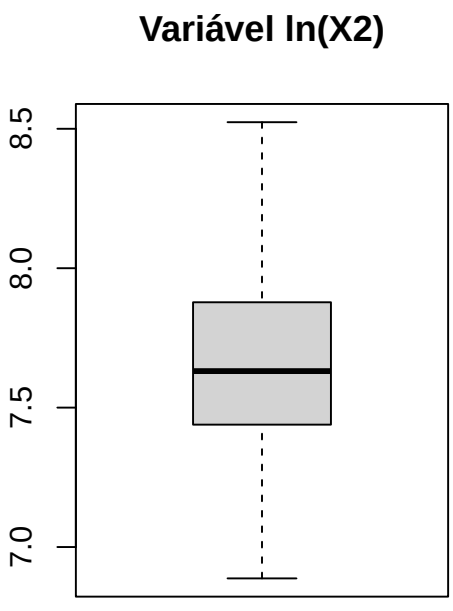
## Descrição das transformações das variáveis numéricas do modelo:

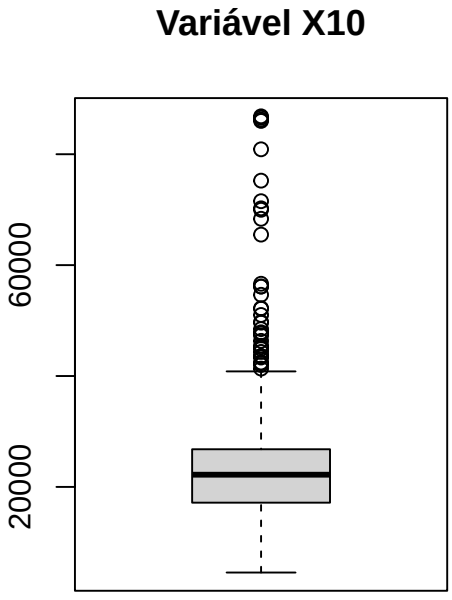
### Variáveis X1 X2 e X10 (numéricas do modelo)

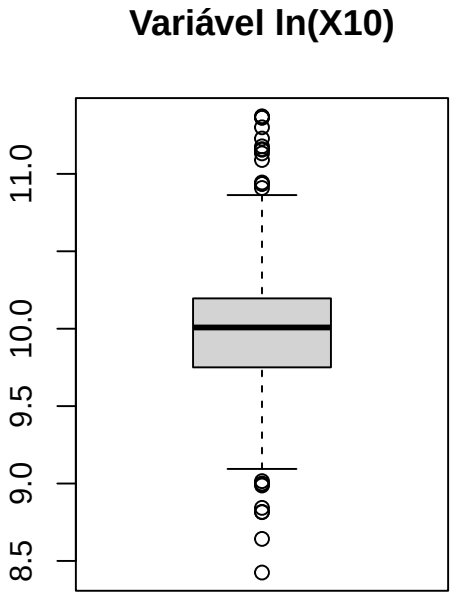
Inicialmente a variável resposta continha uma quantidade considerável de outliers, contudo, depois de aplicarmos o logaritmo natural, se comportou de maneira muito mais estável. O mesmo procedimento foi aplicado tanto para a primeira variável explicativa ($X_2$) quanto para a nova variável explicativa $X_10$ e foram obtidos resultados semelhantes. Houveram transformações que não obtiveram outliers algum.




Variável X2
1000
2000
3000
4000
5000


Variável ln(X2)
7.0
7.5
8.0
8.5


Variável X10
20000
60000


Variável ln(X10)
8.5
9.0
9.5
10.0
11.0

## Variável X9 (Qualidade de construção da casa)

Esta é uma variável que apresenta três valores: 1, 2 e 3. Porém, na realidade, se trata de uma variável qualitativa ordinal, visto que cada valor se refere a um rótulo de construção da casa, sendo 1 - Alta qualidade; 2 - Média qualidade; 3 - Baixa qualidade. A fim de trabalhar num modelo regressivo linear, esta variável será tratada como do tipo Dummy. A análise foi executada utilizando o software R, que para entender a variável como dummy, basta ser lida como variável do tipo *factor*. Entretanto, se estivéssemos trabalhando no SAS, deveríamos realizar uma transformação, criando duas variáveis 0 - 1 que representariam as três categorias de $X_9$.

## Variável X8 - Idade da casa

Conforme explicitado anteriormente, a informação original da variável era o ano de construção da casa, que para trabalhar modificamos para idade da casa.

Além disso, foi considerada a possiblidade de uma transformação quadrática nesta variável, visto que havia a suspeita que tanto casas muito novas teriam um valor maior, quanto casas muito antigas talvez pudessem também ter um valor elevado, considerando o valor histórico/relíquia; levando assim a variável a assumir uma forma de parábola. Entretando, analisando exploratóriamente, este comportamento não foi observado, na realidade notou-se uma relação quase linear no sentido de: quanto mais nova a casa; mais cara ela seria. Portanto, esta transformação foi descartada.

# Análise de regressão linear

# Seleção de variáveis

Após a análise descritiva das variáveis gerar uma breve compreensão de como as variáveis podem se comportar, foram testadas as possibilidades de modelos a partir das combinações e avaliações de como as variáveis contribuem na predição da variável resposta. Ao entrar nessa etapa, foi criado um modelo contendo todas as possíveis variáveis do banco de dados com o objetivo de inicialmente compreender quais são as possíveis variáveis que possuem potencial de explicação.

## Modelo completo

```
##
## Call:
## lm(formula = lny ~ lnX2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
##     lnX10 + X11, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46740 -0.10321 -0.01343  0.09652  0.56170
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.5217425  0.5406030  10.214  < 2e-16 ***
## lnX2         0.6980742  0.0615967  11.333  < 2e-16 ***
## X3           0.0002657  0.0117904   0.023    0.982
## X4           0.0214901  0.0173369   1.240    0.216
## X51          0.0024629  0.0312185   0.079    0.937
## X6           0.0215459  0.0186561   1.155    0.249
## X71          0.0294808  0.0377082   0.782    0.435
## X8           0.0044498  0.0007250   6.138 2.77e-09 ***
## X92         -0.3240606  0.0374184  -8.660 3.42e-16 ***
## X93         -0.3998579  0.0516572  -7.741 1.69e-13 ***
## lnX10        0.1400338  0.0271567   5.157 4.69e-07 ***
## X111        -0.0823877  0.0848600  -0.971    0.332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1671 on 288 degrees of freedom
## Multiple R-squared:  0.8519, Adjusted R-squared:  0.8463
## F-statistic: 150.6 on 11 and 288 DF,  p-value: < 2.2e-16
```

Dado este "primeiro chute" estratégico, foi percebido um valor de $R^2$ *ponderado* $= 0,8463$, o que é muito bom sob o olhar de predição e explicação do modelo. Contudo, duas variáveis, $X_3$ e $X_5$, não se mostraram ter muita influência no resultado de $X_1$ visto que, obtiveram respectivamente os seguintes p-valores, 0,982 e 0,937, que são considerados irrelevantes a quaisquer níveis de significância existentes.

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept) 2.9135   1 104.3267 < 2.2e-16 ***
## lnX2        3.5869   1 128.4366 < 2.2e-16 ***
## X3          0.0000   1   0.0005    0.9820
## X4          0.0429   1   1.5365    0.2161
## X5          0.0002   1   0.0062    0.9372
## X6          0.0372   1   1.3338    0.2491
## X7          0.0171   1   0.6112    0.4350
## X8          1.0520   1  37.6695 2.768e-09 ***
## X9          2.1365   2  38.2504 1.853e-15 ***
## lnX10       0.7426   1  26.5897 4.686e-07 ***
## X11         0.0263   1   0.9426    0.3324
## Residuals   8.0430 288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dessa forma, $X_3$ foi retirado do modelo, e ajustado um outro modelo regressivo sem esta variável.

### Modelo sem a variável X3

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X5 + X6 + X7 + X8 + X9 + lnX10 +
##     X11, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46723 -0.10297 -0.01324  0.09641  0.56210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.520087   0.534663  10.324  < 2e-16 ***
## lnX2         0.698414   0.059622  11.714  < 2e-16 ***
## X4           0.021599   0.016620   1.300    0.195
## X51          0.002518   0.031068   0.081    0.935
## X6           0.021543   0.018623   1.157    0.248
## X71          0.029482   0.037643   0.783    0.434
## X8           0.004449   0.000723   6.154 2.52e-09 ***
## X92         -0.323958   0.037075  -8.738  < 2e-16 ***
## X93         -0.399712   0.051163  -7.813 1.04e-13 ***
## lnX10        0.139994   0.027051   5.175 4.27e-07 ***
## X111        -0.082259   0.084521  -0.973    0.331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1668 on 289 degrees of freedom
## Multiple R-squared:  0.8519, Adjusted R-squared:  0.8468
## F-statistic: 166.2 on 10 and 289 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept) 2.9666   1 106.5936 < 2.2e-16 ***
## lnX2        3.8188   1 137.2170 < 2.2e-16 ***
## X4          0.0470   1   1.6888    0.1948
## X5          0.0002   1   0.0066    0.9355
## X6          0.0372   1   1.3381    0.2483
## X7          0.0171   1   0.6134    0.4342
## X8          1.0539   1  37.8684 2.519e-09 ***
## X9          2.1701   2  38.9874 1.023e-15 ***
## lnX10       0.7454   1  26.7823 4.268e-07 ***
## X11         0.0264   1   0.9472    0.3312
## Residuals   8.0430 289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com a nova versão, notamos que ainda existem variáveis que não se ajustaram ao modelo. Seguindo a parsimônia de eliminá-las uma-a-uma, iremos remover agora a variável $X_7$

### Modelo sem as variáveis X3 e X7

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X5 + X6 + X8 + X9 + lnX10 + X11,
##     data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44213 -0.10536 -0.01065  0.09409  0.55960
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.5686737  0.5306982  10.493  < 2e-16 ***
## lnX2         0.6960474  0.0595060  11.697  < 2e-16 ***
## X4           0.0227446  0.0165450   1.375    0.170
## X51          0.0033876  0.0310277   0.109    0.913
## X6           0.0235023  0.0184421   1.274    0.204
## X8           0.0044053  0.0007203   6.116 3.10e-09 ***
## X92         -0.3259366  0.0369639  -8.818  < 2e-16 ***
## X93         -0.4027508  0.0509815  -7.900 5.83e-14 ***
## lnX10        0.1369554  0.0267536   5.119 5.60e-07 ***
## X111        -0.0844817  0.0844168  -1.001    0.318
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1667 on 290 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.847
## F-statistic: 184.9 on 9 and 290 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept) 3.0602   1 110.1053 < 2.2e-16 ***
## lnX2        3.8028   1 136.8220 < 2.2e-16 ***
## X4          0.0525   1   1.8898    0.1703
## X5          0.0003   1   0.0119    0.9131
## X6          0.0451   1   1.6241    0.2035
## X8          1.0395   1  37.4004 3.103e-09 ***
## X9          2.2097   2  39.7530 5.528e-16 ***
## lnX10       0.7283   1  26.2055 5.600e-07 ***
## X11         0.0278   1   1.0015    0.3178
## Residuals   8.0601 290
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ao analisar o novo modelo, percebemos que ainda existem variáveis não ajustadas. Iremos eliminar agora a variável $X_11$

### Modelo sem as variáveis X3, X7 e X11

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X5 + X6 + X8 + X9 + lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44238 -0.10437 -0.00954  0.09379  0.56208
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.5350688  0.5296362  10.451  < 2e-16 ***
## lnX2         0.6997440  0.0593914  11.782  < 2e-16 ***
## X4           0.0218205  0.0165192   1.321    0.188
## X51          0.0041836  0.0310176   0.135    0.893
## X6           0.0230901  0.0184375   1.252    0.211
## X8           0.0043861  0.0007201   6.091 3.54e-09 ***
## X92         -0.3268638  0.0369524  -8.846  < 2e-16 ***
## X93         -0.4047220  0.0509436  -7.945 4.30e-14 ***
## lnX10        0.1379009  0.0267370   5.158 4.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1667 on 291 degrees of freedom
## Multiple R-squared:  0.8511, Adjusted R-squared:  0.847
## F-statistic: 207.9 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept) 3.0355   1 109.2171 < 2.2e-16 ***
## lnX2        3.8581   1 138.8133 < 2.2e-16 ***
## X4          0.0485   1   1.7448    0.1876
## X5          0.0005   1   0.0182    0.8928
## X6          0.0436   1   1.5684    0.2115
## X8          1.0312   1  37.1011 3.543e-09 ***
## X9          2.2263   2  40.0510 4.317e-16 ***
## lnX10       0.7394   1  26.6016 4.631e-07 ***
## Residuals   8.0879 291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aqui, notamos o mesmo padrão: O $R^2$ ajustado segue bom, porém, ainda há variáveis que não se encaixam no modelo. Iremos remover agora a variável $X_5$

### Modelo sem as variáveis X3, X5, X7 e X11

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X6 + X8 + X9 + lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44207 -0.10455 -0.00847  0.09160  0.55991
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.5397388  0.5276140  10.500  < 2e-16 ***
## lnX2         0.6998833  0.0592825  11.806  < 2e-16 ***
## X4           0.0217831  0.0164891   1.321    0.188
## X6           0.0234049  0.0182585   1.282    0.201
## X8           0.0044054  0.0007045   6.253 1.43e-09 ***
## X92         -0.3265703  0.0368262  -8.868  < 2e-16 ***
## X93         -0.4054371  0.0505817  -8.015 2.66e-14 ***
## lnX10        0.1374732  0.0265036   5.187 4.00e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1664 on 292 degrees of freedom
## Multiple R-squared:  0.8511, Adjusted R-squared:  0.8475
## F-statistic: 238.4 on 7 and 292 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept) 3.0537   1 110.2417 < 2.2e-16 ***
## lnX2        3.8608   1 139.3792 < 2.2e-16 ***
## X4          0.0483   1   1.7452    0.1875
## X6          0.0455   1   1.6432    0.2009
## X8          1.0830   1  39.0990 1.427e-09 ***
## X9          2.2259   2  40.1777 3.856e-16 ***
## lnX10       0.7453   1  26.9046 4.004e-07 ***
## Residuals   8.0885 292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aqui, notamos que o $R^2$ ajustado segue praticamente intacto, porém com um modelo bem mais reduzido e parsimonioso. Ainda há variáveis que sob $\alpha = 0,05$ não se ajustam, porém seus p-valores começam a não diferir muito do aceitável.

Iremos agora testar outra abordagem, que é a de utilizar métodos automáticos para nos ajudar a selecionar as variáveis ideais do modelo à partir daqui.

# Métodos de seleção automática

## Método Backward

Ao contrário do método forward, o método backward começa com o modelo completo e, em cada etapa, remove uma variável por vez com base nos critérios de seleção.

```
## Start:  AIC=-1068
## lny ~ lnX2 + X4 + X6 + X8 + X9 + lnX10
##
##         Df Sum of Sq     RSS      AIC
## - X6     1    0.0455  8.1340 -1068.32
## - X4     1    0.0483  8.1368 -1068.22
## <none>                8.0885 -1068.00
## - lnX10  1    0.7453  8.8337 -1043.56
## - X8     1    1.0830  9.1715 -1032.30
## - X9     2    2.2259 10.3143  -999.08
## - lnX2   1    3.8608 11.9493  -952.93
##
## Step:  AIC=-1068.32
## lny ~ lnX2 + X4 + X8 + X9 + lnX10
##
##         Df Sum of Sq     RSS      AIC
## <none>                8.1340 -1068.32
## - X4     1    0.0569  8.1909 -1068.23
## + X6     1    0.0455  8.0885 -1068.00
## - lnX10  1    0.8175  8.9515 -1041.59
## - X8     1    1.2021  9.3361 -1028.97
## - X9     2    2.5448 10.6787  -990.66
## - lnX2   1    4.0167 12.1506  -949.92
```

```
## Start:  AIC=-1068
## lny ~ lnX2 + X4 + X6 + X8 + X9 + lnX10
##
##         Df Sum of Sq     RSS      AIC
## - X6     1    0.0455  8.1340 -1068.32
## - X4     1    0.0483  8.1368 -1068.22
## <none>                8.0885 -1068.00
## - lnX10  1    0.7453  8.8337 -1043.56
## - X8     1    1.0830  9.1715 -1032.30
## - X9     2    2.2259 10.3143  -999.08
## - lnX2   1    3.8608 11.9493  -952.93
##
## Step:  AIC=-1068.32
## lny ~ lnX2 + X4 + X8 + X9 + lnX10
##
##         Df Sum of Sq     RSS      AIC
## <none>                8.1340 -1068.32
## - X4     1    0.0569  8.1909 -1068.23
## - lnX10  1    0.8175  8.9515 -1041.59
## - X8     1    1.2021  9.3361 -1028.97
## - X9     2    2.5448 10.6787  -990.66
## - lnX2   1    4.0167 12.1506  -949.92
```

```
## Start:  AIC=-1068
## lny ~ lnX2 + X4 + X6 + X8 + X9 + lnX10
```

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X8 + X9 + lnX10, data = dados)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45278 -0.10259 -0.00554  0.09087  0.55528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.462836   0.524767  10.410  < 2e-16 ***
## lnX2         0.708865   0.058931  12.029  < 2e-16 ***
## X4           0.023558   0.016449   1.432    0.153
## X8           0.004567   0.000694   6.580 2.16e-10 ***
## X92         -0.338552   0.035659  -9.494  < 2e-16 ***
## X93         -0.418206   0.049646  -8.424 1.66e-15 ***
## lnX10        0.142436   0.026248   5.427 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1666 on 293 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8472
## F-statistic: 277.2 on 6 and 293 DF,  p-value: < 2.2e-16
```

### Método Forward

Este método começa com um modelo vazio e adiciona uma variável por vez ao modelo, escolhendo a variável que melhora o ajuste do modelo de acordo com os critérios de seleção.

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X6 + X8 + X9 + lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44207 -0.10455 -0.00847  0.09160  0.55991
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.5397388  0.5276140  10.500  < 2e-16 ***
## lnX2         0.6998833  0.0592825  11.806  < 2e-16 ***
## X4           0.0217831  0.0164891   1.321    0.188
## X6           0.0234049  0.0182585   1.282    0.201
## X8           0.0044054  0.0007045   6.253 1.43e-09 ***
## X92         -0.3265703  0.0368262  -8.868  < 2e-16 ***
## X93         -0.4054371  0.0505817  -8.015 2.66e-14 ***
## lnX10        0.1374732  0.0265036   5.187 4.00e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1664 on 292 degrees of freedom
## Multiple R-squared:  0.8511, Adjusted R-squared:  0.8475
## F-statistic: 238.4 on 7 and 292 DF,  p-value: < 2.2e-16
```

### Método Stepwise

Este método realiza um processo iterativo de inclusão e exclusão de variáveis do modelo com base em critérios de seleção, como o valor-p, AIC (Akaike Information Criterion) ou BIC (Bayesian Information Criterion). Ele começa com um modelo inicial e, em cada etapa, adiciona ou remove a variável que resulta no melhor ajuste do modelo com base nos critérios escolhidos.

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X8 + X9 + lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45278 -0.10259 -0.00554  0.09087  0.55528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.462836   0.524767  10.410  < 2e-16 ***
## lnX2         0.708865   0.058931  12.029  < 2e-16 ***
## X4           0.023558   0.016449   1.432    0.153
## X8           0.004567   0.000694   6.580 2.16e-10 ***
## X92         -0.338552   0.035659  -9.494  < 2e-16 ***
## X93         -0.418206   0.049646  -8.424 1.66e-15 ***
## lnX10        0.142436   0.026248   5.427 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1666 on 293 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8472
## F-statistic: 277.2 on 6 and 293 DF,  p-value: < 2.2e-16
```

O novo modelo, que contem todas as variáveis menos X3, X5, X7 e X11 foi realizada uma avaliação por métodos automáticos, visto que ao realizar a ANOVA do modelo novo vemos que todas as variáveis possuem significância ao nível de aproximadamente 20%.

- A análise por backward foi resultou no modelo contendo lnX2, X4, X6, X8, X9 e lnX10
- A análise por forward foi resultou no modelo contendo lnX2, X4, X6, X8, X9 e lnX10
- A análise por stepwise foi resultou no modelo contendo lnX2, X4, X8, X9 e lnX10.

Notamos que para os três métodos de seleção, os resultados de $R^2$ ponderado estão muito próximos, contudo a análise stepwise (mais precisa) foi a única que destoou das demais, removendo uma das variáveis ($X_6$). Portanto, iremos removê-la do modelo também.

## Modelo sem as variáveis X3, X5, X6, X7 e X11

```
##
## Call:
## lm(formula = lny ~ lnX2 + X4 + X8 + X9 + lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45278 -0.10259 -0.00554  0.09087  0.55528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.462836   0.524767  10.410  < 2e-16 ***
## lnX2         0.708865   0.058931  12.029  < 2e-16 ***
## X4           0.023558   0.016449   1.432    0.153
## X8           0.004567   0.000694   6.580 2.16e-10 ***
## X92         -0.338552   0.035659  -9.494  < 2e-16 ***
## X93         -0.418206   0.049646  -8.424 1.66e-15 ***
## lnX10        0.142436   0.026248   5.427 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1666 on 293 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8472
## F-statistic: 277.2 on 6 and 293 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df  F value    Pr(>F)
## (Intercept) 3.0084   1 108.3685 < 2.2e-16 ***
## lnX2        4.0167   1 144.6879 < 2.2e-16 ***
## X4          0.0569   1   2.0512    0.1532
## X8          1.2021   1  43.3022 2.158e-10 ***
## X9          2.5448   2  45.8333 < 2.2e-16 ***
## lnX10       0.8175   1  29.4476 1.205e-07 ***
## Residuals   8.1340 293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Com a nova versão do modelo, sem a variável X6, foi realizada a análise do modelo de regressão e sua respectiva ANOVA. Observa-se que na ANOVA todos os p-valores se mostram significantes, com exceção da variáveL X4. A qualidade de ajuste do modelo segue muito satisfatória, visto que o valor de $R^2$ ponderado observado foi de 0,8472.

### Modelo sem as variáveis X3, X4, X5, X6, X7 e X11

A tabela de regressão linear gerada a partir do modelo anterior, sem o $X_6$, foi notado que a variável $X_4$ parece se destoar muito ao comparar o p-valor dela com as demais. Assim, foi testado retirá-la do modelo para avaliar as possíveis interferências na explicação.

```
##
## Call:
## lm(formula = lny ~ lnX2 + X8 + X9 + lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43646 -0.10136 -0.00669  0.09105  0.56466
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.0501679  0.4393708  11.494  < 2e-16 ***
## lnX2         0.7590029  0.0474916  15.982  < 2e-16 ***
## X8           0.0048405  0.0006683   7.243 3.87e-12 ***
## X92         -0.3346739  0.0356198  -9.396  < 2e-16 ***
## X93         -0.4262013  0.0494188  -8.624 4.08e-16 ***
## lnX10        0.1493101  0.0258516   5.776 1.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1669 on 294 degrees of freedom
## Multiple R-squared:  0.8492, Adjusted R-squared:  0.8466
## F-statistic: 331.1 on 5 and 294 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df F value    Pr(>F)
## (Intercept) 3.6807   1 132.114 < 2.2e-16 ***
## lnX2        7.1160   1 255.419 < 2.2e-16 ***
## X8          1.4615   1  52.459 3.870e-12 ***
## X9          2.5235   2  45.289 < 2.2e-16 ***
## lnX10       0.9294   1  33.358 1.949e-08 ***
## Residuals   8.1909 294
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ao realizar o teste, percebe-se um valor de $R^2$ ponderado muito bom, sendo 0,8466, e a partir dessa medida a variável $X_4$ foi desconsiderada. Com o objetivo de gerar uma maior garantia foi testado uma possível redução do modelo pelo método de avaliação automático stepwise e notou-se um bom comportamente das variáveis do modelo. Por fim, foram validadas as variáveis explitivas para o MRLM.

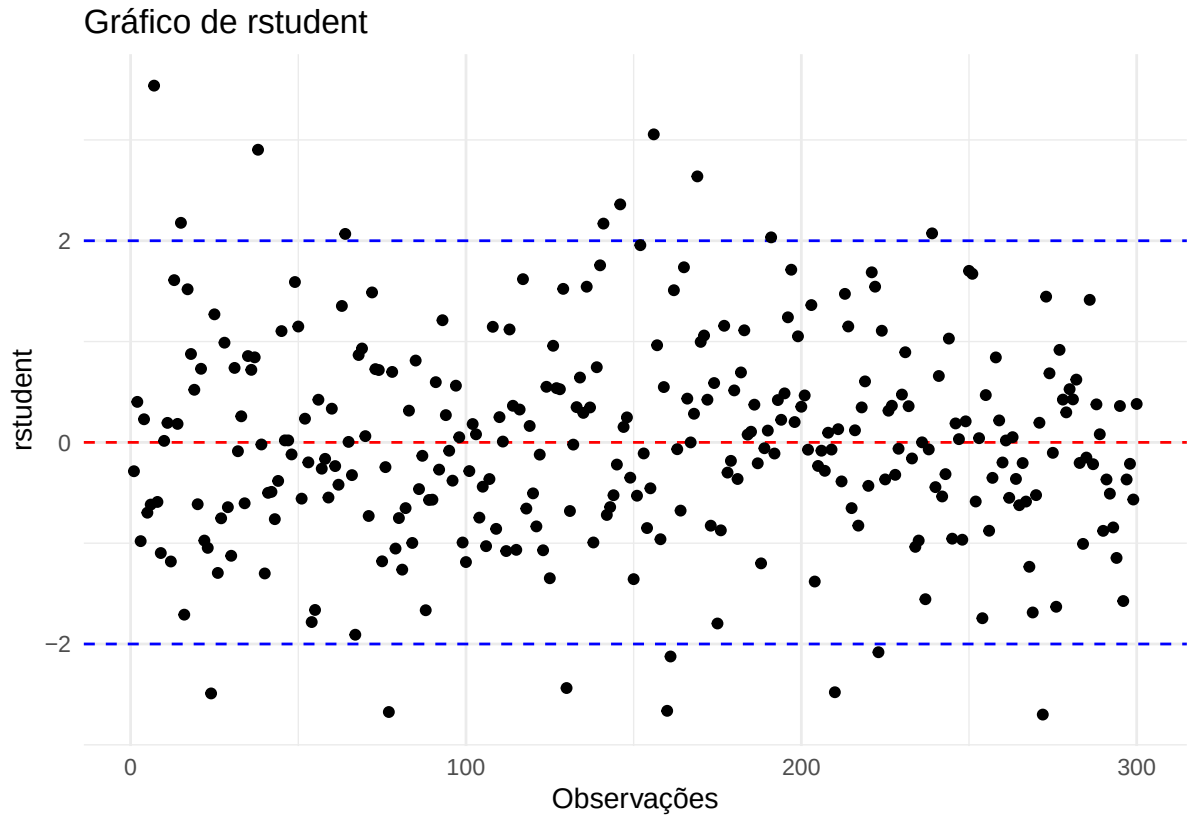
```
## Start:  AIC=-1068.23
## lny ~ lnX2 + X8 + X9 + lnX10
##
##         Df Sum of Sq     RSS      AIC
## <none>                8.1909 -1068.23
## - lnX10  1    0.9294  9.1203 -1037.98
## - X8     1    1.4615  9.6524 -1020.97
## - X9     2    2.5235 10.7144  -991.66
## - lnX2   1    7.1160 15.3069  -882.64
```

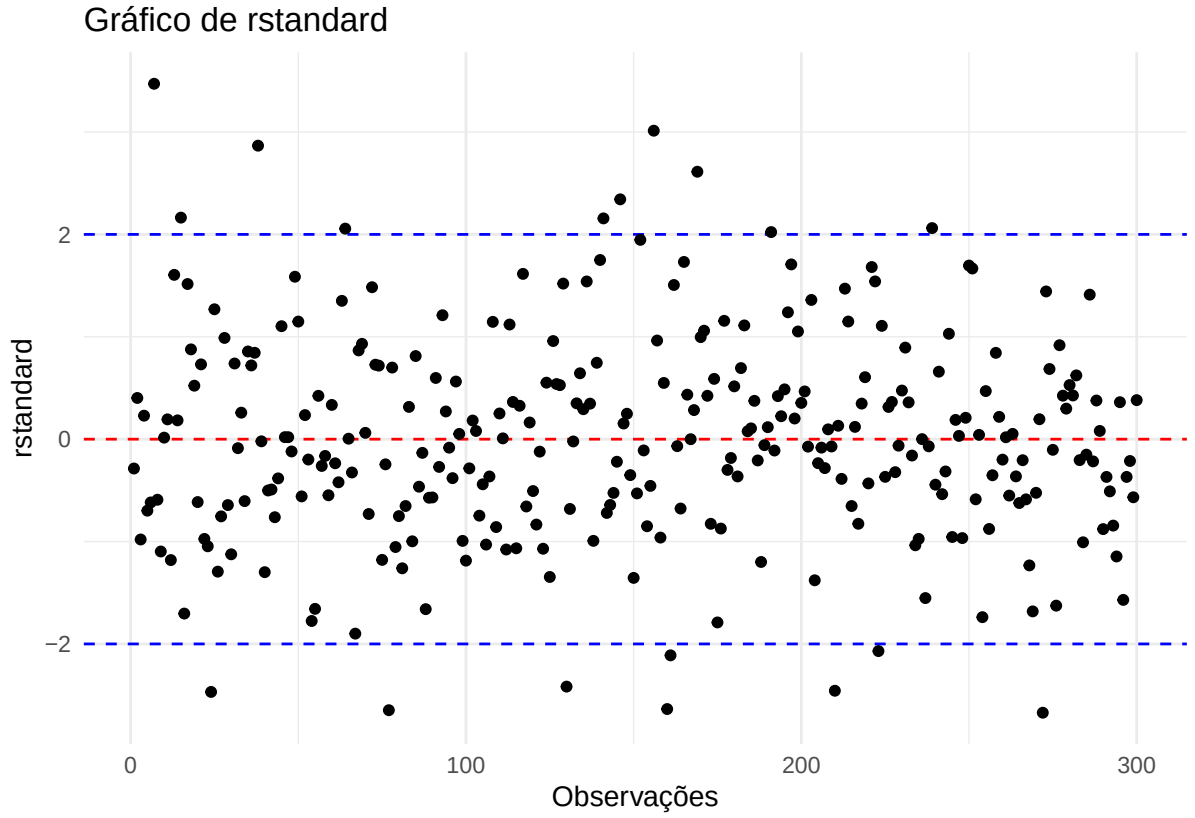
```
##
## Call:
## lm(formula = lny ~ lnX2 + X8 + X9 + lnX10, data = dados)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43646 -0.10136 -0.00669  0.09105  0.56466
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.0501679  0.4393708  11.494  < 2e-16 ***
## lnX2         0.7590029  0.0474916  15.982  < 2e-16 ***
## X8           0.0048405  0.0006683   7.243 3.87e-12 ***
## X92         -0.3346739  0.0356198  -9.396  < 2e-16 ***
## X93         -0.4262013  0.0494188  -8.624 4.08e-16 ***
## lnX10        0.1493101  0.0258516   5.776 1.95e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1669 on 294 degrees of freedom
## Multiple R-squared:  0.8492, Adjusted R-squared:  0.8466
## F-statistic: 331.1 on 5 and 294 DF,  p-value: < 2.2e-16
```

O método stepwise manteve as variáveis exatamente iguais. Portanto, acredito que é seguro afirmar que este é o modelo geral ótimo.

## Análise dos resíduos do modelo selecionado



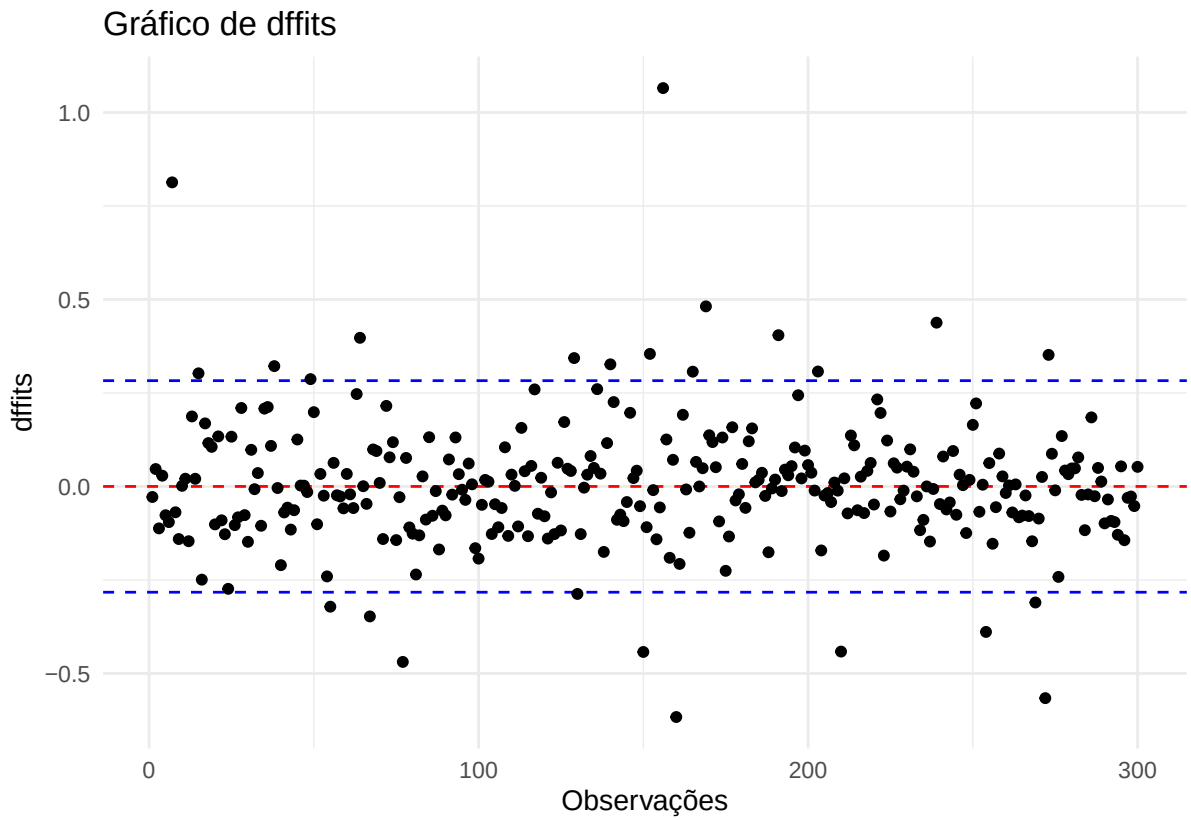
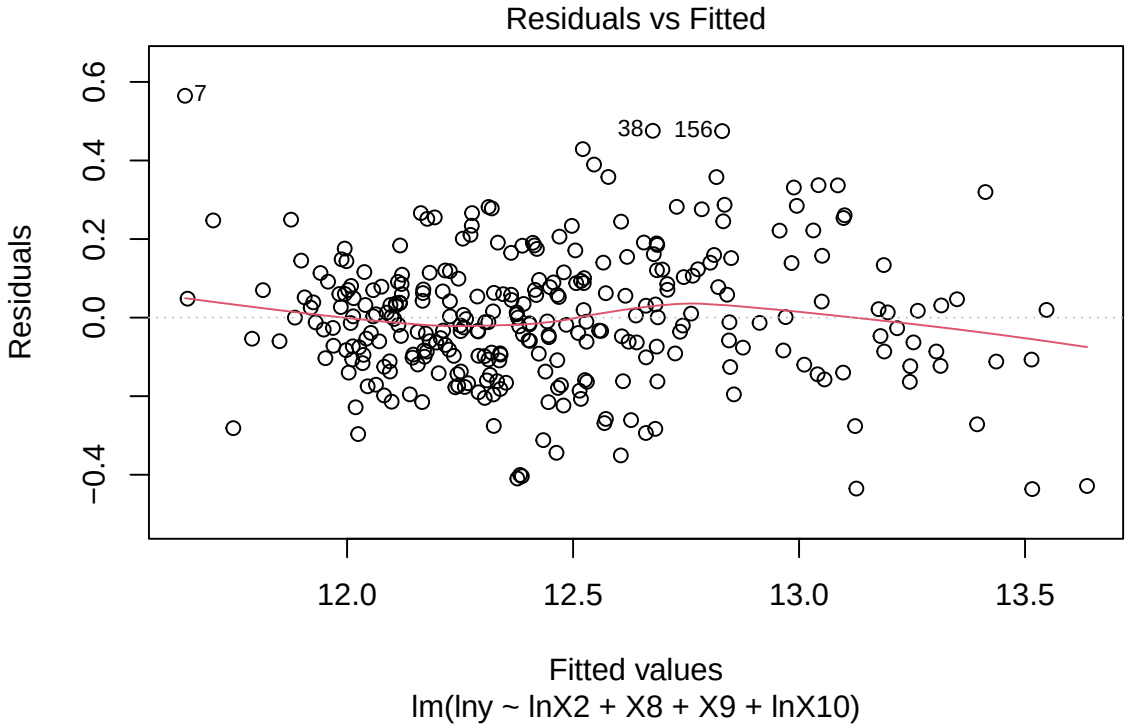
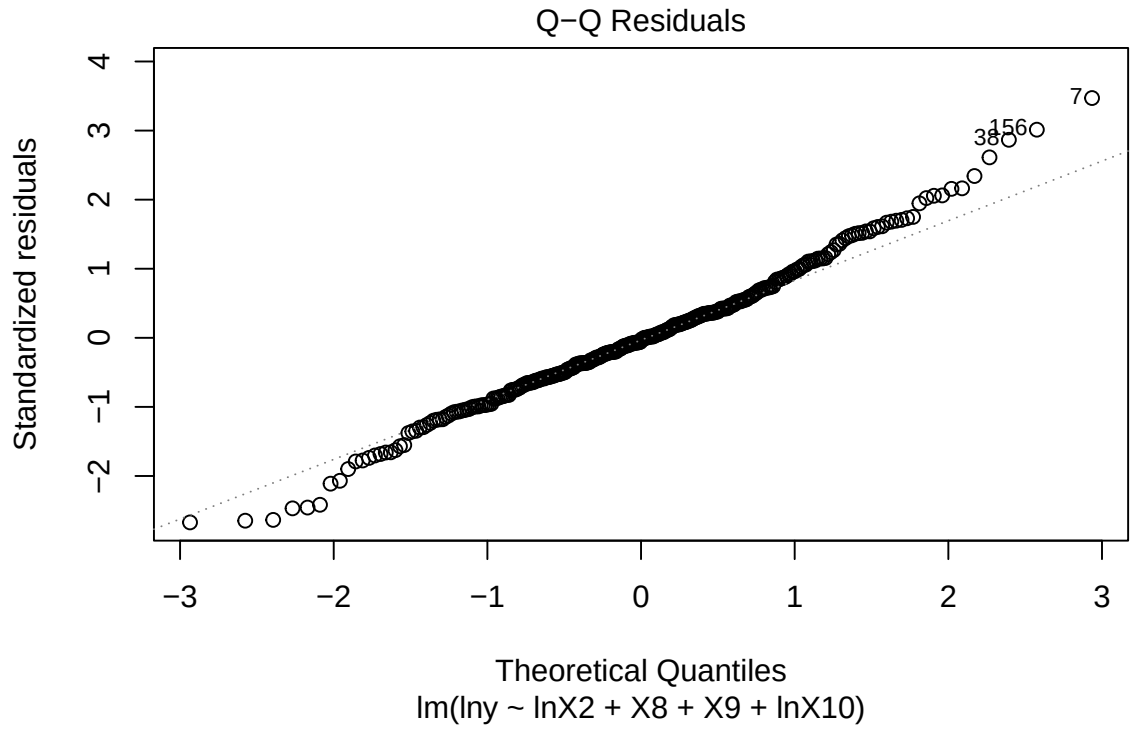
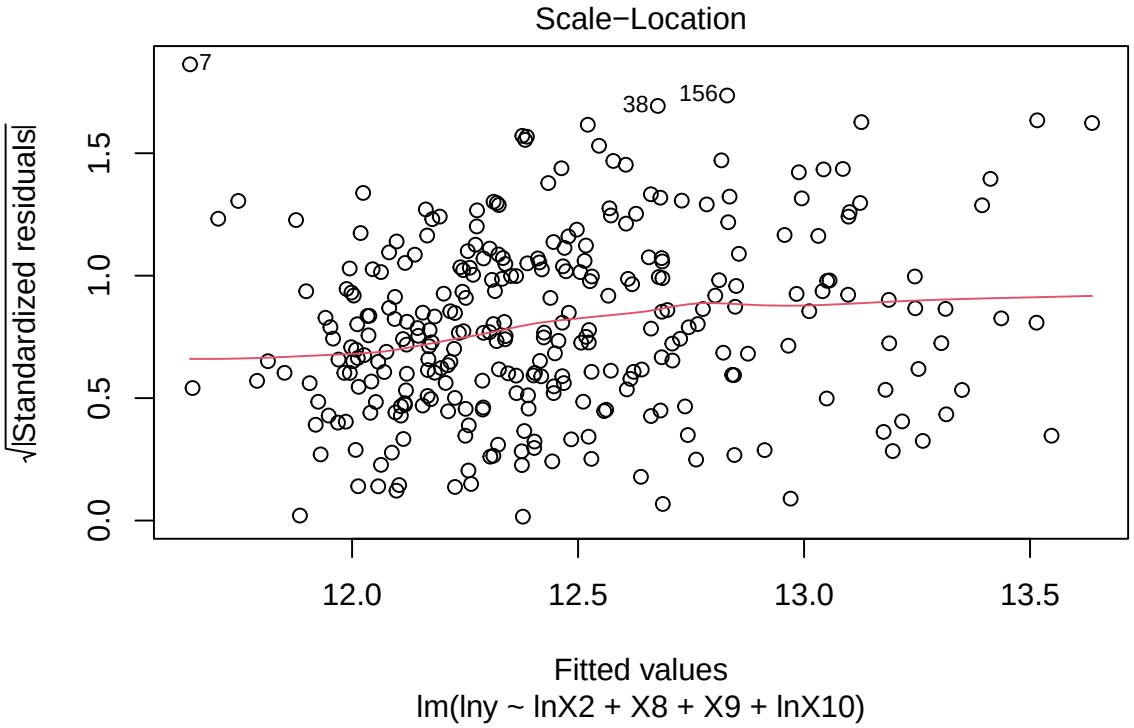
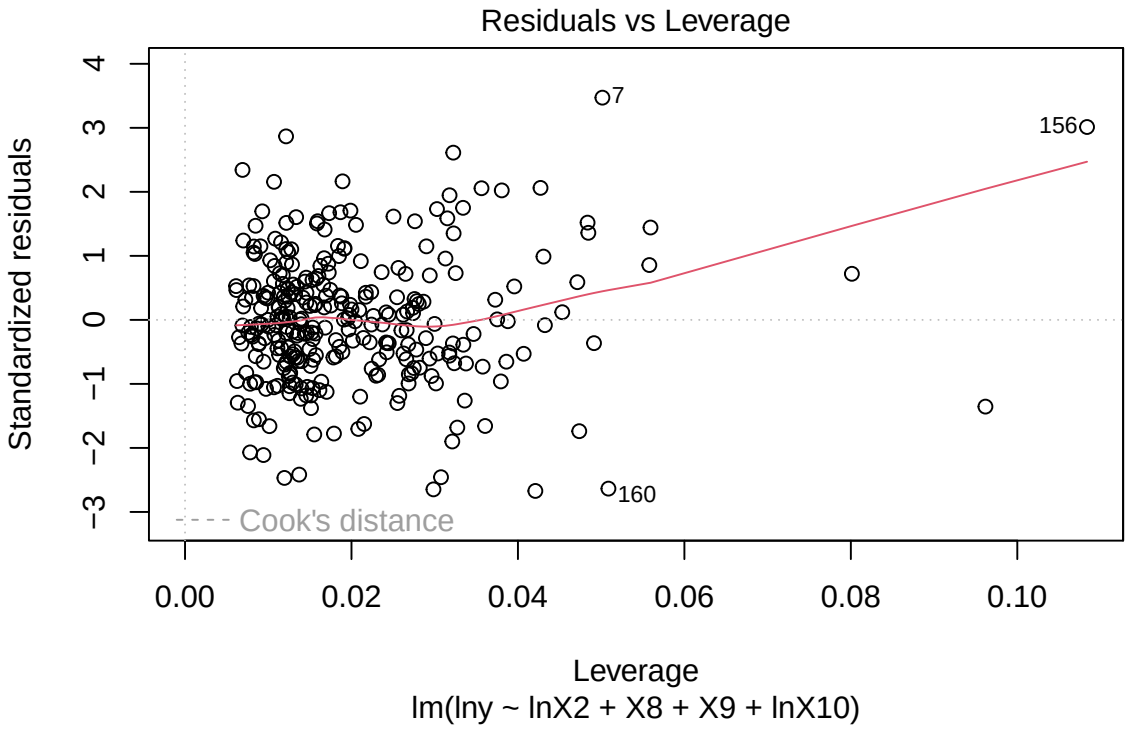


Gráfico de dffits
dffits
1.0
0.5
0.0
−0.5
0
100
200
300
Observações


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.4
7
38
156
12.0
12.5
13.0
13.5
Fitted values
lm(lny ~ lnX2 + X8 + X9 + lnX10)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(lny ~ lnX2 + X8 + X9 + lnX10)
7
156
38


Scale−Location
√|Standardized residuals|
Fitted values
lm(lny ~ lnX2 + X8 + X9 + lnX10)
7
38
156




```
##
##  Shapiro-Wilk normality test
##
## data:  fit7$residuals
## W = 0.9915, p-value = 0.08173
```

Estes gráficos nos ajudam a observar quais valores são possíveis outliers, assim como identificar possíveis valores mais influentes. Além disso, é necessário checar o pressuposto de normalidade dos resíduos do modelo, hipótese esta não rejeitada à $\alpha = 0,05$ (porém bem próximo; seria rejeitada a normalidade à $\alpha = 0,10$).

## Análise de curvatura e suas interações

As etapas realizadas anteriormente nos permitiram selecionar um bom modelo regressivo utilizando muito menos variáveis do que o total coletado, o que facilita e barateia futuras coletas de dados, focando esforço e verba no que realmente interessa. Porém, dado que estamos em posse dessas variáveis, é interessante checar transformações e interações extras sob essas variáveis. Estas não consomem nada a mais de verba e esforço senão do computador (e dos estatísticos encarregados pela análise), portanto, devem ser executadas a fim de tornar o modelo o melhor e mais preciso possível. Uma das coisas importantes que devemos tentar corrigir é a normalidade dos resíduos, visto que pelo teste de Shapiro-Wilk, o p-valor se aproximou bastante da rejeição de $H_0$.

```
##
## Call:
## lm(formula = lny ~ lnX2 * X8 * X9 * lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41056 -0.08628  0.00131  0.08680  0.42567
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -707.85561  399.46697  -1.772   0.0775 .
## lnX2                91.26960   50.47640   1.808   0.0717 .
## X8                   6.90498    3.86085   1.788   0.0748 .
## X92                551.46973  405.41430   1.360   0.1749
## X93                841.54502  406.30804   2.071   0.0393 *
## lnX10               70.56539   38.10167   1.852   0.0651 .
## lnX2:X8             -0.87767    0.48835  -1.797   0.0734 .
## lnX2:X92           -69.63619   51.25372  -1.359   0.1754
## lnX2:X93          -107.52346   51.46696  -2.089   0.0376 *
## X8:X92              -4.95072    3.93755  -1.257   0.2097
## X8:X93              -8.48484    4.03033  -2.105   0.0362 *
## lnX2:lnX10          -8.94830    4.82521  -1.854   0.0647 .
## X8:lnX10            -0.67698    0.36909  -1.834   0.0677 .
## X92:lnX10          -54.70356   38.69481  -1.414   0.1586
## X93:lnX10          -82.44905   38.75533  -2.127   0.0343 *
## lnX2:X8:X92          0.62449    0.49835   1.253   0.2112
## lnX2:X8:X93          1.08647    0.51271   2.119   0.0350 *
## lnX2:X8:lnX10        0.08620    0.04678   1.843   0.0665 .
## lnX2:X92:lnX10       6.91281    4.90249   1.410   0.1596
## lnX2:X93:lnX10      10.53226    4.91947   2.141   0.0332 *
## X8:X92:lnX10         0.48508    0.37676   1.288   0.1990
## X8:X93:lnX10         0.82160    0.38554   2.131   0.0340 *
## lnX2:X8:X92:lnX10   -0.06129    0.04778  -1.283   0.2007
## lnX2:X8:X93:lnX10   -0.10522    0.04914  -2.141   0.0331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1535 on 276 degrees of freedom
## Multiple R-squared:  0.8803, Adjusted R-squared:  0.8703
## F-statistic: 88.23 on 23 and 276 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##                 Sum Sq  Df F value   Pr(>F)
## (Intercept)     0.0740   1  3.1400 0.077498 .
## lnX2            0.0770   1  3.2695 0.071669 .
## X8              0.0754   1  3.1986 0.074798 .
```

```
## X9               0.2609   2  5.5374 0.004387 **
## lnX10            0.0808   1  3.4300 0.065089 .
## lnX2:X8          0.0761   1  3.2300 0.073395 .
## lnX2:X9          0.2555   2  5.4218 0.004903 **
## X8:X9            0.2082   2  4.4183 0.012919 *
## lnX2:lnX10       0.0810   1  3.4391 0.064736 .
## X8:lnX10         0.0793   1  3.3643 0.067701 .
## X9:lnX10         0.2629   2  5.5803 0.004210 **
## lnX2:X8:X9       0.2019   2  4.2848 0.014704 *
## lnX2:X8:lnX10    0.0800   1  3.3951 0.066463 .
## lnX2:X9:lnX10    0.2573   2  5.4611 0.004721 **
## X8:X9:lnX10      0.2058   2  4.3686 0.013556 *
## lnX2:X8:X9:lnX10 0.1996   2  4.2353 0.015428 *
## Residuals        6.5022 276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testando a interação, notamos significância em todos os casos. Para isso, utilizaremos o método automático Stepwise para tentar selecionar as melhores.

```
## Start:  AIC=-1101.49
## lny ~ lnX2 * X8 * X9 * lnX10
##
##                     Df Sum of Sq    RSS     AIC
## <none>                           6.5022 -1101.5
## - lnX2:X8:X9:lnX10  2   0.19955 6.7017 -1096.4
```

```
##
## Call:
## lm(formula = lny ~ lnX2 * X8 * X9 * lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41056 -0.08628  0.00131  0.08680  0.42567
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -707.85561  399.46697  -1.772   0.0775 .
## lnX2                91.26960   50.47640   1.808   0.0717 .
## X8                   6.90498    3.86085   1.788   0.0748 .
## X92                551.46973  405.41430   1.360   0.1749
## X93                841.54502  406.30804   2.071   0.0393 *
## lnX10               70.56539   38.10167   1.852   0.0651 .
## lnX2:X8             -0.87767    0.48835  -1.797   0.0734 .
## lnX2:X92           -69.63619   51.25372  -1.359   0.1754
## lnX2:X93          -107.52346   51.46696  -2.089   0.0376 *
## X8:X92              -4.95072    3.93755  -1.257   0.2097
## X8:X93              -8.48484    4.03033  -2.105   0.0362 *
## lnX2:lnX10          -8.94830    4.82521  -1.854   0.0647 .
## X8:lnX10            -0.67698    0.36909  -1.834   0.0677 .
## X92:lnX10          -54.70356   38.69481  -1.414   0.1586
## X93:lnX10          -82.44905   38.75533  -2.127   0.0343 *
## lnX2:X8:X92          0.62449    0.49835   1.253   0.2112
## lnX2:X8:X93          1.08647    0.51271   2.119   0.0350 *
## lnX2:X8:lnX10        0.08620    0.04678   1.843   0.0665 .
## lnX2:X92:lnX10       6.91281    4.90249   1.410   0.1596
## lnX2:X93:lnX10      10.53226    4.91947   2.141   0.0332 *
## X8:X92:lnX10         0.48508    0.37676   1.288   0.1990
## X8:X93:lnX10         0.82160    0.38554   2.131   0.0340 *
## lnX2:X8:X92:lnX10   -0.06129    0.04778  -1.283   0.2007
```

```
## lnX2:X8:X93:lnX10   -0.10522    0.04914  -2.141   0.0331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1535 on 276 degrees of freedom
## Multiple R-squared:  0.8803, Adjusted R-squared:  0.8703
## F-statistic: 88.23 on 23 and 276 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##                  Sum Sq  Df F value   Pr(>F)
## (Intercept)      0.0740   1  3.1400 0.077498 .
## lnX2             0.0770   1  3.2695 0.071669 .
## X8               0.0754   1  3.1986 0.074798 .
## X9               0.2609   2  5.5374 0.004387 **
## lnX10            0.0808   1  3.4300 0.065089 .
## lnX2:X8          0.0761   1  3.2300 0.073395 .
## lnX2:X9          0.2555   2  5.4218 0.004903 **
## X8:X9            0.2082   2  4.4183 0.012919 *
## lnX2:lnX10       0.0810   1  3.4391 0.064736 .
## X8:lnX10         0.0793   1  3.3643 0.067701 .
## X9:lnX10         0.2629   2  5.5803 0.004210 **
## lnX2:X8:X9       0.2019   2  4.2848 0.014704 *
## lnX2:X8:lnX10    0.0800   1  3.3951 0.066463 .
## lnX2:X9:lnX10    0.2573   2  5.4611 0.004721 **
## X8:X9:lnX10      0.2058   2  4.3686 0.013556 *
## lnX2:X8:X9:lnX10 0.1996   2  4.2353 0.015428 *
## Residuals        6.5022 276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# O modelo final selecionado

Observando os resultados do método Stepwise, notamos que ainda podemos fazer mais alguns ajustes manuais, a fim de chegar ao modelo ideal

```
##
## Call:
## lm(formula = lny ~ lnX2 + X8 + X9 + lnX10 + lnX2:X9 + lnX2:lnX10 +
##     lnX2:X8:X9:lnX10, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39718 -0.09470  0.00694  0.08953  0.44320
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       27.7166409  6.6652651   4.158 4.23e-05 ***
## lnX2              -1.8343157  0.8738610  -2.099  0.03668 *
## X8                -0.0243130  0.0103590  -2.347  0.01960 *
## X92               -7.9489971  1.1347324  -7.005 1.75e-11 ***
## X93               -7.8030809  1.5631612  -4.992 1.04e-06 ***
## lnX10             -1.2100778  0.6428070  -1.882  0.06078 .
## lnX2:X92           0.9162731  0.1455083   6.297 1.13e-09 ***
## lnX2:X93           0.8752089  0.1984819   4.410 1.46e-05 ***
## lnX2:lnX10         0.1495186  0.0864259   1.730  0.08470 .
## lnX2:X8:X91:lnX10  0.0003350  0.0001263   2.653  0.00842 **
## lnX2:X8:X92:lnX10  0.0003692  0.0001302   2.835  0.00491 **
```

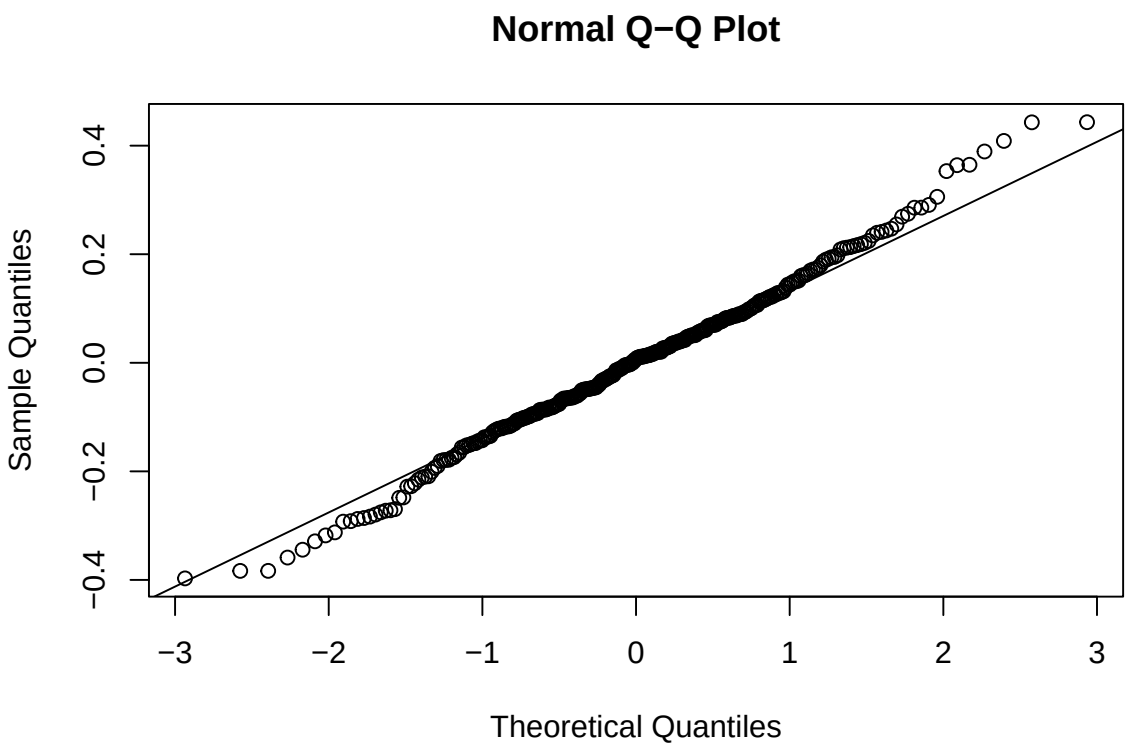
```
## lnX2:X8:X93:lnX10  0.0003841  0.0001354   2.836  0.00489 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1553 on 288 degrees of freedom
## Multiple R-squared:  0.8722, Adjusted R-squared:  0.8673
## F-statistic: 178.6 on 11 and 288 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##                  Sum Sq  Df F value    Pr(>F)
## (Intercept)      0.4169   1 17.2920 4.233e-05 ***
## lnX2             0.1062   1  4.4062   0.03668 *
## X8               0.1328   1  5.5086   0.01960 *
## X9               1.1841   2 24.5597 1.418e-10 ***
## lnX10            0.0854   1  3.5438   0.06078 .
## lnX2:X9          0.9562   2 19.8326 8.523e-09 ***
## lnX2:lnX10       0.0722   1  2.9930   0.08470 .
## lnX2:X8:X9:lnX10 0.2059   3  2.8473   0.03787 *
## Residuals        6.9429 288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Foi testada todas as interações das variáveis selecionadas, lnX2, X8, X9 e lnX10. Algumas intereções se mostraram importantes, como lnX2:X9, lnX2:lnX10 e lnX2:X8:X9:lnX10; e foram deixadas no modelo. Por consequência, o modelo final selecionado é o que contém as variáveis explicativas lnX2, X8,X9 e lnX10; e as interações entre elas lnX2:X9, lnX2:lnX10, lnX2:X8:X9:lnX10.

Com o novo modelo em mãos, foi realizada a análise de regressão linear e sua respectiva ANOVA, resultando em uma aceitação boa de todas as componentes do modelo assim como um acrescimo no valor de $R^2$ ao acrescentar as interações significantes das variávesis, resultando no novo valor de $R^2$ ponderado sendo 0,8673.

```
##
##  Shapiro-Wilk normality test
##
## data:  fit_f$residuals
## W = 0.99433, p-value = 0.3282
```



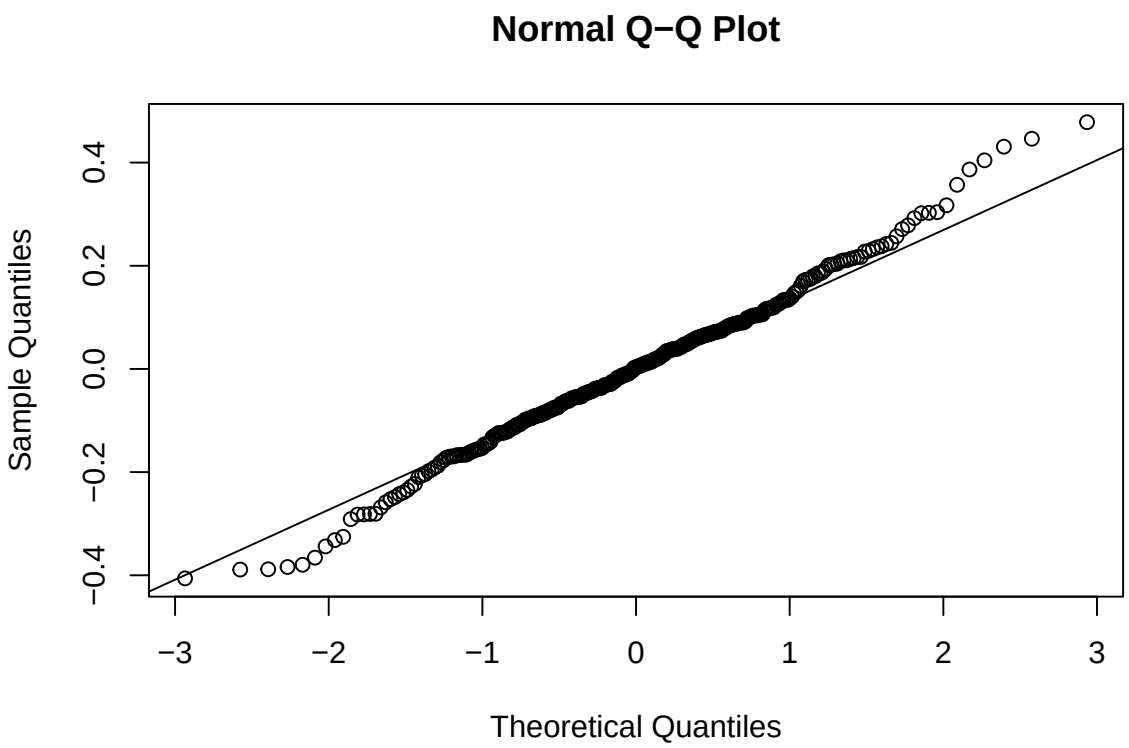

Ao realizar a análise de resíduos do novo modelo, vemos que existe uma aderência a normalidade com p-valor de 0.3282, satisfazendo assim a hipótese de normalidade dos resíduos.

## Testes para possível melhora

```
##
## Call:
## lm(formula = lny ~ lnX2 + X8 + X9 + lnX10 + lnX2:X9 + lnX2:lnX10,
##     data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40600 -0.09328  0.00319  0.08960  0.47826
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.8680441  6.6229007   4.057 6.39e-05 ***
## lnX2        -2.0663250  0.8639439  -2.392 0.017404 *
## X8           0.0048991  0.0006366   7.696 2.20e-13 ***
## X92         -7.1881688  1.0915188  -6.585 2.12e-10 ***
## X93         -5.7193968  1.3693187  -4.177 3.91e-05 ***
## lnX10       -1.4442760  0.6259020  -2.308 0.021728 *
## lnX2:X92     0.8547966  0.1358926   6.290 1.16e-09 ***
## lnX2:X93     0.6488140  0.1760514   3.685 0.000272 ***
## lnX2:lnX10   0.2095383  0.0817713   2.562 0.010896 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1567 on 291 degrees of freedom
## Multiple R-squared:  0.8684, Adjusted R-squared:  0.8648
## F-statistic:   240 on 8 and 291 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: lny
##             Sum Sq  Df F value    Pr(>F)
## (Intercept) 0.4043   1 16.4579 6.395e-05 ***
## lnX2        0.1405   1  5.7204   0.01740 *
## X8          1.4550   1 59.2266 2.199e-13 ***
## X9          1.0807   2 21.9954 1.271e-09 ***
## lnX10       0.1308   1  5.3246   0.02173 *
## lnX2:X9     0.9917   2 20.1835 6.186e-09 ***
## lnX2:lnX10  0.1613   1  6.5664   0.01090 *
## Residuals   7.1488 291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit_f2$residuals
## W = 0.9933, p-value = 0.2022
```




Ainda que na ANOVA tenhamos valores que corroborem para melhor desempenho do modelo acima, os residuos sofrem um leve desgaste, dessa forma, compreendemos que o modelo ajustado anteriormente é o melhor.

# Análise gráfica

## Resíduos, outliers e valores influentes

Na análise de resíduos do modelo, foi contabilizado pela studentização e pela normalização com o objetivo de identificar valores outliers e foram notadas algumas possíveis observações como outliers. Dado o gatilho, foi feita em seguida a análise de valores influentes, que possibilita a identificação das observações que se distanciam da abordagem que o modelo propõe.

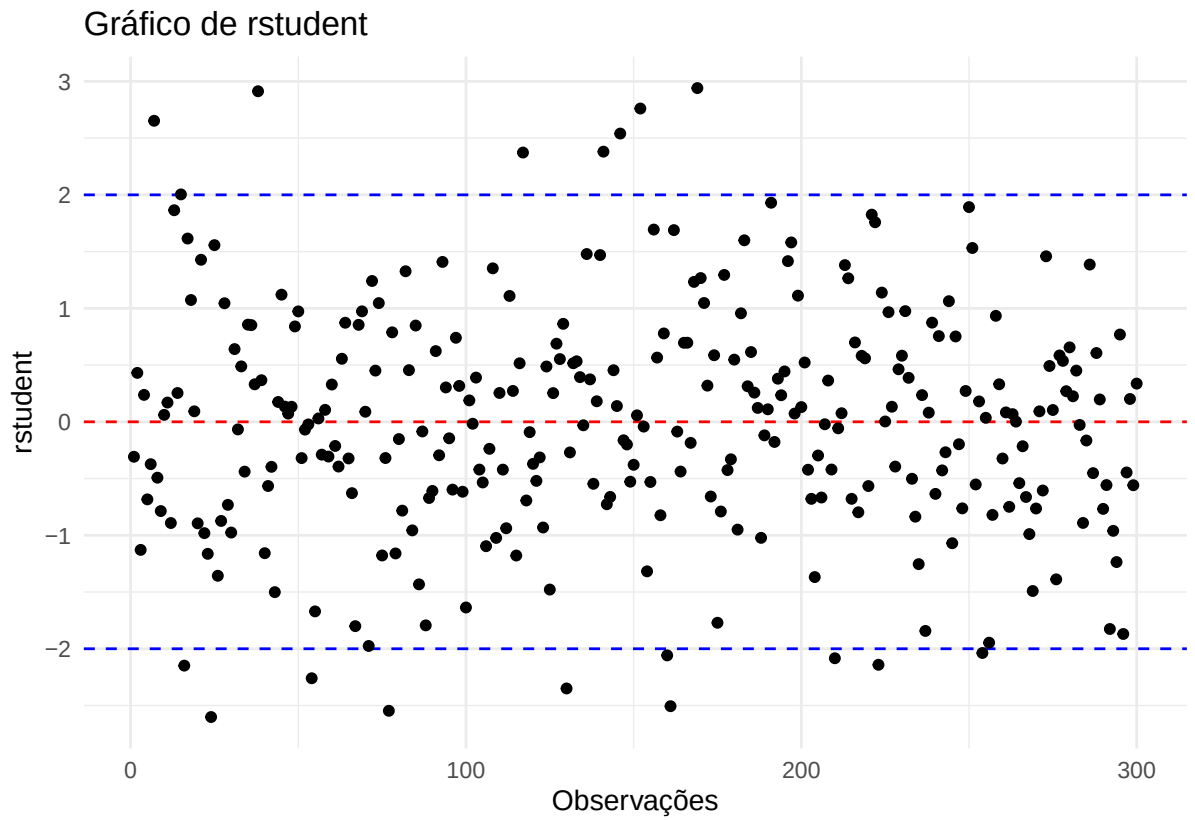


Gráfico de rstudent
rstudent
3
2
1
0
−1
−2
0
100
200
300
Observações

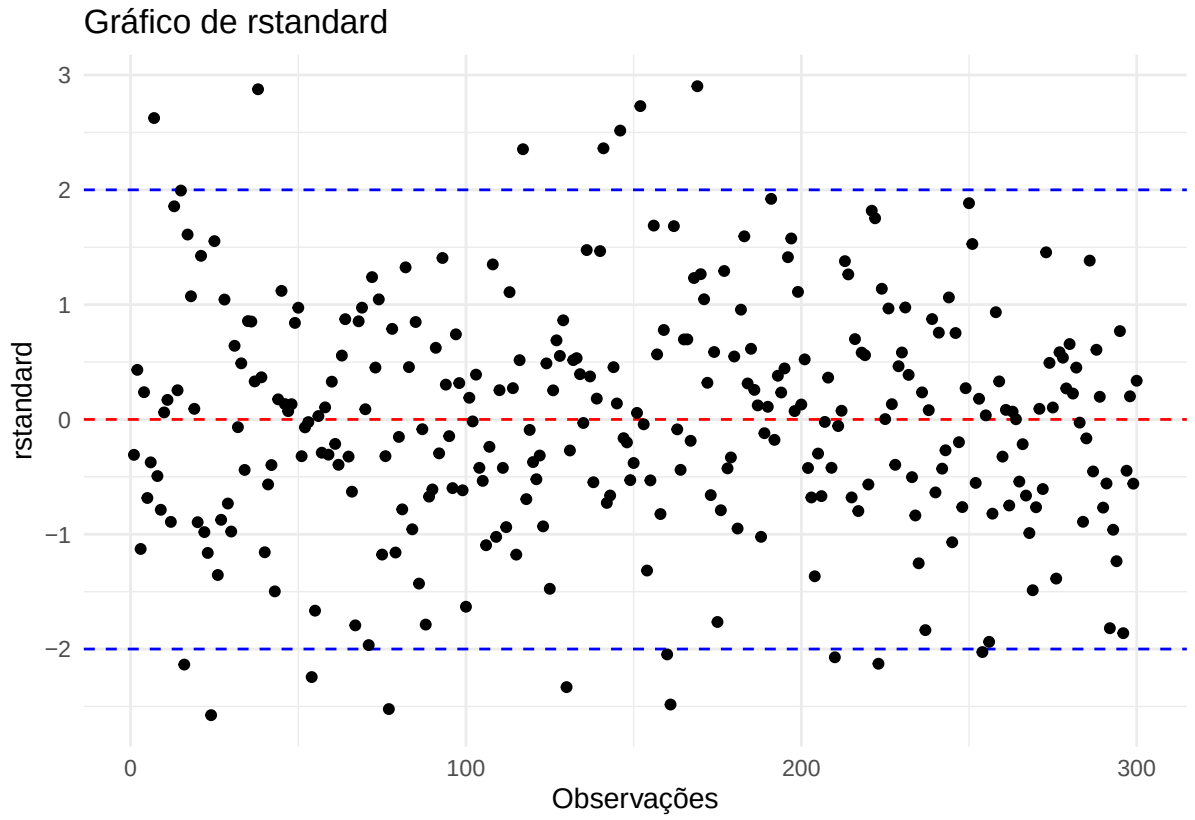
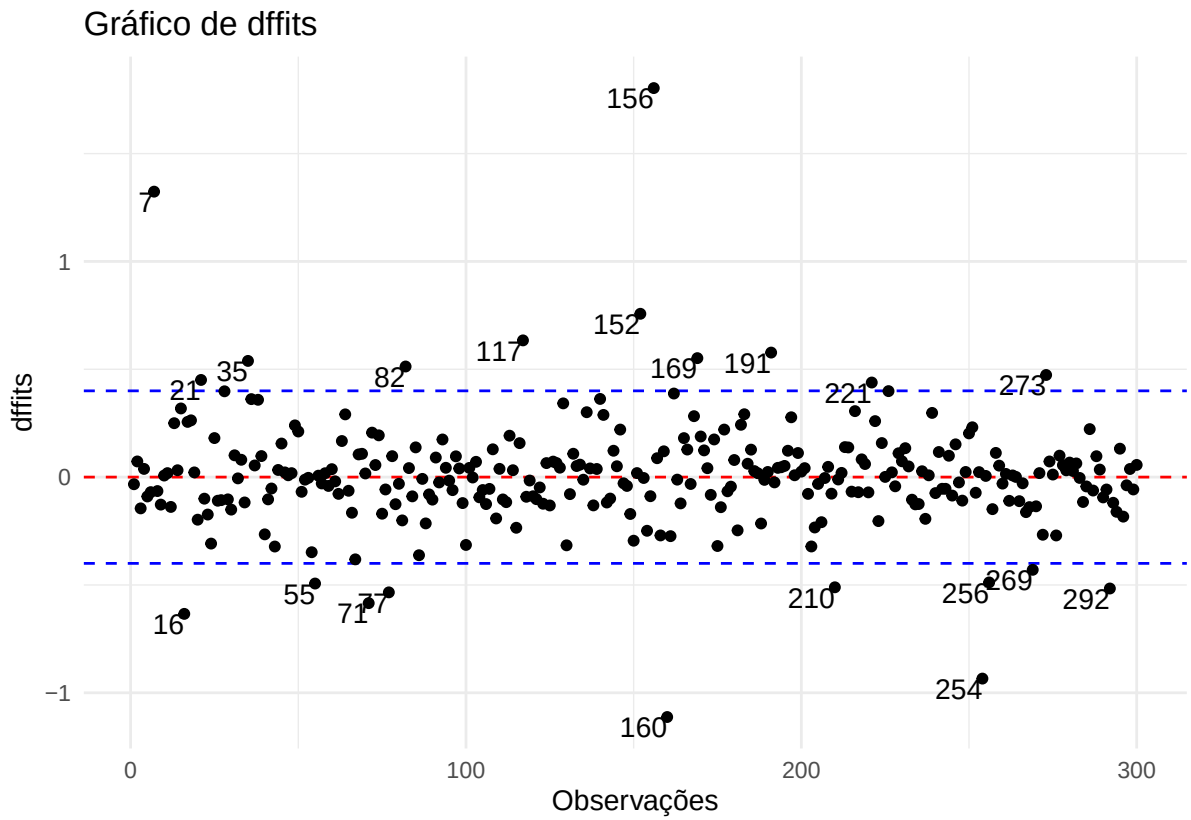
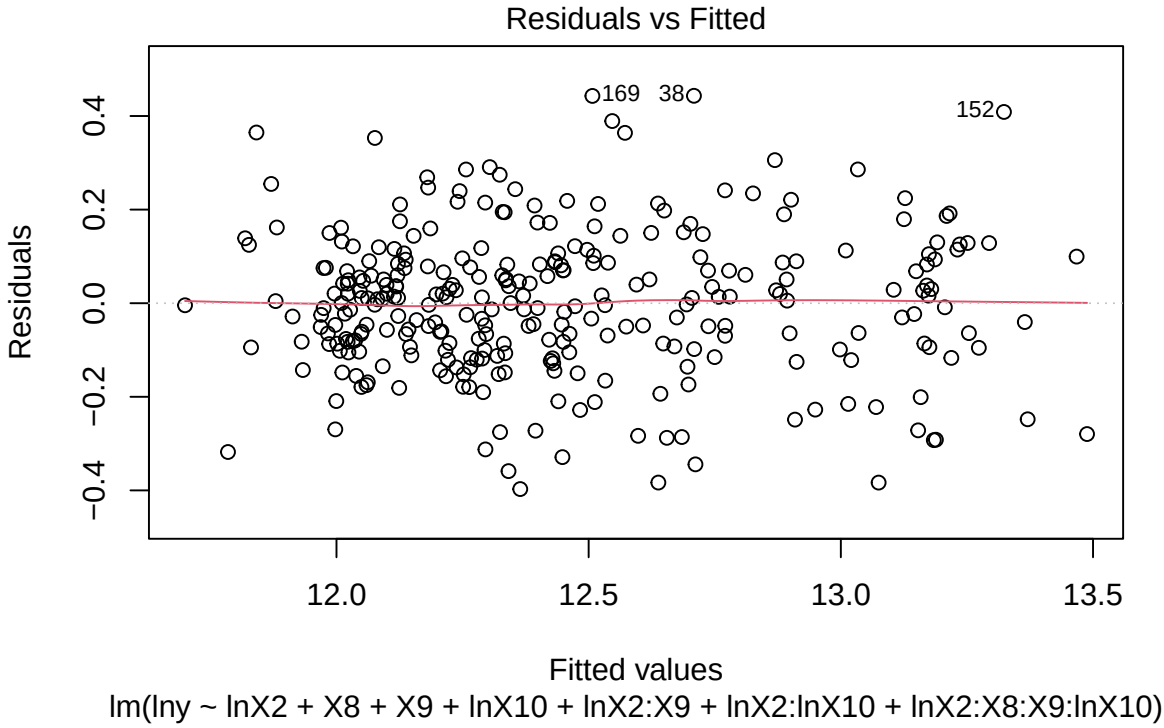
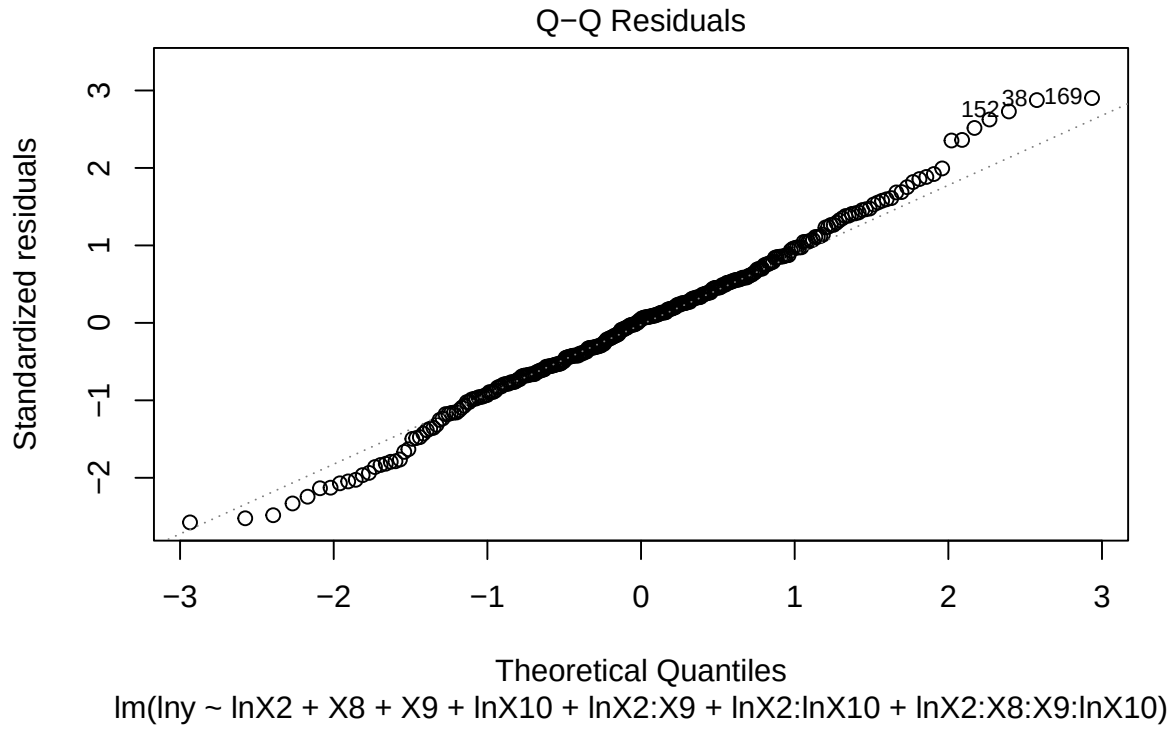


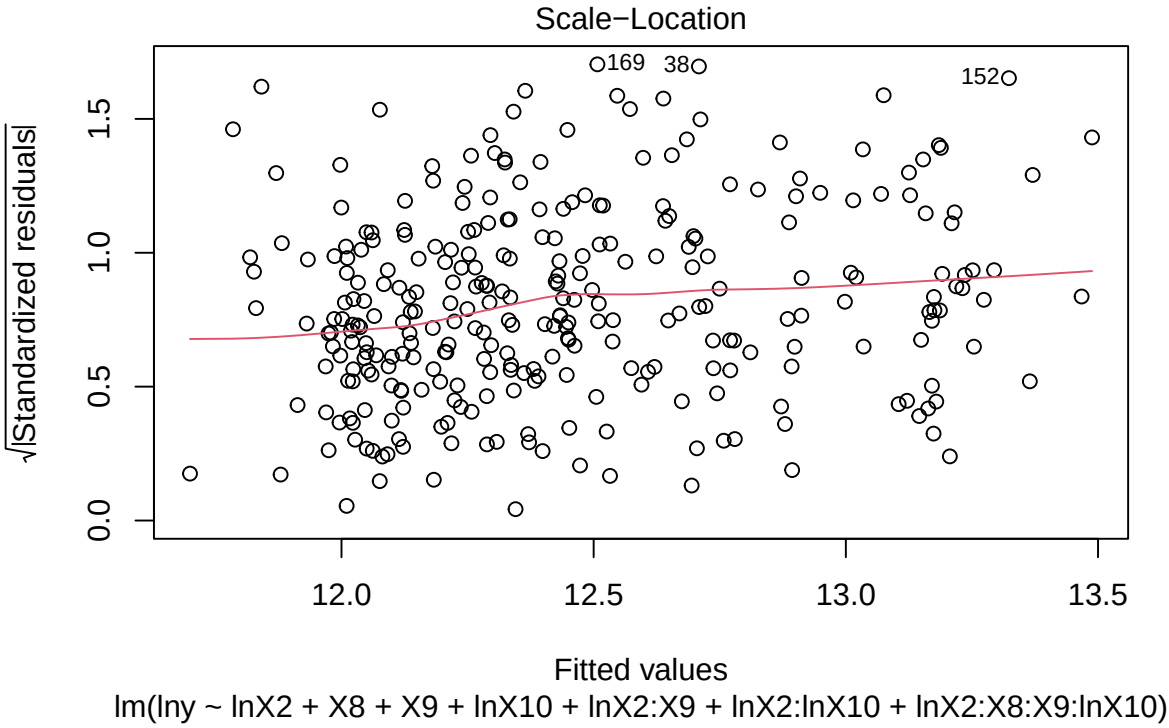
Gráfico de rstandard
rstandard
3
2
1
0
−1
−2
0
100
200
300
Observações

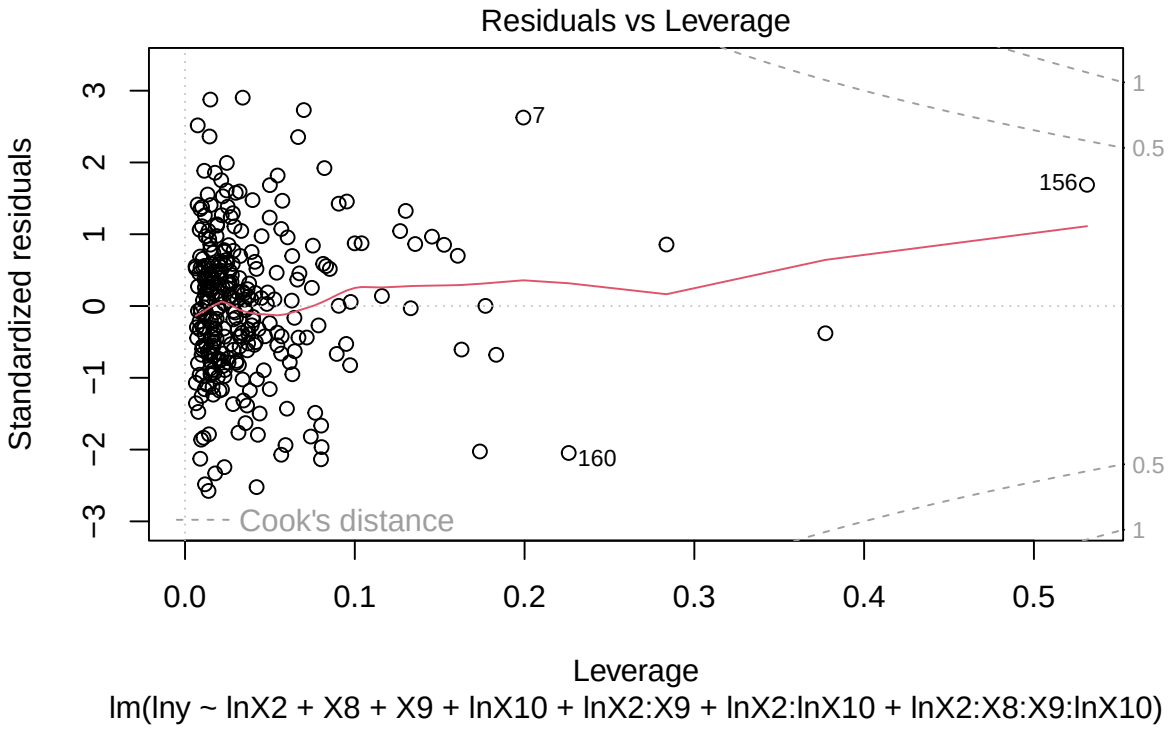



lm(lny ~ lnX2 + X8 + X9 + lnX10 + lnX2:X9 + lnX2:lnX10 + lnX2:X8:X9:lnX10)

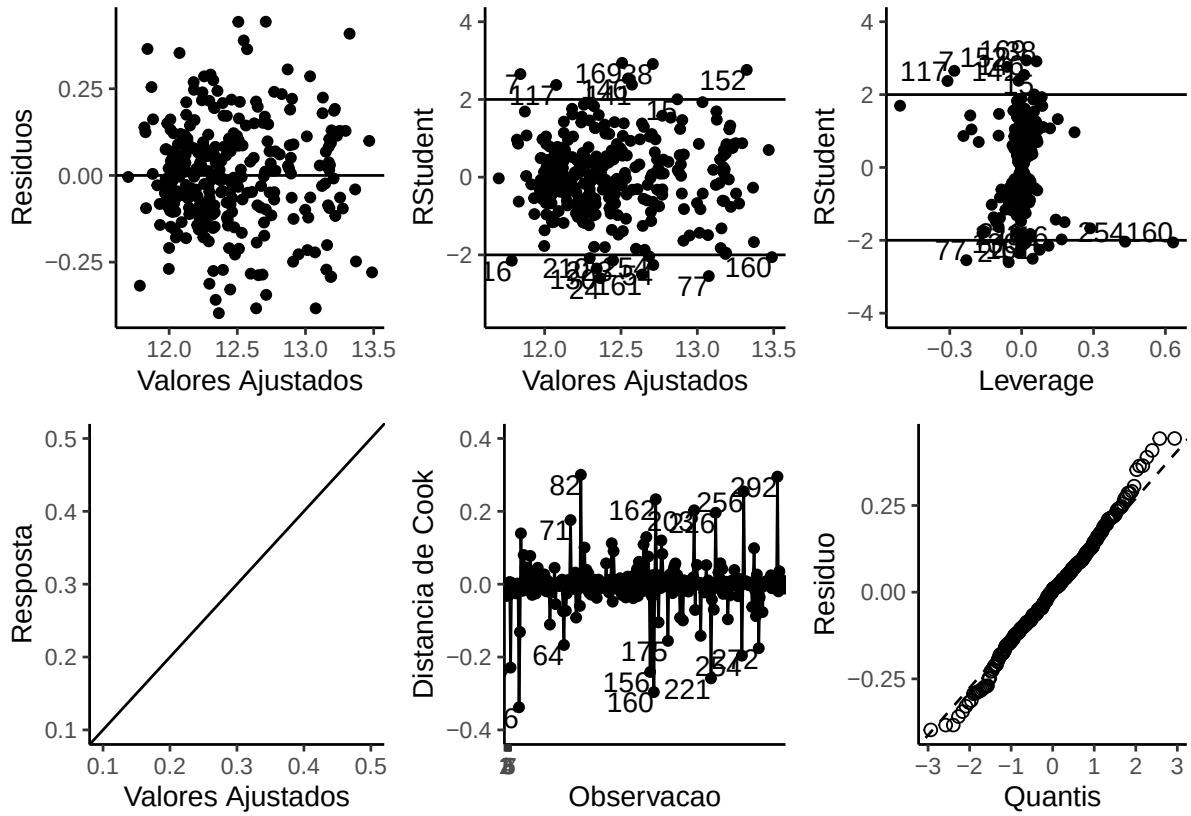

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(lny ~ lnX2 + X8 + X9 + lnX10 + lnX2:X9 + lnX2:lnX10 + lnX2:X8:X9:lnX10)
169
38
152


Scale−Location
√|Standardized residuals|
Fitted values
lm(lny ~ lnX2 + X8 + X9 + lnX10 + lnX2:X9 + lnX2:lnX10 + lnX2:X8:X9:lnX10)
169
38
152




```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## lnX2                 821.1911  1       28.656432
## X8                   447.3133  1       21.149782
## X9                5101962.3926  2       47.526334
## lnX10                816.7249  1       28.578400
## lnX2:X9           4762002.7324  2       46.714038
## lnX2:lnX10           2090.0177  1       45.716711
## lnX2:X8:X9:lnX10   428081.7900  3        8.681361
```




**Intervalo de confiança 95% para os parâmetros do modelo**

```
##                                2.5 %        97.5 %
## (Intercept)          1.459783e+01 40.8354501464
## lnX2                -3.554280e+00 -0.1143517268
## X8                  -4.470198e-02 -0.0039240920
## X92                 -1.018242e+01 -5.7155769338
## X93                 -1.087975e+01 -4.7264120396
## lnX10               -2.475273e+00  0.0551175148
## lnX2:X92             6.298786e-01  1.2026675499
## lnX2:X93             4.845498e-01  1.2658680547
## lnX2:lnX10          -2.058799e-02  0.3196251054
## lnX2:X8:X91:lnX10    8.644369e-05  0.0005834876
## lnX2:X8:X92:lnX10    1.128498e-04  0.0006254943
## lnX2:X8:X93:lnX10    1.175416e-04  0.0006506867
```

## Testando retornar as variáveis originais

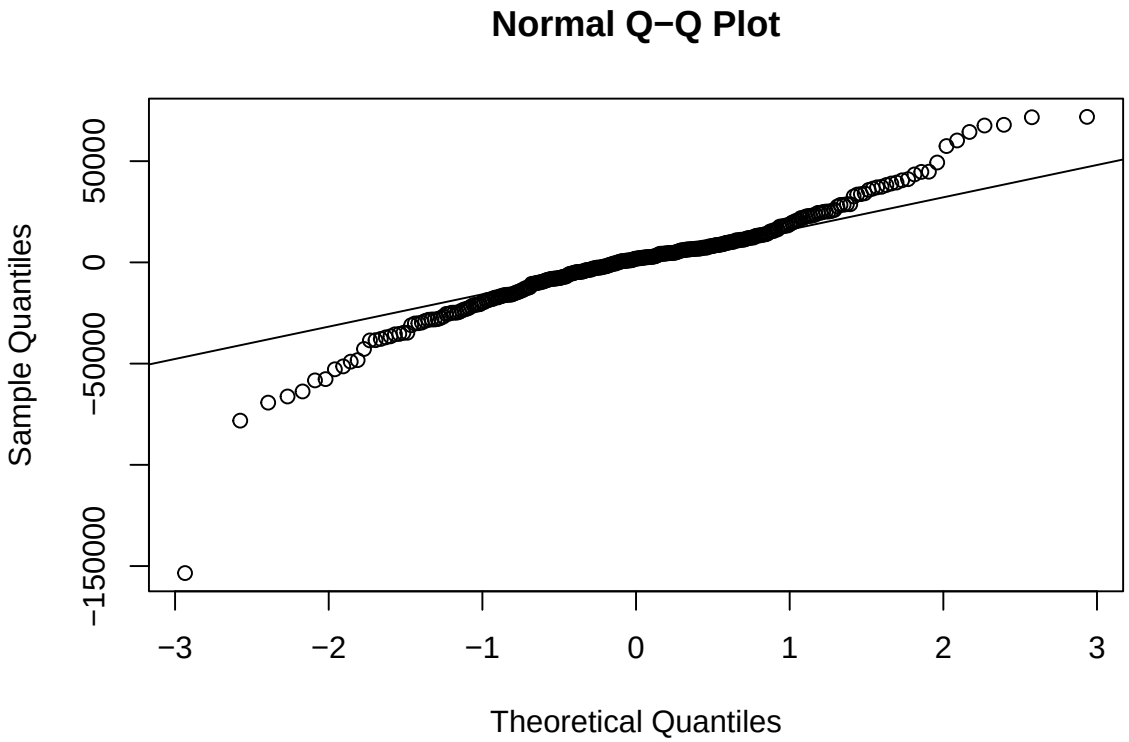
```
##
## Call:
## lm(formula = X1 ~ X2 + X8 + X9 + X10 + X2:X9 + X2:X10 + X2:X8:X9:X1,
##     data = dados)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -153448  -10567    1850   10975   71881
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.244e+05  3.199e+04  22.643  < 2e-16 ***
## X2           -1.247e+02  1.038e+01 -12.008  < 2e-16 ***
## X8           -1.860e+03  1.600e+02 -11.628  < 2e-16 ***
## X92          -3.919e+05  2.338e+04 -16.763  < 2e-16 ***
## X93          -4.760e+05  2.813e+04 -16.917  < 2e-16 ***
## X10           6.807e-01  4.796e-01   1.419    0.157
## X2:X92        8.981e+01  1.043e+01   8.609 4.90e-16 ***
## X2:X93        6.944e+01  1.463e+01   4.746 3.27e-06 ***
## X2:X10        6.995e-06  1.916e-04   0.037    0.971
## X1:X2:X8:X91  2.171e-06  8.558e-08  25.370  < 2e-16 ***
## X1:X2:X8:X92  2.810e-06  1.119e-07  25.112  < 2e-16 ***
## X1:X2:X8:X93  6.481e-06  4.693e-07  13.809  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24980 on 288 degrees of freedom
## Multiple R-squared:  0.9679, Adjusted R-squared:  0.9666
## F-statistic: 788.3 on 11 and 288 DF,  p-value: < 2.2e-16
```

```
## Anova Table (Type III tests)
##
## Response: X1
##                  Sum Sq  Df  F value    Pr(>F)
## (Intercept) 3.1982e+11   1 512.6840 < 2.2e-16 ***
## X2          8.9958e+10   1 144.2039 < 2.2e-16 ***
## X8          8.4347e+10   1 135.2103 < 2.2e-16 ***
## X9          2.0878e+11   2 167.3402 < 2.2e-16 ***
## X10         1.2564e+09   1   2.0140    0.1569
## X2:X9       4.6272e+10   2  37.0872  4.66e-15 ***
## X2:X10      8.3162e+05   1   0.0013    0.9709
## X1:X2:X8:X9 6.6420e+11   3 354.9085 < 2.2e-16 ***
## Residuals   1.7966e+11 288
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## X2            26.62968  1        5.160396
## X8             4.12301  1        2.030520
## X9          1173.03212  2        5.852312
## X10           12.40244  1        3.521709
## X2:X9       1898.24385  2        6.600670
## X2:X10        21.84174  1        4.673515
## X1:X2:X8:X9  841.01729  3        3.072327
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.93376, p-value = 2.589e-10
```




Foi testada a possibilidade de retorno das variáveis do modelo ao seu formato original, uma vez que X2 e X10 foram manipuladas na escala de logaritmo natural. O modelo original possui um ótimo valor de R2 ajustado de 0.9666. Contudo X10 perde sua significância, assim como a interação entre X2 e X10 também.

Porém, a hipótese de normalidade dos resíduos é contundentemente descartada, portanto, é melhor seguir com o modelo com as variáveis numéricas transformadas em logarítmo natural, apenas fazendo a desconversão para apresentar os resultados ao cliente.

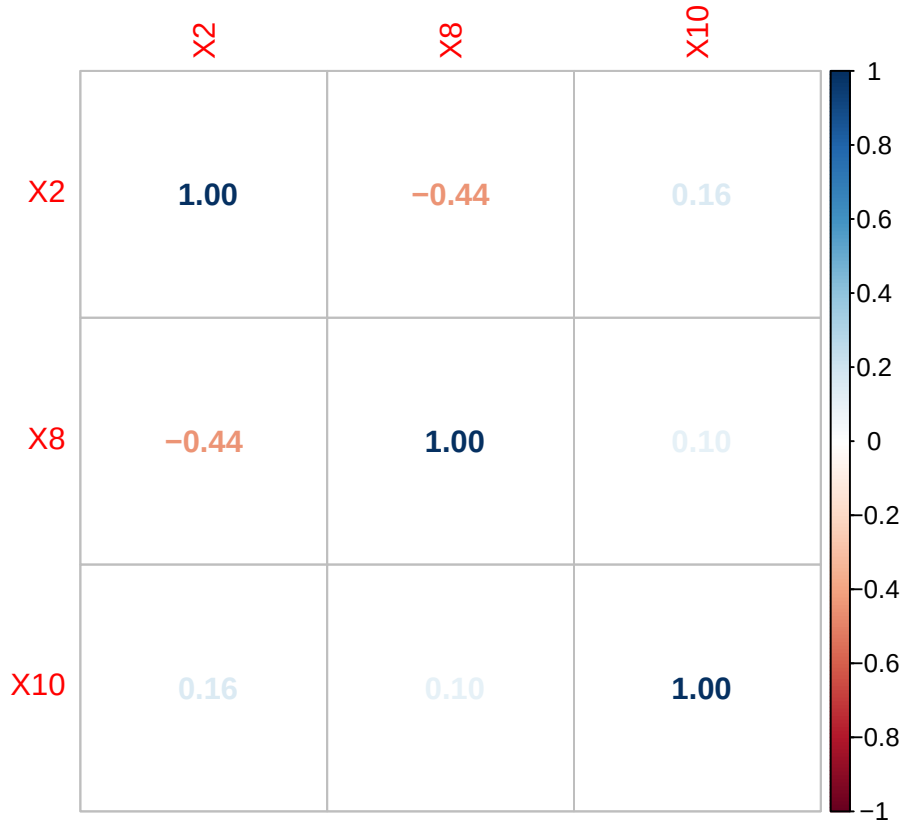
# Multicolinearidade

| | dados.lnX2 | dados.X8 | dados.lnX10 |
|---|---|---|---|
| dados.lnX2 | 1.0000000 | 0.4284081 | 0.2492975 |
| dados.X8 | 0.4284081 | 1.0000000 | -0.0769225 |
| dados.lnX10 | 0.2492975 | -0.0769225 | 1.0000000 |

| | dados.X2 | dados.X8 | dados.X10 |
|---|---|---|---|
| dados.X2 | 1.0000000 | 0.4232893 | 0.2086116 |
| dados.X8 | 0.4232893 | 1.0000000 | -0.1280248 |
| dados.X10 | 0.2086116 | -0.1280248 | 1.0000000 |

### Correlograma das variáveis escolhidas no modelo

Para as variáveis escolhidadas no modelo, tem-se o correlograma a seguir:




Foram geradas as matrizes de correlação entre as variáveis numericas do modelo transformado e o das variáveis originais onde não foi notada grandes evidências de multicolinearidade entre as variáveis. A única diferença seria quanto a correlação entre X8 e lnX10, onde no modelo de variáveis originais ela é um pouco maior, saindo de -0.07 para -0.12, ou seja, ainda que exista diferença não é algo significante. A mesma situação acontece com lnX2 e X8.

# Testando o ajuste do modelo no conjunto de teste

Em ln, o modelo transformado está com um MSE baixíssimo (0.0012098) dos valores preditos pros valores reais. Ao desfazer a transformação e retornar ao modelo original, o MSE sobe para um valor aparentemente alto ($1.1310586 \times 10^7$), mas que deve ser analisado com calma, pela escala da variável.

Concluímos portanto que o modelo treinado com o conjunto de treino de 300 observações prediz bem quando aplicado no conjunto de testes de 222 observações.

## Validação do modelo pelos coeficiente de Mallow.

Foi desenvolvido, para princípio de comparação pelo coeficiente de Mallow 4 modelos comparativos, onde o primeiro é o modelo final gerado com as transformações, o segundo modelo seria o modelo final sem as transformações e, em seguida, os mesmos dois modelos com e sem a transformação, sem as interações das variáveis.

### Coeficiente de Mallow pro modelo sem interações:

61.8879148

### Coeficiente de Mallow pro modelo com as variáveis lnX2, X4, X5, X6, X8, X9 e lnX10:

63.4644446

### Coeficiente de Mallow pro modelo com as variáveis lnX2, X4, X8, X9 e lnX10:

61.4415707

### Coeficiente de Mallow pro modelo com todas as variáveis e todas as interações possíveis entre elas:

134

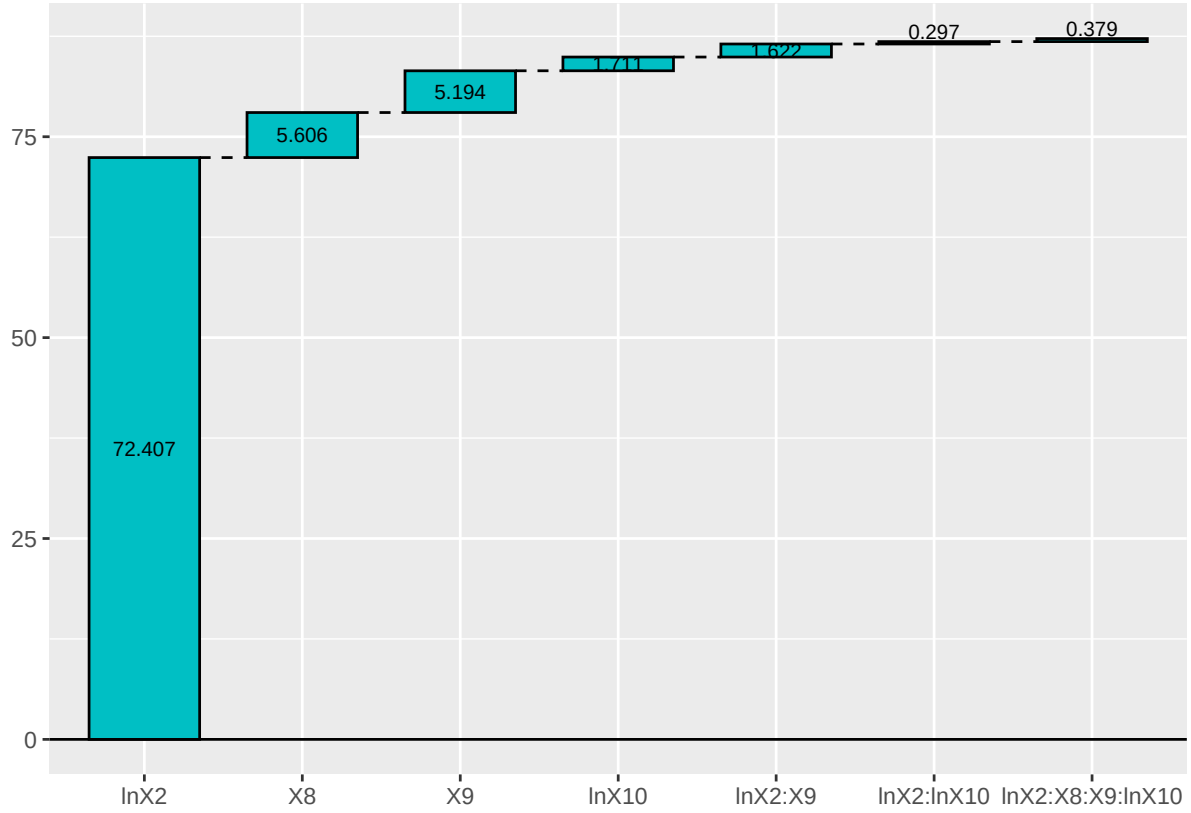
### Coeficiente de Mallow pro modelo selecionado:

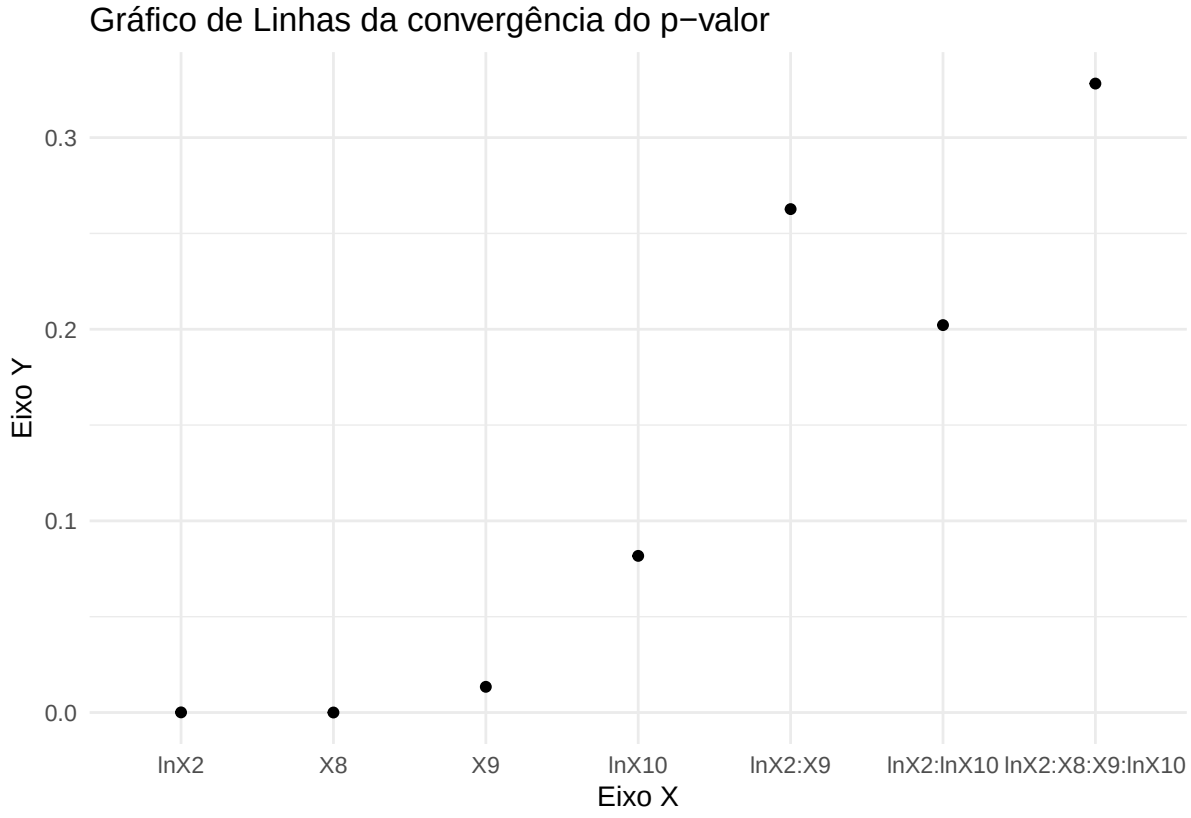
14.2728251

Com isso, concluímos que, dos possíveis modelos analisados, o modelo escolhido é o que apresenta o melhor coeficiente de Mallows, com o valor observado (14.2728251) se aproximando do valor ótimo de igualdade à $p = 12 + 1 = 13$.

## Análise de acrescimo de informação

A partir do gatilho que as interações mostraram ter na análise de validação pelo coeficiente de Mallows, foi análisado o acrescimo de explicação que cada uma das variáveis do modelo podem ter, sendo representada pelo gráfico de cascata abaixo:



Em contrapartida, foi análisado o p-valor do modelo conforme é acrescentada cada uma das variáveis. A partir do momento em que cruzamos as informações dos dois gráficos, vemos que conforme acrescentamos as variáveis obtemos uma explicação maior assim como um maior controle dos resíduos no que diz a convergência a normalidade. A variáveis lnX2 possui uma grande importância na explicação do modelo, contudo seus resíduos são nada controlados. Conforme acrescentamos as variáveis é observado um maior p-valor no teste de normalidade dos resíduos. A partir do momento em que introduzimos as interações das variáveis pertinentes do modelo, vemos que os resíduos se mostram muito mais controlados, demonstrando assim que as interações, apesar de não explicarem muito em termos de R2 ponderado, possuem um poder gigantesco para controlar a normalidade dos resíduos.

# Conclusões

Após testar diversas combinações de variáveis e suas interações para modelagem dos dados, obtivemos um modelo compacto e explicativo para responder o que o exercício nos propunha. Notamos que durante o processo de modelagem, é tentador utilizar o maior número possível de variáveis a fim de se obter a explicação com mais confiabilidade. Entretanto, sabemos que inclusão de variáveis demais leva tanto a problemas de Overfitting, como a adição de variáveis que confundem o modelo. Além disso, é tentador se valer de critérios automáticos para selecionar as variáveis, entretanto, notamos que os métodos automáticos utilizados auxiliaram em alguns momentos na remoção de algumas variáveis, mas sempre foi necessário um olhar estatístico atento às variáveis que sobram pois por vezes estes métodos não utilizam todos os modelos possíveis, e além disso, convém também observar as variáveis removidas sob um olhar humano, visto que nós sabemos a dificuldade e facilidade de coletar cada uma das variáveis, e por vezes temos que optar por variáveis um pouco menos precisas porém de coleta mais simplificada e/ou confiável, do que outras variáveis que por vezes modelos automáticos selecionam mas que são extremamente difíceis de se obter no campo.

No caso do modelo selecionado para esta análise, não foi possível representar graficamente a variável resposta sob as variáveis explicativas, visto que selecionamos 4, mais algumas de suas interações, portanto estaríamos tratando de uma superfície p-dimensional de resposta, o que não é possível representar fidedignamente numa figura de duas dimensões neste documento.

Conclui-se, por fim, que a modelagem regressiva é extremamente útil para obter respostas por vezes difíceis em uma análise menos sofisticada. É bom atentar que, ao longo desta modelagem, fomos limitados a seguir o modelo linear múltiplo e seus pressupostos, o que numa modelagem ainda mais sofisticada, como uma modelagem linear generalizada, não teríamos que seguir.

É importante destacar as limitações e vantagens do modelo. As principais vantagens foi o fato de dispormos de uma gama de variáveis quantitativas que facilita a execução do modelo. Além disso, tivemos acesso aos dados já estruturados, o que dispensa uma etapa importante de ETL. Quanto as limitações, não sabemos exatamente a localidade da coleta dos dados, portanto, tampouco podemos fazer muita inferência sobre o uso deste modelo para predições em geral. Além disso, pelas limitações do modelo regressivo múltiplo, tivemos por vezes que descartar informações que num modelo mais sofisticado, poderiam ser úteis para a inferência dos resultados e sua maior precisão.