Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Relatório Técnico
Análise de Regressão

Adriana Eva Fernandes da Silva

Sumário

Capítulo 1

Contextualização para o estudo em questão

1.1 Introdução

A **Regressão Linear Múltipla** consiste em um conjunto de procedimentos estatísticos que é normalmente utilizado para estudar a relação linear entre um conjunto de covariáveis e uma determinada variável resposta.

É comum utilizar essa técnica estatística para explicar o preço de venda de imóveis a partir de outras variáveis (variáveis explicativas).

O presente estudo tem por objetivo explicar o **preço de venda** de imóveis com base em algumas covariáveis, sendo elas: **Imposto do imóvel**, **Área do terreno**, **Área construída** e **Idade da residência**.

Capítulo 2

Processo de amostragem e coleta de dados

2.1 Amostra

Para esta análise foi considerada uma amostra de **27 imóveis**, onde foram coletadas informações a respeito de 5 variáveis.

Os dados amostrados estão dispostos na Tabela 2.1.

Tabela 2.1: Amostra dos Dados Coletados para o Estudo.

Imposto do imóvel	Área do terreno	Área construída	Idade da residência	Preço de venda do imóvel
4.9176	3.4720	0.9980	42	25.9
5.0208	3.5310	1.5000	62	29.5
4.5429	2.2750	1.1750	40	27.9
4.5573	4.0500	1.2320	54	25.9
5.0597	4.4550	1.1210	42	29.9
3.8910	4.4550	0.9880	56	29.9
5.8980	5.8500	1.2400	51	30.9
5.6039	9.5200	1.5010	32	28.9
15.4202	9.8000	3.4200	42	84.9
14.4598	12.8000	3.0000	14	82.9
5.8282	6.4350	1.2250	32	35.9
5.3003	4.9883	1.5520	30	31.5
6.2712	5.5200	0.9750	30	31.0
5.9592	6.6660	1.1210	32	30.9
5.0500	5.0000	1.0200	46	30.0
8.2464	5.1500	1.6640	50	36.9
6.6969	6.9020	1.4880	22	41.9
7.7841	7.1020	1.3760	17	40.5
9.0384	7.8000	1.5000	23	43.9
5.9894	5.5200	1.2560	40	37.5
7.5422	4.0000	1.6900	22	37.9
8.7951	9.8900	1.8200	50	44.5
6.0931	6.7265	1.6520	44	37.9
8.3607	9.1500	1.7770	48	38.9
8.1400	8.0000	1.5040	3	36.9
9.1416	7.3262	1.8310	31	45.8
12.0000	5.0000	1.2000	30	41.0

A Tabela 2.1 descreve os registros para os 27 imóveis amostrados, onde as carac-

terísticas **Preço de venda**, **Imposto do imóvel**, **Área do terreno**, **Área construída** e **Idade da residência** foram avaliadas.

A Tabela 2.1 sumariza o nome de cada covariável utilizada no modelo, seu tipo, se é qualitativa ou quantitativa e sua descrição

Tabela 2.2: Sumarização dos Dados.

Nome da variável	**Tipo da variável**	**Descrição da variável**
Imposto do imóvel	Quantitativa contínua	Covariável em 100 USD
Área do terreno	Quantitativa contínua	Covariável em 1.000 pés quadrados
Área construída	Quantitativa contínua	Covariável em 1.000 pés quadrados
Idade da residência	Quantitativa discreta	Covariável em anos
Preço de venda do imóvel	Quantitativa contínua	Variável resposta em 1.000 USD

Podemos observar pela Tabela 2.2 que a presente análise possui 4 covariáveis, sendo 3 delas do tipo quantitativa contínua e uma quantitativa discreta, além disso, temos uma variável reposta do tipo quantitativa contínua.

A covariável Imposto do imóvel foi medida em 100 USD, as covariáveis Área do terreno e Área construída foram medidas em 1.000 pés quadrados, a covariável Idade da residência foi medida em anos e a variável Preço de venda do imóvel foi medida em 1.000 USD.

2.2 Busca de dados faltantes

Na descrição do conjunto de dados é realizado o entendimento da base de dados. Aqui é comum entender as covariáveis e verificarmos se existe dados faltantes. Após a busca por estes, percebe-se que não existem covariáveis sem preenchimento.

Capítulo 3

Objetivos da análise estatística

3.1 Objetivos

O objetivo principal desta análise é estudar se existe relação linear entre a variável preço de venda do imóvel (em 1.000 USD) e as covariáveis apresentadas anteriormente. Ou seja, temos interesse em avaliar se a variável resposta preço da venda pode ser explicada pelas covariáveis presentes na base de dados sob estudo e quais delas são significativas para um modelo de regressão linear múltipla.

Capítulo 4

Metodologia estatística

4.1 Metodologia

O presente estudo caracteriza-se pelo desenvolvimento de uma metodologia estatística aplicada na investigação e modelagem da relação linear entre variáveis, essa metodologia consiste na **Regressão Linear Múltipla**.

4.1.1 Análise Descritiva

Para o desenvolvimento desse estudo foi utilizado o *software* estatístico *Rstudio* com intuito de realizar as análises necessárias, além disso, foi utilizado um conjunto de dados (disponibilizado pelo docente) contendo os registros das 5 variáveis em que as unidades amostrais foram avaliadas. Como discutido no Capítulo 2 as variáveis observadas foram: **Preço de venda**, **Imposto do imóvel**, **Área do terreno**, **Área construída** e **Idade da residência**, e as unidades amostrais foram os **27 imóveis**.

Inicialmente, foi realizada a análise descritiva e exploratória dos dados para cada uma das variáveis explicativas, onde as covariáveis quantitativas serão resumidas em média, desvio padrão, mediana e quartis.

Os gráficos utilizados para obtermos indicativos dos resultados sobre cada covariável quantitativa são *boxplots*.

O limite de detecção de outliers é construído utilizando o intervalo interquartílico, dado pela distância entre o primeiro e o terceiro quartil. Sendo assim, os limites inferior e superior de detecção de outlier são dados pelas Equações 4.1 e 4.2:

$$LimiteInferior = PrimeiroQuartil – 1.5 * (TerceiroQuartil – PrimeiroQuartil), \quad (4.1)$$

$$LimiteSuperior = TerceiroQuartil + 1.5 * (TerceiroQuartil – PrimeiroQuartil). \quad (4.2)$$

Os gráficos de pontos são aplicados para verificar a relação das covariáveis em relação a variável resposta com o intuito de analisar se as covariáveis estão bem correlacionadas com o Preço de venda do imóvel.

4.1.2 Análise de Regressão

Em seguida, será realizada a análise de regressão para estudar a relação entre a variável resposta com o respectivo conjunto de covariáveis.

Padronização das Variáveis

O primeiro passo para a análise será verificar se existe a necessecidade da padronização das variáveis caso não possuam a mesma unidade de medida, afim de controlar os erros de arredondamento nos cálculos das equações normais e para permitir as comparações das estimativas dos coeficientes do modelo de regressão. Para tanto, foi utilizada a **transformação correlação** dada pelas Equações 4.3 e 4.4 :

$$Y_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{Y_i - \bar{Y}}{s_Y} \right), i = 1, ..., n, \tag{4.3}$$

e

$$x_i^* = \frac{1}{\sqrt{n-1}} \left(\frac{x_{ik} - \bar{x_k}}{s_k} \right), i = 1, ..., n, k = 1, ..., p-1. \tag{4.4}$$

Multicolinearidade

Para verificar se existe multicolinearidade no modelo será calculada a matriz $X^{*T}X^{*}=$ r_{xx}, que é a matriz de correlação entre as covariáveis, pela Matriz 4.5:

$$r_{xx} = \begin{bmatrix} 1 & r_{12} & ... & r_{1,p-1} \\ r_{21} & 1 & ... & r_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p-1,1} & r_{p-1,2} & ... & 1 \end{bmatrix} \tag{4.5}$$

Em que r_{ij} é a correlação amostral entre variável X_i e X_j sendo $i, j = 1, 2, ..., p-1$, e $X^{*T}y* = r_{yi}$ é correlação amostral entre a variável resposta Y e a covariável X_i, dada pela Equação 4.6 a seguir:

$$r_{YX} = [r_{Y1}, r_{Y2}, ..., r_{Y,p-1}]^T \tag{4.6}$$

A partir da inversa da matriz de correlação entre as covariáveis é possível analisar se há a presença de multicolinearidade no modelo.

Além disso, para ter mais assertividade na análise, será calculado o VIF - Fator de Inflação da Variância, por meio da Equação 4.7, que consiste em um problema comum em regressões, onde as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas. As consequências da multicolinearidade em uma regressão são a de

erros padrão elevados no caso de multicolinearidade moderada ou severa e até mesmo a impossibilidade de qualquer estimação se a multicolinearidade for perfeita.

Para verificar a existência de multicolinearidade, alta correlação entre as variáveis, é utilizado o fator de inflação de variância (VIF). Valores de VIF superiores a 10 indicam presença de multicolinearidade, quando isso ocorre é aconselhável a reformulação do modelo.

Para eliminar a multicolinearidade e evitar estimativas errôneas dos coeficientes, realiza-se uma análise das variáveis com alta correlação, identificando e excluindo-as do ajuste do modelo,

$$VIF_k = \frac{Var(\hat{B}_k^*)}{(\sigma^*)^2}. \tag{4.7}$$

Onde, o menor valor de VIF_k é 1 que indica a completa ausência de multicolinearidade. O máximo VIF_k entre todas as p-1 covariáveis do modelo é usado como indicador da severidade da multicolinearidade. Se $máx(VIF_k)$ for maior que 10, significa que há presença de multicolinearidade.

Análise de Diagnóstico para o modelo completo

Para ser certificado se as suposições do modelo de regressão linear estão sendo satisfeitas, ou seja, linearidade, homocedasticidade e independência dos erros, será feita uma análise nos resíduos. Sendo o resíduo igual ao valor observado menos o valor predito, dado pela Equação 4.8.

$$\epsilon_i = y_i - \hat{y}_i. \tag{4.8}$$

Por meio do Gráfico Qqnorm, será verificado a suposição de normalidade, pelo Gráfico de Preditos *vs* Resíduos pode-se verificar tanto a suposição de homocedasticidade como a de linearidade, e a suposição de independência através do Gráfico da ordem de seleção das observações *vs* os resíduos se esta informação estiver disponível. Além disso, também serão realizados testes de normalidade, sendo estes os testes Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling e Lilliefors. O teste para homocedasticidade que será feito é o teste de Breusch.

Seleção de Covariáveis

Para identificar o menor grupo de covariáveis que são mais significativas para explicar a variável resposta Y, serão utilizados critérios para fazer a seleção do modelo de regressão.

O **Critério** R_p^2 dado pela Equação 4.9 é o coeficiente de determinação do modelo que é calculado para encontrar o conjunto de covariáveis com R^2 suficientemente grande e que seja o ponto onde adicionar mais covariáveis no modelo não é necessário.

$$R^2 = \frac{SQReg}{SQTotal} = 1 - \frac{SQRes}{SQTotal}. \tag{4.9}$$

Em que $SQReg$, $SQTotal$ e $SQRes$ indicam soma dos quadrados da regressão, soma dos quadrados total e soma dos quadrados dos resíduos, respectivamente.

Também será calculado o **Critério** $R^2_{a,p}$ dado pela Equação 4.10 que é o coeficiente de determinação ajustado. Assim como o critério anterior, o objetivo é encontrar o conjunto de covariáveis com R^2_a suficientemente grande ao ponto de que adicionar mais covariáveis no modelo não levam a uma mudança significativa no coeficiente.

$$R^2_a = 1 - \frac{n-1}{n-p} \cdot \frac{SQRes}{SQTotal}. \tag{4.10}$$

Em que n e p indicam o tamanho amostral e o número de parâmetros, respectivamente.

Além dos critérios mencionados anteriormente será utilizado o **Critério de Mallows**, que é descrita pela Equação 4.11.

$$C_p = (n - P)\frac{SQRes_p}{SQRes_P} - (n - 2p). \tag{4.11}$$

Em que P é o número de parâmetros do modelo mais completo com todas as covariáveis disponíveis. Queremos obter o subconjunto de covariáveis cujo modelo possui o menor valor de C_p, sendo que, modelos com C_p pequeno possuem um $SQRes_p$ pequeno. Uma propriedade para a Equação 4.11 é que conforme o $SQRes$ do modelo reduzido for mais próximo do $SQRes$ do modelo completo, o valor de C_p se aproxima de p.

Outro critério de seleção é o **AIC (Akaike's Information Criterion)** e o **BIC (Bayesian Information Criterion)**.

$$AIC_p = n \log SQRes_p - n \log n + 2p, \tag{4.12}$$

e

$$BIC_p = n \log SQRes_p - n \log n + p \log n. \tag{4.13}$$

Por meio dos critérios dado pelas Equações 4.12 e 4.13 será avaliado qual é o modelo com as covariáveis que minimiza o AIC e BIC. Essas medidas também penalizam a adição de novas covariáveis no modelo.

Além dos critérios citados acima será utilizado um procedimento para seleção automática de modelo. Esse método é chamado Backward e, geralmente é empregado para conjuntos de dados com um número pequeno de covariáveis, uma vez que se inicia o teste da seleção das covariáveis com o modelo completo, e vamos eliminando a cada passo as covariáveis que são menos essenciais para o modelo. Sendo assim, quando o número de covariáveis é pequeno, utilizando o método Backward menos passos serão necessários para chegar no modelo selecionado.

Esse método consiste em iniciar com todas as covariáveis (Modelo Completo) e é definido um nível de significância α para a saída de variáveis no modelo (α_s) onde é testado $H_0 : \beta_k = 0$ *versus* $H_1 : \beta_k \neq 0$. Feito isso, são removidas as variáveis menos significativas, que possuem o valor-p mais alto, onde sua eliminação causa a menor queda em R^2 do modelo e o menor aumento no $SQRes$ em comparação com outros preditores.

Essa eliminação é feita até que as covariáveis restantes no modelo tenham um valor-p menor do que α_s pré-estabelecido, sendo este o modelo final.

Análise de Diagnóstico para o modelo selecionado

Após selecionado o modelo, será conduzido novamente uma análise diagnóstica com os resíduos, para ser possível concluir se os erros atendem às suposições do modelo. Logo, será realizado as mesmas análises gráficas e testes feitos para o modelo completo.

4.1.3 Nomenclatura das variáveis

Para efeito de compreensão e facilitação do presente estudo as covariáveis e a variável resposta serão referenciadas, muitas vezes ao longo deste estudo, como X_1, X_2, X_3 e X_4, sendo estas as covariáveis e a variável resposta como Y, como mostra a Tabela 4.1.

Tabela 4.1: Nomenclatura das variáveis.

Nome da Variável	**Nomenclatura**
Imposto do imóvel	X_1
Área do terreno	X_2
Área construída	X_3
Idade da residência	X_4
Preço do imóvel	Y

Capítulo 5

Resultados

Agora, será apresentado os resultados da análise desenvolvida, que foi dividia entre **Análise descritiva e exploratória de dados** e **Análise de Regressão**.

5.1 Análise descritiva e exploratória de dados

Inicialmente, será apresentada a análise descritiva e exploratória do conjunto de dados, a fim de estudar o comportamento das variáveis envolvidas no estudo.

Como etapa inicial da análise descritiva, foi construído o Gráfico 5.1 de correlações das variáveis.

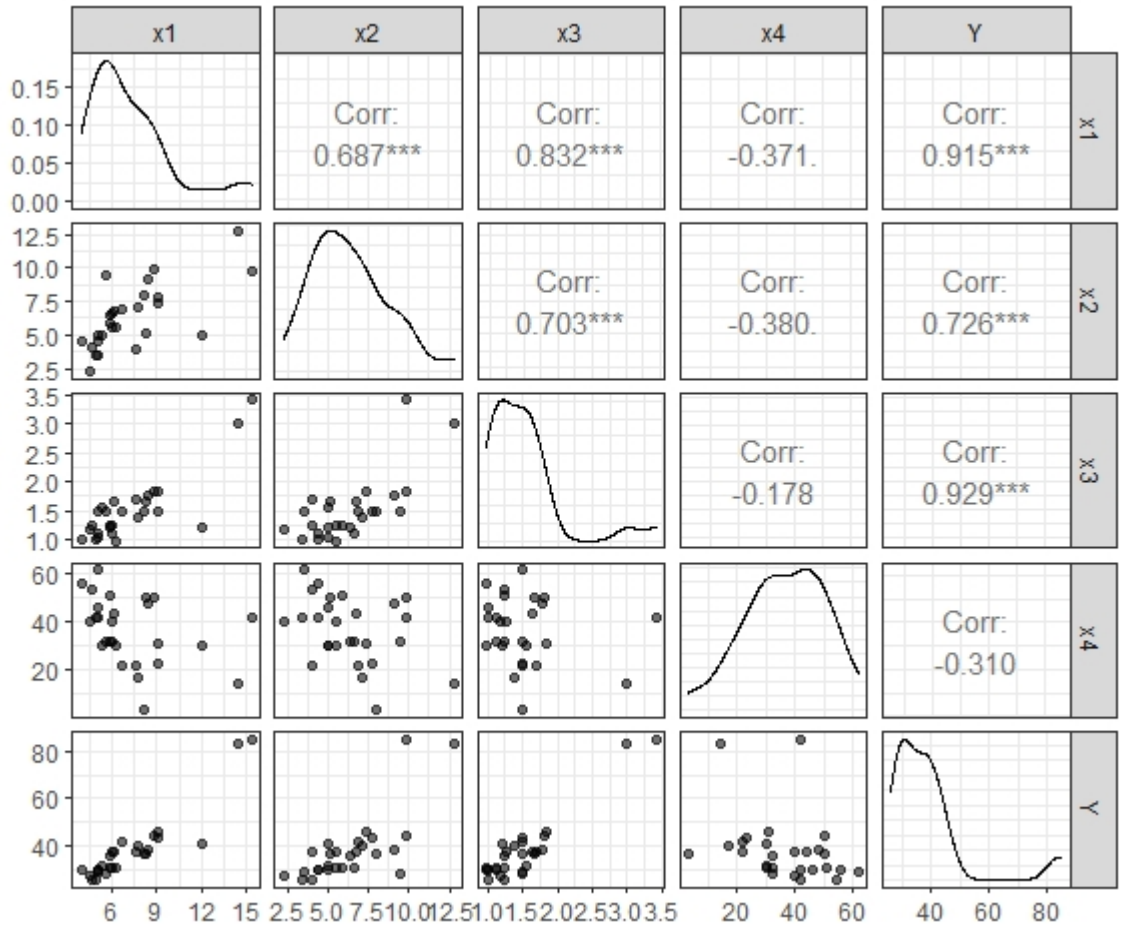

Figura 5.1: Gráfico de correlações.

Pode-se observar pela Figura 5.1 que, no triângulo inferior, estão as correlações das variáveis e, no triângulo superior seus respectivos valores de correlação, na diagonal principal está representada a densidade para cada covariável (X_1, X_2, X_3 e X_4) e a variável resposta Y.

Nota-se que há uma correlação forte de 0.832 entre as covariáveis X_1 (imposto do imóvel) e X_3 (área construída), onde há indícios de que exista multicolinearidade, com base na amostra observada.

Ainda, tem-se um indicativo de que a variável resposta Y (preço do imóvel) e a covariável X_4 (idade da residência) são pouco relacionadas, apresentando uma correlação de -0.310.

Além disso, podemos perceber que X_1 e X_3 possuem forte correlação com a variável resposta Y de 0.915 e 0.929, respectivamente, o que pode ser notado também pelos gráficos de dispersão das mesmas, em que observa-se que a nuvem de pontos está mais próxima de uma reta.

Através dos gráficos de dispersão há indicativos de que todas as covariáveis tem uma relação linear positiva crescente em relação à variável resposta, exceto apenas pela covariável X_4 que aparentemente não tem relação linear com a variável resposta Y.

Agora, será realizada a análise descritiva para cada covariável separadamente, por

meio de gráficos de *boxplots* e tabelas resumindo algumas medidas descritivas.

A seguir está representado o gráfico *boxplot* e sua respectiva tabela resumo para a covariável **Imposto do Imóvel**.

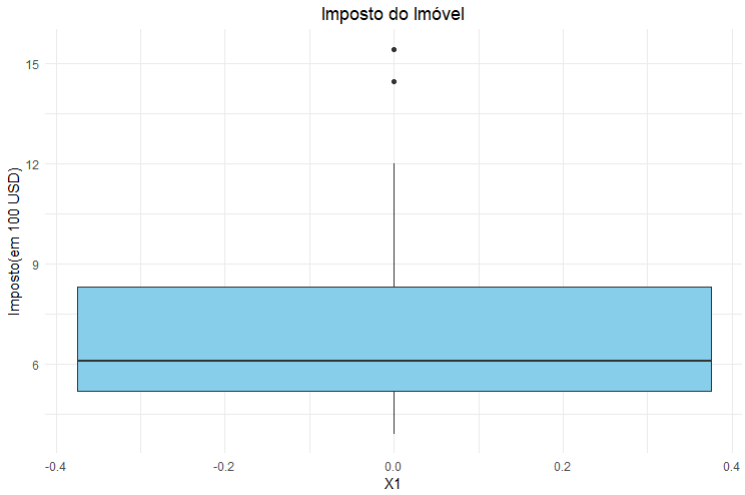

Figura 5.2: Box-plot Imposto do Imóvel.

Tabela 5.1: Tabela resumo Imposto do Imóvel.

Mínimo	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
3.891	5.180	6.093	7.245	8.304	15.420

A partir da Figura 5.2 e sua respectiva Tabela resumo 5.1, verifica-se que o valor médio do imposto do imóvel é de 7.245, ainda pode-se observar que a caixa é assimétrica positiva, uma vez que a linha da mediana está próxima ao primeiro quartil, equivalente a 6.093, comparado aos valores abaixo da mediana. Também vê-se que há 2 outliers presentes na variável.

A seguir está representado o gráfico *boxplot* e sua respectiva tabela resumo para a covariável **Área do terreno**:

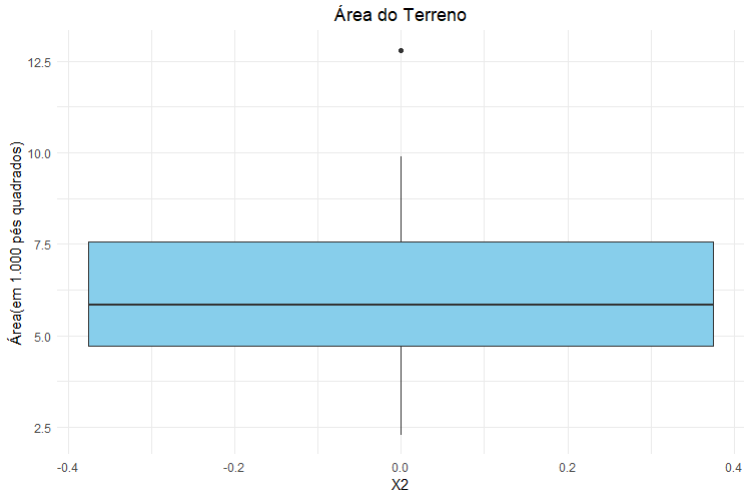

Figura 5.3: Box-plot Área do terreno.

Tabela 5.2: Tabela resumo Área do terreno.

Mínimo	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
2.275	4.722	5.850	6.348	7.563	12.800

A partir da Figura 5.3 e sua respectiva Tabela resumo 5.2, verifica-se que o valor médio da área do terreno é de 6.348, nota-se também que a caixa é quase simétrica uma vez que a linha da mediana está praticamente no centro do retângulo, igual a 5.850, além de apresentar um outlier correspondente ao valor de 12.8.

A seguir está representado o gráfico *boxplot* e sua respectiva tabela resumo para a covariável **Área contruída**:

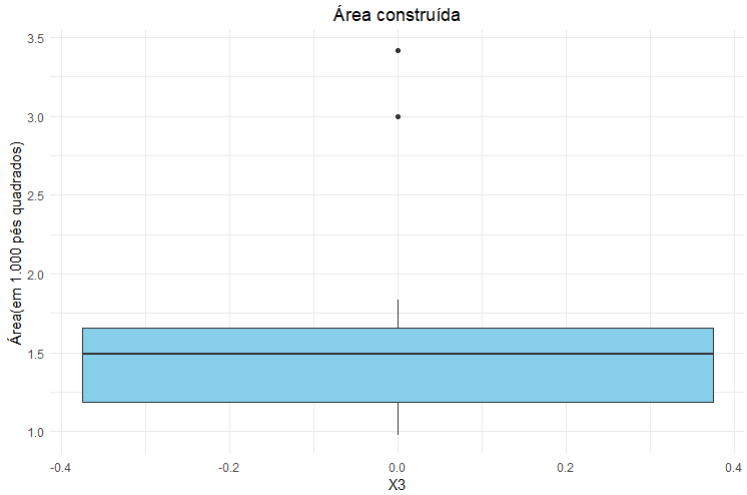

Figura 5.4: Box-plot Área construída.

Tabela 5.3: Tabela resumo Área construída.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0.975	1.188	1.488	1.512	1.658	3.420

A partir da Figura 5.4 e sua respectiva Tabela resumo 5.3, verifica-se que o valor médio da área construída é de 1.512, nota-se também que os dados são assimétricos negativos, com mediana igual a 1.488, além de apresentar dois outliers.

A seguir está representado o gráfico *boxplot* e sua respectiva tabela resumo para a covariável **Idade da resisidência**:

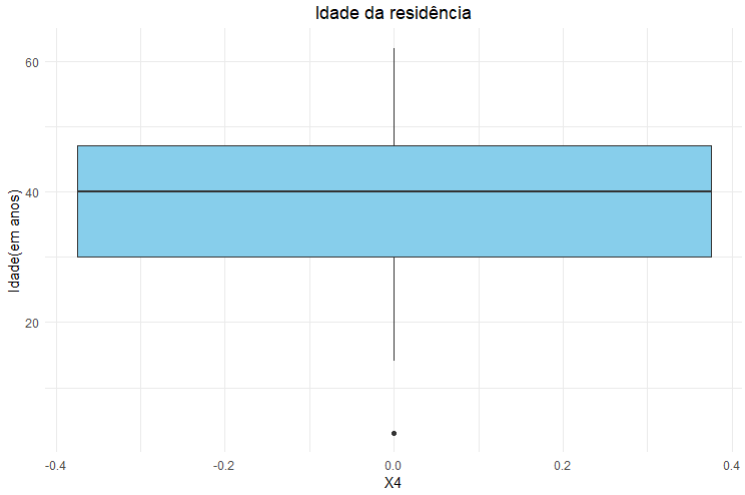

Figura 5.5: Box-plot Idade da residência.

Tabela 5.4: Tabela resumo Idade da residência.

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
3.00	30.00	40.00	36.48	47.00	62.00

A partir da Figura 5.5 e sua respectiva Tabela resumo 5.4, verifica-se que o valor médio da idade da residência é de 36.38, nota-se também que a caixa é assimétrica negativa, com mediana igual a 40, além de apresentar um outlier igual a 3.

A seguir está representado o gráfico *boxplot* e sua respectiva tabela resumo para a covariável **Preço de venda do imóvel**:

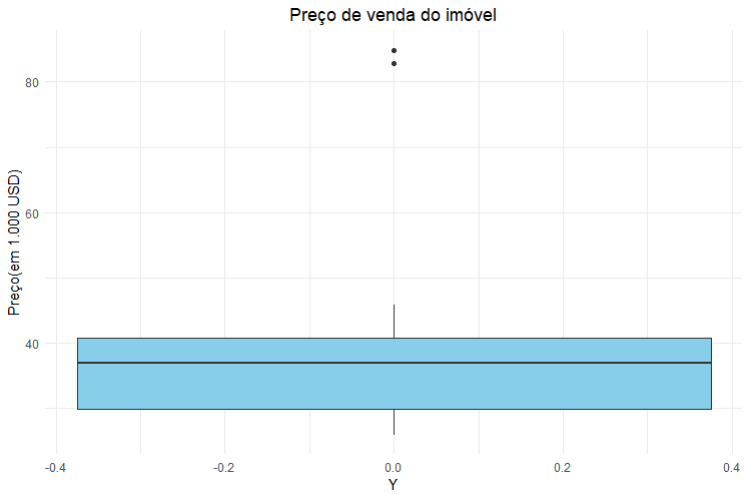

Figura 5.6: Box-plot Preço de venda do imóvel.

Tabela 5.5: Tabela resumo Preço da venda do imóvel.

Mínimo	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
25.90	29.95	36.90	38.50	40.75	84.90

A partir da Figura 5.6 e sua respectiva Tabela resumo 5.5 verifica-se que o valor médio do preço da venda do imóvel é de 38.50, nota-se também que os dados são assimétricos negativos, com mediana igual a 36.90, além de apresentar dois outliers.

A partir das análises univariada de cada variável e bivariada entre a variável resposta e as covariáveis, esperamos que os resultados obtidos confirmem o indicativo de que exista multicolinearidade. Já que X_1 e X_3 apresentaram uma correlação relativamente forte, cremos que alguma delas seja excluída do modelo, pois X_1 contém muito da mesma informação que X_3.

Além disso, ainda esperamos que a covariável X_4 seja excluída na etapa da escolha do modelo, já que ela possui uma correlação baixa com a variável resposta, ou seja, não é significante para explicar a variável resposta.

5.2 Análise de Regressão

Em seguida, foi desenvolvida a análise de regressão para estudar a relação linear entre a variável resposta Y com o conjunto de covariáveis analisado.

5.2.1 Padronização das Variáveis

Uma forma padronizada do modelo de regressão múltipla geral é empregado a fim de controlar erros de arredondamento nos cálculos das equações normais e para permitir comparações das estimativas dos coeficientes do modelo de regressão em uma unidade de medida comum.

Com base no conjunto de dados, observa-se que as covariáveis não possuem a mesma unidade de medida, uma vez que o imposto do imóvel X_1 foi medido em 100 USD, as covariáveis X_2 e X_3, Área do terreno e Área construída, respectivamente, foram medidas em 1.000 pés quadrados, a Idade da residência X_4 representada em anos e a variável resposta Y Preço de venda do imóvel medida em 1.000 USD. Sendo assim, com o intuito de deixar as variáveis em uma unidade de medida comum, realizou-se a padronização das mesmas.

A partir da **transformação correlação** foi gerada a matriz com os valores das covariáveis e da variável resposta padronizada 5.1

$$\begin{bmatrix} X_1 & X_2 & X_3 & X_4 & Y \\ -0.15837018 & -0.234616769 & -0.180591996 & 0.07701286 & -0.172711022 \\ -0.15132751 & -0.229802955 & -0.004241570 & 0.35611986 & -0.123365016 \\ -0.18384626 & -0.332280079 & -0.118412663 & 0.04910216 & -0.145296574 \\ -0.18286641 & -0.187457711 & -0.098388810 & 0.24447706 & -0.172711022 \\ -0.14868056 & -0.154413734 & -0.137382629 & 0.07701286 & -0.117882126 \\ -0.22820486 & -0.154413734 & -0.184104953 & 0.27238776 & -0.117882126 \\ -0.09163835 & -0.040595591 & -0.095578444 & 0.20261101 & -0.104174902 \\ -0.11165041 & 0.258839953 & -0.003890274 & -0.06254064 & -0.131589350 \\ 0.55630069 & 0.281685172 & 0.670246116 & 0.07701286 & 0.636015192 \\ 0.49095017 & 0.526455372 & 0.522701935 & -0.31373695 & 0.608600744 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \quad (5.1)$$

Podemos perceber, pela matriz das variáveis padronizadas 5.1, que os valores das covariáveis e da variável resposta padronizados estão compreendidos no intervalo de -1 e 1.

Assim sendo, com as variáveis na mesma unidade de medida podemos utiliza-lás para as próximas etapas do estudo.

Além disso, vale ressaltar que a matriz 5.1 representa as primeiras 10 linhas, apenas, mas a matriz completa com as variáveis padronizadas pode ser obtida por meio dos códigos processados para as análises.

5.2.2 Análise da Existência de Multicolinearidade

Quando as covariáveis presentes em um modelo de regressão são altamente correlacionadas dizemos que existe multicolinearidade, ou seja, elas explicam a variável resposta da mesma maneira. Nesta etapa é verificado a existência de multicolinearidade no modelo, em que r_{xx}^{-1} é a inversa da matriz de correlação entre as covariáveis. Ao pegarmos

as diagonais dessa matriz encontramos o VIF (fator de inflação da variância) dessas covariáveis.

Para encontrar r_{xx}^{-1} utilizamos o *software* R em que os resultados são apresentados na matriz 5.2

$$r_{xx}^{-1} = \begin{bmatrix} 3.973446 & -0.3864410 & -2.8909027 & 0.8129830 \\ -0.386441 & 2.3194880 & -1.2166129 & 0.5219943 \\ -2.890903 & -1.2166129 & 4.1191236 & -0.8021212 \\ 0.812983 & 0.5219943 & -0.8021212 & 1.3574037 \end{bmatrix} \tag{5.2}$$

VIF_1= 3.973446, VIF_2= 2.319488, VIF_3= 4.119124 e VIF_4= 1.357404.

Com base nos valores da diagonal da Matriz (5.2), verifica-se que VIF_4= 1.357404 é bem próximo de 1, dando indícios de ausência de colinearidade para a variável X_4 idade da residência. Para as demais covariáveis, repara-se que o VIF tem valores maiores, contudo, como nenhum é muito próximo de 10, foi decidido não excluir nenhuma covariável do modelo, porque apesar de ter sido verificado uma correlação significante entre algumas covariáveis pela matriz r_{XX}, obteve-se um baixo valor de VIF para as covariáveis.

5.2.3 Análise de Resíduos no Modelo Completo

Para dar seguimento com a seleção dos modelos, será verificado se as suposições do modelo completo são atendidas. Pois ao fazer a seleção das covariáveis será necessário conduzir testes, os quais são realizados tendo como estatísticas de teste que tem distribuição *t student* e testes com a estatística de teste tendo uma distribuição F, para chegarmos nessas distribuições de estatística é necessário que os erros sejam normais, homocedásticos e independentes. Enquanto que a linearidade deve ser satisfeita já que o modelo utilizado se trata de um modelo de regressão linear.

Dessa forma, será verificado se as suposições de normalidade, homocedasticidade, linearidade e independência dos erros do modelo completo são atendidas através dos resíduos, então foram realizadas análises gráficas e aplicação de testes de hipóteses para comprovar essas suposições.

Normalidade

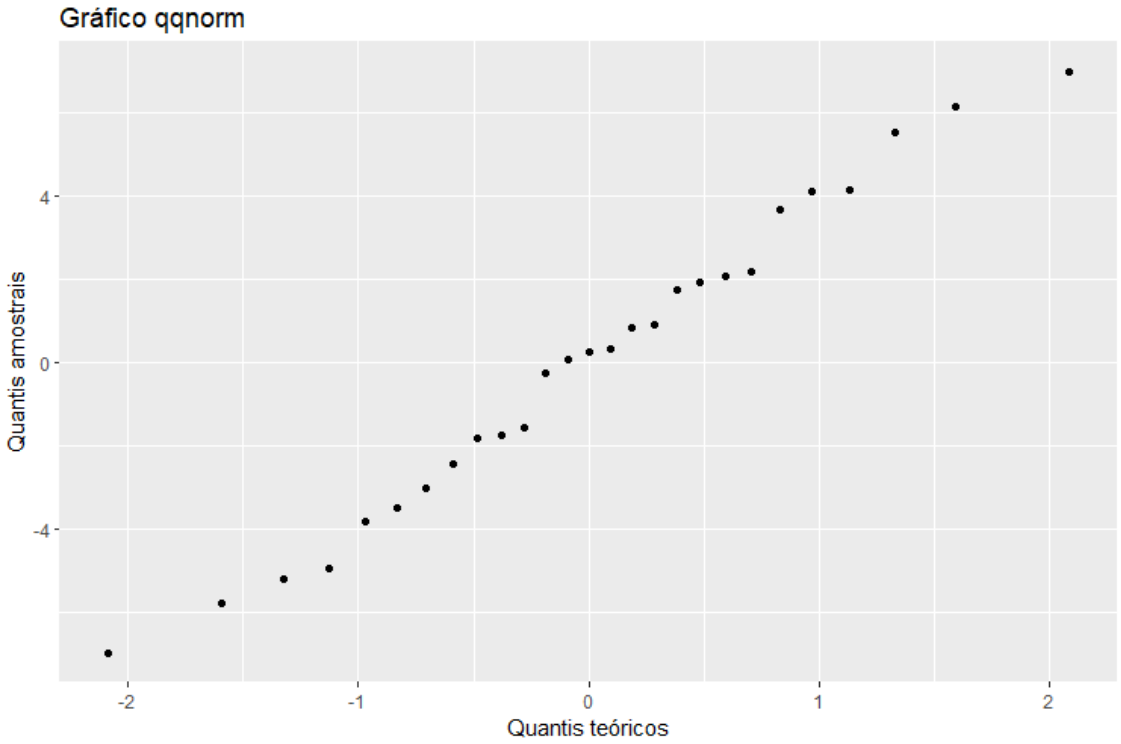

Figura 5.7: Gráfico para Veificar Normalidade.

Pode-se verificar através da Figura 5.7 que os resíduos padronizados observados são distribuídos de maneira a formar uma reta, indicando um formato simétrico da distribuição. Logo, temos um indicativo, com base na amostra observada, de que a hipótese de normalidade é satisfeita.

Tabela 5.6: Testes de Hipótese para Normalidade.

Teste	**Valor - p**	**Conclusão**
Lilliefors	0.984	Não rejeita H_0
Shapiro - Wilk	0.8861	Não rejeita H_0
Kolmogorov - Smirnov	0.9983	Não rejeita H_0
Anderson - Darling	0.953	Não rejeita H_0

Por meio da Tabela 5.6 não rejeitamos a hipótese nula para todos os testes de normalidade ao nível de significância de 5%. Ou seja, há evidências de que a suposição de normalidade dos resíduos é satisfeita.

Homocedasticidade e Linearidade.

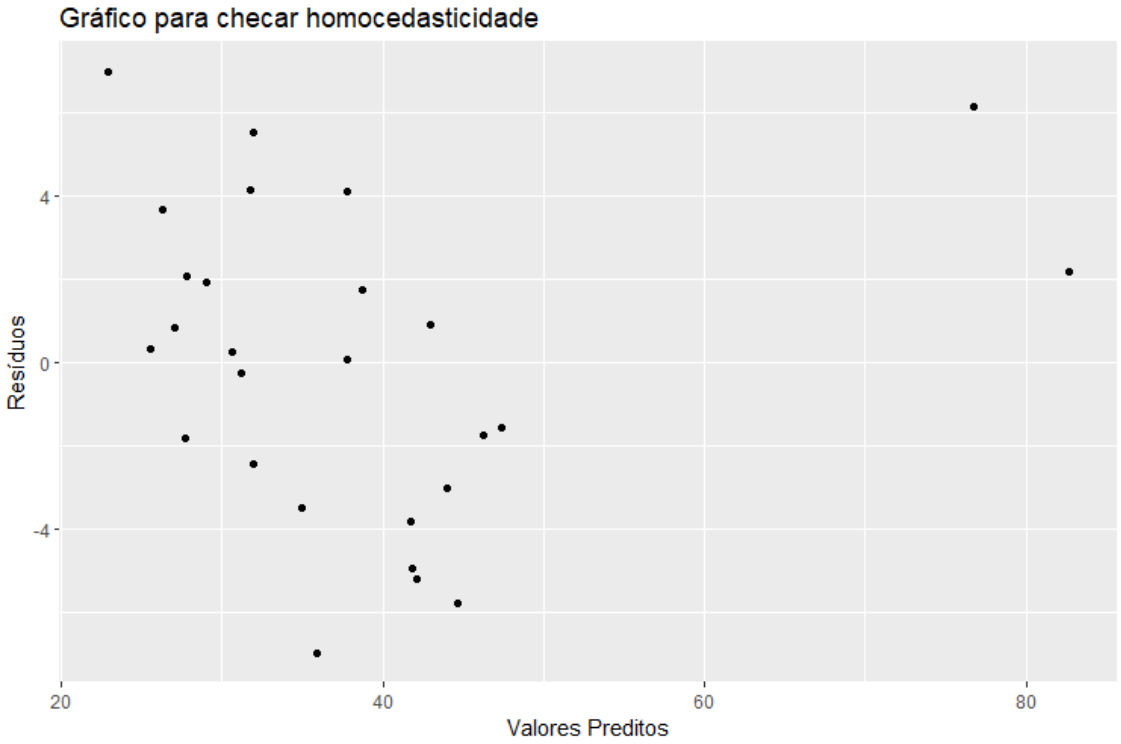

Figura 5.8: Gráfico dos resíduos vs. Valores preditos.

Através da Figura 5.8 tem-se os valores preditos *versus* os resíduos, onde é possível verificar que as observações estão distribuídas em torno do valor 0 sem apresentar oscilações significativas.

Além disso, é possível notar que os pontos não possuem qualquer tendência particular. Assim há indícios de que as hipóteses de homocedasticidade e linearidade dos resíduos não é violada.

Foi realizado ainda, o teste de Breusch - Pagan para a verificação de homocedasticidade, obtendo um valor - p = 0.3131. Ao nível de significância $\alpha = 0.05$, não rejeitamos H_0, portanto há evidências de que a suposição de homocedasticidade não foi violada.

Independência

O banco de dados não traz a informação da ordem seleção dos indivíduos. Logo, não é possível verificar se há independência, portanto será assumido que a suposição de independência dos resíduos está satisfeita.

Como todas as hipóteses sobre os erros estão satisfeitas, então podemos avançar com o processo de seleção das covariáveis.

5.2.4 Selecionando Covariáveis para o Modelo

Nesta seção foram determinados diversos subconjuntos de covariáveis potencialmente úteis, incluindo também covariáveis que sejam essenciais para o modelo, ou seja, iremos selecionar as covariáveis que irão permanecer no modelo segundo alguns critérios e métodos.

Como a base de dados disponibilizada possui um conjunto de covariáveis baixo na nossa amostra, é prudente realizar a seleção utilizando diversos critérios, como: R_p^2, $R_{a,p}^2$, C_p, AIC_p e BIC_p.

Ainda, por causa do baixo número de covariáveis o método automático de seleção de covariáveis aconselhável é o backward ou o backward stepwise. Por motivos de que utilizando este método estaremos mais próximos do modelo final escolhido, e serão realizados menos etapas para alcançar tal desfecho.

Primeiramente vamos utilizar os critérios R_p^2, $R_{a,p}^2$, C_p, AIC_p e BIC_p mencionados para cada combinação de covariáveis possíveis para o nosso modelo. Os resultados obtidos estão dispostos na Tabela 5.7.

Tabela 5.7: Seleção utilizando os critérios: R_p^2, $R_{a,p}^2$, C_p, AIC_p e BIC_p.

Variáveis	R_p^2	$R_{a,p}^2$	C_p	AIC_p	BIC_p
X_3	0.86271916	0.85722792	20.962426	171.6714	93.20787
X_1	0.83770147	0.8312095	8.974019	176.1914	97.17635
X_2	0.52756188	0.50866435	128.292235	205.0401	123.78153
X_4	0.09627993	0.06013112	266.404736	222.5527	140.65899
X_1 X_3	0.92848212	<u>0.92252230</u>	<u>1.902680</u>	<u>156.0652</u>	<u>80.45204</u>
X_3 X_4	0.88440442	0.87477146	16.017998	169.0293	90.63673
X_2 X_3	0.87329747	0.86273892	19.574857	171.5064	92.65278
X_1 X_2	0.85580314	0.84378674	25.177190	174.9985	95.53820
X_1 X_4	0.83870350	0.82526212	30.653132	178.0242	98.07376
X_2 X_4	0.52891996	0.48966329	129.857328	206.9624	124.06553
X_1 X_3 X_4	0.93064383	0.92159738	3.210421	157.2365	82.24687
X_1 X_2 X_3	0.92993961	0.92080130	3.435939	157.2365	82.43320
X_2 X_3 X_4	0.88719183	0.87247773	17.125365	170.3702	91.59958
X_1 X_2 X_4	0.85912725	0.84075254	26.112688	176.3688	96.21114
X_1 X_2 X_3 X_4	<u>0.93130091</u>	0.91881017	5.000000	158.9795	84.52812

Nota-se a partir da Tabela 5.7 que o modelo contendo as covariáveis imposto do imóvel (X_1) e área construída (X_3) possuem os melhores resultados em 4 dos 5 critérios utilizados para a seleção de modelos, pois podemos perceber que, ainda que o R_p^2 de X_1X_3 não tenha sido o maior, vemos que o $R_{a,p}^2$ é o maior, além disso, os valores de C_p, AIC_p e BIC_p são os menores para esse modelo.

O método aplicado para selecionar o modelo mais parcimonioso foi o *backward*. Neste, iniciamos com o modelo completo, ou seja, com todas as covariáveis, e a cada

passo foram retiradas aquelas que são menos essenciais para o modelo, diminuindo assim a sua complexidade.

Ao empregar tal método, os resultados obtidos foram os seguintes:

Tabela 5.8: Método Backward - 1º Passo.

Covariáveis candidatas
X_1
X_2
X_3
X_4

O algoritmo foi iniciado ajustando-se 4 modelos de regressão linear simples, um para cada variável preditora disponível, conforme Tabela 5.8.

Para cada um dos modelos ajustados, testamos $H_0 : \beta_k = 0$ contra $H_1 : \beta_k \neq 0$. O método Backward utiliza o nível de significância $\alpha_s = 0.3$, que será considerado para saída das covariáveis do modelo. Para este teste inicial a covariável X_2 já foi eliminada do modelo, uma vez que para o teste obteve-se um valor-p alto em relação ao nível de significância $\alpha_s = 0.3$, assim não rejeitamos a hipótese nula $H_0 : \beta_k = 0$, portanto, a covariável X_2 pode ser excluída do modelo.

A Tabela 5.9 apresenta o segundo passo do método Backward:

Tabela 5.9: Método Backward - 2º Passo.

Modelo	**Beta**	**Erro.Pad.**	**Beta.Pad**	**t**	**p-valor**	**Lim.If**	**Lim.Sup**
Intercepto	3.210	3.664		0.876	0.390	-4.369	10.789
X_1	2.111	0.539	0.425	3.916	0.001	0.996	3.225
X_3	14.499	2.626	0.566	5.522	0.000	9.067	19.930
X_4	-0.053	0.062	-0.052	-0.847	0.406	-0.181	0.076

Podemos perceber pela tabela 5.9 que por meio do método Backward, no segundo passo, a covariável X_4 foi eliminada do modelo, pois seu p-valor foi maior que o nível de significância $\alpha_s = 0.3$.

Tabela 5.10: Método Backward - 3º Passo.

Modelo	**Beta**	**Erro.Pad.**	**Beta.Pad**	**t**	**p-valor**	**Lim.If**	**Lim.Sup**
Intercepto	0.790	2.279		0.347	0.732	-3.914	5.495
X_1	2.297	0.489	0.463	4.698	0.000	1.288	3.306
X_3	13.933	2.524	0.544	5.519	0.000	8.723	19.143

No 3º Passo do médoto Backward obtemos as covariáveis selecionadas, que neste caso foram X_1 e X_3 definindo o modelo selecionado.

Por fim, consegue-se ajustar um modelo de regressão obtendo o modelo mais parcimonioso, ou seja, o modelo com melhores valores dos critérios propostos e com variáveis significativas, obteve-se um modelo de regressão com as covariáveis X_1 e X_3, ou seja, um modelo com as covariáveis imposto do imóvel e área construída, uma vez que rejeitamos a hipótese $H_0 : \beta_k = 0$ e temos um p-valor obtido para ambas as covariáveis menor que $\alpha_s = 0.3$.

Tabela 5.11: Covariáveis removidas pelo Método Backward.

Cov's.Removidas	**R. Quadrado**	**R.Quadrado.Ajus**	C_p	**AIC**
X_2	0.9306	0.9216	3.2104	157.2365
X_4	0.9285	0.9225	1.9026	156.0652

Assim, pela Tabela 5.11 as variáveis que não foram significativas para o modelo, consequentemente descartadas foram X_2 e X_4, que são a área do terreno e idade da residência, respectivamente

5.2.5 Analisando o Modelo Selecionado

Nesta seção iremos analisar cuidadosamente os efeitos das covariáveis(X_1 e X_3) que foram selecionadas para o modelo, sobre a variável resposta e as possíveis interações entre elas.

Dessa forma tem-se que o modelo de regressão reduzido é dado pela Equação 5.3:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon_i. \tag{5.3}$$

A partir dos valores obtidos na Tabela 5.10, tem-se as estimativas para as covariáveis X_1 e X_3 e o valor do intercepto. Logo, a reta de regressão que melhor se ajusta à amostra observada é dada pela Equação 5.4:

$$\hat{\mu_Y}(x_1, x_3) = 0.790 + 2.297x_1 + 13.933x_3. \tag{5.4}$$

Nota-se que 0 é um valor discrepante em relação às observações da amostra para as covariáveis X_1 (imposto do imóvel) e X_3 (área construída). Dessa forma, $\hat{b_0}$ não é um valor interpretável para esse estudo.

Considerando que a área construída se mantém constante, estima-se que um acréscimo de 2.297 ocorre no valor do preço médio da venda do imóvel quando aumenta-se em uma unidade o imposto do imóvel. Por fim, considerando que o imposto do imóvel se mantém constante, estima-se um aumento de 13.933 no preço médio de venda do imóvel quando aumenta-se em uma unidade de área a área do terreno.

O próximo passo foi calcular a porcentagem da variabilidade das observações do preço de venda do imóvel Y que é explicada pelo modelo de regressão contendo como

covariável apenas o imposto do imóvel(X_1), e depois contendo apenas a covariável área construída(X_3).

$$R^2_{a,1} = \left(\frac{n-1}{n-p}\right)\frac{SQRes}{SQTotal} = \left(\frac{27-1}{27-2}\right)\frac{863.8}{5322.3} = 0.17. \tag{5.5}$$

Por meio da Equação 5.5 pode-se concluir que de acordo com o coeficiente de determinação múltiplo ajustado, aproximadamente 17% da variabilidade das observações do preço de venda do imóvel Y é explicada pelo modelo de regressão considerando apenas a covariável imposto do imóvel X_1.

$$R^2_{a,3} = \left(\frac{n-1}{n-p}\right)\frac{SQRes}{SQTotal} = \left(\frac{27-1}{27-2}\right)\frac{730.7}{5322.3} = 0.14. \tag{5.6}$$

Pela Equação 5.6 pode-se concluir que de acordo com o coeficiente de determinação múltiplo ajustado, aproximadamente 14% da variabilidade das observações do preço de venda do imóvel Y é explicada pelo modelo de regressão considerando apenas a covariável área construída X_3.

Iremos calcular o coeficiente de determinação parcial para cada uma das covariáveis, que mede a redução relativa marginal da variabilidade das observações da variável resposta ao incluir tais covariáveis no modelo.

$$R^2(X_3) = \frac{SQRes(X_1) - SQRes(X_1, X_3)}{SQRes(X_1)} = \frac{863.8 - 380.6}{863.8} = 0.56. \tag{5.7}$$

Temos pela Equação 5.7 há uma redução de aproximadamente 56% na variabilidade das observações do preço do imóvel Y quando incluímos a covariável área construída no modelo que já continha a covariável imposto do imóvel.

$$R^2(X_1) = \frac{SQRes(X_3) - SQRes(X_1, X_3)}{SQRes(X_3)} = \frac{730.7 - 380.6}{730.7} = 0.48. \tag{5.8}$$

Temos pela Equação 5.8 que há uma redução de aproximadamente 48% na variabilidade das observações do preço do imóvel Y quando incluímos a covariável imposto do imóvel no modelo que já continha a covariável área construída.

Conduziremos agora um teste F para decidir se a covariável X_1 pode ser descartada do modelo de regressão dado que X_3 é mantida.

- **Hipóteses:**

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

- **Estatística de teste:** Sob H_0,

$$F = \frac{SQE(R) - SQE(C)}{gl(R) - gl(C)} \div \frac{SQE(C)}{gl(C)} \sim F_{1,n-p}$$

$$F = \frac{SQE(R) - SQE(C)}{1} \div \frac{SQE(C)}{n-p} \sim F_{1,24}$$

- **Regra de decisão:** Rejeitamos a hipótese $H_0 : \beta_1 = 0$ ao nível de significância α = 5% se, e somente se, para amostra observada temos $f \geq f_c$ em q f é estatística F observada e $f_c \in R_+$ é tal que

$$P(F \geq f_c | \beta_1 = 0) = 0.05$$

- **Conclusão:** Como f = 22.07 e $f_c = 4.26$, segue que $f > f_c$ e, portanto, para essa amostra observada, rejeitamos a hipótese $H_0 : \beta_1 = 0$ a um nível de significância de 5%. Em outras palavras, temos evidências de que a covariável X_1 pode ser mantida no modelo de regressão que já possui como covariável X_3 quando consideramos um nível de significância α= 0.05.

Conduziremos agora um teste F para decidir se a covariável X_3 pode ser descartada do modelo de regressão dado que X_1 é mantida.

- **Hipóteses:**

$$\begin{cases} H_0 : \beta_3 = 0 \\ H_1 : \beta_3 \neq 0 \end{cases}$$

- **Estatística de teste:** Sob H_0,

$$F = \frac{SQE(R) - SQE(C)}{gl(R) - gl(C)} \div \frac{SQE(C)}{gl(C)} \sim F_{1,n-p}$$

$$F = \frac{SQE(R) - SQE(C)}{1} \div \frac{SQE(C)}{n-p} \sim F_{1,24}$$

- **Regra de decisão:** Rejeitamos a hipótese $H_0 : \beta_3 = 0$ ao nível de significância α = 5% se, e somente se, para amostra observada temos $f \geq f_c$ em q f é estatística F observada e $f_c \in R_+$ é tal que

$$P(F \geq f_c | \beta_3 = 0) = 0.05$$

- **Conclusão:** Como f = 30.46 e $f_c = 4.26$, segue que $f > f_c$ e, portanto, para essa amostra observada, rejeitamos a hipótese $H_0 : \beta_3 = 0$ a um nível de significância de 5%. Em outras palavras, temos evidências de que a covariável X_3 pode ser mantida no modelo de regressão que já possui como covariável X_1 quando consideramos um nível de significância α= 0.05.

Com os resultados obtidos acima confirmamos a escolha do modelo dada pelos critérios e método utilizados, uma vez que X_1 e X_3 são essenciais para explicar a variável resposta Y (preço de venda do imóvel).

5.2.6 Análise dos Resíduos do Modelo Escolhido

Será verificado se as suposições de normalidade, homocedasticidade, linearidade e independência dos erros do modelo escolhido com as duas covariáveis continuam sendo atendidas através dos resíduos, então foram realizadas análises gráficas e aplicação de testes de hipóteses para comprovar essas suposições.

Normalidade

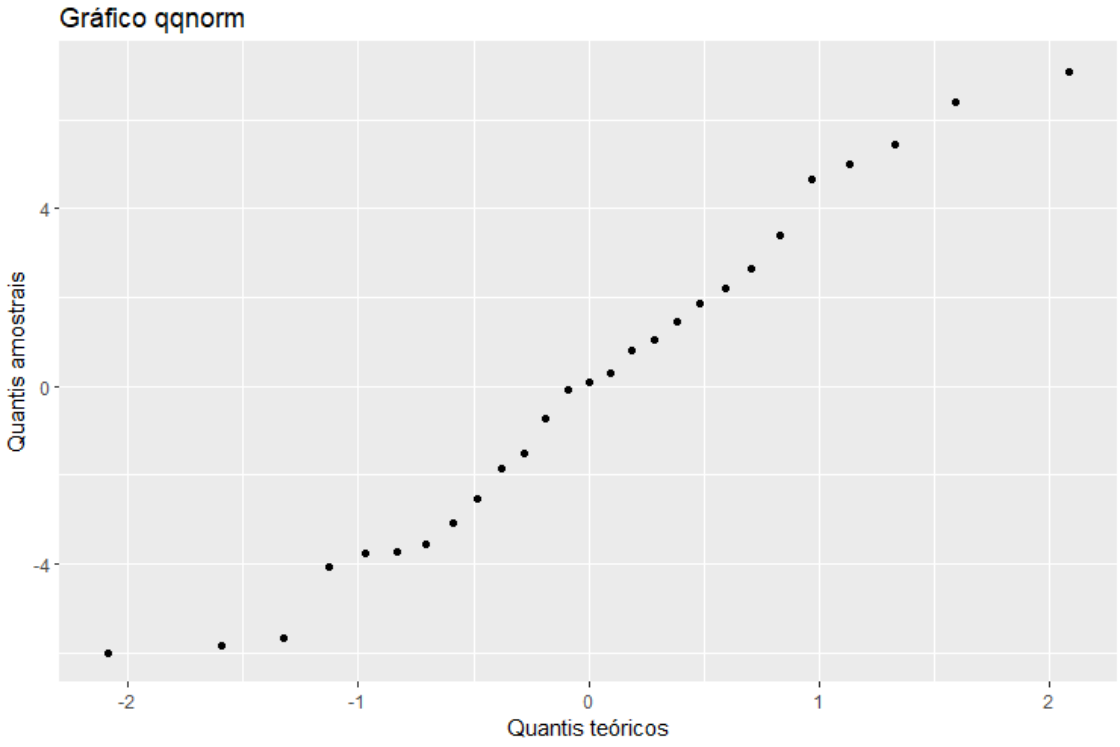

Figura 5.9: Gráfico para Verificar Normalidade.

Pode-se verificar através da Figura 5.9 que os resíduos padronizados observados são distribuídos de maneira a formar uma reta, indicando um formato simétrico da distribuição. Logo, temos um indicativo, com base na amostra observada, de que a hipótese de normalidade é satisfeita.

Tabela 5.12: Testes de Hipótese para Normalidade.

Teste	**Valor - p**	**Conclusão**
Lilliefors	0.8688	Não rejeita H_0
Shapiro - Wilk	0.4903	Não rejeita H_0
Kolmogorov - Smirnov	0.9767	Não rejeita H_0
Anderson - Darling	0.7763	Não rejeita H_0

Por meio da Tabela 5.12 não rejeitamos a hipótese nula para todos os testes de normalidade ao nível de significância de 5%. Ou seja, há evidências de que a suposição

de normalidade dos resíduos é satisfeita.

Homocedasticidade e Linearidade

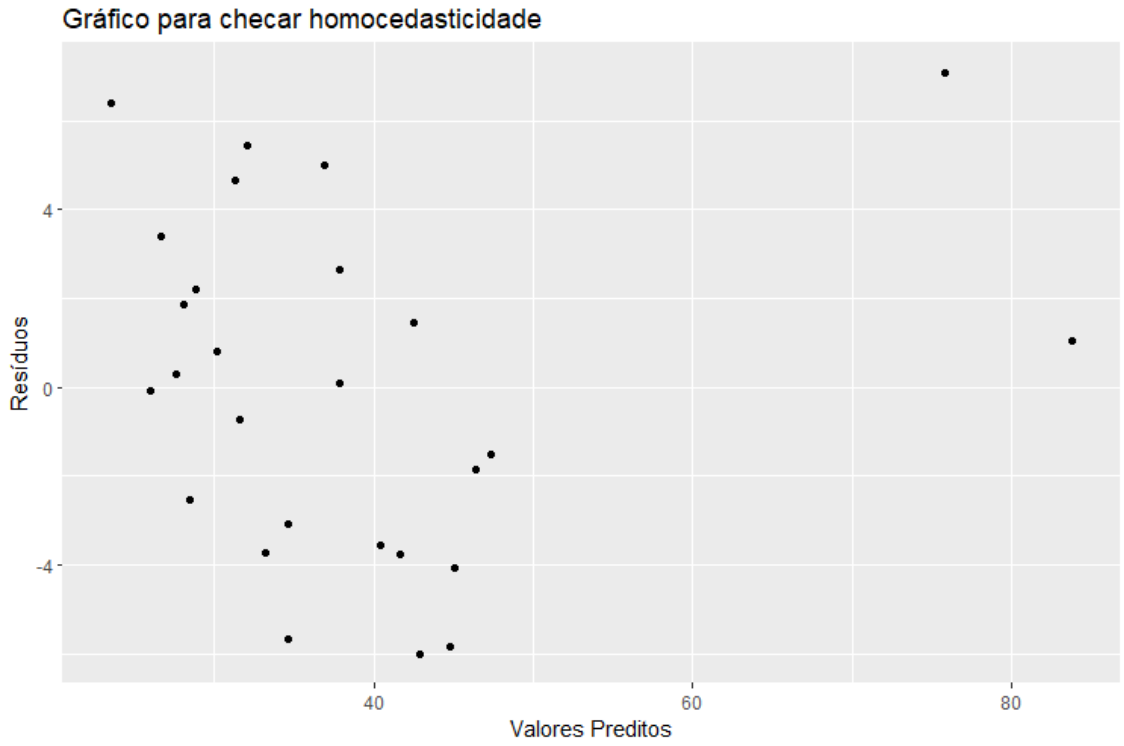

Figura 5.10: Gráfico dos resíduos vs. Valores preditos.

Através da Figura 5.10 tem-se os valores preditos *versus* os resíduos, onde é possível verificar que as observações estão distribuídas em torno do valor 0 sem apresentar oscilações significativas.

Além disso, é possível notar que os pontos não possuem qualquer tendência particular. Assim há indícios de que as hipóteses de homocedasticidade e linearidade dos resíduos não é violada.

Foi realizado ainda, o teste de Breusch - Pagan para a verificação de homocedasticidade, obtendo um valor - p = 0.5951. Ao nível de significância $\alpha = 0.05$, não rejeitamos H_0, portanto há evidências de que a suposição de homocedasticidade não foi violada.

Independência

O banco de dados não traz a informação da ordem seleção dos indivíduos. Logo, não é possível verificar se há independência, portanto será assumido que a suposição de independência dos erros está satisfeita.

Como as hipóteses sobre os erros não foram violadas temos que os resultados obtidos para o modelo selecionado são válidos.

Capítulo 6

Considerações finais

Ao longo do estudo, observamos que as suposições iniciais de significância das covariáveis imposto do imóvel (X_1), área do terreno (X_2), área construída(X_3) e idade da residência (X_4), feitas na análise descritiva para explicar o preço de venda do imóvel (Y), não foram totalmente suportadas na análise de regressão. Ou seja, alguns dos indícios que tivemos na análise descritiva não foram satisfeitos por meio dos procedimentos formais das análises realizadas.

Com base na análise descritiva dos dados, foi verificado através dos gráficos de dispersão e da matriz de correlação, que as covariáveis imposto do imóvel(X_1) e área do terreno(X_3) apresentavam uma alta correlação. Logo, havia um indicativo de que X_1 e X_3 continham muito da mesma informação, ou seja, esperávamos que houvesse multicolinearidade. Contudo, ao usarmos a medida VIF (fator de inflação da variância) obtivemos resultados contrários ao esperado, constatando a ausência de multicolinearidade.

Além disso, esperava-se também que a covariável idade da residência(X_4) fosse excluída do modelo, já que ela possui uma correlação baixa com a variável resposta, o que foi confirmado por meio do método Backward na etapa da seleção das covariáveis para o modelo. Ademais, a covariável área do terreno (X_2) foi descartada do modelo, também pelo método de seleção de variáveis Backward, o que não era previsto, uma vez que analisando o gráfico de correlação entre as variáveis, a covariável X_2 tinha uma correlação consideravelmente alta com Y.

Referências Bibliográficas

[1] NETER, J.; WASSERMAN, W. e KUTNER, M. H. (1983). Applied linear regression models. Irwin.

[2] FERREIRA, R.. Notas de aula da disciplina Análise de Regressão.

[3] Método de Seleção de Variáveis. Disponível em: `https://olsrr.rsquaredacademy.com/articles/variable_selection.html#stepwise-forward-regression`

[4] Teste Breusch-Pagan para homocedasticidade. Disponível em: `https://www.statology.org/breusch-pagan-test-r/`

Capítulo 7

Apêndice: Códigos Utilizados

colorful [frame=lines, baselinestretch=1.2, fontsize=]R

X1¡- as.numeric(imoveis$V1$)$X2 < -as.numeric(imoveis$V2) X3¡- as.numeric(imoveis$V3$)$X4 < -as.numeric(imov
Y¡- as.numeric(imoveis$V5$)$plot(imoveis$)

View(imoveis) library(ggplot2)

ggplot(imoveis, aes(y = X1)) + geom˙boxplot(fill = "skyblue") + labs(x = "X1", y = "Imposto(em 100 USD)", title = "Imposto do Imóvel") + theme˙minimal() + theme(plot.title = element˙text(hjust = 0.5))

ggplot(imoveis, aes(y = X2)) + geom˙boxplot(fill = "skyblue") + labs(x = "X2", y = "Área(em 1.000 pés quadrados)", title = "Área do Terreno")+ theme˙minimal() + theme(plot.title = element˙text(hjust = 0.5))

ggplot(imoveis, aes(y = X3)) + geom˙boxplot(fill = "skyblue") + labs(x = "X3", y = "Área(em 1.000 pés quadrados)", title = "Área construída")+ theme˙minimal() + theme(plot.title = element˙text(hjust = 0.5))

ggplot(imoveis, aes(y = X4)) + geom˙boxplot(fill = "skyblue") + labs(x = "X4", y = "Idade(em anos)", title = "Idade da residência") + theme˙minimal() + theme(plot.title = element˙text(hjust = 0.5))

ggplot(imoveis, aes(y = Y)) + geom˙boxplot(fill = "skyblue") + labs(x = "Y", y = "Preço(em 1.000 USD)", title = "Preço de venda do imóvel")+ theme˙minimal() + theme(plot.title = element˙text(hjust = 0.5))

pairs(+ X1 + X2 + X3 + X4 + Y, data = imoveis, col = c("blue"), pch = 20)

library(RColorBrewer) library(corrplot)

g1¡- pairs(imoveis, col= rainbow(3))

library(GGally)

ggpairs(imoveis, aes(alpha= 0.05))+ theme˙bw()

padronização: pad˙X1= (1/sqrt(27-1))*((X1-mean(X1))/sd(X1)) pad˙X2= (1/sqrt(27-1))*((X2-mean(X2))/sd(X2)
pad˙X3= (1/sqrt(27-1))*((X3-mean(X3))/sd(X3)) pad˙X4= (1/sqrt(27-1))*((X4-mean(X4))/sd(X4)) pad˙Y= (1/sqrt(27-1))*((Y-mean(Y))/sd(Y))

matriz com as covariaveis padronizadas: matriz˙pad= matrix(c(pad˙X1,pad˙X2,pad˙X3,pad˙X4),nrow=27,ncol=4) verifica˙xtx¡- t(matriz˙pad) verifica˙xtx vetor˙corr= (t(matriz˙pad))vetor˙corr

vif(xtx)-1 inversa˙pad=solve(verifica˙xtx) inversa˙pad

vif library(car) ajuste ¡- lm(Y X1+X2+X3+X4) lm(linear model) vif(ajuste)

análise de diagnóstico

checa as suposições res¡-residuals(ajuste) pred¡-fitted.values(ajuste)

checa homocedasticidade library(lmtest) bptest(ajuste) plot(pred,res,xlab='Valores preditos',ylab='Residuos') ggplot(data = imoveis, aes(x = pred, y = res)) + labs(x = "Valores Preditos", y = "Resíduos", title = "Gráfico para checar homocedasticidade") + geom˙point()

checa normalidade qqnorm(res,xlab='Quantis teoricos', ylab='Quantis amostrais') ggplot(imoveis, aes(sample=res)) + labs(x = "Quantis teóricos", y = "Quantis amostrais", title = "Gráfico qqnorm")+ stat˙qq() library(nortest) lillie.test(res) shapiro.test(res) xb ¡- mean(res) sx ¡- sd(res) ks.test(res, "pnorm", xb, sx,alternative='two.sided') ad.test(res)

library("olsrr") library("MASS") library("car")

stepAIC(ajuste, direction = "backward") k1 ¡- ols˙step˙all˙possible(ajuste) k1 plot(k1)

k ¡- ols˙step˙backward˙p(ajuste, details = T) k plot(k)

ajuste ¡- lm(Y X1) lm(linear model) SQReg SQReg ¡- sum((ajuste$fitted.values - mean(Y))^2)$ $SQRegSQRes <- sum((ajuste$residuals)ˆ2) SQRes SQTotal¡- sum((Y- mean(Y))ˆ2) SQTotal

qf(0.05,1,24,lower.tail = F) MODELO ESCOLHIDO

ajuste1 ¡- lm(Y X1+ X3) lm(linear model) summary(ajuste1) análise de diagnóstico

checa as suposições res1¡-residuals(ajuste1) pred1¡-fitted.values(ajuste1)

checa homocedasticidade bptest(ajuste1) plot(pred1,res1,xlab='Valores preditos',ylab='Residuos') ggplot(data = imoveis, aes(x = pred1, y = res1)) + labs(x = "Valores Preditos", y = "Resíduos", title = "Gráfico para checar homocedasticidade") + geom˙point()

checa normalidade qqnorm(res1,xlab='Quantis teoricos', ylab='Quantis amostrais') ggplot(imoveis, aes(sample=res1)) + labs(x = "Quantis teóricos", y = "Quantis amostrais", title = "Gráfico qqnorm")+ stat˙qq() library(nortest) lillie.test(res1) shapiro.test(res1) xb ¡- mean(res1) sx ¡- sd(res1) ks.test(res1, "pnorm", xb, sx,alternative='two.sided') ad.test(res1)