UNIVERSIDADE ESTADUAL DA PARAÍBA - UEPB

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

CURSO DE BACHARELADO EM ESTATÍSTICA

RELATORIO: DADOS DE MODELOS LINEARES GENERALIZADOS – RESIDÊNCIAS EM BOSTON

Lucas Manoel Batista de Albuquerque
Cleanderson Romualdo Fidelis

Modelos Lineares Generalizados

2023

1 INTRODUÇÃO

Este relatório apresenta uma análise abrangente do mercado imobiliário de Boston, focando nas casas disponíveis para venda na cidade. O objetivo é fornecer informações sobre as propriedades disponíveis, os preços médios, as tendências do mercado e outros dados relevantes para compradores em potencial, investidores e profissionais do setor imobiliário.

Além disso, faremos uso de um conjunto de dados abrangente, que contém informações relevantes sobre diversos atributos dos apartamentos em Boston, como taxa de criminalidade per capita, número de quartos, proximidade a comodidades locais e outros fatores influentes nos preços imobiliários. Ao aplicarmos os modelos lineares generalizados a esses dados, poderemos identificar os fatores mais significativos na determinação dos preços dos apartamentos em Boston.

2 METODOLOGIA

2.1 MODELO LINEAR GENERALIZADO

Um modelo linear generalizado é uma extensão do modelo de regressão linear clássico que permite lidar com uma ampla variedade de distribuições de probabilidade e relacionamentos entre variáveis dependentes e independentes. O MLG é composto por três componentes principais: função de ligação, estrutura de distribuição de probabilidade (componente aleatório) e componente sistemático.

- Distribuição de probabilidade (componente aleatório)

Sejam $Y_1, \dots, Y_n$ variáveis aleatórias independentes, cada uma com uma função densidade ou função de probabilidade na forma dada abaixo

$$f(y_i; \theta_i; \emptyset) = \exp\left[\emptyset\{\mathrm{y_i}\theta_i - b(\theta)\} + c(y_i, \emptyset)\right] \quad (1)$$

Com $b(.)$ e $c(.)$ funções conhecidas. Assim, temos que $E(Y_i) = b'(\theta) = \mu_i$ e $Var(Y_i) = \emptyset b''(\theta_i)$

- Componente sistemático

O componente linear é semelhante ao modelo de regressão linear tradicional e é usado para modelar o efeito das variáveis independentes na variável dependente. Consiste em uma combinação linear das variáveis independentes ponderadas por coeficientes. Esses coeficientes são estimados usando métodos de otimização, como a máxima verossimilhança.

As variáveis explanatórias entram em forma de uma soma linear de seus efeitos:

$$n_i = \sum_{i=1}^{d} X_i r B_r = X^T B \quad (2)$$

em que $n_i = \mathbf{X}_i^T \boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ é um vetor de parâmetros desconhecidos a serem estimados e $\boldsymbol{X}_i = (x_{i1}, \dots, x_{ip})^T$ representa o valor das variáveis explicativas.

- Função de ligação

A função de ligação é usada para relacionar a média da variável dependente às variáveis independentes. Ela impõe uma restrição para garantir que a média esteja dentro do intervalo adequado para a distribuição de probabilidade escolhida, isto é

$$n_i = g(\mu_i) \tag{3}$$

Portanto, uma decisão importante na escolha do MLG é definir os termos do trinômio: (1) distribuição da variável resposta; (2) matriz do modelo e (3) função de ligação.

Abaixo na Tabela 1 as funções de ligação canônicas e suas respectivas distribuições:

Tabela 1 – Funções de ligação canônicas

Distribuição	Função de ligação canônica
Normal	Identidade: $\rho = \mu$
Poisson	Logarítmica: $\rho = log(\mu)$
Binomial	Logística: $\rho = log\left(\frac{\pi}{1-\pi}\right)$
Gama	Recíproca: $\rho = \frac{1}{\mu}$
Normal inversa	Recíproca do quadrado: $\rho = \frac{1}{\mu^2}$

Fonte: Autoria própria

2.2 ESCOLHA DO MELHOR MODELO

- Função desvio

A qualidade do ajuste de um MLG é avaliada através da função desvio:

$$D^*(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2\{L(\boldsymbol{y}; \boldsymbol{y}) - L(\widehat{\boldsymbol{\mu}}; \boldsymbol{y})\} \tag{4}$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros) avaliado na estimativa de verossimilhança $\widehat{\boldsymbol{\beta}}$. Um valor pequeno para a função desvio indica que, para um número menor de parâmetros, obtemos um ajuste tão bom quanto o ajuste com o modelo saturado. Denotando por $\hat{\theta}_i = \theta_i(\widehat{\mu}_i)$ e $\tilde{\theta}_i(\tilde{\mu}_i)$ as estimativas de máxima verossimilhança de θ para os modelos com p parâmetros ($p < n$) e saturado ($p = n$), respectivamente, temos que a função $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ fica, alternativamente, dada por:

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} \left\{ y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) + \left(b\left(\hat{\theta}_i\right) - b\left(\tilde{\theta}_i\right)\right)\right\} \tag{5}$$

No caso gama, que é o que será utilizado, $\tilde{\theta}_i = -1/y_i$ e $\hat{\theta}_i = -1/\hat{\mu}_i$. Assim segue que o desvio pode ser expresso na forma:

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\{-log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\} \quad (6)$$

Além disso, contenta-se em testar um MLG comparando-se o valor de $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ com os percentis da distribuição $\chi^2_{n-p;\propto}$. Assim, quando:

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \leq \chi^2_{n-p;\propto},$$

ou seja, $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ é inferior ao valor crítico $\chi^2_{n-p;\propto}$ da distribuição χ^2_{n-p}, pode-se considerar que existem evidências, a um nível aproximado de $100\alpha\%$ de significância, que o modelo proposto está bem ajustado aos dados.

- Análise de resíduos

Uma boa análise de resíduos é necessária para garantir que o modelo se ajuste adequadamente aos dados. Se os resíduos exibirem algum padrão sistemático, isso indica que o modelo não está capturando completamente a estrutura dos dados. Por exemplo, se os resíduos mostrarem uma tendência crescente ou decrescente à medida que os valores preditores aumentam, isso pode indicar uma relação não linear entre as variáveis independentes e dependentes.

Algumas análises de resíduos podem ser realizadas para complementar a avaliação do modelo de regressão linear generalizada. Isso pode incluir gráficos de resíduos versus valores ajustados, gráficos de resíduos versus variáveis independentes, testes formais de normalidade dos resíduos, entre outros.

- Critério de Akaike

O MLG mais adequado será aquele com o menor critério de informação de Akaike AIC, descrito como:

$$AIC = -2log\hat{L} + 2(p) \quad (8)$$

onde $log(\hat{L})$ é o logaritmo da função de máxima verossimilhança. p representa a quantidade de parâmetros do modelo, com a intenção de beneficiar modelos mais parcimoniosos, ou seja, com menos parâmetros.

3 SOBRE OS DADOS

Cada registro no banco de dados descreve uma residência/apartamento num subúrbio, bairro da cidade de Boston. Os dados foram extraídos da Boston Standard Metropolitan Statistical Area (SMSA) em 1970. As variáveis são definidas da seguinte forma:

- CRIM: Taxa de criminalidade per capita
- RM: número médio de quartos por habitação

- DIS: distâncias ponderadas para cinco centros de emprego de Boston
- LSTAT: percentual de pessoas pobres no bairro
- MEDV: Valor do apartamento a cada US$ 1.000,00 (Variável resposta)

Na Tabela 1 vemos uma amostra das residências da cidade de Boston, onde as colunas dessa tabela estão de acordo com a descrição feita anteriormente:

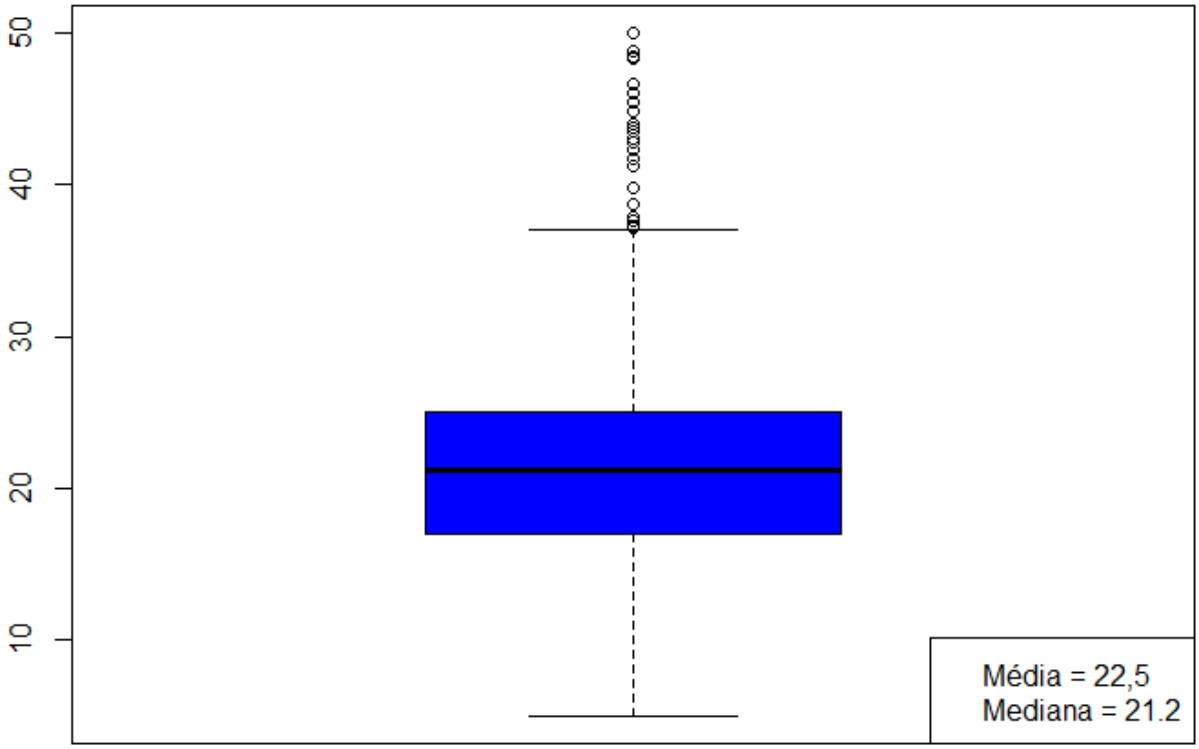
Tabela 2 – Amostra dos dados sobre as residências/apartamentos dos bairros e subúrbios da cidade de Boston

CRIM	RM	DIS	LSTAT	MEDV
0.00632	6.575	4.0900	4.98	24.0
0.02731	6.421	4.9671	9.14	21.6
0.02729	7.185	4.9671	4.03	34.7
0.03237	6.998	6.0622	2.94	33.4
0.06905	7.147	6.0622	5.33	36.2
0.02985	6.430	6.0622	5.21	28.7

Fonte: Autoria própria

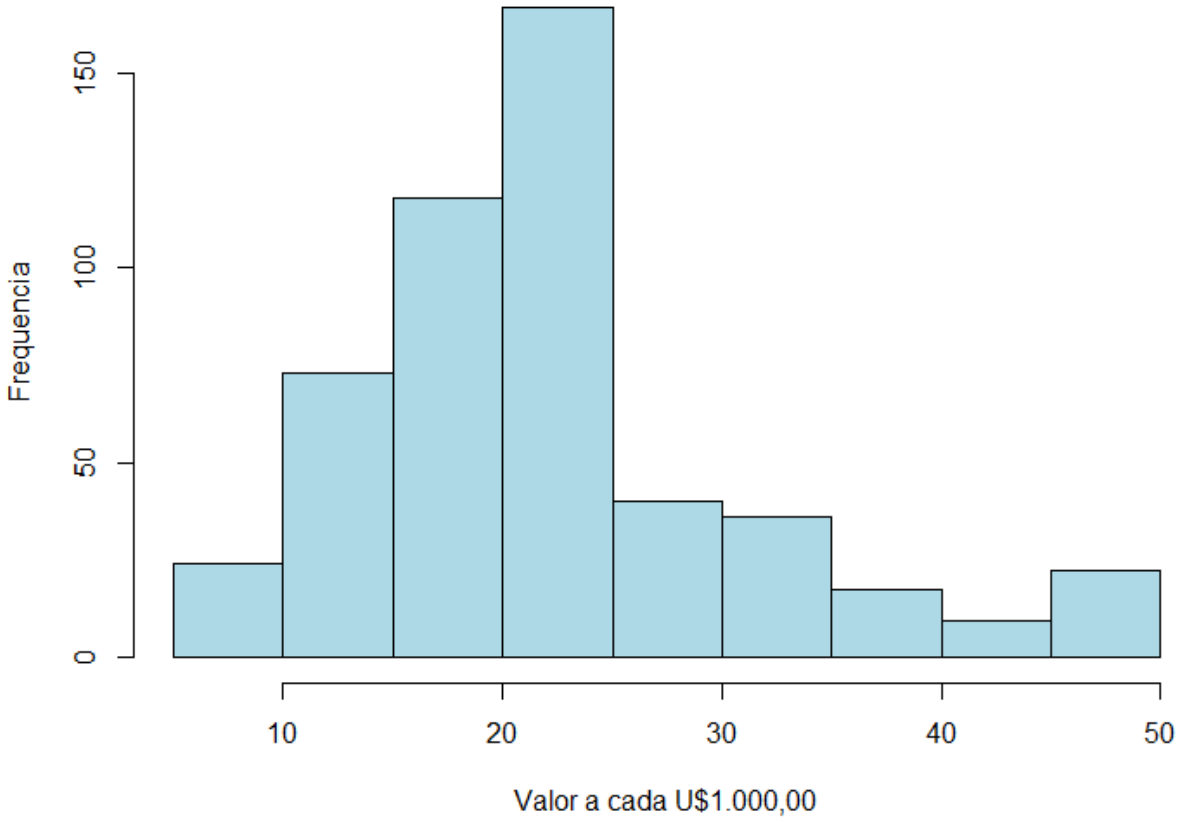
O valor médio das residências/apartamentos, indicado pela variável MEDV, é de 22,5 mil dólares e sua mediana é de 21,2 mil dólares. É possível observar a partir da Figura 1 que há alguns valores que podem ser considerados outliers de acordo com o box-plot, todos eles quando o preço está acima de 36 mil dólares. Pela Figura 2 vemos um histograma assimétrico, com muitos valores acumulados em torno da média. Esse comportamento sugere que os dados sigam uma distribuição gama, que é a que será usada para realizar o nosso modelo linear generalizado com função de ligação reciproca.

Figura 1 - Box-plot do valor das casas/apartamentos por U$1.000,00



Fonte: Autoria própria

Figura 2 - Histograma do valor das casas/apartamentos a cada U$1.000,00



Fonte: Autoria própria.

Podemos observar também algumas estatísticas gerais das variáveis explicativas a partir da Tabela 2, como a média, mediana e valores mínimos e máximos:

Tabela 3 – Estatísticas gerais das variáveis explicativas

--	CRIM	RM	DIS	LSTAT
Mínimo	0,00632	3,56	1,13	1,73
Média	3,61	6,285	3,795	12,65
Mediana	0,256	6,208	3,207	11,36
Máximo	88,97	8,78	12,127	37,97

Fonte: Autoria própria

4 OBJETIVOS

O objetivo deste relatório é observar qual seria a melhor opção para compradores em potencial, investidores e profissionais do setor imobiliário a tomar decisões informadas, com base em dados confiáveis da época.

Além disso, temos como objetivo:

- Realizar um modelo de regressão linear generalizado com a distribuição de probabilidade gama e função de ligação recíproca, onde a variável resposta é o valor do apartamento/residência a cada U$D 1.000 (MEDV)
- Observar quais variáveis podem influenciar a variável resposta MEDV

5 RESULTADOS

5.1 ESTIMAÇÃO DO MODELO DE REGRESSÃO

Para dar início a estimação do modelo de regressão Gama com função de ligação recíproca, utilizando o software Rstudio, foram consideradas as variáveis independentes “CRIM”, “RM”, “DIS” e “LSTAT” com a variável dependente “MEDV”. A Tabela 3 mostra as estimativas dos parâmetros do modelo ajustado junto com seus respectivos p-valores para descrever o preço dos apartamentos da cidade de Boston em relação a taxa de criminalidade per-capita, a quantidade média de quartos, a distância média para cinco centros de emprego em Boston e a porcentagem de pessoas pobres na região.

Tabela 4 – Coeficientes estimados e p-valores correspondentes para um MLG com distribuição Gama e função de ligação recíproca

Variável	Coeficiente (Estimativa)	p-valor
Intercepto	0,049	<0,001
LSTAT	0,0018	<0,001
CRIM	0,0008	<0,001
RM	-0,0048	<0,001
DIS	0,00117	<0,001

*<: menor que

Dado que a função de ligação usada foi a função recíproca, o nosso modelo se apresenta como:

$$\mu_i = \frac{1}{0{,}049 + 0{,}0018 * LSTAT + 0{,}0008 * CRIM - 0{,}0048 * RM + 0{,}00117 * DIS} \quad (9)$$

Vemos que todas as variáveis foram significativas a um nível de 5% de confiança. Além disso, os seus valores fazem sentido para a interpretabilidade do modelo, já que, à medida que se aumenta os valores de LSTAT, CRIM e DIS, há uma diminuição no preço do apartamento. Já para a variável RM (número de quartos), à medida que o número de quartos aumenta, também se aumenta o preço previsto dos apartamentos.

Para esse modelo ajustado, temos um valor de AIC de 2900 junto com um valor de 20,5 para a função desvio com 501 graus de liberdade. Comparando o valor do desvio com o valor tabelado da distribuição qui-quadrado com 501 graus de liberdade ao nível de 5% de significância, vemos que podemos considerar esse um modelo adequado, pois o valor tabelado da qui-quadrado (554,1) é maior do que o valor da função desvio (20,5).

A partir das Figuras 3 e 4 podemos observar como os resíduos do modelo gama com função de ligação recíproca se comportam, e a partir disso verificar quais observações o nosso modelo não consegue generalizar:

Figura 3 - Gráfico de resíduos para cada valor predito

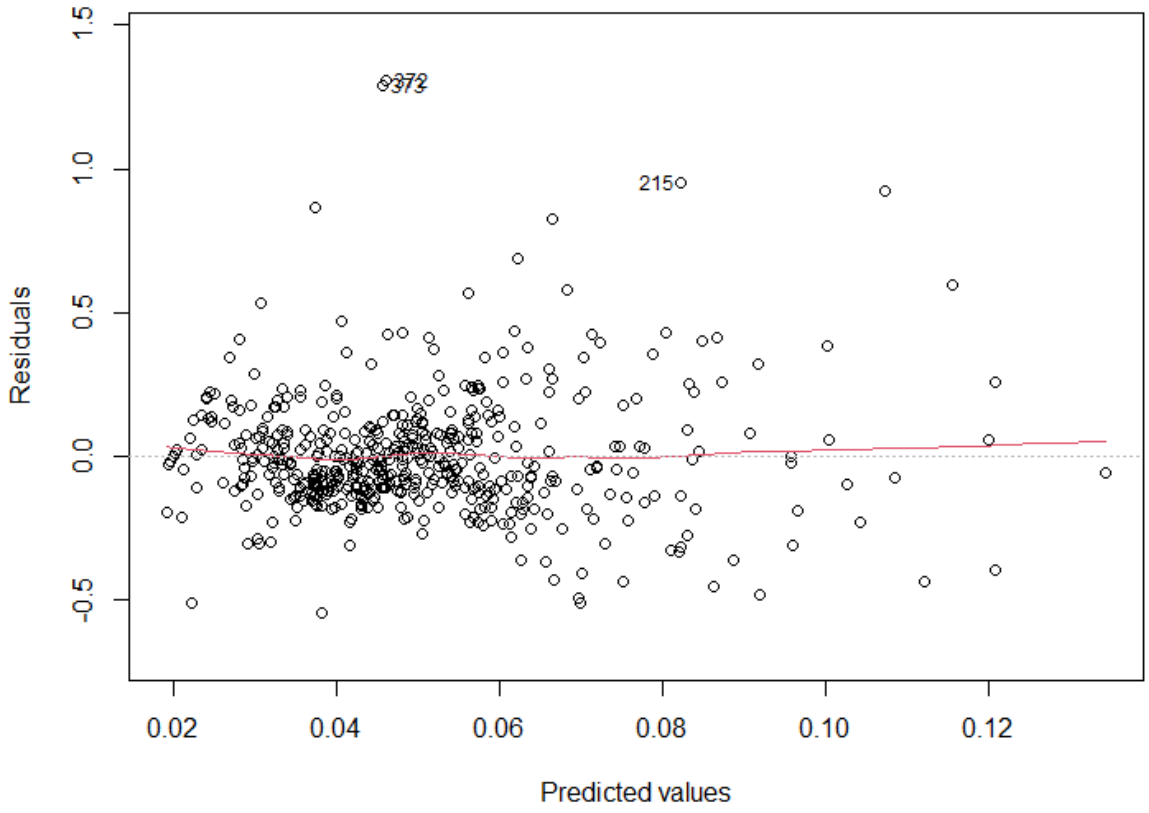
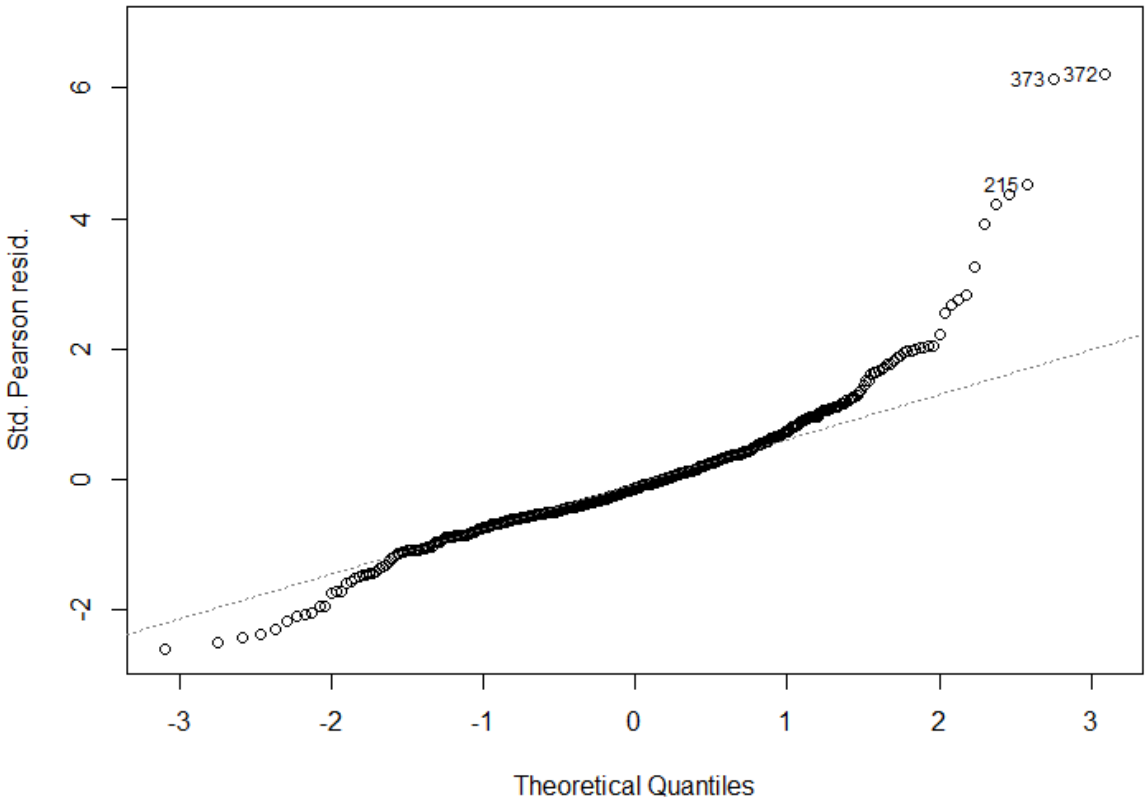


Figura 4 - Gráfico de resíduos comparados aos quantis de uma distribuição normal



Vemos que, a partir dos gráficos inclusos nas Figuras 3 e 4, os resíduos parecem se comportar bem. Ainda assim, há observações que não puderam ser generalizadas pelo nosso modelo. Tais observações estão destacadas nas Figuras e presentes na Tabela 4 para que possamos investigar as causas de erros tão discrepantes:

Tabela 4 – Valores reais e preditos do modelo Gama com função de ligação recíproca e suas respectivas variáveis explicativas

Observação	CRIM	RM	DIS	LSTAT	MEDV	PREDITO
215	0,28	5,41	3,58	29,55	23,7	12,14
365	3,47	8,78	1,90	5,29	21,9	44,84
369	4,89	4,97	1,33	3,26	50	26,76
372	9,23	6,21	1,16	9,53	50	21,67
373	8,26	5,87	1,12	8,88	50	21,84

Fonte: Autoria própria

Para a observação 215 vemos que o modelo previu um valor abaixo do real por causa da covariável LSTAT que nos indica uma porcentagem de pessoas pobres na região de 29,55%, valor muito acima da média de 12,65%. O mesmo se pode afirmar para as observações 369, 372 e 373, mas levando em consideração a variável CRIM, que indica a taxa de criminalidade. Para a observação 365 o modelo previu um valor muito acima do real por conta da variável RM (número médio de quartos do apartamento) ser acima da média, ou seja, apesar de ter um número alto de quartos o valor real do apartamento está relativamente baixo.

De modo geral, o modelo se adequa bem aos dados, seus coeficientes estão de acordo com cada covariável e as exceções, ou seja, observações que não foram captadas de forma ótima pelo modelo devem ser investigadas mais a fundo.

6 REFERÊNCIAS

The Boston Housing Dataset. Disponível em: https://www.kaggle.com/code/prasadperera/the-boston-housing-dataset/notebook .