Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

REGRESSÃO QUANTÍLICA APLICADA AO POTENCIAL DE MERCADO

Leonidas Fioranelli Braga

Uberlândia-MG
2019

Leonidas Fioranelli Braga

REGRESSÃO QUANTÍLICA APLICADA AO POTENCIAL DE MERCADO

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Profa. Dra. Maria Imaculada de Sousa Silva

Uberlândia-MG

2019

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Profa. Dra. Maria Imaculada de Sousa Silva

_____

Prof. Dr. José Waldemar da Silva

_____

Prof. Dr. Janser Moura Pereira

Uberlândia-MG

2019

Agradecimentos

Agradeço à toda minha família, em especial à minha esposa Camila, por me apoiar nesta jornada e ser compreensiva, e às minhas filhas Diana e Luana.

Aos meus colegas de faculdade que participaram do processo de formação.

Aos professores, que contribuíram e estiveram dispostos a ajudar com o meu aprendizado, em particular, à minha orientadora Profa. Dra. Maria Imaculada, que se disponibilizou para ajudar no desenvolvimento deste trabalho, e aos membros da banca Prof. Dr. Janser e Prof. Dr. José Waldemar, que se puseram a disposição para a avaliação do estudo.

À Universidade que disponibilizou as ferramentas necessárias para evoluções dos estudos.

Enfim, a todos que participaram direta ou indiretamente dessa etapa decisiva de minha vida.

Resumo

Um problema atual e recorrente para as empresas é a necessidade de gerar negócios sustentáveis no menor tempo possível, com o menor investimento. Isso sem dúvida demanda atenção e esforço para avaliar o potencial de consumo dos clientes, de forma a realizar ações de marketing e alavancar vendas com maior acurácia. A Regressão Quantílica é uma técnica que vem ganhando espaço na análise de dados econômicos e de mercado, visto que, cada vez mais, há o interesse em entender o comportamento em algum quantil e não apenas na média do conjunto de dados. Sendo assim, a Regressão Quantílica é uma técnica adequada para a estimação do potencial de compra do mercado consumidor, pois o desejo é predizer o quanto um cliente tem de recursos para adquirir os produtos, e não apenas qual seria o seu consumo médio. Neste trabalho, utilizou-se a Regressão Quantílica para estimar o potencial de compra dos clientes de uma grande empresa de tecnologia, modelando o faturamento real da empresa em função de variáveis econômicas relacionadas aos clientes. Analisou-se os resultados do modelo para o quantil 0,80, concluindo-se que o potencial de mercado para esse quantil representa uma possibilidade real dentro da empresa, podendo esses resultados serem usados para a construção de estratégias que possam atrair novos clientes ou estimulá-los a chegarem a esse potencial.

Palavras-chave: Marketing, Planejamento, Potencial de Mercado, Regressão Quantílica.

Abstract

A current and recurring problem for businesses is the need to generate sustainable business in the shortest time and with the least investment. This undoubtedly requeres attention and effort to assess customers' consumption potential, in order to conduct marketing actions and leverage sales more accurately. Quantile Regression is a technique that has been gaining space in the analysis of economic and market data, whereas, increasingly, there is interest in understanding the behavior in some quantile and not only in the dataset average. Thus, Quantile Regression is a suitable technique for estimating the potential of purchasing from the consumer market, since the desire is to predict how much a customer has the resources to acquire the products, and not only what their average consumption would be. In this work, Quantile Regression was used to estimate the customers' purchasing potential of a large technology company, modeling the actual company revenue in function of economic variables related to customers. The results of the model for the 0.80 quantile were analyzed, concluding that the market potential for this quantile represents a real possibility within the company, and these results can be used for the construction of strategies that can attract new clients or stimulate them to reach that potential.

Keywords: Quantile Regression, Market Potential, Marketing Planning.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

A incerteza no mundo dos negócios e a escassez de recursos e tempo, levam as empresas a buscar otimizações para direcionar os esforços em ações que tenham maior retorno financeiro em menor tempo. Uma maneira prática de encontrar respostas é identificando os consumidores que têm maior potencial de compra, ou seja, quais consumidores poderão gerar mais receita, com isso conseguindo um retorno financeiro com tomadas de decisões mais precisas. Todavia, os consumidores que possuem menor potencial de compra não podem ser ignorados, por representarem uma parcela importante do mercado, sendo assim, para estes as empresas têm de elaborar estratégias para atentê-los com menores custos e de forma competitiva no mercado.

Conforme Kotler [17], as empresas têm maiores chances de se saírem bem quando escolhem seus mercados-alvo com cuidado e preparam programas de marketing customizados, ou seja, é importante identificar as oportunidades e preparar as estratégias de marketing para cada perfil de cliente.

Quando se estuda variáveis internas das empresas em conjunto com variáveis externas (o mercado) é possível construir modelos de previsões que auxiliam nas tomadas de decisões e na elaboração de planos estratégicos mais assertivos. Com isso, entender o quanto os clientes têm de potencial de compra propicia uma informação relevante para os estudos de mercado e a preparação das estratégias de marketing.

As análises estatísticas contribuem diretamente com o mapeamento e entendimento do mercado consumidor. Os modelos de regressão são a opção de análise estatística tradicional quando se deseja investigar e modelar a relação entre várias variáveis com uma variável dependente, geralmente com o objetivo de realizar previsões [21] . Utilizando a minimização da soma dos quadrados dos erros, obtém-se a resposta média da variável dependente ou variável resposta, em função de um conjunto de valores particulares das variáveis independentes relacionadas pelo modelo. O método dos mínimos quadrados ordinários (MQO) é o método de estimação dos parâmetros mais comumente utilizado, pelo fato de ter facilidade de implementação, e também porque nas condições exigidas, como normalidade dos erros e homogeneidade das variâncias, o estimador obtido por esse método possui boas propriedades, tais como: os estimadores de MQO para os parâmetros do modelo são não viesados e apresenta a mínima variância. No entanto, existem casos em que normalidade e homogeneidade de variâncias não são atingidos, impossibilitando o uso desta técnica.

De acordo com Hao & Naiman [10], quando a média e a mediana de uma distribuição não coincidem, a mediana pode ser mais apropriada para capturar a tendência central da

distribuição. Assim, é mais efetivo um modelo de regressão para a mediana, que estima o efeito da covariável sobre a mediana condicional, nos casos em que a variável é assimétrica. O método de estimação neste caso, é baseado na minimização dos erros absolutos.

Observa-se que, a Regressão Clássica produz estimadores pouco precisos na presença de valores extremos (*outliers*), uma vez que ela estima para a média condicional, pelos mínimos quadrados ordinários (MQO). Para atenuar esta sensibilidade, Koenker & Bassett [13] propõem a aplicação da Regressão Quantílica, que realiza uma estimativa para cada quantil condicional da variável resposta. Portanto, com a Regressão Quantílica obtém-se uma regressão para cada quantil.

Em casos de assimetria, o modelo de regressão quantílica (MRQ), proposto por Koenker & Basset[13], estima o potencial efeito diferencial de uma covariável sobre os vários quantis da distribuição condicional da variável resposta. O modelo proposto, é uma extensão do modelo linear para estimar a taxa de mudança nas várias partes da distribuição, sendo a aplicação inicial nos exemplos de econometria. Desde então, diversas áreas de aplicação apresentaram este modelo como solução para os problemas envolvendo distribuição assimétrica. Estudos semelhantes foram realizados por Alves Filho [2], Diniz [8], Maciel *et al.* [19], Puiatti [23] e Silva [29].

Conforme abordado por Santos [27] o uso dos MRQ tem sido amplamente reportados na literatura em várias pesquisas para explicar a renda em função de outras variáveis explicativas. Em [27] tem-se um estudo sobre os modelos de Regressão Quantílica, com destaque para a estimação, inferência sobre os parâmetros, testes de adequacidade do ajuste, e uma aplicação à dados da variável renda no Brasil, e sua relação com outras variáveis sócio-econômicas.

Ainda abordando a variável renda, tem-se o estudo do impacto da criação de um terceiro filho na renda dos pais, sendo o impacto em questão estimado em vários pontos da distribuição condicional da variável resposta (rendimento da família) por meio de uma variante da regressão quantílica denominada efeito quantílico de um tratamento [29].

Os dados ecológicos frequentemente têm variação desigual e distribuição assimétrica. Assim, os modelos de Regressão Quantílica apresentam muitas possibilidades para análises estatísticas e interpretação dos resultados desses dados [5]. Cade & Noon [5] apresentam um estudo relacionando diversos trabalhos que utilizam o MRQ com dados ecológicos, discutindo as vantagens do método para esse tipo de dados. O modelo estima as múltiplas taxas de mudanças para a resposta mínima e máxima, por exemplo, fornencendo um quadro mais completo da relação entre as variáveis, em comparação com os métodos de regressão pela média.

Utilizando aplicações a dados ecológicos, Puiatti [23] apresentou um ajuste de MRQ a dados de curva de crescimento, considerando um modelo não linear, apropriado por apresentar parâmetros com interpretação biológica prática. A Regressão Quantílica neste caso, de acordo com o autor, além de contornar o problema da assimetria, permite estimativas em diferentes quantis, gerando resultados mais completos e robustos.

Com o objetivo de analisar teórica e empiricamente a suposta relação positiva existente entre desenvolvimento financeiro e crescimento econômico, Silva & Porto Júnior [30] aplicaram

a técnica de Regressão Quantílica para analisar esses aspectos para dados de 77 países, o que permitiu um mapeamento mais completo do impacto gerado pelas medidas de desenvolvimento financeiro na distribuição condicional da variável resposta.

Dito isso, tem-se que a Regressão Quantílica é aplicada quando a estimativa de interesse não é a média, ou seja, o pesquisador está estudando o comportamento das variáveis nos extremos dos dados. Como exemplo, a projeção do potencial de compra do mercado consumidor, em que a intenção não é entender o consumo médio dos potenciais compradores, mas sim o quanto um cliente com determinadas características pode gastar com os produtos de uma empresa. Com isso, a Regressão Quantílica se mostra apropriada para a estimação do potencial de compra do mercado consumidor, pois há o interesse em otimizar a variável resposta, no lugar de convergir para um valor "médio", como ocorre na regressão linear pelos Mínimos Quadrados Ordinários (MQO).

Desta forma, a Regressão Quantílica, para a análise do perfil dos consumidores de uma empresa, se apresenta como uma técnica de grande utilidade para detectar quais fatores podem influenciar no potencial de compra desses consumidores e proporcionar informações para a tomada de decisões estratégicas.

2. Fundamentação Teórica

Ao estudar vários fenômenos tem-se o anseio de entender a influência que uma ou mais variáveis exercem sobre outra, como, por exemplo, o quanto a renda, a profissão e a quantidade de filhos influenciam no volume de compras de um determinado produto.

Conforme Hoffmann [11] tais relações funcionais podem ser representadas por:

$$Y = f(X_1, X_2, ..., X_k)$$

em que, Y é a variável resposta (ou dependente) e $X_i (i = 1, 2, ..., k)$ são as variáveis preditoras (ou explicativas, independentes, explanatórias).

No entanto, a relação entre as variáveis não é perfeita, com isso há uma diferença entre o valor observado de y e o valor estimado $\widehat{y}$ que é o erro de previsão ε. Segundo Montgomery [21] é conveniente pensar em ε como um erro estatístico, ou seja, ele é uma variável aleatória que explica a falha do modelo em ajustar os dados com exatidão. O erro pode ser composto pelos efeitos de outras variáveis, erros de medição, impossibilidade de mensuração e assim por diante. Assim, um modelo mais plausível é:

$$Y = f(X_1, X_2, ..., X_k) + \varepsilon$$

A análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis, sendo utilizada para vários propósitos, incluindo os seguintes: descrição dos dados, estimação de parâmetros, previsão, estimativa e controle.

2.1 Regressão Linear

A análise de regressão é a técnica estatística utilizada para investigar a relação entre variáveis. Com numerosas aplicações nas mais diversas áreas das ciências, os modelos de regressão são provavelmente a técnica estatística mais utilizada nas pesquisas em geral [21].

O modelo de regressão é projetado para situações em que se acredita que uma variável dependente esteja relacionada com uma ou mais outras medidas feitas geralmente no mesmo objeto [28]. De acordo com Downing & Clack [9], em muitas situações, a variável dependente, que se tem interesse, pode ser afetada por mais de uma variável independente, aplicando-se para tais casos a Regressão Linear Múltipla.

O objetivo da análise é usar os dados observados das variáveis para estimar a forma dessa relação, sendo frequentemente apropriado o seguinte modelo linear para descrever tal relação de uma variável dependente y com k variáveis explicativas X.

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_k X_{kj} + \varepsilon_j$$

em que :
Y_j = variável resposta (dependente)
$X_{ij}, i = 1, ..., k$ = variáveis explicativas (independentes)
β_i = parâmetros associados às variáveis explicativas
ε_j = erro aleatório

A Regressão Linear é ajustada pelo Método dos Mínimos Quandrado (ou Mínimos Quadrados Ordinários (MQO)), que consiste em adotar como estimativas dos parâmetros os valores que minimizam a soma dos quadrados dos desvios. Este método de estimação dos parâmetros é o modelo mais difundido, isso se dá, como cita Santos [27], pelo fato da facilidade computacional para implementar tal cálculo e, além disso, em caso de distribuição normal dos erros do modelo, o estimador obtido por este método possui boas propriedades.

Utilizando a notação matricial o modelo de regressão tem a seguinte equação:

$$y = X\beta + \varepsilon$$

onde:

$$\text{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad e \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Como apresenta Hoffmann [11], ao estabelecer o modelo de Regressão Linear, pressupomos que:

- a variável resposta (Y_j) é função linear das k variáveis explicativas (X_{kj} , j = 1,..., n);
- os valores das variáveis explicativas são fixos;
- $E(\varepsilon_j) = 0$, ou seja, $E(\varepsilon) = 0$, onde 0 representa um vetor de zeros;
- os erros são homocedásticos, isto é, $E(\varepsilon_j^2) = \sigma^2$;
- os erros são não-correlacionados entre si, isto é, $E(\varepsilon_j \varepsilon_h) = 0$ para $j \neq h$;
- os erros têm distribuição normal.

A resposta média estimada, ou valor esperado de Y para um conjunto particular de valores das variáveis preditoras X_{kj} é dada por:

$$E[Y|X] = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_k X_{kj} \ .$$

2.2 REGRESSÃO QUANTÍLICA

Conforme argumentam Koenker & Basset [13], os Mínimos Quadrados Ordinários (MQO) são extremamente sensíveis a valores extremos (*outliers*), como é o caso de distribuições não-gaussianas, produzindo estimadores com pouca precisão, visto que para que a regressão por MQO seja não viesada e necessário que os erros tenham distribuição normal e sejam homocedásticos. Koenker & Basset [13] introduzem a Regressão Quantílica procurando estender a ideia à estimação de funções quantílicas condicionais, modelos nos quais os quantis da distribuição condicional da variável resposta são expressos como funções de covariáveis observadas.

A Regressão Quantílica é um método de estimação motivado pelo interesse em estudar o comportamento de indivíduos "não-médios". Basicamente, a Regressão Quantílica estima várias retas para diferentes quantis associados, em vez de estimar apenas a esperança de Y dado X, como é feito numa Regressão Linear pelo MQO (Mínimos Quadrados Ordinários).

Como abordado por Silva [29], a Regressão Quantílica deve ser vista como uma generalização do modelo de Regressão de Mínimos Desvios Absolutos (MDA), L_1 ou Regressão Mediana para o caso do modelo de Regressão Linear, permitindo estimar não só a mediana, mas também outros quantis da distribuição de Y. Assim, enquanto um modelo especificado com MQO tem a forma $Y = X\beta + \varepsilon$, donde a condição $E[\varepsilon] = 0$ implica que $E[Y|X] = X\beta$, um modelo de Regressão Quantílica irá verificar o efeito que os preditores X terão sobre os quantis de Y, tal que o τ-ésimo quantil da variável Y é definido como

$$Q_\tau = inf\{y : F(y) \geq \tau\}$$

onde $F(y) = P(Y \leq y)$ é a função de distribuição acumulada de Y. Intuitivamente, o τ-ésimo quantil de Y é o valor-limite Q_τ em que há exatamente τ por cento de chance de os valores de Y serem menores que Q_τ . É fácil observar que $0 \leq \tau \leq 1$ (pelo axioma da probabilidade) e que Q_τ é uma função não-decrescente de τ.

Koenker & Hallock [14] apresentam uma definição para quantil, dizendo que um aluno marca no τ-ésimo quantil de um exame padronizado, se ele apresentar um desempenho melhor do que a proporção τ do grupo de referência dos alunos e pior do que a proporção $(1 - \tau)$. Assim, metade dos alunos tem um desempenho melhor que a mediana e metade tem um desempenho pior. Similarmente, os quartis dividem a população em quatro segmentos com proporções iguais da população de referência em cada segmento. Os quintis dividem a população em cinco partes e os decis em dez partes. Os quantis, ou percentis, ou ocasionalmente fractis, referem-se ao caso geral.

Segundo Maciel *et al.* [19], uma importante propriedade de função quantil concerne ao fato que para $-\infty \leq y \leq +\infty$ e $0 \leq \tau \leq 1$, $F(y) \geq \tau$ se e somente se $Q_\tau \leq y$. Assim, tem-se Y identicamente distribuída a Q_τ.

Dessa forma, numa Regressão Quantílica o modelo de regressão será dado por

$$Q_\tau(y|x) = x^T\beta(\tau) = \beta_0(\tau) + x_1\beta_1(\tau) + x_2\beta_2(\tau) + ... + x_k\beta_k(\tau)$$

onde, $\beta(\tau)$ é o efeito marginal das variáveis explicativas X no τ-ésimo quantil de Y, efeito este que pode ser variante a depender do quantil escolhido. Essa abordagem se mostra bastante pertinente para variáveis dependentes cuja distribuição apresenta assimetria, caudas pesadas ou heteroscedasticidade.

Conforme aponta Silva [29] a Regressão Quantílica tem as seguintes vantagens:

- A técnica de Regressão Quantílica permite caracterizar toda distribuição condicional de uma variável resposta a partir de um conjunto de regressores.
- Regressão Quantílica pode ser usada quando a distribuição não é gaussiana.
- Regressão Quantílica é robusta a *outliers*.
- Por utilizar a distribuição condicional da variável resposta, podem se estimar os intervalos de confiança dos parâmetros diretamente dos quantis condicionais desejados.

2.2.1 Estimação dos Parâmetros

Segundo Bloomfield & Steiger [4] a estimação pela Regressão L_1, ou MDA, é recomendada para os casos em que a distribuição dos erros tem caudas pesadas, sendo uma alternativa robusta para a regressão de mínimos quadrados, ou MQO. Isso decorre do fato de os Mínimos Desvios Absolutos possuírem menor sensibilidade às observações cujos resíduos absolutos sejam altos quando comparada aos Mínimos Quadrados Ordinários.

Rodrigues [26] em seu trabalho Diagnóstico em Regressão L_1, descreve muito bem sobre o método de estimação dos parâmetros pelos Mínimos Desvios Absolutos. Koenker [12] mostra que o estimador $\hat{\beta}(\tau)$ pode ser obtido como solução de um problema de programação linear.

Seja Y_j, $(j = 1, 2, ..., n)$ uma amostra aleatória da variável aleatória Y, com distribuição simétrica em torno de $\beta(\tau)$. Para $0 \leq \tau \leq 1$, o τ-ésimo quantil amostral pode ser definido como a solução do problema de minimização dado por:

$$\min_{\hat{\beta}(\tau)\in\mathbb{R}} \left[\sum_{j\in\{j:Y_j\geq\hat{\beta}(\tau)\}} \tau \left| Y_j - \hat{\beta}(\tau) \right| + \sum_{j\in\{j:Y_j\leq\hat{\beta}(\tau)\}} (1-\tau) \left| Y_j - \hat{\beta}(\tau) \right| \right], \tau \in (0,1)$$

Seja $Y = (Y_1, Y_2, ..., Y_n)^T$ o vetor de variáveis resposta do modelo linear, em que Y_j é dado por: $Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_k X_{kj} + \varepsilon_j, j = 1, ..., n$. O estimador $\beta(\tau)$ no modelo de Regressão Quantílica de ordem τ é qualquer solução em $\hat{\beta}(\tau)$ do problema de minimização:

$$\min_{\hat{\beta}(\tau)\in\mathbb{R}^n}\left[\sum_{j\in\{j:Y_j\geq X_j\hat{\beta}(\tau)\}}\tau\left|Y_j-X_j\hat{\beta}(\tau)\right|+\sum_{j\in\{j:Y_j\leq X_j\hat{\beta}(\tau)\}}(1-\tau)\left|Y_j-X_j\hat{\beta}(\tau)\right|\right],\tau\in(0,1)$$

em que, X_j é a j-ésima linha da matriz X.

Conforme Rodrigues [26], o parâmetro τ pondera a Regressão Quantílica, ou seja, é tal que pelo menos de $100\tau\%$ dos valores dos Y_j encontram-se acima do hiperplano de Regressão Quantílica e pelo menos $100(1-\tau)\%$ abaixo.

O método de estimação via mínima soma dos erros absolutos é equivalente ao seguinte problema de programação linear:

$$\min_{\hat{\beta}}\left[1^T\epsilon^+ + 1^T\epsilon^-\right],$$

onde:

1^T = vetor linha de 1s, ou seja, $1^T = (1,\ 1,\ ...,\ 1)$

$\epsilon^+ = (\epsilon_1^+, \epsilon_2^+, ..., \epsilon_n^+)^T$

$\epsilon^- = (\epsilon_1^-, \epsilon_2^-, ..., \epsilon_n^-)^T$,

sendo:

$$\epsilon_j^+ = \begin{cases} \epsilon_j, \text{ se } \epsilon_j > 0 \\ 0, \text{ se } \epsilon_j \leq 0 \end{cases}, j = 1, 2, ..., n$$

$$\epsilon_j^- = \begin{cases} |\epsilon_j|, \text{ se } \epsilon_j < 0 \\ 0, \text{ se } \epsilon_j \geq 0 \end{cases}, j = 1, 2, ..., n;$$

sujeito à condição:

$$Y = X\hat{\beta} + \epsilon^+ - \epsilon^-$$

com ϵ^+ e ϵ^- sendo vetores cujos elementos são todos não negativos.

ALGORITMOS DE PROGRAMAÇÃO LINEAR

Um dos primeiros algoritmos de programação linear para estimar os parâmetros pelo Método dos Mínimos Desvios Absolutos foi proposto por Barrodale & Roberts [3], sendo sua implementação uma adaptação do algoritmo *Simplex* para o problema de minimização de desvios absolutos. Computacionalmente, este algoritmo tem boa aplicação para banco de dados de até 5.000 observados com 50 variáveis, ou seja, para um volume de dados muito grande, este algoritmo pode apresentar limitações, conforme argumenta Chen & Wei [6].

Koenker & d'Orey [16] apresentam uma adaptação desse algoritmo para o problema da Regressão Quantílica, em que a principal modificação da rotina de Barrodale e Roberts é a adição de três novas linhas da matriz que contém o algoritmo *Simplex*, para armazenar a

decomposição do gradiente. A implementação desse algoritmo está detalhada em [16]).

Portnoy & Koenker [22] recomendam, para um banco de dado com muitas observações (grande dimensões), uma algoritmo de programação linear denominado Ponto Interior, que possui uma performance superior ao algoritmo *Simplex*.

A ordem de complexidade computacional está relacionada a quantidade de parâmetros (p) e a quantidade de observações (n), sendo a ordem de complexidade do algoritmo do Método do Mínimo Quadrados de $O(np^2)$ operações, conforme abordado por Portnoy & Koenker [22], ao passo que a ordem de complexidade do algoritmo da Reqressão Quantílica é de $O(n^{\frac{5}{2}}p^3)$, o que evidencia a facilidade computacional e rapidez do Método dos Mínimo Quadrados, sendo assim, um ponto importante para a preferência em sua utilização.

Para reduzir a desvantagem computacional que o Método dos Desvios Absolutos possui, Portnoy & Koenker [22] sugerem uma alteração no algoritmo para a Regressão Quantílica, adicionando uma etapa de pré-processamento no algoritmo. Considere o cálculo de uma estimativa preliminar baseada em uma subamostra de m observações, em que o algoritmo presume que os dados sofreram alguma randomização inicial, então a primeira m observações podem ser considerada representativa da amostra como um todo. Em algumas situações, essa melhoria apresentou desempenho similar ao Método dos Mínimos Quadrados. A implementação do algoritmo do Ponto Interior e do algoritmo com a etapa de pré-processamento está detalhada em [22]).

Implementação no R

As rotinas para a estimação dos parâmetros dos modelos de Regressão Quantílica estão implementados nos principais softwares de estatísticas, mas a referência mais completa é o pacote *quantreg* no software R (Koenker [12] e CRAN R-Project [7]).

No pacote *quantreg* , para o método *Simplex* deve-se usar o argumento $method = "br"$, para o método de Ponto Interior utiliza-se o argumento $method = "fn"$, ou se o interesse for o método com uma etapa de pré-processamento, que melhora a performance do algoritmo, o argumento a ser usado é o $method = "pfn"$.

Esses argumentos são implementados na função rq, que tem o padrão:

$$rq(formula, tau = 0.5, method = "br")$$

A formula tem o seguinte padrão para entrada dos argumentos $Y \sim X_1 + X_1 + ... + X_k$, em que Y é a variável respostas e $X_1 + X_1 + ... + X_k$ são as covariáveis preditoras.

Para obter estimadores para um quantil diferente de $\tau = 0,5$ (mediana) é necessário especificar o τ nos argumentos. É possível gerar estimativas para mais de um quantil na mesma função, sendo que para isso é necessário informar o argumento $\tau = seq(\tau_1, \tau_k, by = \tau_i)$, em que seq é o argumento para uma sequência de quantis, que deve ter como entradas τ_1 o primeiro quantil, τ_k o último quantil e o argumento by, que irá definir τ_i os quantis intermediários ("pulos"). Por exemplo, $\tau = seq(0.15, 0.75, by = 0.15)$ irá gerar as estimativas para os quantis 0,15, 0,30, 0,45, 0,60 e 0,75.

Outro argumento para gerar estimativas para mais de um quantil na mesma função é $\tau = c(\tau_1, \tau_2, ..., \tau_k)$, em que c é o argumento para um vetor de quantis, que deve ter como entradas τ_1 o primeiro quantil até τ_k o último quantil. Por exemplo, $\tau = c(0.2, 0.3, 0.5, 0.7, 0.75, 0.8)$ irá gerar as estimativas para os quantis 0,20, 0,30, 0,50, 0,70, 0,75 e 0,80.

Por padrão o R utiliza o método *Simplex* ($method = "br"$), sendo assim caso o interesse seja utilizar outro método é necessário informar o argumento $method = "fn"$ para o Ponto Interior ou $method = "pfn"$ para o método com uma etapa de pré-processamento.

2.2.2 Interpretação das Estimativas

Tomando um modelo de Regressão Quantílica com apenas uma variável X, tem-se:

$$Y_\tau = \beta_0(\tau) + x_1\beta_1(\tau)$$

Com isso, ao variar 1 unidade em x_1 gera-se o efeito de variação de $\beta_1(\tau)$ no Y_τ estimado, em que τ é o quantil que se está realizando a análise e $\beta_1(\tau)$ é o parâmetro associado à variável independente X_1 no τ-ésimo quantil.

Para a mediana, tem-se que a estimativa do coeficiente é interpretada como a alteração mediana da variável dependente correspondente para uma unidade de mudança da variável independente, ou seja, aumenta-se β para os indivíduos que se encontram no quantil 0,5. De forma análoga, o aumento de uma unidade na variável X irá implicar no aumento de β na variável $Y|\tau$.

A interpretação é similar à da Regressão Linear pelo Método dos Mínimos Quadrados Ordinários (MQO), porém, em vez de a variação ocorrer na média (valor esperado), que será igual para todos os quantis, na Regressão Quantílica a variação ocorre no τ-ésimo quantil que está em análise.

Os coeficientes obtidos na Regressão Quantílica para um quantil particular devem diferir significativamente daqueles obtidos a partir da regressão pelo MQO. Se isso não acontecer o uso da Regressão Quantílica não é justificável. Esta avaliação pode ser feita observando os intervalos de confiança dos coeficientes de regressão das estimativas obtidas a partir de ambas as regressões [18], conforme pode ser observado na Figura 2.1, que traz um exemplo da comparação dos resultados da regressão pelo MQO e a Regressão Quantílica.

A Figura 2.1 foi gerada utilizando o banco de dados do R [25] (swiss {datasets}), que apresenta a medida padronizada de fertilidade e indicadores socioeconômicos para cada uma das 47 províncias francófonas da Suíça por volta de 1888.

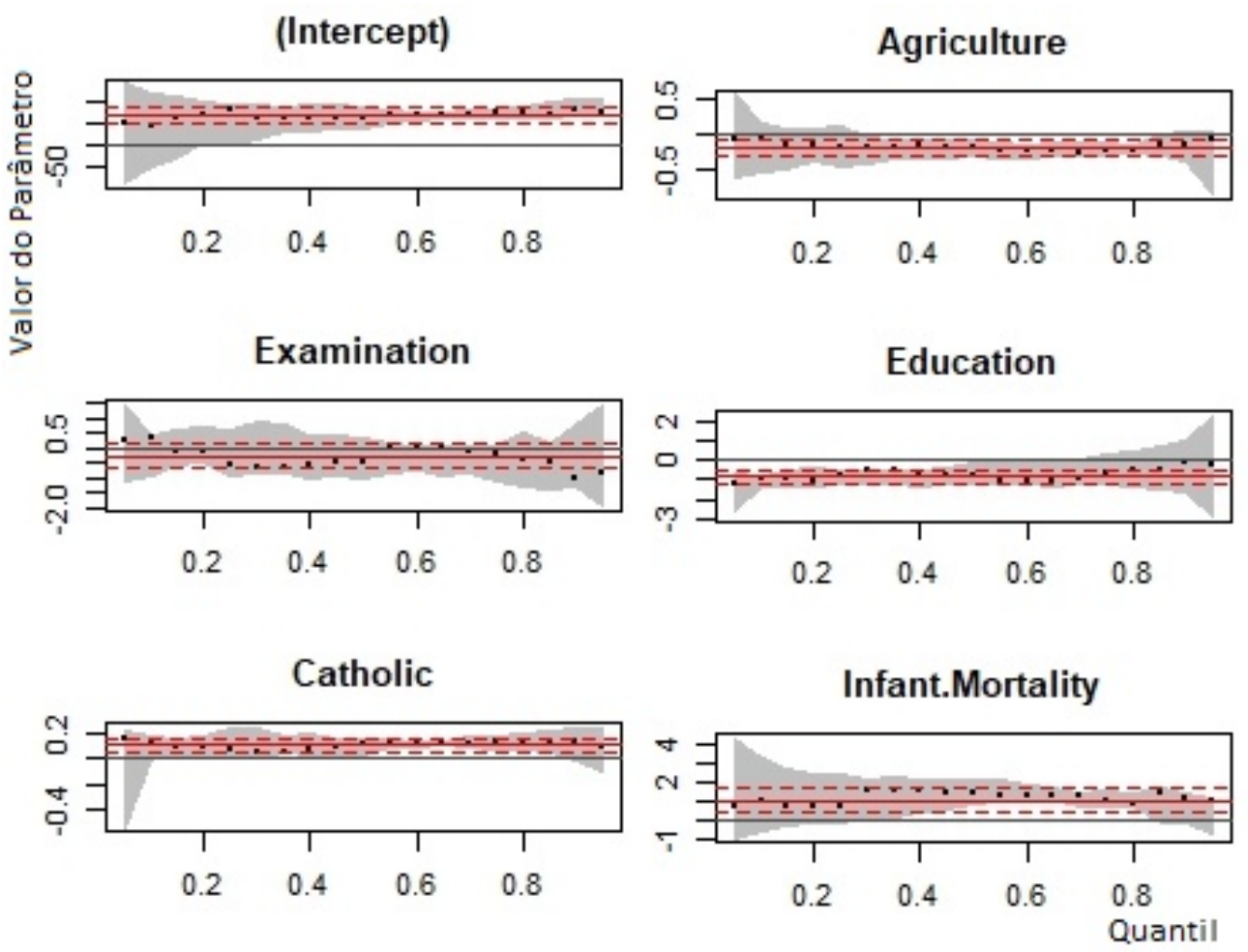

Figura 2.1: Gráfico comparativo da regressão pelo MQO e a Regressão Quantílica

Na Figura 2.1 vários quantis são representados pelo eixo X. A linha central vermelha indica as estimativas dos coeficientes pelos Mínimos Quadrados Ordinários (MQO) e as linhas vermelhas tracejadas são os intervalos de confiança em torno desses coeficientes do MQO. A linha pontilhada preta são as estimativas pelos Mínimos Desvios Absolutos (MDA) e a área cinza é o intervalo de confiança para os coeficientes do MDA para vários quantis. Pode ser visto que, para todas as variáveis, a regressão estimada coincide para a maioria dos quantis. Portanto, o Regressão Quantílica não é justificável para tais quantis. Para que se justifique a utilização da Regressão Quantílica as linhas vermelhas e a área cinza precisam ter a menor área de sobreposição possível.

2.2.3 Avaliação do Ajuste

Koenker & Machado [15] sugerem avaliadores de qualidade de ajuste específicos para Regressão Quantílica. Entre os avaliadores propostos, estão uma medida análoga ao R^2 convencional (coeficiente de determinação), e também processos baseados em Testes de Razão de Verossimilhanças e o Teste de Wald. Estes testes medem a qualidade dos modelos para um quantil em particular e não para distribuição inteira, sendo uma medida para a qualidade de ajuste local

para o quantil em questão.

Coeficiente de Determinação

Pela fácil implementação no R será utilizado a medida análoga ao R^2 (coeficiente de determinação) [1].

O coeficiente de determinação R^2 pode ser interpretado como o percentual da variabilidade da variável resposta que é explicado pelas variáveis preditoras, sendo utilizado muitas vezes como uma medida de qualidade de ajuste.

Koenker & Machado [15] sugerem um critério análogo ao R^2, considerando os parâmetros estimados por um modelo de Regressão Quantílica. Enquanto o R^2 mede o ajuste do modelo em relação à função condicional da média em termos da variância residual, este critério mede a qualidade de ajuste do modelo em um quantil específico em termos do somatório absoluto dos resíduos. Assim, esta medida avalia o ajuste do modelo apenas para um dado quantil, e é calculada por

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)}$$

em que

τ é o quantil do modelo ajustado

$\hat{V}(\tau)$ é a soma dos erros absolutos ponderados do modelo completo em τ

$\tilde{V}(\tau)$ é a soma dos erros absolutos ponderados do modelo reduzido em τ

O valor deste critério está entre 0 e 1. Quanto maior o coeficiente de determinação $R^1(\tau)$, melhor a qualidade do modelo ajustado.

3. Metodologia

A Regressão Quantílica será aplicada na análise dos dados fornecidos por uma empresa de tecnologia para avaliar o potencial de compras de seu mercado consumidor. Sendo assim, será necessário obter dados externos à empresa para ser possível avaliar consumidores que ainda não são clientes, ou seja, com base nas características dos clientes deve ser possível inferir o potencial de compra tanto dos clientes que consomem menos, quanto dos consumidores que ainda não possuem relacionamento com a empresa.

Como a análise será no mercado B2B, ou seja, empresa vendendo para empresa, no trabalho será definido como Empresa a empresa objeto de estudo, e Clientes as empresas que são compradoras dos produtos da empresa pesquisada.

Para a definição do potencial de compra do mercado será avaliado o banco de dados da empresa com todos os seus clientes, em que o faturamento real deles será a variável resposta que o modelo irá predizer, ou seja, baseado no comportamento da base de clientes será elaborado inferência para o comportamento do mercado.

O resultado da modelagem estatística deve permitir a aplicação em clientes que não tem relacionamento com a empresa, ou seja, ainda não compraram os produtos desta. Isso servirá de indicativo para qualificar um cliente antes de ele ser abordado, pela equipe de vendas, tornando este processo mais acurado. Sendo assim, é importante que as variáveis preditoras possam ser avaliadas sem que o cliente tenta que passar quaisquer informação para empresa.

Neste contexto, serão analisadas variáveis de mercado que a empresa possui atualmente, tais como: Atividade Econômica, quantidade de CNPJ (Cadastro Nacional da Pessoa Jurídica) ativos e Receita Presumida, não serão trabalhadas outras informações, devido a necessidade de recursos financeiros para aquisição de outras características de mercado.

A variável Atividade Econômica é o CNAE (Classificação Nacional de Atividades Econômicas) que os clientes declaram para a Receita Federal, sendo está característica representativa do segmento de mercado que uma organização atual, ou seja, através desse código pode-se determinar se o cliente é do ramo: comercio varejista, banco, órgão público, tecnologia, agronegócio, indústria, entre outros.

A variável Receita Presumida é o quanto um cliente poderá receitar em um ano, e esta informação é adquirida de empresas que possuem modelos para calcular o quanto uma organização terá de receita em um ano, com o objetivo de comercializar tal informação. De forma geral, esta característica mensura o quanto um cliente tem de dinheiro para gastar.

A variável quantidade de CNPJ ativos é o número de filiais mais a matriz que um cliente

possui no mercado, por exemplo, a quantidade de lojas de um supermercado, agências de um banco, escritórios de uma empresa de serviços, instalações industriais, etc.

Com isso, estas variáveis (X) serão relacionados com o faturamento real (Y) da base de clientes, em que, pela regressão quantílica será estimado o quanto um cliente tem de potencial para consumir com a empresa pesquisada.

Os dados disponibilizados foram trabalhados em 2 ferramentas: o Excel [20] e o *software* R [24].

As variáveis de Faturamento Real e Receita Presumida possuem uma amplitude considerável, sendo a primeira de R$10,00 a R$ 3.695.188,32, e a segunda de R$70.000,00 aR$1.500.000.000,00. Portanto, ocorrendo a necessidade de transformação das variáveis, pelo logaritmo natural, para redução da variabilidade, o que confere uma melhor ajuste ao modelo.

A variável resposta ficou em função do logaritmo natural do Faturamento Real, sendo assim para se obter o valor do potencial de faturamento real é necessário fazer a transformação inversa, pelo exponencial, para ter-se o valor em R$ do potencial de compra que cada cliente possui.

O IBGE (Instituto Brasileiro de Geografia e Estatística) possui 87 divisões de CNAE, no entanto a empresa pesquisada possui em seu bando de dados clientes de 82 divisões. Com isso, as demais 5 divisões não participarão do estudo, e não terão potencial de consumo calculado, visto que não é possível incluí-las no modelo. Todavia, trabalhar com um modelo de regressão com uma variável categórica que tenha 82 divisões não é parcimonioso.

As 82 divisões CNAE foram agrupadas, utilizando a análise de cluster, conforme o faturamento médio e o 3º quartil do faturamento, dos clientes que pertencem a um CNAE. Primeiramente, foi aplicada uma técnica hierárquicos, utilizando o Método de Ward, para avaliar a quantidade de grupos sugeridos, e posteriormente aplicada o Método de k-means (uma técnica não hierárquica) para melhor definir os agrupamentos. Em que foi obtido 5 Grupos de CNAEs, que foram transformados em 4 variávies Dummies para o estudo do modelo.

Por fim, foi avaliada a variável quantidade de CNPJs, que, também, possui uma amplitude considerável, visto que se tem clientes com 1 CNPJ até um cliente com 5.562. No entanto, a utilização de transformações na variável não foi satisfatório, com isso foi aplicado análises descritivas no comportamento dessa variável em relação a variável resposta (faturamento real), onde pode ser observado que a partir de 9 filiais, não há mudanças significativas nas respostas, portanto para retirar o viés da amplitude da quantidade de CNPJ foi limitado o máximo com a quantidade de 9 CNPJ ativos, ou seja, para os clientes com quantidades maiores que 9 foi adotado o valor igual a 9.

A modelagem foi realizada no R, utilizando o pacote *quantreg* e a qualidade do ajuste avaliada pela medida análoga do Coenficiente de Determinação ($R^1(\tau)$) proposto por Koenker & Machado [15]. Sendo possível observar que o quantil 0,80 gera um resultado satisfatório para a inferência do potencial de compra do mercado.

Os resultados do estudo serão aplicados na obtenção de informações sobre os clientes, possibilitando que a empresa elabore ações para rentabilizar os clientes da base, ações para abordar *prospects* (consumidores que ainda não tem relacionamento (fizeram negócios) com a empresa),

e estudo para expansão de mercado, uma vez que é possível calcular o quanto os clientes de uma nova cidade (que a empresa ainda não atua) tem propensão para gastar com os produtos da empresa estudada.

4. Resultados

As empresas necessitam aumentar o seu faturamento com o menor esforço (custo) possível, sendo assim uma modelagem que prevê o quanto um cliente ou *prospect* (consumidores que ainda não possuem relacionamento com a empresa, ou seja, não adquiram seus produtos) pode consumir dos produtos destas empresas irá auxiliá-las na tomada de decisão sobre as estratégias de mercado para abordar e gerar negócios com os clientes e potenciais clientes.

Com o objetivo de prever o potencial de compra de empresas do mercado consumidor de uma empresa de tecnologia, os dados disponibilizados por ela foram inicialmente analisados por estatísticas descritivas com o objetivo de identificar a melhor maneira de inserir as informações no modelo, de forma a extrair o máximo de informações dos dados com a menor perda possível.

4.1 Caracterização das Variáveis

4.1.1 Faturamento Real

A análise do faturamento real que cada cliente tem com a empresa é muito importante, pois ele será a variável resposta, que irá calibrar o cálculo do potencial de consumo dos clientes com características similares. A modelagem será melhor explorada no desenvolvimento do trabalho, no entanto o objetivo principal do estudo é definir o quanto um cliente (ou *prospect*) poderá faturar. Para isso, serão analisados os faturamentos reais de uma banco de dados com 69.737 clientes, associados com algumas variáveis preditoras, buscando o melhor modelo para prever o potencial de consumo, ou seja, o quanto um consumidor (seja cliente ou não) poderá comprar (faturar) com a empresa pesquisada.

Analisando a referida base de clientes dessa empresa de tecnologia, tem-se a Tabela 4.1, que apresenta a distribuição da base de clientes por faixa de faturamento.

Percentual Clientes	Faturamento Real
74,8%	até R$1.000,00
21,5%	entre R$1.000,01 e R$10.000,00
3,5%	entre R$10.000,01 e R$100.000,00
0,2%	acima de R$100.000,00

Tabela 4.1: Percentual de clientes por faixa de faturamento real

Observa-se ainda que a média de faturamento é R$2.424,08 e a mediana é R$293,98, o

que mostra uma concentração de valores baixos (85% dos faturamentos são menores que a média) e tem-se 3,7% dos valores acima de R$10.000,00, que fazem a média ser muito maior que a mediana. Isso pode ser notado também pela amplitude dos dados, de R$10,00 até R$3.695.188,32, sendo melhor explorado na figura 4.1, o qual apresenta um gráfico de dispersão do faturamento real.

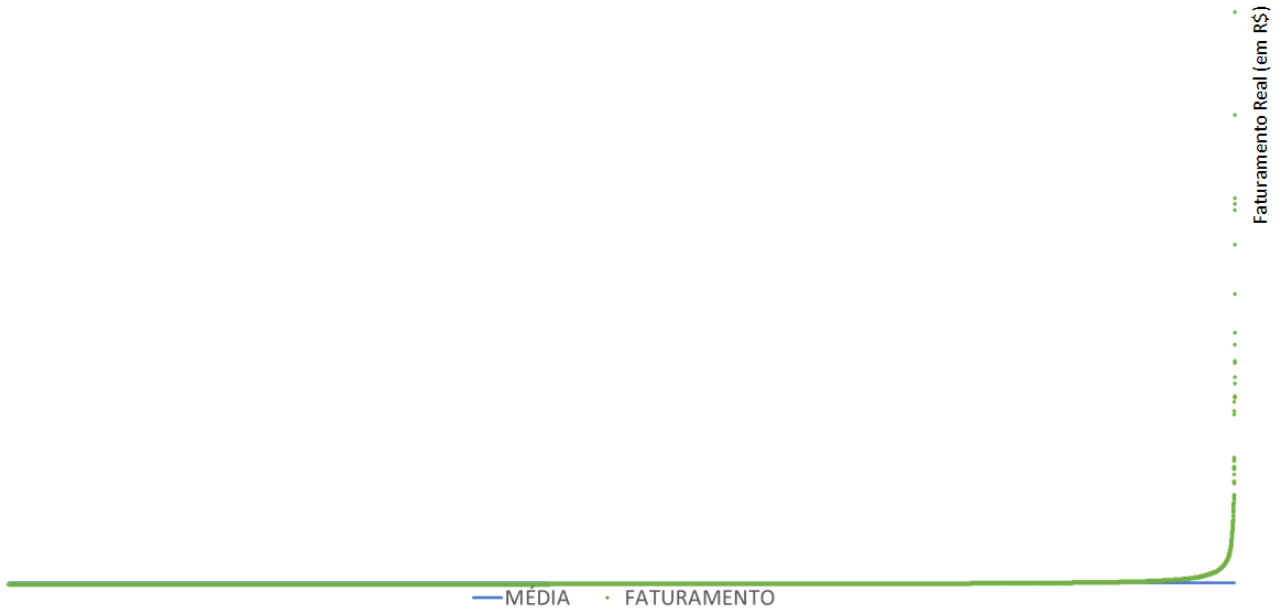

Figura 4.1: Gráfico de dispersão do faturamento real

Como pode ser observado na figura 4.1 tem-se uma concentração muito grande de dados abaixo da média e valores extremos muito acima da média. A construção de um boxplot facilita a observação desses *outliers*, porém devido à amplitude dos dados, a visualização fica comprometida. Para melhorar a visualização, os valores acima de R$10.000,00 (3,7% dos dados)foram omitidos e o boxplot refeito como pode ser observado na figura 4.2. Percebe-se portanto nesta Figura, que mesmo omitindo os $3,7\%$ dos dados mais extremos, existe uma grande quantidade de *outliers*.

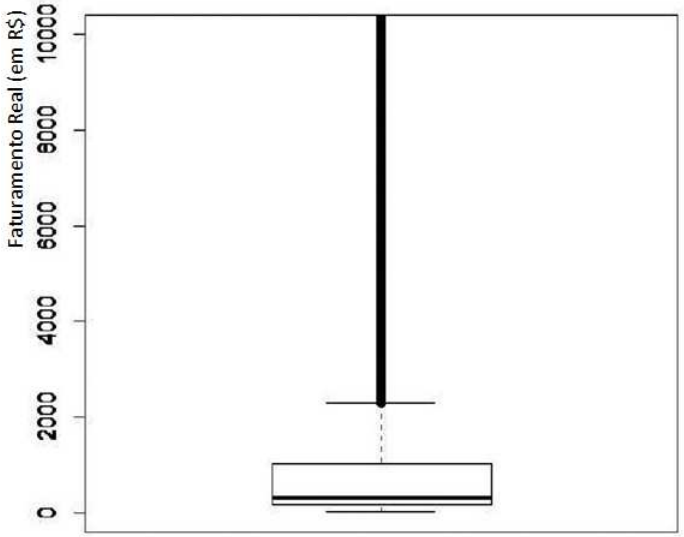

Figura 4.2: Boxplot do faturamento real

Por mais que a Regressão Quantílica seja robusta aos *outliers*, tem-se uma amplitude muito grande, fazendo com que a modelagem fique comprometida. Sendo assim, é necessário a transformação da variável faturamento para se obter uma homogeneidade dos dados. Para tanto, aplicou-se a transformação pelo logarítmo natural, em que se obtém-se uma média ($R\$6,10$), próxima da mediana ($R\$5,68$), sugerindo um conjunto de dados mais homogênio que os valores reais, como pode ser observado na Figura 4.3, que apresenta o gráfico de dispersão para o Logaritmo Natural do Faturamento Real.

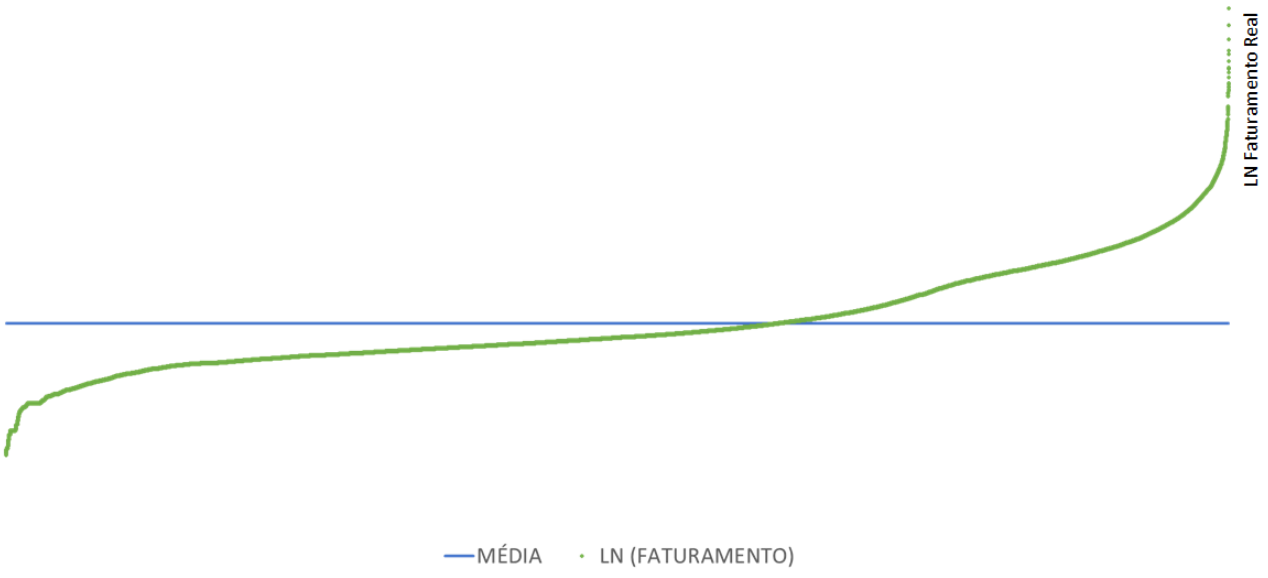

Figura 4.3: Gráfico de dispersão do logarítmo natural do faturamento real

Na Figura 4.4 pode ser observado que houve a redução da quantidade de valores extremos, no entanto, ainda percebe-se presença de alguns. A presença desses *outliers* sugere que algumas empresas faturam acima da média, o que é muito bom para o propósito do trabalho, pois tem-se empresas que possuem um potencial para consumo ainda não explorado. Sendo assim, é

necessário identificar as variáveis que possibilitem o cálculo de um faturamento ótimo a ser alcançado. No decorrer do trabalho será melhor explorado este ponto.

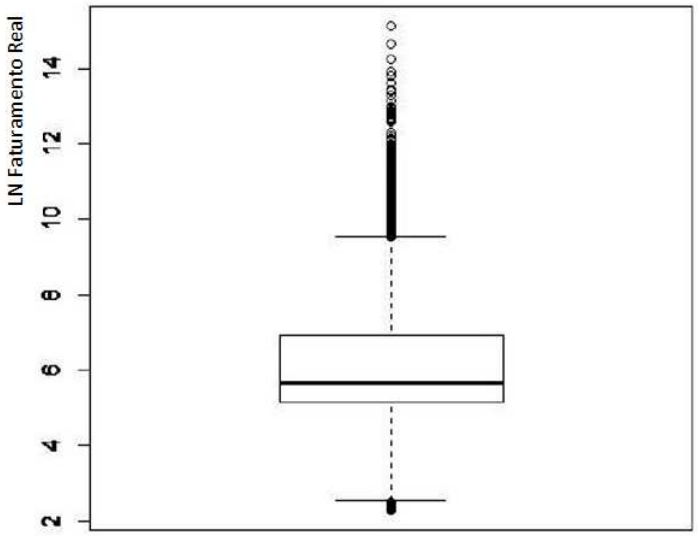

Figura 4.4: Boxplot do logarítmo natural do faturamento real

4.1.2 Atividade Econômica

Uma das variáveis preditoras do modelo é a atividade econômica, que será padronizada pelo CNAE (Classificação Nacional de Atividade Econômica) do IBGE, que todas as empresas possuem em seu cadastro na Receita Federal.

No entanto, o CNAE possi 87 divisão diferentes para definir a atividade econômica de uma empresa, sendo que na base de clientes estudada tem-se 82 dessas divisões, refletindo em uma dispersão muito grande dos dados, além de algumas divisões com pouca representatividade. Com isso, estas divisões foram clusterizadas para se obter perfis semelhantes de atividades econômicas, com o propósito de extrair a maior quantidade de informação dos dados disponíveis.

Na figuras a seguir são apresentadas média do faturamento e a quantidade de clientes das 82 divisões, na qual pode ser observado que algumas divisões de CNAE têm quantidade de eventos pequenos, o que poderia enviesar o estudo estatístico. A média de faturamento entre as divisões de CNAE são similares para alguns e bem distintas com outras. Sendo assim é necessário agrupar as divisões de CNAE por semelhança, de forma a obter grupos que possam ser representativos e tornar o modelo estatístico mais parcimonioso.

A Figura 4.5 apresenta a média de faturamento e a quantidade de clientes para as Divisões de CNAE com mais de 400 clientes, em que é possível observar que o CNAE 47 possui a maior quantidade de clientes e o CNAE 61 a maior média de faturamento, que no gráfico teve que ser cortado para que a escala comportasse as informações dos demais CNAEs.

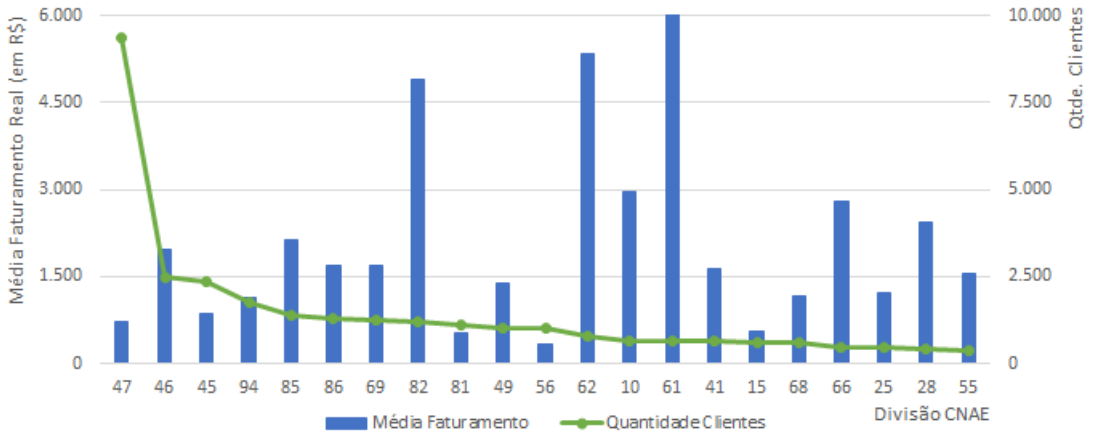

Figura 4.5: Gráfico com a média de faturamento e a quantidade de clientes para as Divisões de CNAE com mais de 400 clientes

A Figura 4.6 apresenta a média de faturamento e a quantidade de clientes para as Divisões de CNAE que tem de 100 a 400 clientes, em que é possível observar que os CNAE 63 e 65 possuem maior média de faturamento em comparação com os demais.

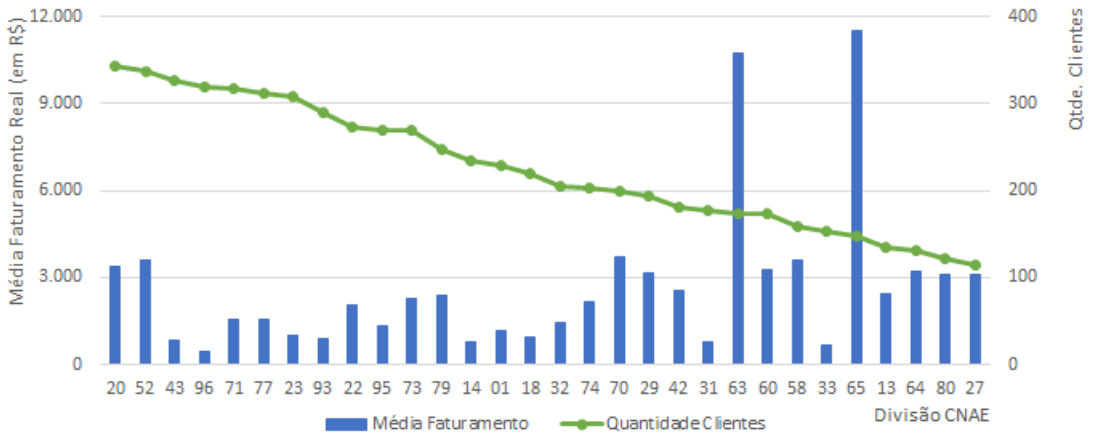

Figura 4.6: Gráfico com a média de faturamento e a quantidade de clientes para as Divisões de CNAE que tem de 100 a 400 clientes

A Figura 4.7 apresenta a média de faturamento e a quantidade de clientes para as Divisões de CNAE com menos de 100 clientes, em que é possível observar que alguns CNAEs tem baixa representatividade de observações.

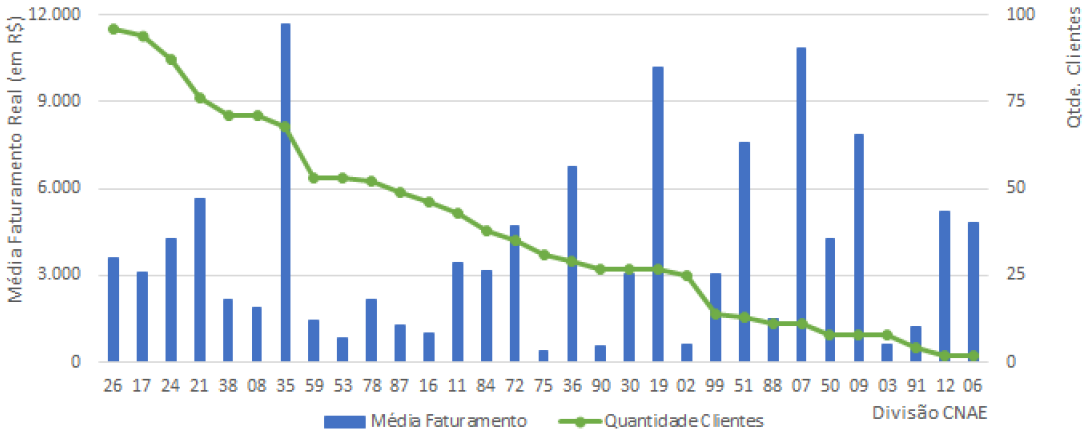

Figura 4.7: Gráfico com a média de faturamento e a quantidade de clientes para as Divisões de CNAE com menos de 100 clientes

Pode ser observado, também, no boxplot da Figura 4.8 a ocorrência sistemática de *outliers*, para várias categorias de CNAE. Ao fazer um agrupamento, espera-se uma redução na quantidade de *outliers*, porém não a eliminação de todos, pois observando as divisões 45, 47, 56, 81, 85, 93, 94, 95 e 96 pode-se ver que elas possuem muito *outliers* que serão carregados para seu agrupamento. Neste momento, é importante comentar que a presença desses pontos não representa algo ruim, pois está se falando apenas de uma variável. Na verdade, espera-se que quando forem inseridas as demais variáveis a composição entre elas deverá reduzir a existência dos *outliers*. Além disso, o objetivo do estudo é explorar uma fatia superior do faturamento (e não um ponto médio, seja a mediana ou a média), e entender quais são as características para elevar os menores faturamentos para um patamar "ótimo", não necessariamente os maiores, pois estes podem ser realmente pontos fora da curva, e deverão ser focos de outros estudos mercadológicos. A Figura 4.8 apresenta o boxplot do Logaritmo Natural do Faturamento Real para cada divisão CNAE.

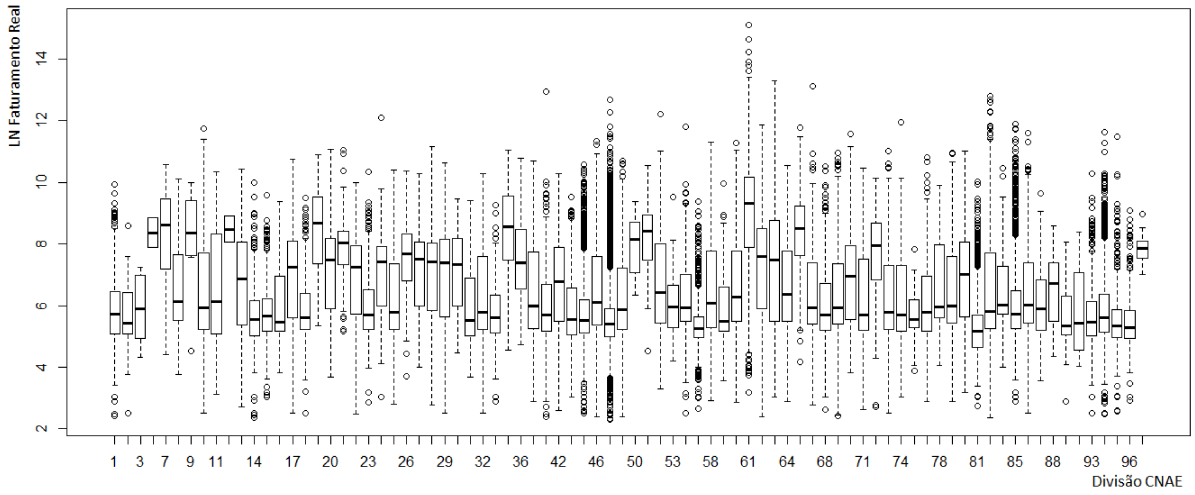

Figura 4.8: Boxplot do logaritmo natural do faturamento real por Divisão CNAE

Para agrupar as atividades econômicas semelhantes, foram utilizadas técnicas para análise

de agrupamento ou de conglomerado (análise de *cluster*), em que o Método de Ward, que é um método hierárquico aglomerativo, foi implementado para estimar a quantidade de *clusters* que devem ser considerados. Nesta técnica, por intermédio de um dendograma é possível observar a quantidade de agrupamentos sugestivos. Posteriormente, utiliza-se o Método de k-médias (técnica não hierárquica)para definir os *clusters*.

Mesmo antes de realizar a análise de *cluster*, é possível observar na estatística descritiva das Figuras 4.5 e 4.8, que a divisão CNAE 61 tem um destaque importante, em que o faturamento médio é 4 vezes maior que o segundo maior faturamento e a mediana e o 3º Quartil são 2 vezes maior que o segundo, sendo assim é de esperar que esta atividade econômica tenha um destaque no modelo de potencial, pois empresas que a possuem possivelmente terá maiores chances de ter um consumo elevado, em comparação com as demais atividades.

Como o objetivo do estudo é estimar o potencial de consumo dos clientes, ou seja, o quanto eles podem aumentar o seu faturamento com a empresa em estudo, a informação utilizada para clusterizar as divisões CNAE foi o faturamento real, sendo que foram consideradas duas variáveis: a média do faturamento e o 3º Quartil do faturamento para cada divisão CNAE. A formação dessas duas variáveis é devido a quantidade de *outliers*. Sendo assim, ao aplicar a análise de *clusters* a inclusão do 3º quartil do faturamento gerou um melhor balanceamento nos grupos.

Analisando a Figura 4.9, que apresenta o gráfico de dispersão dessas duas variáveis propostas, é possível observar quatro agrupamentos. Nesta Figura, identificamos uma divisão CNAE como *outlier*, sendo este correspondente ao CNAE 61, como já citado anteriormente, sendo esta mais uma evidência que ele sozinho irá formar um *cluster*.

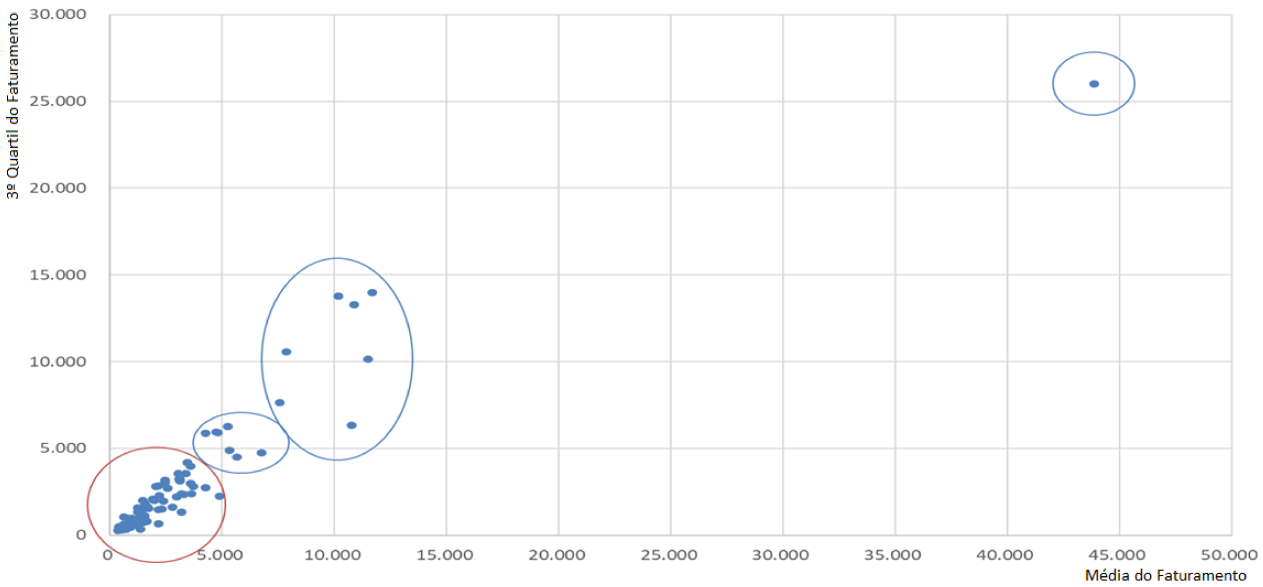

Figura 4.9: Gráfico de dispersão da média de faturamento para cada Divisão CNAE com o valor 3º quartil do faturamento

Na figura 4.10 apresenta-se uma ampliação no gráfico de dispersão, apenas para os casos de CNAEs com média e terceiros quartis menores que 5.000 de faturamento, para se adequar

a escala e possibilitar uma melhor visualização da dispersão. Nesta Figura, a aproximação dos pontos permite verificar que um dos grupos da Figura anterior pode ser dividido em dois grupos, de acordo com os valores da média e terceiro quartil. Assim, há uma sugestão de 5 grupos, em vez de quatro, como sugerido na Figura anterior.

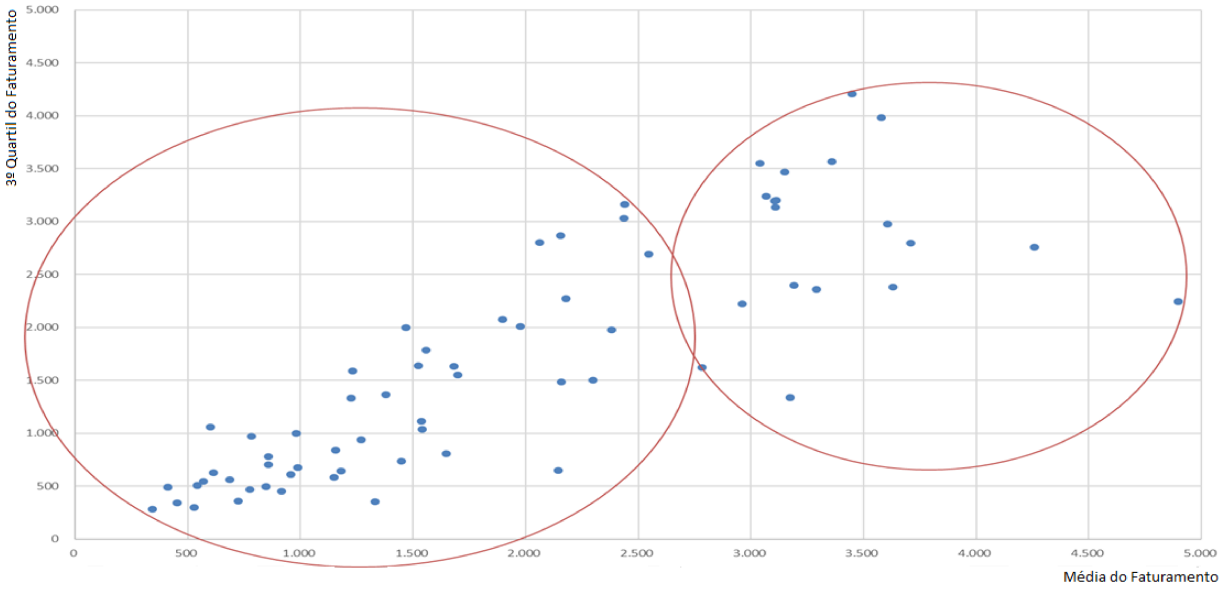

Figura 4.10: Gráfico de dispersão da média de faturamento para cada Divisão CNAE com o valor 3º quartil do faturamento, dos valores inferiores a 5.000

Para avaliar mais criteriosamente o número de grupos, o primeiro passo foi aplicar uma técnica hierárquica, ou seja, utilizar o método de Ward, cosiderando como medida de dissimilaridade a distância euclidiana. Aplicando-se esta técnica, obteve-se o dendograma (Figura 4.11) a partir do qual verifica-se uma sugestão para a quantidade de *clusters*. Em um primeiro nível tem-se a sugestão de 8 grupos e em um segundo momento temos a sugestão de 5 grupos. Sendo assim, para definir o melhor agrupamento será aplicado uma técnica não hierárquica.

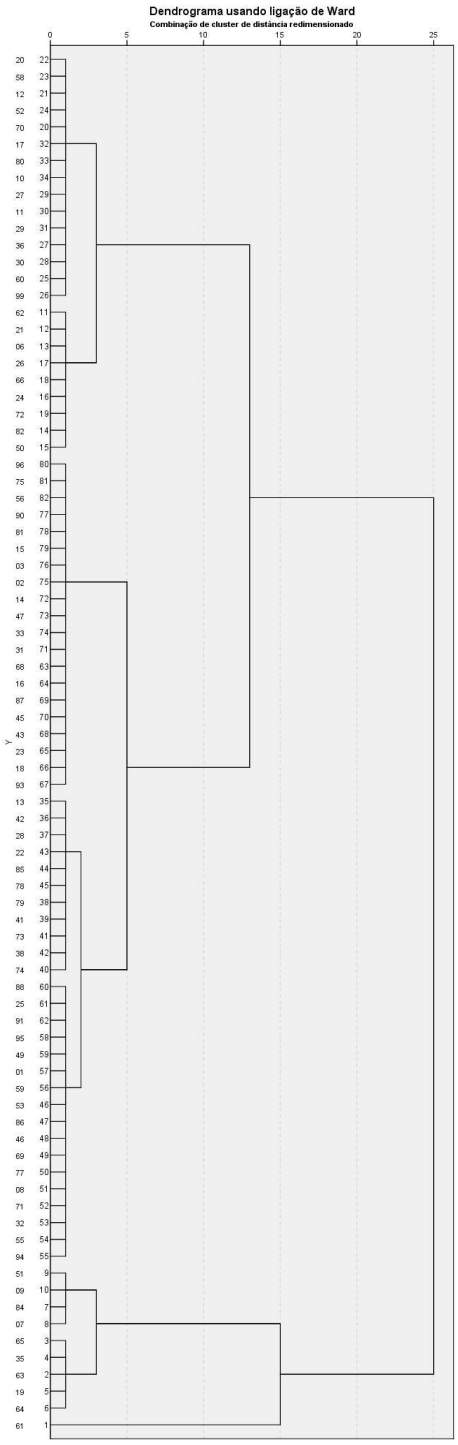

Figura 4.11: Dendograma usando ligação de Ward

Para avaliar a quantidade de grupos, para os CNAEs, foi aplicada a técnica de k-médias para 8 e 5 grupos, utilizando-se em seguida o teste F da Análise de variância (ANOVA) para cada agrupamento, para definir a melhor quantidade de grupos. De acordo com este teste, quanto maior a estatística F melhor é o agrupamento.

Aplicando a técnica de k-médias para 8 *clusters* observa-se a formação de alguns grupos solitários, conforme Tabela 4.2, o que em termos práticos não contribui para o agrupamento das divisões CNAE, no entanto é necessário avaliar e comparar com outra possibilidade de

agrupamento.

Cluster	Quantidade CNAE
1	1
2	3
3	45
4	27
5	1
6	1
7	2
8	2

Tabela 4.2: Formação de 8 *cluster* para CNAE

Utilizado a técnica de k-médias para 5 *clusters* obtém-se a Tabela 4.3, em que pode ser observada uma melhor segmentação das divisões CNAE.

Cluster	Quantidade CNAE
1	6
2	8
3	26
4	1
5	41

Tabela 4.3: Formação de 5 *cluster* para CNAE

Porém, para definir a quantidade de *cluster* mais adequada, foi utilizada a Anova, para comparar a estatística F, sendo o resultado na Tabela 4.4 que quando maior, melhor será o agrupamento.De acordo com o teste F, que quanto maior, melhor será o agrupamento, verifica-se que com 5 *Clusters* obtém-se uma segmentação mais adequada para a divisão CNAE.

Quantidade Cluster	Estatística F
8 Clusters	692,998
5 Clusters	1.195,307

Tabela 4.4: Estatística F para avaliar a quantidade de *clusters*

Para melhor identificar os *clusters*, foi assumida uma classficação, sendo o *cluster* que possui os menores valores, para a média será nomeado como 1, e sucessivamente até o *cluster* 5 que possui a maior média.

Na Tabela 4.5 observam-se algumas estatísticas descritivas do faturamento para os *clusters* formados, sendo possível notar que realmente existem diferenças entre os clusters, com relação à essas medidas. E como já citado anteriormente, um dos *clusters* formados tem apenas uma

divisão CNAE, que é a 61, em que observa-se uma diferença prática maior em relação aos demais *clusters*.

CLUSTER CNAE	MÉDIA	MEDIANA	1ºQUARTIL	3ºQUARTIL
cluster 1	1.121	262	164	570
cluster 2	3.363	612	227	2.717
cluster 3	5.406	2.143	401	5.047
cluster 4	11.076	3.344	784	9.943
cluster 5	43.864	11.000	2.650	26.000
TOTAL	2.424	294	176	1.017

Tabela 4.5: Estatística descritiva do faturamento para os *clusters* de CNAE

Como pode ser observado no *boxplot* da Figura 4.12 houve uma redução na quantidade de *outliers*, e como previsto, um dos agrupamentos, que é o *cluster* 1, continua com vários *outliers*. Isso ocorre devido à quantidade de empresas nos CNAEs agrupados. Como já abordado anteriormente, isso não é necessariamente ruim, pois ao cruzarmos com outras características poder-se-á diluir esses pontos, visto que o faturamento desses clientes pode estar associado a outras variáveis, tais como quantidade de filiais e receita presumida (o quanto um cliente espera ter de receita em um ano).

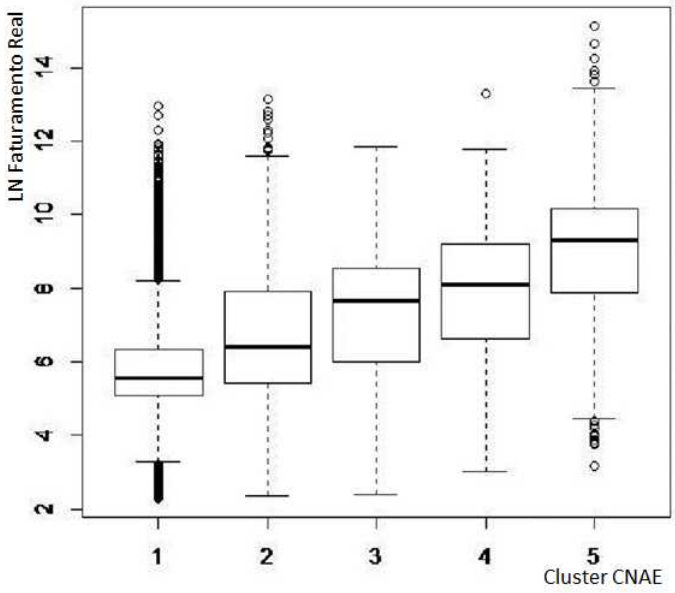

Figura 4.12: Boxplot do logarítmo natural do faturamento real por *cluster* de CNAE

4.1.3 Receita Presumida

A Receita Presumida é outra variável preditora utilizada nesse estudo. Esta variável é incluída no modelo, pois representa o porte econômico dos clientes, que em conjunto com sua atividade econômica, ajuda a descrever o mercado consumidor. A informação sobre essa variável é extraída junto ao Serasa Experian (é uma empresa que reúne informações, faz análises

e pesquisas sobre pessoas física e jurídicas). O Serasa Experian, baseado nas informações que possui sobre os clientes, modela os dados e calcula a provável receita que uma determinada empresa terá em um ano, ou seja, presume a receita anual.

A Receita Presumida é um dado que tem uma amplitude considerável, pois o menor valor é R$70.000,00 e o maior R$1.500.000.000,00, a média de R$26.075.272,38, mediana de R$640.000,00 e o 3º Quartil de R$2.150.000,00. Na Tabela 4.6 pode ser observado a participação de cada faixa de Receita Presumida.

Percentual Clientes	Receita Presumida
62,5%	até R$1.000.000,00 ao ano
25,5%	de R$1.000.000,01 a R$10.000.000,00 ao ano
8,4%	de R$10.000.000,01 a R$100.000.000,00 ao ano
2,5%	de R$100.000.000,01 a R$1.000.000.000,00 ao ano
0,8%	acima de R$1.000.000.000,00

Tabela 4.6: Percentual de clientes por faixa de receita presumida

Assim, facilmente pode ser constatada uma assimetria devido à concentração de valores abaixo da média e *outliers* com valores altos. Os mesmos efeitos foram observados na variável resposta, faturamento real dos clientes.

A construção de um *boxplot* facilita a observação dos *outliers*, porém devido a amplitude dos dados a visualização fica comprometida, sendo uma alternativa para melhorar a visualização construir o *boxplot* omitindo os valores acima de R$12.000.000,00 (11,8% dos dados), como pode ser observado na Figura 4.13.

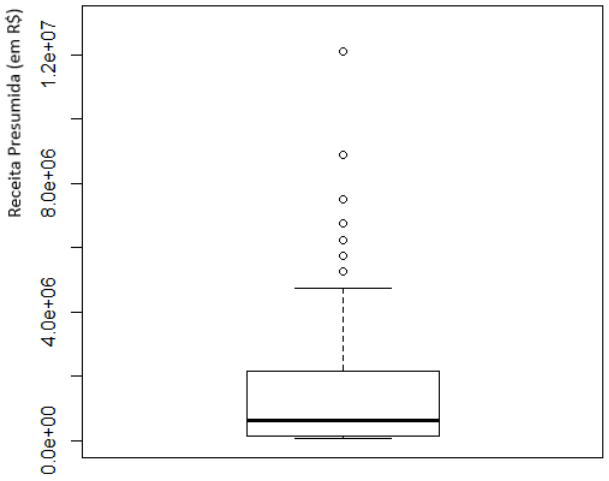

Figura 4.13: Boxplot da receita presumida

Para se obter uma homogeneidade dos dados é aplicada a transformação pelo logaritmo natural na variável Receita Presumida, em que obtém-se uma média (13,68) próxima da mediana (13,37), sugerindo um conjunto de dados mais homogêneo que os valores reais, como pode ser observado na Figura 4.14, com a redução dos *outliers*.

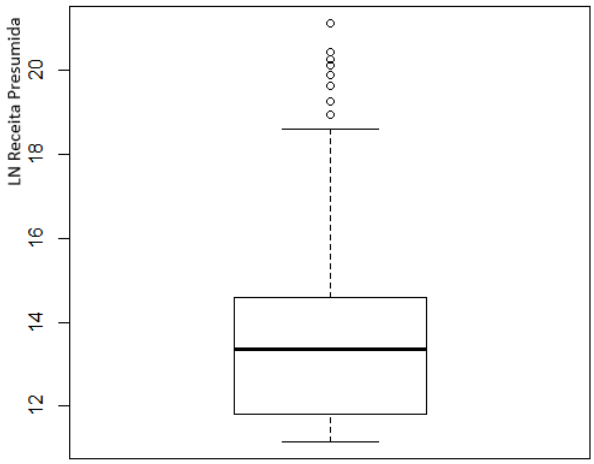

Figura 4.14: Boxplot do logarítmo natural da receita presumida

4.1.4 Quantidade de CNPJ Ativos

A quantidade de CNPJs que uma empresa tem ativos, também, é uma variável explicativa considerada no estudo. Espera-se que quanto mais filiais um cliente tem, maior é a chance de ela comprar mais produtos/serviços. No entanto, para essa variável, ocorre uma amplitude considerável, em que se tem desde clientes com apenas 1 CNPJ (79% dos clientes) até um cliente que possui 5.562 CNPJs, sendo que a média é 3,64 e a mediana é igual a 1,0 que é o mesmo valor do 3º quartil. A transformação pelo logaritmo não é viável, pois 79% das informações seriam transformadas para o valor 0, o que não seria bom para a modelagem estatística. Outras transformações, como por exemplo a raiz quadrada, também, não reduz a amplitude consideravelmente.

Primeiramente, é analisado a distribuição da quantidade de empresas com o número de CNPJ ativos e a média do faturamento real. No entanto, para se ter uma boa visão do comportamento dessa variável é necessário excluir os efeitos das empresas com o CNAE 61, que são as que consomem os produtos/serviços no atacado (em alto volume). Isso é necessário, pois nos eventos em que elas aparecem ocorre um viés na análise, pois esse grupo aumenta a média de forma desproporcional. Como se pode observar quando analisamos a média e o desvio padrão comparativo com e sem o CNAE 61, conforme a Tabela 4.7.

	Média	Desvio Padrão
Excluído o CNAE 61	1.703,93	8.332,79
Todos os CNAEs	2.424,08	28.046,42
Apenas o CNAE 61	43.863,74	200.919,41

Tabela 4.7: Comparação das médias geral e sem o CNAE 61

Avalia-se que a média teve um aumento relativo importante de 42,3%, quando é incluído o CNAE 61, e o desvio padrão ficar 3,4 vezes maior, em relação ao Total.

Para melhor avaliar o comportamento da variável Atividade Econômica é necessário excluir os efeitos dos clientes do CNAE 61, uma vez que ele apresenta um comportamento peculiar. Serão apresentados uma sequência de 3 gráficos para comparar a quantidade de clientes, o faturamento real médio e quantidade de CNPJs que cada cliente possui.

Como 79% dos clientes (29.976) possuem apenas 1 CNPJ a informação gráfica com este evento ficaria distorcida, portanto na Figura 4.15 foi omitida a quantidade de clientes (29.976) com 1 CNPJ, sendo apresentado apenas sua média de faturamento.

A Figura 4.15 apresenta para os clientes com quantidade de CNPJs de 1 a 15 as características da média de faturamento e a quantidade de clientes, em que pode ser observado que a média de faturamento cresce até 9 e 10 CNPJs e depois começa a cair e sobe novamente, apresentando um comportamento sem padrão a partir de 11 CNPJs. Outra característica que se observa é a queda da quantidade de clientes quando vai aumentando a quantidade de CNPJs, com 9 CNPJs tem-se 175 clientes, já com 15 tem-se apenas 43.

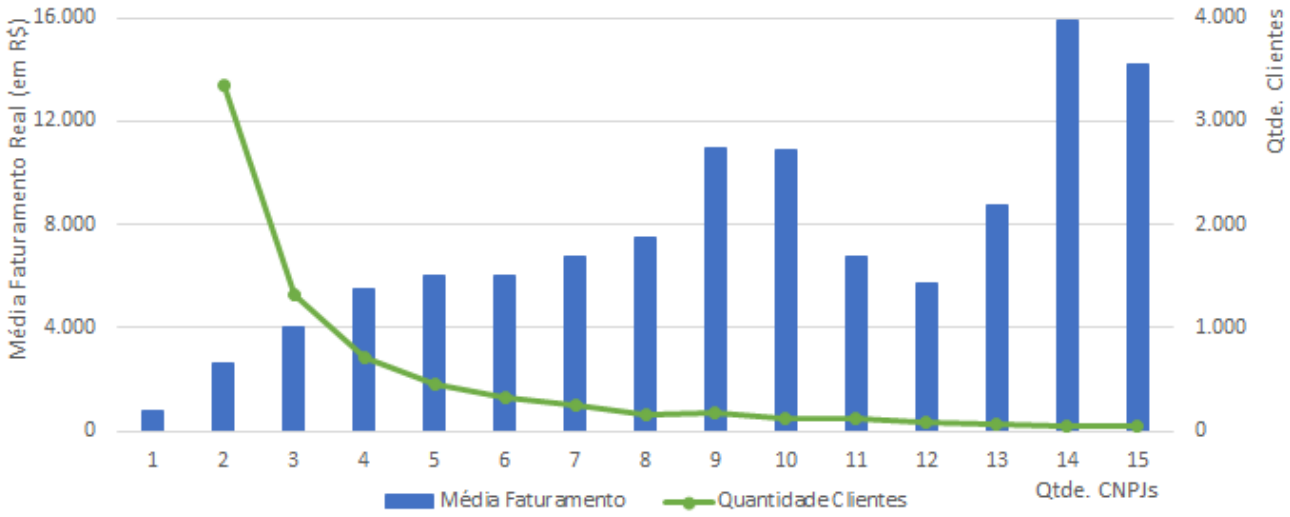

Figura 4.15: Gráfico da quantidade de CNPJ (de 1 a 15) com a quantidade de clientes e média de faturamento real

A Figura 4.16 apresenta para os clientes com quantidade de CNPJs de 16 a 71, as características da média de faturamento e a quantidade de clientes. Foi necessário a apresentação gráficos separados, para ser aplicado uma escala diferente, que possibilite a observação do comportamento das características propostas. Pode ser observado que a média de faturamento oscila muito devido a redução da quantidade de clientes por número de CNPJ, pois em alguns casos tem-se 2 ou 3 clientes, o que não permite uma boa referência para a média.

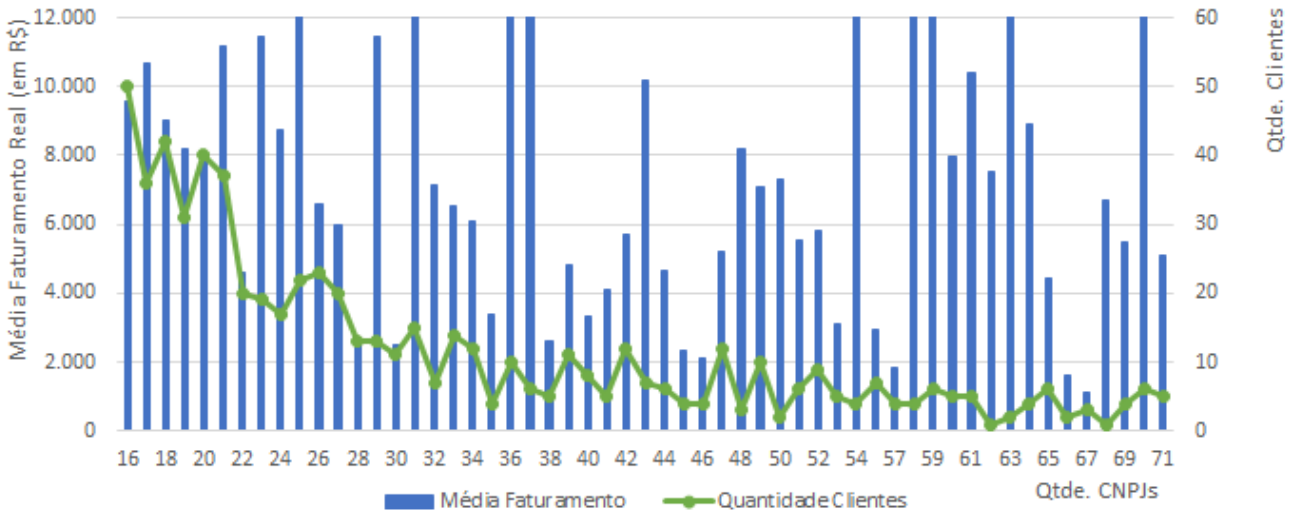

Figura 4.16: Gráfico da quantidade de CNPJ (de 16 a 71) com a quantidade de clientes e média de faturamento real

A Figura 4.17 apresenta para os clientes com quantidade de CNPJs de 72 a 5.562, as características da média de faturamento e a quantidade de clientes. Que pode ser observado que a média de faturamento possui um comportamento sem padrão, o que está relacionado a ocorrência de apenas 1 cliente com aquela determinada quantidade de CNPJs, sendo assim tecnicamente não é uma média, e sim um valor pontual.

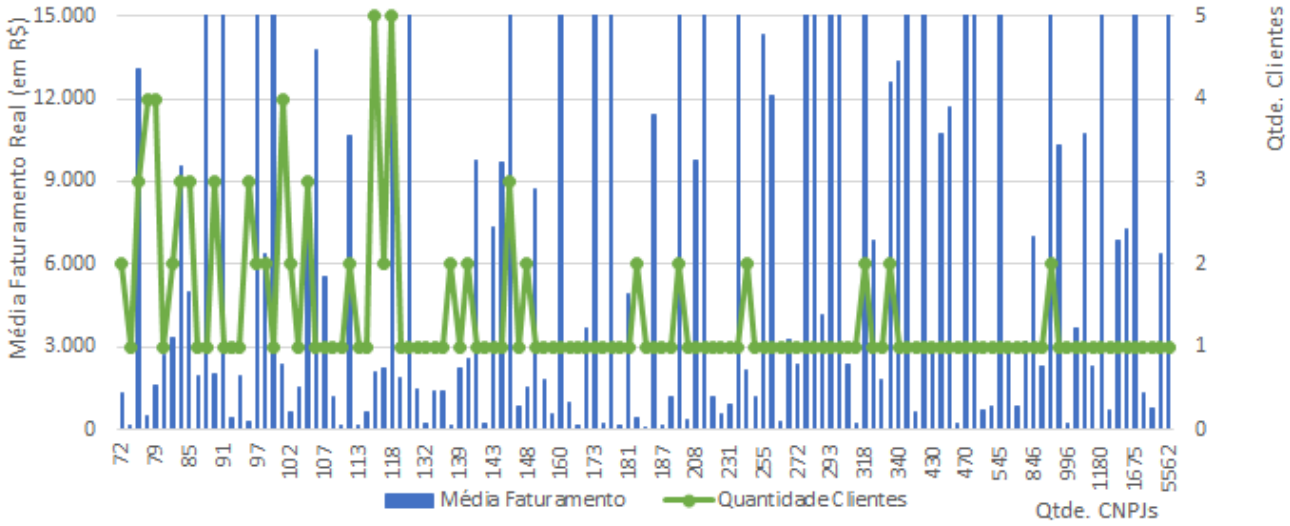

Figura 4.17: Gráfico da quantidade de CNPJ (de 72 a 5.562) com a quantidade de clientes e média de faturamento real

Pode ser observado que a partir de 9 filais a média do faturamento começa a ficar aleatória e a quantidade de clientes que possuem 10 ou mais filiais vai diminuindo e com isso os resultados vão perdendo informação, uma vez que qualquer um evento isolado enviesa as estatísticas

descritivas. Sendo assim, para mitigar estes efeito a quantidade máxima de filiais é limitada a 9, sem perda representativa, visto que a média do faturamento de todas as empresas com 9 ou mais filiais são equivalentes, como pode ser observado na Tabela 4.8.

	Média	Desvio Padrão
Exatos 9 filiais	R$10.975,52	R$47.343,26
Com 9 ou mais filiais	R$10.144,75	R$31.994,04

Tabela 4.8: Média de faturamento real dos clientes com 9 CNPJs e das que possuem 9 ou mais CNPJs

Na Figura 4.18 tem-se o *boxplot* comparativo do logaritmo do faturamento real entre as empresas com 9 filais e as que tem 9 ou mais filais, de forma que é fácil identificar que não há perda de informação ao agrupá-las, sendo possível reduzir o viés das empresas que possuem muitas filiais, porém não tem representatividade para a modelagem.

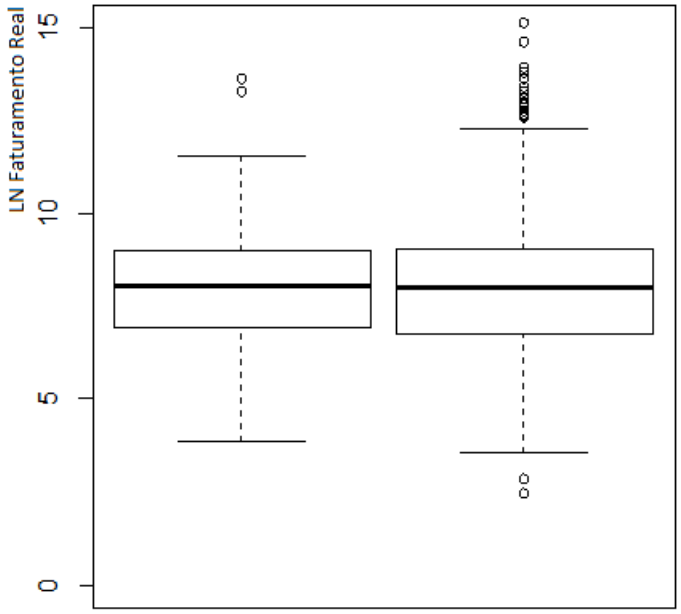

Figura 4.18: Boxplot comparativo do logaritmo natural do faturamento real para os clientes com 9 CNPJs e dos que possuem 9 ou mais

4.2 Modelagem

A Regressão Quantílica é geralmente usada quando se tem *outliers*, alta assimetria e heterocedasticidade nos dados. Essas características podem ser facilmente observadas na análise das variáveis, em que temos uma alta concentração de valores baixos, próximo ao zero, caracterizando uma assimétrica à direita, e por outro lado, vários outliers com valores altos, o que enviesa análises em torno da média.

Na Regressão Linear, predizemos a média da variável dependente para determinadas variáveis independentes. Como a média não descreve toda a distribuição, a modelagem da média não é uma descrição completa de uma relação entre variáveis dependentes e independentes. Assim, podemos usar a Regressão Quantílica, que prevê um quantil (ou percentil) para determinadas variáveis independentes.

Para o modelo de Regressão Quantílica foi utilizado o *software* R, no qual ele executa a função *rq*, que necessita do pacote *"quantreg"* instalado.

A estimativa do potencial de consumo do mercado tem o propósito de otimizar as negociações com os clientes, possibilitando a melhor penetração de produtos que explorem quase todas as necessidades do cliente. Diga-se, quase todas, pois obter tudo que o cliente tem de necessidade é algo utópico no mundo dos negócios, quando não se está falando de monopólio. Sendo assim, o modelo de potencial de mercado irá estimar uma parcela ótima da necessidade dos clientes em que a empresa possa trabalhar nas ofertas e relacionamentos com eles.

Para embasar o uso da Regressão Quantílica, na Figura 4.19, apresentam-se os gráficos com os resultados dos parâmetros para cada variável do modelo, no eixo Y, e cada quantil identificado no eixo X. Nesta figura (4.19) a linha vermelha representa o resultado quando se utiliza o Método dos Mínimos Quadrados (MMQ), com o intervalo de confiança do parâmetro representada pela linha tracejada vermelha. Já a curva preta representa os parâmetros para os diversos quantis (cada ponto na curva é um quantil), do 0,05 ao 0,95 de 0,05 em 0,05, e a área cinza é o intervalo de confiança para os parâmetros dos quantis.

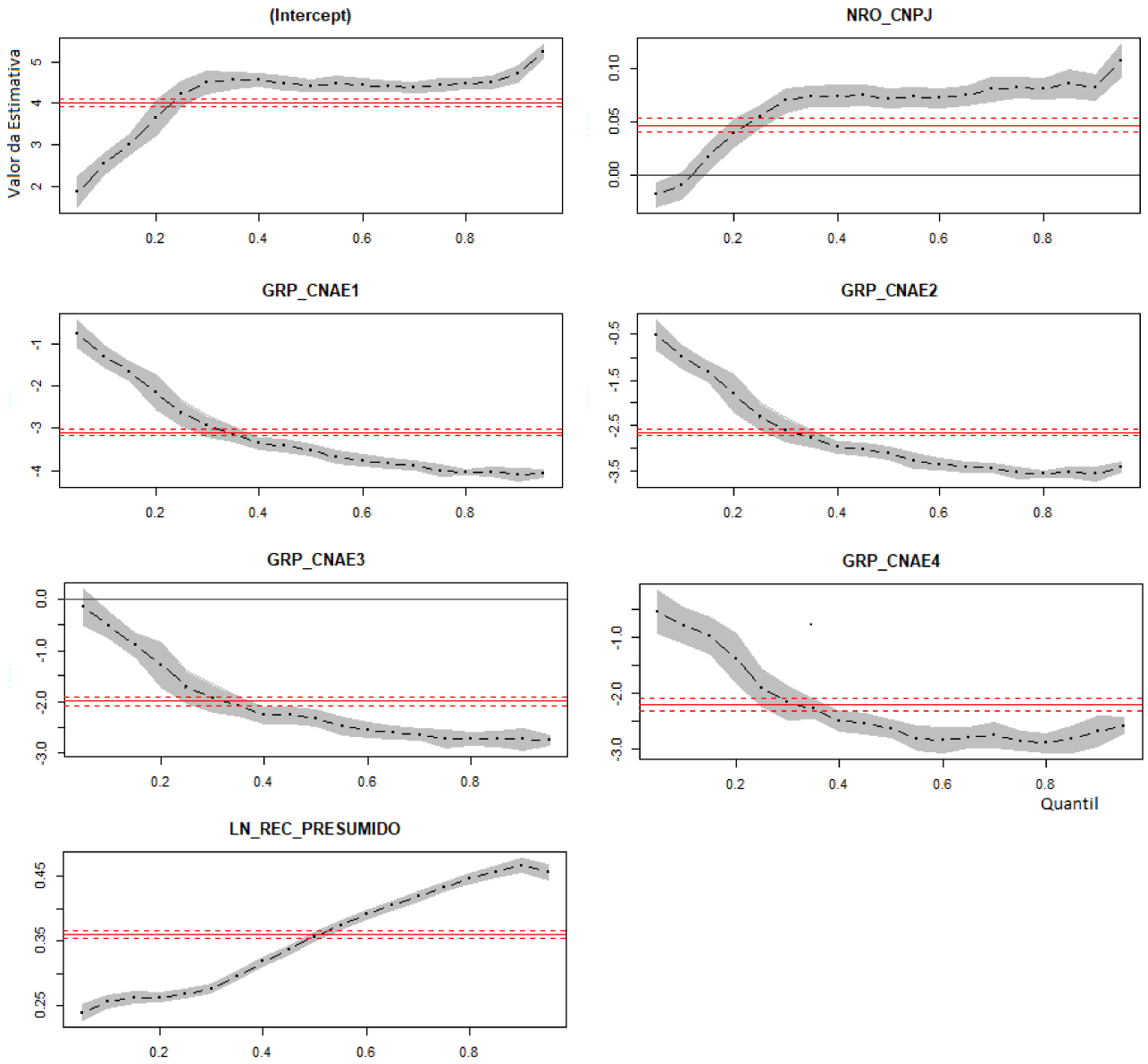

Figura 4.19: Gráfico com a comparação dos parâmetros pela regressão utilizando os MMQ em relação a Quantílica

Neste gráfico (Figura 4.19) espera-se que não ocorra sobreposição das estimativas dos parâmetros, como já abordado na Seção 2.2.2 pela Figura 2.1, visto que se os parâmetros para a Regressão Quantílica forem próximas da Regressão pelo MMQ, não se justificaria a utilização do método quantílico. Portanto, como pode-se observar, os parâmetros para a Regressão Quantílica são diferentes quando comparados com o Método dos Mínimos Quadrados, sendo, então, aplicável o estudo pelos quantis para se definir o potencial de mercado, estimando o quanto um cliente pode gastar com os produtos para empresa estudada.

4.2.1 ESTIMATIVAS PARA OS PARÂMETROS

Os parâmetros foram estimados para os modelos considerando os quantis de 0,60 ao 0,95, pois o propósito é otimizar a rentabilidade dos clientes,sendo assim estimativas acima da mediana

As tabelas a seguir apresentam os valores estimados para os parâmetros e a estatística de Teste t, calculada na função *rq* no R, possibilitando a avaliação dos parâmetros, verificando que a hipótese nula de que o valor do parâmetro, estatisticamente, igual a zero foi rejeitada em todos os casos.

A Tabela 4.9 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,60.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	4,42634	0,09091	48,68759	0,00000
QTDE CNPJS	0,07272	0,00546	13,31428	0,00000
LN REC PRESUMIDO	0,39112	0,00389	100,62688	0,00000
GRP CNAE1	-3,74988	0,07843	-47,81298	0,00000
GRP CNAE2	-3,34000	0,07962	-41,94954	0,00000
GRP CNAE3	-2,55096	0,08716	-29,26719	0,00000
GRP CNAE4	-2,84975	0,13871	-20,54394	0,00000

Tabela 4.9: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,60

A Tabela 4.10 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,65.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	4,40752	0,08232	53,54197	0,00000
QTDE CNPJS	0,07487	0,00555	13,49419	0,00000
LN REC PRESUMIDO	0,40548	0,00417	97,31197	0,00000
GRP CNAE1	-3,81975	0,06597	-57,89902	0,00000
GRP CNAE2	-3,40002	0,06821	-49,84611	0,00000
GRP CNAE3	-2,59882	0,08088	-32,13318	0,00000
GRP CNAE4	-2,80366	0,11093	-25,27412	0,00000

Tabela 4.10: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,65

A Tabela 4.11 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,70.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	4,37464	0,07497	58,35045	0,00000
QTDE CNPJS	0,08052	0,00678	11,88348	0,00000
LN REC PRESUMIDO	0,41935	0,00425	98,77056	0,00000
GRP CNAE1	-3,85903	0,05566	-69,33263	0,00000
GRP CNAE2	-3,41911	0,05838	-58,56520	0,00000
GRP CNAE3	-2,63903	0,06782	-38,91202	0,00000
GRP CNAE4	-2,75879	0,13902	-19,84498	0,00000

Tabela 4.11: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,70

A Tabela 4.12 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,75.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	4,43883	0,09380	47,32173	0,00000
QTDE CNPJS	0,08217	0,00551	14,92680	0,00000
LN REC PRESUMIDO	0,43376	0,00435	99,66331	0,00000
GRP CNAE1	-3,98746	0,07821	-50,98596	0,00000
GRP CNAE2	-3,52697	0,08039	-43,87516	0,00000
GRP CNAE3	-2,70979	0,10341	-26,20370	0,00000
GRP CNAE4	-2,86798	0,10450	-27,44533	0,00000

Tabela 4.12: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,75

A Tabela 4.13 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,80.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	4,46333	0,07066	63,16669	0,00000
QTDE CNPJS	0,08115	0,00605	13,41957	0,00000
LN REC PRESUMIDO	0,44660	0,00496	89,97772	0,00000
GRP CNAE1	-4,02856	0,03975	-101,35102	0,00000
GRP CNAE2	-3,55098	0,04418	-80,36904	0,00000
GRP CNAE3	-2,72068	0,06727	-40,44275	0,00000
GRP CNAE4	-2,90050	0,10618	-27,31758	0,00000

Tabela 4.13: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,80

A Tabela 4.14 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,85.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	4,49746	0,09109	49,37233	0,00000
QTDE CNPJS	0,08596	0,00789	10,89340	0,00000
LN REC PRESUMIDO	0,45714	0,00557	82,09252	0,00000
GRP CNAE1	-4,00857	0,06256	-64,07959	0,00000
GRP CNAE2	-3,51229	0,06620	-53,05322	0,00000
GRP CNAE3	-2,71454	0,09073	-29,91852	0,00000
GRP CNAE4	-2,83004	0,14302	-19,78733	0,00000

Tabela 4.14: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,85

A Tabela 4.15 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,90.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	4,70818	0,12201	38,58990	0,00000
QTDE CNPJS	0,08243	0,00749	11,00083	0,00000
LN REC PRESUMIDO	0,46735	0,00669	69,83417	0,00000
GRP CNAE1	-4,07958	0,09160	-44,53596	0,00000
GRP CNAE2	-3,55050	0,09614	-36,93098	0,00000
GRP CNAE3	-2,73188	0,12782	-21,37272	0,00000
GRP CNAE4	-2,68945	0,17104	-15,72409	0,00000

Tabela 4.15: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,90

A Tabela 4.16 apresenta os resultados do ajuste da regressão quantílica para no Quantil 0,95.

	Valor	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	5,24346	0,10579	49,56394	0,00000
QTDE CNPJS	0,10731	0,00944	11,37186	0,00000
LN REC PRESUMIDO	0,45633	0,00682	66,91283	0,00000
GRP CNAE1	-4,04865	0,05316	-76,15597	0,00000
GRP CNAE2	-3,40325	0,06373	-53,40485	0,00000
GRP CNAE3	-2,74702	0,05935	-46,28374	0,00000
GRP CNAE4	-2,58873	0,09348	-27,69372	0,00000

Tabela 4.16: Estimativas dos parâmetros para o modelo de regressão quantílica considerando o quantil 0,95

4.2.2 Avaliação do Ajuste do Modelo

A Tabela 4.17 apresenta o valores para o Coeficiente de Determinação ($R^1(\tau)$) para a Regressão Quantílica.

Quantil	$R^1(\tau)$
0,60	0,3173646
0,65	0,3434632
0,70	0,3666927
0,75	0,3822920
0,80	0,3835108
0,85	0,3771594
0,90	0,3674549
0,95	0,3515628

Tabela 4.17: $R^1(\tau)$ para os quantis de 0,60 a 0,95

Como tem-se o objetivo de otimizar o valor faturado pelo clientes, espera-se trabalhar na faixa do quantis 0,75 a 0,85, pois tentar adotar uma estratégia de mercado para capturar mais de 90% do potencial de um cliente seria utópico em um mercado de livre concorrência.

Observando os resultados para os vários quantis (Tabela 4.17) tem-se que o quantil 0,80 possui o maior valor para o $R^1(\tau)$, o que sugere um melhor ajuste aos dados. Portanto, serão utilizadas os parâmetros para o quantil 0,80 para descrever o modelo que será aplicado para o potencial de compra do mercado.

4.3 Apresentação do Modelo

Considerando a estimação dos parâmetros, conforme apresentado anteriormente na Tabela 4.13, para o quantil 0,80, o modelo de regressão quantílica é dado pela expressão a seguir:

$$\hat{Y}_{\tau=0,80} = 4,4633 + 0,0812x_1 + 0,4466x_2 - 4,0286x_3 - 3,5123x_4 - 2,7207x_5 - 2,9005x_6$$

em que:
$\hat{Y}_{\tau=0,80}$ = ln do Potencial de Consumo (estimador para o Faturamento Real) no quantil 0,80, sendo ln o logaritmo natural.
x_1 = Quantidade de CNPJs
x_2 = ln da Receita Presumida
x_3 = Grupo de CNAE1
x_4 = Grupo de CNAE2
x_5 = Grupo de CNAE3
x_6 = Grupo de CNAE4

As covariáveis Grupo de CNAE são variáveis *dummies*, em que a classe de referência é o Grupo de CNAE 5, formado pelo CNAE 61, que apresentam características peculiares debatidas anteriormente.

A variável resposta Faturamento Real foi transformada pelo logaritmo natural, para redução da variabilidade e obter uma homogeneidade dos dados. Sendo assim, a resposta do modelo fica em função do logaritmo, portanto há a necessidade de fazer a transformação inversa para obter o valor real do potencial de consumo do mercado, tendo-se portanto,

$$Potencial_{\tau=0,80} = \exp(\hat{Y}_{\tau=0,80})$$

A expressão do modelo pode ser reescrita como:

$$\exp(\hat{Y}_{\tau=0,80}) = \exp(4,4633+0,0812x_1+0,4466x_2-4,0286x_3-3,5123x_4-2,7207x_5-2,9005x_6)$$

No entanto, para facilitar a interpretação do modelo pelas diversas áreas da empresa tomar-se-á:

$$Potencial_{\tau=0,80} = \text{Potencial de Mercado}$$

Sendo assim, tem-se:

$$\begin{gathered}\text{Potencial de Mercado} \\ = \exp(4,4633 + 0,0812x_1 + 0,4466x_2 - 4,0286x_3 - 3,5123x_4 - 2,7207x_5 - 2,9005x_6)\end{gathered}$$

Os parâmetros são interpretados como a taxa de variação no quantil em análise ($\tau = 0,80$) ao se variar o valor da covariável. No entanto, devido às transformações utilizadas, a interpretação não é linear, ou seja, o simples incremento de uma unidade nas covariáveis não tem uma relação direta com a resposta (Potencial de Mercado), pelo fato de a expressão estar contida em um cálculo exponencial. Sendo assim, para se ter a interpretação, é necessário reescrever a equação para que ela demonstre a ideia da relação do incremento que cada covariável irá proporcionar na resposta.

$$\begin{gathered}\text{Potencial de Mercado} = \exp(4,4633) * \exp(0,0812x_1) * \exp(0,4466x_2) * \exp(-4,0286x_3) * \\ \exp(-3,5123x_4) * \exp(-2,7207x_5) * \exp(-2,9005x_6)\end{gathered}$$

Feito isso, é possível demonstrar a interpretação dos efeito das covariáveis na resposta, por exemplo a cada incremento de 1 Nro.CNPJ tem-se o aumento de 0,08115 no cálculo da exponencial que será multiplicado com as demais covariáveis. Para melhor exemplificar segue a Tabela 4.18, que simula um cliente com 1 CNPJ, com uma Atividade Econômica que o classifica no Grupo CNAE 2, e com a Receita Presumida de R$500.000,00, que transformando pelo logaritmo natural temos 13,1224 que será o valor aplicado na equação de cálculo:

	Parâmetro	Valor	Contribuição*
Intercepto	4,4633		86,7760
NRO CNPJ	0,0812	1	1,0845
LN REC PRESUMIDO	0,4466	13,1224	350,8811
GRP CNAE1	-4,0286	0	1
GRP CNAE2	-3,5123	1	0,0298
GRP CNAE3	-2,7207	0	1
GRP CNAE4	-2,9005	0	1
POTENCIAL DE MERCADO			R$985,00

* Contribuição = exp(Parâmetro x Valor assumido pela covariável)

Tabela 4.18: Exemplo de resposta da Regressão Quantílica

Com a Tabela 4.18 é fácil observar que o Potencial de Mercado é formado pela multiplicação das contribuições de cada covariável, em que para cada incremento no valor da covariável tem-se um fator multiplicativo que irá influenciar na resposta.

As covariáveis Número de CNPJ e Receita Presumida apresentam parâmetros positivos, o que realmente faz sentido, pois se espera que quanto mais filiais e maior receita, maior será a capacidade de um cliente comprar os produto e serviços da empresa estudada, ou seja, quanto maior o valor para estas covariáveis maior será o Potencial de Compra desses clientes.

Como esperado para os Grupos (*Clusters*) de CNAE, as 4 variáveis *dummies* inseridas no modelo têm estimativas dos parâmetros negativas. Isso ocorre devido à classe de referência (*cluster* 5) possuir média, mediana, 1º Quartil e 3º Quartil com valores bem superiores aos demais para a variável resposta (faturamento real), como comentado anteriormente e apresentado na Tabela 4.5.

Para evidenciar na prática o impacto do Potencial em cada *Cluster* de CNAE, na Tabela 4.19 tem-se algumas estatísticas descritivas do potencial de compra do mercado, estimado pelo modelo de Regressão Quantílica ajustado, para cada *Cluster*:

CLUSTER CNAE	MÉDIA	MEDIANA	1Q	3Q
Cluster 1	1.499,50	656,30	327,54	1.066,97
Cluster 2	5.157,75	1.678,81	592,15	4.663,91
Cluster 3	11.377,66	4.169,97	2.304,84	10.563,10
Cluster 4	27.685,17	11.304,56	2.829,95	33.429,50
Cluster 5	96.789,80	24.914,23	18.401,35	59.942,37

Tabela 4.19: Análise descritiva do Potencial por *Cluster* de CNAE

É notório que o potencial aumenta conforme o *cluster* de CNAE é alteado, sendo possível perceber que a clusterização dos CNAEs tem um efeito importante para a modelagem.

Quando aplicado o modelo de potencial na base de dados considerada, observou-se que

o menor potencial calculado é R$244,28, que está bem alinhado com a oferta que contém o portfólio de produtos da empresa, sendo o menor valor praticado igual a R$250,00. Portanto, tem-se um ajuste de modelo que atende e está bem balizado com o que a empresa vem praticando no mercado. Por outro lado, o maior potencial calculado é R$2.257.506,04, que, também, está alinhado com os valores encontrados no faturamento real da empresa, pois os clientes que possuem este potencial são os que necessitam e dependem dos produtos fornecidos, pela empresa pesquisada, em grande escala.

Por fim, a aplicação desse modelo pode ajudar a empresa a identificar os consumidores para melhor classificá-los e adotar estratégias de marketing mais adequadas para cada perfil. Podendo cruzar (ou não) o potencial de mercado com outros fatores, internos ou de mercado, para se obter *insigths* (novas ideias) para as tomadas de decisões.

5. Conclusões

O trabalho detalhou uma aplicação prática da Regressão Quantílica, destacando as diferenças e justificativas da implementação em relação à Regressão pelo Método dos Mínimos Quadrados, visto que os resultados apresentaram diferenças para cada quantil analisado, em relação aos resultados da regressão pelos Mínimos Quadrados Ordinários. Portanto, a Regressão Quantílica se apresentou mais robusta para a previsão em dados com alta variabilidade, concentração de valores em uma extremidade e *outliers* em outra (considerando-se o conjunto dos números Reais Positivo), ou seja, foi observada uma assimetria nos dados.

Com isso, foi possível modelar os dados para a obtenção do potencial de consumo do mercado, em que o quantil 0,80 se deu como o mais aderente para maximizar o retorno que a empresa possa esperar de seu clientes. Sendo assim, os resultados do modelo podem ser utilizados para o mapeamento do mercado consumidor, elaboração de planos de ação para rentabilização de clientes da base de dados, construção de estratégias para atrair novos clientes, planejamento para entrada da empresa em novas áreas geográficas, dentre outros tomadas de decisões que permitam o incremento nos resultados financeiros.

Referências Bibliográficas

[1] Acevedo, C., Albano, C. e Eustis, A.: *Quantile Regression*, 2018. https://kevintshoemaker.github.io/NRES-746/QuantileRegression.html, acessado em 03/07/2019.

[2] Alves, C. A.: *Classificação de Ratings, Sustentabilidade e Previsão de Default: uma abordagem utilizando a regressão quantílica*. Dissertação de Mestrado, 2014.

[3] Barrodale, I. e Roberts, F. D. K.: *An Improved Algorithm for Discrete L_1 Linear Approximation*. SIAM Journal on Numerical Analysis, 10(5):839–848, 1973. https://epubs.siam.org/doi/abs/10.1137/0710069.

[4] Bloomfield, P. e Steiger, W. L.: *Least Absolute Deviations: theory, applications and algorithms*. Birkhäuser, 1ª ed., 1983.

[5] Cade, B. S. e Noon, B. R.: *A gentle introduction to quantile regression for ecologists*. Frontiers in Ecology and the Environment, 1(8):412–420, 2003. https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/04-0785.

[6] Chen, C. e Wei, Y.: *Computational Issues for Quantile Regression*. Sankhyä: The Indian Journal of Statistics, 67(2):399–417, 2015. https://www.jstor.org/stable/25053439.

[7] CRAN R-Project: *Package 'quantreg': Quantile Regression*, 2019. https://cran.r-project.org/web/packages/quantreg/quantreg.pdf, acessado em 09/06/2019.

[8] Diniz, S. C.: *Análise do Consumo de Bens e Serviços Artístico-culturais no Brasil Metropolitano*. Dissertação de Mestrado, 2009.

[9] Downing, D. e Clark, J.: *Estatística Aplicada*. Saraiva, 3ª ed., 2011.

[10] Hao, L. e Naiman, D. Q.: *Quantile Regression. Series: Quantitative Applications in the Social*. SAGE Publications, 1ª ed., 2007.

[11] Hoffmann, R.: *Análise de Regressão: uma introdução à econometria*. O Autor, 5ª ed., 2016.

[12] Koenker, R.: *Quantile Regression*. Cambridge University Press, 1ª ed., 2005.

[13] Koenker, R. e Bassett, G. J.: *Regression Quantiles.* Econometrica, 46(1):33–50, 1978. https://www.econometricsociety.org/publications/econometrica/1978/01/01/regression-quantiles.

[14] Koenker, R. e Hallock, K. F.: *Quantile Regression.* Journal of Economic Perspectives, 15(4):143–156, 2001. https://www.aeaweb.org/articles?id=10.1257/jep.15.4.143.

[15] Koenker, R. e Machado, J. A. F.: *Goodness of Fit and Related Inference Processes for Quantile Regression.* Journal of the American Statistical Association, 94(448):1296–1310, 1999. https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473882.

[16] Koenker, R. W. e D'Orey, V.: *Algorithm AS 229: Computing Regression Quantiles.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 36(3):383–393, 1987. https://www.jstor.org/stable/2347802.

[17] Kotler, P.: *Administração de Marketing.* Prentice Hall, 10ª ed., 2000.

[18] Listen Data: *15 Types of Regression you Should Know*, 2019. https://www.listendata.com/2018/03/regression-analysis.html#Quantile-Regression, acessado em 07/04/2019.

[19] Maciel, M. C., Campelo, A. K. e Raposo, M. C. F.: *A dinâmica das Mudanças na Distribuição Salarial e no Retorno à Educação para Mulheres: uma aplicação de regressão quantílica.* Em *Proceedings of the 29th Brazilian Economics Meeting*, p. 14. ANPEC, 2001. https://EconPapers.repec.org/RePEc:anp:en2001:102.

[20] Microsoft: *Microsoft Excel para Office 365.* Microsoft Corporation, Santa Rosa, California, 2018. https://www.microsoft.com/pt-br/.

[21] Peck, D. C. M. E. A. e Vining, G. G.: *Introduction to Linear Regression Analysis.* Wiley, 5ª ed., 2012.

[22] Portnoy, S. e Koenker, R.: *The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators.* Statistical Science, 12(4):279–300, 1997. https://projecteuclid.org/euclid.ss/1030037960.

[23] Puiatti, G. A.: *Regressão Quantílica Não Linear para Descrição de Diferents Níveis de Acúmulo de Matéria Seca em Plantas de Alho.* Tese de Doutorado, 2018.

[24] R Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018. https://www.R-project.org/.

[25] R Datasets: *Swiss Fertility and Socioeconomic Indicators (1888) Data*, 2019. http://127.0.0.1:17855/library/datasets/html/swiss.html, acessado em 07/04/2019.

[26] Rodrigues, K. A. S.: *Diagnóstico em Regressão L_1.* Dissertação de Mestrado, 2019.

[27] Santos, B. R. dos: *Modelos de Regressão Quantílica*. Dissertação de Mestrado, 2012.

[28] Searle, S. R.: *Linear Models*. John Wiley Sons, 1ª ed., 1971.

[29] Silva, E. N. da: *Efeito do Número de Filhos na Distribuição Condicional da Renda Familiar: uma aplicação de variáveis instrumentais para estimar o efeito quantílico de um tratamento*. Dissertação de Mestrado, 2003.

[30] Silva, E. N. da e Silva Porto Júnior, S. da: *Sistema financeiro e crescimento econômico: uma aplicação de regressão quantílica*. Economia Aplicada, 10(3):425–442, 2006. http://dx.doi.org/10.1590/S1413-80502006000300007.