**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

Análise e aplicação de modelos lineares de regressão quantílica

Gabriela Maria Rodrigues

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba
2021**

Gabriela Maria Rodrigues
Licenciada em Matemática

Análise e aplicação de modelos lineares de regressão quantílica

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Profa. Dra. **TACIANA VILLELA SAVIAN**

Dissertação apresentada para obtenção do título de
Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2021

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Rodrigues, Gabriela Maria

Análise e aplicação de modelos lineares de regressão quantílica / Gabriela Maria Rodrigues. -- versão revisada de acordo com a resolução CoPGr 6018 de 2011. -- Piracicaba, 2021 .

77 p.

Dissertação (Mestrado) -- USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Regressão robusta 2. Modelo quadrático 3. Algoritmo *simplex* 4. Distribuição de Laplace assimétrica 5. Dados de crescimento 6. Bezzerro . I. Título.

DEDICATÓRIA

*A Deus,
Aos meus pais, Antônio e Graziella,
À minha irmã Manoella,
Dedico.*

AGRADECIMENTOS

A Deus, todo poderoso, pelo dom da vida, pela oportunidade de realizar esse sonho, por estar sempre presente, por me guiar, proteger, ajudar e dar forças.

Aos meus pais, pelo amor e apoio incondicional, por todo cuidado, dedicação, incentivo e por sempre acreditarem em mim. Vocês são a minha base.

À minha irmã, pelo amor, carinho e por estar sempre ao meu lado.

À Prof. Dra. Taciana Villela Savian, pelos ensinamentos, pela orientação, paciência, disponibilidade e amizade ao longo desta pesquisa.

A todos os professores do programa de pós-graduação em Estatística e Experimentação Agronômica (PPGEEA), que contribuíram para a minha formação acadêmica, pelos valiosos ensinamentos, a minha admiração e obrigada.

Aos colegas e amigos do departamento, pela ajuda nos momentos de dificuldade, conhecimentos compartilhados, momentos de descontração e pela amizade.

A todos os funcionários do PPGEEA, por todo auxílio e disponibilidade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro para realização desta pesquisa.

A todos que contribuíram direta ou indiretamente para a realização deste trabalho.

EPÍGRAFE

O futuro pertence àqueles que acreditam na beleza de seus sonhos.

Eleanor Roosevelt.

Tudo posso naquele que me fortalece.

Filipenses 4:13.

SUMÁRIO

Resumo	7
Abstract	8
Lista de Figuras	9
Lista de Tabelas	10
1 Introdução	11
2 Revisão de Literatura	15
2.1 Regressão linear	15
2.1.1 Método dos mínimos quadrados ordinários	16
2.2 Regressão Quantílica	18
2.2.1 Propriedades	24
2.2.2 Estimação dos parâmetros e inferência	26
2.2.3 Seleção de modelos e diagnósticos	36
3 Estudo de motivação	39
3.1 Material	39
3.2 Métodos	42
4 Resultados e Discussão	47
5 Considerações finais	59
5.0.1 Pesquisas futuras	59
Referências	61
Apêndices	67
5.1 Apêndice II - Principais comandos utilizados no <i>Software R</i>	69

RESUMO

Análise e aplicação de modelos lineares de regressão quantílica

A teoria clássica dos modelos de regressão é baseada na média da distribuição da variável resposta. Em comparação aos métodos convencionais, a regressão quantílica pode caracterizar toda a distribuição da variável dependente, por meio da análise de diferentes quantis. Baseada na minimização dos erros absolutos e pertencente a uma família de modelos robustos, essa metodologia não impõe nenhuma suposição distributiva sobre o erro do modelo, exceto exigir que este tenha um quantil condicional igual a zero, sendo eficiente na estimação dos parâmetros, mesmo em casos de assimetria ou heterogeneidade de variâncias. Os objetivos do presente trabalho foram analisar a metodologia dos modelos de regressão quantílica, bem como avaliar e motivar seu uso em dados de crescimento. Com essa finalidade, foi realizada uma aplicação da metodologia apresentada utilizando dados referentes a pesos em kg de 28 bezerros, por um período de 26 semanas após o nascimento. Esse conjunto apresenta heterogeneidade de variâncias ao longo do tempo e leve assimetria. A regressão quantílica mostrou-se adequada para descrever o peso como uma função do tempo, e para compreender como o tempo afeta sua variação e assimetria. Por meio das curvas de crescimento em diferentes quantis, foi possível classificar os bezerros em grupos de diferentes potenciais de crescimento e identificar animais com padrões de crescimentos distintos, como menores e maiores pesos corporais ao longo do tempo.

Palavras-chave: Regressão robusta, Modelo quadrático, Algoritmo *simplex*, Distribuição de Laplace assimétrica, Dados de crescimento, Bezerro.

ABSTRACT

Analysis and application of linear quantile regression models

The classical theory of regression models is based on the mean of the response variable distribution. In comparison to conventional methods, quantile regression can characterize the entire distribution of the dependent variable, through the analysis of different quantiles. Based on the minimization of absolute errors and belonging to a family of robust models, this methodology does not impose any distributive assumption about the model error, except requiring that it has a conditional quantile equal to zero, being efficient in estimating parameters, even in cases asymmetry or heterogeneity of variances. The objectives of the present work were to analyze the methodology of the quantile regression models, as well as to evaluate and motivate its use in growth data. For this purpose, an application of the presented methodology was carried out using data referring to weights in kg of 28 calves, for a period of 26 weeks after birth. This set presents heterogeneity of variances over time and slight asymmetry. Quantile regression proved to be adequate to describe weight as a function of time, and to understand how time affects its variation and asymmetry. Through growth curves in different quantiles, it was possible to classify calves into groups of different growth potentials and to identify animals with different growth patterns, such as lower and higher body weights over time.

Keywords: Robust regression, Quadratic model, Simplex algorithm, Asymmetric Laplace distribution, Growth data, Calf.

LISTA DE FIGURAS

2.1	Função de perda para diferentes valores de τ	20
2.2	Função de densidade de probabilidade da Laplace assimétrica para $\tau = 0,10$; $0,25$; $0,50$; $0,75$ e $0,90$, $\mu = 0$ e $\sigma = 1$	23
2.3	Ilustração da solução ótima de um problema de programação linear (Adaptado de LIMA JÚNIOR (2015)).	31
3.1	Perfis individuais de crescimento ao longo das semanas.	40
3.2	Perfis individuais de crescimento de cada animal ao longo das semanas.	40
3.3	Gráfico de caixa (<i>boxplot</i>) dos pesos dos bezerros ao longo das semanas.	41
4.1	Estimativas dos parâmetros e respectivos intervalos de confiança, considerando um nível de 5% de significância, ao longo dos quantis.	48
4.2	Gráfico de envelope simulado dos modelos de regressão quantílica para quantis igualmente espaçados no intervalo $[0,10; 0,90]$	49
4.3	Ajuste do modelo de regressão quantílica do quantil condicional de ordem $\tau = 0,45$ e do modelo de MQO.	52
4.4	Gráfico dos resíduos quantílicos vs valores preditos do modelo de regressão quantílica no quantil $\tau=0,45$ (a) e do modelo de MQO (b).	53
4.5	Histograma dos resíduos quantílicos do modelo de regressão quantílica no quantil $\tau=0,45$ (a) e do modelo de MQO (b).	53
4.6	Gráfico de envelope simulado dos resíduos do modelo de regressão quantílica no quantil $\tau=0,45$ (a) e do modelo de MQO (b).	54
4.7	Ajuste dos modelos de regressão quantílica nos quantis condicionais de ordem $\tau = 0,10$; $0,45$ e $0,90$	55
4.8	Perfis individuais de crescimento ao longo das semanas conforme a classificação.	57
4.9	Análise da silhueta para agrupamentos gerados considerando as curvas de crescimento da regressão quantílica.	58

LISTA DE TABELAS

3.1	Médias, desvios padrões, coeficientes de assimetria e coeficientes de variação (CV) dos pesos dos bezerros referentes a cada semana de observação.	41
4.1	Valor-p do teste anovar e do teste F^a , coeficiente de determinação ($R^1(\tau)$ e R^2 ajustado ^b) e critério de informação de Akaike (AIC) para os modelos de regressão quantílica em diferentes quantis e para o modelo de mínimos quadrados ordinários (MQO).	47
4.2	Critério de informação de Akaike (AIC), erro quadrático médio (EQM) e erro absoluto médio (EAM) para os modelos de regressão quantílica em diferentes quantis e para o modelo de mínimos quadrados ordinários (MQO).	50
4.3	Estimativas dos parâmetros para o modelo de regressão quantílica dos quantis condicionais $\tau = 0,30; 0,35; 0,40; 0,45$ e $0,50$ e para o modelo de MQO.	51
4.4	Estimativas dos parâmetros para o modelo de regressão quantílica dos quantis condicionais $\tau = 0,10; 0,45$ e $0,90$.	55
4.5	Distância euclidiana entre as taxas médias de variação das observações de cada bezerro e taxas médias de variação de cada uma das curvas ajustadas pelos três quantis e sua respectiva classificação.	56
4.6	Quantidade de indivíduos classificados em cada subgrupo (quantil $\tau = 0,10; 0,45$ e $0,90$) e seu resumo descritivo dos pesos na última semana.	57
4.7	Quantidade de indivíduos classificados em cada subgrupo (quantil $\tau = 0,10; 0,45$ e $0,90$) e seu resumo descritivo dos pesos na semana inicial.	57
4.8	<i>Silhouette coefficient</i> médio de cada um dos 3 grupos definidos.	58
A.1	Estimativas dos parâmetros para os modelos de regressão quantílica para diferentes quantis.	67
A.2	Estimativas dos parâmetros para o modelo de regressão quantílica do quantil condicional $\tau = 0,45$, considerando a variável dependente em meses.	68
A.3	Dados de crescimento de bezerros (pesos em kg) utilizados como estudo de motivação.	77

1 INTRODUÇÃO

A análise de regressão pode ser descrita como uma técnica estatística utilizada para investigar e modelar o relacionamento entre variáveis (MONTGOMERY *ET AL.*, 2012). A teoria clássica dos modelos de regressão é baseada na regressão na média da distribuição da variável dependente ou variável resposta, em função de uma ou mais variáveis independentes (ou covariáveis, preditoras). Entretanto, em muitas situações práticas, a média pode não ser a melhor medida para explicar a variável resposta, além disso, algumas suposições necessárias para a construção e análise de tais modelos, como normalidade e homogeneidade de variâncias, podem não ser satisfeitas, comprometendo suas propriedades e tornando-os inadequados, sendo necessário então, ir além dessa abordagem.

Na literatura estatística, podem ser encontradas várias propostas de modelos de regressão que excedem a forma de simplesmente modelar a média. Por exemplo, RIGBY e STASINOPOULOS (2005), propõem os modelos aditivos generalizados para locação, escala e forma, como uma abordagem mais ampla, modelando além da média condicional como função de outras variáveis, também outras quantidades de interesse, como parâmetros de escala e curtose. Tais modelos, podem fornecer uma descrição coerente e abrangente, entretanto, mantém a suposição de uma distribuição paramétrica para a variável resposta. Pode ser muito útil abandonar completamente pressupostos distributivos e ainda formular modelos que permitam descrever mais do que a média, como por exemplo, a regressão expectílica (*expectile regression*) (SCHNABEL e EILERS, 2009) e a regressão quantílica (RQ), introduzida por KOENKER e BASSETT JR (1978), que é o principal interesse deste trabalho.

Em comparação com a regressão média convencional, a regressão quantílica é baseada na minimização dos erros absolutos e pode caracterizar toda a distribuição condicional da variável resposta, fornecendo uma avaliação mais completa dos efeitos da covariável, por meio da análise de diferentes quantis. Pertencente a uma família de modelos robustos, a RQ não impõe nenhuma suposição distributiva sobre o erro do modelo, exceto exigir que este tenha um quantil condicional igual a zero, sendo então, uma alternativa ou complemento a teoria dos modelos de regressão usuais, eficiente na estimação dos parâmetros, mesmo em casos de assimetria ou heterogeneidade de variâncias. Além disso, seus modelos são invariantes a transformações monótonas e robustos na presença de pontos atípicos (*outliers*) na variável resposta.

Sua aplicação inicial ocorreu na econometria (BUCHINSKY (1998a); ZIETZ *ET AL.* (2008); DUFRENOT *ET AL.* (2010); FITZENBERGER *ET AL.* (2013)), e desde então, vêm emergindo com uma abordagem inferencial para análise estatística de modelos lineares a não lineares para diversas áreas de conhecimento. BEYERLEIN *ET AL.* (2011) utilizaram a RQ em análises GWAS (*Genome Wide Association Study*) na área de medicina humana, e enfatizaram as vantagens estatísticas e biológicas ao se estimar efeitos de marcadores em

diferentes quantis das distribuições dos fenótipos. Em melhoramento genético vegetal, por exemplo GOURDJI *ET AL.* (2013), foi utilizada para ajustar os rendimentos observados para mudanças nos locais e condições ambientais de ensaios ao longo do tempo, sendo que este, é considerado um passo necessário para avaliar verdadeiros ganhos genéticos em condições teoricamente constantes. Alguns trabalhos também propuseram a regressão quantílica para predição da produção de povoamentos de eucaliptos (CARVALHO (2013); FARIAS *ET AL.* (2018)). Recentemente, PUIATTI *ET AL.* (2018) estudaram o acúmulo de matéria seca em plantas de alho ao longo do tempo, e classificaram seus diferentes acessos de acordo com sua taxa de crescimento e peso assintótico, e SANTOS *ET AL.* (2018), propõe a utilização da RQ para estimação de valores genéticos genômicos para suínos, que possuem fenótipos com distribuições assimétricas.

O estudo das características de crescimento é de grande importância para a produção animal e vegetal. Por exemplo, na área de bovinocultura, a performance das características de crescimento pode ser determinante na lucratividade da produção (BARBOSA *ET AL.* (2017); LAUREANO *ET AL.* (2011)), desta forma, a fim de elevar índices zootécnicos, otimizar o sistema de produção e atender as necessidades econômicas da venda de bovinos em diversas idades, torna-se importante conhecer informações do potencial de ganho de peso (DE REZENDE *ET AL.*, 2014) e da precocidade reprodutiva (SILVEIRA *ET AL.*, 2014), em especial, o estudo da fase inicial é de grande importância na identificação da eficiência de crescimento.

Um aspecto importante que deve ser considerado no ajuste de modelos de regressão a dados de crescimento, é a possível presença de variâncias amostrais heterogêneas. Nesse tipo de ajuste, é comum ocorrer a heterocedasticidade entre as medidas, o que é natural, uma vez que, o crescimento é regulado por fatores extrínsecos ou ambientais e por fatores intrínsecos ou orgânicos. Portanto, conforme o fruto, a planta ou o animal se desenvolvem, a variação do seu tamanho e/ou peso se torna maior (FERNANDES *ET AL.*, 2014). Por exemplo, na identificação de animais com maior eficiência de crescimento é necessário ajustar modelos de regressão por meio de um método robusto à variabilidade dos dados individuais de peso-idade, e que seja capaz de produzir estimativas que representem bem a precocidade dos animais (DA SILVA *ET AL.*, 2006).

Nesse sentido, vários autores utilizaram a regressão quantílica: em estudos de crescimento de plantas, MUGGEO *ET AL.* (2013) estimam curvas de crescimento para alga marinha *Posidonia oceânica*; POLLICE *ET AL.* (2014) estudaram o crescimento de raízes de sorgo; SORRELL *ET AL.* (2012) analisaram o crescimento de três espécies de plantas de zonas úmidas em resposta a profundidade da água. No contexto de crescimento animal, NASCIMENTO *ET AL.* (2019) identifica suínos com diferentes taxas de crescimento e propõe uma classificação desses animais, com base em sua distância euclidiana geral entre cada peso observado e estimado a partir das curvas de crescimento de regressão de quantílica.

O presente trabalho, tem como objetivo apresentar e analisar os modelos de re-

gressão quantílica como uma alternativa mais robusta ou um complemento aos métodos convencionais. Além disso, tem como objetivos específicos: i) verificar as propriedades dos modelos de regressão quantílica considerando assimetria e heterogeneidade de variâncias residuais; ii) avaliar e motivar o uso dos modelos de regressão quantílica para dados de crescimento. Desta forma, é apresentada uma revisão de literatura e um estudo de motivação da metodologia apresentada, utilizando dados reais de crescimento de bezerros.

Esta dissertação está organizada da seguinte forma. A seção 2 apresenta uma breve revisão sobre o modelo clássico de regressão linear, introduz o método da regressão quantílica, apresenta algumas de suas propriedades e seu processo inferencial, que inclui o método de estimação dos parâmetros, intervalos de confiança e testes de hipóteses. Na seção 3 é apresentada uma breve descrição dos dados e dos procedimentos metodológicos para análise. A seção 4 apresenta uma discussão dos resultados empíricos. Por fim, na seção 5, são apresentadas as considerações finais do trabalho com um resumo dos principais resultados obtidos.

2 REVISÃO DE LITERATURA

2.1 Regressão linear

O termo regressão teve sua origem no século XIX, nos trabalhos de Sir Francis Galton, que aplicou conceitos estatísticos a problemas da hereditariedade. Galton estudou a relação entre a altura dos pais e dos filhos, e notou que a altura dos filhos tende à média da altura dos pais. Então, chamou esse evento de *regression toward the mean*, ou seja, existe uma tendência de os dados regredirem à média. Desde então, a análise de regressão vem sendo utilizada em diversas áreas de conhecimento, como: agricultura, economia, administração, biologia etc., com o objetivo principal, segundo MONTGOMERY ET AL. (2012), de investigar e modelar o relacionamento entre a variável dependente, ou variável resposta Y e uma ou mais variáveis independentes X , sendo a teoria clássica de tais modelos, baseada na média da distribuição da variável resposta.

Essa relação pode ser descrita por uma função linear ou não linear, em que o termo linear é utilizado para indicar que o modelo é linear em relação aos seus parâmetros $\beta_0, \beta_1, \dots, \beta_k$, ou seja, as derivadas parciais da função em relação aos parâmetros do modelo não dependem destes. Dessa forma, dado uma amostra de n observações de Y , admitindo que a relação entre Y e X_j , $j = 1, 2, \dots, k$ é linear, pode-se estabelecer uma regressão linear e representar essa relação segundo o modelo estatístico:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i, i = 1, 2, \dots, n$$

em que Y_i é a variável resposta, X_{ij} com $j = 1, \dots, k$, é a variável explicativa, $\beta_0, \beta_1, \dots, \beta_k$ são os parâmetros do modelo e ϵ_i é o erro aleatório.

Ao se estabelecer esse modelo, pressupõe-se que:

- (i) Os valores de X_{ij} são fixos (ou controlados).
- (ii) A média do erro é nula, isto é, $E(\epsilon_i) = 0$.
- (iii) O erro é homocedástico, isto é, para um dado valor de X , a variância do erro ϵ_i é sempre σ^2 : $\text{Var}(\epsilon_i) = E(\epsilon_i^2) - [E(\epsilon_i)]^2 = E(\epsilon_i^2) = \sigma^2$, o que implica em $\text{Var}(Y_i) = E[Y_i - E(Y_i)]^2 = E(\epsilon_i^2) = \sigma^2$.
- (iv) O erro de uma observação é independente do erro de outra observação, ou seja: $\text{Cov}(\epsilon_i, \epsilon'_i) = E(\epsilon_i \epsilon'_i) - E(\epsilon_i)E(\epsilon'_i) = E(\epsilon_i \epsilon'_i) = 0$, para $i \neq i'$.
- (v) Os erros têm distribuição normal.

Então, têm-se que $\epsilon_i \sim N(0, \sigma^2)$ e, portanto, $Y_i \sim N(\beta_0 + \dots + \beta_k X_{ik}, \sigma^2)$.

Escrevendo esse modelo matricialmente, obtêm-se:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \quad (2.1)$$

Que pode ser representado como:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} \text{ e } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

em que $\mathbf{Y}$ é o vetor da variável de observação de dimensão $n \times 1$; $\mathbf{X}$ é a matriz de delineamento, de dimensão $n \times p$ e de posto completo (com exceção de modelos de multicolinearidade) $p = k + 1$, sendo $p = k + 1$ o número de parâmetros; $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos com dimensão $p \times 1$ e $\boldsymbol{\epsilon}$ é o vetor de variáveis aleatórias não observáveis, de dimensão $n \times 1$.

As suposições anteriores sobre ϵ_i podem ser expressas em termos do modelo (2.1):

(i) $E(\boldsymbol{\epsilon}) = \mathbf{0}$ ou $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$;

(ii) $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$ ou $\text{Cov}(\mathbf{Y}) = \sigma^2 \mathbf{I}$, sendo que $\mathbf{I}$ é a matriz identidade de dimensão $n \times n$.

A suposição $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$ inclui as suposições: $\text{Var}(\epsilon_i) = \sigma^2$ e $\text{Cov}(\epsilon_i, \epsilon'_i) = 0$. Portanto, têm-se que: $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$.

2.1.1 Método dos mínimos quadrados ordinários

Para estimação dos parâmetros, o método dos mínimos quadrados ordinários (MQO), publicado em 1805 por Legendre, é o mais utilizado (SEARLE e GRUBER, 1971). Este, consiste em estimar os parâmetros de tal forma que a soma de quadrados dos resíduos (SQR) seja mínima. Considerando o modelo em sua forma matricial (2.1), têm-se: $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$.

A soma de quadrados dos erros é dada por:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^n \epsilon_i^2 = \boldsymbol{\epsilon}' \boldsymbol{\epsilon} \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y}' - \mathbf{X}'\boldsymbol{\beta}') (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \mathbf{X}'\boldsymbol{\beta}'\mathbf{Y} + \mathbf{X}'\boldsymbol{\beta}'\mathbf{X}\boldsymbol{\beta} \\ S(\boldsymbol{\beta}) &= \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \mathbf{X}'\boldsymbol{\beta}'\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Para obter a soma dos quadrados dos resíduos é necessário diferenciar $S(\boldsymbol{\beta})$ em relação a $\boldsymbol{\beta}$:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 - 2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \quad (2.2)$$

Igualando a expressão (2.2) a zero, temos o sistema de equações normais: $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$.

Como $\mathbf{X}$ tem posto coluna completo, $p = r(\mathbf{X})$, então, o sistema é consistente e tem solução única dada por:

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

Com a derivada de segunda ordem de $S(\beta)$, é possível verificar que $S(\beta)$ é mínima para β .

O teorema de Gauss-Markov afirma que, sob a hipótese de veracidade dos pressupostos do modelo, os estimadores obtidos pelo método de mínimos quadrados ordinários serão os melhores estimadores lineares não viesados (MELNV ou BLUE - *Best Linear Unbiased Estimator*), uma propriedade que constitui a justificativa teórica para esse método de estimativa amplamente difundido (HALLIN, 2014). A primeira afirmação diz que $E(\hat{\beta}) = \beta$, enquanto a segunda afirmação: $\text{Var}(\hat{\beta}) \leq \text{Var}(\hat{\beta}')$, qualquer que seja $\hat{\beta}'$. De acordo com RENCHER e SCHAALJE (2008), essas expressões são dadas por:

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' E(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta \\ &= \beta. \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \text{Var}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{Var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sigma^2\mathbf{I}. \end{aligned}$$

Note que, como $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ e $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, então $\hat{\beta}$ é combinação linear de distribuições normais, e temos:

$$\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2\mathbf{I})$$

Algumas limitações levaram à busca por outros métodos de estimação de parâmetros. Por exemplo, a homogeneidade de variâncias é um pressuposto central do modelo de regressão de mínimos quadrados, de acordo com LEWIS-BECK e LEWIS-BECK (2015), a violação desse pressuposto é preocupante, mesmo que as estimativas de MQO continuem a ser não-viesadas, afetam a confiabilidade dos testes de significância e intervalos de confiança. A normalidade também é uma suposição em relação aos resíduos que nem sempre

é satisfeita. Quando os erros estão distribuídos de uma forma assimétrica ou possuem uma cauda mais pesada que a da distribuição normal, então, segundo HAO e NAIMAN (2007) a performance deste método é ruim, podendo ocorrer viés no cálculo de p -valores e acarretando testes de hipóteses inválidos por exemplo.

Outra questão, é a influência que *outliers* exercem nas estimativas dos parâmetros do modelo. Isso faz com que seja necessário, sempre que se utiliza essa técnica, uma criteriosa avaliação de quanto cada ponto influencia no ajuste do modelo, o que pode se tornar bastante trabalhoso, uma vez que, tanto *outliers* na variável resposta quanto nas variáveis preditoras podem atrapalhar na identificação da verdadeira relação entre as variáveis de interesse (SANTOS, 2012). Para contornar essas limitações pode-se utilizar o método de minimização dos erros absolutos, que é a base dos modelos de regressão quantílica, que serão tratados nas próximas seções.

2.2 Regressão Quantílica

Segundo KOENKER e BASSETT JR (1978), a regressão quantílica é uma alternativa mais robusta ao método de estimação de mínimos quadrados. Os autores propõem a classe de modelos no contexto linear, baseada na minimização dos erros absolutos. Essa técnica surgiu em 1757, nos trabalhos ¹ de Boscovich, antes mesmo do surgimento do método dos mínimos quadrados. Boscovich acrescentou ao problema de minimização dos erros absolutos a restrição de que a soma de tais erros seja igual a zero. Inicialmente propôs estimar a mediana para minimizar essa distância absoluta, e essa estimativa foi denominada regressão L_1 . Porém, encontrou dificuldades para computar os valores dos estimadores propostos, e o desenvolvimento dessa técnica foi lento, somente com o avanço dos computadores e a implementação de programação linear em *softwares* estatísticos, principalmente, é que a utilização dessa técnica começou a crescer.

De acordo com KOENKER e BASSETT JR (1978), é possível estimar outros quantis além da mediana, através de uma ponderação na minimização dos erros, e dessa forma, caracterizar toda a distribuição condicional da variável dependente, dado um conjunto de variáveis explanatórias. Assim, a regressão L_1 passou a ser um caso particular da regressão quantílica, se tratando do quantil de ordem $\tau = 0,50$.

Diversamente aos modelos de regressão tradicionais, que estabelecem relações em torno da média condicional, as regressões quantílicas capturam as mudanças na distribuição e em seu formato, a partir de diferentes soluções nos distintos quantis, que podem ser interpretados como diferenças na resposta da variável dependente a mudanças nas covariáveis ao longo dos vários pontos (BUCHINSKY, 1998b). Verificando diferentes regressões quantílicas é possível explorar diferentes partes da distribuição condicional, como a cauda

¹O trabalho de Boscovich consistia em estimar a elipcidade da Terra, para mais detalhes desse trabalho sugere-se KOENKER (2005).

inferior (por exemplo, o quantil $\tau = 0,1$) ou a cauda superior (por exemplo, o quantil $\tau = 0,9$), quando o pesquisador necessita de informações sobre subpopulações específicas. Dessa forma, essa metodologia admite uma visão mais completa do conjunto de dados (HAO e NAIMAN (2007); OLIVEIRA e RIOS-NETO (2006)).

Além disso, essa metodologia possui boas propriedades que justificam sua aplicação, como por exemplo, ser robusta a presença de pontos atípicos, seus modelos são invariantes sob transformações monótonas e são flexíveis com dados incompletos.

Definição de quantis

Para melhor compreensão da regressão quantílica, nessa seção são apresentadas as definições de quantis (KOENKER e BASSETT JR (1978); KOENKER (2005) e SANTOS (2012)).

Uma variável aleatória Y qualquer, com valor real, pode ser caracterizada por sua função de distribuição acumulada:

$$F(y) = P(Y \leq y). \quad (2.3)$$

Utilizando a inversa da função de distribuição no ponto τ , para qualquer $0 < \tau < 1$:

$$F^{-1}(\tau) = \inf \{y : F(y) \geq \tau\}.$$

Dessa forma, definimos o τ -ésimo quantil da variável Y , e podemos enunciar, por exemplo, a mediana $F^{-1}(\frac{1}{2})$, o primeiro quartil $F^{-1}(\frac{1}{4})$, ou o terceiro quartil $F^{-1}(\frac{3}{4})$.

Uma segunda definição é apresentada pelos autores, sendo trivial para a generalização dos quantis para uma classe de modelos. Para isso, adota-se a função de perda (ou *function check*), que atribui pesos diferentes a resíduos positivos e negativos e é definida da seguinte forma:

$$\rho_{\tau}(\varepsilon) = \varepsilon(\tau - I(\varepsilon < 0)) = \begin{cases} \tau \varepsilon_i & \text{se } \varepsilon_i \geq 0 \\ (\tau - 1) \varepsilon_i & \text{se } \varepsilon_i < 0 \end{cases} \quad 0 < \tau < 1 \quad (2.4)$$

em que I é a função indicadora, ε denota o resíduo com quantil de ordem τ igual a 0 e ε_i seus respectivos termos, com $i = 1, 2, \dots, n$.

A função de perda pode ser visualizada na figura 2.1.

O quantil τ é definido como o valor de Y que divide os dados em duas proporções. A função *check* multiplica os resíduos, e faz com que estes, sejam tratados de forma assimétrica, o que resultará em pelo menos $100(\tau)\%$ dos valores de Y acima do hiperplano de regressão quantílica e pelo menos $100(\tau - 1)\%$ abaixo.

Figura 2.1. Função de perda para diferentes valores de τ .

Então, considerando um simples problema da teoria de decisão³, em que o interesse é encontrar $\hat{y}$, um previsor de Y , com função de distribuição F (2.3), de forma que minimize a perda esperada, têm-se:

$$E\rho_{\tau}(Y - \hat{y}) = (\tau - 1) \int_{-\infty}^{\hat{y}} (y - \hat{y})dF(y) + \tau \int_{\hat{y}}^{\infty} (y - \hat{y})dF(y).$$

Diferenciando em relação a $\hat{y}$ e igualando a 0 têm-se:

$$\begin{aligned} \frac{d E\rho_{\tau}(Y - \hat{y})}{d\hat{y}} &= (\tau - 1) \int_{-\infty}^{\hat{y}} \frac{d[ydF(y)]}{d\hat{y}} - \frac{d[\hat{y}dF(y)]}{d\hat{y}} + \tau \int_{\hat{y}}^{\infty} \frac{d[ydF(y)]}{d\hat{y}} - \frac{d[\hat{y}dF(y)]}{d\hat{y}} \\ &= (\tau - 1) \int_{-\infty}^{\hat{y}} -dF(y) + \tau \int_{\hat{y}}^{\infty} -dF(y) \\ &= (1 - \tau) \int_{-\infty}^{\hat{y}} dF(y) - \tau \int_{\hat{y}}^{\infty} dF(y) \\ &= (1 - \tau) \int_{-\infty}^{\hat{y}} dF(y) - \tau \left(1 - \int_{-\infty}^{\hat{y}} dF(y)\right) \\ &= (1 - \tau)F(\hat{y}) - \tau(1 - F(\hat{y})) \\ &= F(\hat{y}) - \tau F(\hat{y}) - \tau + \tau F(\hat{y}) \\ &= F(\hat{y}) - \tau \\ &= F(\hat{y}) - \tau = 0 \\ &= F(\hat{y}) = \tau \end{aligned}$$

³Este é um exercício padrão nos textos da teoria da decisão (por exemplo, FERGUSON (2014)). Outra referência seria FOX ET AL. (1964), que estudaram a admissibilidade do estimador quantil sob essa função de perda.

Como $F(2.3)$ é monótona, qualquer elemento de $\{y : F(y) = \tau\}$ minimiza a perda esperada, ou seja, $\hat{y} = F^{-1}(\tau)$ minimiza a perda esperada para a função de perda definida em (2.4).

Quando F é substituída pela função de distribuição empírica⁴:

$$F_n(y) = n^{-1} \sum_{i=1}^n I(Y_i \leq y), \quad (2.5)$$

em que $I\{y_i \leq y\}$ é a função indicadora do evento $\{y_i \leq y\}$ para $y_{\min} \leq y \leq y_{\max}$, de modo análogo ao anterior, podemos escolher $\hat{y}$ para minimizar a perda esperada:

$$\int \rho_\tau(y - \hat{y}) dF_n(y) = n^{-1} \sum_{i=1}^n \rho_\tau(y_i - \hat{y}). \quad (2.6)$$

Dessa forma, o resultado é o τ -ésimo quantil amostral.

Essa solução pode parecer inerentemente ligada à noção de uma ordem das observações da amostra, mas, o que acontece na verdade, de acordo com KOENKER (2005), foi substituído a classificação pela otimização.

Função de regressão quantílica

Seja uma amostra aleatória de tamanho n , $\mathbf{y} = (y_i, \dots, y_n)'$ a variável resposta (contínua) e $\mathbf{x}$ o vetor p -dimensional de covariáveis. A função que relaciona o τ -ésimo quantil da variável resposta e as covariáveis é chamada função de regressão quantílica, denotada como $Q_\tau(\cdot)$, sob suposição de relação linear, para o τ -ésimo quantil pode ser expressa por:

$$Q_\tau(Y|x) = \mathbf{x}'\boldsymbol{\beta}(\tau) \quad (2.7)$$

sendo $\boldsymbol{\beta}(\tau)$ o vetor p -dimensional de parâmetros associados ao vetor de covariáveis para o τ -ésimo quantil fixo.

Dessa forma, o modelo de regressão quantílica pode ser representado por:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}(\tau) + \varepsilon_i \quad (2.8)$$

em que ε_i é o erro aleatório, com quantil de ordem τ igual a zero, ou seja, possui densidade $f_\tau(\cdot)$, restrita unicamente por acumular probabilidade τ até zero:

$$\int_{-\infty}^0 f_\tau \varepsilon_i d\varepsilon_i = \tau.$$

⁴A função de distribuição acumulada empírica é uma distribuição discreta que atribui peso igual a cada ponto amostral, ou seja, atribui probabilidade $\frac{1}{n}$ a cada uma das n observações originais, $F_n(y)$ corresponde à proporção de valores menores ou iguais a y . Se o tamanho da amostra for suficientemente grande, a lei dos grandes números nos diz que F_n deve se aproximar muito bem de $F(2.3)$, ou seja, F_n desempenha o papel de $F(2.3)$

Porém, além de possuir apenas essa restrição, para o contexto da regressão quantílica, $f_\tau(\cdot)$, não necessita ser determinada de forma explícita e o vetor de parâmetros $\hat{\beta}(\tau)$ pode ser obtido como a solução da minimização:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i' \beta). \quad (2.9)$$

em que $\rho_\tau = \varepsilon(\tau - I(\varepsilon < 0))$ é a função de perda definida anteriormente.

Distribuição de Laplace assimétrica

Na regressão clássica, os estimadores do método de mínimos quadrados coincidem com os estimadores de máxima verossimilhança quando os erros seguem distribuição normal, um resultado bastante conhecido. De forma similar ocorre na regressão quantílica, os estimadores de mínimos desvios absolutos coincidem com os estimadores de máxima verossimilhança, quando o erro do modelo segue distribuição de Laplace assimétrica (DLA). Convenientemente então, esta distribuição passou a ser utilizada como a distribuição do erro do modelo de regressão quantílica, permitindo uma abordagem inferencial mais ampla e estruturas de modelos mais complexas, tornando a RQ uma alternativa muito promissora. Por exemplo, no contexto Bayesiano (YU e MOYEED (2001); YU e ZHANG (2005); YUAN e YIN (2010); HINOSTROZA (2017); SANTOS e BOLFARINE (2018)), na utilização da regressão quantílica com regressão Lasso (LI ET AL., 2010); no ajuste de modelos de mistura (REICH ET AL., 2010) e na modelagem mista (GERACI e BOTTAI (2007); GERACI e BOTTAI (2014) e MORALES (2015)).

Seguindo a formulação de YU e MOYEED (2001), uma variável aleatória Y tem distribuição Laplace assimétrica se sua função de densidade de probabilidade (f.d.p.) for dada por:

$$f(y|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\},$$

em que $-\infty < \mu < \infty$ é o parâmetro de locação, $\sigma > 0$ é o parâmetro de escala, $0 < \tau < 1$ é o parâmetro de assimetria e ρ_τ (2.4) é a função de perda, definida anteriormente.

Se os parâmetros de locação e escala são zero e um, respectivamente, então será chamada de f.d.p. Laplace assimétrica padrão, de forma similar ao que ocorre na distribuição normal. O parâmetro τ é fixado conforme o quantil de interesse, a figura 2.2 ilustra as mudanças ocasionadas por diferentes valores de τ de uma distribuição de Laplace assimétrica padrão. Por exemplo, para $\tau = 0,90$ a maior parte da massa é concentrada em torno da cauda direita, e para $\tau = 0,50$ ambas as caudas têm massa igual, e nesse caso, se reduz a função de densidade de uma distribuição de Laplace simétrica, ou também chamada de exponencial dupla.

Figura 2.2. Função de densidade de probabilidade da Laplace assimétrica para $\tau = 0,10; 0,25; 0,50; 0,75$ e $0,90$, $\mu = 0$ e $\sigma = 1$.

A média e a variância, são dadas por:

$$E(Y) = \mu + \frac{\sigma(1-2\tau)}{\tau(1-\tau)} \text{ e } \text{Var}(Y) = \frac{\sigma^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2}.$$

Dessa forma, considerando o modelo linear de regressão quantílica:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta}(\tau) + \varepsilon_i,$$

já descrito em 2.8, mas agora, ε_i tem distribuição Laplace assimétrica, com quantil de ordem τ igual a zero, ou seja, possui densidade $f_\tau(\cdot)$, restrita unicamente por acumular probabilidade τ até zero.

A função de verossimilhança resultante para n observações será dada por:

$$L(\boldsymbol{\beta}, \sigma) = \frac{\tau^n (1-\tau)^n}{\sigma^n} \exp \left\{ -\frac{1}{\sigma} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i' \boldsymbol{\beta}) \right\}$$

Então:

$$L(\boldsymbol{\beta}) \propto \exp \left\{ -\sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i' \boldsymbol{\beta}) \right\}$$

É fácil notar que, como o expoente é negativo, maximizar a função de verossimilhança dessa densidade com relação a $\boldsymbol{\beta}$ é equivalente a minimizar a soma dos erros absolutos, conforme mencionado anteriormente.

2.2.1 Propriedades

Em teoria, as curvas de quantis condicionais não podem se cruzar, mas na prática isso pode ocorrer, especialmente para pequenos conjuntos de dados. Em muitos casos, isso é apenas um incômodo visual, mas também pode comprometer uma análise posterior, por exemplo, ao estudar distribuições condicionais em valores específicos da variável independente. Recentemente, esse problema tem recebido atenção considerável, sendo possível encontrar na literatura várias abordagens sob o tema. Entre elas, propostas utilizando monotonização natural (CHERNOZHUKOV *ET AL.*, 2010); a estimativa simultânea de curvas de quantis utilizando folhas de quantis (SCHNABEL e EILERS, 2013); técnicas não paramétricas (DETTE e VOLGUSHEV (2008); SHIM *ET AL.* (2009); TAKEUCHI *ET AL.* (2006)); bem como restrições que obrigam o não cruzamento (KOENKER e NG, 2005). Outra regressão alternativa, são os modelos aditivos generalizados para locação, escala e forma, em inglês *Generalized Additive Models for Location, Scale and Shape - GAMLSS* (RIGBY e STASINOPOULOS, 2005), que excedem a forma de simplesmente modelar a média e não tem o problema de cruzar quantis.

De acordo com DAVINO *ET AL.* (2013), o número de quantis distintos que é possível se obter, está relacionado com o número de observações (ou número de covariáveis) e o tamanho da amostra.

Quanto a interpretação dos parâmetros $\beta(\tau)$, similar ao modelo linear, o coeficiente $\beta_j(\tau), j = 1, 2, 3, \dots, p$ pode ser interpretado como a taxa de variação no τ -ésimo quantil da variável resposta Y ao variar-se uma unidade o valor da j -ésima covariável, mantendo-se os valores das demais variáveis fixos. Isto é, seja o τ -ésimo quantil condicional definido em 2.7, o efeito marginal de uma determinada covariável pode ser obtido por meio de suas respectivas derivadas parciais:

$$\frac{\partial Q_\tau(Y|x)}{\partial x_j}.$$

RASTEIRO (2017), observa que os quantis não necessariamente têm a mesma ordenação de Y_i , já que o quantil é condicional às covariáveis (x_i) , por exemplo, se $y_2 > y_1$, seus valores preditos para um determinado quantil, podem ser $\hat{y}_2 < \hat{y}_1$.

Equivariância

Os estimadores de regressão quantílica possuem várias propriedades, que juntas recebem o título de equivariância. Estas, auxiliam a análise dos dados e facilitam os procedimentos computacionais. Quando os dados são alterados de maneiras totalmente previsíveis, espera-se que as estimativas de regressão também mudem de uma maneira que deixe invariante a interpretação dos resultados, isto é, espera-se que essas mudanças não tenham efeito fundamental nas estimativas.

Sendo $\hat{\beta}(\tau)$, o vetor de parâmetros estimados, que depende de $\mathbf{y}$ e da matriz de covariáveis $\mathbf{X}$ observados na amostra, e portanto, descrito como $\hat{\beta}(\tau; \mathbf{y}, \mathbf{X})$, quatro propriedades básicas de equivariância de $\hat{\beta}(\tau; \mathbf{y}, \mathbf{X})$, baseadas em observações $(\mathbf{y}, \mathbf{X})$, são denotadas no teorema a seguir:

Teorema 2.2.1 (KOENKER e BASSETT JR (1978)) *Seja $\mathbf{A}$ matriz não singular de dimensão p , $\gamma \in \mathbb{R}^p$, e $a > 0$. Então, para qualquer $\tau \in [0, 1]$:*

- (i) $\hat{\beta}(\tau; a\mathbf{y}, \mathbf{X}) = a\hat{\beta}(\tau; \mathbf{y}, \mathbf{X})$.
- (ii) $\hat{\beta}(\tau; -a\mathbf{y}, \mathbf{X}) = -a\hat{\beta}(1 - \tau; \mathbf{y}, \mathbf{X})$.
- (iii) $\hat{\beta}(\tau; \mathbf{y} + \mathbf{X}\gamma, \mathbf{X}) = \hat{\beta}(\tau; \mathbf{y}, \mathbf{X}) + \gamma$.
- (iv) $\hat{\beta}(\tau; \mathbf{y} + \mathbf{X}\mathbf{A}, \mathbf{X}) = \mathbf{A}^{-1}\hat{\beta}(\tau; \mathbf{y}, \mathbf{X})$.

As propriedades (i) e (ii) tratam de equivariância de escala, enquanto a propriedade (iii) aborda o contexto conhecido como equivariância de regressão e a (iv) é chamada de equivariância da reparametrização da matriz de planejamento.

Por exemplo, suponha um modelo referente a temperatura de um líquido, y , mas a escala de medições é alterada de Fahrenheit ($^{\circ}\text{F}$) para Celsius ($^{\circ}\text{C}$). Sabendo que $^{\circ}\text{C} = (^{\circ}\text{F} - 32)\frac{5}{9}$ e presumindo que X contenha um intercepto β_0 , sendo então a primeira coluna de X de 1_n , o efeito de nossa mudança na escala de temperatura é simplesmente mudar $\hat{\beta}(\tau; y, X)$ para $\frac{5}{9}(\hat{\beta}(\tau; y, X) - 32\beta_0)$. A primeira coordenada de $\hat{\beta}$ seria deslocada em 32 e todas as coordenadas seriam então redimensionadas pelo fator $\frac{5}{9}$.

Equivariância a transformações monótonas

Para descrever tal propriedade, considera-se primeiramente a regressão linear clássica: quando as suposições do modelo não são verificadas, como a normalidade por exemplo, uma alternativa sugerida por BOX e COX (1964), seria transformar a variável resposta para tentar satisfazer tais suposições. Porém, esse procedimento pode comprometer a interpretação dos parâmetros, fato que pode ser demonstrado pela desigualdade de Jensen

$$E(g(y)) \neq g(E(y)).$$

Uma transformação usada frequentemente é a da função logaritmo. Por exemplo, uma variável Y que não possui distribuição normal, é transformada na variável $W = \log Y$. Conforme exemplificado em SANTOS (2012), é ajustado o seguinte modelo $W_i = \alpha + \beta x_i + \varepsilon_i$ com as suposições usuais e então, temos que $E(W|x) = \alpha + \beta x_i$. Porém, nesse caso $E(W|x) = E(\log Y|x) \neq \log E(Y|x)$, logo, não temos o valor esperado na nossa variável Y inicial.

Por outro lado, a regressão quantílica possui a propriedade da equivariância a transformações monótonas, que pode ser considerada crítica para a compreensão de todo

o seu potencial. KOENKER (2005), descreve como: seja $h(\cdot)$ uma função não decrescente no conjunto $\mathbb{R}$, então, para qualquer variável aleatória Y , e sendo $Q_Y(\tau)$ o quantil de ordem τ dessa variável Y :

$$Q_{h(y)}(\tau) = h(Q_y(\tau))$$

Resultado que segue imediatamente, pelo fato elementar que, para qualquer função h monótona,

$$P(Y \leq y) = P(h(Y) \leq h(y))$$

Essa propriedade pode ser exemplificada de forma simples: sob transformações crescentes, se m é a mediana de uma variável aleatória Y , então, 2^m representa a mediana de 2^y .

Robustez

Conforme mencionado anteriormente, os modelos de regressão quantílica são robustos a presença de *outliers* na variável resposta, propriedade importante que é denominada de robustez. Esse recurso foi formalizado no teorema:

Teorema 2.2.2 (KOENKER e BASSETT JR (1978)) *Seja D uma matriz diagonal com elementos não negativos d_i , com $i=1, \dots, n$, então*

$$\hat{\beta}(\tau; y, X) = \hat{\beta}(\tau; X\hat{\beta}(\tau; y, X) + D\hat{u}, X)$$

em que $\hat{u} = y - X\hat{\beta}(\tau; y, X)$.

KOENKER 2005, explica a robustez do estimador de regressão quantílica de forma mais simplificada: considerando um conjunto de dados, e um ajuste para um quantil condicional de ordem τ , que ele chama de plano, se tomarmos um ponto qualquer y_i , acima desse plano e adicionarmos, ou subtrairmos unidades a ele, de modo que não ultrapasse o plano, a sua contribuição é independente para o ajuste, desde que $(y_i - x'_i \hat{\beta}(\tau))$ não mude. Em outras palavras, pode-se alterar qualquer uma das observações da variável resposta y , para cima e para baixo, desde que não se altere o sinal dos resíduos, sem que a solução inicial se altere.

2.2.2 Estimação dos parâmetros e inferência

Conforme visto na seção 2.2, o vetor de parâmetros $\hat{\beta}(\tau)$, pode ser obtido como a solução da minimização dos erros absolutos ponderados. Esse problema não é resolvido facilmente. Por envolver uma função indicadora, não assume forma analítica fechada, mas pode ser reformulado, segundo KOENKER (2005), para uma representação na forma padrão de um problema de programação linear. Dessa forma, nessa seção, inicialmente serão delineados os principais conceitos que norteiam a programação linear, e que serão úteis para reformulação do problema de RQ.

Problema de programação linear

De acordo com KOLMAN e BECK (1995), um problema de programação linear (PPL) é baseado em três conceitos: função objetivo, variáveis de decisão e restrições. A função objetivo é uma função linear cujo objetivo é otimizá-la, isto é, maximizá-la ou minimizá-la, e é definida como:

$$f(x) = \sum_{i=1}^n c_i x_i = c_1 x_1 + c_2 x_2 + \dots + c_n x_n$$

sendo $c_1, c_2, \dots, c_n$ números reais e $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$.

As variáveis de decisão constituem as incógnitas a serem determinadas pela resolução do modelo e as restrições são condições que limitam os valores que podem ser atribuídos às variáveis.

Um PPL está na forma canônica se suas restrições são desigualdades e na forma padrão quando suas restrições são igualdades. A forma canônica pode ser transformada em sua forma padrão acrescentando variáveis de folgas em suas desigualdades.

A forma canônica, possui a formulação:

$$\begin{aligned} \text{Maximizar } f(x_1, \dots, x_n) &= c_1 x_1 + c_2 x_2 + \dots + c_n x_n \\ \text{Sujeito a: } a_{11}x_1 + \dots + a_{1n}x_n &\leq b_1 \\ a_{21}x_1 + \dots + a_{2n}x_n &\leq b_2 \\ a_{31}x_1 + \dots + a_{3n}x_n &\leq b_3 \\ &\vdots \\ a_{m1}x_1 + \dots + a_{mn}x_n &\leq b_m \end{aligned}$$

Logo, a forma padrão:

$$\begin{aligned} \text{Maximizar } f(x_1, \dots, x_n) &= c_1 x_1 + c_2 x_2 + \dots + c_n x_n \\ \text{Sujeito a: } a_{11}x_1 + \dots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \dots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + \dots + a_{3n}x_n &= b_3 \\ &\vdots \\ a_{m1}x_1 + \dots + a_{mn}x_n &= b_m \end{aligned}$$

em que $k = m + n$; $x_1, x_2, \dots, x_n$ são as variáveis inicialmente observadas no PPL e $x_{n+1}, x_{n+2}, \dots, x_k$ são denominadas variáveis de sobra (ou de folga) e são utilizadas com

o objetivo de transformar cada uma das desigualdades presentes na forma canônica em uma igualdade.

Além disso, o sistema da forma padrão tem as seguintes condições:

- todas as variáveis são não-negativas;
- todos os b_i são não-negativos.

Qualquer variável x_j , não restringida pela condição de não negatividade, pode ser substituída por um par de variáveis não negativas $x_{j'} \geq 0$ e $x_{j''} \geq 0$, fazendo: $x_j = x_{j'} - x_{j''}$, como ocorre com o problema de minimização dos erros absolutos.

A partir desses conceitos iniciais de PPL é possível compreender então, como ocorre a estimação dos parâmetros de regressão quantílica.

Reformulação da minimização dos erros absolutos

Inicialmente, considerando o seguinte modelo de regressão linear quantílica:

$$y_i = \mathbf{x}_i' \boldsymbol{\beta}(\tau) + \varepsilon_i$$

sendo $\mathbf{y} = (y_1, \dots, y_n)$ a variável resposta (contínua) de uma amostra aleatória de tamanho n ; $\mathbf{x}$ o vetor p -dimensional de covariáveis; $\boldsymbol{\beta}(\tau)$ o vetor p -dimensional de parâmetros associados ao vetor de covariáveis $\mathbf{x}$ para o τ -ésimo quantil fixo e ε_i o erro aleatório.

Para estimar $\hat{\boldsymbol{\beta}}(\tau)$ devemos minimizar a soma de erros absolutos ponderados, conforme descrito anteriormente:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}_i' \boldsymbol{\beta}).$$

Que pode ser reescrito como:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left[\sum_{i: y_i \geq \mathbf{x}_i' \boldsymbol{\beta}} \tau(y_i - \mathbf{x}_i' \boldsymbol{\beta}) + \sum_{i: y_i < \mathbf{x}_i' \boldsymbol{\beta}} (1 - \tau)(y_i - \mathbf{x}_i' \boldsymbol{\beta}) \right]. \quad (2.10)$$

Esta função (2.10), é considerada a função objetivo, sob a seguinte restrição:

$$y_i = \sum_{j=1}^p x_{ij} \beta_j(\tau) + \varepsilon_i$$

com β_j e ε_i sem restrição de sinal.

Para a representação na forma padrão de um problema de programação linear, é necessário introduzir $2n$ variáveis artificiais ($\mathbf{u}, \mathbf{v}$: $1, \dots, n$) para representar as partes positivas e negativas do vetor de resíduos, e é necessário alterações em $\beta_j(\tau)$, para que todas as variáveis sejam positivas:

$$\beta_j(\tau) = \beta_j(\tau)^{(1)} - \beta_j(\tau)^{(2)} \quad \text{e} \quad \varepsilon_i = u_i - v_i$$

sendo $\beta_j(\tau)^{(1)} \geq 0$, $\beta_j(\tau)^{(2)} \geq 0$, $u_i \geq 0$ e $v_i \geq 0$, ou seja:

$$u_i \begin{cases} y_i - x'_i \hat{\beta} & \text{se } y_i - x'_i \hat{\beta} > 0 \\ 0, & \text{caso contrário;} \end{cases} \quad \nu_i \begin{cases} y_i - x'_i \hat{\beta} & \text{se } y_i - x'_i \hat{\beta} < 0 \\ 0, & \text{caso contrário;} \end{cases}$$

Então, pode ser descrito um PPL de forma padrão, como:

$$\text{Função objetivo: } \min_{\beta \in \mathbb{R}^p} \left[\sum_{i: y_i \geq x'_i \beta} \tau(u_i) + \sum_{i: y_i < x'_i \beta} (1 - \tau)(v_i) \right]$$

$$\begin{aligned} \text{Restrições: } & y_i = \sum_{j=1}^p x_{ij}(\beta_j(\tau)^{(1)} - \beta_j(\tau)^{(2)}) + (u_i - v_i) \\ & \text{com } \beta_j(\tau)^{(1)} \geq 0, \beta_j(\tau)^{(2)} \geq 0, u_i \geq 0 \text{ e } v_i \geq 0. \end{aligned}$$

Matricialmente o modelo pode ser escrito como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}(\boldsymbol{\tau}) + \boldsymbol{\varepsilon} \quad (2.11)$$

em que $\mathbf{Y}$ é um vetor de dimensão $n \times 1$ de observações; $\mathbf{X}$ é a matriz de delineamento de constantes conhecidas $n \times p$; $\boldsymbol{\beta}(\boldsymbol{\tau})$ é o vetor de dimensão $p \times 1$ de parâmetros desconhecidos, e $\boldsymbol{\varepsilon}$ é o vetor dos erros.

Dessa forma, o problema de programação linear é equivalente a:

$$\min_{(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\nu}) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}} \left\{ \tau \mathbf{1}'_n \mathbf{u} + (1 - \tau) \mathbf{1}'_n \boldsymbol{\nu} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \boldsymbol{\nu} = \mathbf{Y} \right\}$$

em que $\mathbf{1}'_n$ representa um vetor de dimensão $1 \times n$ de valores iguais a 1; $\mathbf{u}$ e $\boldsymbol{\nu}$ são vetores de dimensão $n \times 1$, já definidos.

É importante mencionar que, para todo modelo de programação linear denominado primal, existe um outro modelo chamado dual. Ambos estão relacionados, e são compostos pelos mesmos coeficientes, mas com objetivos diferentes (por exemplo: se um tem o objetivo de maximizar, o outro terá de minimizar). Dessa forma existem soluções compatíveis para os dois, desde que a matriz X das covariáveis possua posto completo, e o teorema do equilíbrio em programação linear garante que pelo menos uma dessas soluções é ótima.

Então, para obter essa solução, podemos citar o algoritmo *simplex*, inicialmente proposto para a regressão L_1 por BARRODALE e ROBERTS (1973), e posteriormente adaptado por KOENKER e D'OREY (1987) para o problema de minimização de desvios absolutos dos demais quantis, o algoritmo ponto interior sugerido por PORTNOY ET AL. (1997) e o algoritmo *smoothing* por CHEN (2007).

Segundo CHEN (2007), ambos tem vantagens e desvantagens, não sendo possível classificar um dominante. O algoritmo *simplex* é mais eficiente e preciso, sempre encontra solução para vários tipos de dados, especialmente aqueles com presença de outliers e pontos de alavancagem, mas, ainda assim é razoável para conjuntos de dados com até 5000

observações e 50 variáveis, enquanto que a técnica do ponto interior, tem performance superior na presença de bancos de dados com muitas observações e pequeno número de covariáveis, mas também apresenta deficiência, já que pode fornecer apenas as soluções aproximadas do problema original, não sendo tão preciso quanto o algoritmo *simplex*. O algoritmo *smoothing* pode apresentar falhas com a presença de pontos discrepantes, mas ganha em velocidade com grande número de covariáveis e também em precisão. Ambos implementados nos principais aplicativos estatísticos, mas a principal referência é o pacote *quantreg* (KOENKER, 2011) no *software* R (R DEVELOPMENT CORE TEAM, 2020) em que podem ser utilizados nas funções `rq` ou `rq.fit`. Pelos motivos mencionados e por este trabalho não apresentar banco de dados com muitas observações, o algoritmo *simplex* foi preferido e será apresentado na próxima seção.

Algoritmo *simplex*

Publicado em 1947 por George B. Dantzing, o *simplex* é um algoritmo desenvolvido para resolver problemas de programação linear, que associa o ponto ótimo da função objetivo a um dos pontos pertencentes ao polígono construído a partir das desigualdades determinadas em sua forma canônica (GOLDBARG e LUNA (2005), LIMA JÚNIOR (2015)).

Esse método, tem seu funcionamento justificado pelo significado geométrico associado ao gradiente da função que mostra o sentido de seu maior crescimento. Por exemplo, o sentido de maior crescimento de $f(x, y) = 3x + 4y$ é $\nabla f = (3, 4)$ e ocorre de forma mais efetiva na direção Oy do que na direção Ox uma vez que, para todo $(a, b) \in \mathbb{R}^n$ tem-se:

$$\frac{\partial f}{\partial y}(a, b) > \frac{\partial f}{\partial x}(a, b)$$

Na forma padrão, a solução do sistema é obtida a partir de escalonamento combinado com testes de otimalidade. No desenvolvimento do processo, as soluções intermediárias são obtidas e são classificadas em básicas e não-básicas, conforme algumas de suas incógnitas tenham valores igualados a zero.

A solução é ótima quando mais nenhuma variável colabora para seu decréscimo (para o caso de RQ), fato este que confirma o final da execução do *simplex* e o mínimo para f .

Em BARRODALE e ROBERTS (1973) e KOENKER e D'OREY (1987), encontra-se detalhes do método, bem como sua implementação para a regressão quantílica.

A solução ótima pode ser representada geometricamente, para problemas de programação linear de modelos que possuam duas variáveis (x_1, x_2 , por exemplo) (LUCHE e SILVA (2016)). Representando os pontos (x_1, x_2) que satisfazem todas as restrições do modelo e formam a região viável, que também é chamada de espaço de soluções factíveis (soluções viáveis) (figura 2.3). A condição de não-negatividade das variáveis para um PPL, implica que os pares (x_1, x_2) de soluções viáveis estão no 1º quadrante dos eixos cartesianos considerados, sendo cada um dos eixos correspondentes a uma variável.

Após todas as restrições incluídas, é fácil identificar o semiplano que contém os pontos de interesse.

Figura 2.3. Ilustração da solução ótima de um problema de programação linear (Adaptado de LIMA JÚNIOR (2015)).

De acordo com LUCHE e SILVA (2016) três tipos de pontos viáveis podem ser identificados:

- **Pontos interiores:** sendo que, nenhum deles determinará a solução ótima;
- **De fronteira:** pode determinar a solução ótima caso alguma equação de restrição for proporcional a equação da função;
- **Vértices:** são os extremos da região viável, por isso, a solução ótima sempre será encontrada em um dos vértices ou em mais de um quando existirem múltiplas soluções ótimas, o que ocorre quando alguma equação de restrição é proporcional a equação da função.

Intervalos de confiança

Após o ajuste do modelo, deve-se verificar se este é adequado para descrever o fenômeno. Para isso, apresenta-se em seguida, a construção de intervalos de confiança e testes de hipóteses para os parâmetros do modelo de regressão quantílica.

No contexto de inferência estatística, para a construção de intervalos de confiança, faz-se necessário a suposição de que uma amostra aleatória, $Y = (Y_1, \dots, Y_n)$, é proveniente de uma população com função de probabilidade ou densidade f e função de distribuição acumulada F . Um método muito utilizado é o da quantidade pivotal. Entretanto, na regressão quantílica, essa suposição nem sempre é satisfeita, isto é, a amostra

não necessariamente proverá de uma variável aleatória absolutamente contínua, como a normal por exemplo, por isso, outros métodos são necessários. São estes: baseado em resultados assintóticos, testes de escores ordinais e reamostragem.

1) Método baseado em resultados assintóticos

Matriz de covariância assintótica

Conforme sugere KOCHERGINSKY *ET AL.* (2005), e explica SANTOS (2012), há duas formas de estimar assintoticamente a matriz de covariância do vetor de estimadores dos parâmetros do modelo de regressão quantílica, com o objetivo de utilizá-la para a construção de intervalos de confiança para os parâmetros desses modelos.

Descreve-se a seguir o teorema de KOENKER e BASSETT JR (1978), que prova que os estimadores dos parâmetros da regressão quantílica são não viesados e assintoticamente seguem distribuição normal. Então, para uma amostra de n observações, considerando $\hat{\beta}(\tau_i)$, a solução do problema da minimização (2.9), para $\tau = \tau_i$:

Teorema 2.2.3 (KOENKER e BASSETT JR (1978)) *Seja $\hat{\beta}(\tau_1), \hat{\beta}(\tau_2), \dots, \hat{\beta}(\tau_m)$, com $0 < \tau_1 < \tau_2 < \dots < \tau_m < 1$, uma sequência de estimadores para os parâmetros do modelo (2.11). Seja $\xi_i(\tau_i) = F^{-1}(\tau_i)$ o quantil de ordem τ_i e assuma que:*

- i) F é contínua e tem densidade f contínua e positiva em ξ_i , para $i = 1, 2, \dots, m$.*
- ii) A matriz X de planejamento tem uma coluna de uns, ou seja, o modelo é ajustado com intercepto.*

iii) $\lim_{n \rightarrow \infty} n^{-1} X' X = Q$, em que Q é matriz positiva definida.

Nessas condições, pode-se mostrar que:

$$\sqrt{n}(\hat{\beta}(\tau_1) - \beta(\tau_1), \hat{\beta}(\tau_2) - \beta(\tau_2), \dots, \hat{\beta}(\tau_m) - \beta(\tau_m)) \xrightarrow{D} N_{m \times p}(\mathbf{0}, V((\tau_1), \dots, (\tau_m)))$$

em que a matriz de covariâncias $V((\tau_1), \dots, (\tau_m))$, pode ser definida como

$$\Omega((\tau_1), \dots, (\tau_m); F) \otimes Q^{-1}$$

sendo $\Omega((\tau_1), \dots, (\tau_m); F)$ a matriz de covariâncias entre m quantis amostrais de amostras aleatórias com distribuição F , e $\otimes$ o produto de Kronecker.

Conforme os autores propõem, supondo τ único, segue-se as duas formas de estimar a matriz de covariância assintótica de $\hat{\beta}(\tau)$. Primeiramente, para o caso em que os erros não são identicamente distribuídos, temos:

$$V(\tau) = \tau(1 - \tau)(X' F X)^{-1} (X' X) (X' F X)^{-1} \quad (2.12)$$

em que $F = \text{diag}(f_1(0), \dots, f_n(0))$ matriz diagonal, e $f_j, j = 1, \dots, n$, é a função densidade dos erros.

Porém, essa função não é sempre determinada de forma explícita, então, uma possibilidade sugerida é substituir o valor de $f_j(0)$ na matriz $(X'FX)$ por uma estimativa assintoticamente não viciada, como:

$$\frac{2h_n}{x_i' \hat{\beta}_{\tau+h_n} - x_i' \hat{\beta}_{\tau-h_n}}.$$

em que $\lim_{n \rightarrow \infty} h_n = 0$. Pode-se verificar métodos para cálculo de h_n em HALL e SHEATHER (1988) e KOENKER (2005).

Para o caso em que os erros são identicamente distribuídos, isso significa que, $f_1(x) = \dots = f_n(x) = f(x)$ em (2.12), temos a seguinte forma de calcular a matriz:

$$V(\tau) = \frac{\tau(1-\tau)}{f^2(0)}(X'X)^{-1}. \quad (2.13)$$

Nessa situação, segundo KOCHERGINSKY *ET AL.* (2005), uma estimativa de $f(0)$ pode ser obtida usando uma diferença de quantis empíricos dos resíduos, com

$$\frac{2h_n}{\hat{F}^{-1}(\tau+h_n) - \hat{F}^{-1}(\tau-h_n)}.$$

Com a estimação da matriz de covariância $V(\tau)$, e sob a distribuição normal assintótica descrita no Teorema (2.2.3), é possível construir os intervalos de confiança para os parâmetros.

2) Método dos testes de escores ordinais

O método dos testes de escores ordinais, proposto inicialmente por GUTENBRUNNER e JURECKOVÁ (1992) e posteriormente KOENKER (2005), utiliza estatísticas de ordem condicionais para calcular a função escore e a respectiva estatística do teste de escore, em seguida, através de um algoritmo de programação linear, verifica-se para quais valores do vetor de estimativas dos parâmetros do modelo a hipótese nula não é rejeitada e ao final do processo iterativo, constrói-se um intervalo de confiança para os parâmetros de interesse. Segundo KOCHERGINSKY *ET AL.* (2005), este método também apresenta dificuldades computacionais, no caso de banco de dados muito grandes, e também, não estima a matriz de variância e covariância dos estimadores dos parâmetros. Para mais detalhes dessa teoria sugere-se KOENKER (2005).

3) Método de reamostragem

O primeiro método de reamostragem proposto no contexto da regressão quantílica, foi o de *bootstrap* não-paramétrico (EFRON e TIBSHIRANI, 1994). Este procedimento é baseado na coleta de amostras com reposição a partir da função de distribuição acumulada empírica $F_n(y)$ (2.5), uma vez que não conhecemos a distribuição dos dados F (2.3). Então, seleciona-se os pares de observações (Y_i, X_i) com probabilidade $1/n$, em que n é

o tamanho da amostra, de forma a construir um novo vetor Y^* com valores da variável resposta e uma nova matriz de planejamento X^* . Esse procedimento é repetido B vezes, e em cada repetição o vetor $\hat{\beta}^*(\tau)$ (réplica *bootstrap* de $\beta(\tau)$) é calculado. Com essas B estimativas para o vetor de parâmetros, é calculado o estimador *bootstrap* do erro-padrão de $\hat{\beta}^*(\tau)$, que é o desvio padrão amostral das réplicas *bootstrap*, $\hat{\beta}_i^*(\tau)$ com $i = 1, \dots, B$, dado pela expressão:

$$E.P.(\hat{\beta}^*(\tau)) = \sqrt{\frac{1}{B-1} \sum_{i=1}^B (\hat{\beta}_i^*(\tau) - \bar{\beta}^*(\tau))^2} \quad (2.14)$$

sendo $\bar{\beta}^*(\tau) = \frac{1}{B} \sum_{i=1}^B \hat{\beta}_i^*(\tau)$ a estimativa *bootstrap*.

Dessa forma, pode-se escrever um intervalo de confiança para $\beta(\tau)$ da seguinte forma:

$$IC(\beta(\tau), 1 - \alpha) = \hat{\beta}(\tau) \pm z_{\alpha/2} \widehat{E.P.(\hat{\beta}^*(\tau))}.$$

em que $z_{\alpha/2}$ é o quantil de ordem $1 - \alpha/2$ da distribuição normal padrão.

Segundo EFRON e TIBSHIRANI (1994), o tamanho de reamostragens utilizadas, variam de acordo com o uso da técnica do *bootstrap*, usualmente são utilizados 50, 200 ou 1000 reamostras. Como é necessário reajustar o modelo para cada reamostra gerada, esse procedimento pode ser bem demorado para número grande de observações ou de variáveis explicativas, sendo computacionalmente exigente. Nesse sentido, KOCHERGINSKY ET AL. (2005) propõem a utilização do método denominado *Markov Chain Marginal Bootstrap (MCMB)* (HE e HU, 2002) para contornar esse problema. Referencia-se a dissertação de SANTOS e ELIAN (2015) para descrição do algoritmo básico da adaptação do *MCMB* para regressão quantílica.

SANTOS e ELIAN (2015) realizaram um estudo de simulação supondo duas situações para a distribuição dos erros do modelo: um caso com erro simétrico (apresentando distribuição normal) e outro assimétrico (com distribuição Gama). Ele destaca alguns resultados importantes: para grandes amostras, ambos os métodos apresentaram resultados satisfatórios, mas, como mencionado anteriormente, o método de *bootstrap* pode ser mais lento do que o *MCMB*. Porém, para pequenas amostras ou quando o interesse está nas caudas da distribuição da variável resposta, deve-se tomar um cuidado maior na escolha do método. De forma geral, os testes de escores ordinais e o método *bootstrap*, apresentaram resultados mais próximos do esperado.

Teste de hipótese

Em modelos de regressão linear clássicos, uma forma usual de verificar a significância da regressão, ou seja, se a variável independente influencia significativamente a variável dependente, é através de testes de hipóteses para os parâmetros.

Por exemplo, uma hipótese de bastante interesse seria:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

contra a hipótese alternativa de que pelo menos um deles seja diferente de zero.

Se a hipótese nula não for rejeitada, significa que as p variáveis independentes não exercem influência na variável dependente, segundo o modelo ajustado, o que permite concluir que o modelo proposto não é adequado para descrever o fenômeno. Este teste pode ser feito pelo método da análise de variância (ANOVA).

Para o caso em que os parâmetros são estimados pelo método da regressão quantílica, como no modelo (2.11), se estimados m diferentes modelos, e sendo $\beta(\tau_j)$ o vetor dos p parâmetros para $\tau = \tau_j (j = 1, 2, \dots, m)$, então a hipótese linear geral do vetor $\zeta = (\beta(\tau_1)', \dots, \beta(\tau_m)')'$, pode ser escrita como:

$$H_0 : R\zeta = \mathbf{r} \quad (2.15)$$

em que R é uma matriz de constantes conhecidas de ordem q e de posto completo, e $\mathbf{r}$ é um vetor de constantes conhecidas, $m \times 1$.

Para testar essa hipótese, a literatura sugere 3 métodos: o método de Wald por KOENKER (2005), o método anovar por CHEN ET AL. (2008) e o método de escores ordinais (rank) por GUTENBRUNNER ET AL. (1993).

Nestas condições, segundo a formulação dos testes do tipo Wald de KOENKER (2005), para testar uma hipótese do tipo $H_0 : \beta_1(\tau) = \beta_2(\tau) = \dots = \beta_p(\tau) = 0$, a estatística de teste, pode ser escrita como:

$$T_n = n \frac{f^2(0)}{\tau(1-\tau)} \sum_{i=1}^p \frac{\beta_i^2(\tau)}{v_{ii}}$$

em que v_{ii} é i -ésimo elemento da diagonal da matriz $(X'X)^{-1}$. Para ser uma estatística, $f(0)$ deve ser substituído por uma estimativa. Essa formulação também pode ser usada para testar hipóteses envolvendo diversos quantis. Segundo KOENKER (2005), a estatística Tn (2.2.2) tem assintoticamente distribuição χ_q^2 sob H_0 , em que q é o posto da matriz R , conforme descrito. Se o interesse fosse uma hipótese do tipo $H_0 : \beta_1(\tau) = \beta_2(\tau) = \dots = \beta_p(\tau) = 0$, em que β_0 seria o intercepto, então a estatística Tn (2.2.2) $\xrightarrow{D} \chi_{p-1}^2$, com p o números de parâmetros do modelo.

CHEN ET AL. (2008), propõem um método que descrevem como similar a uma análise de variância para modelos de regressão L_1 . Com alguns cálculos, os autores utilizam a função *check* (2.4) para escrever a estatística da seguinte forma:

$$M_n = \min_{\beta \in \Omega_0} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i' \beta) - \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}_i' \beta)$$

em que Ω_0 é o espaço paramétrico gerado pela hipótese nula.

Os autores também mostram sob H_0 que

$$M_n \xrightarrow{D} \frac{\chi_q^2}{4f(0)}$$

sendo $f(\cdot)$ a densidade dos erros. De forma similar ao método anterior, se o interesse fosse uma hipótese do tipo $H_0 : \beta_1(\tau) = \beta_2(\tau) = \dots = \beta_p(\tau) = 0$, em que β_0 seria o intercepto, então a estatística M_n (2.2.2) tem assintoticamente distribuição $\frac{\chi^2_{p-1}}{4f(0)}$.

GUTENBRUNNER *ET AL.* (1993) desenvolveram um método a partir de escores ordinais de regressão, referido anteriormente na construção de intervalos de confiança.

Entre os três métodos para testar a hipótese (2.15), SANTOS e ELIAN (2015) apresenta como resultado de seus estudos de simulação, que o teste que considera os desvios absolutos ponderados, anovar, se mostrou o menos poderoso, e os testes Wald e rank, em geral, apresentam comportamento equivalente, não sendo possível determinar qual é o mais poderoso.

2.2.3 Seleção de modelos e diagnósticos

Coefficiente de determinação R^2

Para seleção de modelos na regressão linear clássica, uma medida descritiva muito utilizada é o coeficiente de determinação R^2 :

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

em que $SQRes = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ e $SQTotal = \sum_{i=1}^n (Y_i - \bar{Y})^2$.

R^2 indica a proporção da variação de Y que é explicada pela regressão. Sendo $0 \leq R^2 \leq 1$, quanto mais próximo de 1 for o seu valor, maior será o poder explicativo do modelo estimado.

Como essa medida depende do número de observações da amostra, tende a crescer quando n diminui, então, como alternativa para corrigir esse problema, foi definido outra medida: o coeficiente de determinação ajustado $\bar{R}^2$, dado por:

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

Semelhante ao R^2 da regressão linear clássica, têm-se para aferir os modelos de regressão quantifica o R^2 sugerido por KOENKER e MACHADO (1999).

Inicialmente, particionando a função de regressão quantílica definida em (2.7), um modelo linear para o quantil condicional, com p variáveis explicativas pode ser escrito como:

$$Q_\tau(Y_i|x) = \mathbf{x}'_{i1}\boldsymbol{\beta}_1(\tau) + \mathbf{x}'_{i2}\boldsymbol{\beta}_2(\tau) \quad (2.16)$$

em que x_i é i -ésima linha da matriz X do delineamento, particionada em duas partes $\mathbf{x}_{i1}$ e $\mathbf{x}_{i2}$ de dimensões $p - q$ e q , respectivamente.

Admitindo-se a restrição q -dimensional a esse modelo:

$$H_0 : \beta_2 = 0.$$

É necessário também, considerar um particionamento semelhante para o vetor de parâmetros $\beta(\tau)$. Então, seja $\hat{\beta}(\tau)$ o estimador que minimiza a soma dos erros absolutos ponderados do modelo completo, ou seja,

$$\hat{V}(\tau) = \sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}'_i \hat{\beta}(\tau)).$$

E seja, $\tilde{\beta}(\tau)$, a solução para o modelo restrito imposto pela hipótese acima: $Q_{\tau}(Y_i|x) = \mathbf{x}'_{i1} \beta_1(\tau)$:

$$\tilde{V}(\tau) = \sum_{i=1}^n \rho_{\tau}(y_i - \mathbf{x}'_{i1} \tilde{\beta}_1(\tau)).$$

Então, têm-se o coeficiente de determinação para a regressão quantílica:

$$R^1(\tau) = 1 - \frac{\hat{V}(\tau)}{\tilde{V}(\tau)}. \quad (2.17)$$

É fácil verificar que, como $\tilde{\beta}(\tau)$ é obtido através da restrição imposta a $\hat{\beta}(\tau)$, portanto, $0 \leq R^1(\tau) \leq 1$, assim como R^2 . É importante observar que $R^1(\tau)$ é uma medida local, que mede a qualidade do ajuste correspondente a modelos de regressão quantílica em um quantil específico em função de uma soma ponderada dos resíduos absolutos, conforme explica KOENKER e MACHADO (1999). Enquanto que R^2 é uma medida global que mede o correspondente sucesso sobre toda a distribuição condicional.

Quanto a interpretação dessa estatística, podemos considerar que um modelo completo como em (2.16) é melhor que o modelo reduzido imposto pela restrição, se o valor de $\hat{V}(\tau)$ for significativamente menor que $\tilde{V}(\tau)$, e isso significa que o ajuste do modelo para o quantil condicional de ordem τ , se altera com inclusão das variáveis x_2 no modelo.

Resíduos quantílicos

Para análise dos resíduos dos modelos de regressão quantílica, SANTOS (2012) apresenta técnicas utilizando os resíduos quantílicos aleatorizados, propostos por DUNN e SMYTH (1996). Se os parâmetros do modelo são consistentemente estimados, esse resíduo converge para uma distribuição normal padrão, independente da distribuição da variável resposta e de sua dispersão. Para isso, a suposição de alguma distribuição de probabilidade para a variável resposta é necessária, então, é utilizado o resultado de que desvios absolutos ponderados coincidem com o estimador de máxima verossimilhança, quando os erros do modelo de regressão quantílica seguem distribuição de Laplace assimétrica ($LA(\mu, \sigma, \tau)$). Sua distribuição acumulada é definida por:

$$F(y|\mu, \sigma, \tau) = \begin{cases} \tau \exp \left(\frac{1-\tau}{\sigma}(y - \tau) \right), & \text{se } y \leq \mu \\ 1 - (1 - \tau) \exp \left(-\frac{\tau}{\sigma}(y - \mu) \right), & \text{se } y > \mu \end{cases}$$

Como F é contínua, o teorema da inversa da função distribuição acumulada garante que $F(y; \mu, \sigma, \tau)$ tem distribuição uniforme no intervalo $(0,1)$. Dessa forma, os resíduos quantílicos são definidos por:

$$r_{q,i} = \Phi^{-1}\{F(y_i, \hat{\mu}_i, \hat{\sigma}, \tau)\} \quad (2.18)$$

sendo $\Phi(\cdot)$ é a função de distribuição acumulada da distribuição normal padrão, $\hat{\mu} = \mathbf{x}_i' \boldsymbol{\beta}(\boldsymbol{\tau})$; τ é parâmetro fixado e $\hat{\sigma}$ é o estimador de máxima verossimilhança de σ , de acordo com SANTOS (2012), dado por:

$$\hat{\sigma} = n^{-1} \sum_{i=1}^n \rho_{\tau}(y_i - \hat{y}_i).$$

3 ESTUDO DE MOTIVAÇÃO

3.1 Material

Descrição dos dados

Para ilustrar a metodologia apresentada neste trabalho, utiliza-se como motivação, dados de pesos em kg. de 28 bezerros, durante um período de 26 semanas após o nascimento, isto é, aproximadamente 6 meses. As pesagens foram realizadas quinzenalmente e apresentam dados ausentes. O estudo foi conduzido pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e analisado por SINGER *ET AL.* (2017) e CRUZ (2020). Os dados estão disponíveis na tabela A.3 (em Anexo A) e foram obtidos em SINGER *ET AL.* (2017). Ambas as análises realizadas, combinam técnicas descritivas e de diagnósticos e apresentam um modelo misto com erros aleatórios autorregressivos como adequado para descrever o comportamento do valor esperado do peso, conforme o objetivo do estudo. O objetivo do presente trabalho é apresentar e verificar os modelos de regressão quantílica como uma alternativa mais robusta ou um complemento a outros métodos convencionais apresentados na literatura, em casos de assimetria e heterogeneidade de variâncias residuais.

Inicialmente foi realizado um estudo descritivo. Os perfis individuais de resposta são apresentados nas figuras 3.1 e 3.2. Verifica-se um aumento de peso para todos os indivíduos ao longo do tempo. É possível observar que algumas curvas crescem mais rápido do que outras, demonstrando que existem diferentes velocidades de crescimento entre esses animais. Em geral, a maioria dos animais apresenta uma tendência quadrática nos perfis individuais, porém, alguns apresentam um comportamento linear, por exemplo, o bezerro 20, 23 e 25.

Verifica-se também, que apresentam heterogeneidade de variâncias ao longo do tempo, fato que também pode ser visualizado na figura 3.3, em que os pesos dos bezerros em função das semanas estão representados por meio do gráfico de caixa (*boxplot*). As médias, desvios padrões, coeficientes de assimetria e coeficientes de variação dos pesos de cada semana de observação, são apresentados na tabela 3.1. É evidente que a variação entre os indivíduos é crescente ao longo do tempo e substancialmente menor no início do estudo, o que permite afirmar então, que há uma heterogeneidade de variâncias. A homogeneidade entre as semanas também foi verificada com o teste de Bartlett, sendo $p\text{-valor} < 0,0001$ obtido, ou seja, a hipótese nula de que as variâncias são iguais foi rejeitada considerando um nível de 5% de significância, entretanto, nenhum ponto discrepante é observado nos *boxplots*. Os coeficientes de variação também indicam menor variabilidade na fase inicial, e quanto aos coeficientes de assimetria, observa-se que algumas semanas possuem uma leve assimetria positiva, ou seja, suas observações estão concentradas suavemente do lado esquerdo da distribuição.

Figura 3.1. Perfis individuais de crescimento ao longo das semanas.

Figura 3.2. Perfis individuais de crescimento de cada animal ao longo das semanas.

Figura 3.3. Gráfico de caixa (*boxplot*) dos pesos dos bezerros ao longo das semanas.

Tabela 3.1. Médias, desvios padrões, coeficientes de assimetria e coeficientes de variação (CV) dos pesos dos bezerros referentes a cada semana de observação.

Semanas	Média	Desvio padrão	Assimetria	CV %
0	33,74	4,57	0,56	13,54
2	36,73	4,41	0,53	12,01
4	40,38	5,34	0,20	13,22
6	45,45	6,67	0,23	14,68
8	50,47	8,83	0,42	17,49
10	58,11	10,50	0,43	18,08
12	64,48	13,80	0,56	21,40
14	71,45	14,86	0,40	20,79
16	79,92	16,39	0,57	20,51
18	90,05	17,48	0,48	19,41
20	98,87	18,27	0,37	18,48
22	109,87	20,89	0,20	19,02
24	119,16	21,68	0,06	18,19
26	128,55	23,73	-0,10	18,46

3.2 Métodos

Modelos

Na análise descritiva, observa-se em geral uma tendência quadrática nos perfis individuais, com alguns animais apresentando um comportamento linear (Figura 3.2). Desta forma, para quantis igualmente espaçados no intervalo $[0,10; 0,90]$, e também para a média, considera-se um modelo linear (M_1/M_1^*) e um polinomial de grau 2 (M_2/M_2^*) como candidatos:

Modelos de regressão clássica:

$$M_1: y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}, \quad i = 1, 2, \dots, n, \quad j = 1, \dots, s. \quad (3.1)$$

$$M_2: y_{ij} = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \epsilon_{ij}, \quad i = 1, 2, \dots, n, \quad j = 1, \dots, s. \quad (3.2)$$

sendo y_{ij} os pesos dos animais; x_j são as semanas de observação; β_0 , β_1 e β_2 os parâmetros a serem estimados e ϵ_{ij} o erro aleatório.

Modelos de regressão quantílica:

$$M_1^*: y_{ij} = \beta_0(\tau) + \beta_1(\tau)x_j + \varepsilon_{ij}(\tau), \quad i = 1, \dots, n, \quad j = 1, \dots, s. \quad (3.3)$$

$$M_2^*: y_{ij} = \beta_0(\tau) + \beta_1(\tau)x_j + \beta_2(\tau)x_j^2 + \varepsilon_{ij}(\tau), \quad i = 1, \dots, n, \quad j = 1, \dots, s. \quad (3.4)$$

sendo y_{ij} os pesos dos animais; x_j são as semanas de observação; $\beta_0(\tau)$, $\beta_1(\tau)$ e $\beta_2(\tau)$ são os parâmetros a serem estimados no τ -ésimo quantil e $\varepsilon_{ij}(\tau)$ é o erro aleatório.

Com o interesse em comparar o modelo reduzido (M_1/M_1^*) com o modelo completo proposto (M_2/M_2^*), utiliza-se o teste de anovar para os modelos de regressão quantílica e para os modelos de mínimos quadrados, utiliza-se o teste F, ambos considerando um nível de 5% de significância. Dessa forma, as hipóteses consideradas são:

- H_0 : O efeito do modelo completo é igual ao efeito do modelo reduzido (efeito quadrático não é significativo);
- H_1 : Há diferença significativa entre os efeitos dos modelos propostos.

Se H_0 não for rejeitada, conclui-se que o termo quadrático não foi significativo, logo M_1 é preferível.

Além disso, também são calculados o critério de informação de Akaike (AIC) e o coeficiente de determinação: R^1 para os modelos de regressão quantílica e R^2 para o modelo de regressão usual, descritos anteriormente. Maiores coeficientes indicam melhores ajustes e menores AIC indicam melhores ajustes.

Para estimação dos parâmetros da regressão clássica utiliza-se o método dos mínimos quadrados e para a regressão quantílica o algoritmo *simplex* (KOENKER e D'OREY, 1987).

Teste de hipótese

Verifica-se se as diferenças observadas nos diferentes quantis selecionados são significativas por meio do teste de Wald, considerando um nível de 5% de significância. Têm-se as hipóteses:

- $H_0 : \beta_0(\tau) = \beta_0(\tau'); \beta_1(\tau) = \beta_1(\tau') \text{ e } \beta_2(\tau) = \beta_2(\tau');$
- $H_1 : \beta_0(\tau) \neq \beta_0(\tau') \text{ ou } \beta_1(\tau) \neq \beta_1(\tau') \text{ ou } \beta_2(\tau) \neq \beta_2(\tau'),$

sendo $\tau \neq \tau'$, ou seja, realiza-se o teste simultâneo para múltiplos parâmetros para verificar se a função do τ -ésimo quantil e τ' -ésimo quantil são diferentes.

Técnicas de diagnósticos

A fim de selecionar os quantis mais adequados para explicar a relação do peso com o tempo, verifica-se os efeitos da variável semana ao longo de diferentes pontos da distribuição da variável peso. Inicialmente verifica-se os gráficos dos intervalos de confiança dos parâmetros ao longo dos quantis, igualmente espaçados entre [0.05, 0.95], obtidos pelo método baseado em resultados assintóticos e considerando um nível de 5% de significância. Em seguida, realiza-se uma análise dos resíduos quantílicos de tais modelos, conforme foram definidos anteriormente em 2.18, igualmente espaçados no intervalo [0,10; 0,90], por meio de gráficos de envelope simulado de 95% de confiança.

Também verifica-se a precisão do ajuste dos modelos por meio das medidas estatísticas: erro absoluto médio (EAM) e o erro quadrático médio (EQM), sendo que menores EAM e EQM indicam melhores ajustes. Estes são dados sob as seguintes formulações:

$$\text{EQM} = \frac{\sum_{i=1}^n (\hat{y}_{ij}(\tau) - y_{ij})^2}{n}$$

$$\text{EAM} = \frac{\sum_{i=1}^n \sqrt{(\hat{y}_{ij}(\tau) - y_{ij})^2}}{n}$$

sendo $\hat{y}_{ij}(\tau)$ o valor predito para o τ -ésimo quantil (para a média: $\hat{y}_{ij}$); y_{ij} o valor observado e n o número de observações.

A análise da qualidade do modelo do quantil selecionado e do modelo de regressão clássica, é realizada por meio dos resíduos quantílicos, através de gráficos resíduos vs valores preditos, envelopes simulados e histogramas.

As análises foram realizadas com o *software* R (R DEVELOPMENT CORE TEAM, 2020), principalmente por meio do pacote **quantreg** (KOENKER, 2011). Para todos os testes realizados considerou-se um nível de 5% de significância.

Quantis extremos

Conforme visto anteriormente, é possível verificar o efeito da semana sob o peso em diferentes pontos da distribuição, além da posição central. Com essa finalidade, foram selecionados dois quantis extremos, 0,10 e 0,90 para descrever a cauda inferior e superior da distribuição, respectivamente. Foram calculadas as taxas médias de variação das curvas de cada um desses quantis. Para definição dessa taxa, de modo geral, supondo que y é uma função de x , isto é, $y = f(x)$, então se x variar de x_0 a x_1 , a variação em x será $\Delta x = x_1 - x_0$, e a variação correspondente em y será $\Delta y = y_1 - y_0$. O quociente dessas diferenças é denominado taxa média de variação de y em relação a x no intervalo $[x_0, x_1]$:

$$\frac{\Delta y}{\Delta x} = \frac{y_1 - y_0}{x_1 - x_0}$$

Dessa forma, para cada uma das curvas dos modelos ajustados, calcula-se a taxa média de variação do peso em relação ao tempo (semanas) no intervalo $[0, 26]$.

Após o ajuste dos modelos, os animais foram classificados em subgrupos, sendo cada qual pertencente a um quantil apresentado (posição central e extremos da distribuição). A medida de dissimilaridade utilizada para a classificação, foi a distância euclidiana entre a taxa média de variação de cada curva e a taxa média de variação observada dos pesos de cada bezerro. O animal foi classificado como pertencente ao grupo cuja menor distância foi obtida. A distância euclidiana, que se reduz ao módulo do erro absoluto, é dada por:

$$d_\tau = \sqrt{(t\hat{m}v_{(\tau)} - tmv_i)^2} = |t\hat{m}v_{(\tau)} - tmv_i|$$

sendo $t\hat{m}v_{(\tau)}$ é a taxa média de variação da curva do τ -ésimo quantil e tmv_i é a taxa média de variação observada dos pesos de cada bezerro.

Para verificação dos resultados obtidos no agrupamento, utiliza-se uma das principais medidas de validação interna, o *Silhouette coefficient*. A validação interna, é caracterizada por basear-se exclusivamente nos dados e agrupamentos gerados, sem qualquer informação prévia a respeito de quantos ou quais são os agrupamentos. Assim, este indicador pode auxiliar na determinação do número de grupos adequados ao conjunto de dados em análise (ARBELAITZ ET AL., 2013). De acordo com ROUSSEUW (1987), o Coeficiente *Silhouette* é uma métrica baseada na separação e compactação dos grupos, iniciando-se pelo cálculo de um coeficiente para cada observação i :

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

sendo a_i a distância média entre a observação i e as demais observações do mesmo grupo e b_i a distância média entre a observação i e as demais observações do grupo mais próximo.

O coeficiente s_i é computado para cada observação i , e gera um número entre $[-1, 1]$, sendo que quanto mais próximo a 1, maior a similaridade entre as observações do seu próprio grupo, como também maior a dissimilaridade com as observações de outros agrupamentos. Um s_i em torno de 0 significa que a observação se encontra entre dois grupos, enquanto observações com um s_i negativo, provavelmente foram alocadas no grupo incorreto. A média dos coeficientes de cada observação determina o *Silhouette Score* para o conjunto de dados completo, e indica a qualidade dos agrupamentos gerados, da mesma forma, quanto mais próximo de 1, mais precisa é a formação dos grupos.

4 RESULTADOS E DISCUSSÃO

Conforme descrito anteriormente, inicialmente verifica-se o modelo linear (M_1) e o modelo quadrático (M_2) em alguns quantis igualmente espaçados no intervalo $[0,10; 0,90]$, visto que, em quantis diferentes, diferentes modelos podem ser adequados. Os mesmos modelos também são verificados para a média. A tabela 4.1 apresenta o valor-p do teste anovar, o coeficiente de determinação $R^1(\tau)$ e o critério de informação de Akaike (AIC), comparando os dois modelos de regressão quantílica, e para os modelos de mínimos quadrados, o R^2 ajustado, o AIC e o valor-p do teste F. Verifica-se que o efeito quadrático foi significativo para ambas as metodologias e em todos os quantis, também mostrando-se o mais adequado de acordo com os critérios de seleção de modelos.

Tabela 4.1. Valor-p do teste anovar e do teste F^a, coeficiente de determinação ($R^1(\tau)$ e R^2 ajustado^b) e critério de informação de Akaike (AIC) para os modelos de regressão quantílica em diferentes quantis e para o modelo de mínimos quadrados ordinários (MQO).

valor-p		$R^1/\bar{R}^2$		AIC	
		M ₁	M ₂	M ₁	M ₂
$\tau = 0,10$	0,007	0,47	0,49	3122,75	3086,58
$\tau = 0,20$	<0,001	0,51	0,54	3089,68	3040,60
$\tau = 0,30$	<0,001	0,54	0,58	3084,87	3019,10
$\tau = 0,40$	<0,001	0,56	0,61	3084,66	3006,11
$\tau = 0,50$	<0,001	0,59	0,62	3081,26	3016,80
$\tau = 0,60$	<0,001	0,60	0,63	3093,99	3041,89
$\tau = 0,70$	<0,001	0,61	0,63	3117,41	3078,02
$\tau = 0,80$	<0,001	0,61	0,63	3165,79	3117,13
$\tau = 0,90$	<0,001	0,60	0,62	3257,03	3212,73
MQO	<0,001 ^a	0,79 ^b	0,81 ^b	3087,09	3055,40

^a Valor-p do teste F para os modelos de mínimos quadrados.

^b R^2 ajustado para os modelos de mínimos quadrados.

Definido o modelo quadrático, verifica-se os efeitos da variável semana ao longo de diferentes pontos da distribuição da variável peso na figura 4.1, que apresenta as estimativas de diversos quantis, igualmente espaçados entre $[0,05, 0,95]$ e seus intervalos de confiança, linha azul e a área cinza, respectivamente. As linhas vermelhas indicam as estimativas dos parâmetros de MQO, sendo as pontilhadas seus respectivos intervalos de confiança. Nota-se que as estimativas dos parâmetros β_0 e β_1 são crescentes ao longo dos quantis, indicando que os efeitos da variável semana variam em quantis distintos. O parâmetro β_2 apresenta comportamento quase constante. Ao observar os intervalos de confiança, nota-se que estes são menores para quantis inferiores a $\tau = 0,50$ e pode-se concluir que a variável semana e semana² são estatisticamente significantes para todos os quantis.

Figura 4.1. Estimativas dos parâmetros e respectivos intervalos de confiança, considerando um nível de 5% de significância, ao longo dos quantis.

A fim de verificar os quantis mais adequados para explicar a relação do peso com o tempo, uma análise inicial dos modelos regressão quantílica foi realizada por meio de gráficos de envelope dos resíduos quantílicos (figura 4.2). Algumas conclusões interessantes podem ser obtidas, por exemplo, a partir do quantil de ordem 0,50, o ajuste piora de forma gradativa, sendo que a porcentagem de pontos fora da banda de confiança passa de 30%, o que não ocorre com os quantis inferiores. Esse fato, sugere que a distribuição condicional do peso em função das semanas apresenta uma pequena assimetria a direita, já que, quantis superiores não apresentam ajustes tão bons quanto os quantis inferiores, segundo o gráfico de envelope simulado.

Observa-se também, que alguns quantis podem ser selecionados como candidatos para uma melhor descrição dessa relação, são eles: $\tau = 0,30; 0,35; 0,40$ e $0,45$. Foi verificado se as diferenças observadas nos diferentes quantis selecionados, são significativas por meio do teste de Wald, sendo p-valor $< 0,001$ obtido, rejeitando a hipótese nula, ou seja, pelo menos um dos coeficientes é estatisticamente diferente dos demais, comparando os quantis dois a dois. Outros percentis dentro desse intervalo não são considerados, visto a diferença não significativa entre eles.

Figura 4.2. Gráfico de envelope simulado dos modelos de regressão quantílica para quantis igualmente espaçados no intervalo $[0,10; 0,90]$.

A tabela 4.2 apresenta o critério de informação de Akaike (AIC), o erro quadrático médio (EQM) e o erro absoluto médio (EAM), dos respectivos quantis selecionados ($\tau = 0,30; 0,35; 0,40$ e $0,45$), da mediana ($\tau = 0,50$) e também para o modelo de MQO.

Primeiramente, verifica-se que ambos os modelos quantílicos, obtiveram melhores valores de AIC em comparação com o modelo de MQO. O quantil $\tau = 0,40$ obteve melhor AIC, porém valor aproximado dos demais quantis. Observa-se também, que a mediana obteve o menor EAM e o modelo de MQO obteve melhor EQM.

Tabela 4.2. Critério de informação de Akaike (AIC), erro quadrático médio (EQM) e erro absoluto médio (EAM) para os modelos de regressão quantílica em diferentes quantis e para o modelo de mínimos quadrados ordinários (MQO).

Quantis e MQO	AIC	EQM	EAM
0,30	3019,11	266,29	11,67
0,35	3009,10	251,58	11,27
0,40	3006,12	237,63	10,94
0,45	3010,60	225,38	10,73
0,50	3016,81	218,64	10,64
MQO	3055,40	216,19	10,72

Com base nos resultados obtidos, as estimativas e suas respectivas medidas associadas à inferência, dos referidos quantis são apresentadas na tabela 4.3 (os demais quantis podem ser visualizados na tabela A.1 no apêndice). Observa-se em geral, menores intervalos de confiança e erros padrões para os quantis, sendo que, o quantil 0,45 obteve estimativas ligeiramente mais precisas. Desta forma, na figura 4.3 é possível visualizar este ajuste e o de MQO.

Tabela 4.3. Estimativas dos parâmetros para o modelo de regressão quantílica dos quantis condicionais $\tau = 0,30; 0,35; 0,40; 0,45$ e $0,50$ e para o modelo de MQO.

$\tau = 0,30$	Estimativa	Erro padrão	$\Pr(> t)$	I.C. ¹
β_0	31,0567	0,7384	<0,0001	29,11; 32,53
β_1	0,9200	0,2545	0,0003	0,73; 1,32
β_2	0,1008	0,0143	<0,0001	0,06; 0,11
$\tau = 0,35$				
β_0	31,9000	0,6424	<0,0001	30,02; 33,06
β_1	0,9437	0,2476	0,0002	0,63; 1,17
β_2	0,1008	0,0131	<0,0001	0,09; 0,11
$\tau = 0,40$				
β_0	32,0000	0,5828	<0,0001	31,08; 33,03
β_1	1,0024	0,2440	<0,0001	0,73; 1,38
β_2	0,1024	0,0127	<0,0001	0,08; 0,12
$\tau = 0,45$				
β_0	32,0000	0,3909	<0,0001	31,06; 33,89
β_1	1,2771	0,2276	<0,0001	0,69; 1,65
β_2	0,0935	0,0108	<0,0001	0,08; 0,12
$\tau = 0,50$				
β_0	32,7000	0,4211	<0,0001	31,02; 33,88
β_1	1,3348	0,2427	<0,0001	0,97; 1,87
β_2	0,0945	0,0113	<0,0001	0,07; 0,11
MQO				
β_0	32,8278	1,9984	<0,0001	28,89; 36,75
β_1	1,7121	0,3577	<0,0001	1,00; 2,41
β_2	0,0782	0,0132	<0,0001	0,05; 0,10

¹Intervalos de confiança (I.C.) considerando um nível de 5% de significância.

Figura 4.3. Ajuste do modelo de regressão quantílica do quantil condicional de ordem $\tau = 0,45$ e do modelo de MQO.

Diagnósticos dos resíduos

A finalidade dessa análise de resíduos é a comparação entre o modelo do quantil $\tau = 0,45$ e a média (MQO). Para isto, considera-se os resíduos quantílicos, definidos anteriormente na seção 2.2.3, visto que estes convergem para uma distribuição normal padrão, independente da distribuição da variável resposta e de sua dispersão, desde que os parâmetros do modelo sejam consistentemente estimados. Os quantis extremos podem ser vistos na imagem 4.2, entretanto a comparação com a média não faz sentido.

A figura 4.4, apresenta o gráfico dos resíduos em relação aos valores preditos, observa-se que ambos evidenciam a heterogeneidade de variâncias dos dados e verifica-se uma quantidade similar de pontos fora do intervalo considerado adequado $[-2, 2]$, similar para os dois casos.

Os histogramas dos resíduos e a densidade de uma distribuição normal padrão (figura 4.5), mostram uma melhor simetria para os resíduos quantílicos, indicando que, aparentemente estes seguem distribuição normal, o que foi confirmado com o teste de Shapiro-Wilk, sendo a hipótese nula de normalidade não rejeitada, com p-valor de 0,5086. Para os resíduos do modelo de MQO, a hipótese nula foi rejeitada a um nível de 5% de significância, com p-valor $< 0,0001$ obtido.

No gráfico normal de probabilidade com envelope simulado de 95% de confiança dos resíduos (figura 4.6), verifica-se muitos pontos fora da banda de confiança para ambos os modelos, embora o modelo de RQ apresente melhor resultado.

Esse estudo trata de dados longitudinais, e conforme verificado em SINGER *ET AL.* (2017) e CRUZ (2020), apresenta uma correlação serial nos dados, sugerindo outras es-

truturas de modelos. Embora o modelo de regressão quantílica seja mais robusto a tais suposições, muitos problemas observados na análise de diagnóstico refletem essa necessidade. Os trabalhos mencionados, obtiveram melhores resultados com um modelo misto com erros aleatórios autorregressivos. Para comparação das estimativas dos efeitos fixos, ajustou-se um modelo de regressão quantílica para $\tau = 0,45$ (tabela A.2, no apêndice I), considerando a transformação da variável semana para meses, conforme os autores propõem, observa-se que as estimativas diferem um pouco.

Figura 4.4. Gráfico dos resíduos quantílicos vs valores preditos do modelo de regressão quantílica no quantil $\tau=0,45$ (a) e do modelo de MQO (b).

Figura 4.5. Histograma dos resíduos quantílicos do modelo de regressão quantílica no quantil $\tau=0,45$ (a) e do modelo de MQO (b).

Figura 4.6. Gráfico de envelope simulado dos resíduos do modelo de regressão quantílica no quantil $\tau=0,45$ (a) e do modelo de MQO (b).

Quantis extremos

Vale mencionar que, conforme CADE *ET AL.* (1999), em quantis extremos nem sempre as medidas relacionadas a inferência e diagnóstico serão satisfatórias, entretanto ainda pode existir interesse em quantis específicos. Nesse caso, é interessante caracterizar subpopulações específicas, como animais mais ou menos eficientes em ganho de peso. Desta forma, com a finalidade de verificar o efeito da semana sob o peso em diferentes pontos da distribuição, o modelo quadrático também foi ajustado para dois quantis extremos, 0,10 e 0,90 para descrever a cauda inferior e superior da distribuição, respectivamente.

Tais ajustes são apresentados na figura 4.7 juntamente com o quantil 0,45. É possível notar comportamentos distintos, o que sugere que a relação entre as variáveis de estudo é diferente para τ 's diferentes, resultado que concorda com os gráficos das estimativas ao longo dos quantis, verificados anteriormente na figura 4.1. Foram calculadas as taxas médias de variação das curvas ajustadas dos três quantis selecionados, $\tau = 0,10$; 0,45 e 0,90, sendo elas 2,52, 3,71 e 4,58 respectivamente, o que indica que os animais apresentam variações de crescimento diferentes e crescentes ao longo dos quantis, no intervalo de tempo estudado (0 a 26 semanas). Esse fato, reflete a heterogeneidade de variâncias verificada na análise descritiva dos dados e justifica o ajuste de quantis diferentes, além da posição central ($\tau = 0,45$). Uma homogeneidade de variâncias, possivelmente ocasionaria em taxas de variações aproximadas.

Essas estimativas e suas respectivas medidas associadas à inferência, são apresentadas na tabela 4.4, sendo possível verificar que a hipótese nula de que os valores dos parâmetros sejam iguais a zero, foi rejeitada para os três quantis avaliados. Além disso, ao observar os intervalos de confiança, pode-se concluir que as estimativas são estatisticamente significantes para os quantis selecionados, e comparando tais quantis dois a dois

Figura 4.7. Ajuste dos modelos de regressão quantílica nos quantis condicionais de ordem $\tau = 0,10; 0,45$ e $0,90$.

pelo teste do tipo de Wald, a hipótese nula de igualdade de parâmetros foi rejeitada, com p-valor $<0,0001$.

Tabela 4.4. Estimativas dos parâmetros para o modelo de regressão quantílica dos quantis condicionais $\tau = 0,10; 0,45$ e $0,90$.

	Estimativa	Erro padrão	$\Pr(> t)$	I.C. ¹
$\tau = 0,10$				
β_0	28,1597	1,0522	$<0,0001$	27,03; 29,77
β_1	1,2968	0,3252	0,0001	0,47; 1,49
β_2	0,0471	0,0163	0,0042	0,04; 0,09
$\tau = 0,45$				
β_0	32,0000	0,3909	$<0,0001$	31,06; 33,89
β_1	1,2771	0,2276	$<0,0001$	0,69; 1,65
β_2	0,0935	0,0108	$<0,0001$	0,08; 0,12
$\tau = 0,90$				
β_0	38,5000	0,9770	$<0,0001$	36,11; 41,50
β_1	2,6167	0,4395	$<0,0001$	1,97; 3,40
β_2	0,0755	0,0219	0,0006	0,05; 0,12

¹Intervalos de confiança (I.C.) considerando um nível de 5% de significância

A classificação dos bezerros em subgrupos foi realizada a partir da menor distância euclidiana entre as taxas médias de variação das observações de cada bezerro com as taxas médias de variação de cada uma das curvas ajustadas pelos três quantis (tabela 4).

Tabela 4.5. Distância euclidiana entre as taxas médias de variação das observações de cada bezerro e taxas médias de variação de cada uma das curvas ajustadas pelos três quantis e sua respectiva classificação.

Bezerro	Taxa	Distância $\tau = 0,10$	Distância $\tau = 0,45$	Distância $\tau = 0,90$	Subgrupo
1	2,53	0,01	1,18	2,05	$\tau = 0,10$
2	1,92	0,60	1,78	2,66	$\tau = 0,10$
3	3,82	1,30	0,12	0,76	$\tau = 0,45$
4	3,37	0,84	0,34	1,21	$\tau = 0,45$
5	3,63	1,11	0,08	0,95	$\tau = 0,45$
6	2,85	0,33	0,86	1,73	$\tau = 0,10$
7	4,36	1,84	0,65	0,22	$\tau = 0,90$
8	2,99	0,47	0,72	1,59	$\tau = 0,10$
9	3,73	1,21	0,02	0,85	$\tau = 0,45$
10	3,23	0,71	0,48	1,35	$\tau = 0,45$
11	3,13	0,61	0,57	1,45	$\tau = 0,45$
12	4,40	1,88	0,70	0,18	$\tau = 0,90$
13	4,54	2,02	0,83	0,04	$\tau = 0,90$
14	3,74	1,22	0,03	0,84	$\tau = 0,45$
15	5,22	2,70	1,52	0,64	$\tau = 0,90$
16	4,73	2,21	1,02	0,15	$\tau = 0,90$
17	3,73	1,21	0,03	0,85	$\tau = 0,45$
18	3,88	1,36	0,18	0,70	$\tau = 0,45$
19	2,74	0,22	0,97	1,84	$\tau = 0,10$
20	4,56	2,04	0,85	0,02	$\tau = 0,90$
21	3,66	1,14	0,05	0,92	$\tau = 0,45$
22	1,58	0,94	2,13	3,00	$\tau = 0,10$
23	5,29	2,77	1,58	0,71	$\tau = 0,90$
24	4,56	2,04	0,85	0,02	$\tau = 0,90$
25	4,02	1,50	0,31	0,56	$\tau = 0,45$
26	2,90	0,38	0,81	1,68	$\tau = 0,10$
27	3,93	1,41	0,22	0,65	$\tau = 0,45$
28	2,75	0,23	0,95	1,83	$\tau = 0,10$

Os resultados do peso final de estudo para cada grupo são apresentados na tabela 4.6, que são equivalentes aos seis meses de idade. Em cada quantil $\tau = 0,10$; $0,45$ e $0,90$, foram classificados 8, 12 e 8 animais respectivamente, com pesos médios ao final do estudo de 100,2; 128,75 e 156,58 kg.

Também apresenta-se um resumo descritivo dos pesos iniciais de cada grupo (tabela 4.7), sendo os pesos médios iniciais 33,39; 33,82 e 34,17, para os quantis $\tau = 0,10$; $0,45$ e $0,90$, respectivamente. Esses valores estão próximos, o que é justificado pela variância substancialmente menor nas semanas iniciais de observação, além disso, na figura 4.8, que apresenta as curvas individuais de crescimento de acordo com o grupo, observa-se que os animais transitam entre os quantis até aproximadamente a 20ª semana.

Tabela 4.6. Quantidade de indivíduos classificados em cada subgrupo (quantil $\tau = 0,10$; $0,45$ e $0,90$) e seu resumo descritivo dos pesos na última semana.

Indivíduos	Semana final					
	Min.	1º Qu.	Mediana	Média	3º Qu.	Max.
8 ($\tau = 0,10$)	77,00	93,12	102,80	100,20	106,65	119,70
12 ($\tau = 0,45$)	112,50	126,12	130,85	128,75	132,70	137,10
8 ($\tau = 0,90$)	141,50	151,125	154,15	156,58	159,65	176,80

Tabela 4.7. Quantidade de indivíduos classificados em cada subgrupo (quantil $\tau = 0,10$; $0,45$ e $0,90$) e seu resumo descritivo dos pesos na semana inicial.

	Semana inicial					
	Min.	1º Qu.	Mediana	Média	3º Qu.	Max.
$\tau = 0,10$	28,00	30,00	32,70	33,39	35,50	42,00
$\tau = 0,45$	29,00	30,50	31,00	33,82	36,50	43,00
$\tau = 0,90$	27,00	31,42	32,50	34,17	38,88	41,00

Figura 4.8. Perfis individuais de crescimento ao longo das semanas conforme a classificação.

A fim de verificar quão similar uma observação é do seu próprio grupo em comparação ao grupo mais próximo, na figura 4.9 ilustra-se os resultados da medida de validação interna *Silhouette coefficient*. Observa-se pelo coeficiente de cada observação, que nenhum indivíduo obteve s_i negativo, indicando que todos foram alocados no grupo correto. Verifica-se também *Silhouette Score* (média dos coeficientes da amostra) alto, acima de 0,50, isto é, positivo e suficientemente elevado, assim como a média de cada grupo (tabela 4.8). Além disso, os agrupamentos possuem tamanhos semelhantes e parte significativa dos coeficientes de cada observação acima do coeficiente médio. Pode-se concluir que a medida de validação utilizada corrobora a adequada classificação dos animais. Este resultado, mostra que os dois quantis extremos (0,10 e 0,90) foram selecionados de forma satisfatória para descrever esse conjunto, e as curvas de regressão quantílica podem ser adequadas para identificar animais com diferentes potenciais de crescimento. Por exemplo, no quantil mais baixo ($\tau = 0,10$), menores pesos corporais ao longo do tempo, enquanto no quantil mais alto ($\tau = 0,90$), animais com pesos maiores.

Figura 4.9. Análise da silhueta para agrupamentos gerados considerando as curvas de crescimento da regressão quantílica.

Tabela 4.8. *Silhouette coefficient* médio de cada um dos 3 grupos definidos.

Grupo	Nº de observações	s_i médio
1	8	0,49
2	12	0,60
3	8	0,61

5 CONSIDERAÇÕES FINAIS

Nesta dissertação foram estudados os modelos de regressão quantílica como uma abordagem complementar ou uma alternativa mais robusta à análise de regressão usual. Inicialmente foi realizada uma revisão desses modelos, relacionando-os com os modelos tradicionais de regressão, bem como suas propriedades, métodos de estimação e inferenciais.

Em seguida, um estudo de motivação da metodologia apresentada foi realizado, utilizando dados de crescimento de bezerros. As análises indicaram uma heterogeneidade de variâncias ao longo do tempo, e uma leve assimetria a direita na distribuição do peso em função das semanas de observação, desta forma, algumas suposições necessárias para o modelo de regressão clássica são violadas.

A regressão quantílica mostrou-se adequada para descrever o peso como uma função do tempo, e para compreender como o tempo afeta a variação e a assimetria dos pesos dos bezerros em comparação ao modelo de mínimos quadrados. Por meio de diferentes quantis foi possível identificar animais com padrões de crescimentos distintos, como menores e maiores pesos corporais ao longo do tempo. A classificação dos animais em grupos de diferentes potenciais de crescimento, considerando as curvas de crescimento da regressão quantílica, mostrou-se adequada.

Desta forma, sem desconsiderar os essenciais modelos clássicos de regressão e suas extensões, e tendo em vista os objetivos da pesquisa em questão, sugere-se o uso da metodologia apresentada como uma alternativa mais robusta e complementar para auxiliar estudos de crescimento em geral, de outros animais ou plantas, bem como descrever dados com alta variabilidade ou assimetria.

5.0.1 Pesquisas futuras

Este estudo é uma motivação envolvendo a metodologia da regressão quantílica a dados de crescimento, sendo que estes, são classificados como longitudinais. Dessa forma, sugere-se a análise de outras estruturas de modelos, como não lineares para outros conjuntos, bem como estruturas que considerem a correlação e heterogeneidade de variâncias dentro dos indivíduos.

REFERÊNCIAS

- ARBELAITZ, O., I. GURRUTXAGA, J. MUGUERZA, J. M. PÉREZ, e I. PERONA, 2013 An extensive comparative study of cluster validity indices. *Pattern Recognition* **46**: 243–256.
- BARBOSA, A., P. CARNEIRO, M. REZENDE, I. RAMOS, R. MARTINS FILHO, e C. M. MALHADO, 2017 Parâmetros genéticos para características de crescimento e reprodutivas em bovinos nelore no brasil. *Archivos de zootecnia* **66**: 449–452.
- BARRODALE, I. e F. D. ROBERTS, 1973 An improved algorithm for discrete l_1 linear approximation. *SIAM Journal on Numerical Analysis* **10**: 839–848.
- BEYERLEIN, A., R. VON KRIES, A. R. NESS, e K. K. ONG, 2011 Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children. *Plos one* **6**: e19057.
- BOX, G. E. e D. R. COX, 1964 An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* **26**: 211–243.
- BUCHINSKY, M., 1998a The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of applied econometrics* **13**: 1–30.
- BUCHINSKY, M., 1998b Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources* pp. 88–126.
- CADE, B. S., J. W. TERRELL, e R. L. SCHROEDER, 1999 Estimating effects of limiting factors with regression quantiles. *Ecology* **80**: 311–323.
- CARVALHO, S. D. P. C., 2013 *Estimativa volumétrica por modelo misto e tecnologia laser aerotransportado em plantios clonais de Eucalyptus sp.* Ph.D. thesis, Universidade de São Paulo.
- CHEN, C., 2007 A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics* **16**: 136–164.
- CHEN, K., Z. YING, H. ZHANG, e L. ZHAO, 2008 Analysis of least absolute deviation. *Biometrika* **95**: 107–122.
- CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, e A. GALICHON, 2010 Quantile and probability curves without crossing. *Econometrica* **78**: 1093–1125.
- CRUZ, R. M. D., 2020 *Critérios de informação e seleção de modelos lineares mistos.* Ph.D. thesis, Universidade de São Paulo.

- DA SILVA, N. A. M., L. H. DE AQUINO, F. FONSECA, J. A. MUNIZ, *ET AL.*, 2006 Estudo de parâmetros de crescimento de bezerros nelore por meio de um modelo de regressão linear: uma abordagem bayesiana. *Ciência Animal Brasileira* **7**: 57–65.
- DAVINO, C., M. FURNO, e D. VISTOCCO, 2013 *Quantile regression: theory and applications*, volume 988. John Wiley & Sons.
- DE REZENDE, M., M. DA SILVEIRA, R. DA SILVA, L. DA SILVA, A. GONDO, G. RAMIRES, J. DE SOUZA, *ET AL.*, 2014 Pre and post weaning weight gain in nellore cattle raised in the pantanal, mato grosso do sul, brazil. *Ciência Animal* **24**: 20–27.
- DETTE, H. e S. VOLGUSHEV, 2008 Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**: 609–627.
- DUFRENOT, G., V. MIGNON, e C. TSANGARIDES, 2010 The trade-growth nexus in the developing countries: A quantile regression approach. *Review of World Economics* **146**: 731–761.
- DUNN, P. K. e G. K. SMYTH, 1996 Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**: 236–244.
- EFRON, B. e R. J. TIBSHIRANI, 1994 *An introduction to the bootstrap*. CRC press.
- FARIAS, A. A. *ET AL.*, 2018 Uso de regressão quantílica na predição da produção de povoamentos de eucalipto .
- FERGUSON, T. S., 2014 *Mathematical statistics: A decision theoretic approach*, volume 1. Academic press.
- FERNANDES, T. J., A. A. PEREIRA, J. A. MUNIZ, e T. V. SAVIAN, 2014 Seleção de modelos não lineares para a descrição das curvas de crescimento do fruto do cafeeiro .
- FITZENBERGER, B., R. KOENKER, e J. A. MACHADO, 2013 *Economic applications of quantile regression*. Springer Science & Business Media.
- FOX, M., H. RUBIN, *ET AL.*, 1964 Admissibility of quantile estimates of a single location parameter. *The Annals of Mathematical Statistics* **35**: 1019–1030.
- GERACI, M. e M. BOTTAI, 2007 Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics* **8**: 140–154.
- GERACI, M. e M. BOTTAI, 2014 Linear quantile mixed models. *Statistics and computing* **24**: 461–479.

- GOLDBARG, M. C. e H. P. L. LUNA, 2005 *Otimização combinatória e programação linear: modelos e algoritmos*. Elsevier.
- GOURDJI, S. M., K. L. MATHEWS, M. REYNOLDS, J. CROSSA, e D. B. LOBELL, 2013 An assessment of wheat yield sensitivity and breeding gains in hot environments. *Proceedings of the Royal Society B: Biological Sciences* **280**: 20122190.
- GUTENBRUNNER, C. e J. JURECKOVÁ, 1992 Regression rank scores and regression quantiles. *The Annals of Statistics* pp. 305–330.
- GUTENBRUNNER, C., J. JUREČKOVÁ, R. KOENKER, e S. PORTNOY, 1993 Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics* **2**: 307–331.
- HALL, P. e S. J. SHEATHER, 1988 On the distribution of a studentized quantile. *Journal of the Royal Statistical Society: Series B (Methodological)* **50**: 381–391.
- HALLIN, M., 2014 Gauss–markov theorem in statistics. *Wiley StatsRef: Statistics Reference Online* .
- HAO, L. e D. NAIMAN, 2007 *Quantile regression*, sage publication.
- HE, X. e F. HU, 2002 Markov chain marginal bootstrap. *Journal of the American Statistical Association* **97**: 783–795.
- HINOSTROZA, A. A. A., 2017 *Regressão quantílica bayesiana em modelos de fronteira de produção estocástica*. Ph.D. thesis, Universidade Federal de Rio de Janeiro.
- KOCHERGINSKY, M., X. HE, e Y. MU, 2005 Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics* **14**: 41–55.
- KOENKER, R., 2005 Cambridge university press; cambridge and new york: 2005. *Quantile regression: Econometric society monographs* .
- KOENKER, R., 2011 Quantreg: Quantile regression. r package version 4.71, 2011.
- KOENKER, R. e G. BASSETT JR, 1978 Regression quantiles. *Econometrica: journal of the Econometric Society* pp. 33–50.
- KOENKER, R. e J. A. MACHADO, 1999 Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association* **94**: 1296–1310.
- KOENKER, R. e P. NG, 2005 Inequality constrained quantile regression. *Sankhyā: The Indian Journal of Statistics* pp. 418–440.
- KOENKER, R. W. e V. D’OREY, 1987 Algorithm as 229: Computing regression quantiles. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **36**: 383–393.

- KOLMAN, B. e R. E. BECK, 1995 *Elementary linear programming with applications*. Gulf Professional Publishing.
- LAUREANO, M., A. BOLIGON, R. COSTA, S. FORNI, J. SEVERO, e L. G. D. ALBUQUERQUE, 2011 Estimativas de herdabilidade e tendências genéticas para características de crescimento e reprodutivas em bovinos da raça nelore: Estimates of heritability and genetic trends for growth and reproduction traits in nelore cattle. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia* **63**: 143–152.
- LEWIS-BECK, C. e M. LEWIS-BECK, 2015 *Applied regression: An introduction*, volume 22. Sage publications.
- LI, Q., R. XI, N. LIN, ET AL., 2010 Bayesian regularized quantile regression. *Bayesian Analysis* **5**: 533–556.
- LIMA JÚNIOR, M. E. S. D., 2015 As diferentes interpretações do algoritmo simplex na resolução de problemas de programação linear .
- LUCHE, J. R. D. e A. F. D. SILVA, 2016 Pesquisa operacional i - método simplex (parte i). São Paulo: UNIVESP .
- MONTGOMERY, D. C., E. A. PECK, e G. G. VINING, 2012 *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.
- MORALES, C. E. G., 2015 Quantile regression for mixed-effects models .
- MUGGEO, V. M., M. SCIANDRA, A. TOMASELLO, e S. CALVO, 2013 Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. *Environmental and ecological statistics* **20**: 519–531.
- NASCIMENTO, M., A. NASCIMENTO, J. DEKKERS, e N. SERÃO, 2019 Using quantile regression methodology to evaluate changes in the shape of growth curves in pigs selected for increased feed efficiency based on residual feed intake. *Animal* **13**: 1009–1019.
- OLIVEIRA, A. M. H. C. D. e E. L. G. RIOS-NETO, 2006 Tendências da desigualdade salarial para coortes de mulheres brancas e negras no brasil. *Estudos Econômicos (São Paulo)* **36**: 205–236.
- POLLICE, A., V. M. MUGGEO, F. TORRETTA, R. BOCHICCHIO, e M. AMATO, 2014 Growth curves of sorghum roots via quantile regression with p-splines. In *47th Scientific Meeting of the Italian Statistical Society*, Cooperativa Universitaria Editrice Cagliaritana.
- PORTNOY, S., R. KOENKER, ET AL., 1997 The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science* **12**: 279–300.

- PUIATTI, G. A., P. R. CECOM, M. NASCIMENTO, A. C. C. NASCIMENTO, A. P. S. CARNEIRO, M. PUIATTI, A. C. R. D. OLIVEIRA, *ET AL.*, 2018 Quantile regression of nonlinear models to describe different levels of dry matter accumulation in garlic plants. *Ciência Rural* **48**.
- R DEVELOPMENT CORE TEAM, 2020 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- RASTEIRO, L. R., 2017 *Regressão quantílica para dados censurados*. Ph.D. thesis, Universidade de São Paulo.
- REICH, B. J., H. D. BONDELL, e H. J. WANG, 2010 Flexible bayesian quantile regression for independent and clustered data. *Biostatistics* **11**: 337–352.
- RENCHE, A. C. e G. B. SCHAALE, 2008 *Linear models in statistics*. John Wiley & Sons.
- RIGBY, R. A. e D. M. STASINOPOULOS, 2005 Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**: 507–554.
- ROUSSEUW, P. J., 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**: 53–65.
- SANTOS, B. e H. BOLFARINE, 2018 Bayesian quantile regression analysis for continuous data with a discrete component at zero. *Statistical Modelling* **18**: 73–93.
- SANTOS, B. R. e S. N. ELIAN, 2015 Influence measures in quantile regression models. *Communications in Statistics-Theory and Methods* **44**: 1842–1853.
- SANTOS, B. R. D., 2012 *Modelos de regressão quantílica*. Ph.D. thesis, Universidade de São Paulo.
- SANTOS, P. M. D., A. C. C. NASCIMENTO, M. NASCIMENTO, C. F. AZEVEDO, R. R. MOTA, S. E. F. GUIMARÃES, P. S. LOPES, *ET AL.*, 2018 Use of regularized quantile regression to predict the genetic merit of pigs for asymmetric carcass traits. *Pesquisa Agropecuária Brasileira* **53**: 1011–1017.
- SCHNABEL, S. K. e P. H. EILERS, 2009 Optimal expectile smoothing. *Computational Statistics & Data Analysis* **53**: 4168–4177.
- SCHNABEL, S. K. e P. H. EILERS, 2013 Simultaneous estimation of quantile curves using quantile sheets. *AStA Advances in Statistical Analysis* **97**: 77–87.

- SEARLE, S. R. e M. H. GRUBER, 1971 *Linear models*, volume 10. Wiley Online Library.
- SHIM, J., C. HWANG, e K. H. SEOK, 2009 Non-crossing quantile regression via doubly penalized kernel machine. *Computational Statistics* **24**: 83–94.
- SILVEIRA, M., J. D. SOUZA, L. SILVA, J. FREITAS, A. GONDO, e P. FERRAZ FILHO, 2014 Interação genótipo x ambiente sobre características produtivas e reprodutivas de fêmeas nelore. *Archivos de zootecnia* **63**: 223–226.
- SINGER, J. M., F. M. ROCHA, e J. S. NOBRE, 2017 Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review* **85**: 290–324.
- SORRELL, B. K., C. C. TANNER, e H. BRIX, 2012 Regression analysis of growth responses to water depth in three wetland plant species. *AoB Plants* **2012**.
- TAKEUCHI, I., Q. LE, T. SEARS, A. SMOLA, *ET AL.*, 2006 Nonparametric quantile estimation .
- YU, K. e R. A. MOYEED, 2001 Bayesian quantile regression. *Statistics & Probability Letters* **54**: 437–447.
- YU, K. e J. ZHANG, 2005 A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics—Theory and Methods* **34**: 1867–1879.
- YUAN, Y. e G. YIN, 2010 Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics* **66**: 105–114.
- ZIETZ, J., E. N. ZIETZ, e G. S. SIRMANS, 2008 Determinants of house prices: a quantile regression approach. *The Journal of Real Estate Finance and Economics* **37**: 317–333.

APÊNDICES

Apêndice I - Estimativas dos parâmetros

Tabela A.1. Estimativas dos parâmetros para os modelos de regressão quantílica para diferentes quantis.

$\tau = 0,30$	Estimativa	Erro padrão	$\Pr(> t)$
β_0	31,0567	0,7384	<0,0001
β_1	0,9200	0,2545	0,0003
β_2	0,1008	0,0143	<0,0001
$\tau = 0,35$			
β_0	31,9000	0,6424	<0,0001
β_1	0,9437	0,2476	0,0002
β_2	0,1008	0,0131	<0,0001
$\tau = 0,40$			
β_0	32,0000	0,5828	<0,0001
β_1	1,0024	0,2440	<0,0001
β_2	0,1024	0,0127	<0,0001
$\tau = 0,45$			
β_0	32,0000	0,3909	<0,0001
β_1	1,2771	0,2276	<0,0001
β_2	0,0935	0,0108	<0,0001
$\tau = 0,50$			
β_0	32,7000	0,4211	<0,0001
β_1	1,3348	0,2427	<0,0001
β_2	0,0945	0,0113	<0,0001
$\tau = 0,60$			
β_0	32,7000	1,2112	<0,0001
β_1	1,9325	0,3703	<0,0001
β_2	0,0762	0,0169	<0,0001
$\tau = 0,70$			
β_0	35,0000	1,5472	<0,0001
β_1	2,0315	0,4057	<0,0001
β_2	0,0818	0,0184	<0,0001
$\tau = 0,80$			
β_0	37,2625	1,0903	<0,0001
β_1	2,0917	0,3576	<0,0001
β_2	0,0885	0,0164	<0,0001
$\tau = 0,90$			
β_0	38,5000	0,9770	<0,0001
β_1	2,6167	0,4395	<0,0001
β_2	0,0755	0,0219	0,0006

Tabela A.2. Estimativas dos parâmetros para o modelo de regressão quantílica do quantil condicional $\tau = 0,45$, considerando a variável dependente em meses.

$\tau = 0,45$	Estimativa	Erro padrão	$\Pr(> t)$
β_0	32,00	0,39	<0,0001
β_1	5,47	0,97	<0,0001
β_2	1,71	0,19	<0,0001

5.1 Apêndice II - Principais comandos utilizados no *Software R*

```
rm(list=ls(all=TRUE))
```

```
##Entrada dos dados
```

```
my_file <- read.table(file.choose(), sep=' ', dec=',')
```

```
dados <- my_file
```

```
dados2 <- dados[complete.cases(dados),]
```

```
##Análise descritiva
```

```
medias <- c()
```

```
desvios <- c()
```

```
cv <- c()
```

```
assimetria <- c()
```

```
for(i in 1:14){
```

```
  medias[i] <- mean(na.omit(pesos[seq(i, length(pesos), 14)]))
```

```
  desvios[i] <- sd(na.omit(pesos[seq(i, length(pesos), 14)]))
```

```
  cv[i] <- 100*desvios[i]/medias[i]
```

```
  assimetria[i] <- skewness(na.omit(pesos[seq(i, length(pesos), 14)]))
```

```
}
```

```
descritiva <- cbind(medias, desvios, assimetria, cv)
```

```
xtable(descritiva)
```

```
## Escolha modelo linear ou quadrático
```

```
## Ajuste quantis [0,10/ 0,90] e média
```

```
modelo1 <- rq(pesos ~ semanas, data=dados, tau=1:9/10)
```

```
modelo.mqo1 <- lm(pesos ~ semanas, data=dados)
```

```
modelo2 <- rq(pesos ~ 1+semanas+I(semanas^2),
```

```
      data=dados, tau=1:9/10)
```

```
modelo.mqo2 <- lm(pesos ~ 1+semanas+I(semanas^2),
```

```
      data=dados)
```

```
##Função para cálculo do R1
```

```
r1.rq <- function(modelo){
  if(class(modelo) != "rqs")
  {
    stop("Você deve usar essa função com objetos do tipo rqs")
  }
  taus <- modelo$tau
  rho.c <- modelo$rho
  methods <- modelo$method
  y <- modelo$y
  rho.r <- rq(y~1, tau=taus, method=methods)$rho
  R1 <- 1-rho.c/rho.r
  saida <- data.frame(taus, R1)
  return(saida)
}
```

```
##R1 e R2
```

```
r1.rq(modelo1)
r1.rq(modelo2)
```

```
summary(modelo.mqo1)
summary(modelo.mqo2)
```

```
##AIC
```

```
AIC(modelo.mqo1, modelo.mqo2)
AIC.rq(modelo1)
AIC.rq(modelo2)
```

```
##Teste de hipótese
```

```
taus <- 1:9/10
teste <- list()
for(i in 1:length(taus)){
  modelo1 <- rq(pesos ~ semanas, data=dados, tau=taus[i])
  modelo2 <- rq(pesos ~ 1+semanas+I(semanas^2),
                data=dados, tau=taus[i])
```

```
teste[[i]] <- anova.rq(modelo1, modelo2)
}
```

##Gráfico intervalos de confiança

```
model3 = rq (pesos ~ 1+semanas+I(semanas^2), data = dados,
tau = seq (0.05,0.99, by = 0.05))
quantilplot <- summary(model3)
plot(quantilplot, level=0.95, col=c("blue", "gray"), cex.lab=1.3)
```

Função de distribuição acumulada laplace assimetrica

```
pdla <- function (q, mu=0, sigma=1, tau=0.5){
  acumulada <- vector(length=max(length(q), length(mu)))
  if (length(q) == 0)
    stop("q_deve_ser_fornecido.")
  if (tau >= 1 | tau <= 0)
    stop("tau_deve_ser_um_número_real_entre_(0,1).")
  if (sigma <= 0)
    stop("sigma_deve_ser_um_número_positivo.")
  if (length(q) != length(mu))
  {
    for (k in 1:length(q)){
      if ( ! q[k] > mu)
        acumulada[k] <- tau*exp((1/sigma)*(1-tau)*(q[k]-mu))
      else
        acumulada[k] <- 1-(1-tau)*exp(-(tau/sigma)*(q[k]-mu))
    }
  }
  else
  {
    for (k in 1:length(q))
    {
      if ( ! q[k] > mu[k])
        acumulada[k] <- tau*exp((1/sigma)*(1-tau)*(q[k]-mu[k]))
      else
        acumulada[k] <- 1-(1-tau)*exp(-(tau/sigma)*(q[k]-mu[k]))
    }
  }
}
```



```

}
return( acumulada)
}

```

#Função para gerar números aleatorios laplace assimétrica

```

rdla <- function (n, mu=0, sigma =1, tau =0.5){
  if (sigma <= 0)
    stop("sigma_deve_ser_um_número_positivo")
  if (length(n) == 0)
    stop("O_tamanho_da_amostra_n_deve_ser_fornecido.")
  if (tau >= 1 | tau <= 0)
    stop("tau_deve_ser_um_número_real_entre_(0,1).")
  r <- vector(length=n)
  if (length(mu)==1) mu = rep(mu, n)
  if (length(mu) !=n ) stop ("mu")
  for (k in 1:n){
    u1 <- rexp(1)
    u2 <- rexp(1)
    r[k] <- mu[k] + sigma*(u1/tau - u2/(1-tau))
  }
  return (r)
}

```

Resíduo quantílico

```

res.quantilico <- function(modelo){
  tau <- modelo$tau
  rho <- modelo$rho
  n <- length(residuals(modelo))
  preditos <- fitted(modelo)
  res.quant = qnorm(pdla(q=as.numeric(modelo$y), mu=preditos,
sigma=modelo$rho/n, tau=tau))
  return(res.quant)
}

```

```

#Cálculo AIC, erro quadrático médio(EQM) e erro absoluto médio(EAM)

taus = c(0.30,0.35,0.40,0.45,0.50) #quantis de interesse

##Erro absoluto médio

absoluto <- c()
for(i in 1:length(taus)){
  mod0 <- rq(pesos~1+semanas+I(semanas^2), data=dados2, tau=taus[i])
  diferenca <- abs(mod0$fitted.values-na.omit(pesos))
  soma <- sum(diferenca)/length(na.omit(pesos))
  absoluto [i] <- soma
}

##Erro quadrático médio

desvio_quad <- c()
for(i in 1:length(taus) ){
  mod0 <- rq(pesos~1+semanas+I(semanas^2), data=dados2, tau=taus[i])
  diferenca <- mod0$fitted.values-na.omit(pesos)
  modulo <- (diferenca)^2
  soma <- sum(modulo)/(length(na.omit(na.omit(pesos))))
  desvio_quad[i] <- soma
}

##AIC

model0 <- rq(pesos~1+semanas+I(semanas^2), data=dados, tau=taus)
aic <- AIC.rq(model0)

##Estimativas e suas respectivas medidas associadas à inferência

tau045 <- rq(pesos ~ 1 + semanas + I(semanas^2), tau=0.45)
tau010 <- rq(pesos ~ 1 + semanas + I(semanas^2), tau=0.10)
tau090 <- rq(pesos ~ 1 + semanas + I(semanas^2), tau=0.90)
resumo2 <- summary.rq(tau045, se='nid')
resumo1 <- summary.rq(tau010, se='nid')
resumo3 <- summary.rq(tau090, se='nid')

```

```
## Taxa de variação de cada ajuste
```

```
##tau 10
```

```
func10 <- function(x){28.15974026+1.29675325*x+0.04707792*x^2}
t1 <- (func10(26)-func10(0))/26 ;t1
round(t1,2)
```

```
##tau 45
```

```
func45 <- function(x){32.00+(1.27708333*x) + (0.09348958 *(x^2))}
t2 <- (func45(26)-func45(0))/26 ;t2
round(t2,2)
```

```
##tau 90
```

```
func90 <- function(x){38.50000000+(2.61666667*x)+(0.07552083*(x^2))}
t3 <- (func90(26)-func90(0))/26 ;t3
round(t3,2)
```

```
##Taxas observadas
```

```
bezerros.ind <- split(pesos, ceiling(seq_along(pesos)/14))
semanas2 <- c(0,2,4,6,8,10,12,14,16,18,20,22,24,26)
```

```
taxas_ind <- c()
for(i in 1:28){
  betas <- na.omit(bezerros.ind[[i]])
  min <- betas[1]
  n <- length(betas)
  max <- betas[n]
  taxa <- (max-min)/26
  taxas_ind[i] <- taxa
}
```

```
#taxa do bezerro 2
```

```
betas <- na.omit(bezerros.ind[[2]])
min <- betas[1]
n <- length(betas)
```

```

max <- betas[n]
taxa <- (max-min)/24
taxas_ind[2] <- taxa

##Agrupando pela distância euclidiana das taxas

dist1 <- abs(taxas_ind-t1)
dist2 <- abs(taxas_ind-t2)
dist3 <- abs(taxas_ind-t3)
distancias <- data.frame(dist1 , dist2 , dist3)

clas1 <- c()
for(i in 1:28){
  if (min(distancias[i,]) == distancias[i,1]){
    clas1[i] <- 'tau_10'
  } else if (min(distancias[i,]) == distancias[i,2]){
    clas1[i] <- 'tau_45'
  } else {
    clas1[i] <- 'tau_90'
  }
}

tab12 <- data.frame(taxas_ind , distancias , clas1)
xtable(tab12)

##Cálculo Silhouette coefficient

clas2 <- c()
for(i in 1:28){
  if (min(distancias[i,]) == distancias[i,1]){
    clas2[i] <- 1
  } else if (min(distancias[i,]) == distancias[i,2]){
    clas2[i] <- 2
  } else {
    clas2[i] <- 3
  }
}

```

```
taxas_ind3 <- scale(taxas_ind)
sil <- silhouette(clas2, dist(taxas_ind3))
grafico <- fviz_silhouette(sil)
grafico$labels$title

neg_sil_index <- which(sil[, "sil_width"] < 0)
sil[neg_sil_index, , drop = FALSE] #valores negativos
```

ANEXOS

Anexo A

Dados de crescimento de bezerros

Tabela A.3. Dados de crescimento de bezerros (pesos em kg) utilizados como estudo de motivação.

Bezerros	Semanas													
	0	2	4	6	8	10	12	14	16	18	20	22	24	26
1	28.00	32.70	36.90	41.70	42.80	50.30	53.50	48.20	54.80	61.00	64.40	74.20	85.40	93.70
2			45.20	44.90	47.80	55.50	57.30	64.00	63.70	65.50	70.70	79.20	82.70	91.40
3	31.00	30.60	32.50	32.50	35.20	40.50	47.10	56.20	68.30	80.00	93.00	105.00	118.00	130.40
4	36.00	39.70	42.10	47.20	49.20	51.90	60.80	69.80	81.00	89.50	97.00	105.40	115.60	123.50
5	29.00	32.90	36.60	38.00	44.80	54.20	58.10	65.80	73.20	81.50	92.00	103.70	112.60	123.30
6	28.00	35.10	32.90	41.10	52.00	50.10	51.30	56.80	64.90	72.90	79.00	83.40	93.20	102.00
7	40.00	42.00	46.80	52.10	54.90	65.20				104.70	110.60	124.70	136.80	153.40
8	42.00	45.80		53.50	59.20	71.40	71.20	74.90	83.50	88.50		102.70	111.60	119.70
9	30.00	34.00	35.40	35.90	38.00	48.90	50.00	62.10	70.50	83.00	91.00	103.90	116.50	127.00
10	43.00	43.70	48.20	47.40	48.20	54.70	58.00	62.70	77.40	82.80	97.10	107.10	115.30	127.00
11	31.00	31.60	33.40	40.90	45.90	46.50	55.00	59.80	66.20	76.90	85.80	93.00	103.30	112.50
12	27.00	31.50	38.00	42.60	47.00	56.70	64.80	74.70	84.10	94.90	106.60	122.30	130.60	141.50
13	32.00	35.90	42.50	48.90	54.50	67.80	71.20	84.70	93.90	102.20	116.00	126.20	138.20	150.00
14	34.00	34.40	39.10	41.30	43.30	51.10	54.10	65.40	75.10	87.90	98.00	108.70	124.20	131.30
15	41.00	42.40	47.50	58.30	69.50	79.40	91.70	104.00	119.20	133.10	145.40	156.50	167.20	176.80
16	31.90	34.10	39.70	46.30	50.00	58.70	64.20	81.00	91.50	106.90	117.10	127.70	144.20	154.90
17	40.00	41.80	50.00	55.50	58.60	63.50	70.30	80.00	90.00	102.60	101.20	120.40	130.90	137.10
18	31.00	30.70	34.10	44.30	47.80	59.40	68.30	77.20	84.20	96.20	104.10	114.00	123.00	132.00
19	35.00	38.60	39.70	40.20	37.10	43.10	43.90	48.10	58.30	68.60	78.50	86.80	99.90	106.20
20	33.00	40.70	46.30	55.90	67.20	74.60	87.40	95.40	110.50	122.50	127.00	136.30	144.80	151.50
21	37.00	36.60								103.60	110.70	120.00	126.70	132.20
22	36.00	33.30	35.00	39.90	44.50	46.50	50.50	55.00	59.10	68.90	78.20	75.10	79.00	77.00
23	30.00	35.80	41.20	52.00	64.00	74.90	91.00	95.50				148.00	154.50	167.60
24	38.50	44.50	47.10	54.30	60.90	74.90	83.30	89.70	99.70	110.00	120.80	135.10	141.50	157.00
25	30.00	37.20	43.90	49.00	52.80	60.20	76.30	80.80	94.20	102.60	111.00	115.60	121.40	134.50
26	32.70	35.90		40.20		52.50		61.10		80.90	93.00	100.10	103.20	108.00
27	32.00	35.50	39.20	43.70	51.80	63.40	76.10	81.10	84.60	89.80	97.40	111.00	120.20	134.20
28	32.00	34.70	36.10	39.50	45.20	53.00	56.60	63.70	70.10	74.40	85.10	90.20	96.10	103.60