Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

MODELAGEM DO PREÇO DE LEITE VIA REGRESSÃO LINEAR MÚLTIPLA EM UMA COOPERATIVA DE PRODUTORES

Weila Silva Freitas

Uberlândia-MG
2019

Weila Silva Freitas

MODELAGEM DO PREÇO DE LEITE VIA REGRESSÃO LINEAR MÚLTIPLA EM UMA COOPERATIVA DE PRODUTORES

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

Orientadora: Profª Dra. Mirian Fernandes Carvalho Araújo

Uberlândia-MG

2019

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

Uberlândia, ____ de _____ de 20____

Profª Dra. Mirian Fernandes Carvalho Araújo

Profª Dra. Aurélia Aparecida de Araújo Rodrigues

Profº Dr. José Waldemar da Silva

Uberlândia-MG

2019

Agradecimentos

Comemoro mais esta vitória. Foi preciso muito esforço, determinação, paciência e ousadia para chegar até aqui, não conseguiria nada disso sozinha. Em primeiro lugar sou grata a Deus pelo dom da vida, pelo seu amor infinito e pelo discernimento concedido ao longo dessa jornada. Aos meus pais, que são meus maiores exemplos, me orientando e orando por mim, sempre preocupados, verdadeiros heróis que com todo apoio e incentivo nas horas difíceis, me inspiram sempre a sonhar e lutar pelo que quero. Ao meu filho e ao meu irmão, que com tanto amor me fazem ter motivos para ser uma "grande pessoa". Ao meu companheiro, parceiro em tantos momentos difíceis. Agradeço o apoio de todos os familiares e amigos, sempre superando a distância. Minha eterna gratidão aos meus amigos, colegas, professores da graduação e orientadores que tive, que de alguma forma contribuíram para que este sonho pudesse se concretizar, com eles aprendi amar e construir laços eternos. Pelos momentos de dedicação aos estudos, brincadeiras, pelas trocas de conhecimento e experiências que foram tão importantes na minha vida e que contribuíram para o meu novo olhar profissional. Aos vários amigos que encontrei na caminhada, obrigada pela paciência, pelo sorriso, pela mão estendida, pois essa caminhada não seria a mesma sem vocês.

Resumo

No Brasil a pecuária já foi representada como uma atividade complementar nas fazendas em que o uso dos animais era basicamente como tração nos engenhos. Com o seu desenvolvimento ao longo do tempo notou-se uma expansão dessa atividade comercial com a conquista de novos mercados se tornando uma das grandes atividades industriais responsáveis pela sustentação econômica brasileira. Com essa perspectiva, o objetivo deste trabalho é encontrar um modelo linear múltiplo para estimar a melhor representação para o preço atual do leite (em reais) dos cooperados da COFRUL - Cooperativa Mista dos Produtores Rurais de Frutal LTDA, sendo situada na cidade de Frutal (MG), no Triângulo Mineiro. Além disso, a pesquisa pretende propiciar à empresa que recebe o leite uma estimativa dos períodos de maior e menor influência no preço atual, observando o preço pago dentro do estado de Minas Gerais e total de leite entregue (em litros/mês). Os meses que possuem os menores preços pago ao produtor são: outubro, janeiro e dezembro, e os meses com maiores valores são março e junho. E quanto maior for o total de leite produzido em litros/mês, maior será o preço atual do leite pago ao produtor em reais/litro.

Palavras-chave: Estimativa, Máximo, Mínimo, Variáveis Dummy.

Abstract

In Brazil, livestock has already been represented as a complementary activity on farms where the use of animals was basically traction on the mills. With its development over time, it was noted an expansion of this commercial activity with the conquest of new markets becoming one of the major industrial activities responsible for the Brazilian economic support. With this perspective, the objective of this work is to find a multiple linear model to estimate the best representation for the current price of milk (in reais) of the cooperative members of COFRUL - Cooperativa Mista dos Produtores Rurais de Frutal LTDA, being located in the city of Frutal (MG), in the Triângulo Mineiro. In addition, the research aims to provide the company that receives milk with an estimate of the periods of major and minor influence on the current price, observing the price paid within the state of Minas Gerais and total milk delivered (in liters/month). The months with the lowest prices paid to the producer are: October, January and December, and the months with the highest values are March and June. And the higher the total milk produced in liters/month, the higher the current price of milk paid to the producer in reais/liter.

Keywords: Estimate, Maximum, Minimum, Dummy Variables.

Sumário

Lista de Tabelas

1. Introdução

Por volta do ano de 1532 iniciava uma nova atividade econômica nacional, a produção de leite, que a longo prazo se tornaria uma das atividades agroindustriais mais importantes do país. Essa atividade foi desenvolvida por pequenos e até mesmo por grandes produtores rurais. Os grandes produtores foram caraterizados por possuírem grandes hectares de terra, presença de pouca mão de obra, alta produção com uso de técnicas avançadas em vários setores, rodízio de cultivos e criação de gado leiteiro e de corte. Diferentemente do grande produtor rural, as características observadas do pequeno produtor foram: pequena propriedade para usufruto familiar, plantio de hortaliças, criação de aves e ausência de tecnologias [9].

No começo dessa atividade econômica os recursos eram poucos desde a coleta do leite como por exemplo, a falta de higienização e o transporte adequado do leite até os laticínios. A partir desses fatos em 1950, o governo federal começou a exigir melhores condições sanitárias por parte dos produtores e houve a implementação de fiscalizações para melhor comercialização do leite [10].

Para que o leite tenha condições de ser comercializado ele deve ter origem de animais saudáveis, boas condições de armazenamento e transporte até os laticínios. É de extrema importância o controle efetivo de doenças como por exemplo, a mastite, que é uma inflamação das glândulas mamárias [3].

Atualmente, o cenário da produção leiteira no Brasil possui resoluções com exigências do governo federal que beneficia os grandes produtores e causa um detrimento do pequeno produtor que, por falta de recursos acompanham com muita dificuldade o crescimento da produção leiteira. Devido a isso o pequeno produtor não se mantém dentro da escala produtiva, pois só se mantêm no mercado os produtores que conseguem reduzir os seus custos e aumentam sua produção. Com intuito de acompanharem essa progressão estatística, os pequenos produtores se sujeitam a uma rotina desgastante de trabalho para a melhoria em sua pequena produção, dentre eles pode-se citar algumas como: preservação das pastagens com devido enriquecimento nutricional do solo, compra de rações que estimulam maior produtividade das vacas leiteiras, cultivo de culturas que servirão como alimento nutritivo em período de secas, ordenhamento e vacinação contra possíveis doenças que possam acometer esse rebanho.

Para aumentar seus ganhos e não ser uma atividade sem lucros esses pequenos produtores se juntaram e formaram as associações e cooperativas [1]. Nessas associações os pequenos produtores lutam para aumentar sua visibilidade no mercado de trabalho, aumentando o valor de seus produtos e ganhando espaço para uma concorrência mais justa em relação aos grandes

produtores.

A união entre os pequenos produtores gera condições de compartilhamento de equipamentos como tratores, colheitadeiras, caminhões para transporte. Esses recursos fazem com que os associados saiam lucrando, pelo esforço e pelo benefício comum. A associação fortalece os laços de amizade e solidariedade, luta pela melhoria da comunidade, defende os interesses dos associados melhorando a qualidade de vida, dessa forma produzem e comercializam de forma cooperada.

As cooperativas são uma forma de associativismo que possuem o objetivo de melhorar o preço da produção dos pequenos produtores aumentando sua renda e melhorando sua representatividade. Os agricultores formam as cooperativas com o intuito de colocar seus produtos no mercado, através delas conseguem financiamentos e conquistam materiais como tratores, sementes, colheitadeira, entre outros, também conseguem créditos de tal forma que se colocam numa classe com um lugar representativo no meio dos grandes produtores rurais. Com uma cooperativa os pequenos produtores tornam-se reconhecidos e ganham seu espaço.

Com essa perspectiva, o objetivo deste trabalho é encontrar um modelo linear múltiplo para estimar a melhor representação para o preço atual do leite (em reais) dos cooperados da COFRUL - Cooperativa Mista dos Produtores Rurais de Frutal LTDA, sendo situada na cidade de Frutal (MG), no Triângulo Mineiro. Além disso, a pesquisa pretende propiciar à empresa que recebe o leite uma estimativa dos períodos de maior e menor influência no preço atual, observando o preço pago dentro do estado de Minas Gerais [2] e total de leite entregue (em litros/mês).

Em um estudo [14], utilizando modelo de regressão linear múltipla sobre o preço do leite com informações provenientes de 31 produtores, pertencente a um laticínio do Estado de São Paulo, tendo como objetivo avaliar a relação das variáveis estrato sólido total (EST), contagem de célula somática (CCS) e contagem bacteriana total (CBT) com o preço do leite cru pago ao produtor, os autores concluíram que a quantidade de leite entregue não foi significativa para explicar o preço do leite e essas variáveis foram significativas para explicar o preço do leite.

Em outro artigo [16], utilizando modelos de séries temporais, buscou identificar variações sazonais nos preços recebidos pelos produtores de leite nos estados de Minas Gerais e Bahia, no período de janeiro de 2000 a abril de 2013. Realizou-se previsões futuras dos preços recebidos pelos produtores de leite e pelo modelo X-12 ARIMA, identificou-se sazonalidade presente nas séries, e a partir disso, utilizou-se modelos SARIMA, concluiu que o preço do leite possui sazonalidade e que durante os meses de seca (de maio a setembro) possuia tendência de serem maiores. Em relação aos períodos de chuvas, os preços são mais baixos tendo seus menores pontos no período de dezembro a fevereiro. Portanto há uma certa concordância entre [16] e esse presente estudo, que observou uma variação nos preços em função do clima.

2. Metodologia

2.1 Descrição dos Dados

Os dados foram obtidos via cópias dos relatórios de captação de leite dos cooperados da COFRUL - Cooperativa Mista dos Produtores Rurais de Frutal LTDA, sendo esta situada na cidade de Frutal - MG, no Triângulo Mineiro, referente aos meses de Junho de 2017 a Maio de 2018. Sendo assim, os dados foram organizados em planilhas do excel com as seguintes informações mensais dos produtores cooperados: total de leite (TL) produzido em litros/mês (l/m) é uma variável explicativa e preço atual do leite (PA) pago ao produtor em reais/litro (R$/l) é a variável resposta. A estas duas variáveis, acrescentou-se as informações dos meses em que os dados foram coletados e o preço médio do leite (PMG) no estado de Minas Gerais (dado em R$/l) obtidos no site do CEPEA – ESALQ/USP [2]. O banco de dados é composto por 3300 observações.

2.2 Regressão Linear Múltipla

Inicialmente, realizou-se uma análise estatística descritiva dos dados para compreender o comportamento individual de cada variável em estudo. As variáveis utilizadas para construção do modelo de regressão linear múltipla foram: PA, TL, PMG e os meses de coleta das informações.

Uma regressão linear múltipla [15] é uma técnica multivariada com a finalidade principal de ter uma relação matemática entre uma das variáveis estudadas Y (variável dependente ou resposta) e o restante das variáveis que descrevem o sistema (variáveis independentes, explicativas ou regressoras), e reduzir um grande número de variáveis para poucas dimensões com o mínimo de perda de informação, permitindo a detecção dos principais padrões de similaridade, associação e correlação entre as variáveis. Sua principal aplicação, após encontrar a relação matemática, é produzir valores para a variável dependente quando se têm as variáveis regressoras (cálculo dos valores preditos).

O modelo estatístico de uma regressão linear múltipla com k variáveis regressoras e n observações é expresso por:

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_k X_{kj} + \varepsilon_j \tag{2.1}$$

em que:

Y_j representa a variável resposta, para $j = 1, 2, ..., n$;

X_{ij} representa as variáveis regressoras, para $i = 0, 1, ..., k$;

β_i parâmetros do modelo (ou coeficientes de regressão) a serem estimados;

ε_j são os erros aleatórios do modelo supostos independentes e normalmente distribuídos de média zero e variância σ^2.

O valor esperado na resposta $\mathbf{Y}$ é função linear dos parâmetros β_i por unidade de variação em X_i, quando todas as outras variáveis explicativas forem mantidas constantes.

Em notação matricial, considerando o modelo (2.1), a regressão linear múltipla fica:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

com:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & ... & X_{k1} \\ 1 & X_{12} & X_{22} & ... & X_{k2} \\ ... & ... & ... & ... & ... \\ 1 & X_{1n} & X_{2n} & ... & X_{kn} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix} \text{ e } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

ou ainda,

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & ... & X_{k1} \\ 1 & X_{12} & X_{22} & ... & X_{k2} \\ ... & ... & ... & ... & ... \\ 1 & X_{1n} & X_{2n} & ... & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + ... + \beta_k X_{k1} + \varepsilon_1 \\ \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + ... + \beta_k X_{k2} + \varepsilon_2 \\ ... \\ \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + ... + \beta_k X_{kn} + \varepsilon_n \end{bmatrix}.$$

De forma semelhante em regressão linear simples, têm-se as suposições:

(i) a variável resposta Y é função linear das variáveis explicativas $X_j, j = 1, 2, ..., k$;

(ii) as variáveis explicativas X_j são fixas;

(iii) $E(\varepsilon_i) = 0$, ou seja, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, sendo $\mathbf{0}$ um vetor de zeros de dimensões $n\ x\ 1$;

(iv) os erros são homocedásticos, isto é, $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$;

(v) os erros são independentes, isto é, $Cov(\varepsilon_i, \varepsilon_{i'}) = E(\varepsilon_i \varepsilon_{i'}) = 0, i \neq i'$;

(vi) os erros têm distribuição normal, $\mathrm{N}(0, \sigma^2)$.

Logo, combinando-se (iv) e (v) tem-se $Var(\varepsilon) = E(\varepsilon\varepsilon^T) = \boldsymbol{I}\sigma^2$, sendo $\boldsymbol{I}$ uma matriz identidade, de dimensões $n\ x\ n$. Portanto, considerando-se, também, (vi) tem-se $\varepsilon \sim N(0, \boldsymbol{I}\sigma^2)$ e $\mathbf{Y} \sim N(\mathbf{X}\beta, \boldsymbol{I}\sigma^2)$, pois, $E(\mathbf{Y}) = \mu = \mathbf{X}\beta$ e $Var(\mathbf{Y}) = Var(\varepsilon) = \boldsymbol{I}\sigma^2$. A suposição de normalidade é necessária para a elaboração dos testes de hipóteses e obtenção de intervalos de confiança.

Os parâmetros podem ser estimados por vários métodos, sendo que o mais utilizado é o método de mínimos quadrados (MMQ) e neste estudo a estimação dos parâmetros do modelo de regressão linear múltipla foi realizado pelo MMQ.

O MMQ fundamenta-se em minimizar o erro quadrático médio das medidas, ou seja, consiste em encontrar estimadores para os parâmetros de forma que a soma dos quadrados dos desvios entre os valores estimados pelo modelo e os valores observados seja a menor possível.

Minimiza-se a expressão $\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik})]^2 = (Y - X\beta)'(Y - X\beta)$. Deriva-se simultaneamente em termos de $\boldsymbol{\beta}$, e obtém [11]:

$$\frac{\partial Y'Y}{\partial \beta} - 2\frac{\partial Y'X\beta}{\partial \beta} + \frac{\partial \beta'X'X\beta}{\partial \beta} = 0 - 2(Y'X)' + 2X'X\beta, \tag{2.3}$$

no qual o vetor de dimensão $p + 1$ resulta na expressão $(X'X)^{-1}X'Y = \widehat{\boldsymbol{\beta}}$, em que $\widehat{\boldsymbol{\beta}}$ é o estimador não-viciado do modelo, desde que $(X'X)^{-1}$ exista.

Portanto, o modelo de regressão ajustado e o vetor de resíduos são dados, respectivamente, por:

$$\widehat{Y} = X\widehat{\beta} \text{ e } \varepsilon = Y - \widehat{Y} = Y - X\widehat{\beta}. \tag{2.4}$$

O estimador de σ^2 é dado matricialmente por:

$$QMRes = \frac{Y'Y - \widehat{\beta}'X'Y}{n - p - 1}. \tag{2.5}$$

e a matriz de covariância de $\widehat{\beta}$ é dada pela fórmula:

$$Cov(\widehat{\beta}) = \sigma^2(X'X)^{-1} \tag{2.6}$$

Pode-se também calcular o coeficiente de determinação múltipla R^2, que é um critério utilizado para analisar e comparar os modelos e representa a proporção da variação explicada pelo modelo de regressão, isto é, uma medida de qualidade de ajuste do modelo aos dados sendo calculada por:

$$R^2 = SQReg/SQT = 1 - SQRes/SQT, \tag{2.7}$$

em que:
SQReg é a soma de quadrados da regressão;
SQT é a soma de quadrados total;
SQRes é a soma de quadrados dos erros.

O valor de R^2 tem variação entre 0 e 1, quanto mais próximo de 1 melhor será o ajuste do modelo, ou seja, quanto maior R^2, mais a variação total de Y é explicada pela variáveis explicativas.

Na regressão linear múltipla o coeficiente de determinação tende a aumentar à medida que mais variáveis explicativas são adicionadas ao modelo, independente da variável adicional ser ou não estatisticamente significativa. Este fato leva a um coeficiente que não mede mais a real explicação da variável resposta. Existe uma medida que corrige o coeficiente de determinação pela quantidade de variáveis independentes do modelo, denominado coeficiente de determinação ajustado expresso por:

$$R^2_{aj} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2), \tag{2.8}$$

com $p = k + 1$, isto é, p indica o número de variáveis explicativas mais a constante.

Um aspecto importante para a validação de um ajuste de regressão linear múltipla é a análise de resíduos, que mostra a significância do modelo e estima as contribuições das variáveis regressoras.

Realiza-se a análise de variância com o objetivo de comparar os modelos e estimar a significância da regressão. Considerando o modelo de regressão linear múltipla, definido em (2.2), pode-se construir a tabela ANOVA (análise de variância, corrigida pela média), dada por:

TABELA 2.1: Esquema da ANOVA para o modelo de regressão com p variáveis.

Fonte de variação	Graus de liberdade	Soma de Quadrados	Quadrados Médios	F_{calc}
Regressão	p	SQReg	QMReg	QMRes/QMRes
Erro	$n-p-1$	SQRes	QMRes	
Total	$n-1$	SQT		

F_{calc}:F calculado

em que:

$$\begin{aligned} SQT &= Y'Y - n\overline{y}^2 \\ SQReg &= Y'HY - n\overline{y}^2 \\ SQRes &= Y'Y - \widehat{\beta}'X'Y = SQT - SQReg \\ QMReg &= SQRes/p \\ QMRes &= SQRes/(n-p-1) \end{aligned} \tag{2.9}$$

Na tabela (2.1), as hipóteses testadas são:

$$\begin{aligned} H_0 &: \beta_1 = ... = \beta_p = 0 \\ H_1 &: \text{pelo menos um dos } \beta_s \neq 0 \end{aligned} \tag{2.10}$$

p é o número de variáveis regressoras.

Assim, se a hipótese H_0 for a verdadeira, o modelo não está bem ajustado, ou não há relação da variável resposta com as variáveis preditoras, pois os coeficientes são estatisticamente iguais a zero.

2.3 Análise de Resíduos

Tendo o modelo linear $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, os elementos ε_i do vetor $\boldsymbol{\varepsilon}$ são as diferenças entre os valores observados ($Y_i's$) e aqueles esperados pelo modelo. São chamados de erros, representam a variação dos dados e assume-se que os $\varepsilon_i's$ são independentes e, além disso, $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2)$, isto é, comportam-se como especificado pelas pressuposições. Contudo, nem sempre é o caso e, se as suposições são violadas, têm-se as falhas sistemáticas (não linearidade, não normalidade, heterocedasticidade, não independência dos erros, efeito cumulativo de fatores que não foram consideradas no modelo, etc) e a análise resultante pode levar a conclusões duvidosas. Outro fato bastante comum é a presença de pontos atípicos falhas isoladas), que podem influenciar, ou não, o ajuste do modelo. Elas podem surgir devido a [15]:

- erros grosseiros na variável resposta ou nas variáveis explanatórias, por medidas erradas ou registro da observação, ou ainda, erros de transcrição;

- observação proveniente de uma condição distinta das demais;

- modelo mal especificado (falta de uma ou mais variáveis, modelo inadequado etc);

- escala errada, talvez os dados sejam melhor descritos após uma transformação, do tipo logarítmica ou raiz quadrada;

- distribuição da variável resposta errada, por exemplo, tem uma cauda mais longa do que a distribuição normal.

Em um conjunto de dados ajustando um determinado modelo, para a verificação das pressuposições devem ser considerados como material básico: os valores estimados (ou ajustados), $\widehat{\mu}_i = \widehat{Y}_i$, os resíduos, $r_i = Y_i - \widehat{\mu}_i$, a variância residual estimada, $\widehat{\sigma}^2 = s^2 = QMRes$, e os elementos da diagonal (*leverage*) da matriz de projeção $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ e os valores da diagonal principal são denominadas "valores h", a qual é matriz simétrica e idempotente.

Uma característica muito importante da matriz $\mathbf{H}$ [5] é inerente aos elementos da sua diagonal, em que o elemento h_{ii} (diagonal de $\mathbf{H}$, $i = 1, ..., n$) mede o quão distante a observação y_i

está das demais $(n-1)$ observações no espaço definido pelas variáveis explicativas, isto é, da matriz $\mathbf{X}$ e não envolve as observações em $\mathbf{Y}$.

O elemento h_{ii} representa uma medida de alavancagem da i-ésima observação. Se $h_{ii} \geq 2p/n$, os valores das variáveis explicativas associados a i-ésima observação são atípicos, ou seja, estão distantes do vetor de valores médios das variáveis explicativas. Uma observação com $h_{ii} \geq 2p/n$, poderá ter influência na determinação dos coeficientes da regressão [6].

Um conceito importante é a da deleção (*deletion*), isto é, a comparação do ajuste do modelo escolhido, considerando-se todos os pontos, com o ajuste do mesmo modelo sem os pontos atípicos. As estatísticas obtidas pela omissão de um certo ponto i são denotadas com um índice entre parênteses. Assim, por exemplo, $s^2_{(i)}$ representa a variância residual estimada para o modelo ajustado, excluído o ponto i.

As técnicas usadas para a verificação do ajuste de um modelo a um conjunto de dados podem ser formais e informais. As informais baseiam-se em exames visuais de gráficos para a detecção de padrões, ou então, de pontos discrepantes. As formais envolvem aninhar o modelo sob pesquisa em uma classe maior pela inclusão de um parâmetro (ou vetor de parâmetros) extra. As mais usadas são baseadas nos testes da razão de verossimilhanças e escore. Parâmetros extras podem aparecer devido a:

- inclusão de uma covariável adicional;

- aninhamento de uma covariável X em uma família indexada por um parâmetro γ, sendo um exemplo a família de Box-Cox;

- inclusão de uma variável construída;

- inclusão de uma variável dummy tomando o valor 0 (zero) para a(s) unidade(s) discrepante(s) e 1 (um) para as demais. Isso é equivalente a eliminar essa observação do conjunto de dados, fazer a análise com a(s) observação(ões) discrepante(s) e sem ela(s) e verificar se a mudança no valor SQRes é significativa, ou não. Depende, porém, de localizar o(s) ponto(s) discrepante(s).

Tem-se um excelente método [8] para detectar observações influentes que é a medida da distância desenvolvida por Dennis R. Cook. A distância de Cook é a medida da distância ao quadrado entre a estimativa usual do MMQ de $\boldsymbol{\beta}(\widehat{\boldsymbol{\beta}})$, fundamentado nas n observações, e a estimativa obtida quando o i-ésimo ponto for excluído, denota-se por $\widehat{\boldsymbol{\beta}}_i$. É dada por:

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}})\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}_i - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} \tag{2.11}$$

em que $\widehat{\sigma}^2$ é a estimativa da variância do erro, com $i = 1, 2, ..., n$.

A análise de resíduos é uma das técnicas de diagnóstico do modelo muito utilizada. O resíduo para a i-ésima observação é obtido por meio da função $r_i = Y_i - \widehat{\mu}_i$, que calcula a diferença entre o valor observado (Y_i) e o valor ajustado $(\widehat{Y}_i)$, denominado de resíduo ordinário da variável resposta do modelo. Os resíduos ordinários não são muito informativos, por não

apresentar variância constante $Var(r_i) = \sigma^2(1 - h_{ii})$, pois depende dos valores h_{ii}. A solução é comparar os resíduos de forma padronizada, que é obtido pela expressão:

$$r_i^* = \frac{y_i - \widehat{y}_i}{\sqrt{\sigma^2(1 - h_{ii})}}, i = 1, ..., n. \tag{2.12}$$

Caso o modelo de regressão esteja correto o conjunto de resíduos terão a mesma variância e serão adequados para a verificação de normalidade e homocedasticidade (variância constante) dos erros. As observações que possuírem os valores absolutos dos resíduos padronizados maiores que dois poderão ser considerados pontos aberrantes.

Para este trabalho, considerou-se as técnicas informais, pois há uma amostra grande ($n =$ 3300 observações), que pode levar rejeitar com muita facilidade a validação das pressuposições dos modelos. Os gráficos utilizados foram:

- Resíduos versus valores ajustados: avaliar homogeneidade e independência;

- QQ-Plot dos resíduos: avaliar se os resíduos seguem distribuição normal de probabilidade;

- Resíduos versus medida Leverage: avaliar se há alguma observação influente (outliers).

2.4 MÉTODO STEPWISE

Na construção de modelos estatísticos é importante encontrar o modelo mais parcimonioso que explica os dados. Existem algumas técnicas, como backward, forward e stepwise, que auxiliam na seleção de variáveis explicativas para um modelo de regressão.

Uma das melhores técnicas é o método do passo a frente passo atrás, stepwise, que consiste na mistura dos métodos passo atrás (Backward) e passo a frente (forward). As incorporações ou as eliminações são realizadas passo a passo até que todas as variáveis restantes sejam significativas, ou seja, entra no modelo as variáveis que possui maior coeficiente de correlação com a variável dependente do modelo. Desta forma chega-se no modelo final.

Os procedimentos realizados no método stepwise são [13]:

1. Escolhe-se a variável x_k que possui o maior coeficiente de correlação para entrar no modelo.

2. Uma variável x_i entra no modelo, se o coeficiente de correlação for maior que o anterior, x_i permanece no modelo, caso contrário x_i sai do modelo.

3. x_i sai do modelo se o coeficiente de correlação for menor que o anterior, x_i fica no modelo, caso contrário, x_i permanece fora do modelo. Este passo é repetido até que não tenha mais x_i para sair do modelo. Terminada esta etapa retorna-se ao passo 2 e este passo continua até que não tenham mais variáveis para entrar no modelo.

2.5 Teste de Tukey

Há diversos métodos para realizar comparações múltiplas que sustentam a credibilidade da taxa de erro, um deles é chamado de *procedimentos de Tukey* que permite a formação de intervalos de confiança simultâneos com $100(1-\alpha)$ % para todas as comparações em pares [12]. O método de comparações em pares por Tukey envolve a descoberta de uma diferença mínima significante entre as médias. Esse teste foi utilizado no estudo para comparar os meses com PA equivalentes.

2.6 Variáveis Dummy

Com as comparações feitas pelo Teste de Tukey, encontra-se com a situação em que as variáveis dummy (categóricas) devem ser incorporadas no modelo, essa aplicação da regressão múltipla ocorre quando uma ou mais variáveis regressoras são categóricas. Para que alguma variável possa ser inserida no modelo, é necessário que sejam criadas uma ou mais variáveis assumindo valores numéricos, e que possam representar as categorias da variável nominal [7]. São as variáveis dummy.

As variáveis dummy que venham representar uma variável nominal - A com m categorias, A_1, A_2, ..., A_m - é definido o modelo com $m-1$ termos, D_1, D_2, ..., D_{m-1}, assumindo apenas dois valores: 0 e 1, de forma que para $i = 1, 2, 3, ..., m-1$, têm-se:

$$D_i = \begin{cases} 1, & \text{se a unidade amostral considerada pertence a categoria } A_i \\ 0, & \text{se a unidade amostral pertence a outra categoria } A_j, j \neq i. \end{cases}$$

Da forma como a variável D_i foi definida, obtendo a sequência (1, 0, 0, ..., 0), a unidade amostral em questão estará classificada na categoria A_1 e assim sucessivamente, até (0, 0, 0, ..., 0) categoria m.

Como resultado, o modelo que inclui variáveis dummy essencialmente envolve uma mudança no intercepto, conforme muda-se de uma categoria para a outra.

3. RESULTADOS

Realizou-se, no software R [4], inicialmente as seguintes estatísticas descritivas que são apresentadas na tabela 3.1. Pode-se observar que há uma grande variação no TL ($CV = 112,4\%$), enquanto que o PA e o PMG não teve grandes variações. Nota-se também que no banco de dados há pequenos e grandes produtores (TL mínimo de 34 l/m e TL máximo de 59875 l/m).

TABELA 3.1: Estatística Descritiva do relatório de captação de leite dos cooperados da COFRUL no período de Junho/2017 a Maio/2018.

Variáveis	Média	Mediana	Máximo	Mínimo	D.P.	CV(%)
TL (l/m)	6234	3826	59875	34	7007,01	112,41
PA (R$/l)	1,07	1,06	1,50	0,76	0,17	16,30
PMG(R$/l)	1,24	1,22	1,41	1,11	0,10	8,46

D.P.: Desvio Padrão; CV: Coeficiente de variação; TL: Total de leite entregue; PA: Preço pago aos produtores

PMG: Preço do leite em MG

A análise da variância (ANOVA) representada na tabela 3.2 foi realizada para comparar se há diferença do PA entre os meses estudados. Nota-se que há diferença entre os meses, em relação ao valor pago ao produtor ($p < 0,0001$), com isso a tabela 3.3 apresenta o teste de Tukey para fazer o agrupamento dos meses com preços iguais. Assim, não há diferença significativa entre os meses de maio e abril, entre março e junho, entre julho e fevereiro e entre outubro e janeiro. Já os meses agosto, novembro, setembro e dezembro não foram iguais a nenhum outro mês.

TABELA 3.2: ANOVA para comparação de meses em relação ao PA ao produtor.

Fonte de variação	GL	SQ	QM	F_{calc}	valor p
Meses	11	76,497	6,954	916,5	<0,0001
Resíduos	3288	24,949	0,007		

GL: Graus de liberdade; SQ: Soma de Quadrados; QM: Quadrados Médios

TABELA 3.3: Teste de Tukey para comparação das médias do PA nos meses analisados.

Meses	Média PA	Grupos
Maio	1,32	a
Abril	1,31	a
Junho	1,22	b
Março	1,21	b
Julho	1,07	c
Fevereiro	1,06	c
Agosto	1,02	d
Novembro	0,98	e
Setembro	0,95	f
Outubro	0,93	g
Janeiro	0,92	g
Dezembro	0,87	h

Assim, nas análises subsequentes os meses serão considerados por sete variáveis dummys para representar o efeito dos meses. As variáveis são:

$$D_i = \begin{cases} D_1 = 1, & \text{representa março e junho;} \\ D_2 = 1, & \text{julho e fevereiro;} \\ D_3 = 1, & \text{agosto;} \\ D_4 = 1, & \text{novembro;} \\ D_5 = 1, & \text{setembro;} \\ D_6 = 1, & \text{outubro e janeiro;} \\ D_7 = 1, & \text{dezembro;} \\ D_i = 0, & \text{para } i = 1, ..., 7, \text{maio e abril.} \end{cases}$$

Pela ANOVA do modelo de regressão dos dados originais (tabela 3.4) e ao nível de significância de 0,05, têm-se que, com exceção da variável D4 (mês de novembro), todas as variáveis são significativas ($p < 0,05$), então rejeita H_0, ou seja, os meses, a TL e o PMG, influenciam no PA. Mas é necessário avaliar as pressuposições deste modelo de regressão (Figura 3.1 a Figura 3.3). Utilizou-se os critérios informais (gráficos) pois o tamanho da amostra é muito grande ($n = 3300$ observações) e assim poderá ocorrer, com facilidade, a rejeição da hipótese nula dos testes para avaliar as pressuposições do modelo.

TABELA 3.4: ANOVA para o modelo de regressão com dados originais.

Fonte de variação	GL	SQ	SQM	F_{calc}	valor p
TL	1	3,800	3,800	690,976	<0,0001
PMG	1	50,220	50,220	9132,501	<0,0001
D_1	1	2,501	2,501	454,869	<0,0001
D_2	1	0,030	0,030	5,524	0,0188
D_3	1	1,433	1,433	260,621	<0,0001
D_4	1	0,013	0,013	2,291	0,1302
D_5	1	2,489	2,489	452,549	<0,0001
D_6	1	1,869	1,869	339,792	<0,0001
D_7	1	21,001	21,001	3818,974	<0,0001
Resíduos	3290	18,092	0,005		
Total	3299	0,004			

Na Figura 3.1 é apresentado o gráfico de dispersão dos resíduos do modelo de regressão versus os valores preditos e observa-se que não há nenhuma tendência, ou seja, os pontos estão dispersos em torno do zero indicando que os resíduos apresentam homogeneidade de variância e independência dos resíduos.

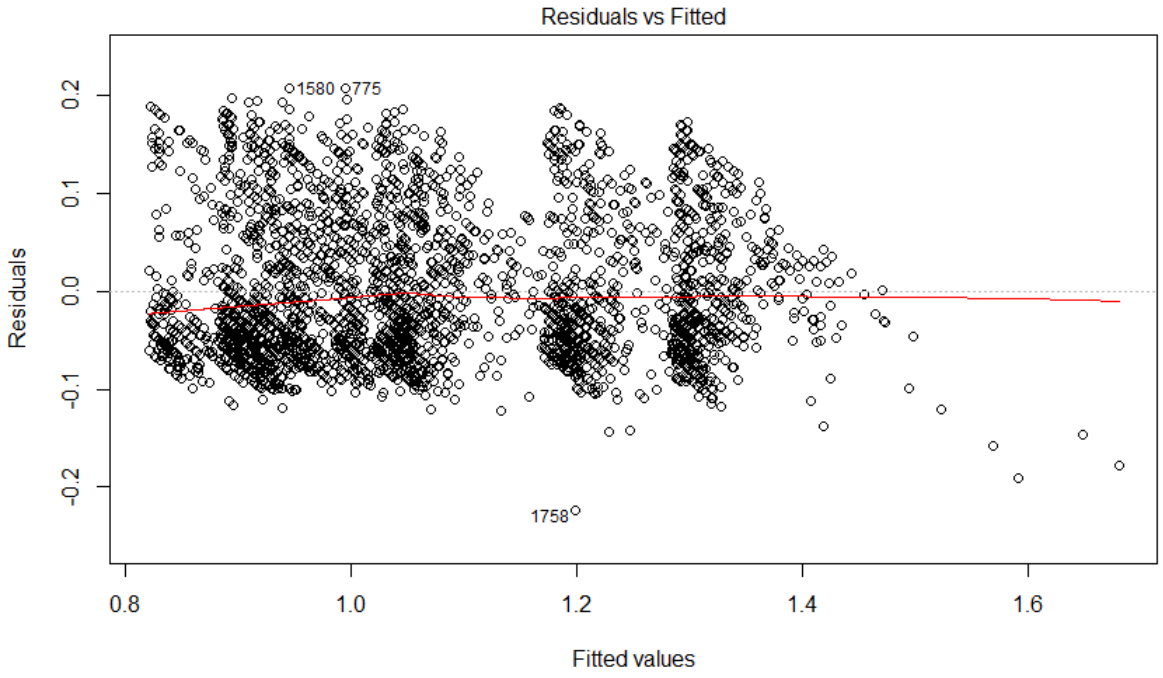

FIGURA 3.1: Gráfico de dispersão dos resíduos versus valores preditos.

Para avaliar se os resíduos seguem distribuição normal de probabilidade, utilizou-se o gráfico QQ-plot (Figura 3.2). Nota-se que os pontos estão distantes (apresentam um formato de "S") de uma reta, portanto, indicando que não é válida a pressuposição de normalidade para os resíduos. Assim, o modelo ajustado não é adequado, sendo necessária uma transformação nos dados e um reajuste do modelo.

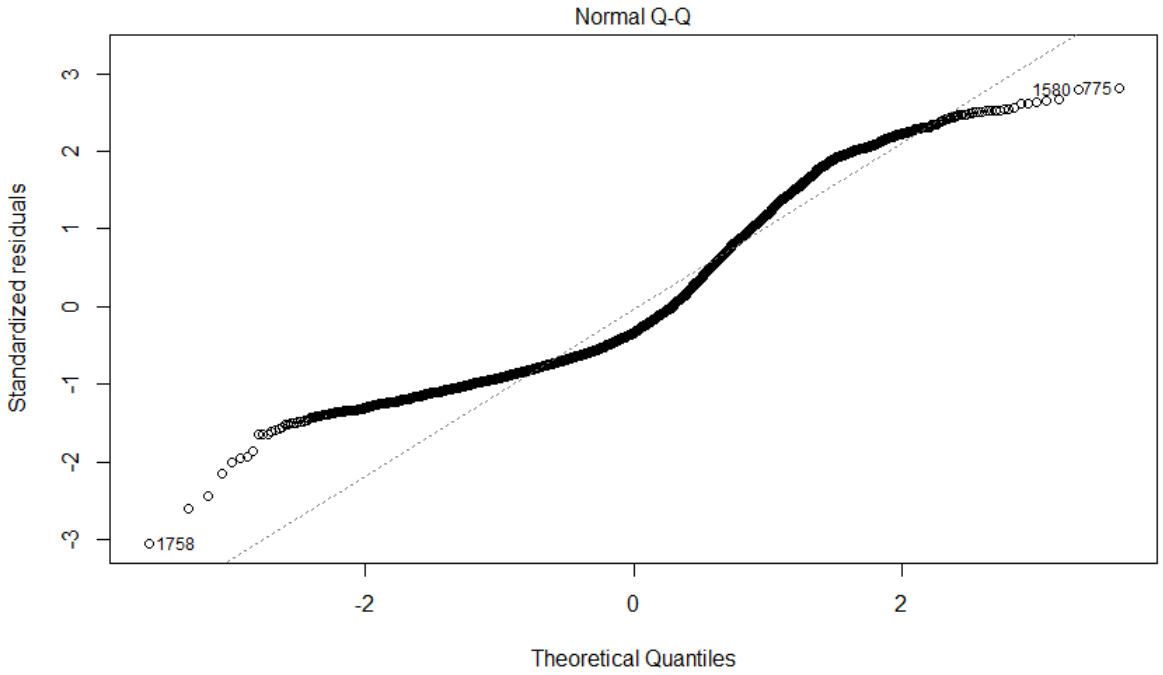

FIGURA 3.2: Q-Qplot para resíduos do modelo ajustado.

A avaliação da ocorrência de valores extremos foi feita pelo gráfico de resíduos versus a medida de Leverage (Figura 3.3). Para que as observações não sejam consideradas outliers, a distância de Cooks, deve ser menor que 1. Pela Figura 3.3 tem-se que para todos os resíduos a distância de Cooks é menor que 1, portanto não há indicativo de valores discrepantes no banco de dados.

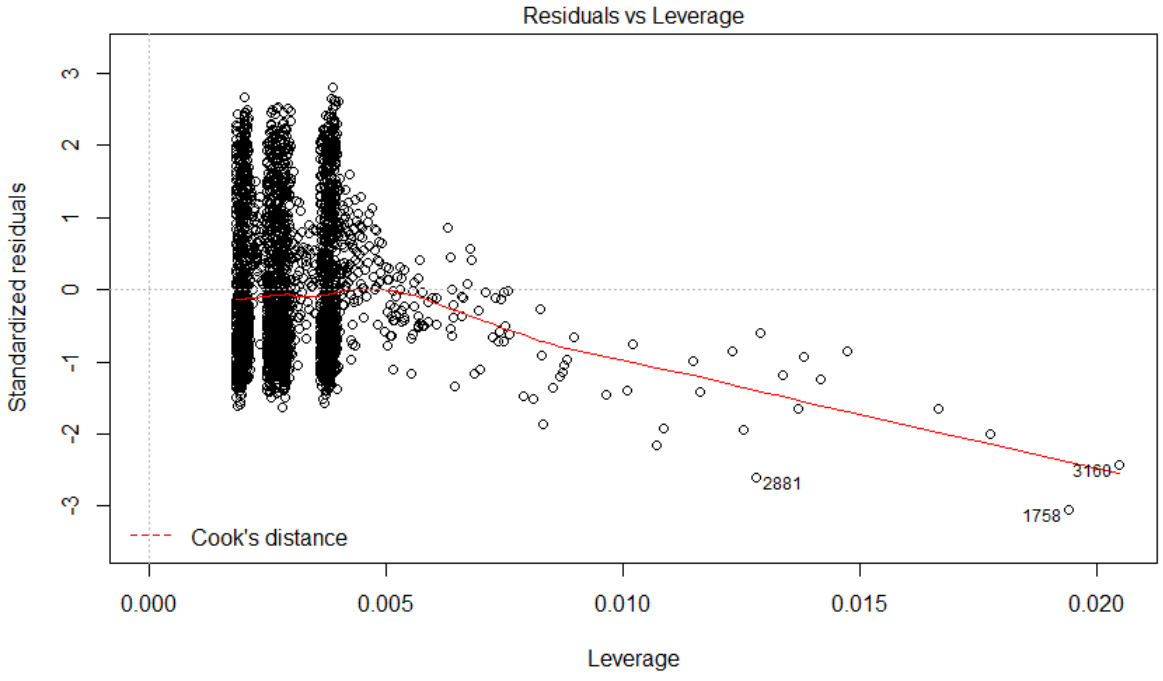

FIGURA 3.3: Gráfico de dispersão dos resíduos versus medida Leverage.

Como citado anteriormente, há a necessidade de transformar os dados para obter um modelo em que as pressuposições sejam verificadas. De acordo com a Figura 3.4, o valor ótimo para a transformação é $\lambda = -0,3$, ou seja, elevar os dados a $-0,3$. Assim, após a transformação sugerida pelo método Box-Cox, refez-se o ajuste do modelo de regressão.

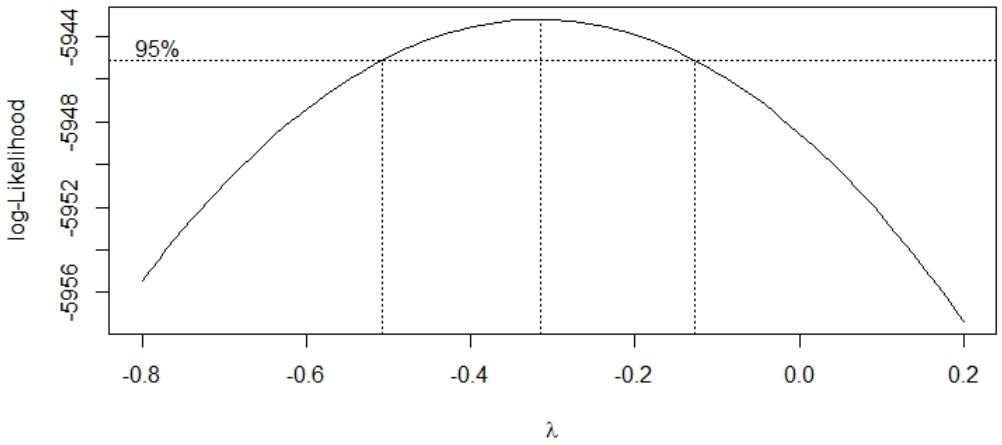

FIGURA 3.4: Transformação Box-Cox do PA do modelo de regressão.

Na Tabela 3.5 é apresentada a ANOVA do modelo de regressão dos dados transformados e ao nível de significância de 0,05, têm-se que todas as variáveis são significativas ($p < 0,05$), inclusive a variável D_4 (mês de novembro), que não foi significativa para os dados sem transformação. Logo, os meses, a TL e o PMG, influenciam no PA.

TABELA 3.5: ANOVA Dados Transformados.

Fonte de variação	GL	SQ	SQM	F_{calc}	valor p
TL	1	0,00000165	0,00000165	6,1946	0,0128
PMG	1	0,00208755	0,00208755	375,6608	<0,0001
D_1	1	0,00009984	0,00009984	375,6608	<0,0001
D_2	1	0,00001061	0,00001061	39,9112	<0,0001
D_3	1	0,00005482	0,00005482	206,2794	<0,0001
D_4	1	0,00000178	0,00000178	6,6915	0,0097
D_5	1	0,00011399	0,00011399	428,8942	<0,0001
D_6	1	0,00014081	0,00014081	529,8118	<0,0001
D_7	1	0,00075496	0,00075496	2840,6140	<0,0001
Resíduos	3290	0,00087439	0,00000027		
Total	3299	0,0041404			

Uma vez que há um modelo de regressão múltipla, há necessidade de fazer a seleção de variáveis, com a finalidade de encontrar um modelo, se possível, mais parcimonioso. Logo, procedeu-se com a aplicação da seleção de variáveis pelo método stepwise e verificou-se que pode ser retirado do modelo de regressão a variável PMG sem trazer grandes prejuízos para a qualidade do modelo final que contará com as variáveis dummys referentes aos meses e a

variável TL. As estimativas dos parâmetros deste modelo encontram-se na tabela 3.6 e, assim, o modelo final ajustado para a variável transformada é:

$$PA^{-0,3} = 0,9301 - 0,00000174TL + 0,02447D_1 + 0,0588D_2 + 0,07606D_3 + 0,08963D_4 + 0,09895D_5 + 0,1091D_6 + 0,1287D_7$$

Por este modelo é possível verificar que a TL produzido tem uma relação negativa (coeficiente negativo) com a variável transformada (PA elevado a $-0,3$). Já as variáveis dummys tem uma relação positiva (coeficientes positivos) com a variável transformada.

Tem-se ainda, que a qualidade do ajuste deste modelo pode ser considerada como boa, pois o $R^2_{aj} = 0,7882$, ou seja, 78,82% da variação do PA é explicado pela TL e os meses a serem considerados.

Colocando a variável resposta do modelo final (PA) na escala original têm-se o seguinte modelo:

$$PA = \frac{1}{(0,9301 - 0,00000174TL + 0,02447D_1 + 0,0588D_2 + 0,07606D_3 + 0,08963D_4 + 0,09895D_5 + 0,1091D_6 + 0,1287D_7)^{3,33}}$$

Assim percebe-se que, na escala original do PA, a variável TL tem uma relação positiva com o PA, ou seja, aumentando a TL, o PA também aumentará (Figura 3.5).

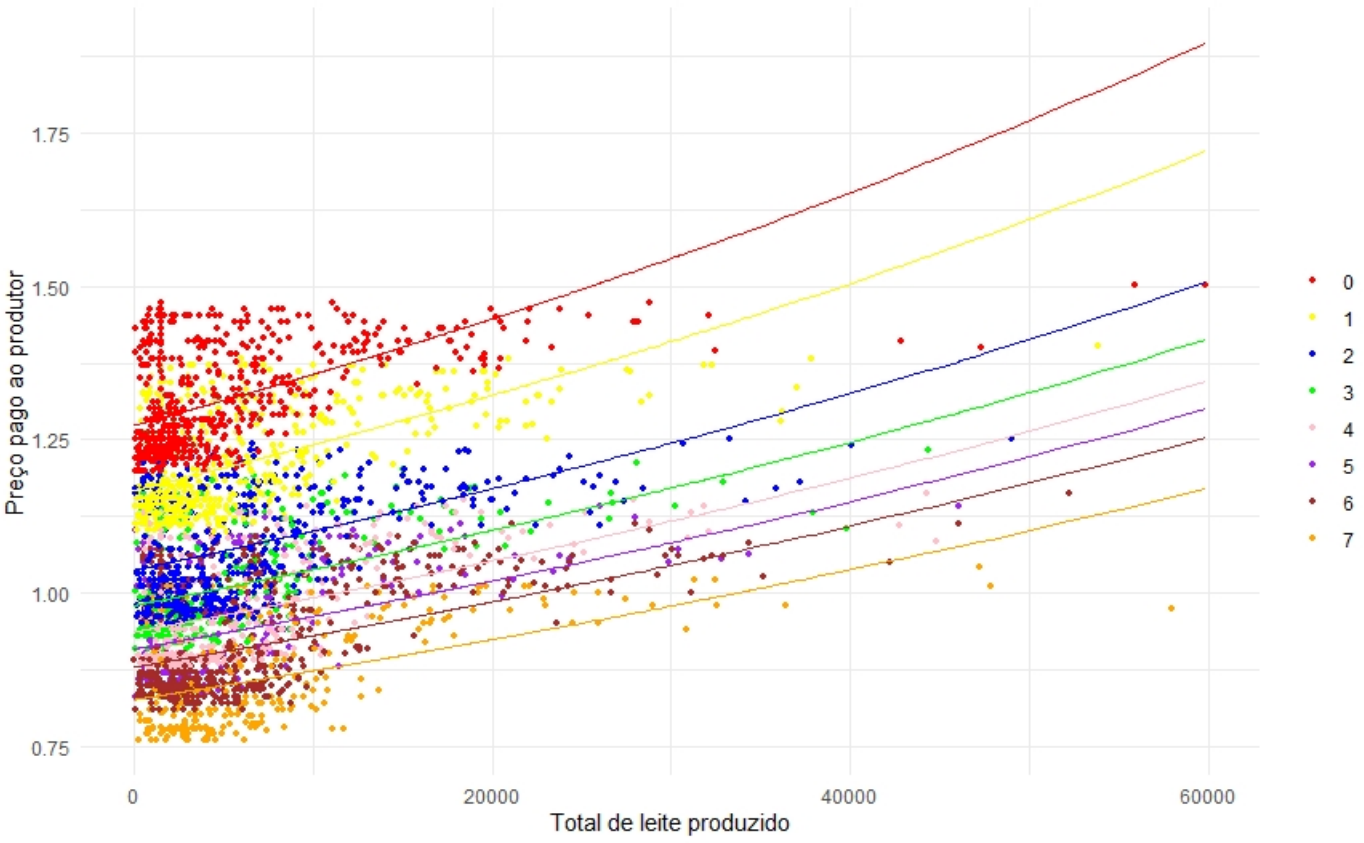

FIGURA 3.5: Gráfico de dispersão do TL versus PA com equações ajustadas para cada variável dummy.

TABELA 3.6: Estimativas dos parâmetros do modelo final.

	β	DP	valor p
Intercepto	0,9301	8,269e-04	<0,0001
TL	-0,00000174	1,371e-07	<0,0001
D_1	0,0244	1,271e-03	<0,0001
D_2	0,0588	1,248e-03	<0,0001
D_3	0,0760	1,520e-03	<0,0001
D_4	0,0896	1,794e-03	<0,0001
D_5	0,0989	1,521e-03	<0,0001
D_6	0,1091	1,293e-03	<0,0001
D_7	0,1287	1,839e-03	<0,0001

A análise de resíduo do modelo final, via critérios informais, são apresentados nas Figuras 3.6 a 3.8. Nota-se que todos os critérios são satisfeitos: homogeneidade e independência dos resíduos (Figuras 3.6), normalidade dos resíduos (Figura 3.7) e não tem observações discrepantes (Figura 3.8).

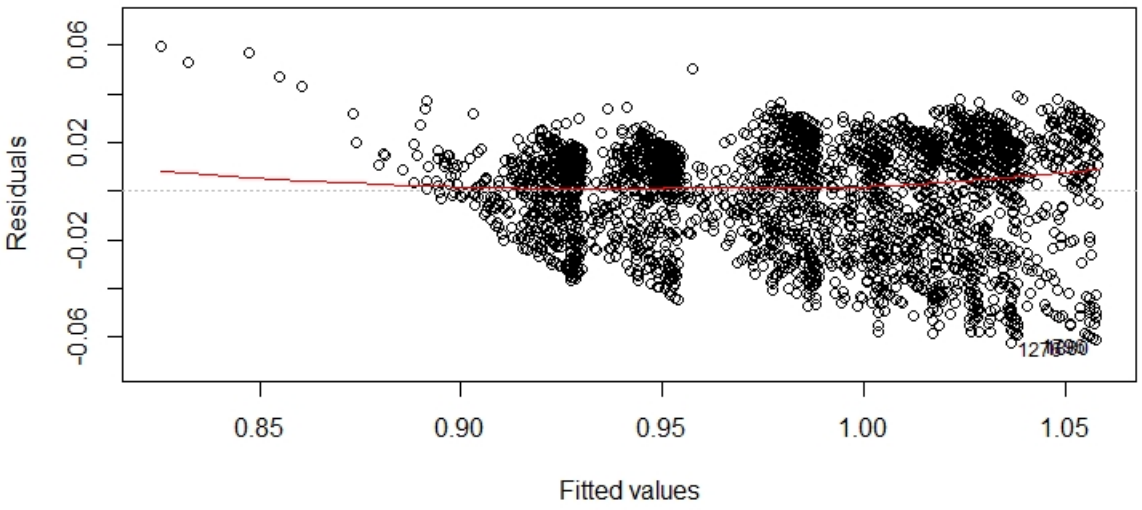

FIGURA 3.6: Gráfico de dispersão dos resíduos versus valores preditos para o modelo final.

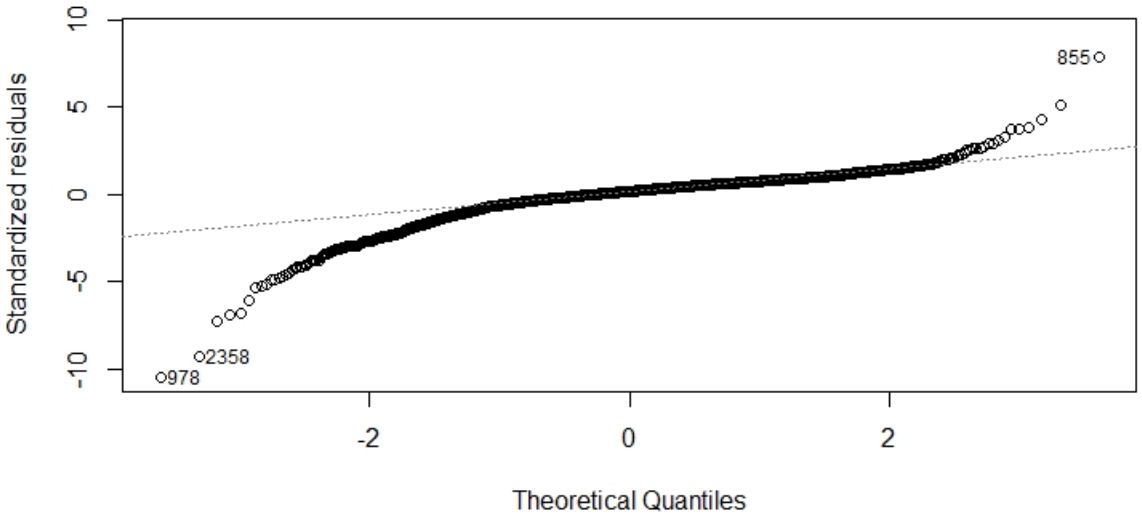

FIGURA 3.7: Q-Qplot para resíduos do modelo final.

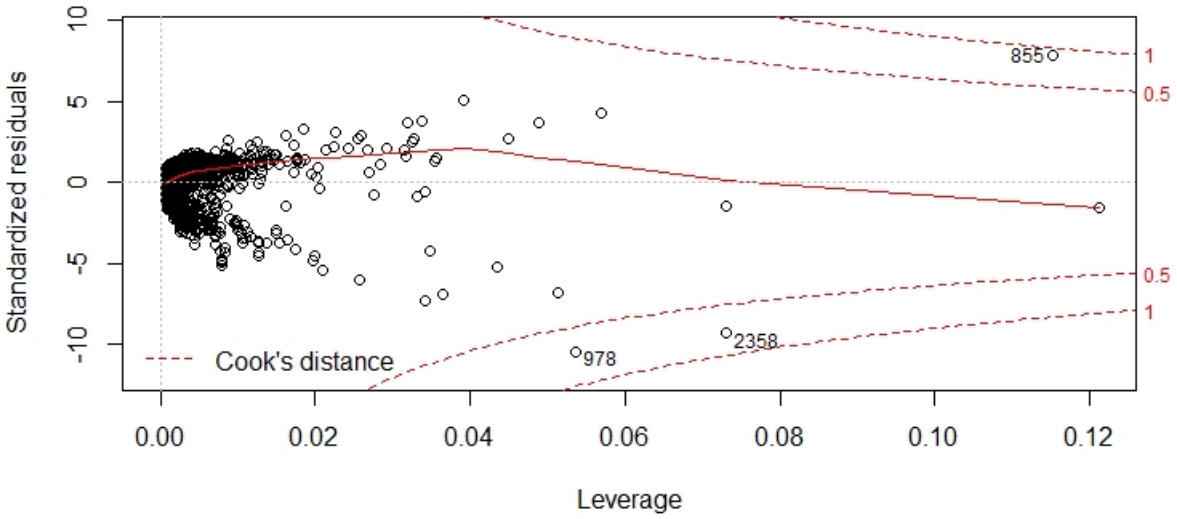

FIGURA 3.8: Gráfico de dispersão dos resíduos versus medida Leverage para o modelo final.

Com estes resultados, observa-se que há um contraste com os resultados obtidos por [14], que estudaram o preço do leite com informações provenientes de 31 produtores. No entanto, os autores concluíram que a quantidade de leite entregue não foi significativa para explicar o preço do leite e as variáveis estrato sólido total, contagem de célula somática e contagem bacteriana total foram significativas para explicar o preço do leite. Além disso, não levaram em conta a época do ano e o modelo ajustado por [14] foi diferente do modelo ajustado neste trabalho.

Já o trabalho [16], concluiu que o preço do leite possui sazonalidade e que durante os meses de seca (de maio a setembro) tendem a ser maiores e em relação aos períodos de chuvas, os preços são mais baixos. Portanto há uma certa concordância entre [16] e esse presente trabalho, que

observou uma variação nos preços em função do clima, ou seja, nos períodos de chuva (dezembro a fevereiro) os preços tiveram uma tendência a serem mais baixos e no período de seca uma tendência de serem mais altos (segundo Tabela 3.3).

4. Conclusões

Neste trabalho ajustou-se o modelo de regressão linear múltipla ao conjunto de dados provenientes da COFRUL. Após o ajuste com todas as variáveis, utilizou-se ferramentas para analisar o ajuste e adequação do modelo de regressão linear múltipla. Foi necessário aplicar uma transformação Box-Cox nos dados (elevou-se os dados a $-0,3$) e as suposições do modelo de regressão foram satisfeitas. O modelo final foi ajustado considerando as variáveis dummys para indicação do mês de coleta e TL. Os meses que possuem os menores preços pago ao produtor são: outubro, janeiro e dezembro, e os meses com maiores valores pagos aos produtores são março e junho. Percebe-se pela expressão do modelo que quanto maior for a TL, maior será o PA.

Referências Bibliográficas

[1] *Associativismo e Cooperativismo.* http://www2.ufersa.edu.br/portal/view/uploads/setores/241/Cartilha%20de%20Associativismo%20e%20Cooperativismo.PET-PROEX.pdf.

[2] *Preço do leite.* https://www.cepea.esalq.usp.br/br/indicador/leite.aspx.

[3] *Um breve história da produção leiteira no Brasil.* https://revistagloborural.globo.com/Noticias/Criacao/Leite/noticia/2018/10/leite-sem-politica.html.

[4] *A Language and Environment for Statistical Computing.* 2019, R.

[5] CORDEIRO, G. M.; LIMA NETO, E. A.: *Modelos paramétricos.* Recife: Universidade Federal Rural de Pernambuco, 2006.

[6] PAULA, G. A.: *Modelos de regressão: com apoio computacional.* São Paulo: IME/USP, 2004.

[7] CHARNET, R.; FREIRE, C. D. L.; CHARNET, E. M. R.; BONVINO, H. et al.: *Análise de modelos de regressão linear com aplicações.* Campinas, SP, Editora da Unicamp, 356p,, 1999.

[8] MONTGOMERY, D. C.; RUNGER, G. C.: *Estatística aplicada e probabilidade para engenheiros.* 2. ed. Rio de Janeiro: LTC, 2008.

[9] KASSAOKA, D.: *Guia do associativismo rural.* São Paulo - Coordenadoria de Desenvolvimento dos Agronegócios, 2017; 52p.

[10] LÍRIO, G.W.; Pierret, D.; PIERRET, V.H.; SOUZA, A.M.; SILVA, W.V. da: *Análise E Previsão Da Série Recebimento E Produção De Leite Da Usina Escola De Laticínios Da Universidade Federal De Santa Maria – RS.* Disponível em: <http://w3.ufsm.br/adriano/revista/cn2003/preleite.pdf>.

[11] SASSI, C.P.; PEREZ, F.G.; MYAZATO, L.; YE, X.; SILVA, P.H.F. e LOUZADA, F.: *Modelos De Regressão Linear Múltipla Utilizando Os Softwares R E Statistica: Uma Aplicação A Dados De Conservação De Frutas.* Disponível em: <http://conteudo.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_377.pd

[12] WALPOLE, R. E. ...[et al.]; [tradução Luciane F Pauleti Vianna]: *Probabilidade e estatística para engenharia e ciências.* –São Paulo: Pearson Prentice Hall, 2009.

[13] ALVES, M.F.; LOTUFO, A.D.P. E LOPES, M.L.M.: *Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas.* 2013.

[14] BARRRETO, V.C.S.; BATISTELA , G.C.; GAIOTTO, M.R. e SIMÕES, D.: *Regressão linear múltipla aplicada ao preço do leite.* 7, dez. 2016.

[15] DEMÉTRIO, C.G.B.; ZOCCHI, S.S.: *Modelos de Regressão.* November 21, 2006.

[16] ALVES, F.F.; ERVILHA, G.T. e TOYOSHIMA, S.H.: *SAZONALIDADE E PREVISÃO DO PREÇO DO LEITE RECEBIDO PELOS PRODUTORES DA BAHIA E DE MINAS GERAIS.*

A. Apêndice

```
    dadostcc<-read.table("dadostcc.txt",h=T, sep="\t", dec=",")
summary(dadostcc)
attach(dadostcc)
##################
# Anova comp. meses
##################
trat<-x1 + x2*2 + x3*3 + x4*4 + x5*5 + x6*6 + x7*7 + x8*8 + x9*9 + x10*10 + x11*11
tapply(Preço.Atual,trat,mean)
require(agricolae)
anova1<-lm(Preço.Atual as.factor(trat), data=dadostcc)
anova(anova1)
HSD.test(anova1,"as.factor(trat)",group=TRUE,console=TRUE)
y1=x1+x10
y2=x2+x9
y3=x3
y4=x6
y5=x4
y6=x5+x8
y7=x7

z=y1+y2*2+y3*3+y4*4+y5*5+y6*6+y7*7
##################
# Reg sem transformar
##################
lm2 <- lm((Preço.Atual) Total.Leite+Preço.MG+y1+y2+y3+y4+y5+y6+y7, weights = 1/Total.Leite)
boxcox(lm2, lambda = seq(-0.6,0,0.1))
##################
# Reg com transformação
##################

lm2_t <- lm((Preço.Atual)^ (-0.3) Total.Leite+Preço.MG+y1+y2+y3+y4+y5+y6+y7, weights
```

```
= 1/Total.Leite)
anova(lm2_t)
summary(lm2_t)
slm2 <- step(lm2_t)
lm3_t <- lm((Preço.Atual)^ (-0.3) Total.Leite+y1+y2+y3+y4+y5+y6+y7, weights = 1/To-
tal.Leite)
summary(lm3_t)
plot(lm3_t)
#######################
library(reshape2)
library(ggplot2)
library("wesanderson")
library("RColorBrewer")

x<-Total.Leite

yp0=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x
yp1=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x+lm3_t$coefficients[3]
yp2=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x+lm3_t$coefficients[4]
yp3=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x+lm3_t$coefficients[5]
yp4=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x+lm3_t$coefficients[6]
yp5=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x+lm3_t$coefficients[7]
yp6=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x+lm3_t$coefficients[8]
yp7=lm3_t$coefficients[1]+lm3_t$coefficients[2]*x+lm3_t$coefficients[9]

yp0_o<-1/(yp0^ (10/3))
yp1_o<-1/(yp1^ (10/3))
yp2_o<-1/(yp2^ (10/3))
yp3_o<-1/(yp3^ (10/3))
yp4_o<-1/(yp4^ (10/3))
yp5_o<-1/(yp5^ (10/3))
yp6_o<-1/(yp6^ (10/3))
yp7_o<-1/(yp7^ (10/3))

g <- ggplot(dadostcc)
g <- g + geom_point(aes(x = Total.Leite, y = Preço.Atual, color = factor(z)),size = 1)
g<-g+ scale_color_manual( values= c("red","yellow","blue ","green","pink","purple","brown","orange")
g<-g+ labs(subtitle = " ", y = "Preço pago ao produtor", x = "Total de leite produzido")
g<-g+ labs( color = " ")
```

```
g<-g+geom_line(aes(x,yp0_o), colour="red")+geom_line(aes(x,yp1_o), colour="yellow")+
geom_line(aes(x,yp2_o), colour="blue")+geom_line(aes(x,yp3_o), colour="green")+
geom_line(aes(x,yp4_o), colour="pink")+geom_line(aes(x,yp5_o), colour="purple")+
geom_line(aes(x,yp6_o), colour="brown")+geom_line(aes(x,yp7_o), colour="orange")
g<-g+theme_minimal()
g
```