Universidade Federal Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Especialização em Estatística

DEST
UFMG

Igor Mazzeto Resende Soares

Estimação dos principais direcionadores dos custos operacionais das empresas brasileiras de transmissão de energia elétrica utilizando modelos de regressão e programação linear

Trabalho de Conclusão de Curso

Belo Horizonte, 2023

Igor Mazzeto Resende Soares

Estimação dos principais direcionadores dos custos operacionais das empresas brasileiras de transmissão de energia elétrica utilizando modelos de regressão e programação linear

Trabalho apresentado ao Departamento de Estatística da Universidade Federal de Minas Gerais como parte dos requisitos para a obtenção do Grau de Especialista em Estatística.

Universidade Federal Minas Gerais

Orientador: Prof. Dr. Marcelo Azevedo Costa

Belo Horizonte
2023

Soares, Igor Mazzeto Resende.

S676e Estimação dos principais direcionadores dos custos operacionais das empresas brasileiras de transmissão de energia elétrica utilizando modelos de regressão e programação linear [recurso eletrônico] / Igor Mazzeto Resende Soares —2023.
1 recurso online (69 f. il, color).

Orientador: Marcelo Azevedo Costa
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística
Referências: 58-63.

1. Estatística. 2. Programação Linear. 3. Energia elétrica, Distribuição de. 4. Energia elétrica – Custos. I. Costa, Marcelo Azevedo. II. Universidade Federal de Minas Gerais. I. Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 - Universidade Federal de Minas Gerais – ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação
Caixa Postal 702
31270-901 Belo Horizonte- MG – Brasil

Telefone (31) 3409-5923
Fax (31) 3499-5924
E-mail: pgest@ufmg.br
WEB: http://www.est.ufmg.br/posgrad/

DEPARTAMENTO DE ESTATÍSTICA – ICEX/UFMG

Igor Mazzeto Resende Soares – 25/08/2023, às 13:30 horas

NOTA DA APRESENTAÇÃO DO TRABALHO DE FIM DE CURSO "ESPECIALIZAÇÃO EM ESTATÍSTICA – UFMG"

BANCA:

PROFESSOR (A)	NOTA(0-100)	ASSINATURA
Prof. Marcelo Azevedo Costa (Orientador)	97	gov.br Documento assinado digitalmente MARCELO AZEVEDO COSTA Data: 25/08/2023 15:31:41-0300 Verifique em https://validar.iti.gov.br
Prof. Tiago Silveira Gontijo	97	gov.br Documento assinado digitalmente TIAGO SILVEIRA GONTIJO Data: 25/08/2023 15:36:46-0300 Verifique em https://validar.iti.gov.br
MÉDIA	**97**	

Trabalho intitulado: "Estimação dos principais direcionadores dos custos operacionais das empresas brasileiras de transmissão de energia elétrica utilizando modelos de regressão e programação linear".

Agradecimentos

Agradeço primeiramente aos meus pais, Valéria e Alberley, pelo amor incondicional e por sempre me apoiarem nas minhas buscas pelo conhecimento e pelos caminhos da vida.

Agradeço à minha noiva, Gabriela, que sempre apoiou as minhas decisões e que sempre esteve ao meu lado. Sem seu amor e carinho nada disso seria possível.

Agradeço ao Professor Marcelo Azevedo Costa, pela paciência e disposição na orientação do trabalho. E por ser um educador exemplar e inspirador, figura rara nos dias atuais.

Por fim, agradeço a Deus, por ter me concedido saúde em tempos tão difíceis. E pela dádiva da curiosidade e da fome de conhecimento, atributos sem os quais provavelmente não teria embarcado em mais uma jornada.

.

"Que haja uma luz nos lugares mais escuros, quando todas as outras luzes se apagarem."(J.R.R. Tolkien)

Resumo

A Agência Nacional de Energia Elétrica (ANEEL) publicou a Tomada de Subsídios – TS nº 14/2022 para a obtenção de informações sobre a base de dados que será utilizada no estudo de *benchmarking* dos custos operacionais regulatórios dos agentes de transmissão. O cálculo desses custos é feito por meio da Análise Envoltória de Dados (*Data Envelopment Analysis*), utilizando o custo operacional como insumo e oito produtos como variáveis explicativas (*drivers*) dos custos. No entanto, devido à forte correlação linear entre as variáveis explicativas, o modelo de regressão pode gerar valores inconsistentes para os coeficientes estimados. Uma alternativa para lidar com esse problema é a utilização de programação linear e restrições nas inequações para estimação dos parâmetros do modelo. Nesse contexto, este trabalho propõe uma metodologia que combina a regressão linear com a programação linear e técnicas de validação cruzada para avaliar os principais direcionadores dos custos operacionais das empresas brasileiras de transmissão de energia elétrica. Os resultados indicam que o modelo de programação linear se mostrou o mais adequado apresentando coeficiente de determinação preditivo de valor igual 0,80 e que apenas cinco entre oito variáveis explicativas foram identificadas relevantes para estimação dos custos operacionais, sugerindo redundância na metodologia atual de cálculo dos custos, podendo comprometer as estimativas de eficiência.

Palavras-chaves: Regressão Linear. Programação Linear. Custos operacionais. Transmissão de energia elétrica. Revisão periódica.

Abstract

The Brazilian Electricity Regulatory Agency (Agência Nacional de Energia Elétrica - ANEEL) has published the subsidy taking document - TS 14/2022 to gather information regarding the database to be used in the regulatory operational cost benchmarking study for transmission agents. The calculation of these costs is carried out using Data Envelopment Analysis (DEA), with operational cost as input and eight variables as explanatory variables (drivers) of costs. However, due to the strong linear correlation among the explanatory variables, a traditional regression model may generate inconsistent values for the estimated coefficients. An alternative approach to address this issue involves the utilization of linear programming and inequalities constraints to estimate the model parameters. In this context, this work proposes a methodology that combines linear regression with linear programming and cross-validation techniques to evaluate the main drivers of operational costs for Brazilian electricity transmission companies. The results indicate that the linear programming model proved to be the most suitable, with a predictive coefficient of determination of 0.80, and that only five out of eight explanatory variables were identified as the primary drivers of operational costs. This suggests redundancy in the current methodology for calculating costs, which could potentially compromise efficiency estimates.

Key-words: Linear regression. Linear programming. Operational costs. Electric power transmission. Periodic review.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANEEL	Agência Nacional de Energia Elétrica
AIC	Critério de informação de Akaike
DEA	*Data Envelopment Analysis*
MEA	Minimização de erros absolutos ponderados
MLG	Modelos lineares generalizados
MVA	Megavolt amperes
Mvar	Megavolt amperes reativo
MQO	Mínimos quadrados ordinários
NT	Nota técnica
PMSO	Contas contábil de pessoal, materiais, serviços de terceiros, seguros, tributos e outros
RAP	Receitas anuais permitidas
REN	Resolução normativa
RTP	Revisão tarifária periódica
SEB	Sistema energético brasileiro
SPE	Sociedade de Propósito Específico
SRM	Superintendência de Regulação Econômica e Estudos do Mercado

Sumário

1 Introdução

O sistema energético brasileiro (SEB) é composto por empresas públicas e privadas que são reguladas pela Agência Nacional de Energia Elétrica (ANEEL). O Art. 2º da Lei nº 9.427 de 26 de dezembro de 1996 dispõe que a ANEEL tem como finalidade regular e fiscalizar a produção, transmissão, distribuição e comercialização de energia elétrica. É também de responsabilidade a ANEEL estabelecer e rever as tarifas segundo do Art. 3º, Incisos XII e XXII. (BRASIL, 1996)

Para os agentes envolvidos no SEB, a revisão de tarifas é de importância central para a determinação das receitas anuais permitidas (RAP) das instalações de transmissão. Segundo Pessanha et al. (2010), a RAP envolve a apuração dos custos anuais dos ativos elétricos, dos custos de administração e custos operacionais, acrescidos dos tributos aplicáveis do setor e de uma parcela de ajuste. Costa et al. (2022) discorrem que tais custos operacionais compreendem uma pequena parte da tarifa de energia e do custo operacional regulatório, dessa forma é crucial que tal parte seja calculada de forma apropriada, para que o operador do sistema possa cobrar o consumidor final de forma adequada.

Lopes et al. (2016) argumentam que uma das dificuldades do setor elétrico no Brasil envolve fornecer aos consumidores finais energia com preços justos com remuneração justa para o distribuidor. Tal remuneração é fortemente afetada pelos custos operacionais do setor, por isso o regulador estimula que os partícipes do sistema promovam redução deles.

Para Pessanha et al. (2010), o desafio do cálculo dos custos operacionais eficientes do setor é a assimetria de informação entre o regulador e os agentes do mercado. Para evitar tal assimetria e definir tais custos de forma procedimental, a reguladora por meio da Resolução Normativa - REN ANEEL nº 257 de 6 de março de 2007 estabeleceu os conceitos gerais e procedimentos para revisão da primeira revisão tarifária periódica (RTP). A partir de então, em cada ciclo de revisão a reguladora divulga documentos normativos e técnicos com orientações para os procedimentos a serem realizados (BRASIL, 2007).

Em fevereiro de 2017, a Superintendência de Regulação Econômica e Estudos do Mercado (SRM) emitiu a Nota Técnica nº 037/2017-SRM/ANEEL por meio da qual apresentou as bases de dados que seriam utilizadas no estudo de eficiência dos custos operacionais das concessionárias de transmissão e que subsidiaria a definição dos custos operacionais regulatórios dos contratos de concessão que passariam por revisão no período de julho de 2018 a junho de 2023 (BRASIL, 2017a).

A ANEEL publicou a Nota Técnica nº 160/2017 (SRM) que dispõe sobre as regras de apuração dos custos operacionais regulatórios. A reguladora explicita no documento publicado que vem adotando a prática de realizar análises comparativas entre os custos operacionais dos operadores do SEB, tornando possível a atribuição de um nível de efi-

ciência a cada empresa e, então, definir os valores de custos operacionais regulatórios associados a uma referência de eficiência. Na nota, fica definido que o insumo utilizado no estudo foi a despesa operacional das transmissoras, composto pelas contas de pessoal, materiais, serviços de terceiros, seguros, tributos e outros (PMSO) (BRASIL, 2017b).

No ano de 2018 foi publicado pela ANEEL Nota Técnica nº 204/2018-SRM/ANEEL que apresenta a metodologia a ser utilizada no cálculo dos custos operacionais regulatórios no âmbito do processo de revisão periódica das receitas anuais permitidas das concessionárias de transmissão de energia elétrica. Conforme descrito no documento, é um ponto central de todo o processo da definição dos custos operacionais o estabelecimento de critério para aferição do nível eficiência de custos das empresas, por isso a reguladora vinha discutindo com a sociedade o modelo de benchmarking a ser utilizado visto que há a necessidade de se levar em consideração as características e particularidades de cada agente. Posteriormente serão discutidos a metodologia, variáveis e demais ajustes definidos no documento (BRASIL, 2018b).

A tomada de subsídio é uma etapa crucial para o ciclo regulatório para a revisão periódica da RAP. Dessa forma a reguladora publicou nota a Nota Técnica nº 97/2018-SRM/ANEEL para consolidação das bases de dados e informações que serão utilizadas para embasar o benchmarking dos custos operacionais regulatórios das concessionárias de transmissão. Visando dar transparência ao processo, foi publicado no Diário Oficial da União (DOU), seção 3, página 106, a abertura da Tomada de Subsídios – TS nº 14/2022 (BRASIL, 2018a), (BRASIL, 2022b).

Em prosseguimento a TS nº 14/2022, a reguladora publicou em 29 de julho de 2022 a Nota Técnica Nº 97/2022–SRM/ANEEL com objetivo de propor instauração de Tomada de Subsídios para apresentar e consolidar as bases de dados que serão utilizadas para subsidiar o estudo de *benchmarking* de eficiência dos custos operacionais regulatórios das concessionárias de transmissão que passarão por revisão periódica da Receita Anual Permitida – RAP em 1º de julho de 2023, conforme data contratual. A NT 97/2022 buscou dar publicidade e transparência para a base de dados que será utilizada, permitindo que a sociedade a avalie, critique, audite e valide, viabilizando a realização de eventuais correções que se mostrem necessárias e conferindo maior transparência ao processo como um todo. A base de dados foi posteriormente publicada conforme descrição das notas técnicas (BRASIL, 2022a).

1.1 Justificativa

Segundo Feitosa Neto (2009), a determinação correta da RAP para as concessionárias atuantes no SEB é de extrema importância para não comprometer a operação do sistema e não onerar excessivamente o consumidor final. Dessa forma, apurar de forma robusta e realista o custo operacional (PMSO) é crucial para o cálculo da RAP e para

manutenção do nível tarifário.

A implementação de uma técnica estatística que visa corrigir distorções matemáticas e qualitativas e que modele as restrições operacionais dos direcionadores de custo é de extrema importância para que se evite o comprometimento das estimativas de eficiência dos custos operacionais. Este trabalho busca contribuir para as discussões acerca do tema e para oferecer ao regulador uma alternativa para o cálculo do parâmetro tão considerável para as definições regulatórias do SEB.

1.2 Objetivos

O presente trabalho tem como propósito implementar um modelo de previsão que respeite as restrições técnicas impostas pela natureza operacional dos direcionadores de custo e avaliar seus resultados, implicações e compará-lo com modelos tradicionais. Portanto, deseja-se que variáveis preditoras (direcionadores de custo) do modelo contribuam de forma positiva ou neutra com a variável resposta (PMSO), eliminando qualquer incongruência ou deturpação da realidade imposta.

São objetivos específicos nessa pesquisa:

- Analisar as correlações existentes entre as variáveis direcionadoras de custo e o custo operacional das concessionárias.
- Definir a representatividade das variáveis direcionadoras de custo nos modelos matemático-estatístico e sua importância para a realização de previsões.
- Definir um modelo de regressão linear múltipla ou de programação linear para o custo operacional que respeite as restrições impostas pela natureza da operação.
- Discutir os resultados encontrados e apresentar uma alternativa ao regulador para o modelo utilizado na definição dos custos operacionais.

2 Referencial teórico

2.1 Regressão Linear

Segundo Upton et al. (2014), regressão linear é o modelo de regressão estatístico mais simples e mais utilizado. Montgomery et al. (2003) definem a análise de regressão como uma técnica estatística utilizada para investigar a relação de uma ou mais variáveis. Para Hoffmann (2016) a análise de regressão permite conhecer os efeitos quer umas variáveis exercem ou que pareçam exercer sobre outras, os autores ainda classificam que tal método é o mais importante para a econometria.

O método de regressão linear foi formalizado por Francis Galton em 1891 em sua publicação Hereditary Genius, onde o autor realiza um estudo da relação entre as alturas de pais e filhos. De forma sucinta, o autor concluiu com os resultados observados por sua técnica que pais altos tendem a ter filhos altos e pais baixos tendem a ter filhos mais baixos. Segundo Stanton (2001), após as contribuições de Karl Pearson em 1896, 1922 e 1930 o método obteve rigor matemático e foi amplamente utilizado.

2.1.1 Regressão Linear Univariada

Montgomery et al. (2003) definem que a regressão linear univariada ou regressão linear simples consiste na relação entre uma variável aleatória Y e outra x de acordo com a expressão abaixo, onde a Y é a variável dependente, x é a variável aleatória independente, β_0 é o coeficiente da regressão chamado de intercepto e β_1 é o coeficiente da regressão chamado de coeficiente de inclinação.

$$E\left(Y|\ x\right) = \ \mu_{Y|x} = \ \beta_0 + \beta_1 x \tag{2.1}$$

Em termos formais, Montgomery et al. (2003) definem que a relação entre Y e x seja uma linha reta onde o valor esperado para cada observação Y em cada nível de x seja uma variável aleatória. Em outros termos, os autores explicam que a maneira apropriada de se generalizar tal afirmação para um modelo linear probabilístico é considerar que o valor esperado de Y seja em função de x, mas que para um valor fixo de x o valor real de Y seja determinado pela função do valor médio mais um termo de erro aleatório, onde e representa tal termo com média zero e variância σ^2:

$$Y_i = \ \beta_0 + \beta_1 x_i + \ e_i \qquad \forall\ i \ = \ 1, 2 \ldots, n \tag{2.2}$$

2.1.2 Método de Mínimos Quadrados e propriedades dos estimadores

Para implementação do modelo linear probabilístico proposto anteriormente, é preciso realizar a estimação do coeficiente angular ou de inclinação (β_1) e o intercepto (β_0). O método dos mínimos quadrados (MQO) é um dos métodos mais utilizados para estimação os coeficientes de regressão linear univariada. Em termos gerais o método consiste na minimização da soma dos quadrados das diferenças entre a variável dependente observada no conjunto de dados de entrada e a saída da função da variável independente, entretanto para que o MQO entregue estimadores consistentes, Lewis-Beck et al. (2015) e Kennedy (2002) defendem que determinados pressupostos devem ser satisfeitos:

- A esperança do termo de erro deve ser igual a zero ($E(\varepsilon) = 0$) e ter distribuição normal com variância σ^2.
- Deve haver homoscedasticidade dos termos de erro, ou seja, a variância do erro é constante para os diferentes valores da variável independente.
- As variáveis independentes não podem apresentar alta correlação entre si, ou seja, não podem apresentar a propriedade de multicolinearidade.
- A variável independente não pode apresentar correlação com o termo de erro

Segundo Montgomery et al. (2003), para estimação dos coeficientes, a soma dos quadrados dos desvios das observações em relação a linha de regressão dada por:

$$L = \sum_{i=1}^{n} \varepsilon_1^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.3}$$

Dessa forma os estimadores de mínimos quadrados devem satisfazer:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \tag{2.4}$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0 \tag{2.5}$$

Após simplificação:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = y_i \tag{2.6}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} y_i x_i \tag{2.7}$$

Finalmente, as estimativas dos coeficientes podem ser descritas como:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.8}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \tag{2.9}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \ , \ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.10}$$

A equação da reta ajustada para o MQO:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.11}$$

2.1.3 Regressão Linear Múltipla

Se na regressão linear univariada tínhamos apenas uma única variável independente para a modelagem, nesse caso o modelo poderá ter n termos que se relacionam com a variável dependente. Milone et al. (1995) disserta que tal método é indicado para realizar investigações em casos em que várias variáveis afetam simultaneamente a variável de interesse, podendo ser uma ferramenta poderosa.

Para representar a o modelo de regressão linear múltipla, podemos utilizar a seguinte equação, onde k representa a quantidade de m variáveis independentes,β_0 representa o intercepto da equação e β_k o efeito da variável independente.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \tag{2.12}$$

Para estimação dos coeficientes da regressão no modelo multivariado é possível utilizar o MQO ajustado conforme descreve Montgomery et al. (2003):

$$L = \sum_{i=1}^{n} \varepsilon_1^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_k x_{ik})^2 \tag{2.13}$$

Minimizando L com relação aos coeficientes, as estimativas têm que satisfazer:

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{k=1}^{m} \hat{\beta}_k x_{ik} \right) x_{ik} = 0 \quad \forall \ k = 1 ... m \tag{2.14}$$

As equações originadas a partir de (2.14) são chamadas de equações normais de mínimos quadrados:

$$
\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} {x_{i1}}^2 + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} y_i x_{i1} \\
&\vdots \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} + ... + \hat{\beta}_1 \sum_{i=1}^{n} {x_{ik}}^2 = \sum_{i=1}^{n} y_i x_{i1}
\end{aligned}
\tag{2.15}
$$

Pode-se reescrever o modelo de regressão múltipla em forma de notação matricial, no qual há k variáveis independentes, n observações. No caso, Y representa o vetor (nx1) de observações da variável dependente, X representa uma matriz (n x m), β o vetor (m+1 x 1) com os coeficientes do modelo:

$$
y = X\beta + \varepsilon \tag{2.16}
$$

Sendo:

$$
\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & ... & X_{1m} \\ 1 & X_{21} & ... & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & ... & X_{nm} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_3 \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_3 \end{bmatrix} \tag{2.17}
$$

2.1.4 Propriedades, pressupostos e adequação dos modelos de regressão linear

Para validar o ajuste de um modelo de regressão linear é preciso verificar sua representatividade e respeitar os pressupostos de normalidade dos resíduos, assumindo que devem apresentar variância constante e média igual a zero.

Sharpe (2000) apresenta algumas suposições e condições para o modelo de regressão linear múltipla. Primeiramente é a suposição de linearidade (significância do modelo), é importante verificar se há uma relação linear entre as variáveis explicativas e a variável resposta. Isso pode ser feito plotando o diagrama de dispersão de y *versus* cada variável X e verificando se há uma relação linear. Também é recomendado plotar o diagrama dos resíduos para detectar violações das condições de linearidade. Tal suposição pode ser verificada pelo teste de significância.

A suposição de independência dos erros do modelo de regressão deve ser verificada para garantir que eles são independentes entre si. Isso significa que os erros não devem estar correlacionados. Além disso, é importante que os dados sejam provenientes de uma amostra aleatória ou experimento aleatório para garantir a condição de aleatoriedade.

A suposição de independência dos erros do modelo de regressão deve ser verificada para garantir que eles são independentes entre si. Isso significa que os erros não devem

estar correlacionados. Além disso, é importante que os dados sejam provenientes de uma amostra aleatória ou experimento aleatório para garantir a condição de aleatoriedade. Outro ponto importante é suposição de igualdade das variâncias dos erros. Como falado anteriormente, tal suposição deve ser verificada para garantir que a variabilidade dos erros é aproximadamente a mesma para todos os valores das variáveis independentes. Por fim, a suposição de normalidade dos erros é importante para garantir que sigam uma distribuição normal

Em relação ao erro aleatório (ε), é preciso realizar a estimação de sua variância para realização dos testes de significância do modelo. Tais testes são válidos para as regressões univariada e multivariada pois tem o mesmo propósito. Montgomery et al. (2003) explicam que os testes de significância são os balizadores para definição se o modelo estimado é adequado ou não. Para estimar o termo da variância do erro, pode-se realizar o cálculo da soma dos quadrados dos resíduos (SQ_E). A SQ_E consiste na soma das diferenças ao quadrado do termo das observações reais (y_i) com os valores estimados ($\hat{y}_i$)

$$SQ_E = \sum_{i=1}^{n} \left(y_i - \hat{Y}_i\right)^2 \tag{2.18}$$

Assim como os resíduos, é possível calcular a soma dos quadrados da regressão (SQ_R) que consiste na soma das diferenças ao quadrado dos valores estimados ($\hat{y}_i$) da média dos valores observados ($\bar{y}$).

$$SQ_R = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{y}\right)^2 \tag{2.19}$$

De posse dos termos calculados, podemos obter a soma dos quadrados totais (SQ_T) da forma abaixo e reescrever quaisquer temos em função de outro.

$$SQ_T = SQ_R + SQ_E \tag{2.20}$$

Segundo Montgomery et al. (2003), reescrevendo a equação (2.18) em termos de SQ_E, temos que $SQ_E = SQ_T - SQ_R$ que é:

$$SQ_E = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \tag{2.21}$$

Tomando a esperança dos termos e reescrevendo a equação, Montgomery et al. (2003) explicam que um estimador não tendencioso para σ^2 é:

$$\sigma^2 = \frac{SQ_E}{n-p} \tag{2.22}$$

2.1.5 Testes de hipótese e análise de resíduos

2.1.5.1 Teste de significância (Teste F)

Conforme descrito anteriormente, Sharpe (2000) explicita a necessidade de verificar as adequações do modelo. O teste de significância, ou teste F, verifica se existe uma relação linear entre a variável resposta e o as variáveis independentes repressoras. O teste de significância envolve o cálculo de um valor-p, que é a probabilidade de obter um resultado igual ou mais extremo do que o observado, assumindo que a hipótese nula é verdadeira. Se o valor-p for menor do que um nível de significância pré-determinado a hipótese nula é rejeitada, indicando que há evidências estatísticas suficientes para suportar a hipótese alternativa de que há uma relação linear significativa entre as variáveis independentes e dependentes. As hipóteses são:

$$\begin{aligned} H_0 &: \beta_1 = \cdots = \beta_k = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{2.23}$$

A estatística F é descrita por Montgomery et al. (2003) como:

$$F_0 = \frac{\frac{SQ_R}{k}}{\frac{SQ_E}{(n-p)}} = \frac{MQ_R}{MQ_E} \tag{2.24}$$

Rejeita-se H_0 quando:

$$F_0 > \quad f_{\alpha,k,n-p} \tag{2.25}$$

2.1.5.2 Análise de resíduos

Pelli Neto (2003) argumenta que a análise de resíduos é uma etapa crucial para avaliar a qualidade de ajuste de um modelo de regressão. Para Sharpe (2000) é necessário avaliar a homoscedasticidade, normalidade e não-correlação dos resíduos. A correlação entre os erros pode indicar que o modelo de regressão não é adequado para representar a relação de dependência entre as variáveis. A análise gráfica dos dados e dos resíduos pode ajudar a detectar essas características, especialmente em dados coletados ao longo do tempo.

São métodos e testes amplamente utilizados para análise de resíduos:

- Gráfico Q-Q Plot para verificação de normalidade.
- Segundo Razali et al. (2011), para dar suporte aos métodos gráficos, métodos mais formais e numéricos devem ser realizados antes de tirar qualquer conclusão sobre a normalidade dos sendo opções os testes de Anderson-Darling, Shapiro-Wilk, Kolmogorov-Smirnov e Lilliefors

- Para teste da independência dos resíduos, o teste de Durbin-Watson deve ser utilizado, caso a independência seja violada, o modelo não representa de forma adequada a linearidade dos dados
- Gráfico de resíduos x valores ajustados para verificar homoscedasticidade
- O pressuposto da homoscedasticidade dos resíduos também pode ser verificado pelos testes de Breusch-Pagan ou Goldfeld-Quandt.

Existem outros testes que podem ser utilizados para verificar a qualidade do ajuste do modelo em relação aos resíduos produzidos. Os descritos anteriormente são os mais utilizados pela literatura e que já podem detectar sua adequação ou não.

2.2 Modelos Lineares Generalizados e regressão não-linear

Modelos Lineares Generalizados (MLG) são uma extensão dos modelos lineares de regressão múltipla que foram propostos por Nelder et al. (1972). Abrangem uma ampla gama de modelos estatísticos aplicáveis em diversas áreas de pesquisa. Os MLG apresentam vantagens em relação aos modelos clássicos, pois permitem a modelagem de variáveis resposta que seguem a família exponencial de distribuições, como por exemplo, distribuição binomial, Poisson e gama.

Outro ponto é que os modelos lineares generalizados fornecem maior flexibilidade na relação funcional entre a média da variável resposta e o preditor linear. Isso é possível devido à introdução de uma função de ligação que pode ser escolhida de acordo com as características do problema em questão. Essa função não precisa ser necessariamente a identidade e pode assumir qualquer forma monótona não-linear.

Os MLG estão dentro de uma classe notável dentro da estatística, a família exponencial. A família exponencial de distribuições é uma classe paramétrica que engloba muitas distribuições conhecidas, como a distribuição normal, binomial, binomial negativa, gama, Poisson, normal inversa, multinominal, beta e logarítmica, entre outras. Essa família de distribuições apresenta propriedades estatísticas importantes, como a propriedade de suficiência estatística, que a tornam amplamente utilizada na teoria estatística.

Na análise de modelos de regressão linear, é comum utilizar distribuições dentro da família exponencial como distribuições para a variável resposta. Isso ocorre devido às propriedades matemáticas convenientes dessas distribuições e à flexibilidade que elas oferecem para modelar a relação entre a média da variável resposta e os preditores.

É importante ressaltar que a família exponencial de distribuições não é restrita apenas a modelos de regressão linear, mas também pode ser utilizada em outras áreas da estatística, como em análise de sobrevivência e modelos de contagem.

Os modelos lineares da família exponencial podem ser escritos de forma generalizada como:

$$f\left(y_i|\theta_i,\phi\right)=\exp(a^{-1}\phi^{-1})\left[y_i\theta_i-b\left(\theta_i\right)\right]+c\left(y_i,\phi\right) \tag{2.26}$$

Onde, $a\left(.\right)$,$b\left(.\right)$e $c\left(.\right)$ são conhecidos, $\phi>0$ é o parâmetro de disperão conhecido e θ_i o parâmetro canônico. Segundo Costa (2019), ϕ está associado unicamente à variância da resposta. Segundo o autor, são propriedades da família exponencial:

$$\begin{aligned} &E\left(Y\right)=b'\left(\theta\right)\\ &VAR\left(Y\right)=a\left(\phi\right).b''\left(\theta_i\right) \end{aligned} \tag{2.27}$$

Segundo Costa (2019) algoritmo para estimação dos MLG foi proposto por McCullagh et al. (1989), e tem objetivo de maximizar a função de verossimilhança. Para análise da qualidade dos ajustes, observa-se a minimização da função Deviance na forma:

$$\begin{aligned} &D_p\ \left(y;\mu\right)=2\left[l\left(y;y\right)-l\left(\mu;y\right)\right]\\ &\mu=E\left(Y\right)\\ &l\left(y;y\right)\left(1\right)\\ &l\left(\mu;y\right)\left(2\right) \end{aligned} \tag{2.28}$$

Onde (1) é função log-verossilmilhança saturada e (2) função log-verossilmilhança. As funções de log-verossimilhança saturada e log-verossimilhança se modificam de acordo com as distribuições como mostra a figura abaixo:

Figura 1 – Funções desvios para alguns modelos

Modelo	Desvio
Normal	$D_p=\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$
Binomial	$D_p=2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)+(m_i-y_i)\log\left(\frac{m_i-y_i}{m_i-\hat{\mu}_i}\right)\right]$
Poisson	$D_p=2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)+(\hat{\mu}_i-y_i)\right]$
Binomial negativo	$D_p=2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)+(y_i+k)\log\left(\frac{\hat{\mu}_i+k}{y_i+k}\right)\right]$
Gama	$D_p=2\sum_{i=1}^{n}\left[\log\left(\frac{\hat{\mu}_i}{y_i}\right)+\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right]$
Normal inverso	$D_p=\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{y_i\hat{\mu}_i^2}$

Fonte: Cordeiro et al. (2008).

Para Costa (2019), é usual a comparação do valor observado da função Deviance com a distribuição Qui-Quadrado com *n-p*graus de liberdade, onde n é o tamanho da amostra e k é o número de parâmetros do modelo (incluindo intercepto). Dessa forma o autor afirma que assintoticamente a estatística D sob hipótese de o modelo é compatível com o comportamento esperado dos dados, é comparado a distribuição Qui-Quadrado com n-p graus de liberdade, na forma:

$$\begin{aligned} D \mid H_0 \quad & \chi^2_{n-p} \\ valor - p = 1 - & \chi^2_{n-p} \end{aligned} \tag{2.29}$$

Para Sousa (2010), em termos gerais, há superdispersão dos dados em relação ao modelo ajustado, ou seja, a variância esperada é maior que a prevista para o modelo, quando o valor-p é 0. A subdispersão é o fenômeno contrário ao da superdispersão e ocorre quando o valor-p é igual a 1. As soluções para os fenômenos citados é a procura de outro modelo que melhor se encaixe na distribuição de dados ou realizar a estimação do parâmetro ϕ no intervalo $0 < \phi < 1$, para ajustar a qualidade do modelo.

2.2.1 Modelo Gama

Dentro da família exponencial, o modelo gama é usado nas situações em que a variância do modelo é proporcional ao quadrado da média ($VAR\,(Y) \propto \mu^2$).

A distribuição Gama tem a seguinte densidade de probabilidade:

$$f\,(y) = \frac{y^{-1}}{\Gamma\,(\nu)} \left(\frac{y\nu}{\mu}\right) exp\left\{\frac{-y\nu}{\mu}\right\} \tag{2.30}$$

Onde ν é o parâmetro de dispersão e Γ é definido por:

$$\Gamma\,(u) = \int_0^\infty x^{u-1} e^x dx \tag{2.31}$$

As propriedades da distribuição Gama são:

$$\begin{aligned} E\,(Y) &= (\mu) \\ VAR\,(Y) &= \nu^{-1}\mu^2 \end{aligned} \tag{2.32}$$

2.3 Regressão Quantílica

Como descrito anteriormente, os estudos estatísticos relacionados a regressão linear em sua maioria buscam estimar as variações de uma variável Y em relação as variáveis dependentes x de acordo com efeito na média, conforme descrito pelo método MQO. Segundo Santos (2012), tal método possui limitações pois está intrinsicamente baseado na assunção da distribuição normal dos erros, dessa forma quando tal condição não é

alcançada, a performance do método é comprometida. Ainda segundo o autor, ainda há o problema da influência que outliers exercem nas estimativas dos parâmetros do modelo, fazendo com que seja necessário um maior cuidado na avaliação na influência dos mesmos sobre o ajuste do modelo.

Uma alternativa ao MQO é a utilização da técnica de minimização de erros absolutos ponderados (MEA) no qual resulta em um modelo de regressão baseado em quantis, em outros termos, uma regressão quantílica. O MEA é um método não-paramétrico pois não depende de uma distribuição pré-determinada e permite investigar os efeitos das variáveis explicativas ao longo da distribuição da variável dependente. De forma geral, a regressão quantílica é uma forma de analisar e estimar modelo para a mediana condicional ou para outros quantis da distribuição da variável dependente Y com relação a regressores x.

Reescrevendo a forma geral da equação de regressão linear condicionada para a média para a forma quantílica, temos a seguinte representação:

$$Q_\tau\left(Y|\ x\right) = \beta_0\left(\tau\right) + \beta_1 x_1\left(\tau\right)\cdots\ + \beta_k\left(\tau\right) \tag{2.33}$$

Onde τ é o quantil condicional e k representa as variáveis dependentes.

Para obter o estimador $\widehat{\beta}\left(\tau\right)$ é preciso minimizar a soma dos erros absolutos ponderados (MEA) representadas pela equação desenvolvida por Koenker (2005) abaixo:

$$min\ \sum_{i=1}^{n} \rho_\tau(y_i\ \ -\ x_i{}'\beta) \tag{2.34}$$

Nessa equação a função ρ_τ representa a função de perda onde é possível obter o valor para um determinado quantil minimizando a perda esperada para Y. Koenker (2005) demonstrou em seus estudos que é possível obter os coeficientes de regressão para a regressão quantílica reformulando o problema acima em formato de programação linear na forma:

$$\begin{aligned} \min \quad & \sum_{j=1}^{n} \tau_j e_{1j} + (1-\tau) e_{2j} \\ \text{s.a.:} \quad & y_j = \alpha_j + \beta_{kj} x_{kj} + e_{1j} + e_{2j} \\ & \beta_k \geq 0 \\ & e_j = e_{1j} + e_{2j} \end{aligned} \tag{2.35}$$

Para o caso em que $\tau = 0.5$ a estimativa é realizada para a mediana.

2.4 Critério de Informação de Akaike

Beatty et al. (2018) desenvolveram uma metodologia que endereça a problemática da seleção de modelos. Segundo Sobral et al. (2016) não existe na literatura uma metodologia única que atue para selecionar modelos matemáticos. No contexto da regressão linear, há a possibilidade de testar diversas configurações de modelos adicionando ou removendo as variáveis dependentes que fazem parte de sua formulação.

Uma das alternativas largamente implementadas é o Critério de Informação de Akaike (AIC). Segundo Sobral et al. (2016) a metodologia consiste na utilização da divergência de Kullback-Lieber, que é uma espécie de medida de distância entre o modelo analisado e outro teórico. Beatty et al. (2018) propuseram uma forma de estimação da distância entre esses dois modelos utilizando a função de verossimilhança de ordem do modelo:

$$AIC = -2 \ln f\left(x \mid \hat{\theta}\right) + 2k \tag{2.36}$$

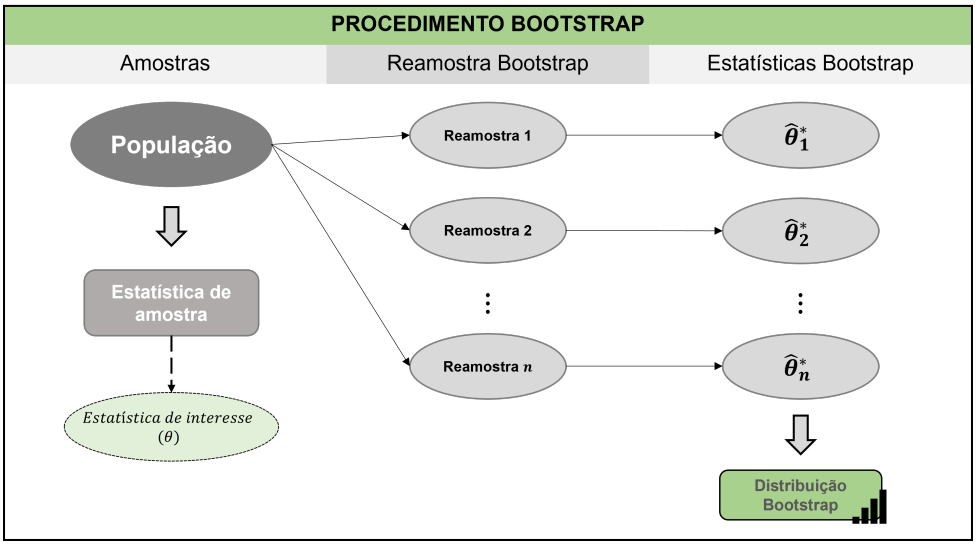
À medida que a verossimilhança aumenta, o termo $-2 \ln f\left(x \mid \hat{\theta}\right)$ decresce, enquanto o termo $2k$ cresce sempre que a ordem do modelo for maior. Dessa forma, o AIC pondera entre a adequação aos dados e a complexidade do modelo. Em outros termos, quando se tem mais de um modelo para comparar, aquele que apresentar menor valor no AIC, será aquele com melhor ajuste.

2.5 *Bootstrap*

Bootstrap é uma técnica estatística desenvolvida por Efron (1979) que consiste em reamostragem aleatória para estimar a distribuição de um conjunto de dados. Inicialmente o método foi desenvolvido como uma alternativa a outra técnica estatística de reamostragem (*jackknife*) e sua utilização se mostrou mais ampla e aplicável.

Segundo o autor, dada uma variável aleatória com distribuição desconhecida e, deseja-se estimar a distribuição dos dados a partir da amostra. É assumido que a amostra pode ser representada como uma pseudopopulação, com características da verdadeira população. Por meio da geração repetida de amostras aleatórias (reamostras) desta pseudopopulação, a distribuição de amostragem de uma estatística pode ser estimada. Dessa forma, o procedimento descrito anteriormente pode ser considerado como *bootstrap* não-paramétrico, já que não é necessário o conhecimento da distribuição dos dados a priori. Cunha et al. (2003) descrevem em seu trabalho a utilização da técnica de *bootstrap* paramétrico, procedimento que tem vasta utilização em diversos campos de estudo. Entretanto, o tema não será abordado neste trabalho. Abaixo a Figura 2 descreve o procedimento *bootstrap*:

Figura 2 – Quadro explicativo da técnica *bootstrap*



Fonte: Autor.

2.5.1 Intervalo de confiança percentílico para *bootstrap*

Segundo Neyman (1937) o termo intervalo de confiança (IC) se refere a um intervalo de valores, calculado a partir de uma amostra aleatória, dentro do qual é provável que o valor de um parâmetro populacional desconhecido esteja contido com um certo nível de confiança. Esse nível de confiança é geralmente expresso em termos de uma porcentagem, como 95%, e indica a probabilidade de que o intervalo de confiança contenha o verdadeiro valor do parâmetro.

Autores como Hung et al. (2014) afirmaram que o intervalo de confiança pode ser usado para estimar o grau de incerteza associado à estimativa do parâmetro e fornecer informações sobre a precisão da estimativa. Elaborando, os intervalos de confiança são importantes para a estatística porque fornecem uma medida da incerteza associada a uma estimativa de parâmetro populacional a partir de uma amostra aleatória. Essa medida de incerteza é crítica para interpretar os resultados de estudos estatísticos e para tomar decisões baseadas em evidências estatísticas.

Para a técnica de bootstrap não-paramétrico é possível utilizar o intervalo de confiança percentílico, que segundo Franco et al. (2005) pode ser descrito da seguinte forma:

1. É gerado uma quantidade B de amostras de *bootstrap* a partir dos dados observados
2. Em seguida o parâmetro de estatístico de interesse estimado em cada uma dessas reamostras *bootstrap* é calculado $\left(\hat{\delta}^*\right)$
3. Os valores do parâmetro estatístico calculados são ordenados em ordem crescente
4. Os percentis de interesse para construir o intervalo de confiança são selecionados (α).

Podemos definir os limites interior e superior de um intervalo de confiança *bootstrap* percentílico ao nível de $(1-\alpha)$ como sendo:

$$IC^*_{1-\alpha} = \left[\hat{\delta}^*_{\alpha/2}; \hat{\delta}^*_{1-(\alpha/2)}\right] \tag{2.37}$$

2.6 Programação Linear

Dantzig (2002) descreve a programação linear como uma técnica matemática utilizada para maximizar ou minimizar uma função linear sujeita a um conjunto de restrições lineares. Essa técnica é amplamente utilizada em problemas de otimização em áreas como economia, engenharia, logística, entre outras. Para Baio et al. (2004), programação linear é uma técnica para o planejamento de atividades para alcançar um resultado ideal, levando em consideração as opções possíveis. Em um problema de programação linear, normalmente existem diversas soluções possíveis, desde que todas as restrições sejam cumpridas, embora haja apenas uma solução ótima.

Frossard (2009) descreve que o objetivo do modelo linear é descobrir o valor mais adequado para uma função, mediante a imposição de um conjunto de restrições lineares de natureza estrita e não estrita. É um modelo matemático de otimização linear, que engloba variáveis de decisão, uma função-objetivo e restrições técnicas expressas por meio de inequações lineares. As inequações lineares, que representam as restrições técnicas, devem ser satisfeitas simultaneamente pelas variáveis de decisão para que a função-objetivo seja maximizada ou minimizada. O modelo linear é amplamente utilizado em diversas áreas para resolver problemas de otimização.

A formulação de um problema de programação linear pode ser escrita da seguinte forma, conforme proposta por Lewis (2008):

$$\begin{array}{lccccccccc}
\text{Minimizar} & c_1x_1 & + & c_1x_1 & + & \cdots & + & c_nx_n & = & z \\
\\
\text{Sujeito a:} & a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & \leq & b_1 \\
 & \vdots & & \vdots & & & & \vdots & \vdots & \vdots \\
 & a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & \leq & b_m \\
 & x_1, & & x_2, & & x_3, & \ldots & \ldots & x_n \geq & 0
\end{array} \tag{2.38}$$

As variáveis x_1,$x_2 \ldots x_n$ são chamados de variáveis de decisão, e seus valores estão sujeitos a $m+1$ restrições. Um conjunto de x_1, $x_2 \ldots x_n$ satisfazendo todas as restrições é chamado de ponto viável e o conjunto de todos esses pontos é chamado de região viável. A solução do programa linear deve ser um ponto $(x_1,\ x_2 \ldots x_n)$ na região viável, ou então nem todas as restrições seriam satisfeitas.

Aplicando a forma de álgebra linear, temos que:

$$\begin{array}{lrcll} \text{Minimizar} & \sum_{j=1}^{n} c_j x_j & = & z & \\ \text{Sujeito a:} & \sum_{j=1}^{n} a_j x_j & \leq & b & \\ & x_j & \geq & 0 & \forall \quad j = 1, \ldots, n \end{array} \tag{2.39}$$

Na forma matricial:

$$\begin{array}{lrcl} \text{Minimizar} & c\mathbf{x} & = & z \\ \text{Sujeito a:} & \mathbf{Ax} & \leq & \mathbf{b} \\ & \mathbf{x} & \geq & 0 \end{array} \tag{2.40}$$

A matriz **A** possui formato mxn cuja j-ésima coluna é a_j . Esta matriz corresponde aos coeficientes em x_1,x_2,... x_n nas restrições de um problema de programação linear. O vetor **x** é um vetor de soluções para o problema, **b** é o vetor do lado direito e **c** é o vetor de coeficientes.

Dantzig (1951) desenvolveu um método para resolver os problemas de programação linear. O Simplex é um algoritmo utilizado para resolver problemas de otimização em que é necessário encontrar a melhor solução entre várias opções. Ele funciona iterativamente, explorando soluções viáveis até encontrar a solução ótima, que é aquela que maximiza ou minimiza a função objetivo sujeita às restrições do problema. O método Simplex é baseado em uma geometria simples, utilizando conceitos de operações matriciais para encontrar as soluções ótimas.

Lima Pinto et al. (2008) descrevem que a ideia geométrica subjacente consiste em percorrer, passo a passo, os pontos extremos adjacente do conjunto viável em um problema de programação linear, aprimorando o valor da função objetivo a cada extremo visitado.

Atualmente o algoritmo Simplex é utilizado computacionalmente de forma ampla por vários programas relacionados a resolução de problemas que envolvem programação linear, Fearnley et al. (2015) afirmam que o método é conhecido por performar bem na prática, apesar de existir métodos mais modernos que reduzem a complexidade computacional necessária para resolver problemas muito grandes.

3 Metodologia

Para a elaboração deste trabalho, optou-se pela abordagem de uma pesquisa descritiva quantitativa e que dispõe de um estudo de caso. Para Silva e Simon (2005) as pesquisas quantitativas são mais adequadas nas ocasiões em que se sabe as qualidades do objeto de estudo e se tem o controle do que vai se pesquisar. Para Silva e Lopes (2014), os dados da abordagem quantitativa têm necessariamente natureza numérica, como grandezas monetárias, físicas ou escala de atitude, portanto é imperativo o uso de tal método.

Godoy (1995) descreve o estudo de caso como um tipo de pesquisa no qual o objeto é uma parte que se analisa de forma profunda. O autor ainda acrescenta que a estratégia do estudo de caso procura responder questões relacionada as razões e modos sobre como certos fenômenos ocorrem e no enfoque exploratório o pesquisador deve estar aberto a novas descobertas.

Para a implementação dos modelos estatísticos, foi utilizado o software R na versão 4.3.1 *"Beagle Scouts"*. O *script* contendo os comandos para implementação dos modelos estatísticos que serão exibidos posteriormente, foram executados em um computador pessoal do modelo *MacBook Pro* da marca *Apple*, contendo processador de 2,4 GHz *Intel Core i5 Quad-Core* e memória *RAM* da especificação 8 *GB* 2133 MHz LPDDR3 no sistema operacional *macOS Ventura* 13.4.1 (22F770820d) (R CORE TEAM, 2023).

Foi escolhida a linguagem R como ferramenta para realização desse estudo de caso pois se trata de uma ferramenta poderosa que permite realizar análises estatísticas complexas e outras técnicas matemáticas com alto custo computacional com uma performance razoável. Segundo Grover et al. (2017), R é uma linguagem de programação e ambiente de software para análise estatística, com recursos para manipulação de dados e visualização gráfica. É de código aberto e suporta várias técnicas estatísticas, como modelagem linear e não linear, testes estatísticos, análise de séries temporais e muito mais. Além disso, R foi aplicado em diversas áreas, incluindo discussões sociais específicas de domínio, como cadeia de suprimentos e internet das coisas, bem como no monitoramento de impactos de eventos em discussões sociais. A literatura também fornece um tutorial sobre R para implementação paralela visando desempenho escalável.

Outro aspecto podereso da linguagem é a possibilidade do emprego de pacotes. O uso de pacotes dentro do R é altamente vantajoso para analistas de dados e cientistas que trabalham com essa linguagem. Esses pacotes são essenciais porque expandem significativamente as capacidades do R, permitindo que os usuários aproveitem uma variedade de funcionalidades pré-construídas para tarefas específicas. Em vez de construir algoritmos sem nenhuma referência, os pacotes oferecem algoritmos, funções e ferramentas de alta qualidade, economizando tempo e esforço. Além disso, a comunidade R é prolífica

na criação de pacotes, o que significa que quase qualquer tarefa de análise de dados já possui uma solução pronta para uso. Isso promove uma abordagem eficiente e colaborativa na resolução de problemas complexos. A flexibilidade e a personalização dos pacotes também permitem que os usuários adaptem o ambiente R às suas necessidades específicas, tornando-o uma ferramenta poderosa e versátil para lidar com uma ampla gama de desafios de análise de dados. Por fim, o uso de pacotes dentro do R não apenas simplifica o processo de análise, mas também amplia o horizonte das possibilidades, tornando-o fundamental para profissionais que trabalham com dados.

Abaixo há uma tabela-resumo contendo os pacotes que foram utilizados na elaboração do *script* do trabalho, bem como as fontes dos seus autores e uma breve descrição do uso:

Tabela 1 – Pacotes do R utilizados no trabalho

Pacote	**Autor**	**Uso no trabalho**
boot	Davison et al. (1997)	Aplicação na técnica de *bootstrap*
dplyr	Wickham, François et al. (2023)	Manipulação de tabelas
exploreR	Coates (2016)	Manipulação rápida de dados
forecast	Hyndman et al. (2008)	Criação de modelos de previsão
ggplot2	Wickham (2016)	Criação de gráficos
ggalt	Rudis et al. (2017)	Geometrias adicionais para criar gráficos
janitor	Firke (2023)	Limpeza e preparação de dados
lpsolve	Berkelaar et al. (2023)	Interface para programação linear
MASS	Venables et al. (2002)	Ajuste de modelo estatístico de regressão
mvshapiro	Gonzalez-Estrada et al. (2013)	Teste de Shapiro
openxlsx	Schauberger et al. (2023)	Leitura de planilhas eletrônicas da extensão *.xlsx*
RcolorBrewer	Neuwirth (2022)	Implementação de paleta de cores nos gráficos
scales	Wickham e Seidel (2022)	Formatação das escalas dos gráficos
tidyverse	Wickham, Averick et al. (2019)	Ecossistema necessário para funcionamento de pacotes

Fonte: Autor.

Por fim, toda a documentação utilizada e produzida por esse trabalho, estão disponíveis no seguinte repositório da plataforma *GitHub*: `https://github.com/igormazzeto/tccufmg`

4 Estudo de caso

4.1 Escolha do tema

A relevância de estudar os custos operacionais das concessionárias que atuam no setor elétrico brasileiro é fundamental para garantir a eficiência e a transparência das operações nesse setor estratégico. Os custos operacionais representam uma parcela significativa dos gastos das concessionárias, e compreender sua composição e evolução ao longo do tempo é essencial para o planejamento, a tomada de decisões e a definição de tarifas justas para os consumidores.

Dentro do processo regulatório de aferição das receitas anuais periódicas pela ANEEL, os custos operacionais desempenham um papel crucial. A agência reguladora utiliza esses custos como base para determinar as tarifas que as concessionárias podem cobrar dos consumidores. Portanto, uma análise precisa e atualizada dos custos operacionais é necessária para garantir que as tarifas sejam justas e suficientes para manter a qualidade e a confiabilidade do fornecimento de energia elétrica.

Uma nova forma de estimação dos custos operacionais que contemple as restrições técnicas inerentes à operação de transmissão de energia elétrica pode ser significativa. Uma mudança nesse sentido poderia afetar os investimentos das concessionárias, a manutenção da infraestrutura e, consequentemente, a qualidade do serviço prestado. Além disso, essa alteração poderia ter implicações nas tarifas pagas pelos consumidores e no planejamento energético do país, uma vez que afetaria a competitividade do setor elétrico e a segurança do abastecimento.

Portanto, estudar os custos operacionais das concessionárias no setor elétrico brasileiro é vital para garantir a eficiência e a equidade desse setor. Os custos operacionais desempenham um papel central no processo regulatório de definição das tarifas e têm impactos profundos no funcionamento do setor e na satisfação dos consumidores. Qualquer proposta de mudança na estimação desses custos deve ser cuidadosamente avaliada para minimizar riscos e assegurar um sistema elétrico confiável e acessível a todos.

4.2 Descrição dos dados

Conforme descrito anteriormente, a base de dados utilizada nessa pesquisa passou por processo de avaliação crítica, auditoria e validação pelos operadores do SEB e pela sociedade. O conjunto está disponível online no portal da reguladora (https://www.gov.br/aneel/pt-br), com todas as informações relativas à tomada de subsídio.

A base de dados foi disponibilizada em forma de planilha eletrônica contendo 125

observações (linhas) e 20 colunas. Os dados coletados são referentes aos dados fornecidos pelas concessionárias e agrupadas pela reguladora dos anos de 2017 a 2021. A planilha contempla as seguintes informações conforme descrito pela NT 97/2022–SRM/ANEEL:

- Concessionária
- Tipo (holding, não-licitada)
- Ano
- Identificação dos agentes
- Custos operacionais contábeis (PMSO)
- Ativos físicos
  - Equipamentos de rede com tensionamento menor de 230 kV
  - Equipamentos de rede com tensionamento maior de 230 kV
  - Equipamentos de subestação com tensionamento menor de 230 kV
  - Equipamentos de subestação com tensionamento maior de 230 kV
  - Módulos de manobra com tensionamento menor de 230 kV
  - Módulos de manobra com tensionamento maior de 230 kV
- Potência total de equipamentos de subestação
- Potência aparente: MVA
- Potência reativa: MVar
- Indisponibilidade de rede (não utilizado na pesquisa)
- Idade Média (não utilizado na pesquisa)
- Adversidade (não utilizado na pesquisa)
- Resultados do modelo de DEA (4 campos não utilizados na pesquisa)

Na NT 97/2022–SRM/ANEEL o regulador explicita que foram identificadas e agrupadas as empresas que efetivamente realizam algum tipo de compartilhamento operacional. A estratégia tem por objetivo mitigar potenciais inconsistências na alocação de custos em algumas transmissoras e Sociedades de Propósito Específico (SPE) que pertencem ao mesmo grupo econômico e eventualmente compartilham custos. Esse agrupamento é importante de se ressaltar pois pode trazer distorções nos dados uma vez que, agrupamentos ou holdings artificiais trazem assimetrias entre agentes que pouco tem semelhanças em termos de custo, ativos físicos e potência de equipamentos. Outro ponto, é que o regulador eliminou da consolidação final de dados com PMSO negativo, PMSO nulo e sem produtos. Dessa forma, na base de dados constam 28 agrupamentos de concessionárias.

4.3 Descrição dos dados

4.3.1 Estatísticas descritivas

Abaixo temos as descrições das variáveis estudadas. Para fins de facilitar a visualização de informações, foi utilizada uma forma reduzida de atribuição enumerando as variáveis de 1 a 8 de acordo com o coeficiente de determinação (R^2) ordenado de maior para menor obtido por meio de modelo de regressão linear multivariado multivariado previamente realizado para exploração de dados:

- X_1 : Extensão de equipamentos de rede com tensionamento maior de 230 kV
- X_2 : Módulos de manobra com tensão igual ou superior a 230 kV
- X_3 : Equipamentos de rede com tensionamento maior de 230 kV
- X_4 : Potência aparente total, em MVA, de equipamentos de subestação
- X_5 : Potência reativa total, em Mvar, de equipamentos de subestação
- X_6 : Equipamentos de subestação com tensão inferior a 230 kV
- X_7 : Módulos de manobra com tensão inferior a 230 kV
- X_8 : Extensão de equipamentos de rede com tensionamento maior de 230 kV

Abaixo temos as principais medidas de estatística descritiva para as variáveis dependentes e independentes utilizadas no estudo:

Tabela 2 – Estatísticas descritivas das variáveis dependente e independentes

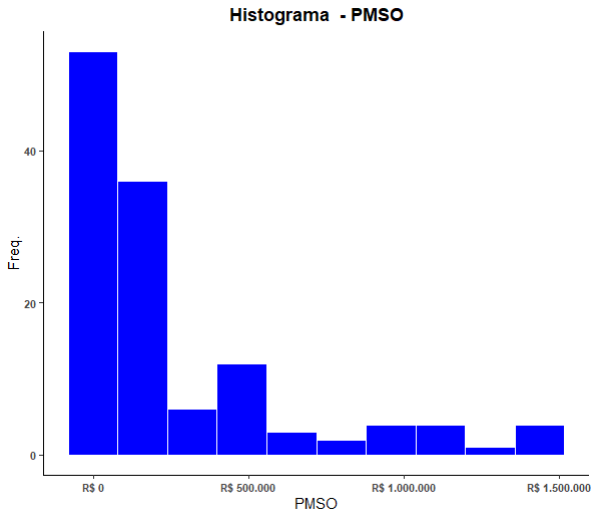
	$PMSO$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Mínimo	1.273,0	0,0	4,0	7,0	0,0	0,0	0,0	0,0	0,0
1º Quartil	30.523,0	773,6	68,0	56,0	2.348,0	977,0	2,0	17,0	0,0
Mediana	103.467,0	3.517,7	138,0	116,0	7.500,0	4.227,0	5,0	43,0	45,2
Média	270.875,0	4.693,6	270,4	209,0	18.441,0	6.467,0	64,8	313,4	787,9
3º Quartil	262.355,0	6.848,9	345,0	275,0	19.527,0	9.485,0	80,0	375,0	500,7
Máximo	1.439.704,0	18.376,7	1.218,0	763,0	98.256,0	41.208,0	346,0	1.841,0	7.297,6

Fonte: Autor.

Pode-se observar na tabela acima que a amplitude dos valores do PMSO é bastante elevada, na ordem de mais de R$ 1,3 milhões entre o menor valor encontrado e o máximo. É importante destacar tal fato pois tamanha divergência pode gerar distorções em quaisquer análises a serem realizadas posteriormente e as mesmas ocorrem devido ao agrupamento realizado pela própria reguladora para, segundo ela, mitigar potenciais inconsistências na alocação de custos em algumas transmissoras.

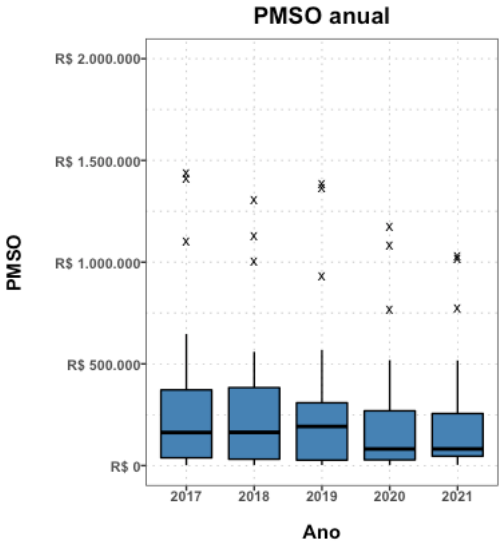
No histograma pode-se verificar que grande parte das transmissoras possuem custos operacionais (PMSO) menores que R$ 500.000,00. Analisando o PMSO entre 2017 e 2021 verifica-se uma consistência nos valores no decorrer dos anos. Em 2020 e 2021 observa-se ligeira redução, principalmente para as empresas registrando custos mais elevados.

Figura 3 – Gráfico de histograma



Fonte: Autor.

Figura 4 – *Boxplot* para os custos anuais

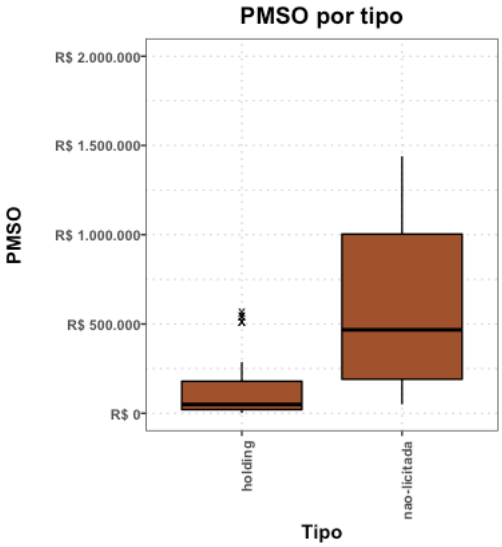


Fonte: Autor.

Quando analisamos o tipo de cada concessionária, verificamos que as não-licitadas mostram maior dispersão do PMSO, com variações relevantes do custo operacional ao longo dos anos. No contexto brasileiro as transmissoras não-licitadas são operadoras do SEB regidas sob concessão pública. São exemplos de transmissoras não-licitadas Companhia de Energia Elétrica de Minas Gerais (CEMIG), Centrais Elétricas do Norte do

Brasil S/A (ELETRONORTE), Furnas Centrais Elétricas S/A (FURNAS) e Companhia Hidroelétrica do São Francisco (CHESF).

Figura 5 – *Boxplot por tipo* por agrupamento de concessionária

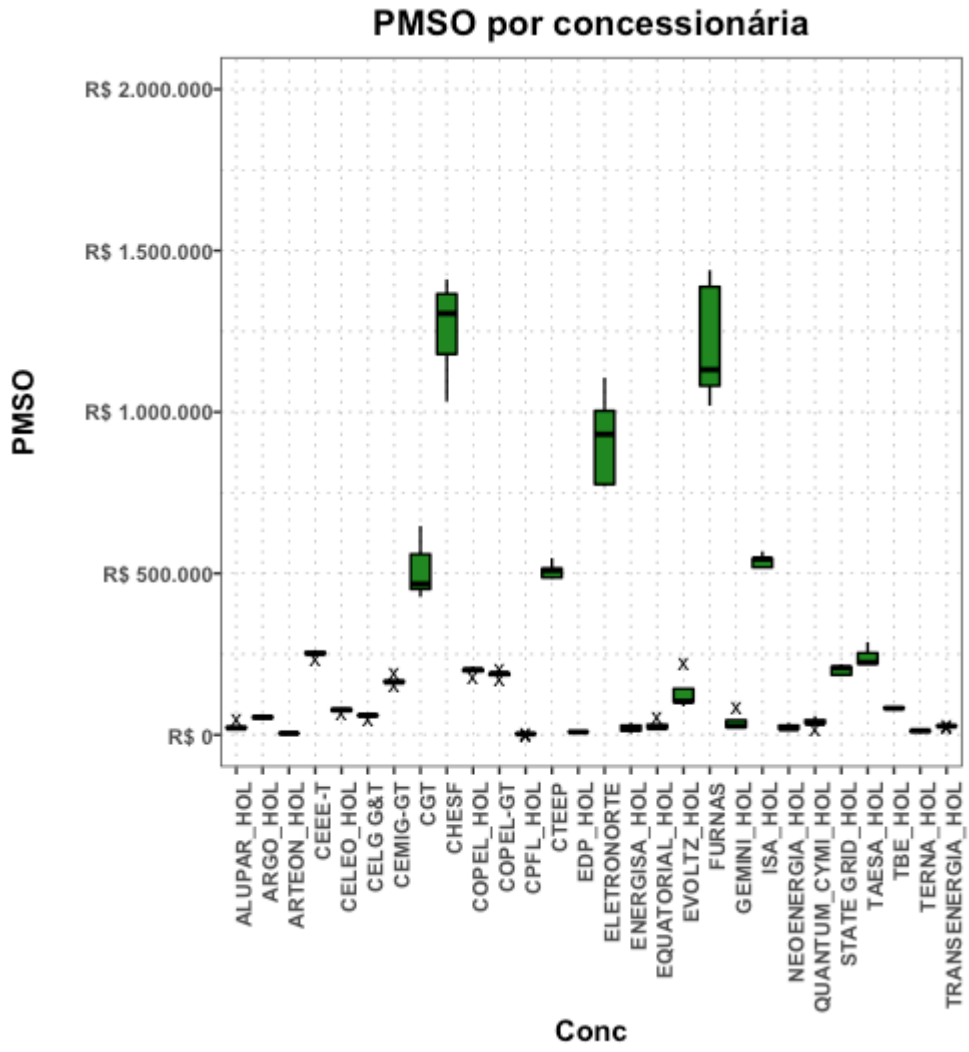


Fonte: Autor.

Em relação às companhias participantes da RAP, três se destacam por apresentaram maiores PMSO e maior dispersão dos custos ao longo do período observado: CHESF, ELETRONORTE e FURNAS. Tais companhias se destacam por pertencer ao Grupo ELETROBRAS, companhia que tem como principal acionista a União. CHESF se caracteriza por possuir ativos em todos os estados da região Nordeste, portanto a quantidade de recursos financeiros para sua operacionalização tende a serem elevados.

FURNAS possui ativos em diversos estados brasileiros como Minas Gerais, São Paulo, Rondônia, Rio Grande do Sul, Rio de Janeiro entre outros. Assim como a CHESF, devido a sua elevada extensão territorial e dispersão de ativos é de se esperar um elevado custo operacional. Por fim a ELETRONORTE possui ativos em 7 estados da região Norte além do estado do Maranhão, enfrentando assim, realidade muito similar as outras duas concessionárias que fazem parte do mesmo grupo econômico. Abaixo, o gráfico mostra o comparativo dos custos operacionais entre os agrupamentos propostos pela ANEEL:

Figura 6 – Custos por concessionária



Fonte: Autor.

4.3.2 Correlação

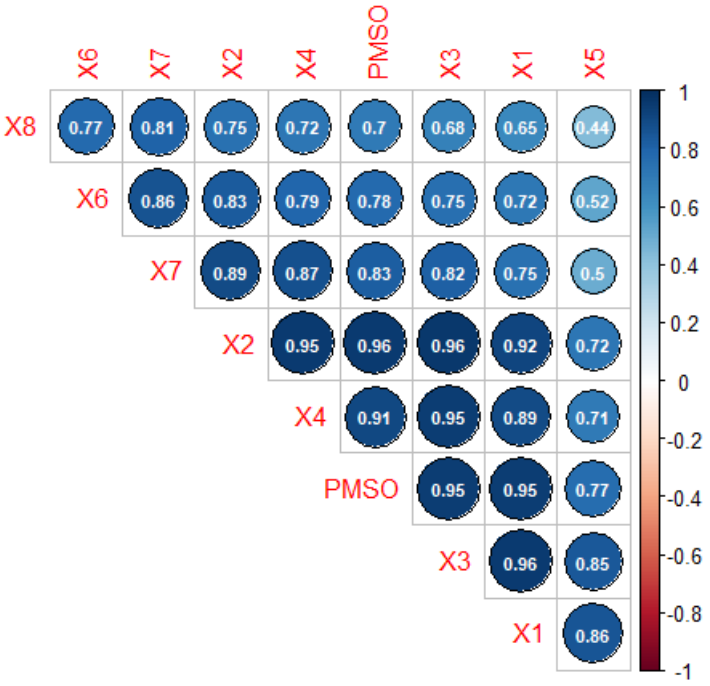
Para a análise de correlação foi utilizada a correlação de Spearman. Quando analisada a correlação entre as variáveis de interesse, nota-se correlação positiva elevada (acima de 0,70) entre a variável PMSO e as demais. Tal comportamento confirma a relação dos ativos dos físicos das concessionárias e seu impacto no custo operacional.

Entretanto, há correlações positivas elevadas entre diversas variáveis explicativas dependentes, mostrando a evidência de multicolinearidade no conjunto de dados estudados. Essa propriedade pode inviabilizar a utilização de modelos linearidades utilizando os pressupostos de MQO.

Analisando a natureza das variáveis aqui apresentadas é esperado que a multicolinearidade seja evidente. Por se tratar de ativos físicos que possuem característica técnicas similares ou até mesmo dependentes umas das outras, seria improvável que uma relação técnica não fosse observada em uma relação estatística. Portanto, a alta correlação entre

as variáveis dependentes era esperada e desejada, pois transmite a realidade operacional para eventuais modelos a serem implementados utilizando o conjunto de dados disponibilizado pela reguladora.

Figura 7 – Matriz de correlação



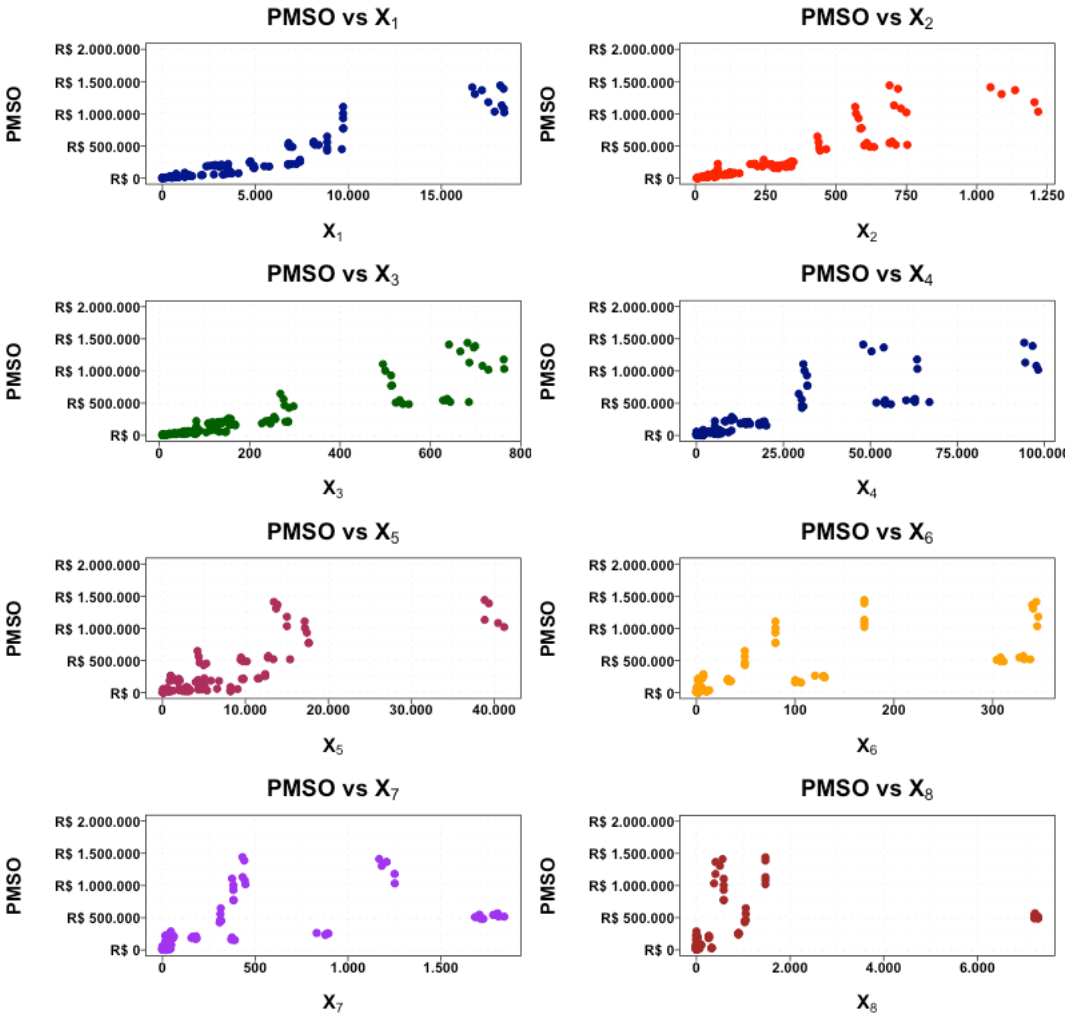
Fonte: Autor.

4.3.3 Dispersão

Analisando a dispersão do custo (PMSO) em função de cada variável dependente, observa-se que há um comportamento aparentemente linear para as variáveis X_1, X_2, X_3 e X_4. As variáveis $X_5 X_6, X_7$ e X_8 apresentam dispersão incompatíveis com um comportamento linear.

A imagem abaixo ilustra a variação do PMSO em relação as variáveis dependentes. É possível observar um comportamento não-linear para variações extremas nas variáveis dependentes:

Figura 8 – Gráficos de dispersão



Fonte: Autor.

4.4 Ajuste do modelo de regressão linear

Após as análises realizadas anteriormente, a fim de verificar a viabilidade da utilização da regressão linear múltipla respeitando as premissas operacionais descritas anteriormente, foi realizado um primeiro ajuste para o modelo.

Vale ressaltar que das 8 variáveis dependentes selecionadas para ajuste do modelo, nenhuma se refere ao modelo DEA previamente aplicado para o conjunto de dados aqui estudado. Abaixo temos os coeficientes obtidos pelo ajuste do modelo realizado no software R:

Tabela 3 – Coeficientes do ajuste do modelo

Regressão linear	
	Coeficientes
β_0	-64.039,21
β_1	13,32
β_2	794,1
β_3	169,74
β_4	**-0,52**
β_5	8,86
β_6	621,45
β_7	**-189,72**
β_8	**-8,01**

Fonte: Autor.

Apesar de ter apresentado um excelente coeficiente de determinação $R^2 = 0.92$, o modelo apresentou coeficientes de regressão (β) com valores negativos, violando a restrição operacional que representa a realidade das concessionárias.

Observando os pressupostos do modelo, os resultados dos testes de hipótese não foram satisfatórios. O a hipótese nula do Teste-F[1] foi rejeitada a um nível de significância de 95%, demonstrando que os coeficientes de regressão são significativos. Entretanto as hipóteses nulas dos testes de homoscedasticidade[2] (variância constante) e autocorrelação[3] foram rejeitadas, demostrando que o modelo não possui variância constante e autocorrelação entre as variáveis dependentes. Por se tratar de dados coletados em painel uma autocorrelação nos dados seria esperada, entretanto para uma questão formal de análise da qualidade do modelo, a aplicação do teste foi mantida.

Tabela 4 – Pressupostos do modelo

Testes dos pressupostos do modelo						
Teste	Pressuposto	Estatística	Hipótese nula	valor-p	Significância	**Veredicto**
Teste-F	Significância	$F = 0,92$	$\beta_1 = ...\beta_k = 0$	$2.20 \cdot 10^{-16}$	0,05	**Rejeitado**
Breusch-Pagan	Homoscedasticidade	$LM = 2,25$	$\delta_1 = ...\delta_k = 0$	$6,50 \cdot 10^{-8}$	0,05	**Rejeitado**
Durbin-Watson	Autocorrelação	$d = 1,07$	Correlação = 0	$2,46 \cdot 10^{-10}$	0,05	**Rejeitado**

Fonte: Autor.

Os testes de resíduos também demonstraram a violação do pressuposto de normalidade[4], conforme exibido pela tabela abaixo:

Tabela 5 – Normalidade do modelo

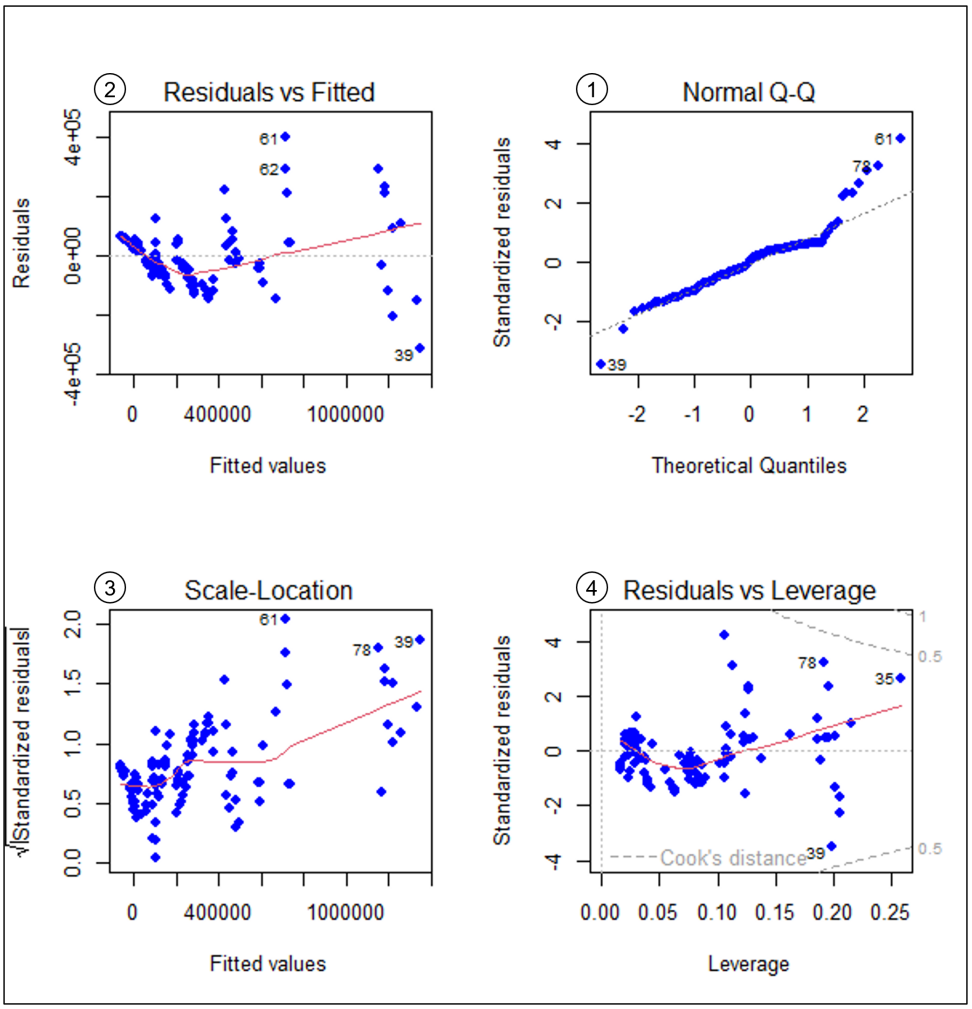
Testes de normalidade para os resíduos					
Teste	Estatística	Hipótese nula	valor-p	Significância	**Veredicto**
Shapiro-Wilk	$W = 0,92$	$H_0 : X \sim N$	$5.54 \cdot 10^{-6}$	0,05	**Rejeitado**
Anderon-Darling	$A = 2,25$	$H_0 : X \sim N$	$9,76 \cdot 10^{-6}$	0,05	**Rejeitado**
Kolmogorov-Smirnov	$D = 0,15$	$H_0 : X \sim N$	$2,03 \cdot 10^{-7}$	0,05	**Rejeitado**

Fonte: Autor.

[1] Sharpe (2000).
[2] Breusch et al. (1979).
[3] White (1992).
[4] Razali et al. (2011).

Por fim, é possível observar a inadequação dos modelos observando os gráficos dos resíduos do ajuste:

Figura 9 – Resultados - ajuste



Fonte: Autor.

No quadrante número 1, é possível verificar a não aderência dos resíduos ao gráfico de probabilidade normal, onde há observações discrepantes da linha de normalidade. Já no quadrante de número 2, tem-se a evidência gráfica da forma de um "funil" irregular, indicando uma variância não constante.

No quadrante 3, o gráfico *Scale-Location* mostra se os resíduos são homoscedáticos ou não. É possível verificar que no modelo ajustado os resíduos não estão espalhados igualmente no gráfico, ficando concentrados entre 0 e 400.000. No quadrante 4, temos o gráfico de *Residuals vs Leverage* que mostra os pontos que mais influentes na regressão. São mais influentes os pontos que estão perto da distância de Cook[5], que no caso são o 78, 35 e 39.

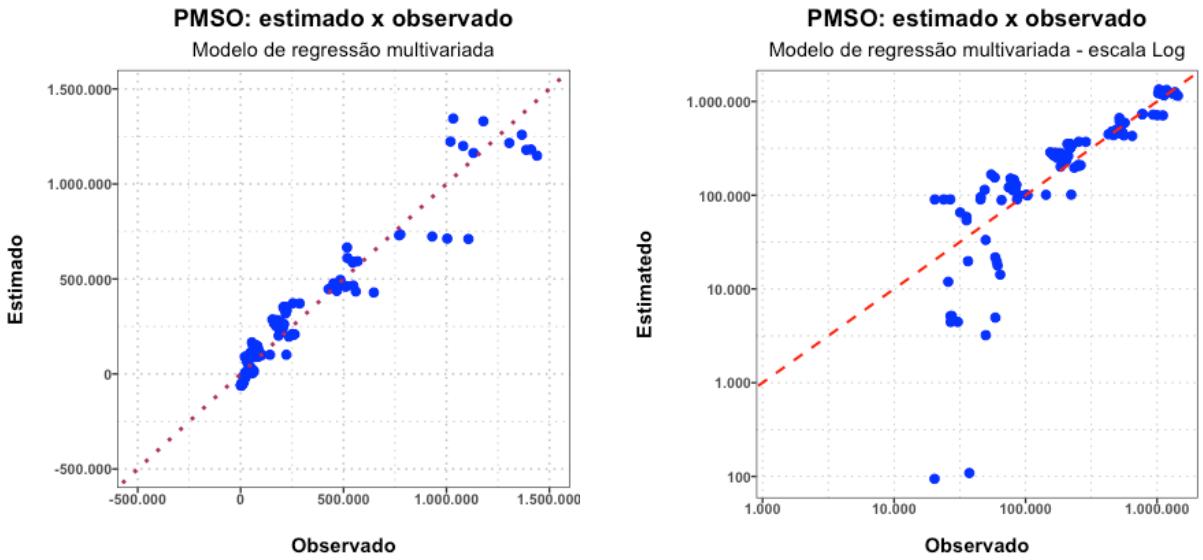
[5] McDonald (2002).

Dessa forma, conclui-se que o ajuste de um modelo de regressão linear múltipla não é adequado para representar o PMSO em função das variáveis aqui selecionadas.

Apesar das inadequações e violação dos pressupostos do modelo, quando se analisa os valores estimados versus os valores observados, obtém-se um resultado interessante, resultado do R^2 elevado.

É importante destacar as diferenças que existem entre companhias que tem maior cobertura, de companhias mais enxutas que atuam no SEB. Tais diferenças são observadas no *boxplot* apresentado anteriormente onde os valores de PMSO ao longo dos anos é apresentado para cada operadora. Dessa forma, para mitigar tais distorções, tomamos o logaritmo dos valores estimados e observados para o cálculo do R^2 preditivo bem como a construção do gráfico contendo tal ajuste:

Figura 10 – Ajustes para modelo de regressão multivariado



Fonte: Autor.

4.4.1 Ajuste de modelo de regressão linear múltiplo *leave-one-out* com validação cruzada

Uma outra forma de explorar a regressão linear múltipla é a implementação das técnicas de validação cruzadas em conjunto com a técnica *leave-one-out*. Nessa técnica, realizaremos os seguintes passos:

1. Criar um novo banco de dados com apenas as informações de uma empresa escolhida, essa será a nossa base de dados de validação

2. Criar outro banco de dados com as informações das empresas remanescentes, esse será a base de dados de treinamento

3. Implementar modelo de regressão linear múltiplo utilizando o banco de dados de treino

4. Implementar a técnica de AIC para escolher o melhor modelo e o conjunto de variáveis

5. Aplicar o modelo de regressão linear múltiplo na base de dados de validação para estimar os valores de PMSO do modelo

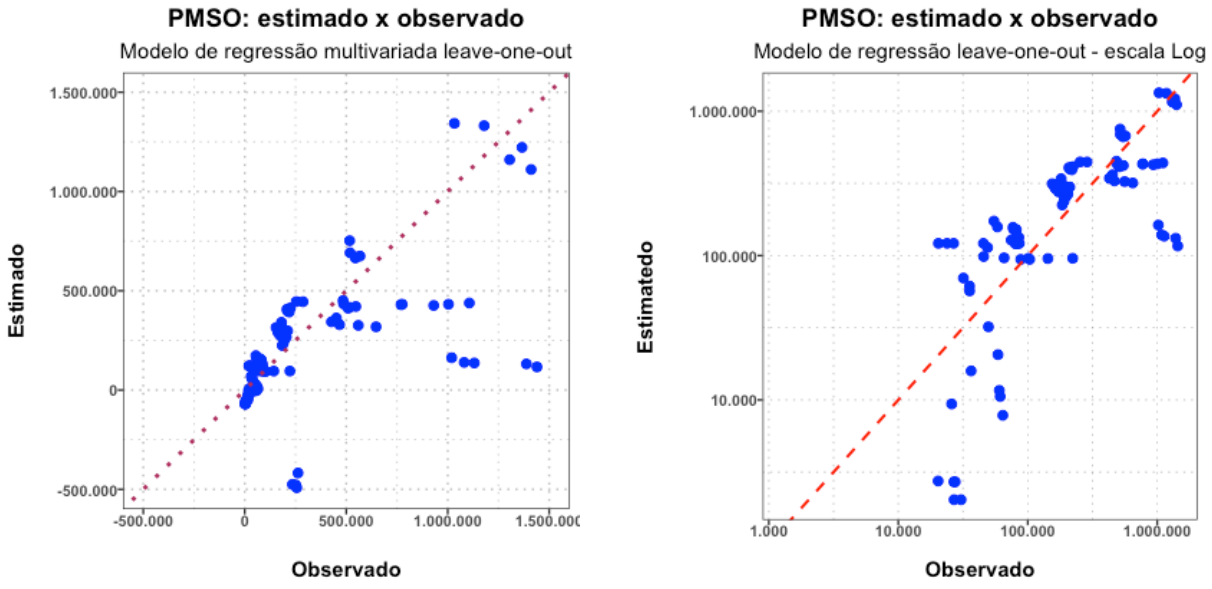
6. Armazenar os dados do PMSO estimado para a empresa do banco de dados de validação

7. Realizar esse procedimento para todas as empresas disponíveis, no caso são 28 empresas

Após a realização do procedimento acima, temos a estimação do PMSO para cada uma das empresas. A aplicação de tal procedimento produziu um R^2 preditivo de 0.32, indicando inadequação para para realizar estimações a posteriori. Repetindo o mesmo procedimento de tomar o logaritmo natural do resultado produzido, obtém-se R^2 preditivo elevado na ordem de 0.93. Tal transformação reduz as assimetrias e distorções já descritas anteriormente

Os gráficos abaixo mostram os valores observados em comparação com os estimados. Para o modelo de validação cruzada implementado para a regressão linear é possível verificar a falta de ajuste evidenciando as distorções. Para o gráfico na escala logarítmica é possível verificar que apesar de pontos dispersos, há um melhor ajuste entre o estimado e o observado.

Figura 11 – Ajustes para modelo de regressão multivariado - *leave-one-out*



Fonte: Autor.

4.4.2 Ajuste de modelo de regressão linear gama

Para resolver os problemas identificados no ajuste anterior, foi proposta a utilização do modelo gama, uma vez que a forma de "funil" foi identificada no gráfico de resíduos do modelo linear multivariado, indicando assim uma possível proporcionalidade quadrática na variância dos resíduos.

Dessa forma, foi realizado o ajuste utilizando a modelagem gama para estimar os coeficientes. Para esse cenário foram realizados dois ajustes:

1. Ajuste com implementação do critério de informação de Akaike (AIC) – Regressão Gama (1)

2. Ajuste com implementação do AIC e partição da base de dados em dois, um conjunto para treino o modelo e outro para aplicação do modelo preditivo (análogo ao realizado na seção anterior) – Regressão gama (2)

A partição foi realizada de forma a inserir e retirar as concessionárias do modelo a fim de se obter o melhor valor para seleção do modelo. É importante ressaltar que, para o ajuste gama, a análise terá foco nos coeficientes obtidos, no coeficiente de determinação preditivo e na qualidade de ajuste fornecida pela deviance.

Tabela 6 – Ajuste do modelo gama

Resultados - ajuste modelo gama		
Coeficientes	(1) Step AIC	(2) Step AIC - Partição
β_0	-1.754,9	-1.273,8
β_1	19,6	17,9
β_2	376,0	431,0
β_3	-	-
β_4	-	-
β_5	-	-
β_6	1.006,3	585,7
β_7	-	-
β_8	-25,5	-

Em uma análise preliminar verifica-se que o modelo (1) viola a restrição operacional, ou seja, o coeficiente de regressão da variável X_8 apresentou valor negativo no ajuste, o que não corresponde com a realidade operacional. Todos os coeficientes de regressão das variáveis escolhidas no ajuste (2) são positivos, o que indica que sua utilização seria possível.

Comparando os valores registrados pelos coeficientes dos modelos, nota-se semelhança na grandeza dos coeficientes β_1 e β_2. O intercepto é negativo em ambos os casos, havendo diferença absoluta significativa entre eles. É importante observar que não há

restrição operacional quanto a este último termo, uma vez ele não diz respeito à natureza da operação, mas sim a uma característica dos modelos.

Em relação ao coeficiente β_6, o modelo (1) apresenta maior valor numérico em relação ao modelo (2), quase o dobro. Pode-se levantar a hipótese que tal superioridade numérica do coeficiente no modelo (1) deve ser compensada pelo termo mais negativo apresentado em β_0.

Abaixo as medidas de ajuste dos modelos:

Tabela 7 – Medidas de ajuste

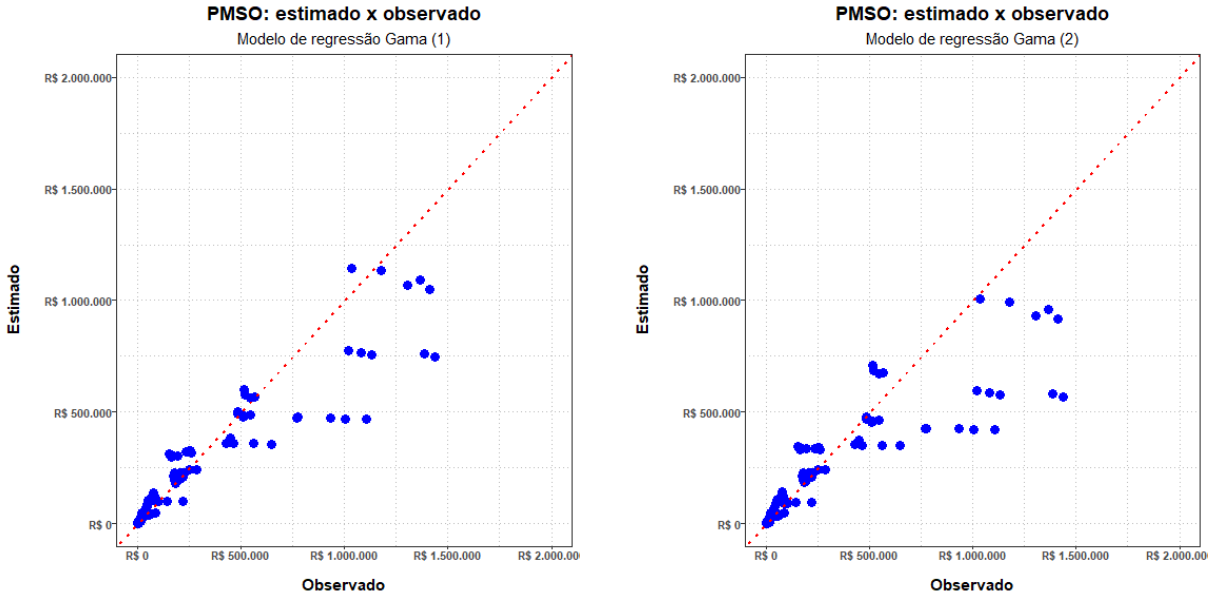
Modelo	R^2	**p-valor (Deviance)**
Regressão Gama (1)	0,82	1,00
Regressão Gama (2)	0,72	1,00

Fonte: Autor.

Na tabela acima o modelo (1) apresenta maior coeficiente de determinação e p-valor calculado a partir da deviance igual a 1, indicando subdispersão da variância. O modelo (2) apresenta menor coeficiente de determinação e deviance igual a 1, indicando também subdispersão nos dados. Como descrito anteriormente, o objetivo deste trabalho é propor uma metodologia para apurar os custos operacionais respeitando as restrições impostas pela natureza operacional das concessionárias. Dessa forma, o modelo (1) se mostra impróprio de ser utilizado devido ao coeficiente negativo e a subdispersão do modelo (2) torna questionável sua aplicação. Ainda que pudesse ser utilizado com a eliminação do fenômeno de subdispersão, a diferença entre os coeficientes R^2 levanta questões se uma outra abordagem poderia levar a um modelo com indicador mais elevado.

Os gráficos abaixo mostram a comparação entre os valores estimados e observados para os dois ajustes, é possível verificar que embora semelhantes, é possível notar algumas diferenças nos modelos, evidenciando a diferença no coeficiente de determinação.

Figura 12 – Ajustes para modelo de regressão gama



Fonte: Autor.

Analisando os resultados na escala logarítmica para novamente reduzir as distorções causadas pelas discrepâncias dos resultados operacionais e pela dimensão de cada companhia, temos os seguintes dados do coeficiente de determinação:

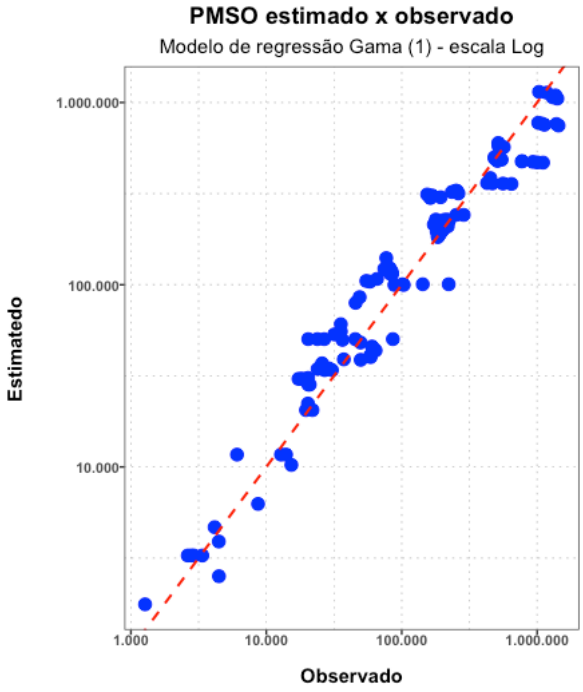
Tabela 8 – Coeficiente de determinação - Logarítmo

Comparativo de R^2	
Modelo	R^2
Regressão Gama (1) - Log	**0,94**
Regressão Gama (2) - Log	**0,92**

Fonte: Autor.

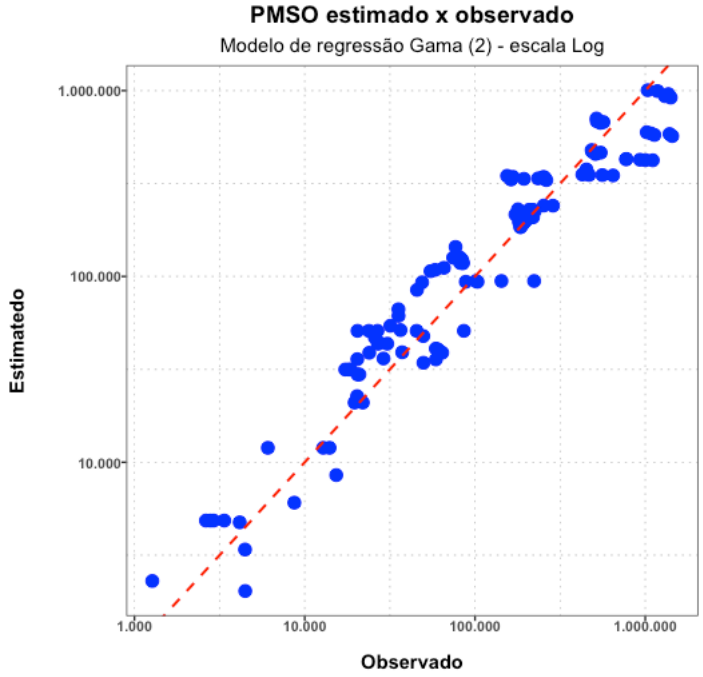
A aplicação da escala logarítmica permite nos permite novamente a observar a distribuição dos valores observados e estimados o efeito das discrepâncias. Os gráficos abaixo mostram o efeito da transformação ao longo da reta tracejada:

Figura 13 – Modelo de regressão Gama (1) - logaritmo



Fonte: Autor.

Figura 14 – Modelo de regressão Gama (2) - logaritmo



Fonte: Autor.

4.5 Modelo de programação linear

Para a estimação de um modelo onde os coeficientes regressores são estritamente positivos, é necessário realizar algumas ponderações:

1. O modelo de programação será implementado para resolver o problema de regressão quantílica, minimizando os erros absolutos e ponderados
2. O a estimação do modelo quantílico será realizado para o quantil da mediana, dessa forma o coeficiente τ assumirá o valor de 0,5
3. A principal restrição do modelo será a positividade dos coeficientes de regressão (β), o termo β_0 será representado pelo termo α.
4. Haverá uma restrição de igualdade entre os termos independentes e dependentes da equação com a presença dos erros

A formulação do modelo descrita a seguir foi uma adaptação do modelo de regressão quantílica proposto por Koenker e Bassett Jr (1978), onde os autores utilizam as técnicas de programação linear para em seus estudos de regressão quantílica. Tal modelo nos permite a flexibilidade de modelar uma regressão para um determinado quantil de uma distribuição não-paramétrica e implementar restrições para os coeficientes de regressão. A formulação do modelo é representada abaixo:

$$\min \quad \sum_{j=1}^{n} \tau_j e_{1j} \quad +(1-\tau)e_{2j}$$

$$\begin{aligned}
\text{s.a.:} \quad & y_j && = && \alpha_j + \beta_{kj} x_{kj} \quad + e_{1j} \quad + e_{2j} \\
& \beta_k && \geq && 0 \\
& e_j && = && e_{1j} + e_{2j} \qquad (4.1) \\
& \forall\ j && = && 1 \ldots N \\
& N && = && \{1 \ldots 128\} \\
& k && = && 1, \ldots, 8 \\
& \tau && = && 0.5
\end{aligned}$$

Para implementação do modelo, foram utilizadas duas estratégias para comparação. A primeira abordagem foi a utilização de todas as variáveis e de todo conjunto de dados disponível, com todas as concessionárias e todas as 128 observações. Na segunda abordagem foi utilizada a estratégia de validação cruzada do tipo *leave-one-out*, algoritmo que consiste nos seguintes passos:

1. Remover as observações para a concessionária na qual queremos estimar o PMSO
2. Resolver o modelo de programação linear utilizando o conjunto de dados com as concessionárias que restaram
3. Com o output do modelo contendo os coeficientes de interesse, estimar o PMSO para a concessionária de interesse
4. Realizar o procedimento 28 vezes, número de concessionárias presentes no estudo

A tabela abaixo mostra os resultados obtidos pela implementação das duas abordagens. É importante ressaltar que, no caso da implementação da técnica *leave-one-out*, foram gerados coeficientes de regressão para todas as 28 iterações do modelo, portanto, foi tirada a média simples dos coeficientes apenas para fim de comparação com a primeira abordagem. O coeficiente de determinação R^2 foi calculado com a compilação dos valores estimados para o PMSO.

Tabela 9 – Resultados - Modelo de programação linear

Programação Linear		
Coeficientes	Modelo completo	*leave-one-out*
β_0	-29.345,2	27.379,5
β_1	20,7	20,9
β_2	450,0	446,3
β_3	-	-
β_4	1,8	1,7
β_5	3,7	3,7
β_6	-	11,9
β_7	-	-
β_8	-	-
R^2	0,88	0,79

Fonte: Autor.

É possível observar comparando os dois modelos que o modelo completo utilizando todas as concessionárias para realizar a estimação dos parâmetros teve desempenho superior no coeficiente de determinação (11% superior). Comparando os coeficientes, $\beta_1, \beta_2, \beta_4$ e β_5 apresentaram valores próximos.

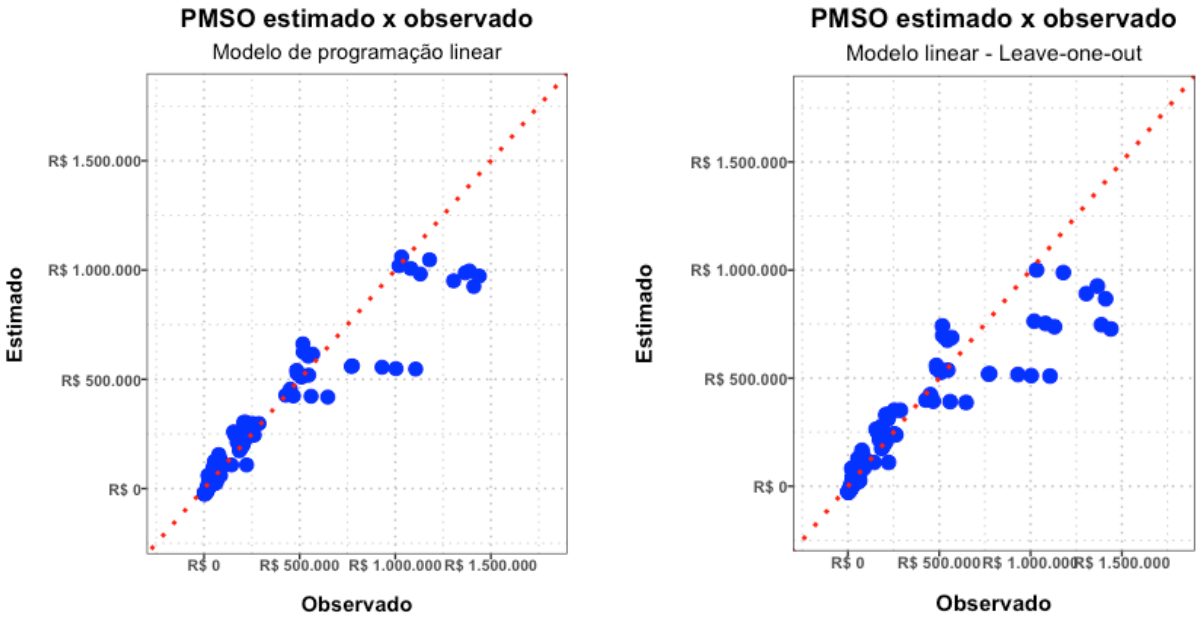
A grande diferença entre os modelos ocorre em relação ao intercepto, chamado de α na formulação e na tabela representada como β_0, onde no modelo completo o coeficiente apresenta valor negativo e na técnica *leave-one-out* apresenta valor positivo. Comparativamente, como se um fosse o oposto do outro. É preciso destacar que para o intercepto não foi aplicada nenhuma restrição quanto a sua positividade. Dessa forma, os modelos balanceram a influencia das variáveis utilizadas com o intercepto.

Como os coeficientes do modelo *leave-one-out* são oriundos de uma média simples das 28 iterações do algoritmo, levanta-se a hipótese de que para as concessionárias com

PMSO extremante elevados, fosse necessário um intercepto mais negativo e o contrário ocorreria para concessionárias com PMSO mais baixos.

É importante destacar que apenas o modelo utilizando a abordagem *leave-one-out* registrou valores para o coeficiente relativo à variável X_6. Tal fato demonstra que há evidências que as variáveis X_1, X_2, X_4 e X_5 podem ser as mais relevantes quando se quer estimar um modelo com as restrições descritas anteriormente.

Figura 15 – Soluções para o modelo de programação linear



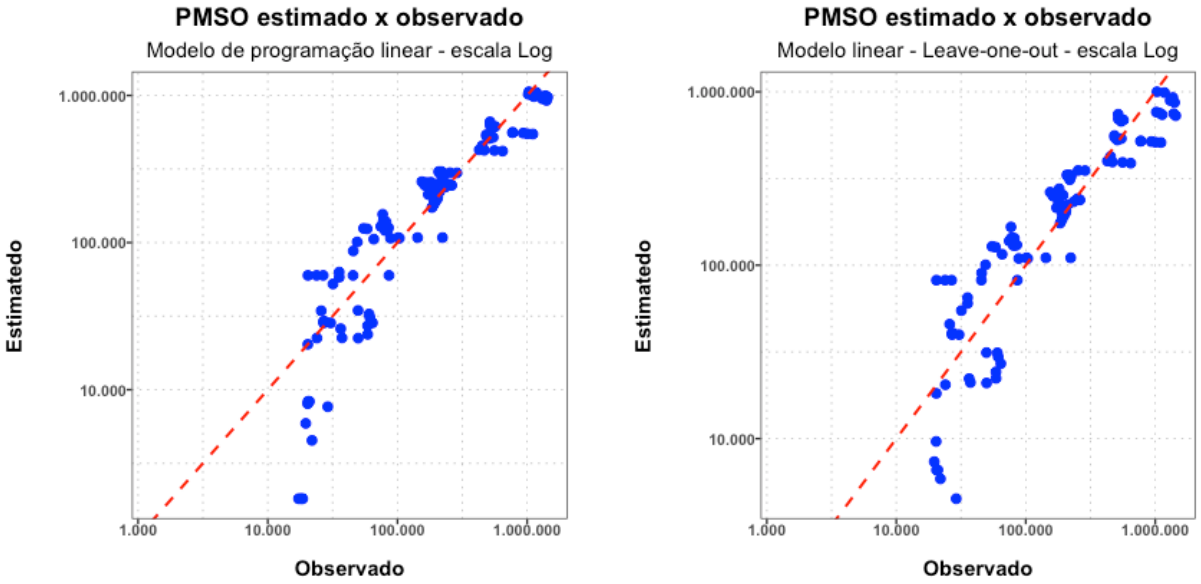
Fonte: Autor.

Os gráficos acima demonstram os valores observados versus estimados para os dois modelos. É possível verificar as diferenças verificando a distância entre os pontos e a reta traçada em vermelho.

Assim como nas demais aplicações anteriores, para visualizar o efeito do modelo implementado sem a influência das distorções presentes no banco de dados, foi aplicado a transformação logarítmica. Dessa forma, obteve-se o coeficiente de determinação preditivo para o modelo linear na ordem de 0.8019 e para a técnica de validação-cruzada com o *leave-one-out* de 0.80.

Abaixo temos a representação dos resultados obtidos nos gráficos:

Figura 16 – Soluções para o modelo de programação linear - logaritmo



Fonte: Autor.

4.6 Modelo de programação linear com aplicação de *bootstrap*

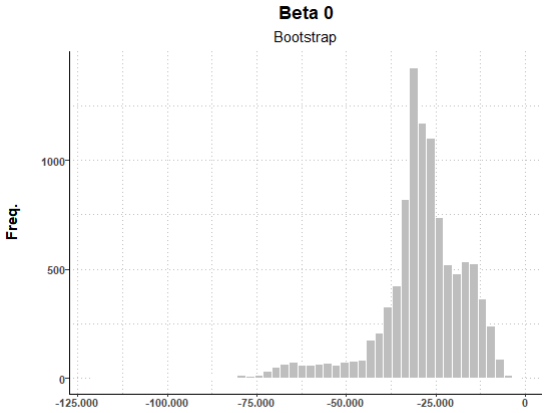
Para verificar se solução do modelo de programação linear completo de fato respeitasse as restrições operacionais, foi realizada uma simulação utilizando a técnica de *bootstrap* e posteriormente construir um intervalo de confiança percentílico. O modelo completo foi escolhido por apresentar maior coeficiente de determinação.

Nesse procedimento foram realizadas 10.000 simulações com observações aleatoriamente no *software* R. Abaixo os passos do procedimento:

1. É gerada uma nova base de dados utilizando a técnica de reamostragem com reposição
2. É implementado o algoritmo SIMPLEX para resolver o modelo com a nova base de dados oriunda da reamostragem
3. A solução contendo os coeficientes de regressão é computada
4. Os passos de 1 a 3 são realizados 10.000 vezes

Abaixo é possível verificar o histograma gerado por cada uma das soluções das simulações realizadas pelo procedimento *bootstrap*:

Figura 17 – Histograma de soluções para β_0



Fonte: Autor.

O histograma gerado para o coeficiente β_0 a partir do procedimento *bootstrap*, reforça a hipótese inicial de que o modelo linear compensa as distâncias dos valores do PMSO entre as concessionárias com o intercepto já que não houve nenhum valor positivo gerado a partir das reamostragens. É importante ressaltar que devido ao Teorema central do limite[6] era esperado um comportamento aproximadamente normal devido ao elevado valor de reamostragens (10.000) geradas, para β_0 o histograma apresentou uma característica mais assimétrica, com concentração de dados a direita do gráfico.

Para os coeficientes β_1 e β_2, obtivemos o resultado esperado para o comportamento do histograma, ou seja, uma aproximação de uma distribuição normal. Vale ressaltar que para β_1 obtivemos um pouco mais de 100 observções com valor nulo, indicando que há uma pequena possibilidade de que a variável X_1 não tenha relevância na estimação dos custos operacionais.

Os resultados dos histogramas dos coeficientes β_3, β_6, β_7 , β_8 foram surpreendentes. É possível observar nos gráficos abaixo que exceto para β_6, em todas as 10.000 simulações, as variáveis X_3, X_7,X_8 se mostraram irrelevantes para a estimação do PMSO.

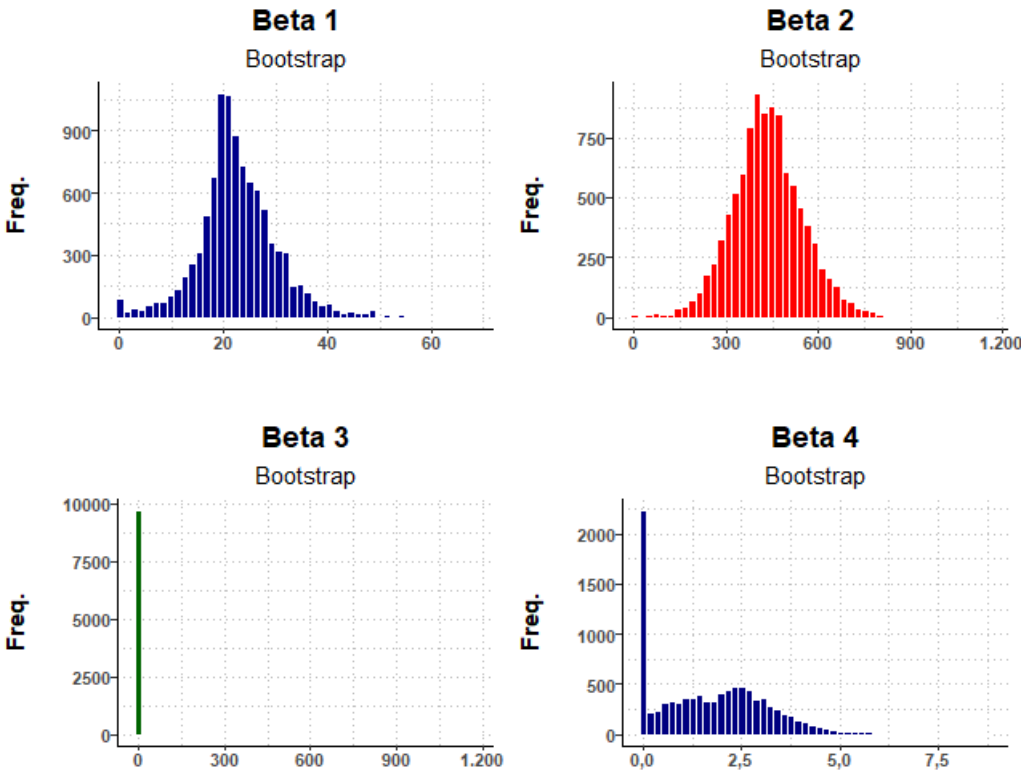
A variável X_6 apresenta baixa probabilidade (menor que 0.2) de ter relevância para a estimação dos custos, entretanto não recomenda-se não desprezar sua importância. De posse dessas evidências, é possível afirmar que as demais variáveis não tem efeito positivo no incremento de custo das concessionárias.

Por fim, para β_4 obteve-se um histograma curioso, onde em algumas situações a variável X_4 não é relevante (coeficiente nulo).

Abaixo, temos as figuras com os resultados dos histogramas gerados para os demais coeficientes da regressão:

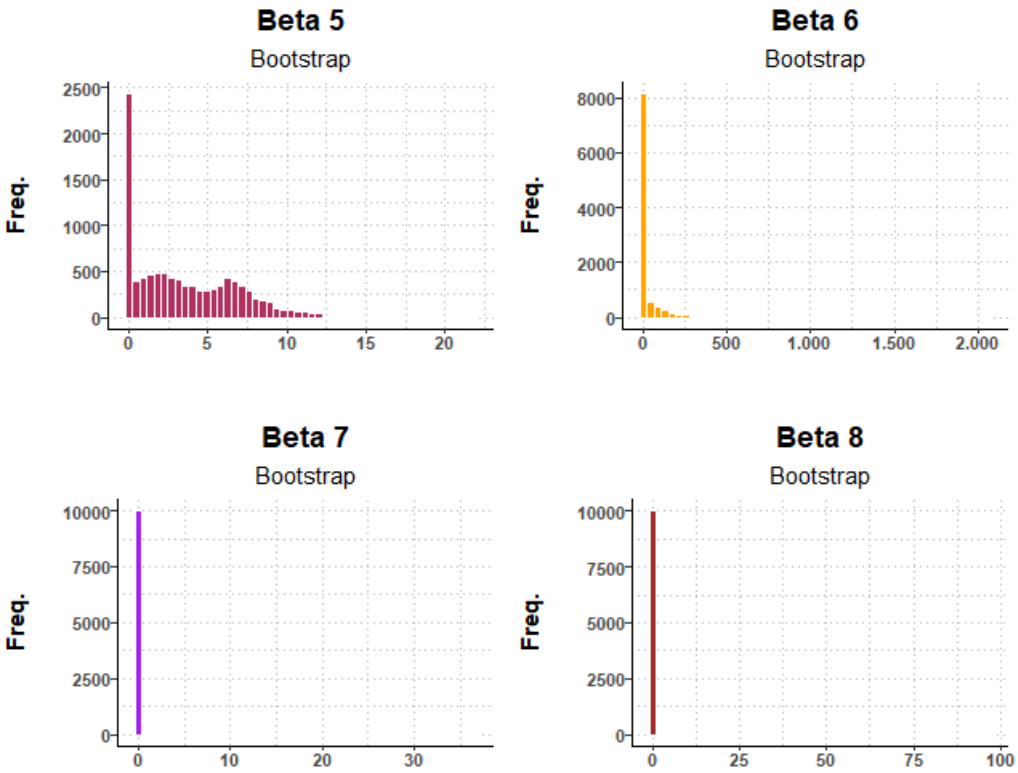
[6] Kwak et al. (2017).

Figura 18 – Histograma de soluções para $\beta_1, \beta_2, \beta_3, \beta_4$



Fonte: Autor.

Figura 19 – Histograma de soluções para $\beta_5, \beta_6, \beta_7, \beta_8$



Fonte: Autor.

O nível de confiança escolhido para construir o intervalo foi de 95%. Dessa forma, obteve-se os seguintes intervalos com seus limites definidos:

Tabela 10 – Intervalo de Confiança Percentílico - *bootstrap*

Intervalo de Confiança Percentílico *bootstrap*		
Coeficientes	2,5%	97,5%
β_0	-64.426,7	-10.021,2
β_1	6,3	40,0
β_2	206,3	672,2
β_3	0,0	51,3
β_4	0,0	4,6
β_5	0,0	11,1
β_6	0,0	281,0
β_7	0,0	0,0
β_8	0,0	0,0

Fonte: Autor.

Analisando os limites do intervalo de confiança percentílico, pode-se notar que para o modelo construído, todos os coeficientes permaneceram dentro dos intervalos. Dessa

forma, é possível afirmar que a um nível de significância de 95%, as soluções encontradas para modelo de programação linear completo são representativas.

5 Conclusão

5.1 Contribuições

O presente trabalho buscou explorar as propriedades e características dos métodos de regressão e programação linear para analisar a correlação entre os custos operacionais (PMSO) das concessionárias de energia elétrica atuantes no SEB e seus principais direcionadores. Dada a importância de tal estudo no contexto da RAP, foi importante destacar a utilização de técnicas as quais respeitavam as restrições operacionais nos quais os operadores estão envolvidos.

Uma grande contribuição do trabalho foi a implantação do modelo de programação linear emulando uma regressão com a aplicação de restrições técnicas. Essa abordagem inovadora integra perfeitamente o poder da programação linear para lidar com complexas restrições inerentes a problemas específicos. Ao incorporar essas restrições, o modelo tem se mostrado notavelmente versátil, acomodando uma ampla gama de cenários do mundo real. Vale ressaltar que essa nova abordagem demonstrou sua eficácia na entrega de resultados robustos e precisos, destacando seu potencial para otimização e melhorias na eficiência em várias áreas.

Além disso, a utilização desse modelo de regressão linear baseado em programação linear mostrou sua capacidade de aprimorar os processos de tomada de decisão. Ao otimizar simultaneamente as variáveis enquanto cumpre as restrições impostas, ele pode potencialmente simplificar a alocação de recursos, minimizar custos e maximizar a eficiência. Essa técnica tem amplas implicações em diversas indústrias, desde a gestão da cadeia de suprimentos até as finanças, onde a previsão precisa de resultados e a capacidade de adaptação a restrições dinâmicas são fundamentais. No geral, esse modelo inovador oferece uma ferramenta poderosa e flexível para abordar problemas complexos com precisão, fornecendo *insights* acionáveis e proporcionando melhorias mensuráveis na tomada de decisões e alocação de recursos.

5.2 Resultados finais

Os direcionadores aqui descritos, sejam eles ativos físicos ou demais equipamentos, apenas agregam custos ao resultado financeiro das concessionárias. Dessa forma seria contraintuitivo ou mesmo ilógico utilizar de modelos para estimação de custo que não obedecessem a natureza financeira e operacional de todo o sistema.

Para comparação dos resultados, foram considerados os modelos implementados com as técnicas de validação cruzada e *leave-one-out* na escala logarítmica para atenua-

ção das distorções presentes no banco de dados aqui descritas. Dessa forma, no estudo, foi possível explorar os resultados dos ajustes de modelos de regressão linear e não-linear (modelo gama) e de programação linear. Por meio do *software* R, os modelos foram implementados e seus resultados colhidos. Foi possível observar a inadequação do modelo de regressão linear multivariado, devido ao descumprimento dos pressupostos do modelo e o não cumprimento da restrição operacional.

Os ajustes do modelo gama tiveram resultado razoável, sendo que a primeira abordagem violou a restrição operacional. A segunda abordagem mostrou certa adequação, mas com o problema da subdispersão, o modelo poderia não captar toda a variabilidade esperada dos dados.

Por fim, o modelo de programação linear teve resultado satisfatório, uma vez que seu projeto foi concebido desde o início para que a restrição operacional fosse respeitada. Diante de tal fato, era preciso verificar a confiabilidade dos resultados do modelo, as simulações realizadas utilizando a técnica *bootstrap*, mostraram que o modelo linear completo, com 95% de confiança, possui resultados robustos, relevantes e com boa capacidade de predição, conforme mostra tabela abaixo, com o comparativo do coeficiente de determinação R^2 de cada modelo abordado.

Tabela 11 – Resultados Finais

Comparativo de R^2	
Modelo	R^2
Regressão linear - *leave-one-out* - Log	**0,93**
Regressão Gama (2) - Log	**0,92**
Modelo Linear - *leave-one-out* - Log	**0,80**

Fonte: Autor.

Verifica-se que a capacidade preditiva do modelo linear completo é semelhante à obtida pelo ajuste produzido pela regressão multivariada, com uma diferença próxima de 5%. É também o modelo com maior capacidade dos demais, evidenciando que a formulação implementada se mostrou adequada. Vale ressaltar que cinco entre as oito variáveis mostraram significância quando aplicado o modelo linear, esse fato implica na redundância dos direcionadores de custo e um eventual comprometimento nas estimativas de eficiência apresentados pela reguladora, quando considerada a metodologia atual de cálculo.

Após os fatos aqui apresentados, conclui-se que o modelo de programação linear é efetivo para realizar estimativas do custo operacional (PMSO) no âmbito da RAP considerando as restrições operacionais observadas para as concessionárias de energia elétrica participantes do SEB. Também é possível concluir que o atual modelo utilizado pela reguladora apresenta redundâncias em sua metodologia, uma vez que foi detectado apenas a significância de quatro entre oito direcionadores de custo.

REFERÊNCIAS

BAIO, Fábio HR et al. Modelo de programação linear para seleção de pulverizadores agrícolas de barras. **Engenharia Agrícola**, SciELO Brasil, v. 24, p. 355–363, 2004.

BEATTY, WS et al. Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. **Comparison of Spring Migration Ecology of American Black Ducks (Anas rubripes) and Mallards (Anas platyrhynchos) in the Montezuma Wetlands Complex**, State University of New York, v. 147, p. 60, 2018.

BERKELAAR, Michel et al. **lpSolve: Interface to 'Lp solve' v.5.5 to Solve Linear-Integer Programs**. 2023. R package version 5.6.18. Disponível em: https://CRAN.R-project.org/package=lpSolve.

BRASIL. **Lei Nº 9.427, de 26 de dezembro de 1996. Institui a Agência Nacional de Energia Elétrica - ANEEL, disciplina o regime das concessões de serviços públicos de energia elétrica e dá outras providências**. 1996. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 27 dez. 1999. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9427cons.htm. Acesso em: 20 ago. 2023.

BRASIL. **Nota Técnica Nº 037/2017– SRM/ANEEL**. 2017. Ministério de Minas e Energia. Agência Nacional de Energia Elétrica. Superintendência de Regulação Econômica e Estudos do Mercado.

BRASIL. **Nota Técnica Nº 097/2018 – SRM/ANEEL**. 2018. Ministério de Minas e Energia, Agência Nacional de Energia Elétrica, Superintendência de Regulação Econômica e Estudos do Mercado.

BRASIL. **Nota Técnica Nº 097/2022– SRM/ANEEL**. 2022. Ministério de Minas e Energia. Agência Nacional de Energia Elétrica. Superintendência de Regulação Econômica e Estudos do Mercado.

BRASIL. **Nota Técnica Nº 160/2017– SRM/ANEEL**. 2017. Ministério de Minas e Energia. Agência Nacional de Energia Elétrica. Superintendência de Regulação Econômica e Estudos do Mercado.

BRASIL. **Nota Técnica Nº 204/2007 – SRM/ANEEL**. 2007. Ministério de Minas e Energia. Agência Nacional de Energia Elétrica. Superintendência de Regulação Econômica e Estudos do Mercado.

BRASIL. **Nota Técnica Nº 204/2018 – SRM/ANEEL**. 2018. Ministério de Minas e Energia, Agência Nacional de Energia Elétrica, Superintendência de Regulação Econômica e Estudos do Mercado.

BRASIL. **Tomada de Subsídios – TS nº 14/2022**. 2022. Diário Oficial [da] República Federativa do Brasil, Poder Executivo, Brasília, DF, 02 ago. 2022. Seção 3, p. 106.

BREUSCH, Trevor S et al. A simple test for heteroscedasticity and random coefficient variation. **Econometrica: Journal of the econometric society**, JSTOR, p. 1287–1294, 1979.

COATES, Michael. **exploreR: Tools for Quickly Exploring Data**. 2016. R package version 0.1. Disponível em: https://CRAN.R-project.org/package=exploreR.

CORDEIRO, Gauss Moutinho et al. Modelos lineares generalizados e extensões. **Piracicaba: USP**, p. 31, 2008.

COSTA, Marcelo Azevedo. Tópicos em ciência dos dados: introdução dos modelos paramétricos e suas aplicações utilizando o R. Universidade Federal de Minas Gerais, 2019.

COSTA, Marcelo Azevedo et al. Stochastic data envelopment analysis applied to the 2015 Brazilian energy distribution benchmarking model. **Decision Analytics Journal**, Elsevier, v. 3, p. 100061, 2022.

CUNHA, Wellington José da et al. Intervalos de confiança bootstrap para modelos de regressão com erros de medida. **Rev. Mat. Estat**, v. 21, n. 2, p. 25–41, 2003.

DANTZIG, George B. Linear programming. **Operations research**, INFORMS, v. 50, n. 1, p. 42–47, 2002.

DANTZIG, George B. Maximization of a linear function of variables subject to linear inequality. 1947. Published in Koopmans TC (ed.): Activity analysis of production and allocation. Wiley & Chapman-Hall, New York-London, 1951.

DAVISON, A. C. et al. **Bootstrap Methods and Their Applications**. Cambridge: Cambridge University Press, 1997. ISBN 0-521-57391-2. Disponível em: http://statwww.epfl.ch/davison/BMA/.

EFRON, B. Bootstrap methods: another look at the jackknife annals of statistics 7: 1–26. **View Article PubMed/NCBI Google Scholar**, v. 24, 1979.

FEARNLEY, John et al. The complexity of the simplex method. In: PROCEEDINGS of the forty-seventh annual ACM symposium on Theory of computing. 2015. P. 201–208.

FEITOSA NETO, Sandoval de Araújo. Um estudo da metodologia de desconto da Receita Anual Permitida (RAP) em função dos atrasos na entrada em operação de empreendimentos de transmissão, 2009.

FIRKE, Sam. **janitor: Simple Tools for Examining and Cleaning Dirty Data**. 2023. R package version 2.2.0. Disponível em: https://CRAN.R-project.org/package=janitor.

FRANCO, Gustavo et al. Intervalos de Confiança Bootstrap para o Parâmetro D em Modelos de Integração Fracionária. v. 37, p. 735–744, 2005.

FROSSARD, Afonso Celso Pagano. Programação linear: maximização de lucro e minimização de custos. **Revista Científica da Faculdade Lourenço Filho**, v. 6, n. 1, p. 1–30, 2009.

GODOY, Arilda Schmidt. Pesquisa qualitativa: tipos fundamentais. **Revista de Administração de empresas**, SciELO Brasil, v. 35, p. 20–29, 1995.

GONZALEZ-ESTRADA, Elizabeth et al. **mvShapiroTest: Generalized Shapiro-Wilk test for multivariate normality**. 2013. R package version 1.0. Disponível em: https://CRAN.R-project.org/package=mvShapiroTest.

GROVER, P. et al. Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature. **Global Journal of Flexible Systems Management**, v. 18, n. 3, p. 203–229, 2017.

HOFFMANN, Rodolfo. Análise de regressão: uma introdução à econometria, 2016.

HUNG, Rupert K et al. Prognostic value of exercise capacity in patients with coronary artery disease: the FIT (Henry Ford ExercIse Testing) project. In: ELSEVIER, 12. MAYO Clinic Proceedings. 2014. v. 89, p. 1644–1654.

HYNDMAN, Rob J et al. Automatic time series forecasting: the forecast package for R. **Journal of Statistical Software**, v. 26, n. 3, p. 1–22, 2008. DOI: 10.18637/jss.v027.i03.

KENNEDY, Peter E. Sinning in the basement: What are the rules? The ten commandments of applied econometrics. **Journal of Economic Surveys**, Wiley Online Library, v. 16, n. 4, p. 569–589, 2002.

KOENKER, Roger. **Quantile regression**. Cambridge university press, 2005. v. 38.

KOENKER, Roger; BASSETT JR, Gilbert. Regression quantiles. **Econometrica: journal of the Econometric Society**, JSTOR, p. 33–50, 1978.

KWAK, Sang Gyu et al. Central limit theorem: the cornerstone of modern statistics. **Korean journal of anesthesiology**, The Korean Society of Anesthesiologists, v. 70, n. 2, p. 144–156, 2017.

LEWIS, Catherine. Linear programming: theory and applications. **Whitman College Mathematics Department**, 2008.

LEWIS-BECK, Colin et al. **Applied regression: An introduction**. Sage publications, 2015. v. 22.

LIMA PINTO, Leizer de et al. Implementação de algoritmos simplex e pontos interiores para programação linear. **Revista EVS-Revista de Ciências Ambientais e Saúde**, v. 35, n. 2, p. 225–246, 2008.

LOPES, Ana Lúcia Miranda et al. Critical evalutaion of the efficient costs assessment model used in the regulation of Brazilian energy distribution service operator Crítica do modelo de cálculo do custo eficiente das empresas brasileiras de distribuição de energia elétrica Evaluación crítica del modelo de evaluación de costos eficiente utilizado en la. **Revista Gestão e Tecnologia**, Universidade Federal de Minas Gerais, v. 16, n. 3, p. 5–30, 2016.

MCCULLAGH, Peter et al. Monographs on statistics and applied probability. **Generalized linear models**, Chapman & Hall, v. 37, 1989.

MCDONALD, Barry. A teaching note on Cook's distance-a guideline. Massey University, 2002.

MILONE, Giuseppe et al. **Estatística aplicada**. Atlas, 1995. v. 1.

MONTGOMERY, Douglas C et al. Estatística aplicada e probabilidade para engenheiros, 2ª. **Ed. Rio de Janeiro: Editora LTC**, p. 416, 2003.

NELDER, John Ashworth et al. Generalized linear models. **Journal of the Royal Statistical Society Series A: Statistics in Society**, Oxford University Press, v. 135, n. 3, p. 370–384, 1972.

NEUWIRTH, Erich. **RColorBrewer: ColorBrewer Palettes**. 2022. R package version 1.1-3. Disponível em: https://CRAN.R-project.org/package=RColorBrewer.

NEYMAN, Jerzy. Outline of a theory of statistical estimation based on the classical theory of probability. **Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences**, The Royal Society London, v. 236, n. 767, p. 333–380, 1937.

PELLI NETO, A. Curso de Engenharia de Avaliação Imobiliária–fundamentos e aplicação da estatística inferencial. **Belo Horizonte, MG:[sn]**, 2003.

PESSANHA, José Francisco Moreira et al. Avaliação dos custos operacionais eficientes das empresas de transmissão do setor elétrico Brasileiro: uma proposta de adaptação do modelo DEA adotado pela ANEEL. **Pesquisa Operacional**, SciELO Brasil, v. 30, p. 521–545, 2010.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: https://www.R-project.org/.

RAZALI, Nornadiah Mohd et al. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. **Journal of statistical modeling and analytics**, v. 2, n. 1, p. 21–33, 2011.

RUDIS, Bob et al. **ggalt: Extra Coordinate Systems, 'Geoms', Statistical Transformations, Scales and Fonts for 'ggplot2'**. 2017. R package version 0.4.0. Disponível em: https://CRAN.R-project.org/package=ggalt.

SANTOS, Bruno Ramos dos. **Modelos de regressão quantílica**. 2012. Tese (Doutorado) – Universidade de São Paulo.

SCHAUBERGER, Philipp et al. **openxlsx: Read, Write and Edit xlsx Files**. 2023. R package version 4.2.5.2. Disponível em: https://CRAN.R-project.org/package=openxlsx.

SHARPE, Norean R. **Estatística aplicada: administração, economia e negócios**. Grupo A-Bookman, 2000.

SILVA, Dirceu da; LOPES, Evandro Luiz. Pesquisa quantitativa: elementos, paradigmas e definições. **Revista de Gestão e Secretariado (Management and Administrative Professional Review)**, v. 5, n. 1, p. 01–18, 2014.

SILVA, Dirceu da; SIMON, Fernanda Oliveira. Abordagem quantitativa de análise de dados de pesquisa: construção e validação de escala de atitude. **Cadernos Ceru**, v. 16, p. 11–27, 2005.

SOBRAL, Thales Lima et al. Utilização dos critérios de informação na seleção de modelos de regressão linear. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 4, n. 1, 2016.

SOUSA, Keliny Martins de Melo. **Modelos lineares generalizados e modelos de dispersão aplicados à modelagem de sinistros agrícolas**. 2010. Tese (Doutorado) – Universidade de São Paulo.

STANTON, Jeffrey M. Galton, Pearson, and the peas: A brief history of linear regression for statistics instructors. **Journal of Statistics Education**, Taylor & Francis, v. 9, n. 3, 2001.

UPTON, Graham et al. **A dictionary of statistics 3e**. Oxford quick reference, 2014.

VENABLES, W. N. et al. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: https://www.stats.ox.ac.uk/pub/MASS4/.

WHITE, Kenneth J. The Durbin-Watson test for autocorrelation in nonlinear models. **The Review of Economics and Statistics**, JSTOR, p. 370–373, 1992.

WICKHAM, Hadley. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. Disponível em: https://ggplot2.tidyverse.org.

WICKHAM, Hadley; AVERICK, Mara et al. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019. DOI: 10.21105/joss.01686.

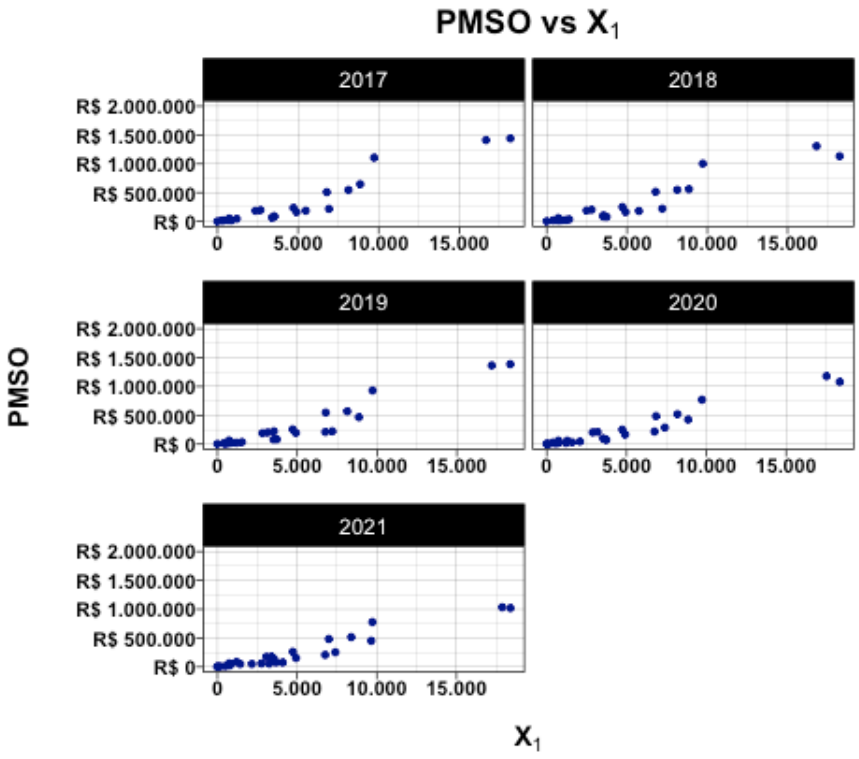
WICKHAM, Hadley; FRANÇOIS, Romain et al. **dplyr: A Grammar of Data Manipulation**. 2023. R package version 1.1.2. Disponível em: https://CRAN.R-project.org/package=dplyr.

WICKHAM, Hadley; SEIDEL, Dana. **scales: Scale Functions for Visualization**. 2022. R package version 1.2.1. Disponível em: https://CRAN.R-project.org/package=scales.

Apêndices

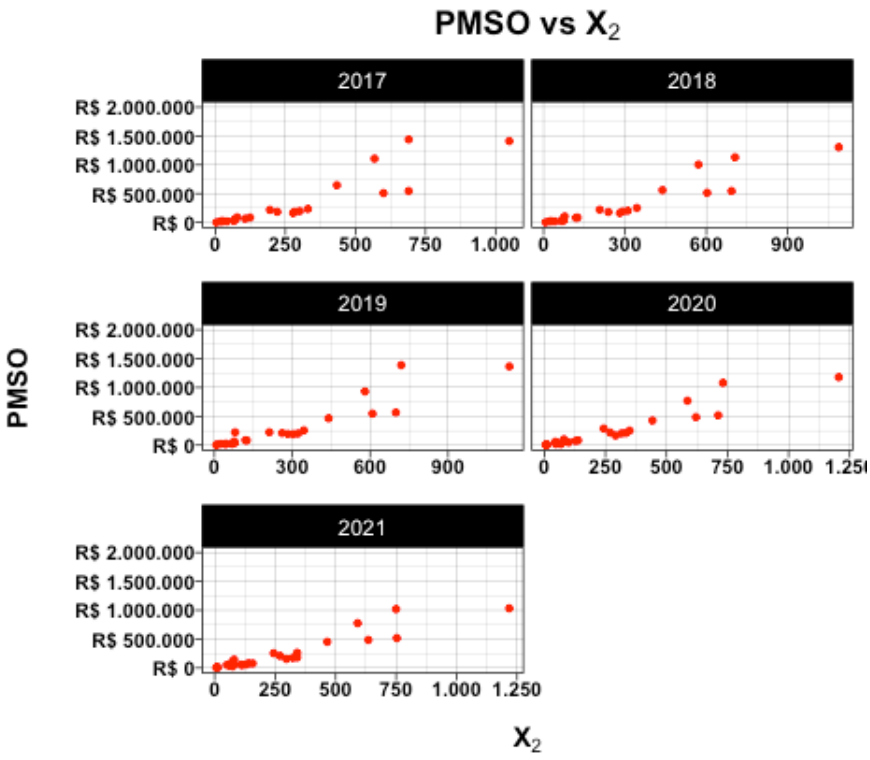
APÊNDICE A – Dispersão PMSO *versus* variáveis por ano

Figura 20 – Dispersão ano - X_1



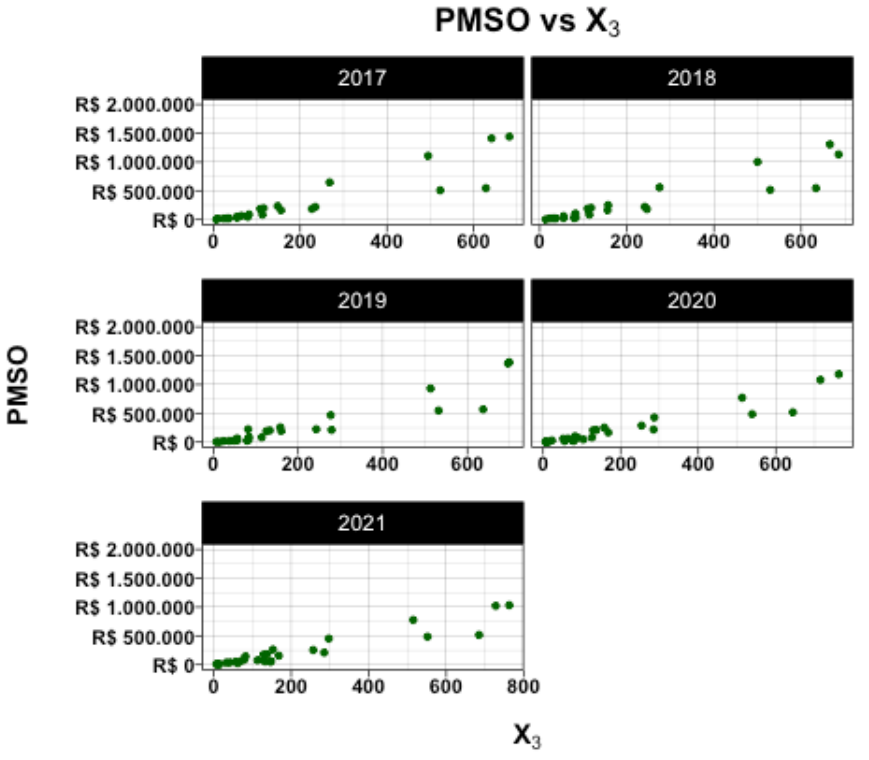
Fonte: Autor.

Figura 21 – Dispersão ano - X_2



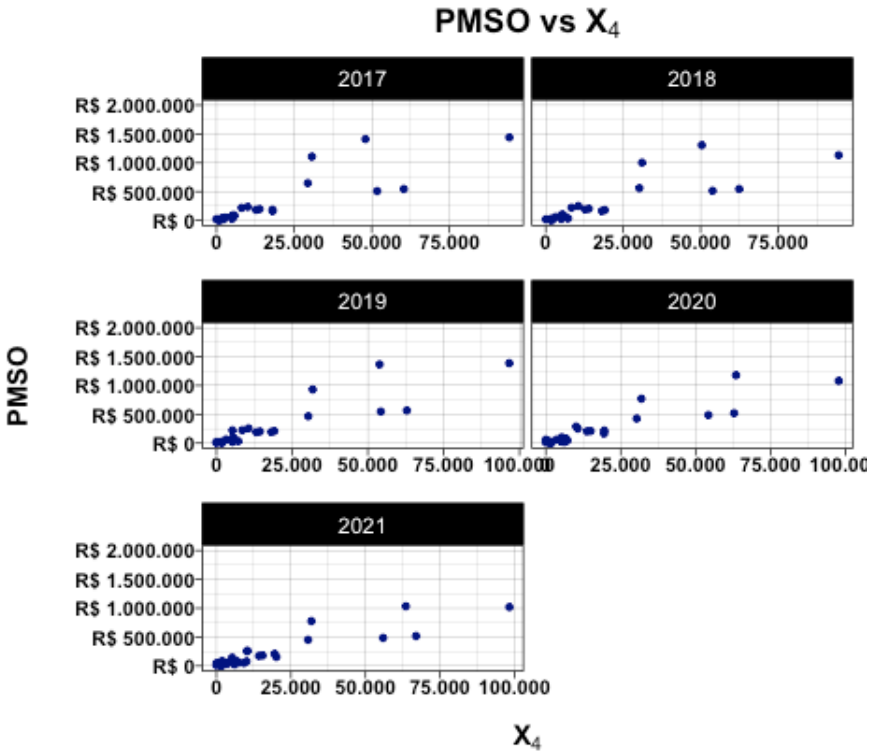
Fonte: Autor.

Figura 22 – Dispersão ano - X_3



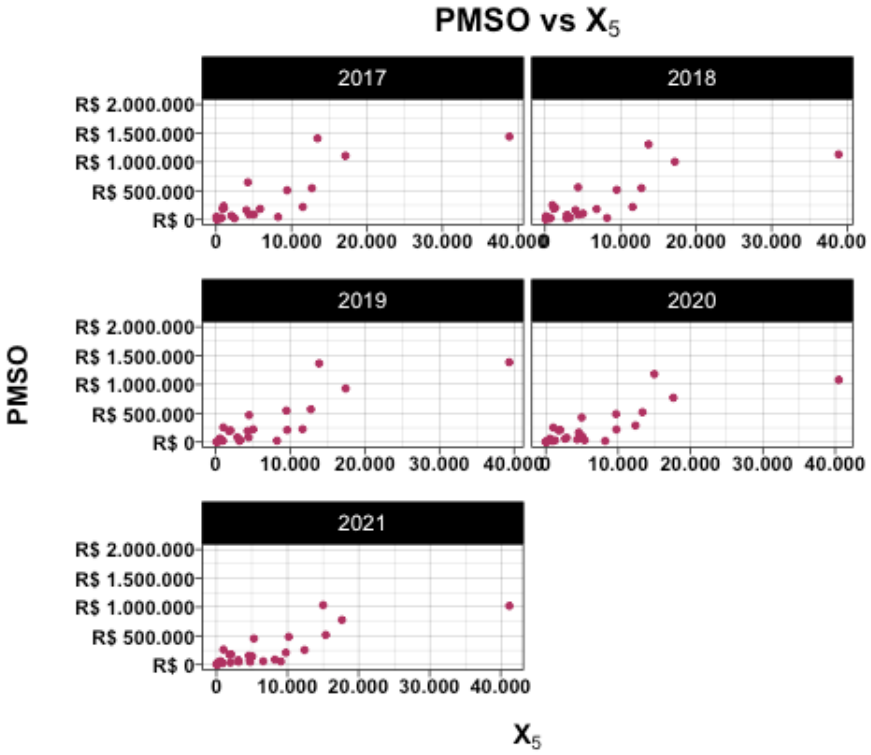
Fonte: Autor.

Figura 23 – Dispersão ano - X_4



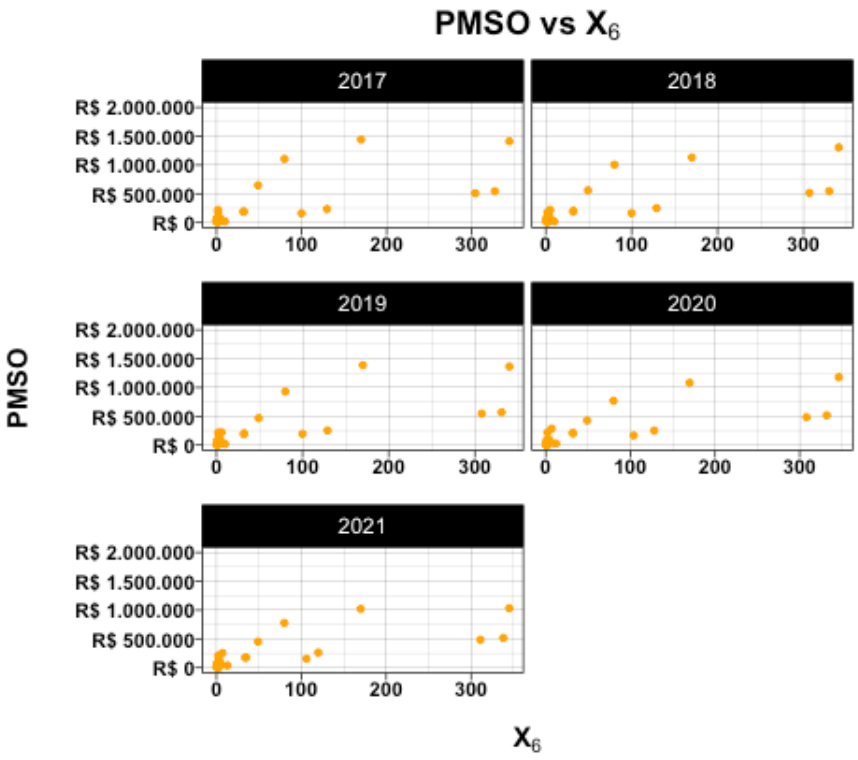
Fonte: Autor.

Figura 24 – Dispersão ano - X_5



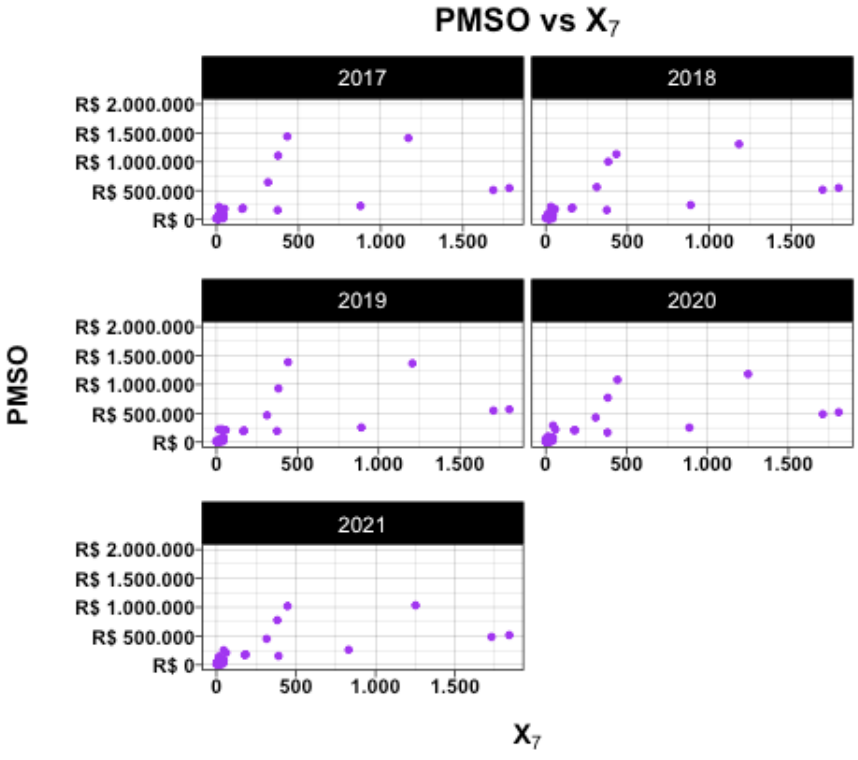
Fonte: Autor.

Figura 25 – Dispersão ano - X_6



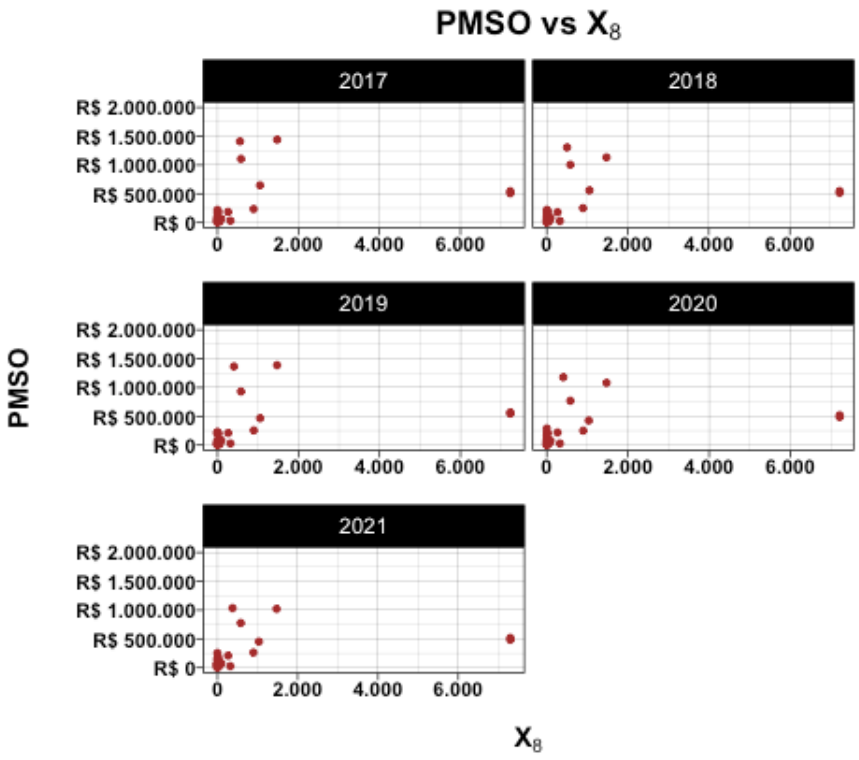
Fonte: Autor.

Figura 26 – Dispersão ano - X_7



Fonte: Autor.

Figura 27 – Dispersão ano - X_8



Fonte: Autor.