UNIVERSIDADE FEDERAL DE MINAS GERAIS
Curso de Especialização em Estatística

ANÁLISE DOS CUSTOS OPERACIONAIS DE CONCESSIONÁRIAS BRASILEIRAS DE ENERGIA ELÉTRICA UTILIZANDO O MÉTODO DE REGRESSÃO MÚLTIPLA

Belo Horizonte
2014

Michele Renata Barbosa da Silva

ANÁLISE DOS CUSTOS OPERACIONAIS DE CONCESSIONÁRIAS BRASILEIRAS DE ENERGIA ELÉTRICA UTILIZANDO O MÉTODO DE REGRESSÃO MÚLTIPLA

Monografia apresentada ao curso de Especialização em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Estatística.

Orientador: Marcelo Azevedo Costa

Belo Horizonte

2014

Michele Renata Barbosa da Silva

ANÁLISE DOS CUSTOS OPERACIONAIS DE CONCESSIONÁRIAS BRASILEIRAS DE ENERGIA ELÉTRICA UTILIZANDO O MÉTODO DE REGRESSÃO MÚLTIPLA

Monografia apresentada ao curso de Especialização em Estatística do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Estatística.

Prof. Marcelo Azevedo Costa (Orientador) – UFMG

Prof.ª Edna Afonso Reis – UFMG

Prof.ª Ilka Afonso Reis – UFMG

Belo Horizonte, 03 de abril de 2014.

Dedico este trabalho aos meus pais, Antonio e Maria, pelo amor, paciência, compreensão e pelos bons valores ensinados, os quais fizeram a diferença na construção da base da minha formação pessoal e profissional.

AGRADECIMENTOS

A Deus, por seu amor incondicional e fidelidade no cumprimento de mais uma promessa na minha vida, concedendo capacidade, sabedoria e cuidando de cada detalhe para a concretização deste sonho.

Aos meus pais e irmão, pelo amor, dedicação, compreensão e pelo exemplo de caráter e integridade.

Ao meu futuro esposo, Davidson Roberto dos Santos Ferreira, pelo incentivo, apoio e compreensão sempre.

Ao meu orientador e professor Marcelo Azevedo Costa, pelos conhecimentos e excelente orientação prestada no desenvolvimento desta pesquisa.

E aos meus queridos amigos pelo companheirismo, que perto ou distantes, estiveram sempre torcendo e me ajudando de alguma forma a alcançar meus objetivos.

Enfim, sou grata a todas as pessoas que de alguma forma contribuíram para o meu desenvolvimento acadêmico e profissional.

“Para avaliar a importância real de uma pessoa, devemos pensar nos efeitos que sua morte produziria”.

François Gaston de Levis

RESUMO

Esta pesquisa trata-se de um estudo descritivo e analítico da modelagem do custo operacional de empresas brasileiras transmissoras de energia elétrica, utilizando o método de regressão linear múltipla. Seu objetivo foi identificar as variáveis fortemente associadas à definição do custo operacional das referidas empresas, bem como comparar os resultados às componentes consideradas no modelo proposto na Nota Técnica nº 383/2012-SRE/ANEEL para definição dos custos de operação. Os dados empregados nesta pesquisa foram obtidos no site da Agência Nacional de Energia Elétrica (ANEEL) e, para fins de estudo, foram também utilizados recursos do pacote estatístico R. Inicialmente foi analisada a correlação entre a variável dependente custo operacional e as possíveis variáveis explicativas e, reconhecidas as variáveis fortemente correlacionadas com a variável custo, foram ajustados modelos de regressão múltipla, a fim de identificar o modelo que melhor define o custo operacional das concessionárias de energia elétrica. Por fim, o modelo escolhido foi comparado ao modelo proposto na Nota Técnica nº 383/2012-SRE/ANEEL.

PALAVRAS-CHAVE: Transmissão de energia elétrica. Custo operacional. Correlação. Variável dependente. Variáveis independentes. Modelo de regressão. Regressão Múltipla.

ABSTRACT

This research it is about a descriptive and analytical study of the modeling of the operating cost of the transmitting electricity Brazilian companies, using the of multiple linear regression method. Their goal was to identify variables strongly associated with the definition of the operational cost of such companies, as well as compare the results to the components considered in the model proposed in the Technical Note nº 383/2012-SRE/ANEEL for definition of operating costs. The data used in this research were obtained from the National Electric Energy Agency (ANEEL) and, for purposes of study, resources were also used statistical package R. Initially was analyzed the correlation between the dependent variable operating cost and possible explanatory variables , and recognized the variables strongly correlated with the variable costs, were adjusted multiple regression models in order to identify the model that best defines the operational cost of utilities electricity. Finally, the chosen model was compared to that proposed in the Technical Note model nº 383/2012-SRE/ANEEL.

KEYWORDS: Electric power transmission. Operating cost. Correlation. Dependent variable. Independent variables. Regression model. Multiple Regression.

LISTA DE FIGURAS

FIGURA 1 – Gráficos de dispersão da variável dependente em função das variáveis independentes (eixo y na escala log) 43

FIGURA 2 – Séries temporais do custo operacional para as transmissoras de energia 44

FIGURA 3 – Gráfico de probabilidade normal dos resíduos..... 50

FIGURA 4 – Resíduos *versus* valores ajustados 51

FIGURA 5 – Gráfico de probabilidade normal dos resíduos..... 53

FIGURA 6 – Resíduos *versus* valores ajustados 54

FIGURA 7 – Gráfico de probabilidade normal dos resíduos..... 57

FIGURA 8 – Resíduos *versus* valores ajustados 57

FIGURA 9 – Gráfico dos resíduos ao longo do tempo para as empresas de transmissão de energia 58

FIGURA 10 – Gráfico de probabilidade normal dos resíduos..... 60

FIGURA 11 – Resíduos *versus* valores ajustados 60

FIGURA 12 – Gráfico dos resíduos ao longo do tempo para as empresas de transmissão de energia 61

LISTA DE TABELAS

TABELA 1 - Análise de Variância para testar a Significância da Regressão..... 25

TABELA 2 - Análise de Variância para testar a Significância da Regressão na Regressão Múltipla..... 36

TABELA 3 - Correlação de *Spearman* entre a variável custo operacional e as variáveis independentes..... 45

TABELA 4 - Correlação de *Spearman* entre as variáveis custo operacional e rede por nível de tensão..... 46

TABELA 5 - Correlação de *Spearman* entre a variável custo operacional e as variáveis de rede..... 47

TABELA 6 - Coeficientes e Estatística de Regressão..... 49

TABELA 7 - Análise residual..... 50

TABELA 8 - Coeficientes e Estatística de Regressão..... 52

TABELA 9 - Análise residual..... 53

TABELA 10 - Coeficientes e Estatística de Regressão..... 55

TABELA 11 - Análise residual..... 56

TABELA 12 - Coeficientes e Estatística de Regressão..... 58

TABELA 13 - Análise residual..... 59

LISTA DE SIGLAS

ANEEL – Agência Nacional de Energia Elétrica

CEEE – Companhia Estadual de Energia Elétrica

CEMIG – Companhia Energética de Minas Gerais

CHESF – Companhia Hidro Elétrica do São Francisco

COPEL – Companhia Paranaense de Energia

CT – Conexão de Transformador

CTEEP – Companhia de Transmissão de Energia Elétrica Paulista

DEA – Data Envelopment Analysis

EL – Entrada de Linha

IB – Interligação de Barramento

SUMÁRIO

1 INTRODUÇÃO 14

1.1 Justificativa 15

1.2 Problema de Pesquisa 15

1.3 Objetivos 16

1.3.1 Objetivo Geral 16

1.3.2 Objetivos Específicos 16

2 REFERENCIAL TEÓRICO 17

2.1 Diagrama de dispersão 17

2.2 Correlação 18

2.2.1 Coeficiente de Correlação de Pearson 18

2.2.2 Coeficiente de Correlação de Spearman 19

2.3 Análise de Regressão 19

2.3.1 Regressão Linear Simples 20

2.3.1.1 Estimativas de Mínimos Quadrados 20

2.3.1.2 Equação final da reta 21

2.3.1.3 Estimativas de Variância 21

2.3.1.4 Propriedades dos estimadores de Mínimos Quadrados 22

2.3.1.5 Análise da adequação do Modelo de Regressão 23

2.3.1.5.1 Testes de Hipóteses 23

2.3.1.5.2 Intervalos de Confiança 25

2.3.1.5.3 Análise Residual 26

2.3.1.5.4 Coeficiente de Determinação (R^2) 27

2.3.1.6 Transformações 28

2.3.2 Regressão Linear Múltipla 29

2.3.2.1 Estimativas de Mínimos Quadrados 29
2.3.2.2 Abordagem Matricial 31
2.3.2.3 Modelo Ajustado 31
2.3.2.4 Estimativas de Variância 32
2.3.2.5 Propriedades dos estimadores de Mínimos Quadrados 32
2.3.2.6 Análise da adequação do Modelo de Regressão 33
2.3.2.6.1 Testes de Hipóteses 34
2.3.2.6.2 Intervalos de Confiança 36
2.3.2.6.3 Análise Residual 37
2.3.2.6.4 Coeficiente de Determinação R^2 e R^2 Ajustado 38
2.3.2.7 Multicolinearidade 39
3 METODOLOGIA 40
3.1 Tipo de pesquisa 40
3.2 Universo e amostra 41
4 ESTUDO DE CASO 42
4.1 Coleta e tratamento dos dados 42
4.2 Análise dos resultados 43
4.2.1 Análise do diagrama de dispersão da variável dependente (custo operacional) versus variáveis independentes 43
4.2.2 Séries temporais do custo operacional para as empresas de transmissão de energia elétrica 44
4.2.3 Análise da correlação pelo método de Spearman 45
4.2.3.1 Variável custo operacional *versus* variáveis independentes 45
4.2.3.2 Variável custo operacional *versus* variáveis de comprimento de rede (km) 46
4.2.4 Ajuste do Modelo de Regressão Linear Múltipla 48
4.2.4.1 Modelos de regressão linear múltipla ajustados considerando as variáveis do modelo DEA 48

4.2.2.1.1 Modelo 1 49
4.2.2.1.2 Modelo 2 51
4.2.4.2 Modelos Propostos 54
4.2.2.2.1 Modelo 3 55
4.2.2.2.2 Modelo 4 58
5 CONCLUSÃO 62
REFERÊNCIAS 63
ANEXOS 64

1 INTRODUÇÃO

O setor elétrico brasileiro possui suas atividades segmentadas em geração de energia elétrica nas usinas, transmissão da energia para as cidades e distribuição da energia para os consumidores. A Agência Nacional de Energia Elétrica (ANEEL) é a responsável pela regulação e fiscalização de todos esses serviços prestados pelas empresas de energia elétrica.

Em 24 de outubro de 2012 a ANEEL publicou a Nota Técnica nº 383/2012-SRE/ANEEL apresentando alternativas para definição das receitas totais de transmissão para as empresas brasileiras de transmissão de energia elétrica. Nessa Nota Técnica é proposto um modelo de *Benchmarking* para definir os custos operacionais a partir de variáveis independentes. O modelo de *Benchmarking* é uma metodologia que consiste numa análise comparativa entre custos operacionais praticados por um conjunto de empresas. A partir desse modelo, é possível calcular um índice de eficiência para cada empresa. Estimados os índices de eficiência, a ANEEL estabelece políticas de redução de tarifas ou de compensações para as empresas de transmissão.

Em particular, a Nota Técnica nº 383/2012-SRE/ANEEL apresenta uma metodologia de *Benchmarking* conhecida como *Data Envelopment Analysis* (DEA): "Em síntese o DEA estima um parâmetro que é o resultado da comparação da empresa em análise com uma Fronteira de Eficiência, formada a partir das empresas consideradas as mais eficientes do setor" (fls. 4 Nota Técnica nº 383/2012-SRE/ANEEL, de 24 de outubro de 2012). Do ponto de vista estatístico, o modelo DEA pode ser caracterizado como um modelo não paramétrico que a partir de variáveis independentes, também definidas como variáveis produto, gera uma fronteira de eficiência. Uma fronteira de eficiência pode ser definida como o limite inferior de uma variável dependente ou variável insumo. Por exemplo, para as diferentes empresas de transmissão de energia, deseja-se estabelecer um limite inferior para o custo operacional dado um conjunto de variáveis insumo: comprimento de rede, número de consumidores, etc.

A Nota Técnica nº 383/2012-SRE/ANEEL indica que as variáveis independentes do modelo DEA são: comprimento de rede (Km), número de módulos, quantidade de

transformadores e potência entregue (MVA). Em particular, a Nota Técnica também apresenta a variável comprimento de rede estratificada segundo 6 (seis) diferentes grupos de tensão.

Neste trabalho foi aplicado o modelo de regressão linear múltipla para identificar as variáveis independentes que caracterizam a variável dependente custo operacional. Desse modo, objetiva-se identificar as componentes fortemente associadas à definição do custo operacional das empresas de transmissão de energia elétrica e comparar os resultados às variáveis independentes (ou produtos) utilizadas no modelo DEA.

1.1 Justificativa

O modelo DEA é utilizado para definir as receitas das concessionárias de energia elétrica. Dessa forma, um modelo que utilize de forma equivocada as variáveis independentes pode, a princípio, estimar um custo operacional maior que o custo real das concessionárias. Como consequência de um modelo irregular, a agência reguladora poderia impor às transmissoras de energia uma receita penalizada, ou seja, abaixo do custo operacional necessário para manter as suas atividades operacionais regulares. Portanto, é importante que as variáveis independentes e o modelo estatístico sejam coerentes com o comportamento do custo operacional.

1.2 Problema de Pesquisa

Toma-se como foco, no presente estudo, a análise da correlação entre a variável dependente custo operacional e as variáveis independentes (ou produtos) consideradas no modelo *Data Envelopment Analysis* (DEA) proposto na Nota Técnica n° 383/212 – SER/ANEEL de 24 de outubro de 2012. Neste sentido, coloca-se em questão: dentre as variáveis independentes utilizadas no modelo DEA, quais são as componentes fortemente associadas à definição do custo operacional das empresas transmissoras de energia elétrica se aplicado o método de regressão linear múltipla?

1.3 Objetivos

1.3.1 Objetivo Geral

A presente pesquisa tem o objetivo de identificar as variáveis independentes ou explicativas fortemente associadas à definição do custo operacional das concessionárias brasileiras de energia elétrica e comparar os resultados às variáveis independentes utilizadas no modelo DEA proposto na Nota Técnica n° 383/212 – SER/ANEEL de 24 de outubro de 2012

1.3.2 Objetivos Específicos

- ✓ Analisar a correlação entre o custo operacional das empresas de transmissão de energia elétrica e as possíveis variáveis independentes ou explicativas;

- ✓ Definir um modelo de regressão linear múltipla para o custo operacional a partir das variáveis independentes.

2 REFERENCIAL TEÓRICO

2.1 Diagrama de dispersão

Conforme afirmam Sharpe, De Veaux e Velleman (2011) o diagrama de dispersão é uma ferramenta gráfica que possibilita entender a relação entre duas variáveis quantitativas.

“Ao analisar o diagrama de dispersão, você consegue ver padrões, tendências, relacionamentos e até mesmo valores incomuns ocasionais, que diferem dos outros”. (SHARPE; DE VEAUX; VELLEMAN, 2011, p.201)

Para a construção do diagrama de dispersão as variáveis que serão imputadas em cada eixo do gráfico devem ser escolhidas com cautela, pois “uma variável tem o papel de explanatória ou variável previsora, enquanto a outra tem o papel de variável resposta. Colocamos a variável explanatória no eixo x e variável resposta no eixo y”. (SHARPE; DE VEAUX; VELLEMAN, 2011, p.201)

Construído o diagrama de dispersão, faz-se necessária a análise de alguns aspectos:

> […] a direção da associação é importante: um padrão que vai do canto superior esquerdo ao canto inferior direito é chamado de negativo. Um padrão que vai na outra direção é chamado de positivo.
>
> O segundo aspecto a ser procurado no diagrama de dispersão é sua forma. Se existe uma relação linear, ela irá aparecer como uma nuvem ou um agrupamento de pontos estendido geralmente numa forma consistente em reta.
>
> A terceira característica a ser observada num diagrama de dispersão é a força da relação.
>
> Finalmente, sempre procure pelo inesperado. […] Um exemplo de tal surpresa é uma observação incomum, ou atípica, fora do padrão geral do diagrama de dispersão.
>
> (SHARPE; DE VEAUX; VELLEMAN, 2011, p.202)

2.2 Correlação

De acordo com Sharpe, De Veaux e Velleman (2011) a correlação mede a força da associação linear entre duas variáveis quantitativas, sendo representada pelo coeficiente de correlação e possuindo como propriedades básicas:

> - **O sinal de um coeficiente de correlação fornece a direção da associação.**
> - **A correlação é sempre um numero entre -1 e 1.** A correlação pode ser exatamente igual a -1,0 ou +1, mas cuidado. Esses valores são incomuns em dados reais, porque significam que todos os pontos dos dados caem exatamente sobre uma linha reta.
> - **A correlação trata x e y simetricamente**. A correlação entre x e y é a mesma correlação entre y e x.
> - **A correlação não tem unidades**. Esse fato é importante quando as unidades dos dados são um tanto vagas (satisfação do cliente, eficiência do trabalhador, produtividade e assim por diante).
> - **A correlação não é afetada por mudanças no centro ou escala de ambas as variáveis.** Mudar as unidades ou a base das variáveis não afeta o coeficiente de correlação, porque a correlação depende somente dos escores – z.
> - **A correlação mensura a força da associação linear entre duas variáveis.** As variáveis podem estar fortemente associadas, mas ainda ter uma pequena correlação se a associação não for linear.
> - **A correlação é sensível às observações incomuns.** Um único valor atípico pode transformar uma correlação pequena em uma grande e vice-versa.
>
> (SHARPE; DE VEAUX; VELLEMAN, 2011, p. 210)

2.2.1 Coeficiente de Correlação de Pearson

A seguir é apresentada a equação de cálculo do coeficiente de correlação amostral entre as variáveis X e Y, segundo Sharpe, De Veaux e Velleman (2011):

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$, i= 1, 2, 3, ..., n.

2.2.2 Coeficiente de Correlação de Spearman

Segundo Pirie (1988) para o cálculo do coeficiente de *Spearman* é considerado os postos ao invés dos valores observados, ou seja, o valor das variáveis é substituído por seu posto no grupo todo. Posto é a posição que um número ocupa dentro de um rol, quando os dados estão classificados em ordem crescente.

O coeficiente de correlação de postos é comumente representado pela letra grega ρ e

> [...] ao contrário do coeficiente de correlação de Pearson, não requer a suposição que a relação entre as variáveis é linear, nem requer que as variáveis sejam medidas em intervalo de classe; pode ser usado para as variáveis medidas no nível ordinal. O ρ é dado por:
>
> $$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n^3 - n)}$$
>
> Onde:
>
> d_i= a diferença entre cada posto de valor correspondentes de x e y, e
> n= o número dos pares dos valores. (PIRIE, 1988, p. 584-587)

2.3 Análise de Regressão

De acordo com Montgomery e Runger (2009), a análise de regressão é uma ferramenta estatística para modelagem e avaliação da relação existente entre duas ou mais variáveis.

Nas seções a seguir, serão conceituados os modelos de regressão linear simples e linear múltipla.

2.3.1 Regressão Linear Simples

Segundo Sharpe, De Veaux e Velleman (2011) um modelo linear consiste em uma equação de linha reta através dos dados em estudo, sendo que “nem todos os pontos no diagrama de dispersão se alinham, mas uma linha reta pode resumir o padrão geral com somente alguns parâmetros”. (SHARPE; DE VEAUX; VELLEMAN, 20011, p.235)

O modelo de regressão é denominado linear simples quando possui apenas uma variável independente ou regressor X e uma variável dependente ou variável de resposta Y, conforme explicado por Montgomery e Runger (2009).

> Suponha que a relação verdadeira entre Y e x seja uma linha reta e que a observação Y em cada nível de x seja uma variável aleatória. O valor esperado de Y para cada valor de x é: $E(Y \mid x) = \beta_0 + \beta_1 x$,
>
> sendo a interseção β_0 e a inclinação β_1 coeficientes desconhecidos da regressão. Consideramos que cada observação, Y, possa ser descrita pelo modelo: $Y = \beta_0 + \beta_1 x + \varepsilon$,
> em que ε é um erro aleatório com média zero e variância σ^2. Os erros aleatórios correspondendo a diferentes observações são também considerados variáveis aleatórias não-correlacionadas.
>
> (MONTGOMERY; RUNGER, 2009, p. 237)

2.3.1.1 Estimativas de Mínimos Quadrados

“As estimativas de β_0 e β_1 devem resultar em uma linha que seja (em algum sentido) o “melhor ajuste” para os dados.” (MONTGOMERY; RUNGER, 2009, p. 237)

Montgomery e Runger (2009) afirmam ainda que o critério para estimar os coeficientes de regressão é denominado método dos mínimos quadrados, sendo que as estimativas de mínimos quadrados da interseção e da inclinação no modelo de regressão linear simples são:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_1^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}},$$

em que $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ e $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$

2.3.1.2 Equação final da reta

De acordo com Montgomery e Runger (2009) a linha estimada ou ajustada da regressão é dada por:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

“Note que cada par de observações satisfaz a relação: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$, i= 1, 2,...,n. Sendo $e_i = y_i - \hat{y}_i$ chamado de resíduo. O resíduo descreve o erro no ajuste do modelo para a i-ésima observação y_i .(MONTGOMERY; RUNGER, 2009, p. 238)

2.3.1.3 Estimativas de Variância

Segundo Montgomery e Runger (2009) a variância (σ^2) do termo do erro ε, é um outro parâmetro não conhecido no modelo de regressão, sendo que os resíduos (e_i) são utilizados no cálculo da estimativa da variância.

Os autores acima afirmam ainda que a soma dos quadrados dos resíduos ou erros é dada por:

$SQ_E = SQ_T - \hat{\beta}_1 S_{xy}$, em que $SQ_T = \sum (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ é a soma total dos quadrados da variável resposta e o valor esperado da soma dos quadrados dos erros é $E(SQ_E) = (n-2)\,\sigma^2$.

Sendo assim, de acordo com Montgomery e Runger (2009), um estimador não tendencioso de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{SQ_E}{n-2}.$$

2.3.1.4 Propriedades dos estimadores de Mínimos Quadrados

As propriedades dos estimadores de mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$ são:

> Temos considerado o termo do erro, ε, no modelo $Y = \beta_0 + \beta_1 x + \varepsilon$ como uma variável aleatória com média zero e variância σ^2. Uma vez que os valores de x são fixos, Y é uma variável aleatória com média $\mu_{Y|x} = \beta_0 + \beta_1 x$ e variância σ^2. Consequentemente os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ dependem dos valores observados dos y's; assim os estimadores dos mínimos quadrados dos coeficientes de regressão podem ser vistos como variáveis aleatórias;
>
> $\hat{\beta}_1$ é um estimador não tendencioso da inclinação verdadeira de β_1;
>
> $\hat{\beta}_0$ é um estimador não tendencioso da interseção β_0;
>
> A covariância das variáveis aleatórias $\hat{\beta}_0$ e $\hat{\beta}_1$ não é zero. Pode ser mostrado que $\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \bar{x} / S_{xx}$; (MONTGOMERY; RUNGER, 2009, p. 243)

De acordo com Montgomery e Runger (2009) em uma regressão linear simples, o erro padrão estimado da inclinação e o erro padrão estimado da interseção são dados, respectivamente, por:

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{e} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

2.3.1.5 Análise da adequação do Modelo de Regressão

Para o ajuste de um modelo de regressão linear faz-se necessária à verificação de certas suposições.

De acordo com Sharpe, De Veaux e Velleman (2011) para o ajuste de um modelo linear os dados devem ser quantitativos, bem como se deve assumir que a relação existente entre as variáveis efetivamente é linear e inexistem pontos que estejam distantes a ponto de distorcer a linha de melhor aderência.

Montgomery e Runger (2009) afirmam que a estimativa dos parâmetros do modelo demanda a suposição de que os erros sejam variáveis aleatórias não correlacionadas, com média zero e variância constante, pois os testes de hipótese e estimação do intervalo requerem que os erros sigam uma distribuição normal.

A fim de verificar a adequação do modelo, ou seja, confirmar se as suposições acima são válidas, são utilizados métodos de análise, conforme apresentados a seguir.

2.3.1.5.1 Testes de Hipóteses

A realização de testes de hipóteses para os parâmetros do modelo faz parte de métodos estatísticos utilizados na verificação da adequação de um modelo de regressão linear.

Segundo Montgomery e Runger (2009) para testar as hipóteses sobre a inclinação e a interseção do modelo faz-se necessário supor que os erros são normais e independentemente distribuídos com média zero e variância σ^2.

Conforme afirmam os autores acima, as hipóteses apropriadas para o teste da significância do modelo são:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Usando o teste t com n-2 graus de liberdade sujeito a H_0, a estatística será definida por:

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

Sendo que rejeitaremos $H_0 : \beta_1 = 0$ se $|t_0| > t_{\alpha/2,n-2}$

De acordo com Montgomery e Runger (2009), não rejeitar a hipótese nula é o mesmo que concluir que não há relação linear entre a variável independente ou explicativa e a variável resposta (Y). Contudo, rejeitar $H_0 : \beta_1 = 0$ pode significar que o modelo de linha reta seja adequado, ou seja, X é importante para explicar a variabilidade de Y.

Outro método usado para testar a significância da regressão é a análise de variância.

> O procedimento divide a variância total na variável de resposta em componentes significantes, como base para o teste. A identidade de análise de variância é dada a seguir:
>
> $$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
>
> Os dois componentes no lado direito da equação medem, respectivamente, a quantidade da variabilidade em y_i, devido à linha de regressão, e a variação residual deixada sem explicação pela linha de regressão. Geralmente chamamos $SQ_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ de soma dos quadrados dos erros e $SQ_R = \sum_{I=1}^{n}(\hat{y}_i - \bar{y})^2$ de soma dos quadrados de regressão. Simbolicamente, a equação pode ser escrita como $SQ_T = SQ_R + SQ_E$, sendo $SQ_T = \sum(y_i - \bar{y})^2$ a soma dos quadrados total corrigida de y. (MONTGOMERY; RUNGER, 2009, p. 245)

Segundo Montgomery e Runger (2009), se a hipótese nula $H_0 : \beta_1 = 0$ for verdadeira, a estatística,

$F_0 = \dfrac{SQ_R / 1}{SQ_E / (n-2)} = \dfrac{MQ_R}{MQ_E}$, segue a distribuição $F_{1,n-2}$ e rejeitaremos a hipótese nula se $f_0 > f_{\alpha;1;n-2}$. As grandezas MQ_R e MQ_E são denominadas média quadrática da regressão e média quadrática do erro, respectivamente.

Abaixo é apresentada a tabela de análise de variância, a qual resume o procedimento de teste:

TABELA 1

Análise de Variância para Testar a Significância da Regressão

Fonte de Variação	Soma Quadrática	Graus de Liberdade	Média Quadrática	F_0
Regressão	$SQ_R = \hat{\beta}_1 S_{xy}$	1	MQ_R	MQ_R / MQ_E
Erro	$SQ_E = SQ_T - \hat{\beta}_1 S_{xy}$	n - 2	MQ_E	
Total	SQ_T	n - 1		

Fonte: MONTGOMERY; RUNGER, 2009

Vale observar que, a interpretação do teste F na regressão linear simples é a mesma do teste t, ou seja, ambos os testes conduzirão as mesmas conclusões.

2.3.1.5.2 Intervalos de Confiança

Os intervalos de confiança também fazem parte de métodos utilizados na verificação da adequação de um modelo de regressão linear. A seguir serão apresentadas técnicas para a construção de intervalos de confiança para a Inclinação e Interseção, bem como para a Resposta Média.

Conforme afirmam Douglas C. Montgomery e George C. Runger (2009), “sob a suposição de que as observações sejam normal e independentemente distribuídas”, na

regressão linear simples o intervalo de confiança de 100 (1-α)% para a inclinação β_1 e para a interseção β_0 são, respectivamente:

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \leq \beta_o \leq \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

Quanto à resposta média, “um intervalo de confiança de 100 (1-α)% para a resposta média no valor de x=x_0, como $\hat{\mu}_{Y|X_O}$, é dado por:

$$\hat{\mu}_{Y|X_O} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \leq \mu_{Y|X_O} \leq \hat{\mu}_{Y|X_O} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Sendo $\hat{\mu}_{Y|X_O} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ calculado a partir do modelo ajustado de regressão”. (MONTGOMERY; RUNGER, 2009, p. 247)

2.3.1.5.3 Análise Residual

O resíduo “é a diferença entre o valor observado e o valor correspondente prevista pelo modelo de regressão”. (SHARPE; DE VEAUX; VELLLEMAN, 2011, p. 250)

Em outras palavras, os resíduos são a parte dos dados que não foi explicada pelo modelo de regressão. A construção do diagrama de dispersão permite a verificação da condição de linearidade e a condição do valor atípico.

> “Um diagrama de dispersão dos resíduos *versus* os valores de x deveria ser um gráfico sem padrões. Ele não deve ter características importantes – nenhuma direção, nenhuma forma. Ele deve se alongar horizontalmente sem mostrar curvas e não deve ter valores atípicos. […] Quanto melhor o modelo se ajusta aos dados, menos os resíduos irão variar em torno da reta.” (SHARPE; DE VEAUX; VELLEMAN, 2011, p.242)

De acordo com Sharpe, Veaux e Velleman (2011), o desvio padrão dos resíduos, s_e, fornece uma medida de quanto os pontos se espalham ao redor da linha de regressão, sendo que é necessário assumir que o desvio padrão em torno da linha reta é o mesmo todas as vezes que o modelo é aplicado (condição da mesma dispersão).

A estimativa do desvio padrão residual é dada por:

$$s_e = \sqrt{\frac{\sum e^2}{n-2}}$$

Uma vez que $\bar{e} = 0$, não é preciso subtrair a média dos resíduos e a subtração de n-2 é devido à subtração de dois parâmetros: inclinação e intercepto.

Outro método de análise dos resíduos, de acordo com Montgomery e Runger (2009), é a plotagem da distribuição normal, a fim de verificar se os resíduos estão normalmente distribuídos. Além disso, pode-se também padronizar os resíduos, calculando $d_i = e_i / \hat{\sigma}^2$, i= 1, 2, n. Se os erros seguem uma distribuição normal, então aproximadamente 95% dos resíduos padronizados devem estar no intervalo (-2, +2), sendo que os que estiverem distantes desse intervalo podem indicar a presença de um *outlier*, ou seja, uma observação atípica ao resto dos dados, devendo este ser analisado com cautela.

2.3.1.5.4 Coeficiente de Determinação (R^2)

O coeficiente de determinação R^2 ou razão da soma dos quadrados é uma medida muito usada no modelo de regressão. Segue abaixo a equação para obtenção do R^2, segundo Montgomery e Runger (2009):

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_e}{SQ_T}$$

Conforme afirmam Montgomery e Runger (2009) a estatística R^2 é constantemente usada para a verificação da adequação do modelo de regressão, sendo interpretada como a quantidade de variabilidade nos dados explicada pelo modelo. Entretanto, deve haver cuidado na utilização do R^2, visto que há várias interpretações incorretas a respeito dessa estatística:

> Em geral, R^2 não mede a magnitude da inclinação da linha de regressão.
>
> Um grande valor de R^2 não implica uma inclinação pronunciada.
>
> O R^2 não mede a adequação do modelo, uma vez que ele pode ser artificialmente aumentado através da adição, ao modelo, de termos polinomiais de ordens superiores em x. Mesmo se y e x estiverem relacionados de uma maneira não linear, R^2 será frequentemente grande.
>
> Muito embora R^2 seja grande, isso não implica necessariamente que o modelo de regressão forneça previsões exatas de futuras observações.
>
> (MONTGOMERY; RUNGER, 2009, p. 243)

Dessa forma, o modelo de regressão deve ser escolhido com base na análise conjunta de todos os métodos apresentados anteriormente.

2.3.1.6 Transformações

Há casos em que se constata que o modelo de regressão não é adequado porque a função verdadeira de regressão não é linear. No entanto, em alguns desses casos, “uma função não-linear pode ser expressa como uma linha reta, usando uma transformação adequada. Tais modelos não-lineares são chamados de intrinsecamente lineares”. (MONTGOMERY; RUNGER, 2009, p. 256)

Como exemplos de modelo não lineares que sejam intrinsecamente lineares:

> Considere a função exponencial $Y = \beta_0 e^{\beta 1x}\varepsilon$. Essa função é intrinsecamente linear, uma vez que ela pode ser transformada em uma linha reta por uma transformação logarítmica $\ln Y = \ln \beta_0 + \beta_1 x + \ln \varepsilon$. Essa transformação requer que os termos transformados do erro, $\ln \varepsilon$, sejam normal e independentemente distribuídos, com média 0 e variância σ^2.

> Uma outra função intrinsecamente linear é $Y = \beta_0 + \beta_1\left(\frac{1}{x}\right) + \varepsilon$. Usando a transformação recíproca z=1/x, o modelo é linearizado para $Y = \beta_0 + \beta_1 z + \varepsilon$.
>
> (MONTGOMERY; RUNGER, 2009, p. 256)

2.3.2 Regressão Linear Múltipla

De acordo com Montgomery e Runger (2009) o modelo de regressão múltipla é aquele que possui mais de um regressor.

> Em geral, a variável dependente ou de resposta, y, pode estar relacionada a *k* variáveis independentes ou regressoras.
> O modelo $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$ é chamado de modelo de regressão linear múltipla com *k* variáveis regressoras. Os parâmetros $\beta_j, j = 0,1,...,k$, são chamados de coeficientes de regressão. Esse modelo descreve um hiperplano no espaço *k*-dimensional das variáveis regressoras $\{x_j\}$. O parâmetro β_j representa a variação esperada na resposta Y por unidade de variação unitária em x_j, quando todos os outros regressores restantes $x_i (i \neq j)$ forem mantidos constantes. (MONTGOMERY; RUNGER, 2009, p. 266)

Montgomery e Runger (2009) afirmam ainda que comumente os modelos de regressão linear múltipla são usados como funções de aproximações, ou seja, "a verdadeira relação funcional entre Y e $x_1, x_2, ..., x_k$ é desconhecida, porém em certas faixas das variáveis independentes, o modelo de regressão linear é uma aproximação adequada".

2.3.2.1 <u>Estimativas de Mínimos Quadrados</u>

Conforme afirmam Montgomery e Runger (2009) a função dos mínimos quadrados é dada por:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{k} \beta_j x_{ij} \right)^2$$

“Queremos minimizar *L* com relação a $\beta_0, \beta_1, ..., \beta_k$. As estimativas de mínimos quadrados de $\beta_0, \beta_1, ..., \beta_k$ têm de satisfazer

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_O, \hat{\beta}_1, ..., \hat{\beta}_K} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) = 0$$

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_O, \hat{\beta}_1, ..., \hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right)x_{ij} = 0 \qquad \text{j = 1, 2,..., k.”}$$

(MONTGOMERY; RUNGER, 2009, p. 267)

Montgomery e Runger (2009) explicam ainda que, com a simplificação das equações acima são originadas as chamadas equações normais de mínimos quadrados:

$$n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_{1i} + \hat{\beta}_2\sum_{i=1}^{n}x_{2i} + ... + \hat{\beta}_k\sum_{i=1}^{n}x_{ki} = \sum_{i=1}^{n}y_i$$

$$\hat{\beta}_0\sum_{i=1}^{n}x_{i1} + \hat{\beta}_1\sum_{i=1}^{n}x_{i1}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{i1}x_{i2} + ... + \hat{\beta}_k\sum_{i=1}^{n}x_{i1}x_{ik} = \sum_{i=1}^{n}x_{i1}y_i$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\hat{\beta}_0\sum_{i=1}^{n}x_{ik} + \hat{\beta}_1\sum_{i=1}^{n}x_{ik}x_{i1} + \hat{\beta}_2\sum_{i=1}^{n}x_{ik}x_{i2} + ... + \hat{\beta}_k\sum_{i=1}^{n}x_{ik}^2 = \sum_{i=1}^{n}x_{ik}y_i$$

De acordo com Montgomery e Runger (2009) há *p*=k+1 equações normais, ou seja, uma para cada um dos coeficientes desconhecidos de regressão, sendo que a solução para as referidas equações normais serão os estimadores de mínimos quadrados de cada coeficiente de regressão.

2.3.2.2 Abordagem Matricial

Suponha que haja *k* variáveis regressoras e *n* observações, (x_{i1}, x_{i2}, ..., x_{ik}, y_i), i=1,2,n, e que o modelo relacionando os regressores à resposta seja $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$ i= 1, 2, ..., n. Segundo Montgomery e Runger (2009), o referido modelo pode ser expresso na notação matricial como $y = X\beta + \varepsilon$:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} \cdots x_{1k} \\ 1 & x_{21} & x_{22} \ldots x_{2k} \\ \vdots & \vdots & \vdots \quad \vdots \\ 1 & x_{n1} & x_{n2} \cdots x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{e} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

"Em geral, y é um vetor (*n* x 1) das observações, X é uma matriz (*n* x *p*) dos níveis das variáveis independentes, β é um vetor (*p* x 1) dos coeficientes de regressão e ε é um vetor (*n* x 1) dos erros aleatórios". (MONTGOMERY; RUNGER, 2009, p. 269)

Montgomery e Runger (2009) afirmam que as equações normais de mínimos quadrados na forma matricial são dadas por $X'X\hat{\beta} = X'y$, e que para resolvê-las é necessário multiplicar ambos os lados da equação anterior pelo inverso de $X'X$. Desta forma, a estimativa de mínimos quadrados de β é:

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

Ressalta-se que na prática os cálculos de regressão múltipla são geralmente realizados com a utilização de softwares estatísticos.

2.3.2.3 Modelo Ajustado

De acordo com Montgomery e Runger (2009) o modelo estimado ou ajustado da regressão é dado por:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \beta_j x_{ij} \qquad \text{i= 1, 2, ..., n}$$

Na notação matricial, o modelo ajustado é $\hat{y} = X\hat{\beta}$. Por fim, o vetor (*n* x 1) dos resíduos é denotado por $e = y - \hat{y}$. (MONTGOMERY; RUNGER, 2009, p. 271)

2.3.2.4 Estimativas de Variância

Segundo Montgomery e Runger (2009) na regressão linear múltipla com *p* parâmetros, um estimador não tendencioso de σ^2 é:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SQ_E}{n-p}$$

Na equação acima “o numerador é chamado de erro ou de soma dos quadrados dos resíduos e o denominador $n-p$ é chamado de graus de liberdade do erro ou do resíduo”. (MONTGOMERY; RUNGER, 2009, p. 271)

O SQE ou soma dos quadrados dos resíduos também pode ser obtido pela seguinte fórmula:

$$SQE = \frac{y'y - \hat{\beta}'X'y}{n-p}$$

2.3.2.5 Propriedades dos estimadores de Mínimos Quadrados

As propriedades dos estimadores de mínimos quadrados $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ são:

> Consideramos que os erros, $\in_i$ sejam estatisticamente independentes, com média zero e variância σ^2. Sob essas suposições, os estimadores de mínimos quadrados $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ são estimadores não tendenciosos dos coeficientes de regressão $\beta_0, \beta_1, ..., \beta_k$.
>
> As variâncias dos $\hat{\beta}'s$ são expressas em termos dos elementos da inversa matriz X'X. A inversa de X'X vezes a constante σ^2 representa a matriz de covariâncias dos coeficientes de regressão $\hat{\beta}$. Os elementos da diagonal de $\sigma^2(X'X)^{-1}$ são as variâncias de $\beta_0, \beta_1, ..., \beta_k$ e os elementos fora da diagonal dessa matriz são as covariâncias. Em geral, a matriz de covariâncias de $\hat{\beta}$ é uma matriz simétrica (*p* x *p*), cojo *jj*-ésimo elemento é a variância de $\hat{\beta}_j$ e cujo *i,j*-ésimo elemento é a covariância entre $\hat{\beta}_i$ e $\hat{\beta}_j$, ou seja, $\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2 C$. As estimativas das variâncias desses coeficientes de regressão são obtidas trocando σ^2 por sua estimativa. (MONTGOMERY; RUNGER, 2009, p. 271)

Conforme afirmam Douglas C. Montgomery e George C. Runger (2009), "quando σ^2 for trocado por sua estimativa $\hat{\sigma}^2$, a raiz quadrada da variância estimada do *j*-ésimo coeficiente de regressão é chamada de erro padrão estimado de $\hat{\beta}_j$ ou $ep(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$". Os referidos erros são uma medida da precisão de estimação para os coeficientes de regressão, sendo que erros pequenos indicam boa precisão.

2.3.2.6 Análise da adequação do Modelo de Regressão

No modelo de regressão múltipla as suposições são aproximadamente as mesmas da regressão simples. Os autores Sharpe, De Veaux e Velleman (2011) apresentam como suposições e condições da regressão múltipla:

- Suposição de linearidade: para verificar se há uma relação linear subjacente no modelo de regressão múltipla, é necessário verificar a condição de linearidade para cada um das variáveis explicativas através do diagrama de dispersão de y *versus* cada variável X. Assim como na regressão simples, também se sugere plotar o diagrama dos resíduos, a fim de constatar possíveis violações das condições de linearidade.

- Suposição de independência: deve-se verificar se a suposição é válida, ou seja, se os erros do verdadeiro modelo de regressão são independentes um dos outros. Além disso, os dados devem ser provenientes de uma amostra aleatória ou um experimento aleatório (condição de aleatoriedade);
- Suposição de igualdade das variâncias: a variabilidade dos erros deve ser aproximadamente a mesma para todos os valores de cada variável independente ou explicativa. Para ver se tal suposição é válida, sugere-se construir o diagrama de dispersão e verificar a condição da mesma dispersão;
- Suposição de normalidade: assume-se que os erros seguem uma distribuição Normal, sendo que a condição de normalidade é verificada no histograma ou diagrama da probabilidade Normal dos resíduos.

Assim como na regressão linear simples, com o intuito de verificar a adequação do modelo, na regressão múltipla também são utilizados métodos de análise, conforme apresentados a seguir.

2.3.2.6.1 Testes de Hipóteses

Conforme já relatado anteriormente, certos testes de hipóteses referentes aos parâmetros do modelo são válidos na verificação da adequação deste.

Segundo Montgomery e Runger (2009), como no caso da regressão linear simples, testes de hipóteses requisitam que os erros sejam normais e independentemente distribuídos com média zero e variância σ^2.

Conforme afirmam os autores acima, o teste da significância do modelo permite verificar se existe uma relação linear entre a variável resposta e um subconjunto de regressores x_1, x_2, ..., x_k, sendo que as hipóteses apropriadas são:

$H_0 : \beta_1 = \beta_2 = \ldots . \beta_k = 0$

$H_1 : \beta_j \neq 0$, para no mínimo um j.

Rejeitar a hipótese nula H_0 significa dizer que pelo menos uma das variáveis regressoras contribui significativamente para o modelo.

De acordo com Montgomery e Runger (2009), o teste para a significância da regressão "é uma generalização do procedimento usado na regressão linear simples. A soma total dos quadrados SQ_T é dividida na soma dos quadrados devido à regressão e na soma dos quadrados devido ao erro, $SQ_T = SQ_R + SQ_E$". (MONTGOMERY; RUNGER, 2009, p. 278)

A estatística de teste para $H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$ é:

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

Onde:

$$SQ_E = y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} - \left[\hat{\beta}' X' y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] \text{ ou } SQ_E = SQ_T - SQ_R$$

$$SQ_R = \hat{\beta}' X' y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

"Devemos rejeitar H_0 se o valor calculado da estatística de teste, f_0, for maior do que $f_{\alpha,k,n-p}$". (MONTGOMERY; RUNGER, 2009, p. 278)

O procedimento de teste é resumido em uma tabela de variância, conforme apresentado a seguir:

TABELA 2

Análise de Variância para Testar a Significância da Regressão na Regressão Múltipla

Fonte de Variação	Soma Quadrática	Graus de Liberdade	Média Quadrática	F_0
Regressão	SQ_R	k	MQ_R	MQ_R / MQ_E
Erro ou resíduo	SQ_E	n – p	MQ_E	
Total	SQ_T	n – 1		

Fonte: MONTGOMERY; RUNGER, 2009

2.3.2.6.2 Intervalos de Confiança

Em modelos de regressão múltipla, é útil construir estimativas de intervalos de confiança para os coeficientes de regressão. “O desenvolvimento de um procedimento para obter esses intervalos requer que os erros $\{\varepsilon_i\}$ sejam normal e independentemente distribuídos, com média zero e variância σ^2”. (MONTGOMERY; RUNGER, 2009, p. 283)

Segundo Montgomery e Runger (2009) um intervalo de confiança de 100 (1-α)% para o coeficiente de regressão $\beta_j, j = 0,1,\ldots,k$ no modelo de regressão linear múltipla é dado por:

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

Sendo, $\sqrt{\hat{\sigma}^2 C_{jj}}$ o erro padrão do coeficiente de regressão $\hat{\beta}_j$.

Relativamente à resposta média, de acordo com os autores acima, no modelo de regressão linear múltipla um intervalo de confiança de 100 (1-α)% para a resposta média no ponto x_{01}, x_{02}, x_{0k} é:

$$\hat{\mu}_{Y|X_O} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x'_0 (X'X)^{-1} x_0} \le \mu_{Y|X_O} \le \hat{\mu}_{Y|X_O} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x'_0 (X'X)^{-1} x_0}$$

2.3.2.6.3 Análise Residual

Na regressão múltipla, a análise dos resíduos, dados por $e_i = y_i - \hat{y}_i$, é importante para a verificação da adequação do modelo, conforme afirma Montgomery e Runger (2009). Para tal análise, a construção de gráficos de resíduos é útil, bem como plotar os resíduos contra variáveis que não façam parte do modelo, mas que sejam possíveis candidatas à inclusão no modelo. Os autores afirmam ainda que é possível verificar padrões de comportamento nesses gráficos.

De acordo com Montgomery e Runger (2009), o resíduo na forma de *Student* é o cálculo de resíduos escalonados (possuem desvio padrão aproximadamente igual a um) mais utilizados, pois com ele possíveis *outliers* ou observações não usuais ficam mais óbvios nos gráficos residuais. É dado por:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \quad \text{i= 1, 2, ..., n.}$$

Em que h_{ii} é o i-ésimo elemento da diagonal da matriz $H = X(X'X)^{-1}X'$. A matriz H em algumas vezes é chamada de matriz “chapéu”, pois $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$.

“Sob as suposições usuais de que os erros do modelo são independentemente distribuídos, com média zero e variância σ^2, podemos mostrar que a variância do i-ésimo resíduo e_i é $V(e_i) = \sigma^2(1-h_{ij})$, i= 1, 2,..., n.” (MONTGOMERY; RUNGER, 2009, p. 287)

Montgomery e Runger (2009) afirmam que, quando aplicamos a regressão múltipla, casualmente encontramos algum subconjunto de observações excepcionalmente influentes. Tais pontos devem ser avaliados, a fim de verificar se os mesmos controlam muitas propriedades do modelo. Se esses pontos influentes forem “ruins” ou errôneos de algum modo, então deverão ser eliminados. Caso nada haja de errado com os pontos, deve-se ao menos verificar se eles produzem ou não resultados consistentes com o restante dos dados.

2.3.2.6.4 Coeficiente de Determinação R^2 e R^2 Ajustado

Segundo Sharpe, De Veaux e Velleman (2011) na regressão múltipla o coeficiente de determinação R^2 é interpretado como o percentual da variabilidade de Y que é explicada pelo modelo com todas as variáveis independentes incluídas, sendo dado por:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

Conforme afirmam Montgomery e Runger (2009), o fato de o R^2 sempre aumentar quando um regressor é adicionado ao modelo torna difícil verificar se o aumento está nos dizendo algo útil acerca do novo regressor. Por tal motivo, é preferível a utilização da estatística R^2 Ajustada, visto que a mesma somente aumentará quando uma nova variável for adicionada ao modelo se a nova variável reduzir a média quadrática do erro.

Acerca do R^2 Ajustado, é possível afirmar que o mesmo:

> [...] impõe uma "penalidade" para cada termo novo que é adicionado ao modelo, na tentativa de fazer modelos de tamanhos diferentes (número de variáveis previsoras) comparáveis. Ele difere do R^2 porque pode diminuir quando uma variável previsora é adicionada ao modelo de regressão ou crescer quando um previsor é removido, se ele não contribuir para melhorar o modelo. Na verdade, ele pode até mesmo ser negativo. (SHARPE; DE VEAUX; VELLEMAN, 2011, p.564)

De acordo com Sharpe, De Veaux e Velleman (2011), em uma regressão múltipla com *k* variáveis previsoras e *n* casos, o R^2 Ajustado é definido como:

$$R^2_{ajust} = 1 - (1 - R^2)\frac{n-1}{n-k-1} = 1 - \frac{SQE/(n-k-1)}{SQT/(n-1)}$$

2.3.2.7 Multicolinearidade

Nos modelos de regressão múltipla esperamos encontrar dependência entre a variável resposta e as variáveis independentes. Porém, na maioria dos casos também é constatada dependências entre as variáveis regressoras. De acordo com Montgomery e Runger (2009), quando essas dependências são fortes, dizemos que existe multicolinearidade, a qual pode gerar sérios efeitos nas estimativas dos coeficientes de regressão e na aplicabilidade geral do modelo estimado.

Segundo Sharpe, De Veaux e Velleman (2011), a estatística que mensura o grau de colinearidade do um determinado regressor com os outros é chamada de FIV (Fator de Inflação de Variância e é determinada como:

$$FIV_j = \frac{1}{1 - R_j^2} \quad \text{j=1, 2,..., k.}$$

O R^2 mostra, neste caso, quão bem o $j^{ésimo}$ regressor pode ser previsto por outros regressores. O termo $1 - R^2$ mensura o que aquele previsor deixou de trazer para o modelo de regressão.

"O FIV diz quanto a variância do coeficiente foi inflacionada devido a esta multicolinearidade. Quanto maior o FIV, maior o erro padrão do seu coeficiente e menos ele pode contribuir para o modelo de regressão". (SHARPE; DE VEAUX; VELLEMAN, 2011, p.609)

Segundo Montgomery e Runger (2009) alguns autores sugeriram que se qualquer fator de inflação da variância exceder 10, então a multicolinearidade será um problema. Já outros autores consideram esse valor muito liberal e sugerem que os fatores de inflação não devem exceder 4 ou 5.

- A colinearidade de qualquer previsor com os outros do modelo pode ser mensurada com o Fator de Inflação da Variância.
- A alta colinearidade leva a uma estimativa pobre do coeficiente e a um grande erro padrão (e, consequentemente, a uma estatística *t* pequena). O coeficiente pode parecer ter o tamanho errado ou apresentar até mesmo o sinal errado.

> ♦ Consequentemente, se um modelo de regressão múltipla tem um R^2 alto e um *F* grande, mas as estatísticas *t* individuais não são significativas, você deve suspeitar da colinearidade.
> ♦ A colinearidade é mensurada em termos do R_j^2 entre um previsor e *todos* os outros previsores no modelo. Ela não é mensurada em termos da correlação entre quaisquer dois previsores. É claro, se dois previsores estão altamente correlacionados, então o R_j^2 com ainda mais previsores deve ser pelo menos do mesmo valor e normalmente maior. (SHARPE; DE VEAUX; VELLEMAN, 2011, p.610)

De acordo com Montgomery e Runger (2009), para resolver o problema de multicolinearidade, “frequentemente sugere-se aumentar os dados com novas observações especificamente projetadas para fragmentar as dependências lineares aproximadas que existem corretamente”. Outra alternativa é remover certas variáveis do modelo, porém há a desvantagem de descartar a informação contida nas variáveis excluídas.

3 METODOLOGIA

Neste capítulo, será descrita a metodologia empregada no presente estudo, ou seja, os diversos recursos utilizados na elaboração deste trabalho, objetivando apresentar de forma sistemática o universo de pesquisa, possibilitando ao leitor um melhor entendimento do estudo.

3.1 Tipo de pesquisa

Segundo a classificação de pesquisa proposta por Vergara (2003) quanto aos fins e meios de investigação utilizados, tem-se que o presente estudo é classificado conforme a seguir:

a) Quanto aos fins, trata-se de uma pesquisa descritiva, uma vez que visa modelar o custo operacional gerado pelas transmissoras de energia elétrica de acordo com as variáveis: potência (MVA), quantidade de transformadores, módulos e comprimento de rede (km). Nesse sentido, Sylvia Constant

Vergara (2003) explicita que a pesquisa descritiva “expõe características de determinada população ou de determinado fenômeno. Pode também estabelecer correlações entre variáveis e definir sua natureza”.

b) Quanto aos meios de investigação utilizados, a presente pesquisa trata-se de um estudo de caso, pois tem como centro o custo operacional de transmissoras brasileiras de energia elétrica. Sylvia Constant Vergara (2003) afirma que estudo de caso é “o circunscrito a uma ou poucas unidades, entendidas essas como pessoa, família, produto, empresa, órgão público, comunidade ou mesmo país. Tem caráter de profundidade e detalhamento”.

3.2 Universo e amostra

Delimitar o universo de pesquisa ou população consiste em definir o “conjunto de elementos (empresas, produtos, pessoas, por exemplo) que possuem as características que serão objeto de estudo”. (VERGARA, 2003, p.50)

Neste trabalho tomar-se-á como universo de pesquisa todas as medidas, presentes e futuras, das transmissoras brasileiras de energia elétrica: FURNAS, CTEEP, CHESF, ELETRONORTE, ELETROSUL, CEMIG, COPEL e CEEE.

Conforme define Marconi e Lakatos (2003) a amostra trata-se de uma parte da população convenientemente escolhida, ou seja, é um subconjunto do universo.

A amostra utilizada neste estudo possui 39 observações e consiste em um conjunto de dados composto pelos ativos físicos e custos operacionais anuais contabilizados no período de 2007 a 2011 pelas empresas acima citadas.

4 ESTUDO DE CASO

4.1 Coleta e tratamento dos dados

Os dados empregados nesta pesquisa estão disponíveis on-line (*aneel.gov.br*). Nesse domínio são disponibilizadas informações relativas à energia elétrica no Brasil.

A base de dados utilizada possui 39 observações, sendo composta pelos ativos físicos e custos operacionais anuais contabilizados no período de 2007 a 2011 pelas transmissoras brasileiras de energia elétrica: FURNAS, CTEEP, CHESF, ELETRONORTE, ELETROSUL, CEMIG, COPEL e CEEE.

Neste estudo, a variável dependente é o custo operacional contabilizado anualmente para cada empresa. E as variáveis independentes são: comprimento de rede (km), módulos, quantidade de transformadores e potência (MVA).

Segundo a Nota Técnica n° 383/212 – SER/ANEEL de 24 de outubro de 2012, as variáveis comprimento de rede (linhas de transmissão), módulos (somatório do número de módulos EL – Entrada de Linha, CT – Conexão de Transformador e IB – Interligação de Barramento), quantidade de transformadores e potência (capacidade instalada de transformação – MVA) são as unidades modulares consideradas como representativas do produto das transmissores de energia elétrica. Tais variáveis de produto são consideradas na modelagem do custo operacional, uma vez que refletem as instalações oferecidas por cada transmissora.

A variável comprimento de rede (km) é apresentada na base de dados com o valor da rede total (km) disponibilizada anualmente por cada transmissora, bem como é apresentada desagregada por nível de tensão, sendo os níveis: 765 a 600 kV, 525 a 440 kV, 345 kV, 230 kV, 138 kV e 88 a 69 kV.

Para uma melhor visualização e compreensão dos resultados, os dados referentes à variável dependente custo operacional foram divididos por um milhão (1.000.000), bem como os dados relativos às variáveis preditoras: potência (MVA), módulos e comprimentos de rede (km) foram divididos por mil (1.000).

4.2 Análise dos resultados

Nesta subseção serão apresentadas as análises descritivas dos dados e as observações acerca dos resultados obtidos com a aplicação do método de regressão múltipla.

4.2.1 Análise do diagrama de dispersão da variável dependente (custo operacional) versus variáveis independentes

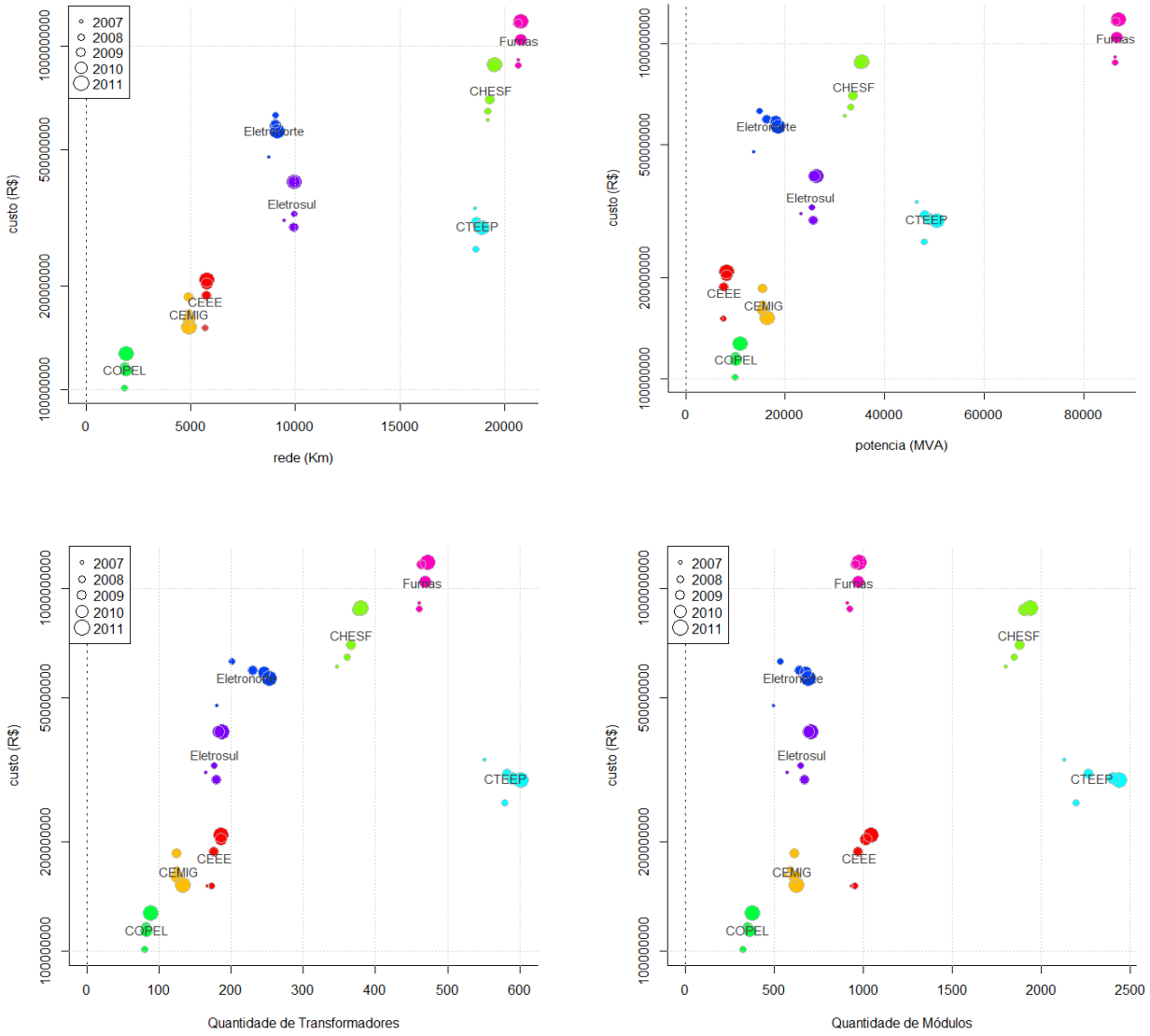

Figura 1 - Gráficos de dispersão da variável dependente em função das variáveis independentes (eixo y na escala log)

Analisando-se o gráfico de dispersão é possível verificar visualmente a existência de uma relação linear entre a variável custo operacional e as variáveis independentes, sendo que a intensidade dessa correlação é apresentada posteriormente através do cálculo do coeficiente de correlação.

4.2.2 Séries temporais do custo operacional para as empresas de transmissão de energia elétrica

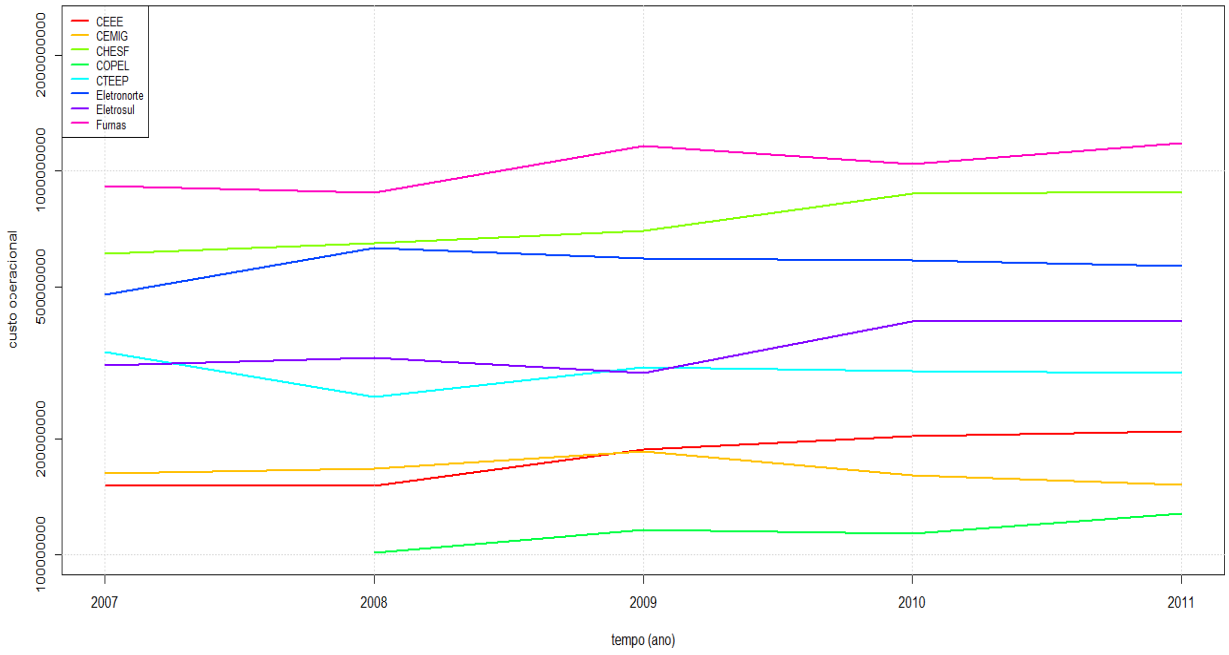

Figura 2 - Séries temporais do custo operacional para as transmissoras de energia

No gráfico acima se observa o comportamento do custo operacional de cada empresa transmissora de energia elétrica no período de 2007 a 2011.

4.2.3 Análise da correlação pelo método de Spearman

4.2.3.1 Variável custo operacional *versus* variáveis independentes

Na tabela apresentada a seguir é verificado o coeficiente de correlação de *Spearman* entre a variável dependente custo operacional e as variáveis independentes ou explicativas: comprimento de rede (km), módulos, quantidade de transformadores e potência (MVA).

TABELA 3
Correlação de *Spearman* entre a variável custo operacional e as variáveis independentes

	Custo Operacional	Potência (MVA)	Transformadores	Módulos	Rede Total
Custo Operacional	1	0,7365171	0,7241329	0,4291715	0,884735
Potência (MVA)	0,7365171	1	0,8196838	0,5700396	0,889171
Tranformadores	0,7241329	0,8196838	1	0,8354945	0,880945
Módulos	0,4291715	0,5700396	0,8354945	1	0,705615
Rede Total	0,8847349	0,8891707	0,8809452	0,7056145	1

Fonte: Elaborada pela autora

Analisando-se os resultados verifica-se que, com exceção da variável módulos, todas as outras variáveis estão correlacionadas fortemente e positivamente com a variável custo operacional, destacando-se a variável rede total, que possui a correlação mais forte (ρ = 0,8847349) com a variável dependente. A variável módulos está correlacionada fortemente e positivamente com a variável quantidade de transformadores (ρ = 0,8354945), o que pode gerar problemas de multicolinearidade no modelo de regressão.

4.2.3.2 Variável custo operacional *versus* variáveis de comprimento de rede (km)

Em decorrência da forte correlação existente entre as variáveis custo operacional e rede total, foi calculada a correlação da variável custo operacional com as variáveis de rede segregadas por grupo de tensão: 765 a 600 kV, 525 a 440 kV, 345 kV, 230 kV, 138 kV e 88 a 69 kV. Na tabela abaixo são apresentados os resultados:

TABELA 4
Correlação de *Spearman* entre as variáveis custo operacional e rede por nível de tensão

	custo	rede.765.600	rede.525.440	rede.345	rede.230	rede.138	rede.88.69
custo	1	0,572432	0,684172	0,196873	0,546415	0,430940	0,163272
rede.765.600	0,572432	1	0,240098	0,663153	-0,102459	0,411070	-0,500305
rede.525.440	0,684172	0,240098	1	0,305612	0,147979	0,541870	0,528511
rede.345	0,196873	0,663153	0,305612	1	-0,672489	0,245226	-0,459386
rede.230	0,546415	-0,102459	0,147979	-0,672489	1	0,042736	0,412472
rede.138	0,430940	0,411070	0,541870	0,245226	0,042736	1	0,339009
rede.88.69	0,163272	-0,500305	0,528511	-0,459386	0,412472	0,339009	1

Fonte: Elaborada pela autora

Pelos resultados obtidos constata-se que a variável rede 525 a 440 kV é a que apresenta correlação mais forte com a variável custo operacional, apresentando um coeficiente de correlação de ρ = 0,684172. Já as variáveis: rede 138 kV, rede 345 kV e rede 88 a 69 kV possuem fraca correlação com a variável custo operacional, sendo a última a variável com menor correlação (ρ = 0,163272).

Na sequência, foi avaliada a correlação de *Spearman* entre a variável custo operacional e variáveis formadas pela soma de diferentes grupos de tensão. Vale ressaltar que, para cada novo grupo foi adicionado na combinação o nível de tensão vizinho. Como exemplo das combinações realizadas tem-se:

Rede 765 a 600 kV + rede 525 a 440 kV = rede 765 a 440 kV;
Rede 765 a 600 kV + rede 525 a 440 kV + rede 345 kV = rede 765 a 345 kV;
Rede 765 a 600 kV + ... + rede 88 a 69 kV = rede 765 a 69 kV;
Rede 525 a 440 kV + rede 345 kV = rede 525 a 345 kV.

E assim sucessivamente, até serem obtidas todas as combinações possíveis para os níveis de tensão disponíveis na base de dados.

A seguir é apresentada a tabela contendo a correlação entre a variável custo operacional e a nova segregação da variável comprimento de rede:

TABELA 5

Correlação de Spearman entre a variável custo e as variáveis de rede

Grupo de tensão	ρ
rede.765.440	0,804399
rede.765.345	0,651782
rede.765.230	0,949076
rede.765.138	0,884735
rede.765.69	0,884735
rede.525.345	0,651782
rede.525.230	0,927186
rede.525.138	0,765162
rede.525.69	0,765162
rede.345.230	0,855870
rede.345.138	0,749963
rede.345.69	0,749963
rede.230.138	0,459208
rede.230.69	0,459208
rede.138.69	0,432643

Fonte: Elaborada pela autora

Os resultados mostram que a soma das redes para os grupos de tensão: 765 a 230 kV e 525 a 230 kV apresentou forte correlação com a variável dependente custo operacional, ρ = 0,949076 e ρ = 0,927186, respectivamente, superiores à correlação da variável rede total (ρ = 0,8847349), o que indica que tais grupos são variáveis fortemente associadas à definição do custo operacional das concessionárias brasileiras de energia elétrica.

4.2.4 Ajuste do Modelo de Regressão Linear Múltipla

Reconhecidas as variáveis independentes fortemente correlacionadas individualmente com a variável dependente custo operacional, foram ajustados modelos de regressão múltipla, a fim de identificar o modelo que melhor define o custo operacional das concessionárias de energia elétrica.

No que tange a variável comprimento de rede, para fins de comparação, foram ajustados modelos considerando os grupos de tensão utilizados no modelo DEA e modelos considerando separadamente os grupos de tensão que apresentaram forte correlação com a variável dependente, conforme exposto na seção anterior.

4.2.4.1 **Modelos de regressão linear múltipla ajustados considerando as variáveis do modelo DEA**

O modelo DEA, do ponto de vista estatístico, pode ser caracterizado como uma metodologia não paramétrica que a partir de variáveis independentes ou produto, gera uma fronteira de eficiência, a qual consiste no limite inferior de uma variável dependente ou insumo. De acordo com a Nota Técnica nº 383/2012-SRE/ANEEL, no modelo DEA a variável dependente ou insumo é o custo operacional e as variáveis independentes ou produto são: comprimento de rede (Km), número de módulos, quantidade de transformadores e potência entregue (MVA). Cumpre ressaltar que a variável comprimento de rede é apresentada estratificada nos níveis de tensão: rede 765 a 600 kV, rede 525 a 440 kV, rede 345 kV, rede 230 kV, rede 138 kV e rede 88 a 69 kV.

Como uma metodologia alternativa, para a definição do custo operacional, foram ajustados modelos de regressão linear múltipla contemplando as mesmas variáveis utilizadas no modelo DEA. Para tal, foi utilizado o pacote estatístico R, sendo estimados os modelos apresentados a seguir.

4.2.2.1.1 Modelo 1

Custo Operacional = 137,590 + 80,435 Rede Total (km) – 4,245 Potência (MVA) – 0,017 Quantidade de transformadores – 459,781 Módulos

Na tabela abaixo, são apresentados os resultados do modelo de regressão múltipla ajustado:

TABELA 6
Coeficientes e Estatística de Regressão

	Estimate	Std. Error	t value	Pr(>|t|)	FIV
(Intercept)	137,590	39,650	3,470	0,001434 **	
Rede.Total	80,435	9,550	8,422	7,8e-10 ***	11,29
Potencia.MVA	-4,245	2,874	-1,477	0,148844	13,65
Qtd.TR	-0,017	0,529	-0,032	0,974711	20,58
Modulos	-459,781	106,448	-4,319	0,000129 ***	12,48
Signif. Codes	0 '***'	0,001 '**'	0,01 '*'	0,05 '.' 0,1''	1
Residual standard error: 120,2 on 34 degrees of freedom					
Multiple R-squared: 0,8699 Adjusted R-squared: 0,8546					
F-statistic: 56,85 on 4 and 34 DF, p-value: 1,377e-14					

Fonte: Elaborada pela autora

O teste t dos coeficientes, considerando um nível de significância de 5% mostra que são significativas apenas as variáveis rede total e módulos, visto que foram as únicas que apresentaram no teste um p-valor menor ou igual a 0,05.

O resultado apresentado (p-valor menor ou igual a 0,05) no teste F para a significância do modelo de regressão indica que pelo menos uma variável independente é significativa.

O modelo explica 85,46% (R^2 Ajustado) da variabilidade dos dados em relação ao modelo da média.

Os resultados elevados (superiores a 10) do Fator de Inflação da Variância (FIV) indicam a existência de multicolinearidade no modelo de regressão.

Na análise descritiva dos resíduos observa-se que 50% dos valores situam-se acima de 2,89 e 50% abaixo. O menor valor de resíduo encontrado é -222,16, sendo que 25% estão dispostos abaixo de –83,53 e 25% acima de 72,91 e o valor máximo é igual a 226,59.

TABELA 7
Análise residual

Estatísticas Descritivas					Teste de normalidade	
Mín.	**1° quartil**	**Mediana**	**3° quartil**	**Máx.**	**W**	**p-valor**
-222,16	-83,53	2,89	72,91	226,59	0,9820	0,7753

Fonte: Elaborada pela autora

O teste de normalidade aplicado (*Shapiro-Wilk*) indica que os resíduos apresentam distribuição normal, o que se pode verificar mediante o valor da estatística p-valor, superior a 0,05.

A seguir é apresentado o gráfico de probabilidade normal dos resíduos do modelo de regressão ajustado.

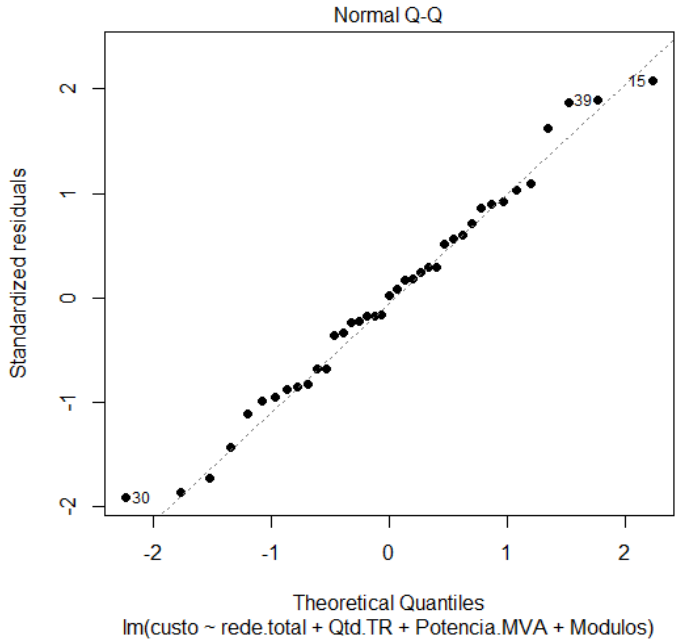

Figura 3 – Gráfico de probabilidade normal dos resíduos

No gráfico dos resíduos *versus* valores ajustados (Figura 3) tem-se a evidência visual da forma de um “funil”, o que indica que os resíduos são heterocedásticos, ou seja, os pontos apresentam variâncias desiguais e não estão distribuídos aleatoriamente em torno do eixo y=0 (Residuals=0).

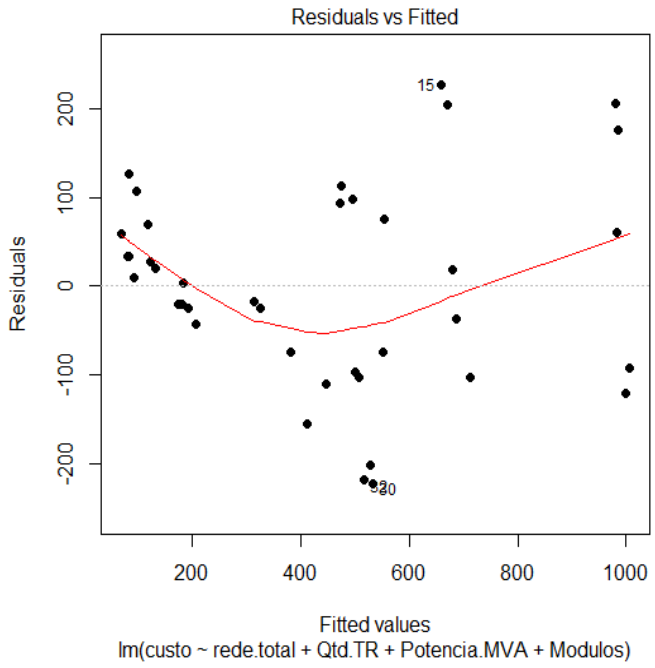

Figura 4 – Resíduos *versus* valores ajustados

4.2.2.1.2 Modelo 2

Custo Operacional = -44,735 + 3,464 Potência (MVA) + 2,993 Quantidade de transformadores – 450,274 Módulos – 29,468 rede 765.600 + 4,319 rede 525.440 + 7,274 rede 345 + 33,393 rede 230 – 64,103 rede 138 – 15,920 rede 88.69.

Na tabela a seguir, são apresentados os resultados do modelo de regressão múltipla ajustado.

TABELA 8
Coeficientes e Estatística de Regressão

	Estimate	Std. Error	t value	Pr(>|t|)	FIV
(Intercept)	-44,735	120,133	-0,372	0,7123	
Potencia.MVA	3,464	19,620	0,177	0,8611	1646,84
Qtd.TR	2,993	1,304	2,296	0,0291 *	323,13
Modulos	-450,274	44,791	-1,022	0,3155	553,96
rede.765.600	-29,468	262,034	-0,112	0,9112	1882,03
rede.525.440	4,319	112,420	0,038	0,9696	384,55
rede.345	7,274	145,428	0,050	0,9604	613,10
rede.230	33,393	49,650	0,673	0,5066	256,05
rede.138	-64,103	34,271	-1,870	0,0715 .	68,84
rede.88.69	-15,920	805,161	-0,020	0,9844	542,23
Signif. Codes	0 '***'	0,001 '**'	0,01 '*'	0,05 '.' 0,1' '	1
Residual standard error: 74,68 on 29 degrees of freedom					
Multiple R-squared: 0,9571 Adjusted R-squared: 0,9438					
F-statistic: 71,97 on 9 and 29 DF, p-value: < 2,2e-16					

Fonte: Elaborada pela autora

O teste t dos coeficientes considerando um nível de significância de 5% mostra que são significativas apenas as variáveis módulos e rede 138 kV, visto que foram as únicas que apresentaram no teste um p-valor menor ou igual a 0,05.

O resultado apresentado (p-valor menor ou igual a 0,05) no teste F para a significância do modelo de regressão indica que pelo menos uma variável independente é significativa.

O modelo explica 94,38% (R^2 Ajustado) da variabilidade dos dados em relação ao modelo da média.

Os resultados elevados (superiores a 10) do Fator de Inflação da Variância (FIV) indicam a existência de multicolinearidade no modelo de regressão.

Na análise descritiva dos resíduos observa-se que 50% dos valores situam-se acima de -3,25 e 50% abaixo. O menor valor de resíduo encontrado é -153,74, sendo que 25% estão dispostos abaixo de -24,59 e 25% acima de 26,58 e o valor máximo é igual a 140,73.

TABELA 9
Análise residual

Estatísticas Descritivas					Teste de normalidade	
Mín.	**1° quartil**	**Mediana**	**3° quartil**	**Máx.**	**W**	**p-valor**
-153,736	-24,586	-3,252	26,581	140,729	0,9626	0,2181

Fonte: Elaborada pela autora

O teste de normalidade aplicado (*Shapiro-Wilk*) indica que os resíduos apresentam distribuição normal, o que se pode verificar mediante o valor da estatística p-valor, superior a 0,05.

A seguir é apresentado o gráfico de probabilidade normal dos resíduos do modelo de regressão ajustado.

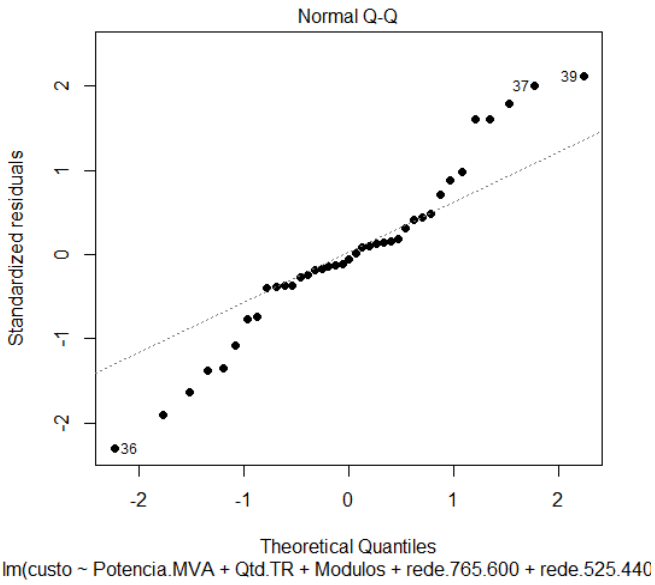

Figura 5 – Gráfico de probabilidade normal dos resíduos

No gráfico dos resíduos *versus* valores ajustados (Figura 5) tem-se a evidência visual da forma de um “funil”, o que indica que os resíduos são heterocedásticos, ou seja, os pontos apresentam variâncias desiguais e não estão distribuídos aleatoriamente em torno do eixo y=0 (Residuals=0).

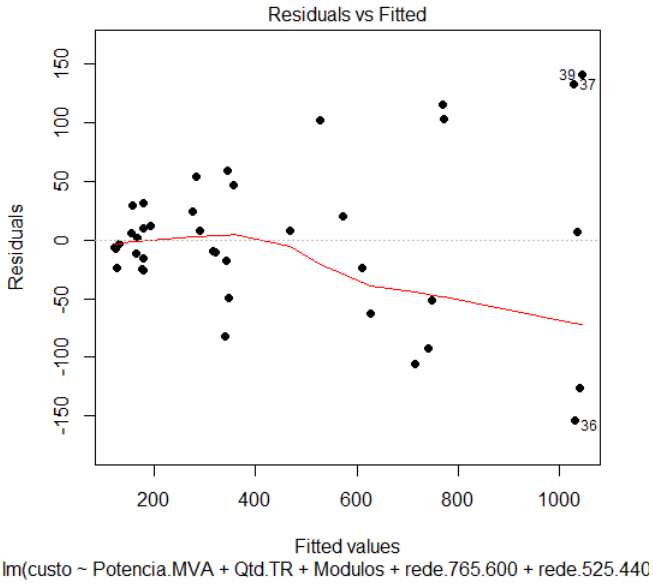

Figura 6 – Resíduos *versus* valores ajustados

4.2.4.2 Modelos Propostos

Embora os modelos ajustados na seção anterior expliquem 85,46% (Modelo 1) e 94,38% (Modelo 2) da variabilidade dos dados, foi verificado no teste t para os coeficientes, que nem todas as variáveis independentes são significativas, bem como foi constatada a existência de multicolinearidade nos modelos e heterocedasticidade na análise dos resíduos. A fim de estabilizar a variância foi aplicada transformação na variável resposta, porém algumas variáveis explicativas permaneceram não significativas nos modelos.

Por tal motivo, foram ajustados outros modelos de regressão, desta vez considerando a análise feita da correlação entre as variáveis explicativas e a variável dependente custo operacional.

Primeiramente, representando a variável comprimento de rede, foi considerado o grupo de tensão 765 a 230 kV no ajuste dos modelos, em virtude da forte correlação (ρ=0,949076) apresentada com a variável resposta. Em seguida, foram ajustados outros modelos, desta vez considerando o segundo grupo de tensão, 525 a 230 kV, fortemente correlacionado com a variável custo operacional (ρ = 0,927186).

Comparados os resultados, foi verificado que aplicando-se o método de regressão linear múltipla, as variáveis explicativas que melhor definem o custo operacional das empresas de transmissão de energia elétrica, compõem os dois modelos estimados apresentados na sequência.

4.2.2.2.1 Modelo 3

Log (Custo Operacional) = 4,786 + 0,149 rede 230.765 + 0,007 Quantidade de transformadores – 0,031 Potência (MVA) – 1,319 Módulos.

Na tabela a seguir, são apresentados os resultados do modelo de regressão múltipla ajustado:

TABELA 10
Coeficientes e Estatística de Regressão

	Estimate	Std. Error	t value	Pr(>|t|)	FIV
(Intercept)	4,7855643	0,0590521	81,040	< 2e-16 ***	
rede.230a765	0,1491290	0,0082712	18,030	< 2e-16 ***	3,10
Qtd.TR	0,0071368	0,0008178	8,727	3,38e-10 ***	24,88
Potencia.MVA	-0,3052750	0,0039938	-7,644	6,96e-09 ***	13,35
Modulos	-1,3193174	0,1430799	-9,221	8,93e-11 ***	11,42
Signif. Codes	0 '***'	0,001 '**'	0,01 '*'	0,05'.'	0,1 ' ' 1
Residual standard error: 0,1689 on 34 degrees of freedom					
Multiple R-squared: 0,9525 Adjusted R-squared: 0,9469					
F-statistic: 170,6 on 4 and 34 DF, p-value: < 2,2e-16					

Fonte: Elaborada pela autora

Pelo teste t dos coeficientes é verificado que todas as variáveis independentes são significativas, visto que a hipótese nula foi rejeitada ($H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$) ao nível de significância de 5%.

O resultado apresentado (p-valor menor ou igual a 0,05) no teste F para a significância do modelo de regressão indica que pelo menos uma variável independente é significativa.

O modelo explica 94,69% (R^2 Ajustado) da variabilidade dos dados em relação ao modelo da média.

O valor obtido para o FIV da variável independente quantidade de transformadores (FIV= 24,88) indica a existência de multicolinearidade no modelo, porém, no teste t para os coeficientes individuais a hipótese de nulidade foi rejeitada para todos os coeficientes, bem como no teste F a hipótese nula foi rejeitada, indicando que ao menos uma dos coeficientes das variáveis independentes é estatisticamente diferente de zero. Por tal motivo, optou-se por manter no modelo de regressão todas as variáveis independentes.

Na análise descritiva dos resíduos observa-se que 50% dos valores situam-se acima de -0,11 e 50% abaixo. O menor valor de resíduo encontrado é -0,26, sendo que 25% estão dispostos abaixo de - 0,11 e 25% acima de 0,10 e o valor máximo corresponde a 0,41.

TABELA 11
Análise residual

Mín.	1° quartil	Mediana	3° quartil	Máx.	W	p-valor
-0,26021	-0,11245	0,00054	0,10179	0,40988	0,9486	0,07384

Fonte: Elaborada pela autora

O teste de normalidade aplicado (*Shapiro-Wilk*) indica que os resíduos apresentam distribuição normal, o que se pode verificar mediante o valor da estatística p-valor, superior a 0,05.

A seguir é apresentado o gráfico de probabilidade normal dos resíduos do modelo de regressão ajustado.

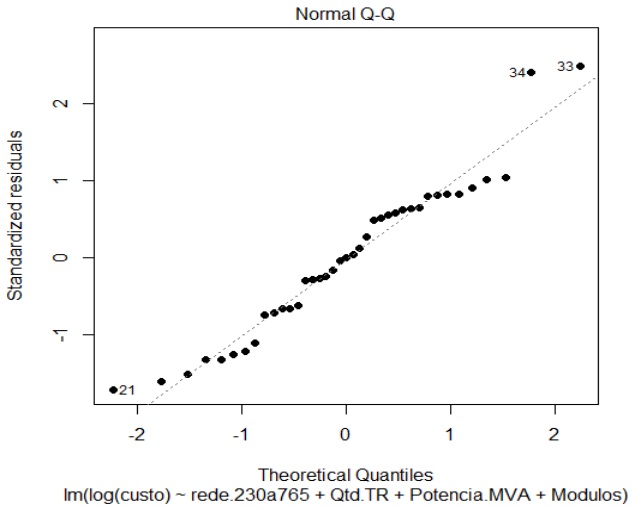

Figura 7 – Gráfico de probabilidade normal dos resíduos

No gráfico dos resíduos *versus* valores ajustados observa-se que os resíduos se encontram distribuídos aleatoriamente em torno do eixo y=0, indicando que são homocedásticos, ou seja, apresentam variância constante.

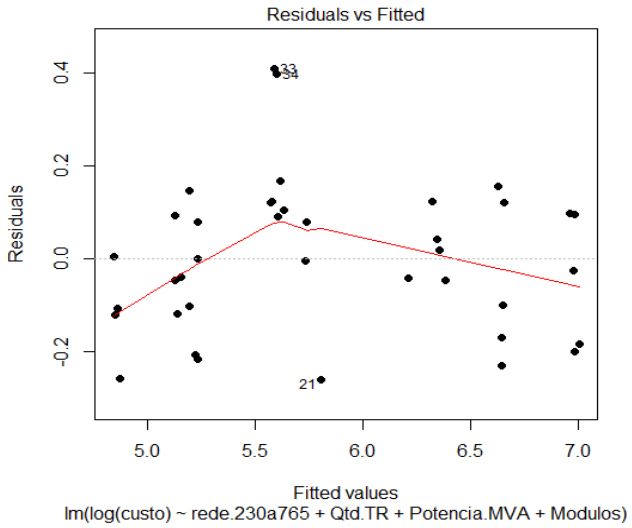

Figura 8 – Resíduos *versus* valores ajustados

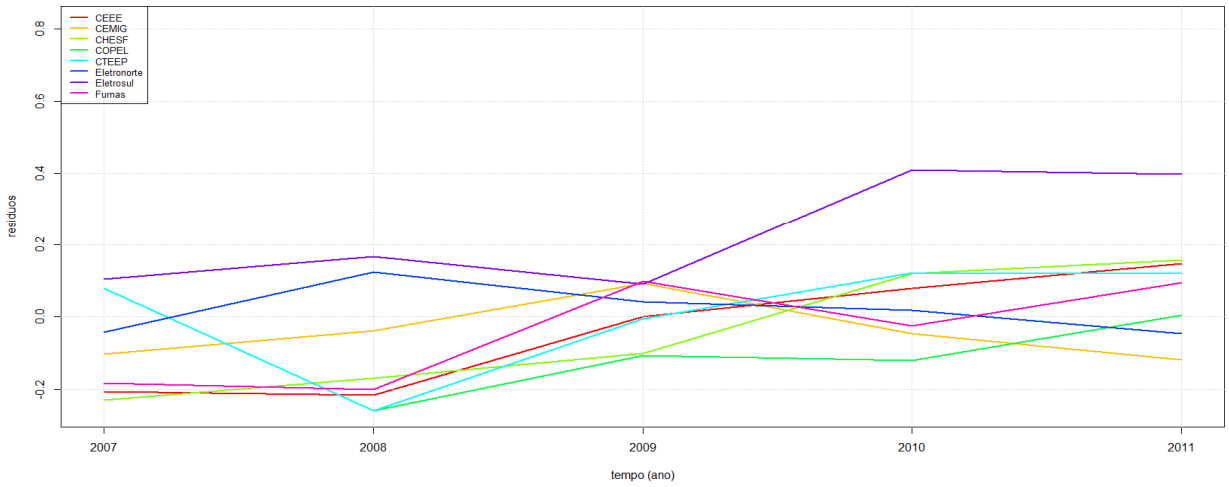

Figura 9 - Gráfico dos resíduos ao longo do tempo para as empresas de transmissão de energia

4.2.2.2.2 Modelo 4

Log (Custo Operacional) = 4,734 + 0,151 rede 230.525 + 0,007 Quantidade de transformadores – 0,017 Potência (MVA) – 1,489 Módulos.

Na tabela abaixo, são apresentados os resultados do modelo de regressão múltipla ajustado.

TABELA 12
Coeficientes e Estatística de Regressão

	Estimate	Std. Error	t value	Pr(>|t|)	FIV
(Intercept)	4,7335632	0,0557075	84,972	< 2e-16 ***	
rede.230a525	0,1513420	0,0077306	19,577	< 2e-16 ***	2,19
Qtd.TR	0,0068071	0,0007507	9,068	1,34e-10 ***	24,36
Potencia.MVA	-0,0170615	0,0032839	-5,196	9,58e-06 ***	10,49
Modulos	-1,4887331	0,1368593	-10,878	1,30e-12 ***	12,14
Signif. Codes	0 '***'	0,001 '**'	0,01 '*'	0,05'.' 0,1' '	1
Residual standard error: 0,1566 on 34 degrees of freedom					
Multiple R-squared: 0,9591 Adjusted R-squared: 0,9543					
F-statistic: 199,6 on 4 and 34 DF, p-value: < 2,2e-16					

Fonte: Elaborada pela autora

Pelo teste t dos coeficientes é verificado que todas as variáveis independentes são significativas, visto que a hipótese nula foi rejeitada ($H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$) ao nível de significância de 5%. O modelo explica 95,43% (R^2 Ajustado) da variabilidade dos dados em relação ao modelo da média.

O resultado apresentado (p-valor menor ou igual a 0,05) no teste F para a significância do modelo de regressão indica que pelo menos uma variável independente é significativa.

O valor obtido para o FIV da variável independente quantidade de transformadores (FIV= 24,36) indica a existência de multicolinearidade no modelo, porém no teste t para os coeficientes individuais, a hipótese de nulidade foi rejeitada para todos os coeficientes, bem como no teste F a hipótese nula foi rejeitada, indicando que ao menos uma dos coeficientes das variáveis independentes é estatisticamente diferente de zero. Por tal motivo, optou-se por manter no modelo de regressão todas as variáveis independentes.

Na análise descritiva dos resíduos observa-se que 50% dos valores situam-se acima de -0,02 e 50% abaixo. O menor valor de resíduo encontrado é -0,31, sendo que 25% estão dispostos abaixo de - 0,10 e 25% acima de 0,10 e o valor máximo corresponde a 0,31.

TABELA 13
Análise residual

Estatísticas Descritivas					Teste de normalidade	
Mín.	**1° quartil**	**Mediana**	**3° quartil**	**Máx.**	**W**	**p-valor**
-0,30941	-0,10538	-0,01966	0,10187	0,31567	0,9841	0,8468

Fonte: Elaborada pela autora

O teste de normalidade aplicado (*Shapiro-Wilk*) indica que os resíduos apresentam distribuição normal, o que se pode verificar mediante o valor da estatística p-valor, superior a 0,05.

A seguir é apresentado o gráfico de probabilidade normal dos resíduos do modelo de regressão ajustado.

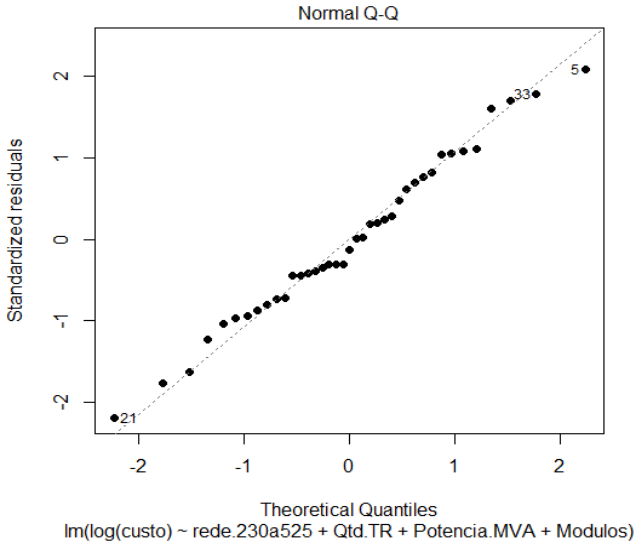

Figura 10 – Gráfico de probabilidade normal dos resíduos

No gráfico dos resíduos *versus* valores ajustados observa-se que os resíduos se encontram distribuídos aleatoriamente em torno do eixo y=0, indicando que são homocedásticos, ou seja, apresentam variância constante.

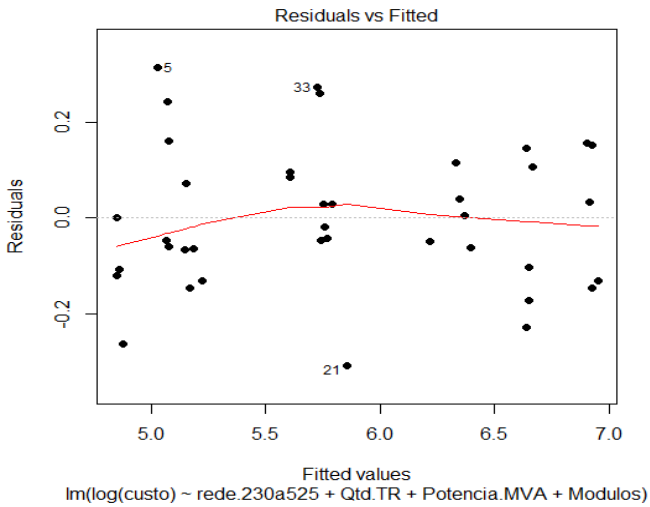

Figura 11 – Resíduos *versus* valores ajustados

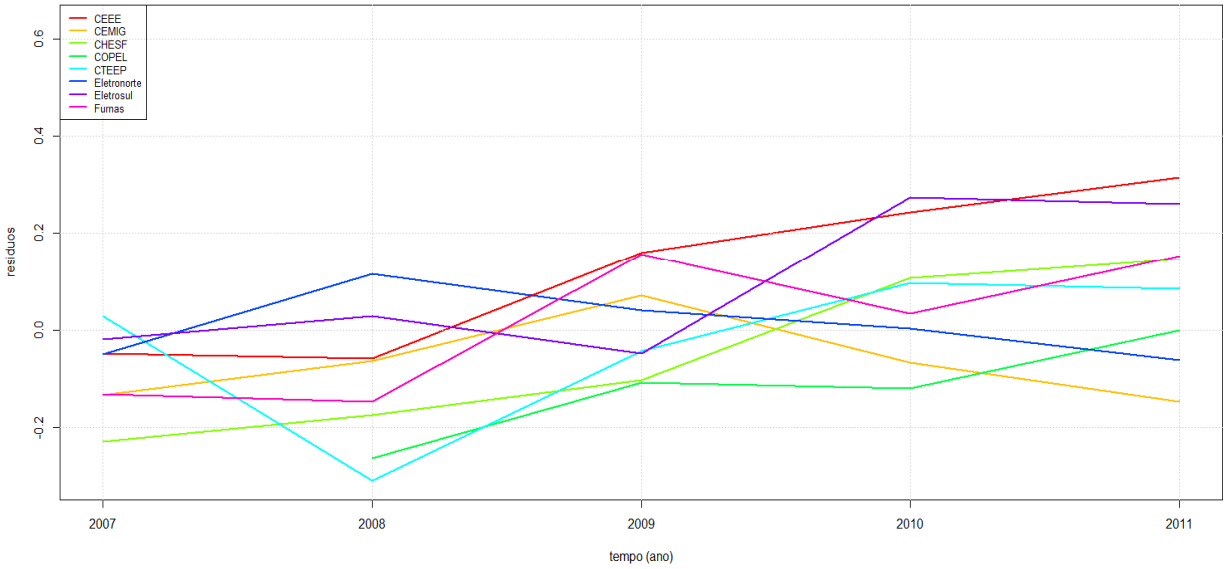

Figura 12. Gráficos dos resíduos ao longo do tempo para as empresas de transmissão de energia

5 CONCLUSÃO

A metodologia aplicada no presente trabalho foi eficaz no alcance do objetivo de identificação das variáveis independentes fortemente associadas à definição do custo operacional das concessionárias brasileiras de energia elétrica.

A aplicação da técnica de regressão linear múltipla possibilitou a estimativa de dois modelos capazes de definir o custo operacional das transmissoras de energia elétrica, a partir das variáveis independentes contempladas no modelo DEA, conforme Nota Técnica nº 383/2012 – SER/ANEEL.

Na análise da correlação foi constatada que a variável independente rede total é a mais fortemente correlacionada com a variável dependente custo operacional, sendo que na Nota Técnica supracitada a variável de rede é também apresentada segregada por grupos de tensão. No entanto, tais grupos são combinações de redes de baixa tensão com outras redes também de baixa tensão, redes de média tensão com outras redes de média tensão e redes de alta tensão com outras redes também de alta tensão, ou seja, são grupos de baixa, média e alta tensão.

Analisando-se as variáveis de rede, foram feitas combinações diferentes das compostas originalmente na Nota Técnica, contemplando em um mesmo grupo, redes de baixa, média e alta tensão, e, analisada a correlação dos novos grupos de tensão com a variável custo operacional, foram constatados dois grupos fortemente correlacionados com a variável resposta custo operacional, sendo tal correlação superior às apresentadas para os grupos originais de tensão, apresentados na Nota Técnica nº 383/2012 – SER/ANEEL.

A partir dos novos grupos de tensão foram obtidos através de ajustes dois modelos de regressão capazes de explicar 94,69% (modelo 3) e 95,43% (modelo 4) do custo operacional.

Sendo assim, pelos resultados obtidos e como a pesquisa possui caráter investigativo, conclui-se que o método de regressão linear múltipla é aplicável na identificação de variáveis relevantes na definição dos custos operacionais.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA. Nota Técnica nº 383/2012 – SER/ANEEL. MME. ANEEL, 2012.

PIRIE, W. **Spearman rank correlation coefficient**. In Kots S, Johnson NL, Read CB (Eds) Encyclopedia of statistical sciences.New York. Wiley, 1988. p 584-587. 8 v.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica.** 5. ed. São Paulo: Atlas, 2003. 311p.

MINGOTI, Sueli Aparecida. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: Ed. UFMG, 2005. 295p. (Didática ;8) ISBN 857041451X

MONTGOMERY, Douglas C.; RUNGER, George C. **Estatística aplicada e probabilidade para engenheiros.** 4. ed. Rio de Janeiro: LTC - Livros Técnicos e Científicos, c2009. xvi, 493 p.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS. Pró-Reitoria de Graduação. Sistema de Bibliotecas. **Padrão PUC Minas de normalização:** normas da ABNT para apresentação de trabalhos científicos, teses, dissertações e monografias. Belo Horizonte, 2010. 52p. Disponível em < http://www.pucminas.br/documentos/normalizacao_projetos.pdf>. Acesso em: 15 ago. 2011.

SHARPE, Norean R.; DE VEAUX, Richard D.; VELLEMAN, Paul F. **Estatística Aplicada:** Administração, Economia e Negócios. Tradução Lori Viali, Dr. Porto Alegre: Bookman, 2011. 871 p.

VERGARA, Sylvia Constant. **Projetos e relatórios de pesquisa em administração.** 4. ed. São Paulo: Atlas, 2003. 96p.

ANEXOS

BASE DE DADOS

NOTA TÉCNICA Nº 383/2012 – SER/ANEEL, DE 24/10/2012.

empresa	ano	custo	Potencia.MVA	Qtd.TR	Modulos	rede.765-600	rede.525-440	rede.345	rede.230	rede.138	rede.88-69	rede.total
CEEE	2007	151157279,99	7160	167	935	0	0	0	4687	760	227	5674
CEEE	2008	151267020,10	7592	173	951	0	0	0	4705	760	227	5692
CEEE	2009	188029163,08	7717	176	971	0	0	0	4768	760	227	5755
CEEE	2010	203542426,10	8268	186	1013	0	0	0	4772	760	227	5759
CEEE	2011	209292493,03	8268	186	1043	0	0	0	4772	760	227	5759
CEMIG	2007	162815754,82	15425	125	562	0	2169	1967	757	0	0	4893
CEMIG	2008	167259060,94	15425	125	591	0	2169	1967	768	0	0	4904
CEMIG	2009	185489048,75	15425	125	613	0	2169	1967	768	0	0	4904
CEMIG	2010	160853856,88	15425	125	615	0	2169	1967	768	0	0	4904
CEMIG	2011	151997424,83	16375	133	625	0	2169	1967	768	0	0	4904
CHESF	2007	608845846,00	31902	347	1801	0	5120	0	13185	572	316	19193
CHESF	2008	648816210,42	33185	361	1845	0	5120	0	13185	572	316	19193
CHESF	2009	698012424,23	33641	367	1878	0	5120	0	13308	572	316	19316
CHESF	2010	874906955,72	34952	376	1906	0	5120	0	13431	572	325	19448
CHESF	2011	884022468,19	35292	379	1939	0	5120	0	13472	572	325	19489
COPEL	2008	100991372,70	9962	80	324	0	157	0	1511	138	13	1819
COPEL	2009	116121430,35	10112	83	352	0	157	0	1566	138	13	1874
COPEL	2010	113612951,87	10112	83	362	0	157	0	1596	138	13	1904
COPEL	2011	127772954,52	10862	88	378	0	157	0	1596	138	13	1904
CTEEP	2007	336779500,29	46442	551	2125	0	6303	715	1323	9128	1085	18554
CTEEP	2008	256523749,97	47901	579	2192	0	6303	715	1323	9168	1085	18594
CTEEP	2009	307278468,03	48044	582	2266	0	6303	715	1364	9169	1085	18636
CTEEP	2010	299521609,59	49115	590	2401	0	6303	715	1364	9285	1099	18766
CTEEP	2011	296774468,35	50466	601	2435	0	6305	715	1364	9423	1099	18906
Eletronorte	2007	476618240,28	13720	180	495	0	3230	0	4887	366	248	8731
Eletronorte	2008	629078665,79	14783	201	535	0	3230	0	5201	366	248	9045
Eletronorte	2009	593092687,18	16301	231	644	0	3230	0	5201	366	248	9045
Eletronorte	2010	586353955,76	18039	246	676	0	3230	0	5201	366	248	9045
Eletronorte	2011	565010021,54	18539	253	693	0	3230	0	5295	366	248	9139
Eletrosul	2007	310752080,89	23122	165	574	0	2957	0	4648	1822	24	9451
Eletrosul	2008	324949737,14	25290	176	650	0	2962	0	5113	1828	24	9927
Eletrosul	2009	297116179,43	25723	180	674	0	2962	0	5135	1828	24	9949
Eletrosul	2010	403126384,96	25823	182	693	0	2962	0	5135	1828	24	9949
Eletrosul	2011	402804356,54	26189	187	703	0	2962	0	5135	1828	24	9949
Furnas	2007	913513906,96	86288	461	909	5922	4572	6121	1883	2148	0	20646
Furnas	2008	878362600,09	86288	461	926	5922	4572	6121	1883	2148	0	20646
Furnas	2009	1161128251,63	86438	464	955	5922	4572	6131	1883	2149	0	20657
Furnas	2010	1043720973,01	86723	469	973	5922	4572	6221	1883	2149	0	20747
Furnas	2011	1185290053,84	86948	472	978	5922	4572	6221	1883	2149	0	20747

COMANDOS UTILIZADOS NO SOFTWARE ESTATÍSTICO R PARA LEITURA DA BASE DE DADOS, ANÁLISE DA CORRELAÇÃO E APLICAÇÃO DO MÉTODO DE REGRESSÃO LINEAR MÚLTIPLA

Leitura da base de dados "transmissao.csv":

```
dados <- read.csv( file.choose() )
```

****

Tratamento da base de dados:

```
dados[,"custo"]        <- dados[,"custo"]/1e6
dados[,"rede.total"]   <- dados[,"rede.total"]/1000
dados[,"rede.765.600"] <- dados[,"rede.765.600"]/1000
dados[,"rede.525.440"] <- dados[,"rede.525.440"]/1000
dados[,"rede.345"]     <- dados[,"rede.345"]/1000
dados[,"rede.230"]     <- dados[,"rede.230"]/1000
dados[,"rede.138"]     <- dados[,"rede.138"]/1000
dados[,"rede.88.69"]   <- dados[,"rede.88.69"]/1000
dados[,"Potencia.MVA"] <- dados[,"Potencia.MVA"]/1000
dados[,"Modulos"]      <- dados[,"Modulos"]/1000
```

****

Análise da correlação de *Spearman*:

```
cor(dados[,c("custo", "Potencia.MVA", "Qtd.TR", "Modulos", "rede.total")], method = "spearman")

cor(dados[,c("custo", "rede.765.600", "rede.525.440", "rede.345", "rede.230", "rede.138", "rede.88.69")], method = "spearman")
names(dados)
cor(cbind( dados[,"custo"], rowSums(dados[,c(9:10)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(9:11)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(9:12)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(9:13)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(9:14)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(10:11)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(10:12)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(10:13)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(10:14)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(11:12)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(11:13)])), method = "spearman")
```

```
cor(cbind( dados[,"custo"], rowSums(dados[,c(11:14)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(12:13)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(12:14)])), method = "spearman")
cor(cbind( dados[,"custo"], rowSums(dados[,c(13:14)])), method = "spearman")
```

****

Criando uma nova coluna:

```
dados[,"rede.230a765"] <- rowSums(dados[,c(9:12)])
dados[,"rede.230a525"] <- rowSums(dados[,c(10:12)])
```

****

Gráficos de dispersão:

```
color <- rainbow(8)
dados$cor <- color[dados$empresa]
dados$tam <- 2.3*(dados$ano - 2006)/5 + 0.3

plot(custo ~ rede.total, data=dados, pch=19, cex=tam, col=color[empresa], bty="l",
xlim=c(0,max(rede.total)), log="y", xlab="rede (Km)", ylab= "custo (R$)") points(custo
~ rede.total, data=dados, pch=21, cex=tam, col="gray", lwd=0.1) abline(v=0, lty=2)
grid()
legend("topleft", legend=as.character(unique(dados$ano)), pch=21,
pt.cex=unique(dados$tam),bg="white")
d <- aggregate(cbind(custo, rede.total) ~ empresa, data = dados, mean)
text(d$rede.total, d$custo, as.character(d$empresa), cex=0.9, col=gray(0.2))
savePlot("figura.png", type="png")

plot(custo ~ Potencia.MVA, data=dados, pch=19, cex=tam, col=color[empresa], bty="l",
xlim=c(0,max(Potencia.MVA)), log="y", xlab="potencia (MVA)", ylab= "custo (R$)")
points(custo ~ Potencia.MVA, data=dados, pch=21, cex=tam, col="gray", lwd=0.1)
abline(v=0, lty=2)
grid()
legend("topleft", legend=as.character(unique(dados$ano)), pch=21,
pt.cex=unique(dados$tam),bg="white")
d <- aggregate(cbind(custo, Potencia.MVA) ~ empresa, data = dados,
mean)text(d$Potencia.MVA, d$custo, as.character(d$empresa), cex=0.9, col=gray(0.2))
savePlot("figura.png", type="png")

plot(custo ~ Qtd.TR, data=dados, pch=19, cex=tam, col=color[empresa], bty="l",
xlim=c(0,max(Qtd.TR)), log="y", xlab="Quantidade de Transformadores", ylab= "custo
(R$)") points(custo ~ Qtd.TR, data=dados, pch=21, cex=tam, col="gray", lwd=0.1)
abline(v=0, lty=2)
grid()
```

```
legend("topleft", legend=as.character(unique(dados$ano)), pch=21,
pt.cex=unique(dados$tam),bg="white")
d <- aggregate(cbind(custo, Qtd.TR) ~ empresa, data = dados, mean)
  text(d$Qtd.TR, d$custo, as.character(d$empresa), cex=0.9, col=gray(0.2))
  savePlot("figura.png", type="png")

plot(custo ~ Modulos, data=dados, pch=19, cex=tam,col=color[empresa], bty="l",
xlim=c(0,max(Modulos)), log="y",xlab="Quantidade de Módulos", ylab= "custo (R$)")
points(custo ~ Modulos, data=dados, pch=21, cex=tam, col="gray", lwd=0.1)
abline(v=0, lty=2)
grid()
legend("topleft", legend=as.character(unique(dados$ano)), pch=21,
pt.cex=unique(dados$tam),bg="white")
d <- aggregate(cbind(custo, Modulos) ~ empresa, data = dados, mean)
text(d$Modulos, d$custo, as.character(d$empresa), cex=0.9, col=gray(0.2))
savePlot("figura.png", type="png")

*******************************************************************************
```

Série temporal para cada empresa:

```
series <- matrix(c(dados$custo[1:15],NA,dados$custo[16:39]), nrow = 5, ncol = 8, byrow
= FALSE)
colnames(series) <-levels(dados$empresa)

plot(2007:2011, series[,1], type="l", lwd=2, col=color[1], ylab="custo operacional",
xlab="tempo (ano)", ylim=c(min(series,na.rm=TRUE),2*max(series,na.rm=TRUE)), log="y");
grid()
for(cont in 2:8)
lines(2007:2011, series[,cont], lwd=2, col=color[cont])
legend("topleft", legend=levels(dados$empresa), lty=1, lwd=2, col=color, bty="o",
cex=0.8 )
savePlot("figura.png", type="png")

*******************************************************************************
```

Ajuste do modelo de regressão:

```
modelo.linear <- lm(custo ~ rede.total + Potencia.MVA + Qtd.TR + Modulos, data=dados)
summary(modelo.linear)
plot(modelo.linear, pch=19, col="black")
```

```
modelo.linear <- lm(custo ~ Potencia.MVA + Qtd.TR + Modulos + rede.765.600 + rede.525.440 + rede.345 + rede.230 + rede.138 + rede.88.69, data=dados)
summary(modelo.linear)
plot(modelo.linear, pch=19, col="black")

modelo.linear <- lm(log(custo) ~ rede.230a765 + Qtd.TR + Potencia.MVA + Modulos, data=dados)
summary(modelo.linear)
plot(modelo.linear, pch=19, col="black")

modelo.linear <- lm(log(custo) ~ rede.230a525 + Qtd.TR + Potencia.MVA + Modulos, data=dados)
summary(modelo.linear)
plot(modelo.linear, pch=19, col="black")

********************************************************************************
```

ANOVA:

```
summary.aov(modelo.linear)

********************************************************************************
```

Teste de normalidade:

```
shapiro.test( residuals(modelo.linear))

********************************************************************************
```

Análise residual:

```
plot(modelo.linear, pch=19, col="black")

********************************************************************************
```

Série temporal dos resíduos para cada modelo de regressão proposto:

```
series <- matrix(c(res[1:15],NA,res[16:39]), nrow = 5, ncol = 8, byrow = FALSE)
colnames(series) <-levels(dados$empresa)

plot(2007:2011, series[,1], type="l", lwd=2, col=color[1], ylab="residuos", xlab="tempo (ano)", ylim=c(min(series,na.rm=TRUE),2*max(series,na.rm=TRUE)));

grid()
```

```
for(cont in 2:8) lines(2007:2011, series[,cont], lwd=2, col=color[cont])
legend("topleft",  legend=levels(dados$empresa),  lty=1,  lwd=2,  col=color,  bty="o",
cex=0.8 )
savePlot("figura.png", type="png")

*******************************************************************************
```

Escreve a base de dados em arquivo:

```
write.csv(dados, "MeusDados.csv")
*******************************************************************************
```