UNIVERSIDADE FEDERAL DE MINAS GERAIS

Programa de Pós-Graduação em Estatística

Gláucia Sibele de Paula da Silva Faria

TESTE DE HIPÓTESES UTILIZANDO A METODOLOGIA BOOTSTRAP NÃO PARAMÉTRICO PARA IDENTIFICAÇÃO DAS VARIÁVEIS SIGNIFICATIVAS PARA O TEMPO DE CHEGADA DE TRENS AO SEU DESTINO

Belo Horizonte

Dezembro / 2022

Gláucia Sibele de Paula da Silva Faria

TESTE DE HIPÓTESES UTILIZANDO A METODOLOGIA BOOTSTRAP NÃO PARAMÉTRICO PARA IDENTIFICAÇÃO DAS VARIÁVEIS SIGNIFICATIVAS PARA O TEMPO DE CHEGADA DE TRENS AO SEU DESTINO

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: Prof. Roberto da Costa Quinino

Belo Horizonte
Dezembro / 2022

Faria, Gláucia Sibele de Paula da Silva.

F224t Teste de hipóteses utilizando a metodologia bootstrap não paramétrico para identificação das variáveis significativas para o tempo de chegada de trens ao seu destino [manuscrito] / Gláucia Sibele de Paula da Silva Faria. — 2022.
45 f. il.

Orientador: Roberto da Costa Quinino.
Monografia (especialização) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: 40

1. Estatística. 2. Análise de regressão. 3. Variáveis de probabilidade. 4. Bootstrap (Estatística). I. Quinino, Roberto da Costa. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III.Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irenquer Vismeg Lucas Cruz CRB 6/1510 - Universidade Federal de Minas Gerais - ICEx

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

ATA DO 261 ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE GLÁUCIA SIBELE DE PAULA DA SILVA FARIA.

Aos quatorze dias do mês de dezembro de 2022, às 17:00 horas, com utilização de recursos de videoconferência a distância, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pela Comissão do Curso de Especialização em Estatística, para julgar a apresentação do trabalho de fim de curso da aluna **Gláucia Sibele de Paula da Silva Faria**, intitulado: "Teste de hipóteses utilizando a metodologia bootstrap não paramétrico para identificação das variáveis significativas para o tempo de chegada de trens ao seu destino", como requisito para obtenção do Grau de Especialista em Estatística. Abrindo a sessão, o Presidente da Comissão, Professor Roberto da Costa Quinino – Orientador, após dar conhecimento aos presentes do teor das normas regulamentares, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Após a defesa, os membros da banca examinadora reuniram-se sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação: a candidata foi considerada Aprovada condicional às modificações sugeridas pela banca examinadora no prazo de 30 dias a partir da data de hoje por unanimidade. O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 14 de dezembro de 2022.

Roberto da Costa Quinino:80871291720

Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2022.12.14 18:19:25 -03'00'

Prof. Roberto da Costa Quinino (Orientador)
Departamento de Estatística / ICEX / UFMG

DANILO GILBERTO DE OLIVEIRA VALADARES:06707566601

Danilo Gilberto de Oliveira Valadares
Departamento de Estatística / ICEX / UFMG

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

E-mail: pgest@ufmg.br
Tel: 3409-5923 – FAX: 3409-5924

DECLARAÇÃO DE CUMPRIMENTO DE REQUISITOS PARA CONCLUSÃO DO CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA.

Declaro para os devidos fins que Gláucia Sibele de Paula da Silva Faria, número de registro 2019705189, cumpriu todos os requisitos necessários para conclusão do curso de Especialização em Estatística e que me entregou a versão final corrigida. O trabalho foi apresentado no dia 14 de dezembro de 2022 com o título "*Teste de hipóteses utilizando a metodologia bootstrap não paramétrico para identificação das variáveis significativas para o tempo de chegada de trens ao seu destino*".

Belo Horizonte, 21 de dezembro de 2022

Roberto da Costa Quinino:80871291720
Assinado de forma digital por Roberto da Costa Quinino:80871291720
Dados: 2022.12.21 22:49:43 -03'00'

Prof. Roberto da Costa Quinino
Coordenador da Comissão do Curso de Especialização em Estatística

Prof. Roberto da Costa Quinino
Coordenador do curso de Especialização em Estatística
Departamento de Estatística / UFMG

AGRADECIMENTOS

Agradeço aos meus filhos pela compreensão devido às ausências nas brincadeiras.
Ao meu esposo pelo apoio e incentivo.

RESUMO

Conhecer o tempo gasto para um trem chegar ao seu destino proporciona benefícios, tendo em vista que permite que a operação planeje a movimentação dos trens, sendo o carregamento e descarregamento de produtos transportados para diferentes clientes. Para que isso seja possível, nos deparamos, contudo, com diversos fatores que podem influenciar neste tempo, sendo necessário a identificação das variáveis que são significativas para este problema. Diante disso, este projeto propõe, através da utilização da técnica de Bootstrap não Paramétrico e do modelo de regressão múltipla com interações, a identificação das variáveis significativas para a chegada do trem ao seu destino. Identificação que se mostrou, neste contexto, satisfatória, garantindo insumos para novos estudos.

Palavras-chave: Variáveis significativas, Bootstrap não Paramétrico, Modelo de regressão múltipla com interações

ABSTRACT

Knowing the time taken for a train to reach its destination provides benefits for the operation to plan the movement of trains, with the loading and unloading of products transported to different customers. For this to be possible, however, we are faced with several factors that can influence this time, making it necessary to identify the variables that are significant for this problem. Therefore, this project proposes, through the use of the non-parametric Bootstrap technique and the multiple regression model with interactions, the identification of the significant variables for the arrival of the train at its destination. Identification that proved to be, in this context, satisfactory, guaranteeing inputs for new studies.

Key Words: Significant variables, non-parametric Bootstrap, Multiple regression model with interactions

LISTA DE FIGURAS

LISTA DE TABELAS

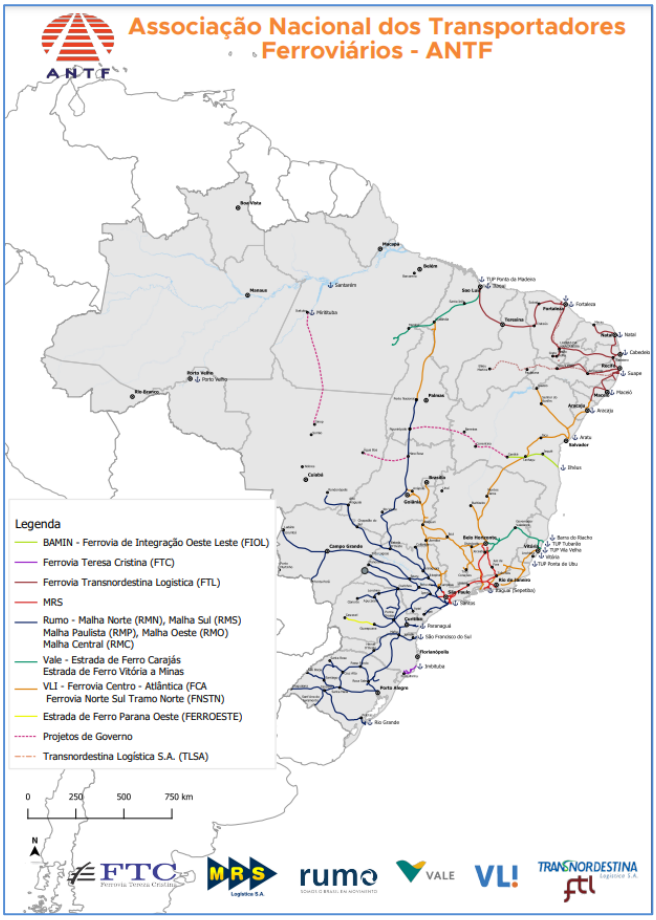
SUMÁRIO

1 INTRODUÇÃO

A operação ferroviária no Brasil tem como objetivo realizar o transporte de mercadorias de uma origem a um destino para um determinado cliente. A origem e o destino podem ser de fazendas a portos de acordo com as necessidades dos clientes. No Brasil ela conta com a ANTT (Agência Nacional de Transportes Terrestres) órgão responsável por acompanhar, fiscalizar e regulamentar as operações das empresas quer realização o transporte, no caso são as concessionárias.

A malha ferroviária hoje concedida pelo Governo Federal é de 29.320 km. As mercadorias são escoadas de vários setores da economia como agrícola, siderúrgico para os portos ou recebimento de produtos importados como fertilizantes. O transporte ferroviário é um dos transportes em que o custo é menor comparado a outros como exemplo, o transporte rodoviário, pois o frete é menor e transporta uma quantidade maior de carga. Abaixo temos o mapa ferroviário brasileiro e as empresas operadoras.

Figura 1: Mapa Ferroviário Brasileiro



Fonte: https://www.antf.org.br

De acordo com os dados recebidos das concessionárias pela ANTT, observou-se um crescimento de 30,1% no transporte de mercadorias na comparação de março de 2021 com o mesmo período do ano de 2020. No relatório de Produção da ANTF de janeiro de 2022, registrou-se um crescimento em setores de graneis agrícolas e combustíveis, (142,4%) e (9,1%) respectivamente. E devido a interrupção de diversos fluxos de transporte devido às chuvas, o volume transportado pelas concessionárias associadas á ANTF apresentaram uma queda de 3,8%.

O tempo de chegada dos trens em seu destino, potencializa a operação nos terminais e portos, pois é possível realizar diferentes ações que mitigam riscos, melhoram a organização dos pátios, a saída, distribuição e a circulação dos trens nos terminais e portos e a programação da partida, inclusive devido a quantidade de vagões e locomotivas que atendem ao que se deseja transportar.

A identificação das variáveis significativas, objeto desse estudo, permitirá posteriores trabalhos.

2 FUNDAMENTAÇÃO TEÓRICA

O presente trabalho pretende apresentar uma modelagem utilizando o método de Bootstrap não Paramétrico e regressão múltipla com interações técnicas auxiliares. Portanto, é importante estabelecer os conceitos relacionados com a implementação ou que auxiliam a compreensão dos resultados.

2.1 Regressão Linear Múltipla

Segundo (MONTGOMERY; PECK; VINING, 2012), uma das técnicas mais utilizadas para analisar dados com muitos fatores é a análise de regressão devido ao seu processo de utilizar uma equação para expressar as relações que existem entre uma variável de interesse (resposta) e um conjunto de variáveis preditoras (explicativas).

É uma técnica estatística empregada em uma grande variedade de aplicações e áreas como engenharia, ciências químicas e físicas entre outras. (MONTGOMERY; PECK; VINING, 2012).

O modelo de regressão múltipla pode ser descrito pela equação abaixo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

Onde:

Y = variável dependente ou resposta
$\beta_j, j = 0,1,2,3, \ldots . k$ = são os coeficientes de regressão
$x_i, i = 0,1,2,3, \ldots k$ = variáveis regressoras ou preditoras
ε = componente erro aleatório

2.2 Regressão Linear Múltipla com Interação

Um modelo de regressão múltipla também pode conter efeitos de interação, que de forma simples, pode ser descrito como o produto entre as variáveis do modelo (MONTGOMERY; RUNGER, 2016).

Considerando três variáveis independentes, a equação do modelo de regressão múltipla com interações pode ser descrita:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon \quad (2)$$

Onde temos que $x_1 x_2$ é o produto entre as variáveis presentes no modelo.

Entende-se que o efeito gerado pela variação da variável x_1 depende do nível de variação da variável x_2 e assim por seguinte. Para um maior número de variáveis independentes temos que considerar as interações de todas as ordens.

2.2.1 *Método dos Mínimos Quadrados*

Para se estimar os coeficientes de regressão, um dos métodos que pode ser utilizado é o dos mínimos quadrados (MMQ) (MONTGOMERY; RUNGER, 2016).

Esse método proposto pelo alemão Karl Gauss (1777-1855) consiste em minimizar a soma dos quadrados dos desvios das observações.

A função dos mínimos quadrados é dada pela equação

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_1 - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2 \quad (3)$$

O objetivo é minimizar L com relação aos coeficientes $\beta_0, \beta_{1,\ldots,} \beta_k$, de maneira que as estimativas de MMQ satisfaça

$$\frac{\partial L}{\partial \beta_j}_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = 2 \sum_{i=1}^{n} (y_1 - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) = 0 \quad (4)$$

2.2.2 *Abordagem Matricial*

De acordo com (MONTGOMERY; RUNGER, 2016) as operações matemáticas de um ajuste de modelo de regressão múltipla são mais convenientes de serem expressas utilizando notação matricial. Suponha que exista *k* variáveis regressoras e *n* observações e o modelo apresentado com as regressoras e suas respectivas respostas seja

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad \mathbf{i = 1, 2, \ldots, n} \tag{5}$$

O modelo acima é considerado um sistema de n equações, o qual pode ser expresso na notação matricial

$$y = X\beta + \epsilon \tag{6}$$

Sendo

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \; e \; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{7}$$

Segundo (MONTGOMERY; RUNGER, 2016), de maneira geral, y é um vetor (n x 1) das observações e X é uma matriz (n x p) dos níveis das variáveis independentes, β é um vetor (p x 1) dos coeficientes de regressão e $\in$ é um vetor (n x 1) dos erros aleatórios. A matriz X é frequentemente chamada de matriz modelo.

Tem-se como objetivo encontrar o vetor dos estimadores de mínimos quadrados, $\hat{\beta}$, que minimiza:

$$L = \sum_{i=0}^{n} \in_i^2 = \in' \in = (y - X\beta)'(y - X\beta) \tag{8}$$

O estimador $\hat{\beta}$ de mínimos quadrados é a solução para β nas equações

$$\frac{\partial \mathrm{L}}{\partial \beta} = 0 \tag{9}$$

As equações resultantes e que precisam ser resolvidas são equações normais de mínimos quadrados na forma matricial, conforme abaixo

$$X'X\hat{\beta} = X'y \tag{10}$$

Com o objetivo de resolver as equações normais, multiplique ambos os lados das equações $X'X\hat{\beta} = X'y$ pelo inverso de $X'X$.

Por conseguinte, a estimativa de mínimos quadrados de β é

$$\hat{\beta} = (X'X)^{-1}X'y \tag{11}$$

Note que existe $p = k + 1$ equações normais para $p = k + 1$ incógnitas. Ainda a matriz $X'X$ sempre será não singular, como foi considerado anteriormente, de maneira que, para inverter essas matrizes, os métodos apresentados sobre determinantes e matrizes, podem ser usados para encontrar $(X'X)^{-1}$. Normalmente a forma matricial das equações normais é idêntica à forma escalar. Escrevendo a equação (13) em detalhes, obtemos (15):

O modelo de regressão ajustado é dado por (12)

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij}, \quad i = 1,2,\dots,n \tag{12}$$

E na notação matricial é

$$\hat{y} = X\hat{\beta} \tag{13}$$

Têm se como a diferença entre a observação y_i e o valor ajustado $\hat{y}_i$ um resíduo, $e_i = y_i - \hat{y}_i$, cujo vetor (n x 1) é dado por (14)

$$e = y - \hat{y} \tag{14}$$

Notação matricial em detalhes da equação (13)

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \dots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x^2{}_{i1} & \sum_{i=1}^{n} x_{ik}x_{i1} & \dots & \sum_{i=1}^{n} x_{ik}x_{i1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \dots & \sum_{i=1}^{n} x^2{}_{ik} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix} \quad (15)$$

2.2.3 *Testes de Hipóteses* - Teste para a Significância da Regressão

Em problemas de regressão linear múltipla, alguns testes de hipóteses relativos aos parâmetros dos modelos são importantes de acordo com a adequação do modelo (MONTGOMERY; RUNGER, 2016).

Um teste para determinar se há uma relação linear entre a variável resposta y e um subconjunto de regressores x_1, x_2, ..., x_k é o teste para significância da regressão.

Têm-se como hipóteses coerentes

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: caso\ contrário \end{cases}$$

A rejeição de $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ sugere que uma ou mais variáveis regressoras $x_1, x_2, \dots, x_k$ contribui significativamente para o modelo considerando o nível e confiança adotado.

Esse teste é considerado uma generalização do procedimento utilizado na regressão linear múltipla. A soma total dos quadrados SQ_T é dividida em uma soma dos quadrados devido à regressão SQ_R e em uma soma dos quadrados devido ao erro SQ_E

$$SQ_T = SQ_R + SQ_E \quad (16)$$

Caso, $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ seja uma hipótese verdadeira, temos que a soma dos quadrados da regressão/desvio padrão será uma variável aleatória qui-quadrado, com k graus de liberdade.

A estatística de teste para H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ é dada pelo teste ANOVA

$$F_0 = \frac{\frac{SQ_R}{k}/k}{SQ_E/(n-p)} = \frac{MQ_R}{MQ_E} \tag{17}$$

Rejeita-se H_0 caso o valor calculado da estatística de teste seja maior que $f_{\alpha,k,n-p}$, obtido em uma tabela da distribuição F de Fisher.

Em sua grande maioria, o nível de significância do teste comumente adotado é de 5% representado por α. A **Tabela 1** ilustra o resultado da análise de variância (ANOVA).

Tabela 1: Análise de Variância para testar a significância da regressão

Fonte de Variação	***Soma dos Quadrados***	***Graus de Liberdade***	***Média Quadrática***	***F0***
Regressão	SQ_R	k	MQ_R	$\frac{MQ_R}{MQ_E}$
Erro ou resíduo	SQ_E	n - p	MQ_E	
Total	SQ_T	n - 1		

Fonte: MONTGOMERY e RUNGER (2016)

O coeficiente de determinação múltipla R^2 é utilizado como uma estatística para avaliar o ajuste do modelo,

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T} \tag{18}$$

O seu valor responde ao percentual da variabilidade na resposta do modelo, ou seja, se o valor de $R^2 = 0{,}62$ têm-se que 62% da variabilidade dos dados é explicado pelo modelo. Porém à medida que se acrescenta uma nova variável ao modelo, o valor de R^2 aumenta também, sendo uma problemática como uma medida da qualidade do ajuste do modelo.

A fim de evitar incoerência com o resultado de R^2, usualmente, utiliza-se o coeficiente $R^2\ ajustado$.

$$R^2\text{ajustado} = 1 - \frac{SQ_E/(n-p)}{SQ_T/(n-p)} = 1 \tag{19}$$

$R^2\ ajustado$ aumentará somente se a variável acrescentada ao modelo reduzir a média quadrática do erro, uma vez que $SQ_E/(n-p)$ é uma média quadrática do erro ou do resíduo e $SQ_T/(n-1)$ é uma constante (MONTGOMERY; RUNGER, 2016).

2.2.4 *Testes para os Coeficientes Individuais de Regressão*

Os testes para os coeficientes individuais são utilizados para verificar se cada uma das variáveis presentes no modelo de regressão são significativas, determinando se incluindo ou retirando uma variável regressora o modelo pode se tornar mais robusto.

A hipótese de que um coeficiente individual de regressão β_j é igual a zero é dada por

$$H_0: \beta_j = 0 \text{ versus } H_1: \beta_j \neq 0$$

Caso a hipótese H_0: $\beta_j = 0$ não seja rejeitada, temos que o regressor x_j é um candidato a ser retirado do modelo, ou seja, ele não é significativo ao nível de confiança considerado. A inclusão de variáveis ao modelo de regressão múltipla contribui para o aumento da soma dos quadrados da regressão e diminui a soma dos quadrados dos erros. Sempre avaliar se realmente a inclusão de novas variáveis contribui para um modelo melhor ajustado. O teste de hipóteses é realizado considerando a equação (20) com $\beta_j = 0$.

2.2.5 *Intervalos de Confiança*

Segundo (MONTGOMERY; RUNGER, 2016), a construção de estimativas de intervalos de confiança para modelos de regressão múltipla se faz útil, pois é possível desenvolver procedimentos capazes de obter os intervalos de confiança dado que os erros $\{\varepsilon_i\}$ sejam normais e independentemente distribuídos, com média zero e variância σ^2.

Dado que $\hat{\beta}$ estimador de mínimos quadrados é uma combinação linear das observações, têm-se que ele é normalmente distribuído com vetor médio $\boldsymbol{\beta}$ e matriz de covariância $\sigma^2(X'X)^{-1}$. Logo temos uma distribuição t, com n-p graus de liberdade, sendo Cjj, o jj-ésimo elemento da matriz $(X'X)^{-1}$ e σ^2 a estimativa da variância do erro dada por $\frac{e'e}{n-p}$. A equação (20) descreve a estatística usada para construção do intervalo de confiança.

$$T = \frac{\widehat{\beta_j} - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad \text{j= 0,1, ..., k} \qquad (20)$$

Logo, temos o intervalo de confiança 100(1 – α)% para o coeficiente de regressão βj, j = 0,1, ..., k é expresso pela equação (21)

$$IC_{\beta_j}^{100(1-\alpha)\%} = [\hat{\beta}_j \pm t_{1-\frac{\alpha}{2};n-p}\,] \qquad (21)$$

Se o intervalo de confiança contiver o valor zero então não rejeitamos H_0: $\beta_j = 0$.

2.2.6 *Análise de Resíduos*

Análise de resíduos é um método utilizado para averiguar o comportamento do modelo utilizando o conjunto de dados, e as diferenças entre os valores observados y_i e $\hat{y}_i$ os valores ajustados (resíduos) do modelo de regressão com o intuito de averiguar se as suposições feitas para o desenvolvimento do modelo são satisfeitas (MORETTIN; BUSSAB, 2010).
Supõem-se que os erros sejam distribuídos de forma aproximadamente normal, com variância constante. (MONTGOMERY; RUNGER, 2016).

A verificação das suposições feitas para os erros do modelo de regressão é realizada utilizando o gráfico de probabilidade normal dos resíduos, gráfico de dispersão entre resíduos e variáveis do modelo e gráfico de dispersão entre resíduos e valores preditos pelo modelo. Este último gráfico também é utilizado para verificação da suposição da linearidade entre a variável resposta e as variáveis explicativas.

2.3 Bootstrap Não Paramétrico

Método de simulação baseado em dados ao qual utiliza-se do computador e é aplicado em diferentes problemas (EFRON; TIBSHIRANI, 1993).

Utilizado em diversos campos de atuação a sua aplicação permite o tratamento de distribuições não conhecidas em modelos mais realísticos e destinados a simplificar o cálculo de inferências produzindo-as de forma automática (EFRON; TIBSHIRANI, 1993).

O Bootstrap não paramétrico, pelo método empírico, propõe probabilidades iguais n^{-1} para cada valor da amostra y_j(DAVIDOS; HINKLEY, 1997).

Efron e Tibshirani descreve:

> "As inferências bootstrap não paramétricas são assintoticamente eficiente. Ou seja, para grandes amostras, eles fornecem dados precisos, independentemente da população subjacente."

A amostra de Bootstrap não Paramétrico consiste em elementos escolhidos com reposição a partir da amostra original contendo o mesmo tamanho da amostra original n sendo a quantidade de reamostras possíveis n^n.

Seja F a função de distribuição empírica;

$$\hat{F}(y) = \frac{A\{y_j \leq y\}}{n} \tag{22}$$

Onde A é o número de vezes que o evento ocorre.

O procedimento de Bootstrap consiste nas etapas abaixo (EFRON, TIBSHIRANI, 1993):

1. Considera uma amostra contendo observações iid $x_i, i = 1, \dots, n$
2. A partir da distribuição empírica dos dados, tem-se que para cada valor x_i há igual probabilidade $\frac{1}{n}$ de ocorrer
3. Calcule a função F_n^*
    - Estimativa Bootstrap de T(F)
4. Repete-se os passos anteriores B vezes

2.3.1 *Estimativa Erro Padrão*

Segundo (EFRON, TIBSHIRANI, 1993) o erro padrão é frequentemente utilizado para atribuir valores aproximados de intervalo de confiança para um parâmetro $\emptyset$ de interesse. A partir de uma estimativa B e um erro padrão.

A estimativa do erro padrão na distribuição via Bootstrap é dado pelo desvio padrão amostral das estimativas $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{B}$

$$se(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{b-1}^{B}(\hat{\theta}^b - \overline{\hat{\theta}^*})} \quad (23)$$

Onde:

$\hat{\theta}^*$= Vetor contendo a estatística para cada reamostra;

B = Número de reamostragens

$\hat{\theta}$ = Estatística da amostra original.

2.3.2 *Estimativa de Viés*

Uma estatística utilizada para estimar um parâmetro contêm um viés quando a distribuição amostral não estiver centrada no verdadeiro valor do parâmetro. A técnica permite avaliar o vício verificando se a distribuição de Bootstrap da estatística está centrada na estatística da amostra mestre (MONTGOMERY; RUNGER,2016).

O estimador do viés Bootstrap é dado por:

$$Vies = \overline{\hat{\theta}^*} - \hat{\theta} \quad (24)$$

Valores positivos de viés indicam que, em média, tende a sobrestimar $\hat{\theta}$.

2.3.3 *Intervalo de Confiança*

Na literatura podemos identificar diversas abordagens para cálculo do intervalo de confiança (IC). Para esse trabalho será apresentado o IC percentílico que foi objeto desse estudo.

O Intervalo de Confiança de Bootstrap percentílico utiliza a distribuição empírica das estimativas de Bootstrap como distribuição de referência.

A função de distribuição $\hat{\theta}^*$, o intervalo de confiança percentílico $1-2\alpha$ é dado pelos percentis $\alpha\ e\ \alpha-1$, onde podemos considerar 97% e 2,5% e escrita conforma abaixo (EFRON, TIBSHIRANI, 1993):

$$\hat{\theta}_{\%,lo,}\hat{\theta}_{\%,up} = [\hat{\theta}^{*\alpha}, \hat{\theta}^{*(1-\alpha)}] \quad (25)$$

3 DESCRIÇÃO DA BASE DE DADOS

O trabalho foi realizado a partir de uma base de dados de uma empresa de logística que opera em diferentes setores da economia, a qual subsidia outras empresas operadoras de transporte intermodal pelo Brasil.

O estudo de caso tem por objetivo verificar a partir do Bootstrap não Paramétrico quais as variáveis utilizadas são significativas ao nível de confiança 95%.

A base coletada para esse estudo consta com 6833 registros do período de agosto/2021 a março/2022.

A variável preditora e as variáveis explicativas são descritas respectivamente nas **Tabelas 2** e **3**.

Tabela 2: Variável Preditora

Notação	***Descrição***
Tempo	Tempo gasto para sair de uma estação inicial a estação final

Fonte: Elaborado pela autora

Tabela 3: Variáveis Explicativas

Notação	***Descrição***
ComprimentoRota	Comprimento da distância percorrida
QtdVagoes	Total de vagões Transportados
ToneladaBruta	Tonelada bruta total puxada pelas locomotivas
qtdLocomotivas	Quantidade de locomotivas tracionando

Fonte: Elaborado pela autora

Realizado uma análise das estatísticas descritivas da base de dados, observa-se na **Tabela 4** que em média um trem gasta 1831 minutos para sair de uma estação A para chegar até a estação B, o equivalente a aproximadamente 1,27 dia. O maior valor observado na amostra foi de 48466, equivalente a aproximadamente 33,65 dias e o menor tempo foi de 1 minuto.

Tabela 4: Estatísticas Descritivas

Variável	**Mínimo**	**1º Q**	**Mediana**	**Média**	**3º Q**	**Máximo**	**SD**
Tempo	1	347	1157	1974	2475	38972	2394.668
ComprimentoRota	5000	67000	200598	324095	578870	1663074	325717.8
QtdVagoes	2.00	18.00	40.00	46.87	77.00	197.00	31.71594
ToneladaBruta	44000	1275660	3545288	4391189	7617583	24934819	3499421
qtdLocomotivas	1.000	2.000	3.000	3.193	4.000	12.000	1.808416

Fonte: Elaborado pela autora

4 ANÁLISE DOS DADOS DO EXEMPLO NUMÉRICO

Para o presente estudo será realizada o modelo de regressão múltipla com interações relacionando a variável resposta e variáveis explicativas.

A **Tabela 5** apresenta as interações das variáveis explicativas.

Tabela 5: Interações da Variáveis

Interações	**Variáveis**
ComprimentoRota * QtdVagoes	23
ComprimentoRota * ToneladaBruta	24
ComprimentoRota * qtdLocomotivas	25
QtdVagoes * ToneladaBruta	34
QtdVagoes * qtdLocomotivas	35
ToneladaBruta * qtdLocomotivas	45
ComprimentoRota * QtdVagoes* ToneladaBruta	234
ComprimentoRota * QtdVagoes* qtdLocomotivas	235
ComprimentoRota * ToneladaBruta * qtdLocomotivas	245
QtdVagoes * ToneladaBruta * qtdLocomotivas	345
ComprimentoRota * QtdVagoes* ToneladaBruta * qtdLocomotivas	2345

Fonte: Elaborado pela autora

Utilizando Software R, foi ajustado um modelo de regressão múltipla com interação. A **Figura 2** apresenta a saída do modelo.

Figura 2: Summary Modelo de Regressão

```
Call:
lm(formula = Tempo ~ ., data = Trem)

Residuals:
   Min     1Q Median     3Q    Max
 -9632   -458   -145    273  34344

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.543e+01  9.988e+01   0.955 0.339364
ComprimentoRota  2.057e-03  2.709e-04   7.593 3.54e-14 ***
QtdVagoes       -2.179e+01  5.754e+00  -3.787 0.000154 ***
ToneladaBruta    2.131e-04  7.454e-05   2.859 0.004257 **
qtdLocomotivas   9.304e+01  3.989e+01   2.332 0.019704 *
`23`             2.870e-04  1.281e-05  22.408  < 2e-16 ***
`24`            -2.641e-09  1.509e-10 -17.506  < 2e-16 ***
`25`             2.976e-04  6.495e-05   4.582 4.68e-06 ***
`34`            -1.063e-06  7.360e-07  -1.444 0.148813
`35`             6.051e+00  1.594e+00   3.795 0.000149 ***
`45`            -9.074e-05  2.075e-05  -4.373 1.24e-05 ***
`234`            3.354e-12  1.389e-12   2.415 0.015765 *
`235`           -5.574e-05  3.267e-06 -17.061  < 2e-16 ***
`245`            6.953e-10  4.017e-11  17.311  < 2e-16 ***
`345`            6.356e-07  2.307e-07   2.755 0.005886 **
`2345`          -1.947e-12  4.311e-13  -4.515 6.43e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1362 on 6817 degrees of freedom
Multiple R-squared:  0.6773,    Adjusted R-squared:  0.6766
F-statistic: 953.8 on 15 and 6817 DF,  p-value: < 2.2e-16
```

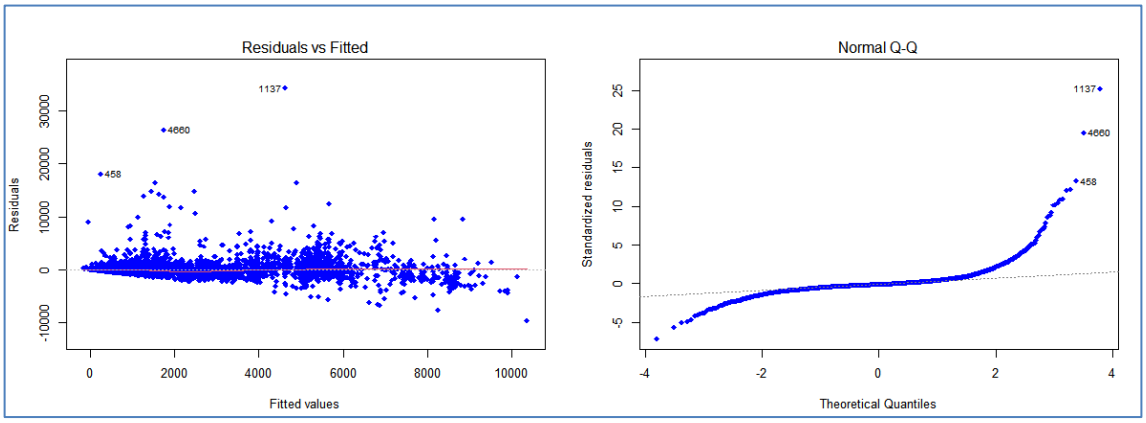
Fonte: Elaborado pela autora

Observando a saída do modelo de regressão múltipla com interação pela **Figura 2**, verifica-se que o Intercepto, a variável quantidade de locomotivas e a interação 34 e a 234 não são significativas. O R^2 ajustado é de 67,66% e p-value < 2.2e-16. No entanto, antes de avaliarmos os resultados é necessário verificar as hipóteses relativas à componente erro por meio da avaliação dos resíduos.

Pela **Figura 3** observamos pelo gráfico Resíduos x Preditos que não há indicação de heterocedasticidade (variância não constante) dos resíduos uma vez que parece não haver relação entre os resíduos e a variável $\hat{y}$.

Pelo gráfico Normal- QQ observa-se que não há normalidade nos erros, onde se vê que os dados não seguem uma tendência linear.

Figura 3: Resíduos



Fonte: Elaborado pela autora

Além da verificação da normalidade dos resíduos por meio normal plot desta verificação foi realizado o teste de hipóteses para verificar se os resíduos seguem uma distribuição normal.

Teste de hipóteses:

$$\begin{cases} H_0: \text{ Resíduos normais} \\ H_1: \textit{Resíduos não normais} \end{cases}$$

Por meio do teste de Anderson Darling, ver **Figura 4**, o valor obtido de p-valor é < 2.2e-16, indicando que se rejeita H_0 ao nível de 95% de confiança, ou seja, os resíduos não seguem uma distribuição normal.

Figura 4: Teste Anderson Darling

```
> ad.test(ModLinear$residuals)

        Anderson-Darling normality test

data:  ModLinear$residuals
A = 509.84, p-value < 2.2e-16
```

Fonte: Elaborado pela autora

Devido a não normalidade dos resíduos e não encontrarmos transformações adequadas para torná-los normais, um dos recursos possíveis para realização dos testes de hipóteses é o Bootstrap não Paramétrico e deste modo poderemos identificar quais as variáveis são significativas ao nível de confiança 95%.

Neste trabalho, o procedimento de Bootstrap não Paramétrico utilizado segue as seguintes etapas:

Etapa 1: Considere a base de dados de 6833 observações do sistema que se deseja verificar quais as variáveis são consideradas significativas por meio da regressão múltipla. A base é constituída da variável resposta, das 4 variáveis iniciais e as 11 variáveis adicionais relativo as interações. Assim, temos uma matrix M com 6833 linhas e 16 colunas. Faça i=1;

Etapa 2: Gere um novo banco de dados (Z) amostrando com reposição 6833 linhas da matriz M;

Etapa 3: Com a matriz Z obtenha os estimadores de mínimos quadrados e arquive os resultados na i-ésima linha de um vetor R;

Etapa 4: Faça i=i+1. Se i=10000 então finalize e caso contrário vá para Etapa 2.

A partir do data-frame gerado (matriz R) pelo procedimento de Bootstrap foi calculado o intervalo de confiança percentílico de 95% para cada parâmetro (LI e LS) de cada variável e realizado o Teste de Hipóteses para identificação de quais regressoras são significativas para o presente estudo. Se o intervalo de confiança conter o zero então a variável é julgada não significativa.

Resultado dos intervalos de confiança é apresentado na **Tabela 6**.

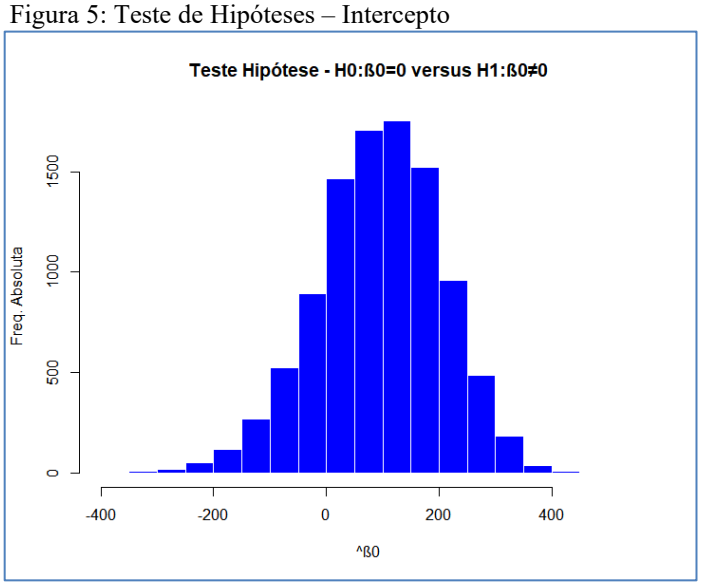
Tabela 6: Teste de Hipóteses

Variáveis	**Coeficientes**	**LI**	**LS**	**Decisão**
Intercepto	$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$	-134,568	296,7394947	$Não\ Rejeito\ H_0: \beta_0 = 0$
ComprimentoRota	$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$	0,001225	0,002796866	$Rejeito\ H_0: \beta_1 = 0$
QtdVagoes	$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$	-37,765	-8,09148	$Rejeito\ H_0: \beta_2 = 0$
ToneladaBruta	$\begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$	0,000005559	0,000451138	$Rejeito\ H_0: \beta_3 = 0$
qtdLocomotivas	$\begin{cases} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{cases}$	-12,319	229,9686411	$Não\ Rejeito\ H_0: \beta_4 = 0$
23	$\begin{cases} H_0: \beta_5 = 0 \\ H_1: \beta_5 \neq 0 \end{cases}$	0,000246	0,000331	$Rejeito\ H_0: \beta_5 = 0$
24	$\begin{cases} H_0: \beta_6 = 0 \\ H_1: \beta_6 \neq 0 \end{cases}$	-3,11E-09	-2,19E-09	$Rejeito\ H_0: \beta_6 = 0$
25	$\begin{cases} H_0: \beta_7 = 0 \\ H_1: \beta_7 \neq 0 \end{cases}$	0,000124	0,0005	$Rejeito\ H_0: \beta_7 = 0$
34	$\begin{cases} H_0: \beta_8 = 0 \\ H_1: \beta_8 \neq 0 \end{cases}$	-2,82E-06	7,09E-07	$Não\ Rejeito\ H_0: \beta_8 = 0$
35	$\begin{cases} H_0: \beta_9 = 0 \\ H_1: \beta_9 \neq 0 \end{cases}$	1,860059	10,92076174	$Rejeito\ H_0: \beta_9 = 0$
45	$\begin{cases} H_0: \beta_{10} = 0 \\ H_1: \beta_{10} \neq 0 \end{cases}$	-0,00018	-1,50E-05	$Rejeito\ H_0: \beta_{10} = 0$
234	$\begin{cases} H_0: \beta_{11} = 0 \\ H_1: \beta_{11} \neq 0 \end{cases}$	-7,14E-14	6,68E-12	$Não\ Rejeito\ H_0: \beta_{11} = 0$
235	$\begin{cases} H_0: \beta_{12} = 0 \\ H_1: \beta_{12} \neq 0 \end{cases}$	-6,85E-05	-4,41E-05	$Rejeito\ H_0: \beta_{12} = 0$
245	$\begin{cases} H_0: \beta_{13} = 0 \\ H_1: \beta_{13} \neq 0 \end{cases}$	5,60E-10	8,39E-10	$Rejeito\ H_0: \beta_{13} = 0$
345	$\begin{cases} H_0: \beta_{14} = 0 \\ H_1: \beta_{14} \neq 0 \end{cases}$	1,26E-08	1,28E-06	$Rejeito\ H_0: \beta_{14} = 0$
2345	$\begin{cases} H_0: \beta_{15} = 0 \\ H_1: \beta_{15} \neq 0 \end{cases}$	-3,12E-12	-7,19E-13	$Rejeito\ H_0: \beta_{15} = 0$

Fonte: Elaborado pela autora

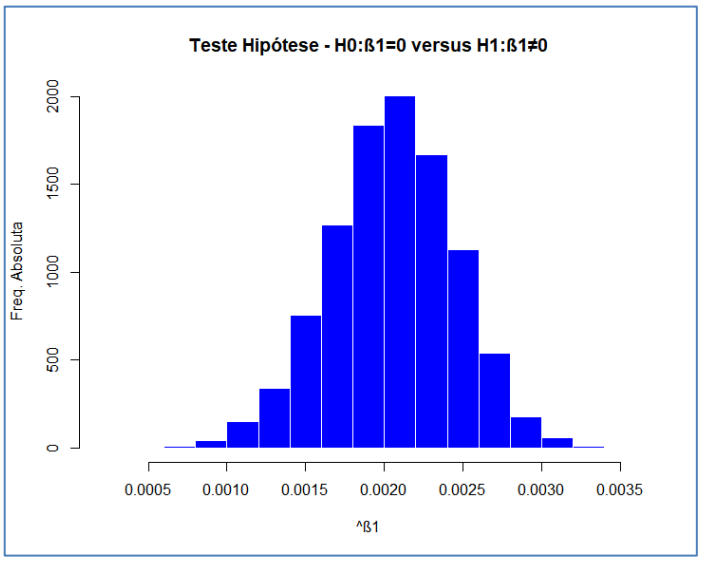
Nas **Figura 5-11** são apresentados histogramas com a indicação dos limites percentílicos de confiança de 95% para melhor visualização do leitor dos resultados dos testes de hipóteses.

Figura 5: Teste de Hipóteses – Intercepto



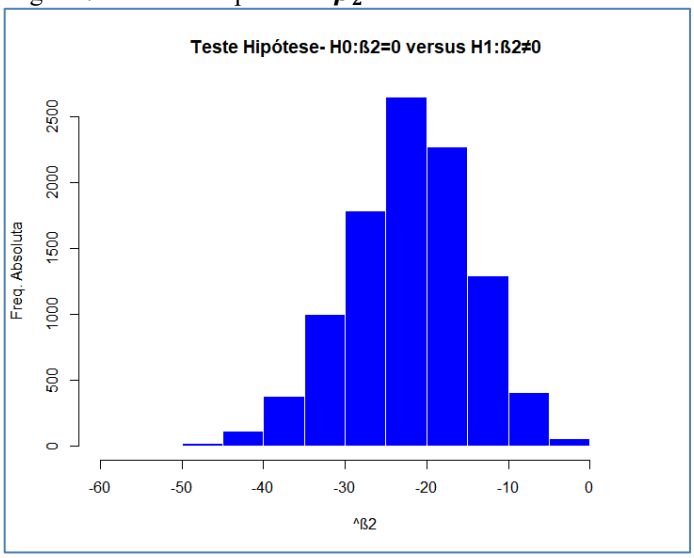
Fonte: Elaborado pela autora

Figura 6: Teste de Hipóteses - $\boldsymbol{\beta_1}$



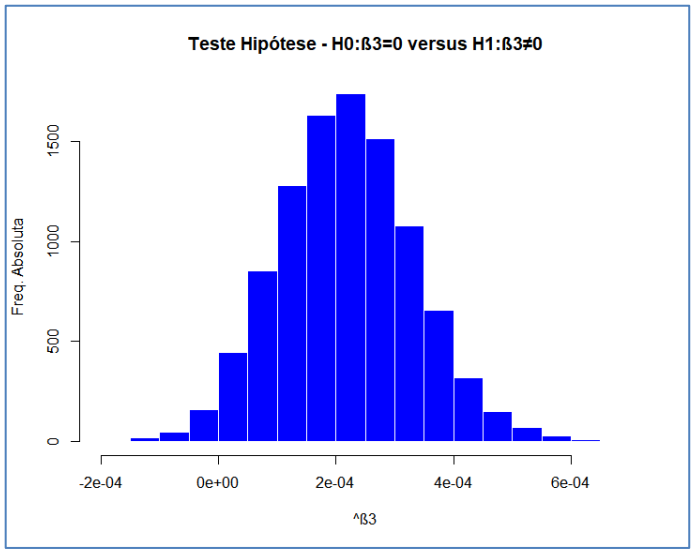
Fonte: Elaborado pela autora

Figura 7: Teste de Hipóteses - $\boldsymbol{\beta_2}$



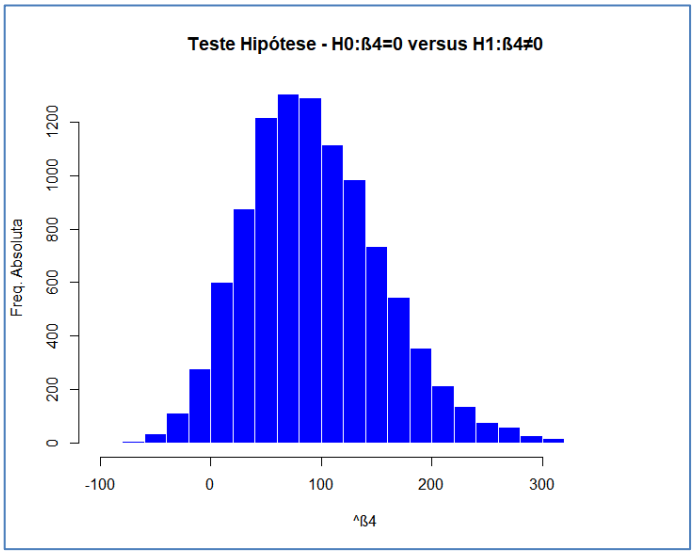
Fonte: Elaborado pela autora

Figura 8: Teste de Hipóteses - $\boldsymbol{\beta_3}$



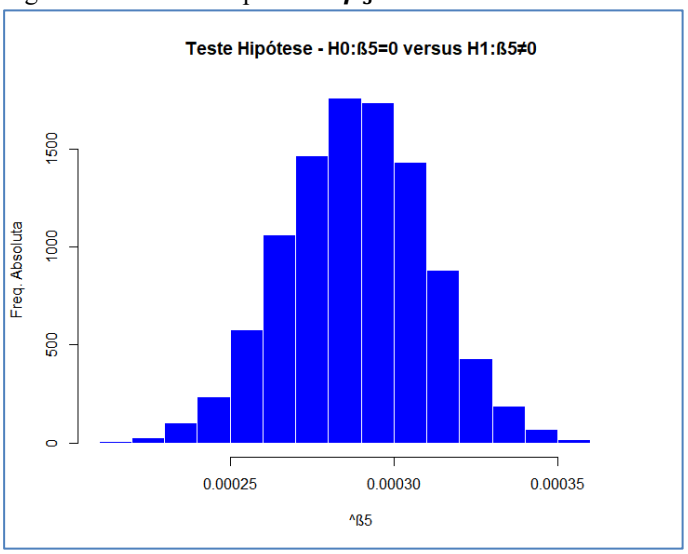
Fonte: Elaborado pela autora

Figura 9: Teste de Hipóteses - $\boldsymbol{\beta_4}$



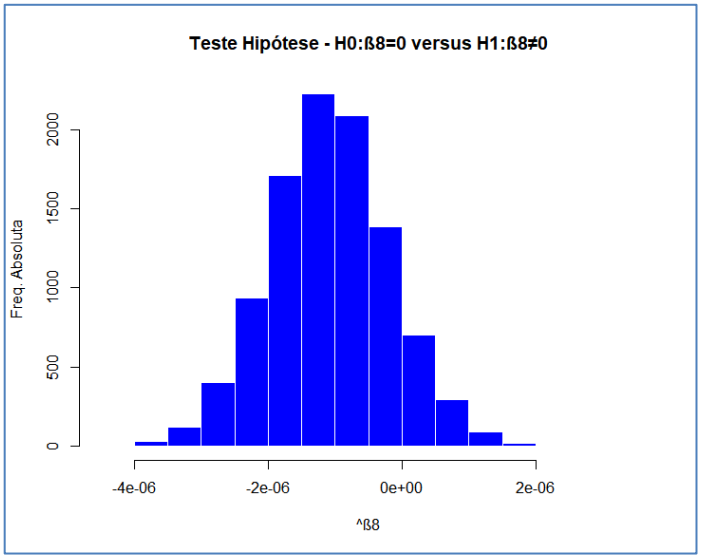
Fonte: Elaborado pela autora

Figura 10: Teste de Hipóteses - $\boldsymbol{\beta_5}$



Fonte: Elaborado pela autora

Figura 11: Teste de Hipóteses - $\boldsymbol{\beta_8}$



Fonte: Elaborado pela autora

Por meio dos Testes de Hipóteses identificou-se que o Intercepto não é significativo e será removido do modelo. A variável Quantidade de Locomotivas mesmo não apresentando ser significativa permanecerá no mesmo, pois as interações de ordem 3 e 4 se mostraram significativas. As interações 34 (QtdVagoes* ToneladaBruta) e 234 (ComprimentoRota * QtdVagoes* ToneladaBruta) não se mostraram significativas e foram removidas do modelo. Enfatizamos que o termo significativo usado neste texto está relacionado ao teste de hipóteses com nível de confiança 95%.

Nova execução das etapas do Bootstrap não Paramétrico foi realizada após remoção das variáveis não significativas e suas interações e as análises para identificação das variáveis relevantes ao nível de confiança de 95% são apresentados na **Tabela 7** e ilustrados pelos gráficos de histogramas nas **Figuras 12-17**.

Portanto, as variáveis significativas com nível de confiança obtidas foram:

- ComprimentoRota,
- QtdVagoes,
- ToneladaBruta

as interações foram:

- 23 - ComprimentoRota * QtdVagoes

- 24 - ComprimentoRota * ToneladaBruta
- 25 - ComprimentoRota * qtdLocomotivas
- 35 - QtdVagoes * qtdLocomotivas
- 45 - ToneladaBruta * qtdLocomotivas
- 235 - ComprimentoRota * QtdVagoes* qtdLocomotivas
- 245 - ComprimentoRota * ToneladaBruta * qtdLocomotivas
- 345 - QtdVagoes * ToneladaBruta * qtdLocomotivas
- 2345 - ComprimentoRota * QtdVagoes* ToneladaBruta * qtdLocomotivas

A variável quantidade Locomotivas mesmo não apresentando ser significativa permanecerá no mesmo, pois as interações de ordem 3 e 4 se mostraram significativas.

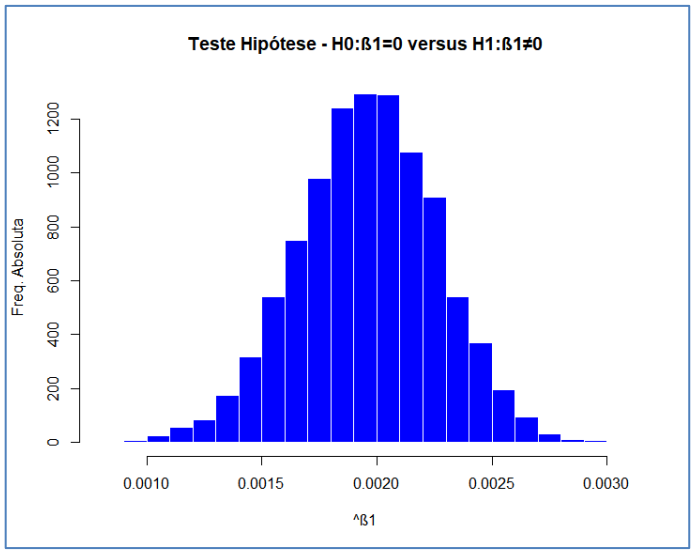
O resultado dos intervalos de confiança é apresentado na **Tabela 7**.

Tabela 7: Modelo Final - Testes de Hipóteses

Variáveis	Coeficientes	LI	LS	Decisão
ComprimentoRota	$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$	0,001349	0,002534	$Rejeito\ H_0: \beta_1 = 0$
QtdVagoes	$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases}$	-37,6929	-3,58526	$Rejeito\ H_0: \beta_2 = 0$
ToneladaBruta	$\begin{cases} H_0: \beta_3 = 0 \\ H_1: \beta_3 \neq 0 \end{cases}$	-4,08E-05	0,000323	$Não\ Rejeito\ H_0: \beta_3 = 0$
qtdLocomotivas	$\begin{cases} H_0: \beta_4 = 0 \\ H_1: \beta_4 \neq 0 \end{cases}$	80,26958	190,9696	$Rejeito\ H_0: \beta_4 = 0$
23	$\begin{cases} H_0: \beta_5 = 0 \\ H_1: \beta_5 \neq 0 \end{cases}$	0,000244	0,000328	$Rejeito\ H_0: \beta_5 = 0$
24	$\begin{cases} H_0: \beta_6 = 0 \\ H_1: \beta_6 \neq 0 \end{cases}$	-2,77E-09	-1,98E-09	$Rejeito\ H_0: \beta_6 = 0$
25	$\begin{cases} H_0: \beta_7 = 0 \\ H_1: \beta_7 \neq 0 \end{cases}$	0,000233	0,000443	$Rejeito\ H_0: \beta_7 = 0$
35	$\begin{cases} H_0: \beta_8 = 0 \\ H_1: \beta_8 \neq 0 \end{cases}$	1,049317	11,27841	$Rejeito\ H_0: \beta_8 = 0$
45	$\begin{cases} H_0: \beta_9 = 0 \\ H_1: \beta_9 \neq 0 \end{cases}$	-0,00015	-1,48E-05	$Rejeito\ H_0: \beta_9 = 0$
235	$\begin{cases} H_0: \beta_{10} = 0 \\ H_1: \beta_{10} \neq 0 \end{cases}$	-6,88E-05	-4,53E-05	$Rejeito\ H_0: \beta_{10} = 0$
245	$\begin{cases} H_0: \beta_{11} = 0 \\ H_1: \beta_{11} \neq 0 \end{cases}$	5,13E-10	7,62E-10	$Rejeito\ H_0: \beta_{11} = 0$
345	$\begin{cases} H_0: \beta_{12} = 0 \\ H_1: \beta_{12} \neq 0 \end{cases}$	2,21E-07	5,81E-07	$Rejeito\ H_0: \beta_{12} = 0$
2345	$\begin{cases} H_0: \beta_{13} = 0 \\ H_1: \beta_{13} \neq 0 \end{cases}$	-1,39E-12	-6,41E-13	$Rejeito\ H_0: \beta_{13} = 0$

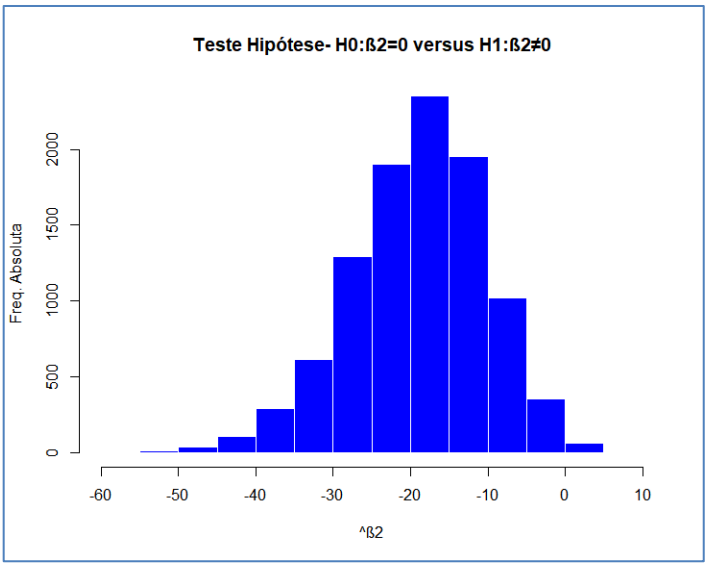
Fonte: Elaborado pela autora

Figura 12: Modelo Final - Teste de Hipóteses - $\boldsymbol{\beta_1}$



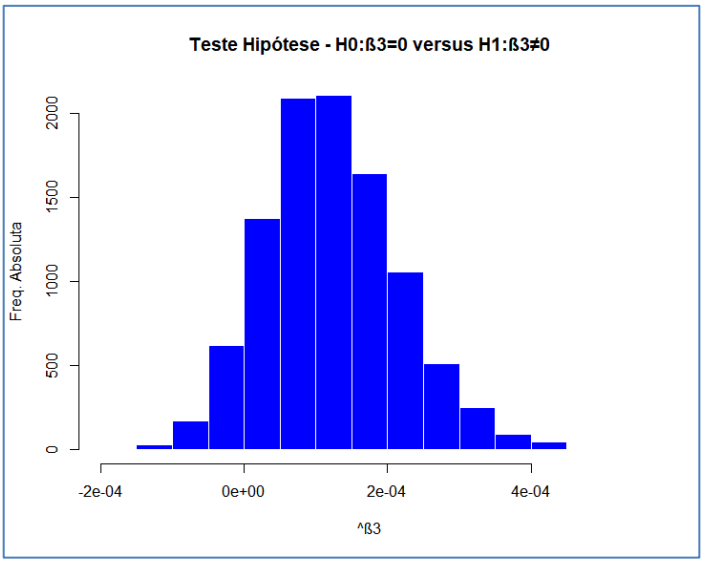
Fonte: Elaborado pela autora

Figura 13: Modelo Final - Teste de Hipóteses – $\boldsymbol{\beta_2}$



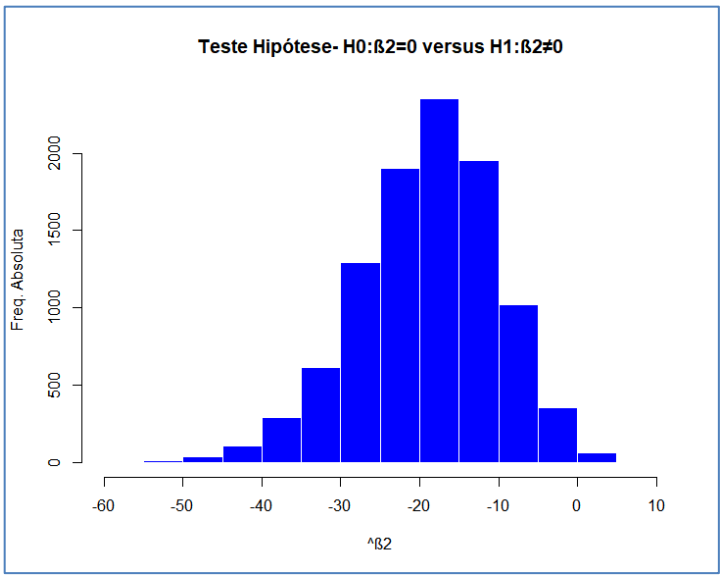
Fonte: Elaborado pela autora

Figura 14: Modelo Final - Teste de Hipóteses - $\boldsymbol{\beta_3}$



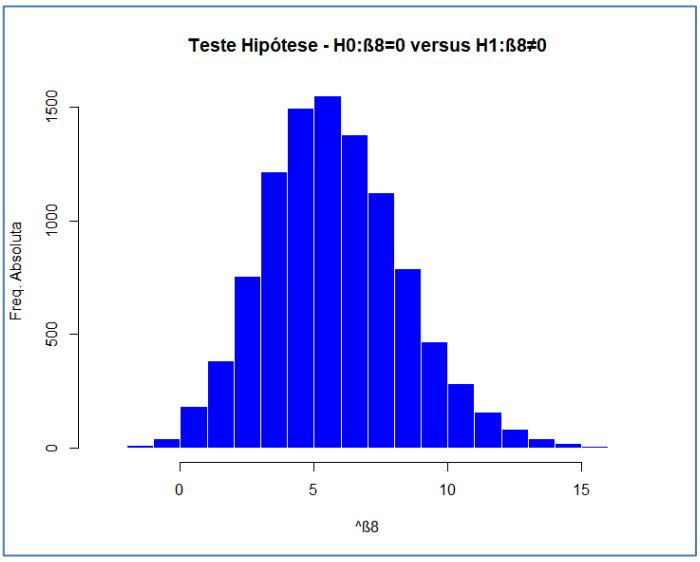
Fonte: Elaborado pela autora

Figura 15: Modelo Final - Teste de Hipóteses – $\boldsymbol{\beta_4}$



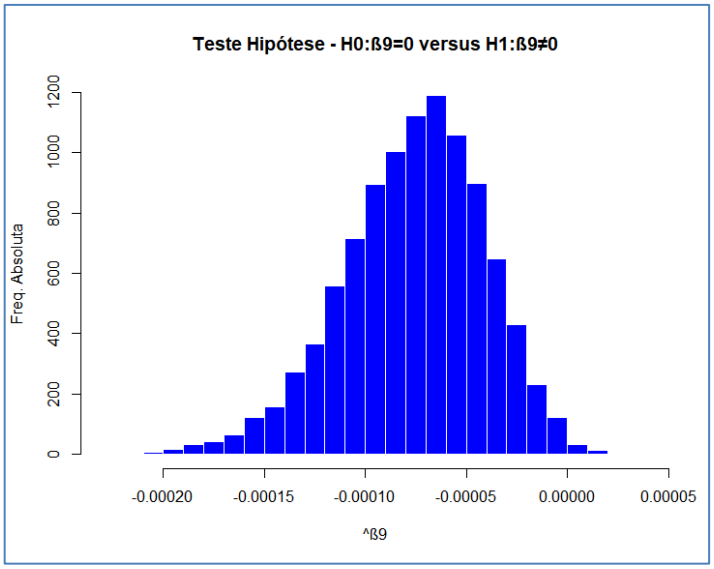
Fonte: Elaborado pela autora

Figura 16: Modelo Final - Teste de Hipóteses - $\boldsymbol{\beta_8}$



Fonte: Elaborado pela autora

Figura 17: Modelo Final - Teste de Hipóteses - $\boldsymbol{\beta_9}$



Fonte: Elaborado pela autora

5 CONCLUSÕES E SUGESTÃO DE ESTUDOS FUTUROS

Esse trabalho teve como objetivo verificar quais as variáveis são significativas (com confiança 95%) para um modelo de regressão múltipla com interação utilizando Bootstrap não Paramétrico para que em trabalhos futuros seja realizado a predição do tempo de chegado do trem ao seu destino.

Diante dos resultados apresentados, pode-se concluir que as variáveis significativas para posterior avaliação do tempo de chegada dos trens ao seu destino foram:

- Comprimento Rota
- Quantidade de Vagões
- Quantidade de Locomotivas

e as interações:

Tabela 8: Interações Significativas

Interações	**Variáveis**
ComprimentoRota * QtdVagoes	23
ComprimentoRota * ToneladaBruta	24
ComprimentoRota * qtdLocomotivas	25
QtdVagoes * qtdLocomotivas	35
ToneladaBruta * qtdLocomotivas	45
ComprimentoRota * QtdVagoes* qtdLocomotivas	235
ComprimentoRota * ToneladaBruta * qtdLocomotivas	245
QtdVagoes * ToneladaBruta * qtdLocomotivas	345
ComprimentoRota * QtdVagoes* ToneladaBruta * qtdLocomotivas	2345

Fonte: Elaborado pela autora

Apesar da variável Tonelada Bruta não apresentar significância para o modelo na segunda execução do procedimento de Bootstrap não Paramétrico, ela permaneceu, pois nas variáveis em que ocorre interação com ela, os dados foram relevantes.

REFERÊNCIAS

ANTF. INVESTIMENTOS ULTRAPASSAM R$ 141,9 BI. **ANTF.ORG**, 2022. Disponível em: <https://www.antf.org.br/wp-content/uploads/2022/03/Relat%C3%B3rio-deProdu%C3%A7%C3%A3o-Janeiro-de-2022-2.pdf>. Acesso em 25 de agosto de 2022.

Anuário do Setor Ferroviário. **GOV.BR**. Disponível em: <https://www.gov.br/antt/pt-br/assuntos/ferrovias/anuario-do-setor-ferroviario> Acesso em: 07 de out. de 2022.

BUSSAB, W.O. e Morettin, P.A. **Estatística Básica**. São Paulo: Atual, 1987.

CARGA GERAL: EXPANSÃO ANUAL MÉDIA CHEGA A 4,2%. **ANTF.ORG**, 2022. Disponível em: https://www.antf.org.br/relea. es/carga-geral/> Acesso em 20 de agosto de 2022.

DAVISON, A.C; HINKLEY, D.V; **Bootstrap methods and their application**. Cambridge University Press. Cambridge, 1997.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. Chapman & Hall, 1993.

Ferrovias Brasileiras. **GOV.BR**, 2022. Disponível em: <https://www.gov.br/infraestrutura/pt-br/assuntos/transporte-terrestre/ferrovias-brasileiras> Acesso em 20 de agosto de 2022.

MAPA FERROVIÁRIO. **ANTF.ORG**, 2021. Disponível em: <https://www.antf.org.br/wp-content/uploads/2021/06/Mapa-Site-VF.pdf> Acesso em: 07 de out. de 2022.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística e Probabilidade para Engenheiros**. 2. ed. Rio de Janeiro: LTC, 2016.

MONTGOMERY, Douglas. C.; PECK; Elizabeth A.; VINNING; G. Geofrrey. **Introduction to linear regression analysis**, 2. ed. New York: John Wiley and Sons, 1992.

ANEXO A – Programa R

```
library(pracma)
library(readxl)
library(tidyverse)
require(nortest)
library(dplyr)
library(magrittr)
#Leitura da Base de dados

Trem <- read_excel('TTrem.xlsx')

#str(dados)   #Examina a estrutura do data frame "dados"
str(Trem)

#Análise Descritiva
summary(Trem)
sd(Trem$Tempo)
sd(Trem$ComprimentoRota)
sd(Trem$QtdVagoes)
sd(Trem$ToneladaBruta)
sd(Trem$qtdLocomotivas)

## Armazenamos o número de linhas no dataframe
n <- nrow(Trem)

## Criar interações
D23 <- Trem$ComprimentoRota*Trem$QtdVagoes
D24 <- Trem$ComprimentoRota*Trem$ToneladaBruta
D25 <- Trem$ComprimentoRota*Trem$qtdLocomotivas
D34 <- Trem$QtdVagoes*Trem$ToneladaBruta
D35 <- Trem$QtdVagoes*Trem$qtdLocomotivas
D45 <- Trem$ToneladaBruta*Trem$qtdLocomotivas
D234 <- Trem$ComprimentoRota*Trem$QtdVagoes*Trem$ToneladaBruta
D235 <-Trem$ComprimentoRota*Trem$QtdVagoes*Trem$qtdLocomotivas
D245 <-Trem$ComprimentoRota*Trem$ToneladaBruta*Trem$qtdLocomotivas
D345 <-Trem$QtdVagoes*Trem$ToneladaBruta*Trem$qtdLocomotivas
D2345 <- Trem$ComprimentoRota*Trem$QtdVagoes*Trem$ToneladaBruta*Trem$qtdLocomotivas

Trem <- within(Trem, '23' <- c(D23))
Trem <- within(Trem, '24' <- c(D24))
Trem <- within(Trem, '25' <- c(D25))
Trem <- within(Trem, '34' <- c(D34))
Trem <- within(Trem, '35' <- c(D35))
Trem <- within(Trem, '45' <- c(D45))
Trem <- within(Trem, '234' <- c(D234))
Trem <- within(Trem, '235' <- c(D235))
Trem <- within(Trem, '245' <- c(D245))
Trem <- within(Trem, '345' <- c(D345))
Trem <- within(Trem, '2345' <- c(D2345))

## Ajuste Modelo de Regressão Múltipla com Interações
ModLinear <- lm(Tempo ~ ., data = Trem)
summary(ModLinear)
par(mfrow=c(2,2))
plot(ModLinear, pch=18, col="blue")

#Teste de normalidade Anderson-Darling
ad.test(ModLinear$residuals)

#Bootstrap
n=6833
Trem <- Trem[,1:5]
Trem=data.matrix(Trem)
corridas=10000
Resultados=matrix(0,corridas,16)
#Guarda a amostra das linhas que serão sorteadas
TremBoot=matrix(0,n,16)
for (i in 1:corridas){
  A=sample(1:n,n,replace=TRUE)
  for (j in 1:n){
    A1=A[j]
    TremBoot[j,1]=Trem[A1,1] #1
```

```
    TremBoot[j,2]=Trem[A1,2] #2
    TremBoot[j,3]=Trem[A1,3] #3
    TremBoot[j,4]=Trem[A1,4] #4
    TremBoot[j,5]=Trem[A1,5] #5
    TremBoot[j,6]=Trem[A1,2]*Trem[A1,3] #6
    TremBoot[j,7]=Trem[A1,2]*Trem[A1,4] #7
    TremBoot[j,8]=Trem[A1,2]*Trem[A1,5] #8
    TremBoot[j,9]=Trem[A1,3]*Trem[A1,4] #9
    TremBoot[j,10]=Trem[A1,3]*Trem[A1,5] #10
    TremBoot[j,11]=Trem[A1,4]*Trem[A1,5] #11
    TremBoot[j,12]=Trem[A1,2]*Trem[A1,3]*Trem[A1,4] #12
    TremBoot[j,13]=Trem[A1,2]*Trem[A1,3]*Trem[A1,5] #13
    TremBoot[j,14]=Trem[A1,2]*Trem[A1,4]*Trem[A1,5] #14
    TremBoot[j,15]=Trem[A1,3]*Trem[A1,4]*Trem[A1,5] #15
    TremBoot[j,16]=Trem[A1,2]*Trem[A1,3]*Trem[A1,4]*Trem[A1,5] #16

  }
TremBoot=data.frame(TremBoot)
R<-glm(TremBoot$X1~TremBoot$X2+TremBoot$X3
         +TremBoot$X4+TremBoot$X5+TremBoot$X6+TremBoot$X7+TremBoot$X8
         +TremBoot$X9+TremBoot$X10+TremBoot$X11+TremBoot$X12
         +TremBoot$X13+TremBoot$X14+TremBoot$X15+TremBoot$X16)

Resultados[i,1]=R$coefficients[1]
Resultados[i,2]=R$coefficients[2]
Resultados[i,3]=R$coefficients[3]
Resultados[i,4]=R$coefficients[4]
Resultados[i,5]=R$coefficients[5]
Resultados[i,6]=R$coefficients[6]
Resultados[i,7]=R$coefficients[7]
Resultados[i,8]=R$coefficients[8]
Resultados[i,9]=R$coefficients[9]
Resultados[i,10]=R$coefficients[10]
Resultados[i,11]=R$coefficients[11]
Resultados[i,12]=R$coefficients[12]
Resultados[i,13]=R$coefficients[13]
Resultados[i,14]=R$coefficients[14]
Resultados[i,15]=R$coefficients[15]
Resultados[i,16]=R$coefficients[16]


}

RF = data.frame(Resultados)
writexl::write_xlsx(RF,"RF_Inicial.xlsx")

RF <- read_excel("RF_Inicial.xlsx")

# Histograma
par(mfrow=c(8,2))

##Intercepto
hist(RF$X1,breaks = 20, main = "Teste Hipótese - H0:??0=0 versus H1:??0???0", xlab = "^??0",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##ComprimentoRota
hist(RF$X2,breaks = 20, main = "Teste Hipótese - H0:??1=0 versus H1:??1???0", xlab = "^??1",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##QtdeVAgoes
hist(RF$X3,breaks = 20, main = "Teste Hipótese- H0:??2=0 versus H1:??2???0", xlab = "^??2",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##ToneladaBruta
hist(RF$X4,breaks = 20, main = "Teste Hipótese - H0:??3=0 versus H1:??3???0", xlab = "^??3",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##QtdLocomotivas
hist(RF$X5,breaks = 20, main = "Teste Hipótese - H0:??4=0 versus H1:??4???0", xlab = "^??4",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##23
hist(RF$X6,breaks = 20, main = "Teste Hipótese - H0:??5=0 versus H1:??5???0", xlab = "^??5",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##24
hist(RF$X7,breaks = 20, main = "Teste Hipótese - H0:??6=0 versus H1:??6???0", xlab = "^??6",
```

```
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##25
hist(RF$X8,breaks = 20, main = "Teste Hipótese - H0:???7=0 versus H1:???7??0", xlab = "^???7",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##34
hist(RF$X9,breaks = 20, main = "Teste Hipótese - H0:???8=0 versus H1:???8??0", xlab = "^???8",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##35
hist(RF$X10,breaks = 20, main = "Teste Hipótese - H0:???9=0 versus H1:???9??0", xlab = "^???9",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##45
hist(RF$X11,breaks = 20, main = "Teste Hipótese - H0:??10=0 versus H1:??10??0", xlab =
"^??10",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##234
hist(RF$X12,breaks = 20, main = "Teste Hipótese - H0:??11=0 versus H1:??11??0", xlab =
"^??11",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##235
hist(RF$X13,breaks = 20, main = "Teste Hipótese - H0:??12=0 versus H1:??12??0", xlab =
"^??12",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##245
hist(RF$X14,breaks = 20, main = "Teste Hipótese - H0:??13=0 versus H1:??13??0", xlab =
"^??13",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##345
hist(RF$X15,breaks = 20, main = "Teste Hipótese - H0:??14=0 versus H1:??14??0", xlab =
"^??14",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
##2345
hist(RF$X16,breaks = 20, main = "Teste Hipótese - H0:??15=0 versus H1:??15??0", xlab =
"^??150",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

#Intervalo de Confiança
LI = matrix(0,16,2)
IC = matrix(0,16,2)
RF=data.matrix(RF)

for (j in 1:16) {

  LI[j,1] <- quantile(RF[, j], 0.025)
  LI[j,2] <- quantile(RF[, j], 0.975)
}
#Formatando o dado
IC[,1]= scales::number(LI[,1], accuracy =0.0000000000000000001, big.mark = ".", decimal.mark
= ",")
IC[,2]= scales::number(LI[,2], accuracy =0.0000000000000000001, big.mark = ".", decimal.mark
= ",")
IC

#2ª Execução retirando variáveis não significativas

#Bootstrap
Trem=data.matrix(Trem)
corridas=10000
Resultados=matrix(0,corridas,14)
#Guarda a amostra das linhas que serão sorteadas
TremBoot=matrix(0,n,14)
for (i in 1:corridas){
  A=sample(1:n,n,replace=TRUE)
  for (j in 1:n){
    A1=A[j]
    TremBoot[j,1]=Trem[A1,1] #1
    TremBoot[j,2]=Trem[A1,2] #2
    TremBoot[j,3]=Trem[A1,3] #3
    TremBoot[j,4]=Trem[A1,4] #4
    TremBoot[j,5]=Trem[A1,5] #5
    TremBoot[j,6]=Trem[A1,2]*Trem[A1,3] #6
    TremBoot[j,7]=Trem[A1,2]*Trem[A1,4] #7
```

```
    TremBoot[j,8]=Trem[A1,2]*Trem[A1,5] #8
    TremBoot[j,9]=Trem[A1,3]*Trem[A1,5] #9
    TremBoot[j,10]=Trem[A1,4]*Trem[A1,5] #10
    TremBoot[j,11]=Trem[A1,2]*Trem[A1,3]*Trem[A1,5] #11
    TremBoot[j,12]=Trem[A1,2]*Trem[A1,4]*Trem[A1,5] #12
    TremBoot[j,13]=Trem[A1,3]*Trem[A1,4]*Trem[A1,5] #13
    TremBoot[j,14]=Trem[A1,2]*Trem[A1,3]*Trem[A1,4]*Trem[A1,5] #14

  }
  TremBoot=data.frame(TremBoot)
  #Removido o intercepto
  R<-lm(TremBoot$X1~-1+TremBoot$X2+TremBoot$X3
        +TremBoot$X4+TremBoot$X5+TremBoot$X6+TremBoot$X7+TremBoot$X8
        +TremBoot$X9+TremBoot$X10+TremBoot$X11+TremBoot$X12
        +TremBoot$X13+TremBoot$X14)

  Resultados[i,1]=R$coefficients[1]
  Resultados[i,2]=R$coefficients[2]
  Resultados[i,3]=R$coefficients[3]
  Resultados[i,4]=R$coefficients[4]
  Resultados[i,5]=R$coefficients[5]
  Resultados[i,6]=R$coefficients[6]
  Resultados[i,7]=R$coefficients[7]
  Resultados[i,8]=R$coefficients[8]
  Resultados[i,9]=R$coefficients[9]
  Resultados[i,10]=R$coefficients[10]
  Resultados[i,11]=R$coefficients[11]
  Resultados[i,12]=R$coefficients[12]
  Resultados[i,13]=R$coefficients[13]

}

RF = data.frame(Resultados)
writexl::write_xlsx(RF,"RF_Final.xlsx")

RF <- read_excel("RF_Final.xlsx")

# Histograma
par(mfrow=c(8,2))


hist(RF$ComprimentoRota,breaks = 20, main = "Teste Hipótese - H0:??1=0 versus H1:??1???0",
xlab = "^??1",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$QtdVagoes,breaks = 20, main =  "Teste Hipótese- H0:??2=0 versus H1:??2???0", xlab =
"^??2",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$ToneladaBruta,breaks = 20, main = "Teste Hipótese - H0:??3=0 versus H1:??3???0",
xlab = "^??3",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$qtdLocomotivas,breaks = 20, main = "Teste Hipótese - H0:??4=0 versus H1:??4???0",
xlab = "^??4",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`23`,breaks = 20, main = "Teste Hipótese - H0:??5=0 versus H1:??5???0", xlab = "^??5",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`24`,breaks = 20, main = "Teste Hipótese - H0:??6=0 versus H1:??6???0", xlab = "^??6",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`25`,breaks = 20, main = "Teste Hipótese - H0:??7=0 versus H1:??7???0", xlab = "^??7",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`35`,breaks = 20, main = "Teste Hipótese - H0:??8=0 versus H1:??8???0", xlab = "^??8",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`45`,breaks = 20, main = "Teste Hipótese - H0:??9=0 versus H1:??9???0", xlab = "^??9",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)
```

```
hist(RF$`235`,breaks = 20, main = "Teste Hipótese - H0:??10=0 versus H1:??10???0", xlab =
"^??10",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`245`,breaks = 20, main = "Teste Hipótese - H0:??11=0 versus H1:??11???0", xlab =
"^??11",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`345`,breaks = 20, main = "Teste Hipótese - H0:??12=0 versus H1:??12???0", xlab =
"^??12",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)

hist(RF$`2345`,breaks = 20, main = "Teste Hipótese - H0:??13=0 versus H1:??13???0", xlab =
"^??13",
     ylab = "Freq. Absoluta", col = c("blue"), border = FALSE)


#Intervalo de Confiança
LI = matrix(0,13,2)
IC = matrix(0,13,2)
RF=data.matrix(RF)

for (j in 1:13) {

  LI[j,1] <- quantile(RF[, j], 0.025)
  LI[j,2] <- quantile(RF[, j], 0.975)
}
#Formatando o dado
IC[,1]= scales::number(LI[,1], accuracy =0.0000000000000000001, big.mark = ".", decimal.mark
= ",")
IC[,2]= scales::number(LI[,2], accuracy =0.0000000000000000001, big.mark = ".", decimal.mark
= ",")
IC
```