UNIVERDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Especialização em Estatística

Mirelle Rachel de Sales Castor

BOOTSTRAP NÃO PARAMÉTRICO APLICADO AO MODELO DE REGRESSÃO LINEAR MÚLTIPLA COM USO DA PLANILHA DO EXCEL

Belo Horizonte
2020

Mirelle Rachel de Sales Castor

BOOTSTRAP NÃO PARAMÉTRICO APLICADO AO MODELO DE REGRESSÃO LINEAR MÚLTIPLA COM USO DA PLANILHA DO EXCEL

Versão Final

Monografia de especialização apresentada ao Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Especialista em Estatística Aplicada.

Orientador: Prof. Dr. Roberto da Costa Quinino

Belo Horizonte
2020

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6ª Região nº 1510

Castor, Mirelle Rachel de Sales

C354b Bootstrap não paramétrico aplicado ao modelo de regressão linear múltipla com uso da planilha do excel / Mirelle Rachel de Sales Castor.— Belo Horizonte, 2020. n.p.. il.; 29 cm.

Especialização (monografia) - Universidade Federal de Minas Gerais – Departamento de Estatística.
Orientador: Roberto da Costa Quinino.

1. Estatística. 2. Análise de regressão. .3. Bootstrap (Estatística). 4. Excel (Programa de computador). I. Orientador. II. Título.

CDU 519.6 (043)

Universidade Federal de Minas Gerais
de Ciências Exatas
Departamento de Estatística

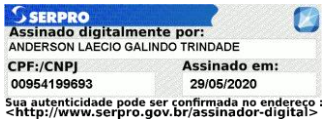
Programa de Pós-Graduação / Especialização
Av. Pres. Antônio Carlos, 6627 - Pampulha
31270-901 – Belo Horizonte – MG

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

ATA DO 207ª. TRABALHO DE FIM DE CURSO DE ESPECIALIZAÇÃO EM ESTATÍSTICA DE MIRELLE RACHEL DE SALES CASTOR.

Aos vinte e sete dias do mês de maio de 2020, às 08:00 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 **Mirelle Rachel de Sales Castor**, intitulado: **"**Bootstrap Não Paramétrico Aplicado ao Modelo de Regressão Linear Múltipla com uso da Planilha do Excel**"**, 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, 27 de maio de 2020.

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



Prof. Anderson Laécio Galindo Trindade
DEP / UFMG

p/ 

Daniela Carneiro Tibo
CEMIG SAÚDE

AGRADECIMENTOS

Agradeço ao professor Roberto, pela disponibilidade e apoio para a conclusão deste trabalho.

Agradeço ao Vinícius, por todo apoio e compreensão pelas minhas constantes ausências.

Agradeço as minhas irmãs por estarem ao meu lado desde sempre.

Agradeço aos amigos do trabalho, que tanto me ouviram falar das diversas versões desta monografia.

E por fim, agradeço aos meus amigos que me incentivaram a finalizar mais esta etapa. Em especial agradeço a Giselle, pela amizade, carinho e atenção ao ler este trabalho.

RESUMO

O objetivo da análise de regressão é estudar o relacionamento entre uma variável Y, denominada dependente ou resposta, e uma ou mais de variáveis independentes ou regressoras. Ou seja, compreender como determinadas variáveis influenciam no comportamento de outra variável. Entende-se que a regressão possui como principal objetivo fornecer uma equação que relaciona a variável dependente com as variáveis independentes consideradas no modelo, possibilitando também fazer predições sobre o comportamento do fenômeno estudado, ajustar parâmetros ou o modelo e realizar inferências sobre elas. Muitas aplicações da análise de regressão envolvem situações em que há mais de uma variável regressora. Neste trabalho aplicamos o método do Bootstrap por pares na utilização dos testes de significância necessários no modelo de regressão múltipla. O objetivo foi flexibilizar a necessidade da componente erro ($\in$) do modelo possuir distribuição normal e variância constante. O procedimento foi implementado na ferramenta xxcel e não demandou programação. Todas as funções necessárias foram descritas em detalhes para uso em um ambiente empresarial com uso da planilha Excel, favorecendo também a compreensão do método do bootstrap.

Palavras-chaves: Regressão linear múltipla; Bootstrap não paramétrico; Excel.

ABSTRACT

The purpose of regression analysis is to study the relationship between a variable Y, called a dependent or response, and one or more of independent or regressing variables. That is, to understand how certain variables influence the behavior of another variable. It is understood that the main objective of regression is to provide an equation that relates the dependent variable with the independent variables considered in the model, also making it possible to make predictions about the behavior of the studied phenomenon, adjust parameters or the model and make inferences about them. Many applications of regression analysis involve situations in which there is more than one regressor variable. In this work we apply the Bootstrap method in pairs in the use of the tests of significance required in the multiple regression model. The objective was to relax the need for the error component ($\in$) of the model to have normal distribution and constant variance. The procedure was implemented in the Excel spreadsheet and did not require programming. All the necessary functions were described in detail for use in a business environment using the Excel spreadsheet, also favoring the understanding of the bootstrap method.

Keywords: Multiple linear regression; Non-parametric bootstrap; Excel.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A análise de regressão estuda o relacionamento entre uma variável Y, denominada dependente, e uma ou várias variáveis independentes $X_1, X_2, \dots, X_p$. Caso se considere apenas uma variável independente denominamos de análise de regressão simples, caso usemos duas ou mais variáveis, de análise de regressão múltipla.

A importância do estudo da análise de regressão advém da necessidade do entendimento de determinados fenômenos nas Ciências da Natureza como na Física, Biologia, Química, dentre outras, nas Ciências Sociais, nas Ciências da Saúde, e na Engenharia que podem ser explicados pelo relacionamento linear entre uma variável dependente em função das variáveis independentes ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$). Embora que operacionalmente simples, existem aspectos relacionados a significância dos parâmetros que precisam de um cuidado especial. Tradicionalmente a verificação da significância dos parâmetros é realizado adotando as hipóteses de que a componente erro ($\in$) possui distribuição normal, variância constante e erros não correlacionados. Em situações práticas é comum verificar situações em que tais hipóteses não são satisfeitas.

Neste trabalho apresentaremos uma metodologia denominada Bootstrap não paramétrico por pares que apresenta bons resultados mesmo quando a componente erro é não normal e/ou não possui variância constante. O objetivo básico do bootstrap é reamostrar de um conjunto de dados, diretamente ou via um modelo ajustado, a fim de criar réplicas dos dados, a partir das quais podemos avaliar a variabilidade de quantidades de interesse, sem usar cálculos analíticos. O método bootstrap obtém sua amostra via amostragem com reposição da amostra original. A chave é a substituição das observações após a amostragem, o que permite ao pesquisador criar tantas amostras quanto necessárias. Permitindo assim estimarmos o erro padrão da estatística bem como construir intervalos de confiança ou realizar testes de hipóteses sobre parâmetros de interesse. Em geral o uso do Bootstrap demanda o uso de pacotes estatísticos específicos, mas em geral boa parte das empresas não o possuem ou o acham de difícil utilização quando comparado com o Excel. Neste sentido, toda a metodologia será explicada com uso do Excel por meio de planilha e sem a necessidade de programação em *softwares* estatísticos.

2 OBJETIVO GERAL

O objetivo desta monografia é apresentar a metodologia bootstrap não paramétrico para ser usado em modelos de regressão linear múltipla em que a componente erro é não normal e/ou não possui variância constante, com o uso da ferramenta excel.

3 FUNDAMENTAÇÃO TEÓRICA

Em diversas situações nos campos de estudo da área médica, industrial, biológica, química, entre outras, torna-se fundamental verificar se duas ou mais variáveis estão relacionadas de alguma forma. Para expressar esta relação precisamos estabelecer um modelo matemático. Este tipo de modelagem é chamado de regressão e possibilita compreender como determinadas variáveis influenciam em outra variável, ou seja, como o comportamento de uma variável pode mudar o comportamento de outra. (MONTGOMERY; PECK; VINNING, 1992b)

A regressão é uma técnica estatística que Milone e Angelini (1995) segundo permite construir os modelos e avaliar sua qualidade na chamada análise de regressão, sendo baseadas em técnicas de amostragem. Desta forma entende-se que a regressão tem como função básica fornecer uma equação que relaciona a variável dependente com as variáveis independentes consideradas no modelo, possibilitando fazer predições sobre o comportamento do fenômeno estudado, auxiliar no processo de seleção das variáveis que impactam significativamente na variação do que se está sendo estudado, estimar parâmetros ou ajustar um modelo e realizar inferências sobre eles, tais como, testes de hipóteses e intervalos de confiança.

Chamamos de regressão linear simples, quando há apenas uma variável resposta ou dependente e chamamos de regressão linear múltipla, quando há mais de uma variável regressora ou independente. (MONTGOMERY; PECK; VINNING, 1992b)

3.1 Modelo de regressão linear simples

O modelo de regressão linear simples relaciona uma variável aleatória Y, denominada variável resposta ou dependente, com uma variável X, denominada de variável regressora ou independente.

A análise de regressão parte de um conjunto de n observações pareadas (x, y), relativas às variáveis X e Y . Diz-se que um determinado valor de y depende, em parte, do correspondente valor de x.

Esta correspondência é descrita por uma relação linear entre x e y, expressa na equação (1):

$$Y = \beta_0 + \beta_1 X + \in \quad (1)$$

Na equação (1) a componente ∈ representa o efeito aleatório, isto é, o efeito de uma infinidade de fatores que estão afetando a observação y de forma aleatória e não somos capazes de controlar. Os erros aleatórios devem ser independentes e identicamente distribuídos com distribuição normal e variância constante. Os parâmetros β_0 e β_1 do modelo são desconhecidos e devem ser estimados.

No método de estimação dos parâmetros do modelo, o objetivo esperado é encontrar uma reta que passe mais próximo possível dos pontos observados, utilizando-se o critério dos mínimos quadrados (BARBETTA, 1998). As estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ são dadas respectivamente por (2) e (3):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (2)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \quad (3)$$

O Método dos Mínimos Quadrados (MMQ) é uma eficiente estratégia de estimação dos parâmetros da regressão e sua aplicação não é limitada apenas às relações lineares.

Existem casos em que o estudo com modelo de regressão linear simples, ou seja, um modelo no qual a variável dependente é relacionada a uma única variável independente, não é suficiente para representar a realidade dos fenômenos. Em busca da melhor alternativa recorre-se à análise com várias variáveis independentes.

3.2 Regressão linear múltipla

A regressão linear múltipla pode ser usada nos estudos de fenômenos que são representados por funções de mais de uma variável independente. (MILONE; ANGELINI, 1995)

Conforme a significância relativa do conjunto de variáveis independentes, a estimativa da variável dependente baseada numa única variável independente pode ser consideravelmente imprecisa. Com o objetivo de melhorar a capacidade de predição do modelo, utiliza-se outras variáveis independentes, considerando principalmente as mais significativas.

A equação da regressão múltipla tem a forma expressa em (4):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_\kappa X_{i\kappa} + \in_i \qquad (4)$$

Onde:

Y_i: variável dependente medida no elemento i da amostra;

$X_{i1}, X_{i2}, X_{i\kappa}$: variáveis independentes;

β_0 : intercepto, também conhecido como média geral;

$\beta_k, k = 1, \ldots, k$: determina o efeito da variável independente;

k o número de variáveis explicativas do modelo;

$\in_i$: erro aleatório associado ao elemento i da amostra, i =1,2,..., n.

O termo linear é usado porque a equação (4) é uma função linear dos parâmetros desconhecidos $\beta_0, \beta_1, \ldots, e\, \beta_k$. Estes modelos de regressão linear múltiplas são frequentemente utilizados como funções de aproximações, ou seja, a verdadeira relação funcional entre Y e $X_1, X_2, \ldots, X_k$ é desconhecida, porém em certas faixas das variáveis independentes o modelo de regressão linear é uma aproximação adequada. (MONTGOMERY; RUNGER, 2012a).

3.3 Estimação de Mínimos Quadrados dos Parâmetros

De acordo com Montgomery e Runger (2012a), o método dos mínimos quadrados (MMQ) pode ser usado para estimar os coeficientes de regressão múltipla.

A função dos mínimos quadrados é dada por (5):

$$L = \sum_{i=1}^{n} \in_1^2 = \sum_{i=1}^{n} \left(Y_\text{i} - \beta_0 - \sum_{j=1}^{k} \beta_j X_\text{ij}\right)^2 \tag{5}$$

Ao minimizarmos *L* com relação a $\beta_0, \beta_1, \ldots, \beta_k$. As estimativas de mínimos quadrados de $\beta_0, \beta_1, \ldots, \beta_k$ têm que satisfazer (6):

$$\frac{\partial L}{\partial \beta_0}_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_\text{i} - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j X_\text{ij} \right) = 0 \tag{6}$$

Simplificando a equação (6), obtemos as equações normais de mínimos quadrados (7):

$$\begin{aligned}
& n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_\text{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_\text{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_\text{ik} = \sum_{i=1}^{n} y_\text{i} \\
& \hat{\beta}_0 \sum_{i=1}^{n} x_\text{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x^2{}_\text{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_\text{i1} x_\text{i2} \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_\text{i1} x_\text{ik} = \sum_{i=1}^{n} x_\text{i1} y_i \\
& \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
& \hat{\beta}_0 \sum_{i=1}^{n} x_\text{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_\text{ik} x_\text{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_\text{ik} x_\text{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik} y_\text{i}
\end{aligned} \tag{7}$$

A solução para as equações normais serão os estimadores de mínimos quadrados dos coeficientes de regressão, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

3.4 Abordagem matricial para regressão múltipla

Ao realizarmos o ajuste de um modelo de regressão múltipla, indicamos a expressão das operações matemáticas utilizando notação matricial. Com isso, suponhamos que haja k variáveis independentes e n observações $(X_{i1}, X_{i2}, \dots, X_{ik}, Y_i), i = 1,2, \dots, n,$ e que o modelo relacionando as variáveis independentes às dependentes seja dado pela por (8):

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \in_i \qquad i = 1,2, \dots, n \tag{8}$$

O modelo expresso em (8) é um sistema de n equações, que pode ser expresso na notação matricial como (9):

$$y = X\beta + \in \tag{9}$$

Sendo (10):

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_3 \end{bmatrix} \quad e \quad \in = \begin{bmatrix} \in_1 \\ \in_2 \\ \vdots \\ \in_3 \end{bmatrix} \tag{10}$$

Em geral, y é um vetor (n x 1) das observações considerada variável dependente, X é uma matriz (n x k) dos níveis das variáveis independentes, β é um vetor (k + 1 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.

Deseja-se encontrar o vetor dos estimadores de mínimos quadrados, $\hat{\beta}$, que minimiza (11):

$$L = \sum_{i=1}^{n} \in_1^2 = \in' \in = (y - X\beta)'(y - X\beta) \tag{11}$$

Os estimadores de β_j são os valores que minimizam L expresso em (11). Eles são dados por (12):

$$\frac{\partial L}{\partial \beta} = 0 \tag{12}$$

As equações resultantes que têm de ser resolvidas são (13):

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

As equações (13) são normais de mínimos quadrados na forma matricial, sendo idênticas à forma escalar das equações normais dadas em (7). Com o objetivo de resolvermos as equações normais, deve-se multiplicar ambos os lados das equações (13) pelo inverso $X'X$. Consequentemente, a estimativa de mínimos quadrados de β é (14):

$$\hat{\beta} = (X'X)^{-1}X'y \tag{14}$$

Observe que há $p = k + 1$ equações normais para $p = k + 1$ incognitas. Além disso, a matriz $X'X$ é sempre não singular, como foi considerado anteriormente, de modo que, para inverter essas matrizes, os métodos descritos 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):

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

Podemos reescrever o modelo ajustado de regressão (16):

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij} \qquad i = 1,2,\ldots,n \quad (16)$$

Na notação matricial, o modelo ajustado é (17):

$$\hat{y}_i = X\hat{\beta}_j \quad (17)$$

A diferença entre a observação y_i e o valor ajustado $\hat{y}_i$ é um resíduo, $e_i = y_j - \hat{y}_j$. O vetor (*n* x 1) dos resíduos é denotado por (18):

$$e = y - \hat{y} \quad (18)$$

3.5 Teste de significância para a regressão

O Teste para a significância da regressão é um teste para determinar se existe uma relação linear entre a variável de resposta y e um subconjunto de regressores x_1, x_2, ..., x_{1k}. As hipóteses apropriadas são expressas por (19):

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: \beta_j \neq 0 \; para \; no \; mínimo \; um \; j \end{cases} \quad (19)$$

A rejeição de $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ implica que no mínimo uma das variáveis regressoras x_1, x_2, ..., x_{1k} contribui significativamente para o modelo.

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 em uma soma dos quadrados devido à regressão e em uma soma dos quadrados devido ao erro como descrito em (20):

$$SQ_T = SQ_R + SQ_E \quad (20)$$

A soma dos quadrados devido ao erro, SQ_E é dada por (21):

$$SQ_E = SQ_T - \hat{\beta}_1 S_{xy} \quad (21)$$

A soma dos quadrados devido a regressão, SQ_R é dada por (22):

$$SQ_R = \hat{\beta}_1 S_{xy} \quad (22)$$

Considerando $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ como verdadeiro, então SQ_R/σ^2 será uma variável aleatória qui-quadrado com k graus de liberdade. Note que o número de graus de liberdade para essa variável qui-quadrado é igual ao número de variáveis independente no modelo. É possível mostrar que SQ_E/σ^2 é uma variável aleatória qui-quadrado, com $n-k-1$ graus de liberdade, e que SQ_E e SQ_R são independentes. A estatística de teste para $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ é dada por (23):

$$F_0 = \frac{SQ_R/k}{SQ_E/\ (n-k-1)} = \frac{MQ_R}{MQ_E} = \frac{V(\hat{Y})}{V(Y)} \quad (23)$$

Devemos rejeitar H_0 se o valor calculado da estatística de teste (23) for maior que $f_{1-\alpha;\,k;\,n-k-1}$, obtido em uma tabela da distribuição F de Fisher. O valor α indica o nível de significância do teste, usualmente adotado como 5%.

3.6 Testes para os coeficientes individuais de regressão

Segundo Montgomery e Runger (2012a), frequentemente testamos hipóteses para os coeficientes individuais de regressão. Estes testes são úteis na determinação do valor potencial de cada uma das variáveis independentes no modelo de regressão. Desta forma, o modelo pode ser mais efetivo com a inclusão de variáveis adicionais ou com a retirada de alguma variável independente atualmente inserida no modelo.

A hipótese para testar se um coeficiente individual de regressão, como β_j, é igual a zero como expresso em (24):

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases} \tag{24}$$

A estatística de teste para essa hipótese é dada por (25):

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{25}$$

Em que C_{jj} é o elemento da diagonal de $(X'X^{-1})$ correspondente a $\hat{\beta}_j$. O denominador da equação é o erro-padrão do coeficiente $\hat{\beta}_j$. A hipótese nula $H_0: \beta_j = 0$ será rejeitada se $|t_0| > t_{1-\frac{\alpha}{2},n-k-1}$. Denominamos como teste parcial ou marginal, pois o coeficiente de regressão $\hat{\beta}_j$ depende de todos as outras variáveis independentes x_i $(i \neq j)$ que estão no modelo.

Se $H_0: \beta_j = 0$ não for rejeitada, indicará que a variável independente x_i pode ser retirada do modelo com o nível de significância especificado. A inclusão de uma variável

a um modelo de regressão sempre aumenta a soma dos quadrados da regressão e diminuí a soma dos quadrados do erro.

3.7 Análise dos resíduos do modelo de regressão ajustado

Segundo NETO (2003), a análise dos resíduos é uma das etapas mais importantes na definição da qualidade de ajuste de um modelo de regressão. É necessário avaliar se os resíduos são homocedásticos, ou seja, tenham mesma variância, se são provenientes de uma distribuição normal com média zero e se são não-correlacionados. Quando os erros ou desvios do modelo são correlacionados, o modelo de regressão não é o modelo adequado para traduzir a relação de dependência. A correlação entre os erros aparece com frequência associada a dados coletados ao longo do tempo. Por isso, é conveniente proceder a uma análise gráfica dos dados e também dos resíduos, que possa detectar uma tendência a distribuição normal.

A análise gráfica da dispersão resíduos dos dados em torno da média zero é importante para verificar se os dados se distribuem aleatoriamente, e se a distribuição é homogênea, visualizando a presença de homocedasticidade no modelo. A normalidade dos dados pode ser avaliada através do gráfico histograma e de testes estatísticos específicos para testar a aderência da distribuição normal aos resíduos.

Os resíduos do modelo de regressão múltipla, definidos por $e_{\mathrm{i}=} y_i - \hat{y}_i$, desenvolvem um importante papel no julgamento da adequação do modelo, da mesma forma que para a regressão linear simples. Os gráficos de resíduos são úteis para a interpretação do modelo. Torna-se necessário também analisar os gráficos dos resíduos das variáveis que não estão presentes do modelo, mas que sejam possíveis candidatas à inclusão no modelo. Padrões de comportamento nos gráficos de resíduos indicam que o modelo por ser melhorado pela adição das variáveis candidatas.

Tanto na regressão linear simples quanto na regressão múltipla, as suposições do modelo ajustado precisam ser validadas para que os resultados sejam confiáveis. Chamamos de análise dos resíduos um conjunto de técnicas utilizadas para investigar a adequabilidade de um modelo de regressão com base nos resíduos. O resíduo (e_{i}) é

dado pela diferença entre a variável resposta observada (Y_i) e a variável resposta estimada ($\hat{Y}_i$), isto é (26):

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_p x_{ki} \qquad i = 1, \ldots, n \tag{26}$$

O objetivo principal da análise dos resíduos é que, se o modelo for apropriado, os resíduos devem refletir as propriedades impostas pelo termo de erro do modelo. Tais suposições são $Y = X\beta + \in$, em que $\in = (\in_1, \in_2, \ldots, \in_n)$, com:

i. $\in_i \ e \ \in_j$ são independentes (i $\neq$ j)
ii. $Var(\in_i) = \sigma^2$ (constante)
iii. $\in_i \sim N(0, \sigma^2)$ (normalidade)
iv. Modelo é linear
v. Não existir outliers (pontos atípicos) influentes.

Na regressão múltipla, além das suposições listadas acima, precisamos diagnosticar colinearidade e multicolinearidade entre as variáveis de entrada para que a relação existente entre elas não interfira nos resultados, causando inferências errôneas ou pouco confiáveis.

As técnicas utilizadas para verificar as suposições descritas acima podem ser informais, utilizando gráficos, ou formais, utilizando testes. As técnicas gráficas, por serem visuais, podem ser subjetivas e por isso técnicas formais são mais indicadas para a tomada de decisão. O ideal é combinar as técnicas disponíveis, tanto formais quanto informais, para o diagnóstico de problemas nas suposições do modelo. Algumas técnicas gráficas para análise dos resíduos são:

- **Gráfico dos resíduos versus valores ajustados:** verifica a homoscedasticidade do modelo, isto é, σ^2constante.
- **Gráfico dos resíduos versus a ordem de coleta dos dados:** avalia a hipótese de independência dos dados.
- **Papel de probabilidade normal:** verifica a normalidade dos dados.

- **Gráfico dos resíduos studentizados versus valores ajustados:** verifica se existem outliers em Y.
- **Gráfico dos resíduos padronizados versus valores ajustados:** verifica se existem outliers em Y.
- **Gráfico do leverage (diagonal da matriz H):** verifica se existem outliers em X.

Para a análise formal dos resíduos, podemos realizar os seguintes testes:

- **Testes de normalidade**: em que detalhes estão contidos no conteúdo de Inferência.
- **Teste de Durbin-Watson:** para testar independência dos resíduos.
- **Teste de Breusch-Pagan e Goldfeld-Quandt**: para testar se os resíduos são homoscedásticos.
- **Teste de falta de ajuste**: para verificar se o modelo ajustado é realmente linear.

Maiores detalhes sobre a análise de resíduos pode ser observada em Gujarati (2004). Dado a dificuldade de que todas as hipóteses da componente erro sejam satisfeitas, iniciaremos na próximo seção uma metodologia que poderá ser aplicada em situações em que os resíduos são não normais e/ou a variância é heterocedástica.

4. BOOTSTRAP NÃO PARAMÉTRICO POR PARES

Na estatística o bootstrap é um método de reamostragem proposto por Bradley Efron em 1979. Este método é utilizado para aproximar a distribuição de uma estatística baseado em reamostragens de uma amostra aleatória e permite assim estimarmos o erro padrão de estatísticas bem como construir intervalos de confiança ou realizar testes de hipóteses sobre parâmetros de interesse. Segundo Chernick e LaBudde (2011) o impacto de Efron (1979) é melhor expressa em Davison e Hinkley (1997) que escreveu:

> "A publicação em 1979 de Bradley Efron foi o primeiro artigo sobre métodos bootstrap, o que foi um grande acontecimento na estatística que ao mesmo tempo sintetizava algumas das primeiras ideias de reamostragem e estabelecia uma nova estrutura para análise estatística baseada em simulação. A ideia de substituir aproximações complicadas e muitas vezes imprecisas a viciadas, variância e outras medidas de incerteza por meio de simulação computacionais captou a imaginação de pesquisadores teóricos e usuários de métodos estatísticos".

O procedimento Bootstrap utilizado neste trabalho apresenta as seguintes etapas:

Etapa 1: Considere uma amostra aleatória de tamanho *n* da variável resposta Y e das variáveis independentes X em um estudo de regressão linear múltipla com *k* variáveis independentes. Insira os dados de Y e X em uma matriz Z. Crie uma variável contadora j sendo inicialmente igual a 1,e outra variável B sendo o número de reamostras que será realizado. No caso do presente trabalho o número B de reamostragem será igual a 1000;
Etapa 2: Gere uma amostra aleatória de tamanho *n* das *n* linhas da matriz Z com reposição. Com esta amostra estime por mínimos quadrados ou máxima verossimilhança os parâmetros da Regressão múltipla, isto é, $\hat{\beta}_i;\ \mathrm{i} = 0,1,\dots,\mathrm{k}.$
Etapa 3: Arquive $\hat{\beta}_k$ na j-ésima linha de uma matriz M que contém $k+1$ colunas;
Etapa 4: Se $j < B$ então vá para Etapa 2 e caso contrário vá para Etapa 5.

Etapa 5: Calcule os percentis 2,5% e 97,5% de cada coluna da matriz M. Estes representarão o intervalo de 95% de confiança para $\hat{\beta}_k$. Se o interval de confiança conter o valor zero então não poderemos rejeitar, ao nível de confiança 95%, que $\beta_i = 0$.

4.1. Bootstrap não Paramétrico aplicado ao Modelo de Regressão utilizando o Excel

Nesta seção explicaremos o uso da metodologia Bootstrap não paramétrico aplicado ao modelo de regressão múltipla. Para melhor entendimento do leitor adaptamos o exemplo 12-14 Qualidade do Vinho constante no livro Estatística e Probabilidade para Engenheiros (MONTGOMERY; RUNGER, 2003). A variável resposta é y (qualidade) e desejamos encontrar a melhor equação de regressão que relaciona qualidade aos três outros parâmetros, x_2 - aroma, x_4 - sabor, x_5 - afinação e está descrito na Tabela 1. Observe na Figura 1 que as probabilidades de significância para o parâmetro de todas as variáveis são significantes, isto é, rejeitamos ao nível de confiança 95% que são nulos. Além disso, a Figura 2 retrata a análise de resíduos indicando que os testes de hipóteses descritos na Figura 1 são aceitáveis. Observe que os resultados na Figura 1 só são corretos se a análise de resíduos indica que a componente erro do modelo possui distribuição normal, variável constante e erros não correlacionados. Na próxima etapa iremos verificar a significância dos parâmetros das variáveis sem a necessidade da validação de que a componente erro possua distribuição normal, variância constante utilizando o método Bootstrap.

Tabela 1: Dados da Qualidade do Vinho

Ordem	Y	x2	x4	x5
1	9,8	3,3	3,1	4,1
2	12,6	4,4	3,5	3,9
3	11,9	3,9	4,8	4,7
4	11,1	3,9	3,1	3,6
5	13,3	5,6	5,5	5,1
6	12,8	4,6	5	4,1
7	12,8	4,8	4,8	3,3
8	12	5,3	4,3	5,2
9	13,6	4,3	3,9	2,9
10	13,9	4,3	4,7	3,9
11	14,4	5,1	4,5	3,6
12	12,3	3,3	4,3	3,6
13	16,1	5,9	7	4,1
14	16,1	7,7	6,7	3,7
15	15,5	7,1	5,8	4,1
16	15,5	5,5	5,6	4,4
17	13,8	6,3	4,8	4,6
18	13,8	5	5,5	4,1
19	11,3	4,6	4,3	3,1
20	7,9	3,4	3,4	3,4
21	15,1	6,4	6,6	4,8
22	13,5	5,5	5,3	3,8
23	10,8	4,7	5	3,7
24	9,5	4,1	4,1	4
25	12,7	6	5,7	4,7
26	11,6	4,3	4,7	4,9
27	11,7	3,9	5,1	5,1
28	11,9	5,1	5	5,1
29	10,8	3,9	5	4,4
30	8,5	4,5	2,9	3,9
31	10,7	5,2	5	6
32	9,1	4,2	3	4,7

33	12,1	3,3	4,3	4,5
34	14,9	6,8	6	5,2
35	13,5	5	5,5	4,8
36	12,2	3,5	4,2	3,3
37	10,3	4,3	3,5	5,8
38	13,2	5,2	5,7	3,5

Fonte: Tabela elaborada pelo autora

Figura 1: Saída do Minitab após análise de regressão dos dados contidos na tabela 1

Regression Analysis: Y versus x2; x4; x5

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	108,935	36,312	26,92	0,000
x2	1	6,603	6,603	4,90	0,034
x4	1	25,689	25,689	19,05	0,000
x5	1	6,999	6,999	5,19	0,029
Error	34	45,853	1,349		
Total	37	154,788			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1,16131	70,38%	67,76%	63,79%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6,47	1,33	4,85	0,000	
x2	0,580	0,262	2,21	0,034	2,21
x4	1,200	0,275	4,36	0,000	2,19
x5	-0,602	0,264	-2,28	0,029	1,04

Regression Equation

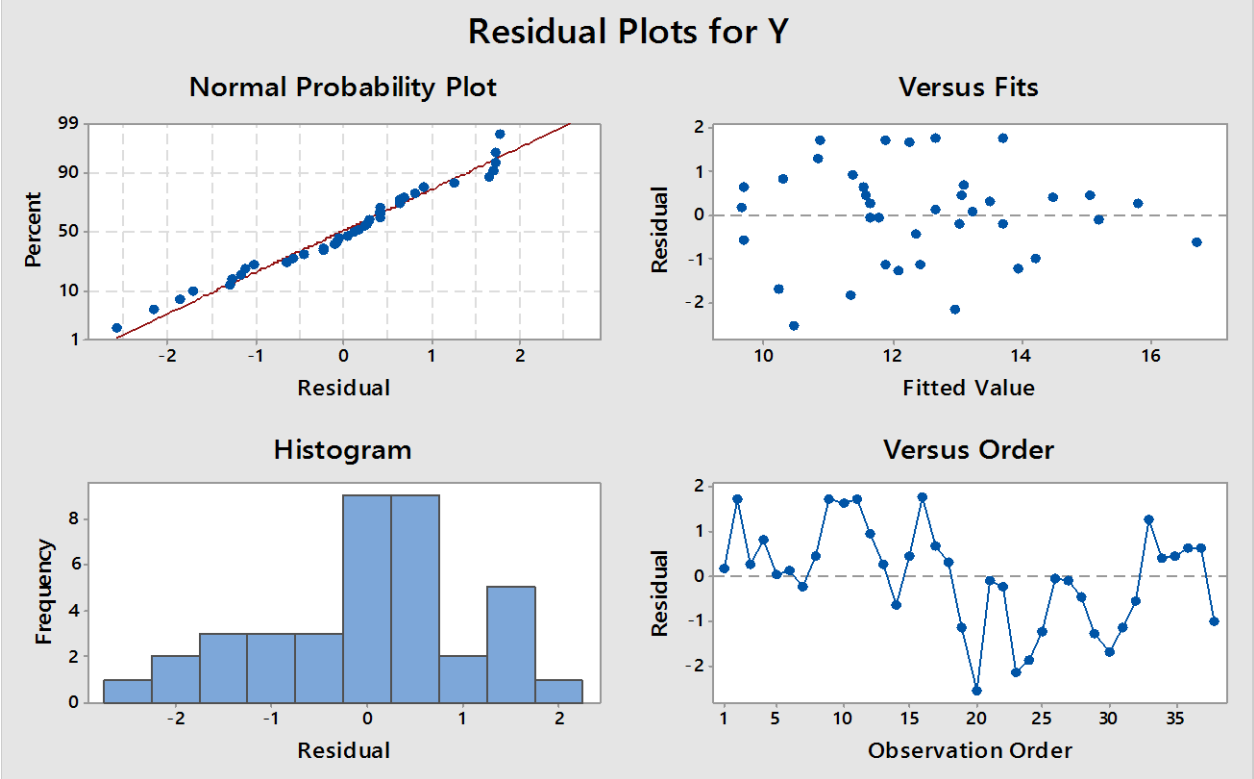
Y = 6,47 + 0,580 x2 + 1,200 x4 - 0,602 x5

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	Resid	Std Resid	
20	7,900	10,471	-2,571	-2,34	R

R Large residual

Figura 2: Resíduos dos dados contidos na tabela após análise de regressão



Observamos que no gráfico de probabilidade podemos supor que os resíduos estão distribuídos normalmente. No histograma dos resíduos não observamos outliers. No gráfico resíduos versus valores ajustados podemos entender que há uma variância constante. No gráfico resíduos versus ordem de dados, podemos entender que os resíduos não são correlacionados, pois os pontos não parecem ter uma tendência e por isso temos indícios de independência dos erros.

4.2 Segunda etapa: Gerando as reamostras

Retiramos da amostra inicial (Tabela 1) 1000 amostras (linhas) de tamanho 38. Para fazer isto, geramos na coluna A do excel 38.000 valores entre 1 e 38. Os primeiros 38 valores gerados indicarão a primeira amostra, os seguintes a segunda amostra e assim por diante. Como ilustrado na Figura 3, utilizamos a função ALEATORIOENTRE do Excel na coluna B, denominada simulação dos blocos, para que seja possível

selecionar de forma aleatória uma das 38 observações dos dados da Tabela 1, Dados da Qualidade do Vinho.

Para obtermos os dados das observações da Tabela 1, que também estão contidos na segunda aba do Excel utilizado, denominada Exercício Montgomery, utilizamos a função DESLOC nas colunas de C a F, ilustrado na Figura 4, que possui como objetivo retornar uma referência para um intervalo, que é um número especificado de linhas e colunas de uma célula ou intervalo de células.

Figura 3: Demonstração da utilização da função ALEATORIOENTRE

A	B	C	D	E	F
		Simulação			
Ordem	Simulação Blocos	Y	x2	x4	x5
1	17	=DESLOC('Exercicio Montgomery'!$C$1;B3;0)			
2	27	11,7	3,9	5,1	5,1
3	31	10,7	5,2	5	6
4	13	16,1	5,9	7	4,1
5	27	11,7	3,9	5,1	5,1
6	35	13,5	5	5,5	4,8
7	1	9,8	3,3	3,1	4,1
8	27	11,7	3,9	5,1	5,1
9	11	14,4	5,1	4,5	3,6
10	31	10,7	5,2	5	6
11	20	7,9	3,4	3,4	3,4
12	4	11,1	3,9	3,1	3,6

Figura 4: Demonstração da utilização da função DESLOC

A	B	C	D	E	F
		Simulação			
Ordem	Simulação Blocos	Y	x2	x4	x5
1	17	=DESLOC('Exercicio Montgomery'!$C$1;B3;0)			
2	27	11,7	3,9	5,1	5,1
3	31	10,7	5,2	5	6
4	13	16,1	5,9	7	4,1
5	27	11,7	3,9	5,1	5,1
6	35	13,5	5	5,5	4,8
7	1	9,8	3,3	3,1	4,1
8	27	11,7	3,9	5,1	5,1
9	11	14,4	5,1	4,5	3,6
10	31	10,7	5,2	5	6
11	20	7,9	3,4	3,4	3,4
12	4	11,1	3,9	3,1	3,6

4.3 Terceira etapa: Preparação das 1000 amostras

Na Figura 5 temos a coluna G que indica qual é a simulação entre as 1000 realizadas. As colunas H e I indicam respectivamente o início e fim de cada amostra obtidas nas colunas C a F.

Figura 5: Demonstração da utilização da separação da simulação em blocos

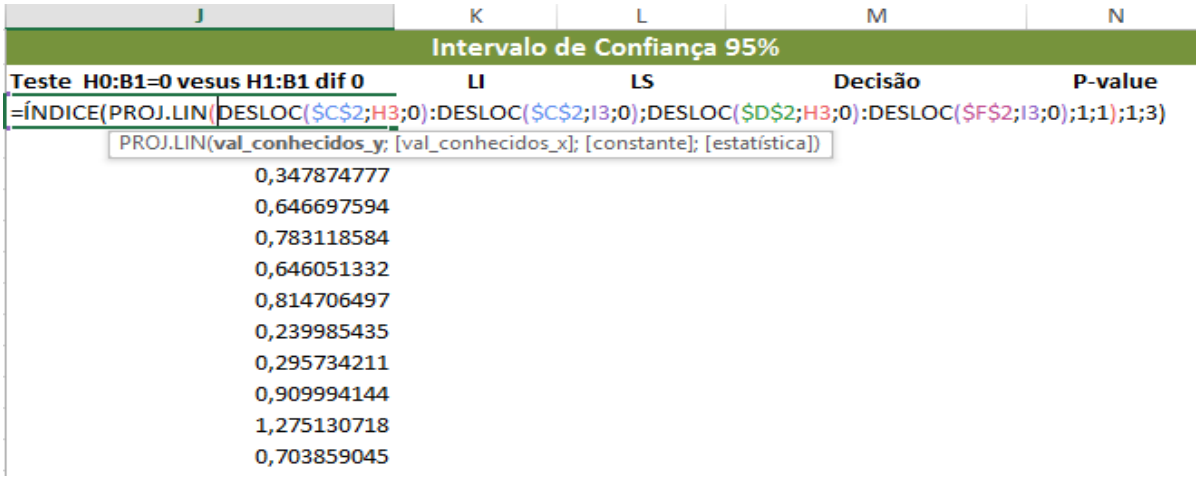
G	H	I
	Bloco	
Simulação	Início	Fim
1	1	38
2	39	76
3	77	114
4	115	152
5	153	190
6	191	228
7	229	266
8	267	304
9	305	342
10	343	380
11	381	418
12	419	456

4.4 Quarta etapa: Gerando as estimativas $\widehat{\boldsymbol{\beta}}$ e realizando teste de hipóteses

Nesta etapa testaremos a hipótese $H_0: \hat{\beta}_1 = 0$ vesus $H_1: \hat{\beta}_1 \neq 0$ .Na coluna J arquivamos as estimativas $\hat{\beta}_1$ para as 1000 amostras definidas nas colunas H e I. Para esta etapa utilizamos a função PROJ.LIN em conjunto com a função Índice do excel.

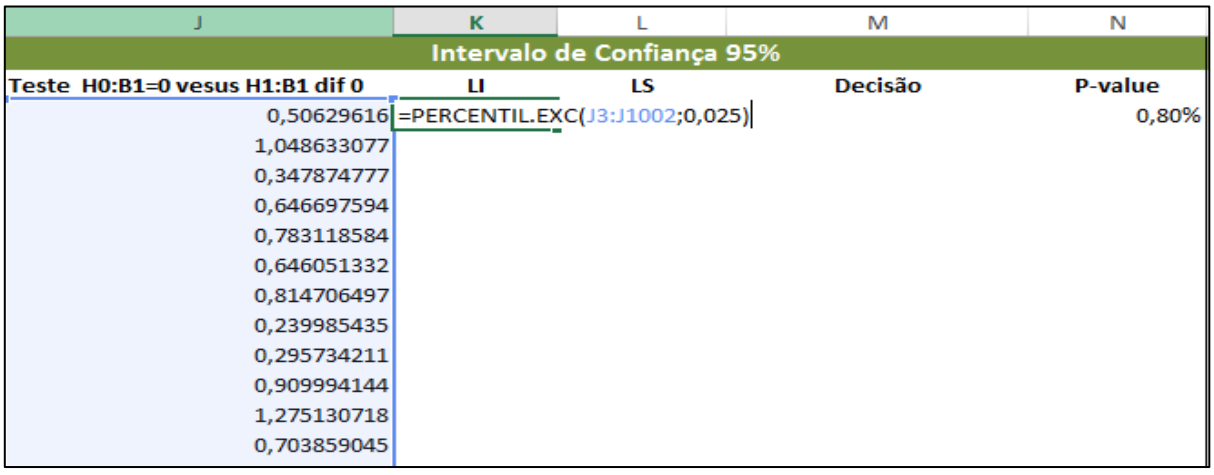
A função PROJ.LIN calcula as estimativas usando o método quadrados mínimos e como trata-se de uma função matricial precisamos da função Índice para indicar o elemento desejado. No caso do exemplo discutido a estimativa $\hat{\beta}_1$ encontra-se na primeira linha e terceira coluna.

Figura 6: Demonstração da utilização das funções ÍNDICE e PROJ.LIN



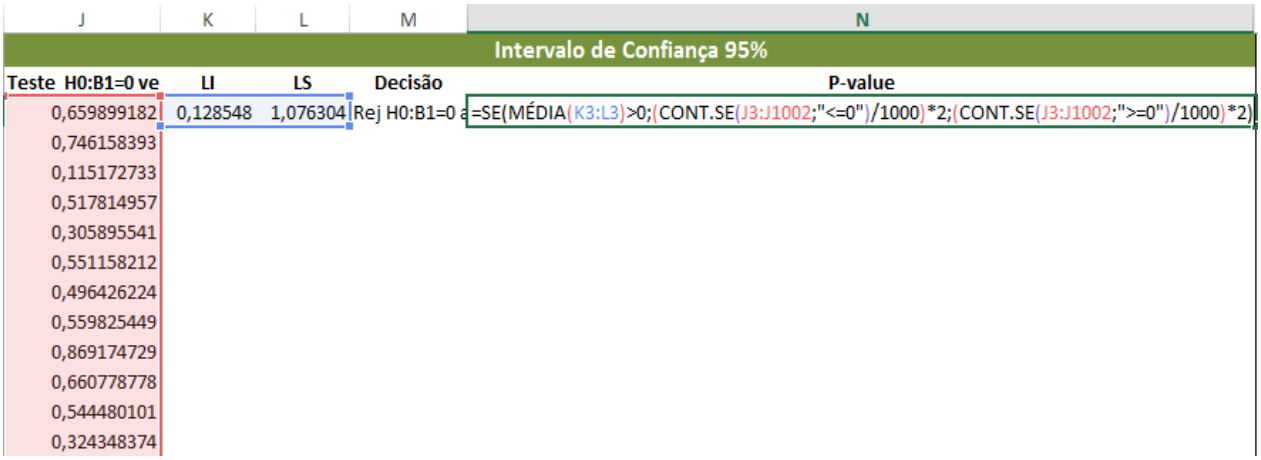
Com as estimativas de $\hat{\beta}_1$ arquivadas na coluna J podemos realizar o teste de hipóteses considerando que o limite superior de confiança é o percentil 97,5% e o limite inferior de confiança é o percential 2,5%. A função PERCENTIL.EXC retorna o k-ésimo percentil de valores em um intervalo. Utilizamos esta função para estabelecer um limite de confiança, ou seja, definir o limite inferior(LI) =PERCENTIL.EXC(J3:J1002;0,025), e o limite superior(LS) =PERCENTIL.EXC(J3:J1002;0,975). Se o intervalo conter o zero então não rejeitamos $H_0: \hat{\beta}_1 = 0$ e caso contrário rejeitamos.

Figura 7: Demonstração da utilização da função PERCENTIL.EXC



Como obtivemos LI=0,191654675 e LS=1,052709972, rejeitamos $H_0: \hat{\beta}_1 = 0$ ao nível de confiança 95%, conforme demonstrado na figura 8.

Figure 8: Cálculo do P-value



J	K	L	M	N
		Intervalo de Confiança 95%		
Teste H0:B1=0 ve	LI	LS	Decisão	P-value
0,659899182	0,128548	1,076304	Rej H0:B1=0 a	=SE(MÉDIA(K3:L3)>0;(CONT.SE(J3:J1002;"<=0")/1000)*2;(CONT.SE(J3:J1002;">=0")/1000)*2)
0,746158393				
0,115172733				
0,517814957				
0,305895541				
0,551158212				
0,496426224				
0,559825449				
0,869174729				
0,660778778				
0,544480101				
0,324348374				

O valor de p (p-value) é calculado em relação à média da distribuição das estimativas arquivadas na coluna J. Se a média da coluna J for maior que zero então o valor de p será igual ao percentual de dados da coluna J inferiores a zero multiplicado por dois. Caso contrário será igual ao percentual de valores da coluna J superiores a zero multiplicado por dois. Para o cálculo do valor de p, construímos uma regar condicional no excel com a utilização das funções SE, MÉDIA e CONT.SE.

4.5 Quinta etapa

Seguimos os mesmos passos da quarta etapa para testar as hipóteses $H_0: \beta_2 = 0$ *versus* $H_1: \beta_2 \neq 0$ e $H_0: \beta_3 = 0$ *versus* $H_1: \beta_3 \neq 0$.

5 ANÁLISE DOS RESULTADOS

5.1 $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$

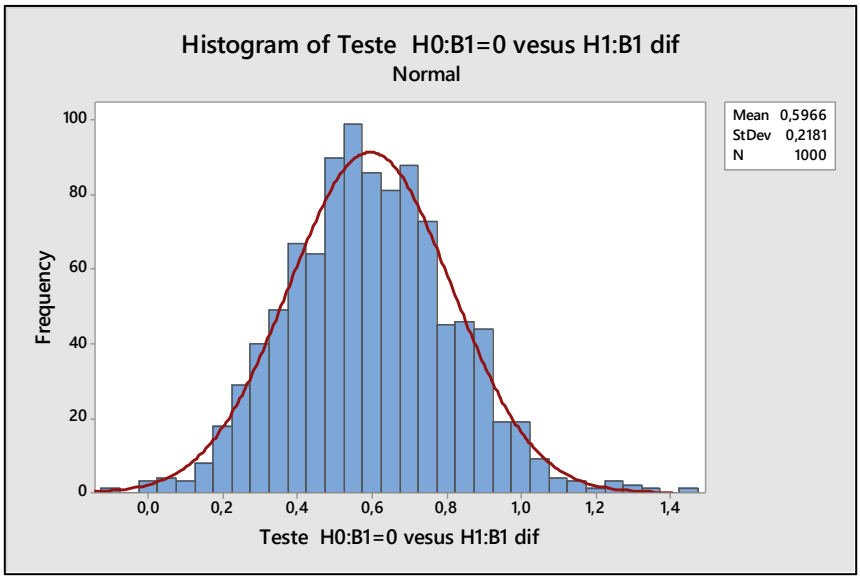
Para otimizarmos o processo de análise do resultado, criamos uma regra no excel onde a decisão será demonstrada na coluna M. Considerando os dados apresentados e analisando o resultado, LI = 0,188524 e LS = 1,023529, rejeitamos $H_0: \hat{\beta}_1 = 0$ ao nível de confiança 95%, conforme demonstrado na figura 9 .

Figura 9: Resultado do teste de hipótese $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$

J	K	L	M	N
Intervalo de Confiança 95%				
Teste H0:B1=0 vesus H1:B1 dif 0	**LI**	**LS**	**Decisão**	**P-value**
0,635038206	0,188524	1,023529	Rej H0:B1=0 ao nível 95%	0,40%

O histograma foi gerado no Minitab apenas para possibilitar uma melhor visualização do gráfico, apesar de ser possível realizar o gráfico no excel, o mesmo não apresenta um recurso visual adequado para este trabalho.

Figura 10: Histograma gerado no Minitab - $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$



Analisando o gráfico das amostras, podemos perceber que há normalidade nos dados apresentados.

5.2 $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$

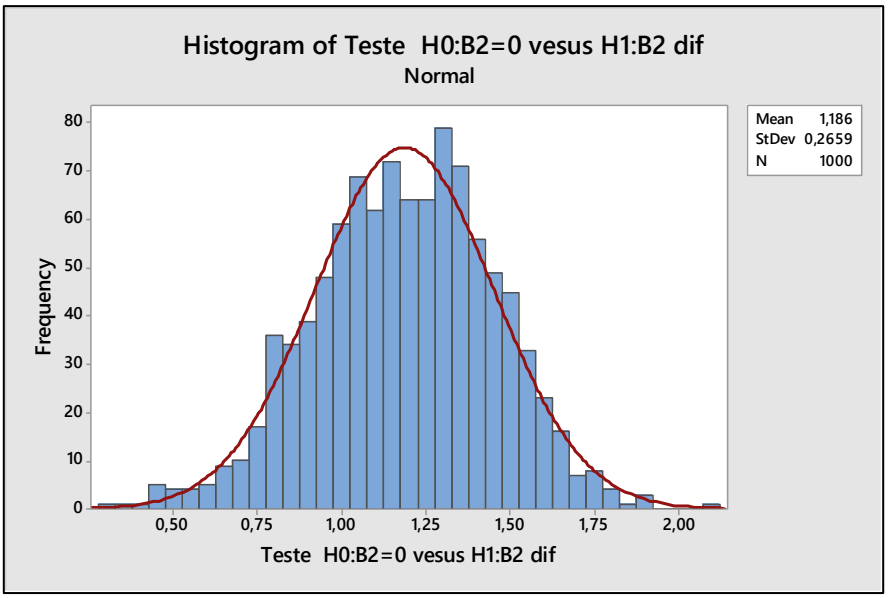
Considerando os dados apresentados e analisando o resultado, LI = 0,63951597 e LS = 1,673164505, rejeitamos $H_0: \hat{\beta}_2 = 0$ ao nível de confiança 95%, conforme demonstrado na figura 11.

Figura 11: Resultado do teste de hipótese $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$

O	P	Q	R	S
Intervalo de Confiança 95%				
Teste H0:B2=0 vesus H1:B2 dif 0	**LI**	**LS**	**Decisão**	**P-value**
1,016960364	0,63951597	1,673164505	Rej H0:B2=0 ao nível 95%	0,00%

Analisando o gráfico das amostras, podemos perceber que há normalidade nos dados apresentados.

Figura 12: Histograma gerado no Minitab - $H_0: \beta_2 = 0$ versus $H_1: \beta_2 \neq 0$



5.3 $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$

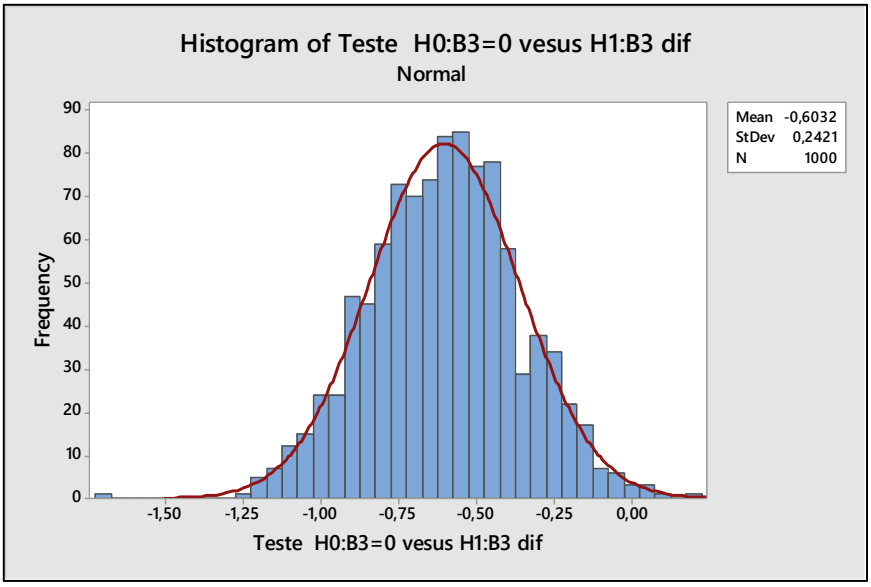
Considerando os dados apresentados e analisando o resultado, LI = -1,081026 e LS = -0,143359, rejeitamos $H_0: \hat{\beta}_3 = 0$ ao nível de confiança 95%, conforme demonstrado na figura 13.

Figura 13: Resultado do teste de hipótese $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$

T	U	V	W	X
Intervalo de Confiança 95%				
Teste H0:B3=0 vesus H1:B3 dif 0	LI	LS	Decisão	P-value
-0,878688564	-1,081026	-0,143359	Rej H0:B3=0 ao nível 95%	1,40%

Analisando o gráfico das amostras, podemos perceber que há normalidade nos dados apresentados.

Figura 14: Histograma gerado no Minitab - $H_0: \beta_3 = 0$ versus $H_1: \beta_3 \neq 0$



6 TESTE PARA VERIFICAR $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Os testes realizados na seção 5 só fazem sentido se rejeitarmos $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Nesta seção realizaremos o teste em uma perspectiva não paramétrica e utilizando o conceito do procedimento Bootstrap. Se $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ for verdadeira então os valores de Y não serão influenciados pelas variáveis independentes. Neste cenário podemos alocar amostras aleatórias de Y, sendo que no trabalho utilizamos 38 valores, as 1.000 amostras de variáveis independentes obtidas com o procedimento explicado na seção 5. Em seguida calculamos o valor de Fcal, utilizando a expressão citada em (23), para cada uma das 1.000 amostras das variáveis independentes e o respectivo vetor Y obtido aleatoriamente. O valores de F's obtidos constituem-se na distribuição de F com a hipótese de que $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ é verdadeiro. Com tais valores calculamos o limite superior (LS) de forma que o percentual de dados acima desse valor seja 5%, por exemplo.

Figura 15: Resultado do teste de hipótese $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

Teste F Não Paramétrico

Aleatório	Y aleatório	Teste H0:B1=B2=B3=0 vesus cc	Fcal	LS	Decisão	P-value
34	14,9	0,068985254	26,92486	2,977435428	Rej H0:B1=B2=B3=0 ao nível 95%	0,00%
13	16,1	0,606735852				
24	9,5	0,64030155				
6	12,8	2,029081231				
2	12,6	1,282952275				
20	7,9	1,379815597				
30	8,5	0,186437763				
11	14,4	1,849087382				
29	10,8	0,105352006				
16	15,5	1,187346809				
10	13,9	1,018341332				
10	13,9	1,179555704				
25	12,7	0,756942622				
19	11,3	2,765861497				
35	13,5	2,014288494				
22	13,5	0,102168854				
20	7,9	1,103688553				
10	13,9	0,823341482				

Aleatório	ALEATÓRIOENTRE(1;38)
Y aleatório	DESLOC('Exercicio Montgomery'!$C$1;Y3;0)
Teste H0:B1=B2=B3=0 vesus cc	ÍNDICE(PROJ.LIN(DESLOC($Z$2;H3;0):DESLOC($Z$2;I3;0);'Exercicio Montgomery'!$D$2:$F$39;1;1);4;1)
Fcal	ÍNDICE(PROJ.LIN('Exercicio Montgomery'!C2:C39;'Exercicio Montgomery'!D2:F39;1;1);4;1)
LS	PERCENTIL.EXC(AA3:AA1002;0,95)
Decisão	SE(AB3>=AC3;"Rej H0:B1=B2=B3=0 ao nível 95%";"Não Rej H0:B3=0 ao nível
P-value	CONT.SE(AA3:AA1002;">="&AB3)/1000

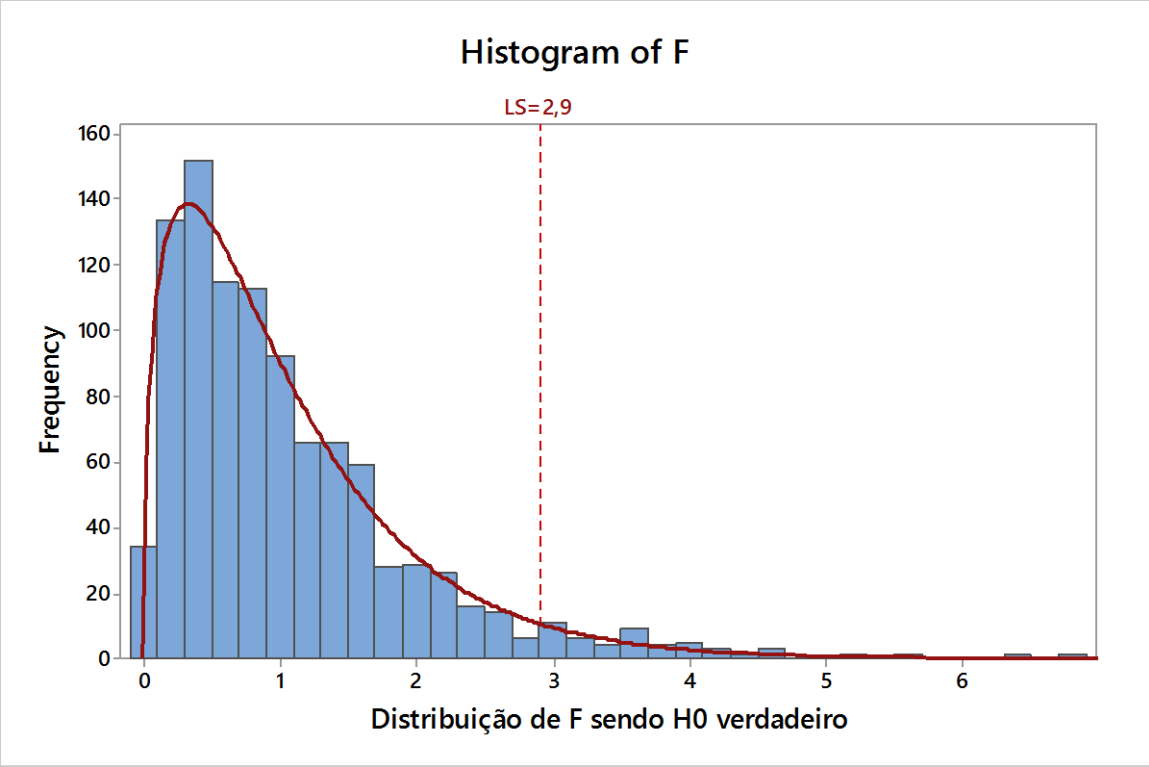
Finalmente calculamos o valor de Fcal, utilizando a expressão citada em (23), para os dados originais. Se Fcal > LS então rejeitamos $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ ao nível

de significância de 5%. Caso contrário não rejeitamos $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ ao nível de significância de 5% e concluímos que pelo menos uma variáveis interfere na variável independente.

A Figura 15 ilustra todo o procedimento feito no Excel com as respectivas funções necessárias na coluna AF. A Figura 16 mostra a distribuição dos F's obtidos na coluna AC.

Para este teste, o histograma também foi gerado no Minitab, também com o objetivo de obter uma melhor visualização dos dados. Neste gráfico, podemos verificar o ponto citado do limite superior (LS) e o comportamento dos demais dados, confirmado os resultados obtidos com as fórmulas do excel.

Figura 16: Histograma gerado no Minitab - $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$



7 CONCLUSÃO

O presente trabalho buscou exemplificar a utilização do método bootstrap não paramétrico com o objetivo de proporcionar uma melhor compreensão do método de forma didática pelo uso da ferramenta excel. A utilização da ferramenta excel está presente no cotidiano das empresas para diversas atividades de controle, que necessitam de registrar diversas informações e compilar dados para arquivamento ou análise e tomada de decisão, como também para setores mais estratégicos que precisam acompanhar a evolução do cenário em tempo mais real. Estas atividades muitas vezes exigem programação e pode ser um dificultador para muitos profissionais. No entanto, o Excel é uma ferramenta que pode permitir que tais profissionais consigam realizar suas tarefas mesmo sem um grande conhecimento de programação. Dessa forma, entendemos que diversos segmentos do mercado estão habituados com a utilização da ferramenta Excel, sendo parte integrante da rotina de trabalho diário. Neste sentido, o trabalho apresentou um procedimento realizado na ferramenta excel que permite implementar uma metodologia denominada bootstrap que usualmente demanda programação.

Além disso, o bootstrap demanda o uso de pacotes estatísticos específicos muitas vezes não presentes nas empresas. Assim, a implementação do bootstrap não paramétrico no excel viabiliza o aumento da utilização destes método, considerando que é necessário apenas o conhecimento de fórmulas do excel. Ou seja, não foi necessário instalar nenhum programa específico para suporte neste trabalho, diferentemente dos *softwares* estatísticos que em sua maioria possuem custo para as organizações e/ou demandam capacidade técnica mais especializada para o processo de programação necessário. Desta forma, utilizar-se do excel para tal funcionalidade possibilita ampliar o acesso dos profissionais neste recurso que possui menor custo financeiro e baixa complexidade de conhecimento.

Neste trabalho utilizamos o bootstrap não paramétrico para uso da técnica estatística de regressão linear múltipla. A vantagem é que a metodologia funciona bem mesmo que a componente erro do modelo não possua distribuição normal e tenha

variância constante. A implementação do método no excel é de baixa complexidade e todas as fórmulas necessárias foram explicitadas no trabalho.

Torna-se pertinente realizar a avaliação de regressão linear múltipla utilizando os testes por meio do método tradicional e paralelamente com o uso do bootstrap não paramétrico. Se os resultados forem similares então não existem problemas e podemos adotar o método tradicional. Se forem discrepantes então o método bootstrap não paramétrico seria mais adequado para a análise de regressão múltipla pois não demanda as hipóteses de normalidade e variância constante.

REFERÊNCIAS

BARBETTA, Pedro Alberto. **Estatística aplicada às ciências sociais**. 7. ed. Florianópolis: Editora da UFSC, 2007. 315 p.

CHARNET, R.; et al. **Análise de modelos de regressão linear com aplicações**. 1. ed. Campinas: Editora da Unicamp, 1999. 356 p.

CHERNICK, Michael R.; LABUDDE, ROBERT A. **An introduction to bootstrap methods with applications to R**. Nova Jersey: WILEY, 2011.

DAVISON, A.C; HINKLEY, D.V; **Bootstrap methods and their application**. Cambridge University Press. Cambridge, 2013.

EFRON, B. Bootstrap methods: another look at the jackknife. **The Annals of Statistics. Stanford University**. v. 7. n. 1. p. 1–26. maio. 1979.

GUJARATI, D.N., **Econometria Básica**. 3. ed. São Paulo: Markon Books, 2004.

MARTINS, G. **A. Estatística Geral e Aplicada.** 3. Ed. São Paulo: Atlas,2005.

MILONE, Giuseppe; ANGELINI, Flavio. **Estatística aplicada**. São Paulo: Atlas, 1995. 286 p.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística e Probabilidade para Engenheiros**. 2. ed. Rio de Janeiro: LTC, 2012.

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

NETO, A. P. **Curso de engenharia de avaliação imobiliária**: fundamentos e aplicação da estatística inferencial, Belo Horizonte/MG, 2003.