UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA

CLOVIS LUIZ BAIERLE

MÉTODOS BOOTSTRAP EM REGRESSÃO LINEAR SIMPLES

TRABALHO DE CONCLUSÃO DE CURSO

TOLEDO
2019

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA

CLOVIS LUIZ BAIERLE

MÉTODOS BOOTSTRAP EM REGRESSÃO LINEAR SIMPLES

Trabalho de Conclusão de Curso apresentado ao Curso de Licenciatura em Matemática da Universidade Tecnológica Federal do Paraná, Câmpus Toledo, como requisito parcial à obtenção do título de Licenciado em Matemática.

Orientador(a): Gustavo Henrique Dalposso

TOLEDO
2019

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
COORDENAÇÃO DO CURSO DE LICENCIATURA EM MATEMÁTICA

TERMO DE APROVAÇÃO

O Trabalho de Conclusão de Curso intitulado “Método Bootstrap em Regressão Linear Simples” foi considerado **APROVADO** de acordo com a ata nº __ de __/__/____.

Fizeram parte da banca examinadora os professores:

_____
Dr. Gustavo Henrique Dalposso (orientador)

_____
Dra. Daniela Trentin Nava

_____
Dra. Regiane Slongo Fagundes

TOLEDO
2019

AGRADECIMENTOS

Aos meus pais por estarem sempre por perto, me motivando e suportando meus momentos de estresse e impaciência. Obrigada por tudo, eu amo vocês.

A minha namorada Aline, por todo o apoio neste processo, me dando confiança e força para seguir em frente, dia após dia, e por ter sido parceira e paciente o tempo todo.

Ao meu amigo Bruno pelas conversas e trabalhos realizados nestes quatros anos de companheirismo. Sua motivação e insistência para comigo foram essenciais nessa jornada. Você é incrível!

A minha amiga Camila, por estes quatros anos de amizade, de apoio e motivação, a qual, não cansam de me lembrar do meu potencial. Levarei nossa amizade para o resto da vida. Gratidão sempre!

A minha amiga Tawine pelo apoio nesta jornada, nos trabalhos e conversas realizadas. Você é incrível!

Minha dupla de Estágio, Iasmim. Superamos dificuldades, choramos, rimos, amadurecemos. Sua força de vontade e determinação são contagiantes. Obrigada pelo companheirismo.

Ao meu orientador e exemplo de Professor, Gustavo Henrique Dalposso. Agradeço primeiramente por ter feito parte da minha caminhada acadêmica. Seus ensinamentos serão sempre levados por mim. Gratidão por ter aceito me orientar. A alegria que senti foi enorme e é uma honra poder carregar seu nome nesse trabalho tão importante. Obrigada pelas orientações, compreensão e paciência. Você é extraordinário!

Obrigada a minha banca, professoras Daniela Trentin Nava e Regiane Slongo Fagundes, por terem aceito o convite e por contribuírem com meu trabalho.

Aos demais professores da Universidade que também são inspiradores. Aprendi com vocês não somente dentro, mas também fora da sala de aula, obrigada por cada ensinamento.

Minha eterna gratidão a todos que de alguma outra forma contribuíram.

RESUMO

O ato de estabelecer uma equação matemática linear (uma reta) que descreva a relação entre duas variáveis se torna fundamental em diversas situações, como por exemplo, no caso em que duas variáveis medem aproximadamente a mesma coisa, mas uma delas é relativamente dispendiosa, ou difícil de lidar, enquanto que a outra não. Outra questão importante que pode ser abordada utilizando esta relação entre duas variáveis refere-se a predizer valores futuros de uma variável. Neste contexto, um método tradicional muito utilizado é conhecido como regressão linear, que através da minimização da soma dos quadrados dos resíduos, permite determinar estimativas do coeficiente angular e linear da reta procurada. Em uma análise de regressão linear, geralmente o pesquisador tem interesses em realizar inferências acerca dos parâmetros do modelo, no entanto, para fazer isto, é necessário requerer pressupostos que muitas vezes não podem ser assumidos. Dentre os principais pressupostos podemos citar o formato da distribuição dos erros do modelo, que deve ser normal ou também a existência de dados atípicos pois neste caso, o método de regressão atribui muito peso a estes dados, o que leva a mudança da orientação da reta e distorcendo as estimativas, acarretando em erros nos resultados esperados. Uma alternativa para os métodos tradicionais é o método de reamostragem bootstrap, o qual realiza inferências através de reamostras com reposição obtidas da amostra original. Sem a necessidade de assumir pressupostos, o método bootstrap permite quantificar a incerteza calculando os erros padrões e intervalos de confiança dos parâmetros desconhecidos. Nos estudos de regressão linear podem ser considerados dois métodos bootstrap: o método bootstrap dos pares e o método bootstrap dos resíduos. Neste contexto, este trabalho tem como objetivo comparar os intervalos de confiança dos parâmetros obtidos com métodos bootstrap com os obtidos utilizando a estatística clássica. Utilizando quatro conjuntos de dados obtidos na literatura, utilizou-se o software R para realizar os ajustes das retas de regressão e calcular os intervalos de confiança. Os resultados mostraram que os métodos bootstrap são uma viável alternativa para realizar inferências em modelos de regressão sem a necessidade de se verificar os pressupostos

Palavras-chave: Regressão linear. Intervalo de confiança. Reamostragem.

ABSTRACT

The act of establishing a linear mathematical equation (a line) that describes the relationship between two variables becomes fundamental in many situations, such as when two variables measure roughly the same thing, but one of them is relatively expensive, or hard to deal with, while the other doesn't. Another important issue that can be addressed using this relationship between two variables is to predict future values of one variable. In this context, a widely used traditional method is known as linear regression, which by minimizing the sum of the squares of the residuals, allows to determine estimates of the angular and linear coefficient of the sought line. In a linear regression analysis, the researcher is usually interested in making inferences about the model parameters, however, to do so requires assumptions that often cannot be assumed. Among the main assumptions we can mention the error distribution format of the model, which should be normal or also the existence of atypical data because in this case, the regression method gives a lot of weight to these data, which leads to a change in the orientation of the line. and distorting estimates, leading to errors in expected results. An alternative to traditional methods is the bootstrap resampling method, which makes inferences through replacement resamples obtained from the original sample. Without assuming assumptions, the bootstrap method allows quantifying uncertainty by calculating the standard errors and confidence intervals of unknown parameters. In linear regression studies two bootstrap methods can be considered: the pair bootstrap method and the residual bootstrap method. In this context, the objective of this work is to compare the confidence intervals of the parameters obtained with bootstrap methods with those obtained using the classical statistics. Using four data sets obtained in the literature, software R was used to perform regression line adjustments and to calculate confidence intervals. The results showed that bootstrap methods are a viable alternative to make inferences in regression models without the need to verify the assumptions.

Keywords: Linear Regression. Confidence Interval. Resampling.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1 INTRODUÇÃO

O ato de antever o futuro sempre encantou a humanidade. Saber o que vai acontecer antes de algo ocorrer pode propiciar melhor aproveitamento ou uma preparação antecipada de problemas. Talvez até mais importante que antecipar os resultados seja reconhecer o que pode interferir, permitindo se planejar. Por exemplo na epidemiologia, a necessidade de prever o futuro e, com base nisso, intervir nos processos do presente é mais que mera curiosidade. É, de fato, assunto de vida ou morte, pois a redução da carga de doenças na população depende da efetividade desse esforço (ANTUNES; CARDOSO, 2015).

Também, considerando a globalização e a competitividade no mercado financeiro, ocorre diariamente a busca por táticas que facilitam atingir os objetivos de forma a estarem sempre à frente de seus concorrentes. Focando sempre a redução dos custos de produção de forma que a qualidade dos produtos permaneça a mesma (FILHO, 2002).

Para solucionar estes problemas, a utilização de regressão linear é uma solução, pois a regressão é uma técnica utilizada na investigação da relação entre variáveis que surgem em problemas das mais variadas áreas da ciência. De uma forma geral, o investigador procura aferir a influência de uma variável explicativa X sobre o valor esperado de uma variável de resposta denominada habitualmente por Y (SILVA, 2016).

O qual, a partir de uma série de pontos representativos das variáveis que compõem um determinado fenômeno, uma função que o expresse matematicamente. A função obtida deve permitir com satisfatória segurança a realização de análises e projeções sobre o fenômeno estudado (FRANCO, 2006).

Portanto, determina-se intervalos de confiança (IC), o qual, permitem ter uma ideia indireta da qualidade da regressão. Para além de uma validação geral do modelo obtido, os IC podem servir para confirmar hipóteses de valores particulares para os parâmetros, estabelecidas por via teórica ou em anteriores experiências (MATOS, 1995). Deste modo, a utilização de intervalo de confiança, nos garante uma porcentagem de confiança dos valores obtidos ou analisados. O intervalo de confiança é estimado de um parâmetro de interesse de uma população. Desta maneira, ao invés de se estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. O quanto estas estimativas são prováveis será determinado pelo nível de confiança para Intervalos de confiança, usados para indicar a confiabilidade de uma estimativa (DALPOSSO, 2017). Cada intervalo de confiança está associado a um nível de confiança (1-α), que é um número que nos diz qual a probabilidade de o valor real estar dentro de intervalo.

O nível de confiança e a proporção de todas as amostras possíveis para as quais o intervalo de confiança abrange o valor real (NAVIDI, 2010). Dependendo do tamanho amostral, ou de pressupostos não possíveis de verificar, os intervalos de confiança podem ser viesados e desta forma, os métodos bootstrap constituem uma alternativa eficiente para a teoria usual. Visto que, além de ser livre de complexidades algébricas, possibilita a obtenção de intervalos de confiança sem a necessidade de pressupostos sobre a distribuição do estimador (CHERNICK; LABUDDE, 2011).

O método bootstrap é um procedimento de reamostragem bastante utilizada nas mais diversas situações da estatística. A expressão bootstrap está relaciona ao texto "*pulling oneself up by one's bootstrap*", uma frase usada aparentemente pela primeira vez no livro "As viagens singulares, campanhas e aventuras do Barão de Munchausen", de Rudolph Erich Raspe em 1786 (CHERNICK; LaBUDDE, 2011). O termo bootstrap faz alusão as histórias de que o Barão de Munchausen era capaz de se erguer do pântano puxando os cadarços das próprias botas, ou seja, saindo de situações difíceis utilizando os próprios esforços. Em estatística, bootstrap refere-se a fazer inferências acerca de parâmetros desconhecidos, utilizando reamostragem com reposição do conjunto amostral. Cada reamostragem permite calcular uma nova estatística e o conjunto de todas estas estimativas permite elaborar uma distribuição empírica, a qual é utilizada nas inferências (DALPOSSO, 2017).

O ano crítico para o bootstrap foi quando Brad Efron publicou um artigo em 1979 no periódico *Annals of Statistics*, o qual Efron tinha como objetivo explicar o método de reamostragem *jackknife* utilizando um método mais primitivo, sendo este o bootstrap, onde o autor argumenta que o método é amplamente aplicável e mais confiável em relação ao método *jackknife*. Como a metodologia bootstrap envolve amostragem com reposição n vezes para uma amostra de tamanho n, existe n^n possíveis amostras (CHERNICK; LaBUDDE, 2011). Neste sentido, bootstrap refere-se a um caso de simulação de Monte Carlo que trata a amostra original como a pseudo-população ou estimativa da população.

Existem dois procedimentos para se estimar os coeficientes do modelo de regressão utilizando a técnica de bootstrap: o método bootstrap dos resíduos e o método bootstrap dos pares.

O método bootstrap dos resíduos consiste em estimar os coeficientes de regressão para os dados originais e assim gerar os respectivos resíduos para as n observações realizadas. Estes resíduos formarão a amostra mestre. Deve-se então gerar as reamostras a partir destes resíduos. Para cada reamostra é calculada as estimativas dos coeficientes de regressão (RYZZO;

CYMROT, 2006).

O método bootstrap dos pares consiste em estimar os coeficientes de regressão utilizando os próprios dados originais como amostra mestre. Estes dados originais (que são vetores) devem ser reamostrados. Para cada reamostra são estimados os coeficientes da regressão linear para os dados da reamostra transformados (RYZZO; CYMROT, 2006).

2 OBJETIVO

Ajustar modelos de regressão em quatro conjuntos de dados e comparar os intervalos de confiança obtidos através da estatística clássica e utilizando os métodos bootstrap. Para isto utilizou-se o *software R*, com auxílio do pacote car (Fox J.; Weisberg, 2011).

3 MATERIAL E MÉTODOS

3.1 MATERIAL

Como base deste estudo, foram utilizados quatro problemas de regressão linear, o qual, o comportamento podemos verificar na Figura 1.

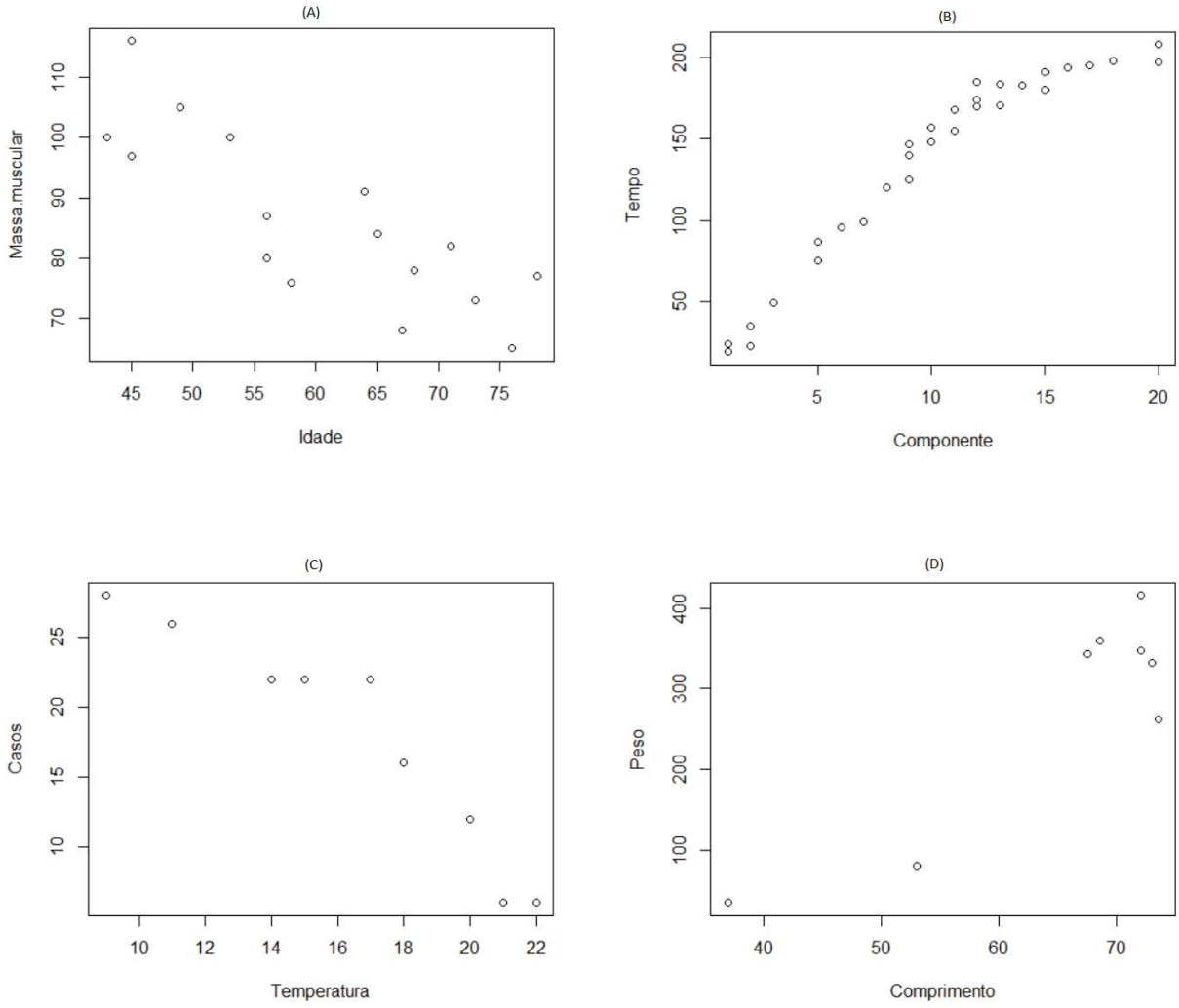

Figura 1: Gráficos de dispersão dos dados analisados.

Figura 1 (A): É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação, uma nutricionista selecionou 18 mulheres, com idade entre 40 e 79 anos, e observou em cada uma delas a idade (X) e a massa muscular (em quilograma) (Y) (RUGGIERO; LOPES, 1996).

Figura 1 (B): Uma empresa que presta serviço de manutenção de computadores coletou dados referentes a tempos de reparo (em minutos) (Y) e número componentes reparados ou substituídos em 30 computadores (X) (LOPES, 2003).

Figura 1 (C): Mostra os resultados de uma pesquisa realizada durante o mês de julho em um hospital pediátrico no qual foram apreciados a temperatura média do dia (X) e números de atendimento de casos com problema respiratório (Y) (MARIANO *et al*., 2006)

Figura 1 (D): Pesquisadores pretendem encontrar o peso de um urso de maneira mais fácil, para isto, através de seu comprimento (em polegadas) (X), os pesquisadores pretendem encontrar o Peso (em libras) (Y) do urso (TRIOLA, 2008).

3.2 MÉTODOS

Regressão linear:

É um procedimento que consiste em determinar estimativas do coeficiente angular e linear da reta procurada, a partir de uma série de pontos representativos das variáveis que compõem um determinado fenômeno, utilizando a minimização da soma dos quadrados dos resíduos. A função obtida deve permitir com satisfatória segurança a realização de análises e projeções sobre o fenômeno estudado (FRANCO, 2006).

Defini-se neste trabalho a regressão linear do tipo:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

onde β_0 e β_1 são parâmetros, x é a variável explanatória e Y é a variável dependente, ε_i denominado erro aleatório ou estocástico, onde se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X. Através da minimização da soma dos quadrados dos resíduos, serão obtidas as estimativas b_o e b_1.

Método bootstrap:

É usado para fornecer uma maneira de obter estimativas para o caso de a distribuição de erros ser desconhecida e não normal, independendo do método para estimar estes parâmetros. Portanto temos duas abordagens básicas para utilizar o método bootstrap em regressão, ambas podendo ser usadas em regressões lineares ou não lineares:

Bootstrap dos resíduos:

Algoritmo 1: Bootstrap dos resíduos (SHERMAN; SASKIA, 1997).

a) Após determinado $\boldsymbol{b}$ calcule os resíduos $\varepsilon_i = y_i - x_i^T b$, $i = 1, \dots, n$;
b) Obtenha uma reamostra com reposição ε_i^*, $i = 1, \dots, n$ de $\varepsilon_i = 1, \dots, n$;
c) Calcule $y_i^* - x_i^T b + \varepsilon_i^*$;
d) Calcule $\boldsymbol{b}^{*(1)}$ a partir de $[\boldsymbol{y}^*, \boldsymbol{X}]$ da mesma maneira que $\boldsymbol{b}$ é calculado a partir dos dados originais $[\boldsymbol{y}, \boldsymbol{X}]$;
e) Obtenha a distribuição bootstrap calculando $\boldsymbol{b}^{*(r)}$, $r = 1, \dots, R$. Sendo R o tamanho da reamostras.

Bootstrap dos pares:

Algoritmo 2: Bootstrap dos pares (DAVISON; HINKLEY, 1997).

a) Considere o conjunto de dados originais $[\boldsymbol{Y}, \boldsymbol{X}]$;
b) Obtenha uma nova matriz $\left[\boldsymbol{Y}^{*(1)}, \boldsymbol{X}^{*(1)}\right]$ a partir de uma reamostragem com reposição das linhas da matriz $[\boldsymbol{Y}, \boldsymbol{X}]$;
c) Calcule $\widehat{\boldsymbol{b}}^{*(1)}$ da mesma forma que foi calculado $\boldsymbol{b}$;
d) Obtenha a distribuição bootstrap calculando $\widehat{\boldsymbol{b}}^{*(b)}$, $b = 1, \dots, B$. Sendo B o tamanho da reamostras.

Intervalo de confiança:

É qualquer intervalo construído em torno de um estimador, que tem a probabilidade de conter o verdadeiro valor do correspondente parâmetro de uma população (DODGE, 2008).

A cada intervalo de confiança está associado um nível de confiança $(1 - \alpha)$, que é um número que representa a probabilidade de o valor real estar dentro de intervalo. O nível de confiança é a proporção de todas as amostras possíveis para as quais o intervalo de confiança abrange o valor real. Por exemplo, um intervalo de confiança de 95% é calculado a partir de um processo que terá êxito na abrangência do parâmetro populacional para 95% das amostras que poderiam ser extraídas (NAVIDI, 2010).

Em uma regressão linear do tipo $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ o intervalo de confiança para β_0 e β_1 é determinado por:

$$b_n - t_0 s(b_n) < \beta_n < b_n + t_0 s(b_n)$$

em que b_né a estimativa obtida, t_0 é o valor critico da distribuição t de Student e $s(b_n)$ é a estimativa do desvio padrão de b_n (HOFFMANN, 2006).

Intervalo de confiança bootstrap percentil de Efron:

A utilização do intervalo de confiança percentil é a maneira mais simples segundo Efron de construir um intervalo de confiança para um parâmetro com base em estimativas bootstrap. A maneira mais adequada para escolher um intervalo de confiança de 95%, é ordenar as estimativas obtidas crescentemente e excluir 2,5% das menores estimativas e 2,5% das maiores estimativas.

Recursos computacionais:

Neste trabalho utilizou-se o *Software R* (R Core Team, 2019), como auxilio para os devidos cálculos utilizamos do pacote car (Fox J.; Weisberg, 2011).

4 RESULTADOS E DISCUSSÕES

Dos problemas apresentados na Figura 1, realizou-se a regressão linear e os resultados são apresentados na Figura 2.

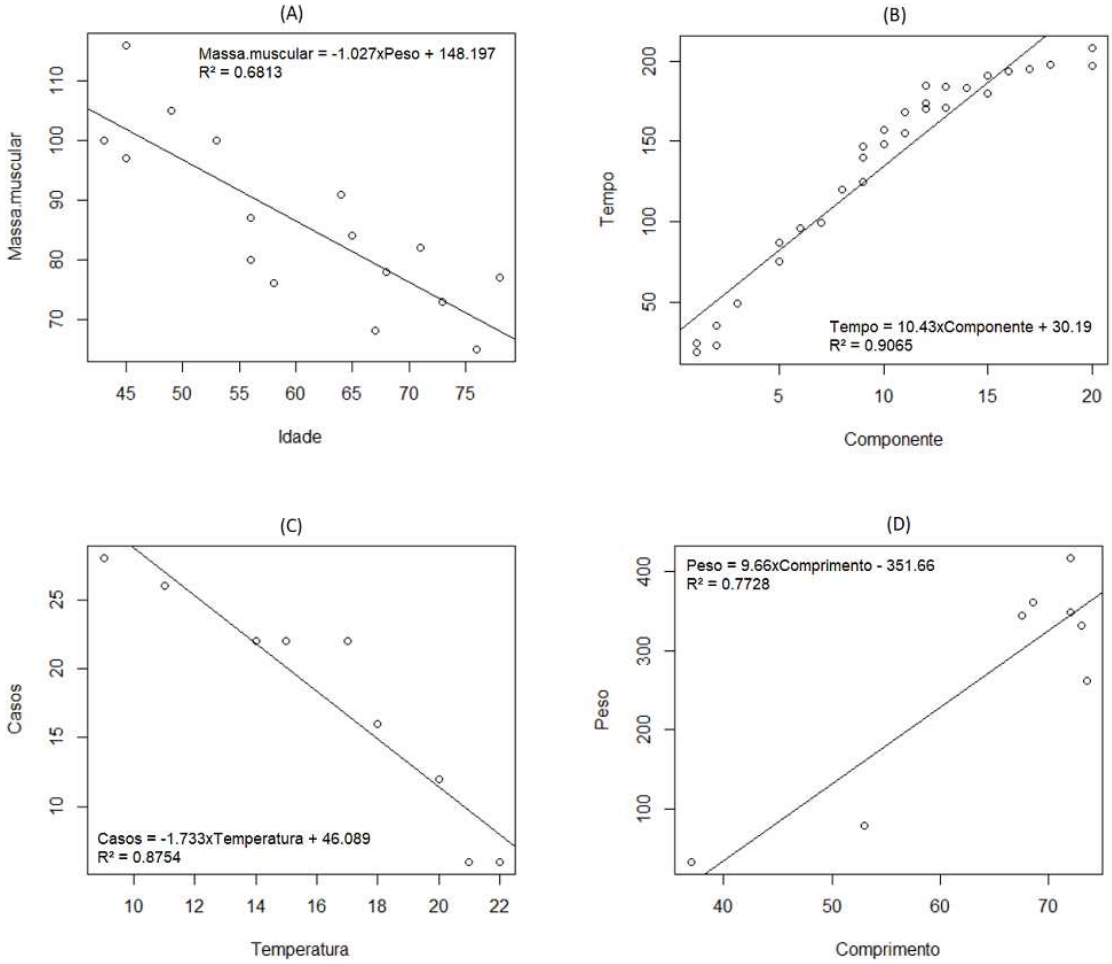

Figura 2: Modelos ajustados por regressão linear.

Fonte: Compilação do autor.[1]

Observando modelo destacado na Figura 2(A), é possível fazer inferências acerca da massa muscular de mulheres, por exemplo, pode-se estimar que uma mulher de 60 anos terá uma

[1] (A): É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação, uma nutricionista selecionou 18 mulheres, com idade entre 40 e 79 anos, e observou em cada uma delas a idade (X) e a massa muscular (em quilograma) (Y) (RUGGIERO; LOPES, 1996). (B): Uma empresa que presta serviço de manutenção de computadores coletou dados referentes a tempos de reparo (em minutos) (Y) e número componentes reparados ou substituídos em 30 computadores (X) (LOPES, 2003). (C): Mostra os resultados de uma pesquisa realizada durante o mês de julho em um hospital pediátrico no qual foram apreciados a temperatura média do dia (X) e números de atendimento de casos com problema respiratório (Y) (MARIANO *et al*., 2006) (D): Pesquisadores pretendem encontrar o peso de um urso de maneira mais fácil, para isto, através de seu comprimento (em polegadas) (X), os pesquisadores pretendem encontrar o Peso (em libras) (Y) do urso (TRIOLA, 2008).

massa muscular de 86,58 Kg. Pelo modelo indicado na Figura 2(B) um funcionário da empresa que presta serviço de manutenção em computadores pode afirmar a um cliente que para substituir uma placa de vídeo em um computador não serão gastos mais de 41 minutos. De acordo com o modelo apresentado na Figura 2 (C) em um dia bem frio no hospital pediátrico, digamos que faça zero graus, espera-se atender cerca de 46 casos com problemas respiratórios. Já o modelo apresentado na Figura 2 (D) indica aproximadamente que um urso que apresente um comprimento de 50 polegadas terá um peso de 131 libras.

Pode-se observar que os coeficientes de determinação (R^2) dos modelos ajustados são satisfatórios, pois quanto mais próximo de 1 o coeficiente de determinação, melhor é o ajuste. Uma provável causa do coeficiente de determinação dos modelos da Figura 2 (A) e Figura 2(D) serem menores, pode ser justificada por variáveis que não foram consideradas no modelo. Por exemplo, o coeficiente de determinação 0,6813 da Figura 2(A) indica que aproximadamente 68% da variabilidade da massa muscular está sendo explicada pela variabilidade da idade. Pode-se tentar obter um coeficiente de determinação mais alto incluindo outras variáveis explanatórias e implementando um modelo de regressão linear múltipla, porém, este não é o foco deste trabalho.

Embora a Figura 2 tenha fornecido as estimativas pontuais dos parâmetros dos modelos, uma simples estimativa pontual pode não ser suficiente para fornecer evidências, que de fato auxiliem em suas deduções. São necessárias também medidas da precisão desta estimativa, como por exemplo a utilização de intervalos de confiança para os parâmetros. A Figura 3 apresenta os histogramas das estimativas bootstrap dos parâmetros destacando-se os intervalos de confiança clássico, intervalos de confiança bootstrap e estimativa pontual. Na Figura 3: (A), temos destacados os intervalos de confiança do modelo: $Massa.muscular = 148{,}197 - 1{,}027 \times Peso$. A amplitude do intervalo de confiança clássico do intercepto (148,197) foi 44,54, um valor levemente inferior à amplitude do intervalo bootstrap dos pares (44,59) e superior à amplitude do intervalo bootstrap dos resíduos (41,51). Como uma das utilidades dos intervalos é fornecer a ideia de dispersão ou variabilidade das estimativas podemos concluir que para o intercepto deste modelo, o intervalo de confiança bootstrap dos resíduos foi o mais preciso.

O mesmo ocorre para o coeficiente angular (-1,027) onde à amplitude do intervalo clássico (0,712) é levemente superior à amplitude dos intervalos bootstrap dos pares (0,665) e bootstrap dos resíduos (0,656). Desta forma, podemos concluir que para o coeficiente angular, o intervalo de confiança bootstrap dos resíduos foi o mais preciso.

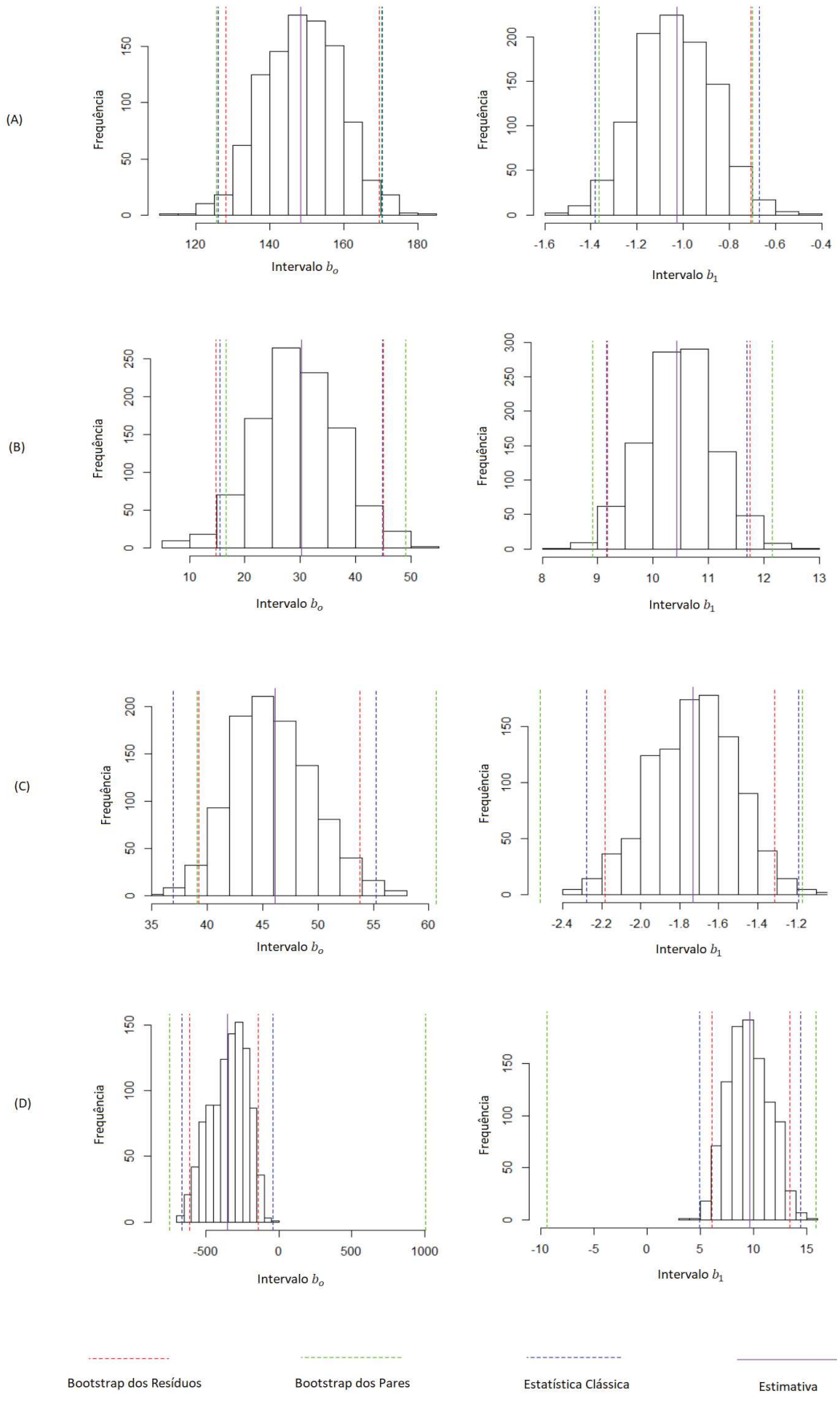

Figura 3: Histograma dos intervalos de confiança dos parâmetros.

Na Figura 3(B) destaca-se os intervalos de confiança do modelo: $Tempo = 30{,}19 + 10{,}43 \times Componete$. Analisando o intercepto (30,19) temos que a amplitude do intervalo de confiança clássico (15,58; 44,86) é inferior à amplitude do intervalo bootstrap dos pares (16,56; 49,10) e levemente inferior à amplitude do intervalo bootstrap dos resíduos (14,87; 44,95). O mesmo ocorre para o coeficiente angular (10,43), onde a amplitude do o intervalo de confiança clássico (9,16; 11,70) é inferior à amplitude do intervalo bootstrap dos pares (8,91; 12,15) e levemente inferior à amplitude do intervalo bootstrap dos resíduos (9,17; 11,74), ou seja, a amplitude do intervalo de confiança clássico apresenta uma leve diferença em relação a amplitude do intervalo de confiança bootstrap dos resíduos, garantindo que ambos tenham uma boa precisão.

Na Figura 3(C) destaca-se os intervalos de confiança do modelo: $Casos = 46{,}089 - 1{,}733 \times Temperatura$. Podemos verificar que tanto para o intercepto (46,089) quanto para o coeficiente angular (-1,733) a amplitude do intervalo de confiança do bootstrap dos resíduos (14,515 para intercepto e 0,867 para o coeficiente angular) é inferior a amplitude do intervalo de confiança clássico (18,286 para o intercepto e 1,084 para o coeficiente angular) e do intervalo de confiança bootstrap dos pares (21,557 para o intercepto e 1,344 para o coeficiente angular), ou seja, o intervalo de confiança dos resíduos neste caso é mais preciso.

Na Figura 3(D) destaca-se os intervalos de confiança do modelo: $Peso = -351{,}66 + 9{,}66 \times Comprimento$. A amplitude dos intervalos de confiança do bootstrap dos pares apresentam muita variabilidade, pois analisando o intercepto $(-351{,}66)$ o intervalo varia de -748,82 a 1008,14 e analisando o coeficiente angular (9,66) o intervalo varia de -9,42 a 15,84, ou seja, os intervalos de confiança bootstrap dos pares apresentam falta de precisão, pois além de apresentarem grande variabilidade, os mesmos contém o zero, tornando os parâmetros não significativos. Por este motivo neste caso se descarta a utilização do intervalo bootstrap dos pares para novos estudos. Mas ao contrário, temos que no intercepto a amplitude do intervalo bootstrap dos resíduos (468,47) é mais preciso que a amplitude do intervalo de confiança clássico (623,52), e o mesmo ocorre no coeficiente angular, onde a amplitude do intervalo bootstrap dos resíduos (7,311) é inferior a amplitude do intervalo de confiança clássico (9,49), garantindo que o intervalo bootstrap dos resíduos é mais preciso.

5 CONSIDERAÇÕES FINAIS

Conclui-se que a utilização de recursos computacionais é essencial para o desenvolvimento deste estudo, pois devido aos cálculos e estimativas a serem obtidas, manualmente se tornaria algo inviável.

Dos resultados obtidos, temos que os métodos bootstrap são uma excelente alternativa para os métodos tradicionais, pois em três dos quatros estudos de casos as amplitudes dos intervalos de confiança bootstrap dos resíduos foram inferiores às amplitudes dos intervalos de confiança clássico, tornando muito mais preciso, além de, não ser necessário à verificação de pressupostos, como por exemplo o formato da distribuição dos erros do modelo, que deve ser normal .

Como trabalhos futuros, pretende-se utilizar os métodos bootstrap para verificar os pressupostos que a regressão linear impõem ao utilizá-la, como a existência de dados atípicos no conjunto amostral e utilizar os métodos bootstrap na existência de novas variáveis, como no caso de regressão linear múltipla, afim de verificar, se com novas variáveis, os resultados se tornam mais precisos.

REFERÊNCIAS

ANTUNES, José Leopoldo Ferreira; CARDOSO, Maria Regina Alves. Uso da análise de séries temporais em estudos epidemiológicos. **Epidemiologia e Serviços de Saúde**, [s.l.], v. 24, n. 3, set. 2015. Instituto Evandro Chagas. Disponível em: <http://dx.doi.org/10.5123/s1679-49742015000300024>. Acesso em: 10 de Out. de 2019.

BIMBO, Alberto del. **Visual information retrieval.** San Francisco: Morgan Kaufmann Publishers Inc, 1999. (ISBN: 9781558606241).

CHERNICK, Michael R.; LABUDDE, Robert A. **An Introduction to Bootstrap Methods with Applications to R.** New Jersey: John Wiley & Sons Inc, 2011. (ISBN: 9780470467046).

DALPOSSO, Gustavo Henrique. **Método Bootstrap na agricultura de precisão.** 2017. 90 f. Tese (Doutorado - Programa de Pós-graduação em Engenharia Agrícola) - Universidade Estadual do Oeste do Paraná, Cascavel, 2017. Disponível em <http://tede.unioeste.br/handle/tede/3075>. Acesso em: 10 de Out. de 2019.

DAVISON, Anthony C.; HINKLEY, David Victor. **Bootstrap Methods and Their Application.** Cambridge: Cambridge University Press, 1997. (ISBN: 9780511802843).

DODGE, Yadolah. **The Concise Encyclopedia of Statistics.** New York: Springer, 2008. (ISBN: 9780387328331).

FILHO, Miguel Lopes de Oliveira. **A Utilização da Regressão Linear Como Ferramenta Estratégica Para a Projeção dos Custos Produção.** IX Congresso Brasileiro de Custos - São Paulo, Brasil, 13 a 15 de outubro de 2002. Disponível em: <https://anaiscbc.emnuvens.com.br/anais/article/viewFile/2762/2762>. Acesso em: 8 de Out. de 2019.

FRANCO, Neide Maria Bertoldi. **Cálculo Numérico.** São Paulo, Universidade de São Paulo - Instituto de Ciências Matemáticas e de Computação, [2006].

John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

LOPES, Luiz Felipe Dias. **Apostila Estatística.** UFSM - Universidade Federal de Santa Maria, 2003. Disponível em: <http://www.inf.ufsc.br/~vera.carmo/LIVROS/LIVROS/Luis%20Felipe%20Dias%20Lopes.pdf>. Acesso em: 10 de Out. de 2019.

MARIANO, Mirtes Vitória; LAURICEL, Christiane Mazur; FRUGOLI, Alexandre Daliberto. **Estatística Indutiva:** Teoria, Exercícios Resolvidos Tarefas. [s.i.]: do Autor, [2006].

MATOS, Manuel. **Manual Operacional para a Regressão Linear.** FEUP - Faculdade de Engenharia da Universidade do Porto, 1995.

NAVIDI, William. **Probabilidade e Estatística para Ciências Exatas.** New York: Mc Graw Hill, 2010. (ISBN: 9788580550733).

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

RIZZO, Ana Lucia Tucci; CYMROT, Raquel. **Estudo e aplicações da técnica Bootstrap.** In: II Jornada de Iniciação Científica PIBIC e PIVIC, 2006. São Paulo. Anais... São Paulo, Universidade Presbiteriana Mackenzie, 2006. Disponível em: <http://meusite.mackenzie.com.br/raquelc/ana_lucia.pdf>. Acesso em: 9 de Out. De 2019.

RUGGIERO, Márcia A. Gomes; LOPES, Vera Lúcia da Rocha. **Cálculo Numérico:** aspectos teóricos e computacionais. 2. ed. São Paulo: Makron Books, 1996. (ISBN: 9788534602044)

SHERMAN, Michael; CESSIE, Saskia Le. A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. **Communications In Statistics - Simulation And Computation**, [s.l.], v. 26, n. 3, 1997. http://dx.doi.org/10.1080/03610919708813417.

SILVA, João Pedro Bento Clemente da. **Modelos de regressão linear e logística utilizando o software R.** 2016. 146 f. Tese (Mestrado em Estatística, Matemática e Computação), Universidade Aberta, Lisboa, 2016. Disponível em: <http://hdl.handle.net/10400.2/6129>. Acesso em: 12 de Out. de 2019.

TRIOLA, Mario F. **Introdução a estatística.** 10. ed. [s.i]: Ltc, 2008. (ISBN: 9788521615866).

APÊNDICE

Estudo de caso 1 (Script R)

#É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação, uma nutricionista selecionou 18 mulheres, com idade entre 40 e 79 anos, e observou em cada uma delas a idade (X) e a massa muscular (Y).

```
dados<-
matrix(c(71.0,64.0,43.0,67.0,56.0,73.0,68.0,56.0,76.0,65.0,45.0,58.0,45.0,53.0,49.0,78.0,73.0,
68.0,82.0,91.0,100.0,68.0,87.0,73.0,78.0,80.0,65.0,84.0,116.0,76.0,97.0,100.0,105.0,77.0,73.0
,78.0),nrow=18,ncol=2)
Idade<-dados[,1]
Idade
Massa.muscular<-dados[,2]
Massa.muscular
#Gráfico
plot(Idade,Massa.muscular)
#Modelo
modelo<-lm(Massa.muscular~Idade)
modelo
#Informação do modelo
summary(modelo)
#semente aleatória
set.seed(1)
############################################################
#Bootstrap dos resíduos
betahat.boot<-Boot(modelo,R=1000,method="residual")
#réplica do intercepto
betahat.boot$t[,1]
#réplica do coeficiente angular
betahat.boot$t[,2]
#Intervalo de confiança
confint(betahat.boot,type="perc")
```

```
#########################################################
#Bootstrap dos Pares
betahat.boot1<-Boot(modelo,R=1000,method="case")
#réplica do intercepto
betahat.boot1$t[,1]
#réplica do coeficiente angular
betahat.boot1$t[,2]
#Intervalo de confiança
confint(betahat.boot1,type="perc")
confint(modelo)

hist(betahat.boot$t[,1])
abline(v=128.10107,col="red",lty=2)
abline(v=169.6151731,col="red",lty=2)
abline(v=125.591211,col="Green",lty=2)
abline(v=170.1805602,col="Green",lty=2)
abline(v=125.927935,col="blue",lty=2)
abline(v=170.4659704,col="blue",lty=2)
abline(v=148.197,col="Purple")
sd(betahat.boot$t[,1])
mean(betahat.boot$t[,1])

hist(betahat.boot$t[,2])
abline(v= -1.36358,col="red",lty=2)
abline(v= -0.7072482,col="red",lty=2)
abline(v= -1.364576,col="Green",lty=2)
abline(v= -0.6995062,col="Green",lty=2)
abline(v= -1.382844,col="blue",lty=2)
abline(v=-0.6704864,col="blue",lty=2)
abline(v= -1.027,col="Purple")
sd(betahat.boot$t[,2])
mean(betahat.boot$t[,2])
```

Estudo de caso 2 (Script R)

Uma empresa que presta serviço de manutenção de computadores coletou dados referentes a tempos de reparo (em minutos) e número componentes reparados ou substituídos em 30 computadores. Os dados são apresentados na tabela abaixo.

```
dados<-
matrix(c(1,1,2,2,3,5,5,6,7,8,9,9,9,10,10,11,11,12,12,12,13,13,14,15,15,16,17,18,20,20,19,24,2
3,35,49,75,87,96,99,120,125,147,140,148,157,155,168,170,185,174,171,184,183,180,191,194
,195,198,197,208),nrow=30,ncol=2)
Componente<-dados[,1]
Componente
Tempo<-dados[,2]
Tempo
#Gráfico
plot(Componente,Tempo)
#Modelo
modelo<-lm(Tempo~Componente)
modelo
#Informação do modelo
summary(modelo)
#semente aleatória
set.seed(1)
############################################################
#Bootstrap dos resíduos
betahat.boot<-Boot(modelo,R=1000,method="residual")
#réplica do intercepto
betahat.boot$t[,1]
#réplica do coeficiente angular
betahat.boot$t[,2]
#Intervalo de confiança
confint(betahat.boot,type="perc")
#########################################################
#Bootstrap dos Pares
```

```
betahat.boot1<-Boot(modelo,R=1000,method="case")
#réplica do intercepto
betahat.boot1$t[,1]
#réplica do coeficiente angular
betahat.boot1$t[,2]
#Intervalo de confiança
confint(betahat.boot1,type="perc")
confint(modelo)

hist(betahat.boot$t[,1])
abline(v=14.869689,col="red",lty=2)
abline(v=44.95255,col="red",lty=2)
abline(v=16.556666 ,col="Green",lty=2)
abline(v=49.09787,col="Green",lty=2)
abline(v=15.527136,col="blue",lty=2)
abline(v=44.86065,col="blue",lty=2)
abline(v=30.194 ,col="Purple")
sd(betahat.boot$t[,1])
mean(betahat.boot$t[,1])

hist(betahat.boot$t[,2])
abline(v= 9.171157,col="red",lty=2)
abline(v= 11.74630,col="red",lty=2)
abline(v= 8.909311,col="Green",lty=2)
abline(v= 12.15160,col="Green",lty=2)
abline(v= 9.156615,col="blue",lty=2)
abline(v=11.70079,col="blue",lty=2)
abline(v= 10.429 ,col="Purple")
sd(betahat.boot$t[,2])
mean(betahat.boot$t[,2])
```

Estudo de caso 3 (Script R)

A tabela abaixo mostra os resultados de uma pesquisa realizada durante o mês de julho em um hospital pediátrico no qual foram apreciados: temperatura média do dia e números de atendimento de casos com problema respiratório.

```
dados<-matrix(c(9,11,14,15,17,18,20,21,22,28,26,22,22,22,16,12,6,6),nrow=9,ncol=2)
Temperatura<-dados[,1]
Temperatura
Casos<-dados[,2]
Casos
#Gráfico
plot(Temperatura,Casos)
#Modelo
modelo<-lm(Casos~Temperatura)
modelo
#Informação do modelo
summary(modelo)
#semente aleatória
set.seed(1)
############################################################
#Bootstrap dos resíduos
betahat.boot<-Boot(modelo,R=1000,method="residual")
#réplica do intercepto
betahat.boot$t[,1]
#réplica do coeficiente angular
betahat.boot$t[,2]
#Intervalo de confiança
confint(betahat.boot,type="perc")
#########################################################
#Bootstrap dos Pares
betahat.boot1<-Boot(modelo,R=1000,method="case")
#réplica do intercepto
betahat.boot1$t[,1]
#réplica do coeficiente angular
```

```
betahat.boot1$t[,2]
#Intervalo de confiança
confint(betahat.boot1,type="perc")
confint(modelo)

hist(betahat.boot$t[,1])
abline(v=39.255810,col="red",lty=2)
abline(v=53.771473,col="red",lty=2)
abline(v=39.12365 ,col="Green",lty=2)
abline(v=60.68074,col="Green",lty=2)
abline(v=36.945875,col="blue",lty=2)
abline(v=55.231903,col="blue",lty=2)
abline(v= 46.0889 ,col="Purple")
sd(betahat.boot$t[,1])
mean(betahat.boot$t[,1])

hist(betahat.boot$t[,2])
abline(v= -2.182012,col="red",lty=2)
abline(v= -1.314598,col="red",lty=2)
abline(v= -2.51403,col="Green",lty=2)
abline(v= -1.17053,col="Green",lty=2)
abline(v= -2.275342,col="blue",lty=2)
abline(v=-1.191325,col="blue",lty=2)
abline(v= -1.7333 ,col="Purple")
sd(betahat.boot$t[,2])
mean(betahat.boot$t[,2])
```

Estudo de caso 4 (Script R)

```
dados<-
matrix(c(53.0,67.5,72.0,72.0,73.5,68.5,73,37,80,344,416,348,262,360,332,34),nrow=8,ncol=2
)
```

```
comprimento<-dados[,1]
comprimento
peso<-dados[,2]
peso
#Gráfico
plot(comprimento,peso)
#Modelo
modelo<-lm(peso~comprimento)
modelo
#Informação do modelo
summary(modelo)
#semente aleatória
set.seed(1)
############################################################
#Bootstrap dos resíduos
betahat.boot<-Boot(modelo,R=1000,method="residual")
#réplica do intercepto
betahat.boot$t[,1]
#réplica do coeficiente angular
betahat.boot$t[,2]
#Intervalo de confiança
confint(betahat.boot,type="perc")
#########################################################
#Bootstrap dos Pares
betahat.boot1<-Boot(modelo,R=1000,method="case")
#réplica do intercepto
betahat.boot1$t[,1]
#réplica do coeficiente angular
betahat.boot1$t[,2]
#Intervalo de confiança
confint(betahat.boot1,type="perc")
confint(modelo)
```

```
hist(betahat.boot$t[,1])
abline(v=-610.482442,col="red",lty=2)
abline(v=-142.01211,col="red",lty=2)
abline(v=-748.821328 ,col="Green",lty=2)
abline(v=1008.14421,col="Green",lty=2)
abline(v=-663.421144,col="blue",lty=2)
abline(v=-39.89864,col="blue",lty=2)
abline(v= -351.660 ,col="Purple")
sd(betahat.boot$t[,1])
mean(betahat.boot$t[,1])

hist(betahat.boot$t[,2])
abline(v= 6.132723,col="red",lty=2)
abline(v= 13.44368,col="red",lty=2)
abline(v= -9.415568 ,col="Green",lty=2)
abline(v= 15.84061,col="Green",lty=2)
abline(v= 4.914137,col="blue",lty=2)
abline(v= 14.40543,col="blue",lty=2)
abline(v=  9.660 ,col="Purple")
sd(betahat.boot$t[,2])
mean(betahat.boot$t[,2])
```