MÉTODOS BOOTSTRAP EM REGRESSÃO

Paulo Bessa e Romário Adrián

Resumo

O trabalho descreve a aplicação do *Bootstrap* nos processos de estimação, análise de variabilidade e construção de intervalos de confiança para os parâmetros de um modelo de regressão linear. Partindo da rejeição de normalidade para a distribuição dos resíduos, pretendeu-se investigar a performance dos métodos *Bootstrap* em comparação às técnicas tradicionais de estimação. Observou-se que todos os métodos aplicados apresentaram uma maior performance em relação ao procedimento padrão, gerando estimadores com menor variabilidade e estimativas intervalares mais coerentes e precisas. Neste sentido, dentre os métodos utilizados, destaca-se a reamostragem dos resíduos como aquele que obteve a melhor performance geral. Portanto, trata-se de uma verificação da relevância dos métodos de reamostragem para questões inerentes à Estatística Aplicada.

Palavras-chave : Regressão linear. Bootstrap. Intervalos de confiança.

1 INTRODUÇÃO

Em se tratando de modelos de regressão lineares com as suposições verificadas, o Teorema de Gauss-Markov fornece os melhores estimadores lineares não viesados para os coeficientes. Métodos alternativos são aplicados à medida que tais suposições não se mostram válidas, a exemplo dos métodos de reamostragem como o *Bootstrap*. Para um conjunto de dados financeiros hipotéticos de uma empresa, ajustou-se um modelo de regressão linear com o intuito de prever o faturamento mensal em função do investimento em propaganda. Rejeitou-se a suposição de normalidade devido à observação de assimetria para a distribuição dos resíduos. Dessa forma, comparou-se as estimativas intervalares tradicionais para os parâmetros com os intervalos de confiança oriundos de métodos *Bootstrap* : reamostragem dos resíduos, reamostragem dos pares e *Bootstrap* selvagem ou ponderado.

2 MÉTODOS DE REAMOSTRAGEM

Dois dos maiores problemas em Estatística Aplicada envolvem a obtenção de um estimador para um parâmetro de interesse e a análise de sua acurácia. Existe uma extensa teoria fundamentada em métodos paramétricos para a estimação de parâmetros. Esses métodos são baseados em suposições de distribuições probabilísticas conhecidas, tais como normal, t de Student, qui-quadrado e F de Snedecor. Entretanto, quando essas suposições não são verificadas, os métodos paramétricos podem apresentar baixa performance e gerar resultados de baixa qualidade. Há, portanto, a necessidade de técnicas alternativas, a exemplo dos métodos não paramétricos.

As vantagens dos métodos não paramétricos residem no fato de que estes podem ser aplicados sem a necessidade da suposição de uma distribuição de probabilística (livres de distribuição), além de demonstrarem uma menor sensibilidade a valores extremos ou *outliers*. Dentre tais técnicas, ressalta-se a classe dos métodos de reamostragem : utilizam da distribuição empírica e da amostra original para criar novas amostras (reamostras). Com o decorrer dos anos os métodos de reamostragem ganharam destaque e notoriedade na comunidade estatística devido em grande parte à sua grande flexibilidade e extensa gama de aplicações. Exemplos destes métodos são o *Jackknife*, a Validação Cruzada e o *Bootstrap*.

Primeiramente implementado por R.E. von Mises (1883 - 1953) e depois desenvolvido por Quenouille (1924 - 1973) e Turkey (1915 - 2000) na década de 1950, o *Jackknife* consiste na permutação da amostra original ao se retirar uma observação por vez. É um método de rápido processamento computacional com a capacidade de produzir estimadores consistentes ; no entanto, apresenta baixa eficácia para a construção de intervalos de confiança precisos, baseando-se em aproximações grosseiras. Suas limitações acarretaram no desenvolvimento de técnicas mais gerais e eficientes, como o *Bootstrap*.

2.1 Bootstrap

Popularizado e unificado por Efron (1938-) com base em métodos anteriormente desenvolvidos, o *Bootstrap* é um dos mais célebres e divulgados métodos de reamostragem. A princípio criado para a estimação da variabilidade de estimadores, o *Bootstrap* adquiriu uma vasta gama de aplicações, tais como :

1. Estimação de parâmetros ;
2. Análise da eficiência de estimadores ;
3. Correção de viés ;
4. Construção de intervalos de confiança ;
5. Regressão ;
6. Análise de séries temporais.

As estimativas *Bootstrap* são geralmente produzidas via simulação de Monte Carlo, ou seja, por meio da geração de um número massivo de reamostras com reposição a partir da amostra original. As estimativas obtidas são utilizadas, por exemplo, para a estimação do erro padrão e análise de viés de estimadores, bem como a construção de intervalos de confiança. Vale ressaltar que se recomenda a geração de pelo menos 5000 reamostras para a obtenção via *Bootstrap* de intervalos de confiança com certo grau de precisão.

Com relação à aplicação em regressão, o *Bootstrap* representa, por exemplo, um meio de se obter as estimativas dos coeficientes quando as premissas do modelo não são válidas ou quando rejeita-se a suposição de normalidade. No final da década de 1970, Efron desenvolveu duas abordagens distintas para se estimar os parâmetros de regressão via *Bootstrap* :

1. Reamostragem dos resíduos ;
2. Reamostragem dos pares.

Tais abordagens diferem quanto à geração das reamostras e às suposições realizadas.

2.1.1 *Reamostragem dos resíduos*

A reamostragem dos resíduos pressupõe que o modelo seja homoscedástico com distribuição probabilística desconhecida. De forma generalizada, um modelo de regressão pode ser escrito da seguinte forma

$$y_i = g_i(\beta) + e_i \;\; \forall i = 1, 2, \cdots, n$$

em que

1. $y = (y_1, \cdots, y_n)^T$ é o vetor de observações da variável resposta

2. $\beta = (\beta_0, \beta_1, \cdots, \beta_n)^T$ é o vetor de parâmetros

3. $g(\beta)$ é uma função de β

4. e representa a fonte de variação

Para a aplicação do método, a priori define-se uma função que mensura a distância entre os valores observados e os valores preditos, a exemplo da função utilizada para a estimação via mínimos quadrados.

$$Q(\beta) = \sum_{i=1}^{n} (y_i - g_i(\beta))^2$$

A minimização dessa função sob o modelo linear fornece $\hat{\beta} = (X^T X)^{-1} X^T y$ como estimativa pontual para β. Considere $\hat{\epsilon} = (\hat{\epsilon_1}, \cdots, \hat{\epsilon_n})^T$ o vetor dos resíduos ordinários tal que

$$\hat{\epsilon_i} = y_i - g_i(\hat{\beta}) \;\; \forall i = 1, 2, \cdots, n$$

Calcula-se o resíduo ordinário para cada observação y_i correspondente. A réplica *Bootstrap* $\hat{\epsilon}^* = (\hat{\epsilon_1}^*, \cdots, \hat{\epsilon_n}^*)^T$ é gerada a partir da reamostragem com reposição de $\hat{\epsilon}$. Os novos valores y_i^* são tais que

$$y_i^* = g_i(\hat{\beta}) + \epsilon_i^* \;\; \forall i = 1, 2, \cdots, n$$

Dessa maneira, a estimativa *Bootstrap* para o vetor de parâmetros é obtida a partir da geração de uma quantidade n de resíduos *Bootstrap*. Sob o modelo linear, tal estimativa é dada por

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

Via simulação de Monte Carlo, é possível produzir um número R de vetores de resíduos *Bootstrap*. Dessa forma, são calculadas R estimativas *Bootstrap* para β. A distribuição dessas estimativas é analisada para a construção de intervalos de confiança e testes de hipóteses.

Pelo exposto, o método dos resíduos para o modelo linear pode ser descrito conforme as etapas abaixo.

1. Obtém-se o vetor $\hat{\beta}$ via mínimos quadrados;

2. Para cada observação, calcula-se o resíduo $\hat{\epsilon}_i = y_i - g_i(\hat{\beta})$;

3. Reamostra-se com reposição o vetor de resíduos e avalia-se

$$y_i^{*} = g_i(\hat{\beta}) + \epsilon_i^{*}$$

4. Para cada y_i^{*}, calcula-se a estimativa *Bootstrap*

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

5. Repetem-se os passos 3 e 4 em um total de R vezes, gerando o vetor de estimativas *Bootstrap*.

2.1.2 ***Reamostragem dos pares***

Seja $x = (x_1, \cdots, x_n)^T$ o vetor de valores da variável explicativa em um modelo de regressão linear simples. Sendo a distribuição probabilística do modelo desconhecida, a reamostragem dos pares assume que os pares (x, y) são independentes e identicamente distribuídos. O método consiste em reamostrar com reposição tais vetores de pares, produzindo as amostras *Bootstrap* (x^{*}, y^{*}). Por consequência, considerando o MRLS, a nova matriz de especificação e o vetor de observações da variável resposta são dados por

$$X_* = \begin{bmatrix} 1 & x_1^* \\ 1 & x_2^* \\ \vdots & \vdots \\ 1 & x_n^* \end{bmatrix}$$

$$y^* = (y_1^*, \cdots, y_n^*)^T$$

Deste modo, a estimativa *Bootstrap* para o vetor de parâmetros é $\hat{\beta}^* = (X_*^T X_*)^{-1} X_*^T y^*$. De forma análoga à reamostragem dos resíduos, pode-se gerar R estimativas via simulação de Monte Carlo caso o procedimento seja repetido em um total de R vezes. A seguir estão ordenadas as etapas que descrevem a reamostragem dos pares para o modelo linear.

1. Reamostra-se com reposição os pares (x, y) ;

2. Define-se os novos vetores com os valores das variáveis resposta e explicativa, bem como a nova matriz de especificação ;

3. Calcula-se a estimativa *Bootstrap* $\hat{\beta}^*$ com esses novos vetores ;

4. Repetem-se os passos anteriores em um total de R vezes, gerando o vetor de estimativas *Bootstrap*.

2.1.3 *Resíduos vs pares*

Um ponto relevante se trata de um critério de escolha entre a reamostragem dos resíduos e a reamostragem dos pares. Efron (1979) argumenta que tal escolha depende da natureza do modelo trabalhado, embora cada método possua suas próprias vantagens e desvantagens. De uma forma geral, o método dos pares é mais seguro pois é menos sensível às suposições : independe da homoscedasticidade do modelo e tende a ser mais robusto quando este se mostra heterocedástico ; contudo, pode produzir intervalos de confiança de maiores comprimentos e não apresentar uma boa performance na presença de distribuições com alto grau de assimetria. Por sua vez, o método dos resíduos é preferível ao dos pares quando o tipo do modelo é especificado e tende a produzir intervalos de confiança de menores comprimentos

2.1.4 *Bootstrap selvagem*

No ano de 1986, Wu propôs uma técnica alternativa àquelas desenvolvidas por Efron para estimação em modelos de regressão. Seu intuito era construir um procedimento a ser utilizado tanto sob homoscedasticidade quanto sob heteroscedasticidade. A técnica foi denominada *Bootstrap* selvagem, também conhecida como *Bootstrap* ponderado. Consiste em gerar multiplicadores (pesos) para as variáveis latentes e_i por meio da geração de variáveis aleatórias independentes e identicamente distribuídas. Os passos referentes ao *Bootstrap* selvagem para o modelo linear seguem abaixo.

1. Extrai-se uma quantidade n de variáveis iid $V_1, \cdots, V_n$ a partir de uma distribuição especificada. Essas variáveis terão média nula e variância unitária ;

2. Reamostra-se com reposição o vetor dessas variáveis ;

3. Estima-se o vetor dos parâmetros $\hat{\beta}$ pelos métodos usuais ;

4. Para cada variável V_i, avaliamos y_i^* tal que

$$y_i^* = g_i(\hat{\beta}) + V_i \cdot e_i$$

5. Para cada y^*, calcula-se a estimativa *Bootstrap*

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

6. Repetem-se os passos 2, 4 e 5 em um total de R vezes, gerando o vetor de estimativas *Bootstrap*.

A técnica possui muitas variantes encontradas na literatura, que se diferenciam quanto à distribuição das variáveis geradas no primeiro passo, a exemplo da normal padrão. O *Bootstrap* selvagem tende a ser mais performático em relação ao método dos pares sob a suposição de heteroscedasticidade. Por fim, vale ressaltar que sob as condições de regularidade o método produz estimadores assintoticamente consistentes.

2.1.5 *Matriz de variância-covariância*

Em todas as técnicas *Bootstrap* apresentadas, a análise da variabilidade dos estimadores foi realizada segundo a matriz de variância-covariância proposta por Efron.

$$\sum{}^{*} = \frac{1}{R-1} \cdot \sum_{j=1}^{R} (\hat{\beta}_j^* - \beta^*) \cdot (\hat{\beta}_j^* - \beta^*)^T$$

em que

1. R é quantidade de reamostras fabricadas;

2. $\hat{\beta}_j^*$ é a j-ésima estimativa *Bootstrap* para β;

3. $\beta^* = \dfrac{1}{R} \cdot \displaystyle\sum_{j=1}^{R} \hat{\beta}_j^*$

2.1.6 *Intervalos de confiança Bootstrap*

A estimativa intervalar tradicional para os coeficientes de regressão foi comparado com os intervalos de confiança oriundos dos métodos *Bootstrap*. Dentre os procedimentos aplicados para a construção destes intervalos, destacam-se o intervalo percentil e o intervalo baseado na suposição de normalidade para as estimativas geradas pelas reamostras.

O intervalo percentil fundamenta-se nas separatrizes do conjunto de estimativas *Bootstrap*. Por exemplo, um intervalo com nível de confiança de 95% é aquele que compreende 95% das estimativas. De forma geral, o intervalo percentil com nível de confiança $(1-2\,\alpha)\cdot 100\%$ para um parâmetro é definido como se segue.

$$IC_{(1-2\,\alpha)}(\theta) = \left[P_{\alpha}(\hat{\theta}^*)\,;\, P_{1-\alpha}(\hat{\theta}^*) \right]$$

em que

1. θ é o parâmetro de interesse;

2. $\hat{\theta}$ é a estimativa *Bootstrap* para θ;

3. $P_{\alpha}(\hat{\theta}^*)$ é a separatriz de ordem $(\alpha)\cdot 100$ das estimativas *Bootstrap*;

4. $P_{1-\alpha}(\hat{\theta}^*)$ é a separatriz de ordem $(1-\alpha)\cdot 100$ das estimativas *Bootstrap*.

Caso seja verificado normalidade para as estimativas geradas, se mostra possível a construção de intervalos baseados nos pontos críticos da distribuição normal padrão.

$$IC_{(1-\alpha)}(\theta) = \left[\theta^* \pm z_{1-\frac{\alpha}{2}} \cdot EP(\hat{\theta}^*)\right]$$

em que

1. θ^* é a média das estimativas *Bootstrap* para θ ;

2. $z_{1-\frac{\alpha}{2}}$ é o quantil de $N_{(0,1)}$;

3. $EP(\hat{\theta}^*)$ é o erro padrão das estimativas *Bootstrap*.

3 RESULTADOS

A priori verificou-se a adequação do modelo linear simples (MRLS) ao conjunto de dados. O coeficiente de correlação de Pearson calculado foi cerca de 0,98, indicando uma forte correlação linear entre as variáveis. O ajuste do modelo foi realizado no *software* RStudio, fornecendo $R^2 = 0,97$ e as seguintes estimativas para os parâmetros.

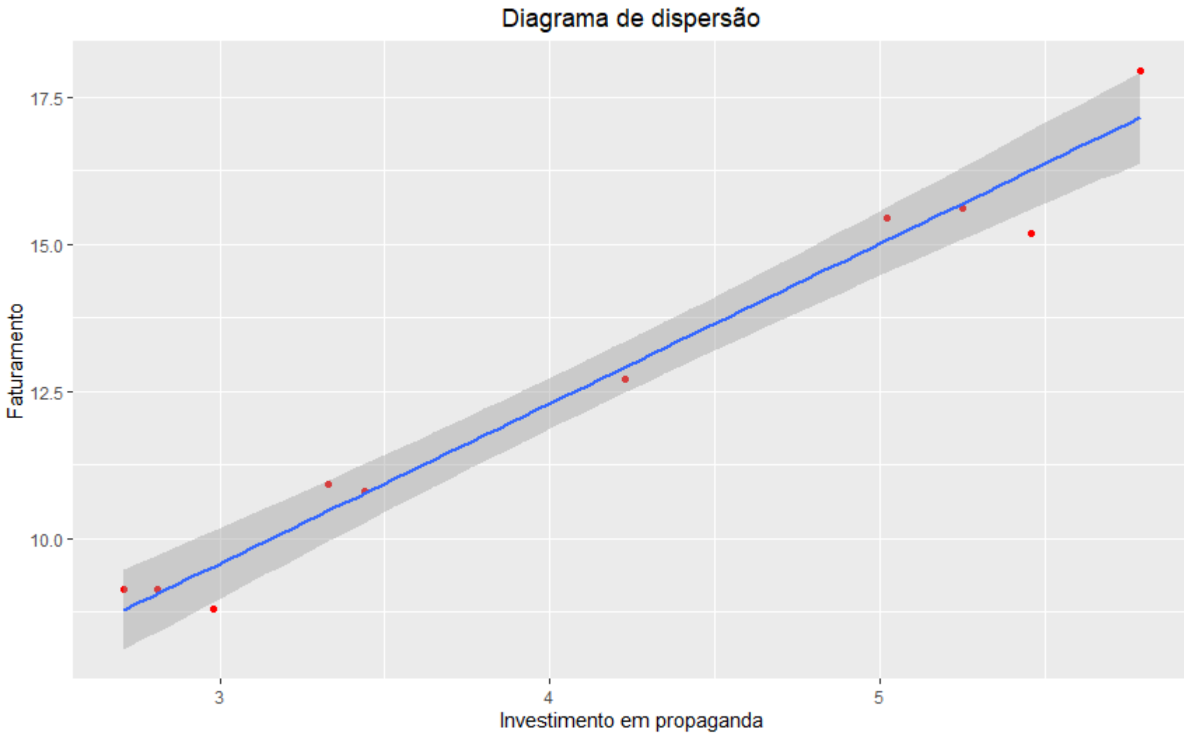
Quadro 1 - Estimativas sob o MRLS

MRLS		
Parâmetro	Estimativa	Erro padrão
β_0	1,418	0,702
β_1	2,718	0,165

Fonte: Elaborado pelo autor

A obtenção das estimativas usuais permitiu a construção de um gráfico de dispersão com a função de regressão ajustada. O gráfico de fato demonstra a validade de um modelo linear da forma $y_i = \beta_0 + \beta_1\, x_i + e_i$.

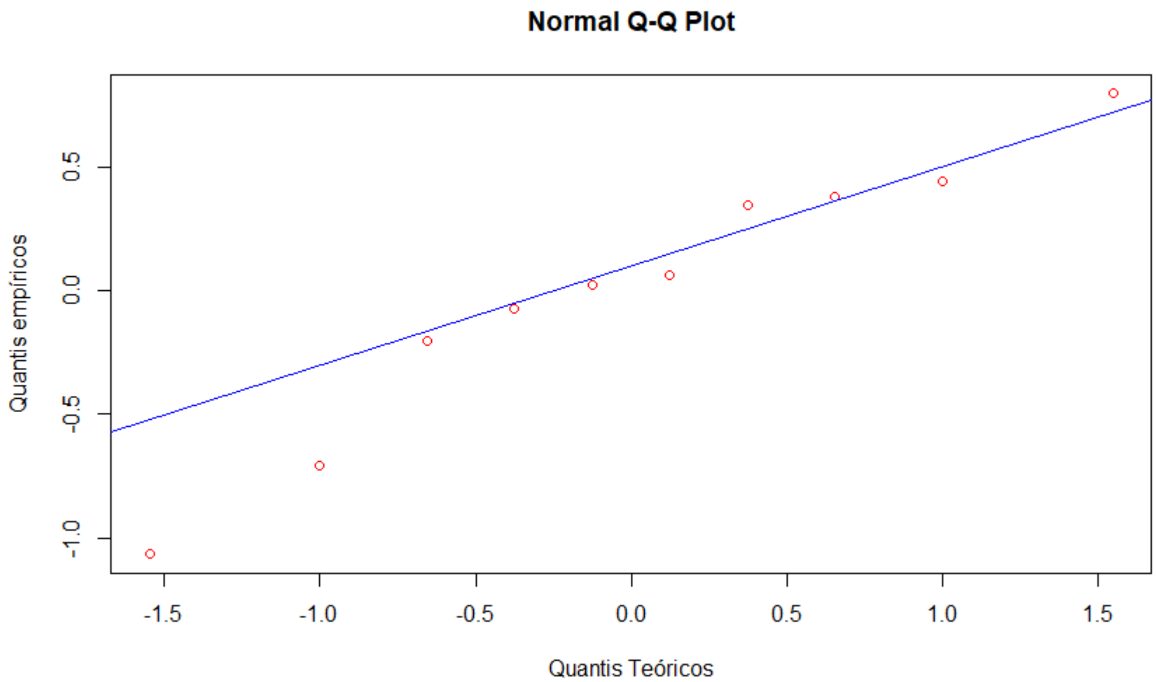
Figura 1 - Reta de regressão ajustada para o conjunto de dados



Fonte: Elaborado pelo autor

Percebeu-se por meio de um gráfico quantil-quantil que a distribuição dos resíduos ordinários obtidos pelo ajuste apresenta características assimétricas, portanto sendo inválida a suposição de normalidade. Tal fato pode causar impactos para a realização de procedimentos inerentes à inferência de segunda ordem sob técnicas tradicionais, como a construção de intervalos de confiança para os coeficientes.

Figura 2 - Gráfico quantil-quantil para os resíduos

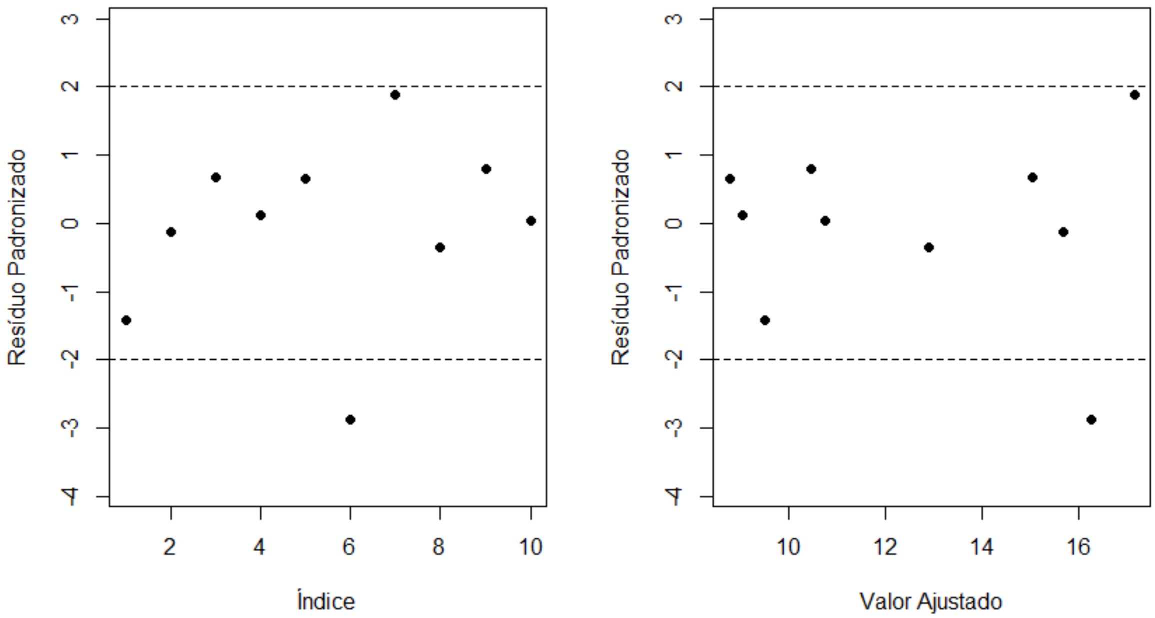


Fonte: Elaborado pelo autor

A figura acima sinaliza assimetria para a distribuição dos resíduos pela fato de que que a dispersão dos pontos no espaço cartesiano não se mostra propício para a formação de um segmento de um reta , implicando que a distribuição destes mesmos pontos possui certa curvatura.

Além disso, averiguou-se a plausibilidade da suposição de homoscedasticidade para o modelo por meio da análise gráfica dos resíduos padronizados (Figura 3) : estes se mostraram aleatoriamente dispersos em volta da origem. Vale ressaltar que a verificação de homoscedasticidade viabiliza a aplicação da reamostragem dos resíduos como um meio de estimação dos parâmetros e construção de intervalos de confiança.

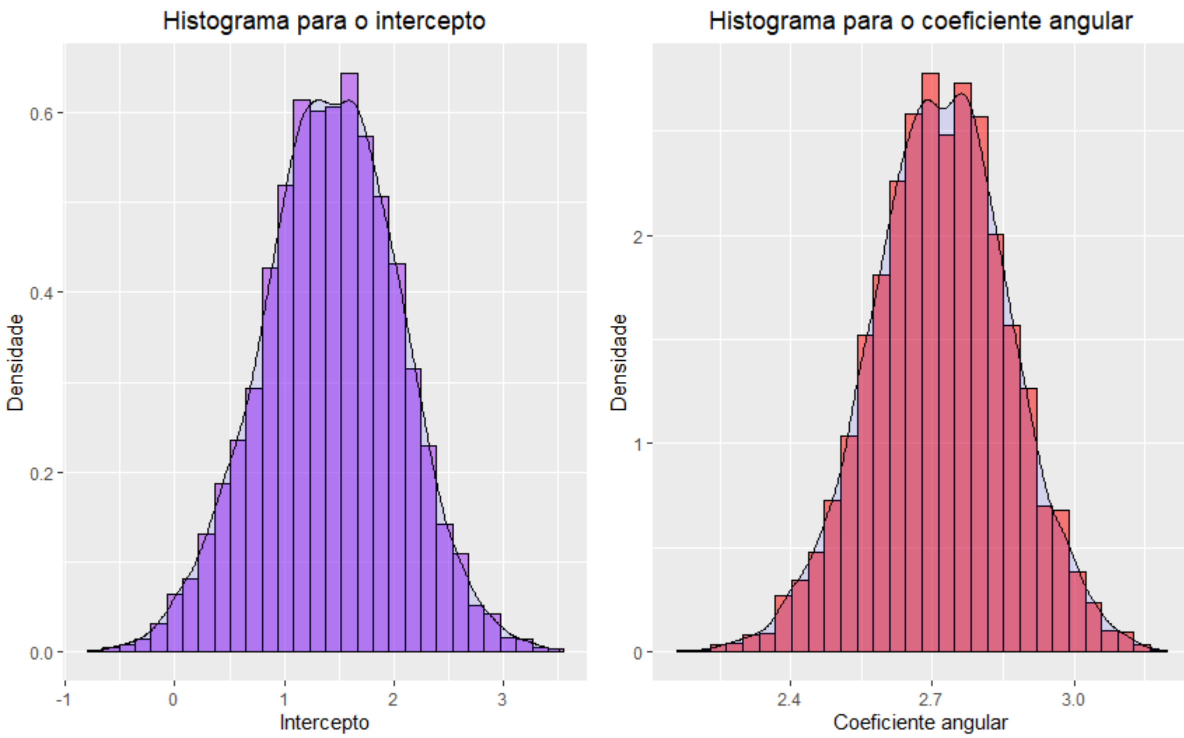
Figura 3 - Dispersão dos resíduos padronizados



Fonte: Elaborado pelo autor

A reamostragem dos resíduos foi aplicada a partir do pacote *lmboot* do *software* RStudio, com a geração de 5000 amostras *Bootstrap*. Analisou-se por meio de histogramas a distribuição das estimativas *Bootstrap* para cada parâmetro (Figura 4) : os gráficos obtidos foram bastante similares e indicaram simetria em relação à distribuição das estimativas. Dessa forma, diagramas quantil-quantil foram construídos com o intuito de embasar a suposição de normalidade, já que a verificação desta suposição permite a construção de intervalos de confiança fundamentados nos pontos críticos da normal padrão.

Figura 4 - Estimação pela reamostragem dos resíduos

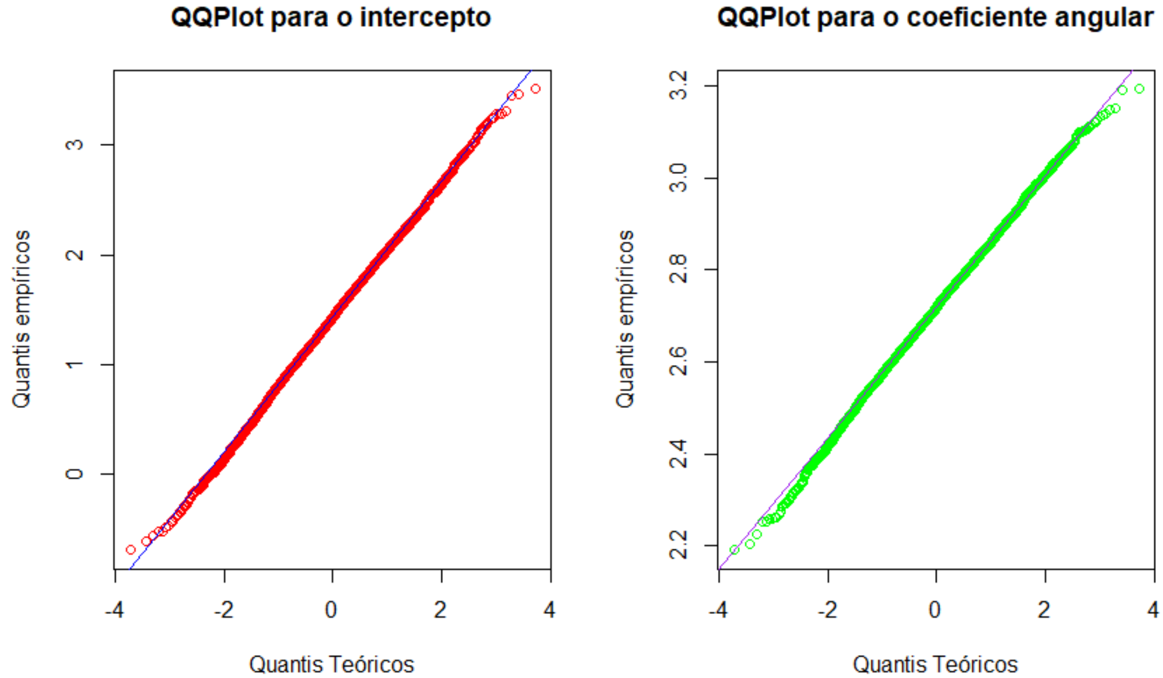


Fonte: Elaborado pelo autor

Nota-se pelos dois histogramas uma grande concentração das estimativas em intervalos específicos: para o intercepto houve uma grande concentração de valores entre 1 e 2; em relação ao coeficiente angular, os valores concentraram-se entre 2,55 e 2,85, aproximadamente. Estes intervalos contêm as estimativas obtidas pelo método usual de mínimos quadrados, acarretando em um baixo viés das estimativas *Bootstrap* em relação às tradicionais.

Por fim, foi averiguado que de fato a distribuição das estimativas para cada parâmetro é normal uma vez que a dispersão dos pontos no espaço cartesiano tornou viável a formação de um segmento de um reta de ligação entre os pontos (Figura 5).

Figura 5 - Gráficos quantil-quantil para os coeficientes



Fonte : Elaborado pelo autor

Com base na matriz de variância-covariância estimada pelo método da reamostragem dos resíduos, nota-se que os estimadores *Bootstrap* apresentam maior eficiência em relação aos estimadores padrão : para β_0, o erro padrão obtido foi cerca de 0,622 ; já para β_1 observou-se uma variabilidade de cerca de 0,145.

$$\sum\nolimits_R^* = \begin{bmatrix} 0,387 & -0,087 \\ -0,087 & 0,021 \end{bmatrix}$$

A obtenção da matriz acima possibilita a construção de intervalos de confiança baseados na distribuição normal padrão. Os Quadros 2 e 3 comparam o intervalo de confiança usual com os intervalos *Bootstrap* obtidos para cada coeficiente, ao nível de significância de 5%. De uma forma geral, os intervalos *Bootstrap* apresentaram as menores amplitudes, demonstrando assim maior precisão em relação ao método tradicional.

Quadro 2 - Resíduos : Intervalos para o intercepto

Comparação de IC's para β_0			
Método	LI	LS	Amplitude
Usual	-0,201	3,037	3,238
Normal	0,202	2,641	2,439
Percentil	0,157	2,617	2,460

Fonte : Elaborado pelo autor

Quadro 3 - Resíduos : Intervalos para o coeficiente angular

Comparação de IC's para β_1			
Método	LI	LS	Amplitude
Usual	2,337	3,098	0,761
Normal	2,432	3,002	0,570
Percentil	2,423	3,001	0,578

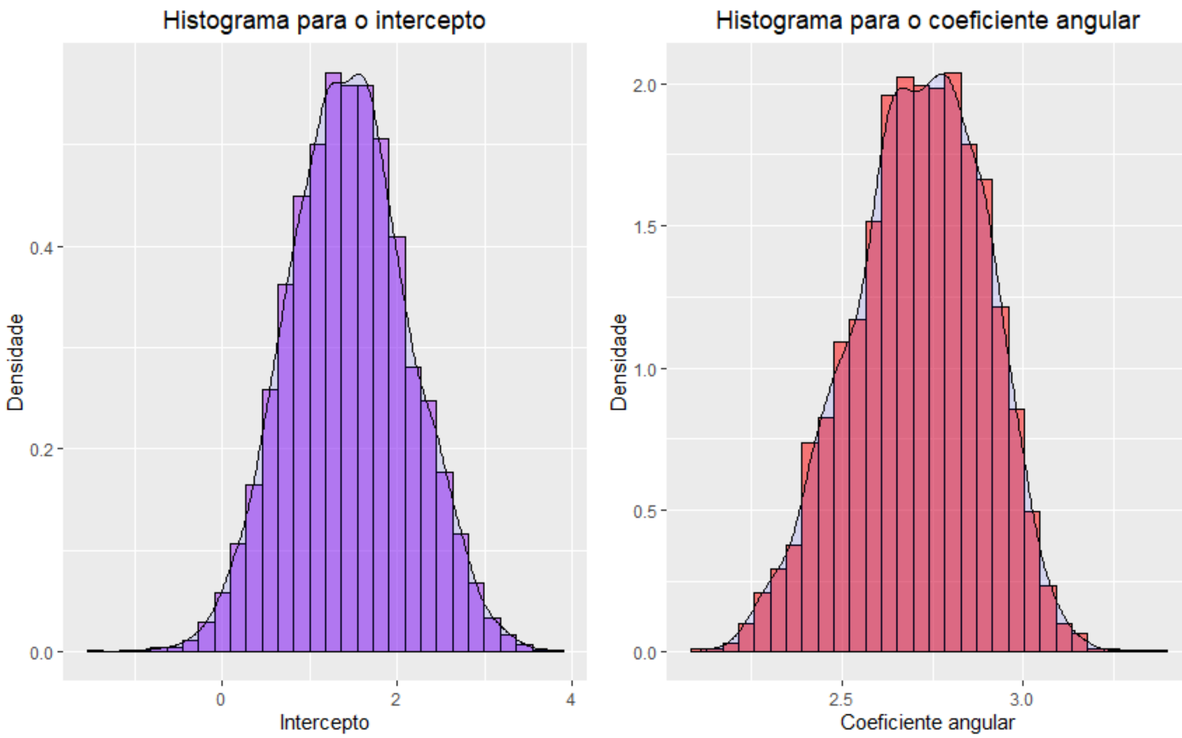
Fonte : Elaborado pelo autor

Enfatiza-se que o intervalo usual para o intercepto engloba valores negativos, causando incoerência em sua interpretação pela natureza dos dados. De forma análoga à reamostragem dos resíduos, ajustou-se um modelo de regressão linear simples por meio da reamostragem dos pares com a geração de 5000 amostras *Bootstrap* para o vetor de parâmetros a partir do pacote *lmboot*. O novo método produziu estimadores com maior variabilidade em comparação à reamostragem dos resíduos, fato comprovado por meio da matriz de variância-covariância estimada.

$$\sum\nolimits_P^* = \begin{bmatrix} 0,467 & -0,121 \\ \\ -0,121 & 0,033 \end{bmatrix}$$

Assim como a técnica anterior, os histogramas para cada coeficiente foram bastante similares, sendo observado normalidade para a distribuição das estimativas.

Figura 6 - Estimação pela reamostragem dos pares



Fonte : Elaborado pelo autor

Ao nível de significância de 5%, os intervalos de confiança construídos com a aplicação da reamostragem dos pares foram mais precisos em relação ao tradicional, no entanto apresentaram maior amplitude em comparação à reamostragem dos resíduos.

Quadro 4 - Pares : Intervalos para o intercepto

Comparação de IC's para β_0			
Método	LI	LS	Amplitude
Usual	-0,201	3,037	3,238
Normal	0,102	2,782	2,680
Percentil	0,137	2,795	2,658

Fonte : Elaborado pelo autor

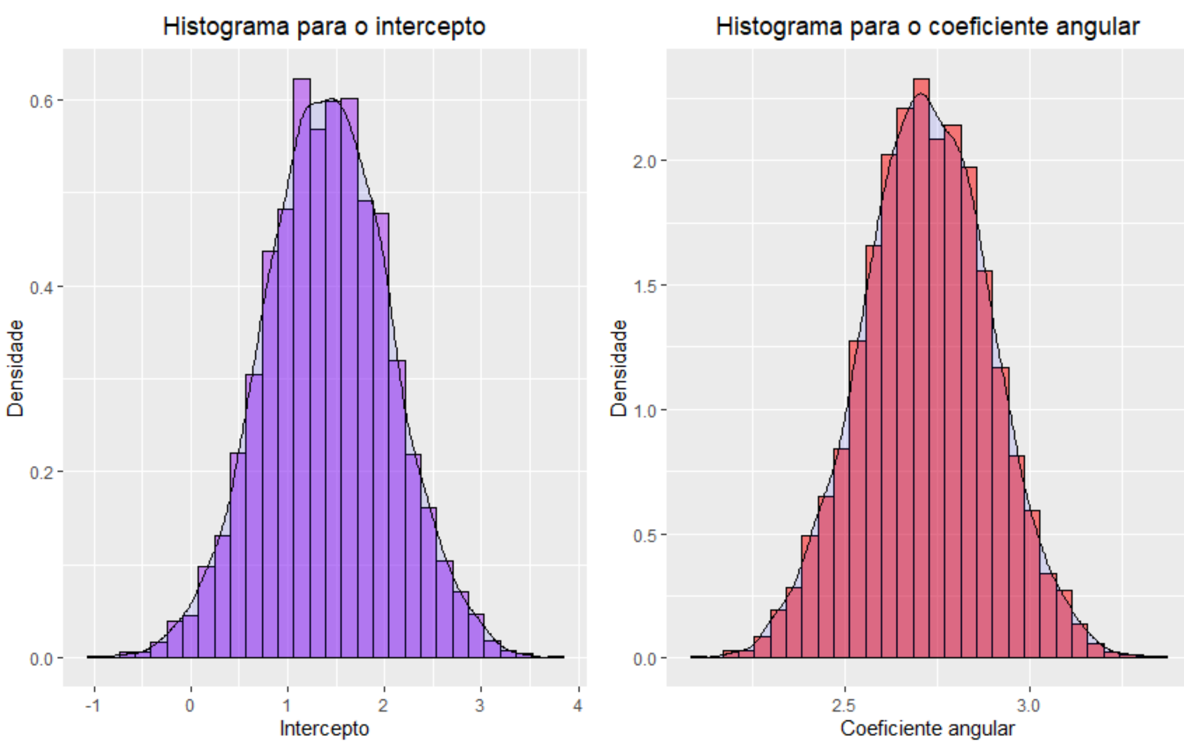
Quadro 5 - Pares : Intervalos para o coeficiente angular

Comparação de IC's para β_1			
Método	LI	LS	Amplitude
Usual	2,337	3,098	0,761
Normal	2,354	3,068	0,714
Percentil	2,331	3,028	0,697

Fonte : Elaborado pelo autor

Por último, o *Bootstrap* selvagem foi utilizado para o ajuste do modelo linear, com a geração de 5000 reamostras e a extração das variáveis aleatórias a partir da normal padrão. Da mesma forma que as técnicas anteriores, ocorreu a verificação de normalidade em relação à distribuição das estimativas.

Figura 7 - Estimação pelo *Bootstrap* selvagem



Fonte : Elaborado pelo autor

O *Bootstrap* selvagem gerou estimadores com menor variabilidade em comparação à reamostragem dos pares, porém menos eficientes em relação ao método dos resíduos, como se observa pela matriz de variância-covariância estimada.

$$\sum_S^* = \begin{bmatrix} 0,411 & -0,108 \\ & \\ -0,108 & 0,030 \end{bmatrix}$$

Por último, os intervalos calculados com nível de confiança de 95% gerados a partir do *Bootstrap* ponderado foram mais precisos em relação ao tradicional. Vale ressaltar que o método teve uma melhor performance geral em comparação à reamostragem dos pares, pois produziu intervalos com menores amplitudes.

Quadro 6 - *Bootstrap* ponderado: Intervalos para o intercepto

Comparação de IC's para β_0			
Método	LI	LS	Amplitude
Usual	-0,201	3,037	3,238
Normal	0,154	2,670	2,516
Percentil	0,144	2,690	2,546

Fonte: Elaborado pelo autor

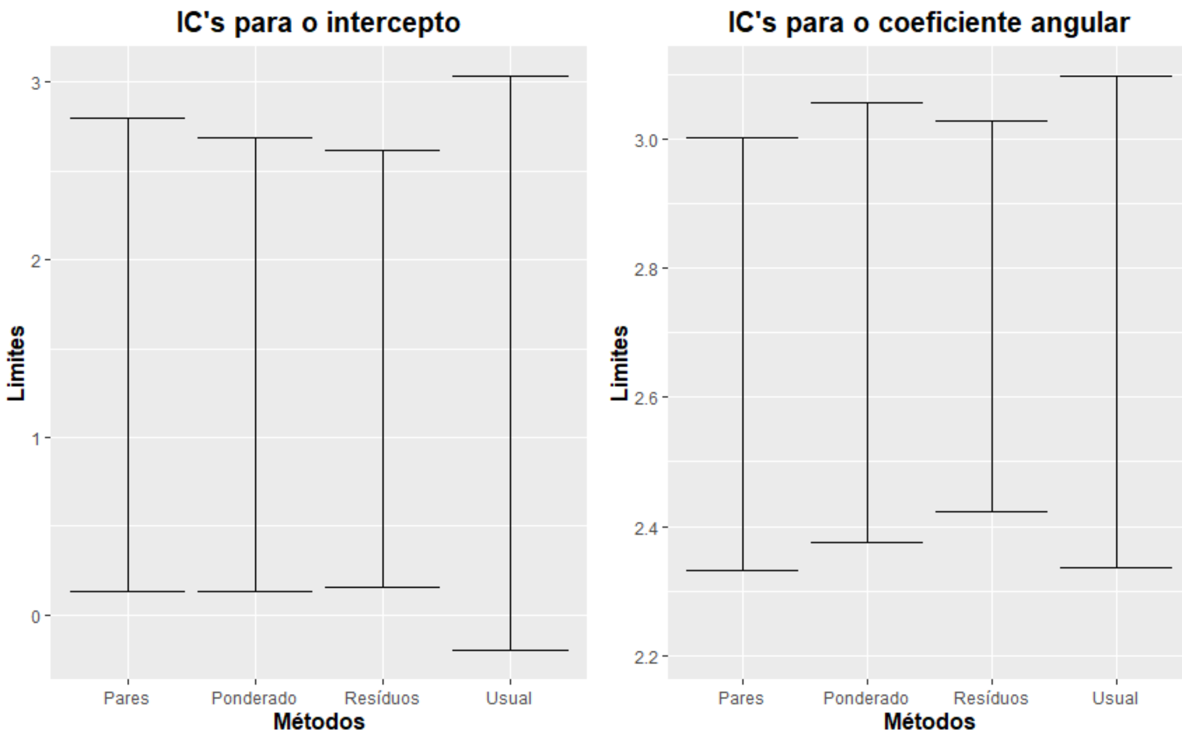
Quadro 7 - *Bootstrap* ponderado: Intervalos para o coeficiente angular

Comparação de IC's para β_1			
Método	LI	LS	Amplitude
Usual	2,337	3,098	0,761
Normal	2,381	3,057	0,676
Percentil	2,376	3,056	0,680

Fonte: Elaborado pelo autor

Diante dos resultados obtidos com a aplicação dos três métodos, a figura abaixo realiza uma comparação entre o intervalo usual e os intervalos percentis conforme o método para cada parâmetro.

Figura 8 - Comparação de intervalos de confiança



Fonte : Elaborado pelo autor

Pelo exposto, nota-se que :

1. Todos os três métodos produziram melhores estimativas se comparados ao procedimento padrão.

2. A reamostragem dos resíduos apresentou a melhor performance geral, produzindo os estimadores mais eficientes e os intervalos mais precisos.

3. A reamostragem dos pares, dentre os métodos *Bootstrap*, gerou os estimadores de maior variabilidade e os intervalos com maiores amplitudes.

4. O *Bootstrap* ponderado demonstrou uma melhor performance em relação à reamostragem dos pares, porém produziu intervalos mais amplos em comparação à reamostragem dos resíduos.

4 CONCLUSÃO

As características assimétricas observadas referentes à distribuição probabilística do faturamento mensal de fato impactaram na realização de procedimentos inerentes à inferência de segunda ordem sob técnicas tradicionais, a exemplo da construção de intervalos de confiança. O fato de que os métodos *Bootstrap* apresentados se mostraram mais performáticos, gerando estimativas intervalares mais coerentes e precisas, torna plausível a preferência destes métodos em relação aos usuais. Isso enfatiza a utilidade, praticidade e importância do *Bootstrap* para questões referentes à Estatística Aplicada, principalmente em situações mais gerais e de difícil tratamento caracterizadas pela ausência de informações a respeito da natureza probabilística dos dados.

REFERÊNCIAS

BUSSAB, Wilton de O. ; MORETTIN, Pedro A. **Estatística Básica**. 9 ed. São Paulo : Saraiva, 2017.

CHERNICK, Michael R. **Bootstrap Methods**. 2. ed. Newtown : John Wiley & Sons, 2008.

CHERNICK, Michael R ; LABUDDE, Robert A.**An introduction to bootstrap methods with applications to R**. Newtown : John Wiley & Sons, 2011.

DAVISON, Anthony Christopher ; HINKLEY, David Victor. **Bootstrap methods and their application**. Cambridge : Cambridge University Press, 1997.

EFRON, Bradley. Bootstrap Methods : Another Look at the Jackknife.**The Annals of Statistics**, California, v.7, p.1-26, jan. 1979.

EFRON, Bradley ; TIBSHIRANI, Rob.**An introduction to the bootstrap**. New York : CRC Press, 1993.

HELWIG, Nathaniel E. Bootstrap Resampling. 04 jan. 2017. Apresentação do Power Point. Disponível em : http://users.stat.umn.edu/ helwig/notes/boot-Notes.pdf.

INÁCIO, Felipe Chaves. **Bootstrap ponderado : uma avaliação númerica**. 2004. Dissertação (Mestrado) - Curso de Estatística, Universidade Federal do Pernambuco, Recife, 2004.Disponível em : https://repositorio.ufpe.br/bitstream/123456789/6597/1/arquivo7260_1.pdf.
Acesso em : 05 out. 2022.

RSTudio Team (2020). Rstudio : Integrated Development for RStudio, PBC, Boston, MA. Disponível em : http://www.rstudio.com/. Acesso em : 24 out.2022.

APÊNDICE A - Algoritmos na linguagem R

```
#Dados
x = c(2.98,5.25,5.02,2.81,2.71,5.46,5.79,4.23,3.33,3.44) # investimento com propaganda
y =c(8.81,15.61,15.44,9.12,9.13,15.19,17.95,12.71,10.91,10.79) # faturamento
dados = cbind(x,y)
summary(x)
summary(y)
cor(x,y) # Valor alto indicando forte correlação linear positiva
# Gráfico de dispersão
require(ggplot2)
require(gridExtra)
ggplot(data.frame(dados), aes(x=x, y=y)) + geom_point(col='red')+geom_smooth(method=lm)+
xlab("Investimento em propaganda")+ylab("Faturamento")+ggtitle("Diagrama de dispersão")
+theme(plot.title = element_text(hjust = 0.5))

# Ajuste do MRLS
modelo = lm(y ~ x)
summary(modelo)
confint(modelo)
```

```
# Os resíduos são normais?
qqnorm(modelo$residuals,xlab = ”Quantis Teóricos”,ylab=”Quantis empíricos”,col = ’red’)
qqline(modelo$residuals,col = ’blue’)

# O modelo é homoscedástico?
fit.model = modelo
X = model.matrix(fit.model)
nl = nrow(X)
p = ncol(X)
H = X%*%solve(t(X)%*%X)%*%t(X)
h = diag(H)
lms = summary(fit.model)
s = lms$sigma
r = resid(lms)
ts = r/(s*sqrt(1-h))
di = (1/p)*(h/(1-h))*(ts)*(ts)
si = lm.influence(fit.model)$sigma
tsi = r/(si*sqrt(1-h))
a = max(tsi)
b = min(tsi)
par(mfrow=c(1,2))
plot(tsi,xlab=”Índice”, ylab=”Resíduo Padronizado”, ylim=c(b-1,a+1), pch=16) abline(2,0,lty=2)
abline(-2,0,lty=2) plot(fitted(fit.model),tsi,xlab=”Valor Ajustado”, ylab=”Resíduo Padro-
nizado”, ylim=c(b-1,a+1), pch=16) abline(2,0,lty=2) abline(-2,0,lty=2)
# Coeficientes de regressão
beta0 = as.numeric(modelo$coefficients[1])
beta1 = as.numeric(modelo$coefficients[2])
```

```
# Reamostragem dos resíduos
require(lmboot)
set.seed(123)
n = length(y)
B = 5000
residual_obj = residual.boot(y ~ x,B=B,seed=123)
plot1=ggplot(data.frame(residual_obj$bootEstParam[,1]),aes(x=residual_obj$bootEstParam[,1]))
+geom_histogram(aes(y=..density..),colour=1,fill="purple",alpha=0.5,position="identity")
+geom_density(alpha=.1,fill='blue')+xlab("Intercepto")+ylab("Densidade")+ggtitle("Histograma
para o intercepto")+theme(plot.title = element_text(hjust = 0.5))
plot2 = ggplot(data.frame(residual_obj$bootEstParam[,2]),aes(x=residual_obj$bootEstParam[,2]))
+geom_histogram(aes(y=..density..),colour=1,fill="red",alpha=0.5,position="identity")
+geom_density(alpha=.1,fill='blue')+xlab("Coeficiente angular")+ylab("Densidade")+
ggtitle("Histograma para o coeficiente angular")+theme(plot.title = element_text(hjust =
0.5))
grid.arrange(plot1, plot2, ncol=2)

par(mfrow = c(1,2))
qqnorm(residual_obj$bootEstParam[,1],xlab = "Quantis Teóricos",ylab="Quantis empíricos",col
= 'red',main = "QQPlot para o intercepto")
qqline(data.frame(residual_obj$bootEstParam[,1]),col='blue')
qqnorm(residual_obj$bootEstParam[,2],xlab = "Quantis Teóricos",ylab="Quantis empíricos",col
= 'green',main = "QQPlot para o coeficiente angular")
qqline(data.frame(residual_obj$bootEstParam[,2]),col='purple')
beta0_boot = residual_obj$bootEstParam[,1]
beta1_boot = residual_obj$bootEstParam[,2]
```

```
# Matriz de variância-covariância estimada
mboot = as.matrix((1/B)*colSums(residual_obj$bootEstParam),nrow=1)
lista = list()
for(i in 1:B) {
lista[[ i ]] =as.matrix((residual_obj$bootEstParam[i,] - mboot) %*% t(residual_obj$bootEstParam[i,]
- mboot),nrow=2,ncol=2)
}
mvc = 1/(B-1)*Reduce("+", lista)
colnames(mvc)=NULL
rownames(mvc)=NULL
mvc

# Intervalos de confiança
# Percentil
r0=sort(beta0_boot)
ICp_beta0 = c(r0[round(0.025*B)], r0[round(0.975*B)])
ICp_beta0
r1=sort(beta1_boot)
ICp_beta1 = c(r1[round(0.025*B)], r1[round(0.975*B)])
ICp_beta1
# Paramétrico
mb0 = mean(beta0_boot)
se0 = sqrt(mvc[1,1])
Icn_beta0 = c(mb0-1.96*se0,mb0+1.96*se0)
Icn_beta0
mb1= mean(beta1_boot)
se1 = sqrt(mvc[2,2])
Icn_beta1 = c(mb1-1.96*se1,mb1+1.96*se1)
Icn_beta1
```

```
# Bootstrap ponderado
pond_obj = wild.boot(y ~ x,B=B,seed=123)
plot1=ggplot(data.frame(pond_obj$bootEstParam[,1]),aes(x=pond_obj$bootEstParam[,1]))
+geom_histogram(aes(y=..density..),colour=1,fill="purple",alpha=0.5,position="identity")
+geom_density(alpha=.1,fill='blue')+xlab("Intercepto")+ylab("Densidade")+ggtitle("Histograma
para o intercepto")+theme(plot.title = element_text(hjust = 0.5))
plot2 = ggplot(data.frame(pond_obj$bootEstParam[,2]),aes(x=pond_obj$bootEstParam[,2]))
+geom_histogram(aes(y=..density..),colour=1,fill="red",alpha=0.5,position="identity")
+geom_density(alpha=.1,fill='blue')+xlab("Coeficiente angular")+ylab("Densidade")+
ggtitle("Histograma para o coeficiente angular")+theme(plot.title = element_text(hjust =
0.5))
grid.arrange(plot1, plot2, ncol=2)

par(mfrow = c(1,2))
qqnorm(pond_obj$bootEstParam[,1],xlab = "Quantis Teóricos",ylab="Quantis empíricos",col
= 'red',main = "QQPlot para o intercepto")
qqline(data.frame(pond_obj$bootEstParam[,1]),col='blue')
qqnorm(pond_obj$bootEstParam[,2],xlab = "Quantis Teóricos",ylab="Quantis empíricos",col
= 'green',main = "QQPlot para o coeficiente angular")
qqline(data.frame(pond_obj$bootEstParam[,2]),col='purple')
beta0_boot = pond_obj$bootEstParam[,1]
beta1_boot = pond_obj$bootEstParam[,2]

# Matriz de variância-covariância estimada
mboot = as.matrix((1/B)*colSums(pond_obj$bootEstParam),nrow=1)
lista = list()
for(i in 1:B) {
lista[[ i ]] =as.matrix((pond_obj$bootEstParam[i,] - mboot) %*% t(pond_obj$bootEstParam[i,]
- mboot),nrow=2,ncol=2)
```

```
}
mvc = 1/(B-1)*Reduce("+", lista)
colnames(mvc)=NULL
rownames(mvc)=NULL
mvc

# Intervalos de confiança
# Percentil
r0=sort(beta0_boot)
ICp_beta0 = c(r0[round(0.025*B)], r0[round(0.975*B)])
ICp_beta0
r1=sort(beta1_boot)
ICp_beta1 = c(r1[round(0.025*B)], r1[round(0.975*B)])
ICp_beta1

# Paramétrico
mb0 = mean(beta0_boot)
se0 = sqrt(mvc[1,1])
Icn_beta0 = c(mb0-1.96*se0,mb0+1.96*se0)
Icn_beta0

mb1= mean(beta1_boot)
se1 = sqrt(mvc[2,2])
Icn_beta1 = c(mb1-1.96*se1,mb1+1.96*se1)
Icn_beta1
```

```
# Reamostragem dos pares
pairs_obj = paired.boot(y ~ x,B=B,seed=123)
plot1=ggplot(data.frame(pairs_obj$bootEstParam[,1]),aes(x=pairs_obj$bootEstParam[,1]))
+geom_histogram(aes(y=..density..),colour=1,fill="purple",alpha=0.5,position="identity")
+geom_density(alpha=.1,fill='blue')+xlab("Intercepto")+ylab("Densidade")+ggtitle("Histograma
para o intercepto")+theme(plot.title = element_text(hjust = 0.5))
plot2 = ggplot(data.frame(pairs_obj$bootEstParam[,2]),aes(x=pairs_obj$bootEstParam[,2]))
+geom_histogram(aes(y=..density..),colour=1,fill="red",alpha=0.5,position="identity")
+geom_density(alpha=.1,fill='blue')+xlab("Coeficiente angular")+ylab("Densidade")+
ggtitle("Histograma para o coeficiente angular")+theme(plot.title = element_text(hjust =
0.5))
grid.arrange(plot1, plot2, ncol=2)

par(mfrow = c(1,2))
qqnorm(pairs_obj$bootEstParam[,1],xlab = "Quantis Teóricos",ylab="Quantis empíricos",col
= 'red',main = "QQPlot para o intercepto")
qqline(data.frame(pairs_obj$bootEstParam[,1]),col='blue')
qqnorm(pairs_obj$bootEstParam[,2],xlab = "Quantis Teóricos",ylab="Quantis empíricos",col
= 'green',main = "QQPlot para o coeficiente angular")
qqline(data.frame(pairs_obj$bootEstParam[,2]),col='purple')
beta0_boot = pairs_obj$bootEstParam[,1]
beta1_boot = pairs_obj$bootEstParam[,2]

# Matriz de variância-covariância estimada
mboot = as.matrix((1/B)*colSums(pairs_obj$bootEstParam),nrow=1)
lista = list()
for(i in 1:B) {
lista[[ i ]] =as.matrix((pairs_obj$bootEstParam[i,] - mboot) %*% t(pairs_obj$bootEstParam[i,]
- mboot),nrow=2,ncol=2)
```

```
}
mvc = 1/(B-1)*Reduce("+", lista)
colnames(mvc)=NULL
rownames(mvc)=NULL
mvc

# Intervalos de confiança
# Percentil
r0=sort(beta0_boot)
ICp_beta0 = c(r0[round(0.025*B)], r0[round(0.975*B)])
ICp_beta0
r1=sort(beta1_boot)
ICp_beta1 = c(r1[round(0.025*B)], r1[round(0.975*B)])
ICp_beta1

# Paramétrico
mb0 = mean(beta0_boot)
se0 = sqrt(mvc[1,1])
Icn_beta0 = c(mb0-1.96*se0,mb0+1.96*se0)
Icn_beta0

mb1= mean(beta1_boot)
se1 = sqrt(mvc[2,2])
Icn_beta1 = c(mb1-1.96*se1,mb1+1.96*se1)
Icn_beta1
```

# Comparação dos intervalos de confiança
intervals_b0 = data.frame(Método = c(”Usual”,”Resíduos”,”Pares”,”Ponderado”),
LI = c(-0.201,0.157,0.137,0.134),LS = c(3.037,2.617,2.795,2.69))
plot_bo = ggplot(intervals_b0, aes(x=intervals_b0[,1],y = runif(4,-0.5,3)))+geom_errorbar
(aes(ymin = LI, ymax = LS))+xlab(”Métodos”)+ylab(”Limites”)+ggtitle(”IC’s para o intercepto”)+theme(plot.title = element_text(hjust = 0.5,face = ’bold’, size=’15’),axis.title.x = element_text(face = ’bold’, size=’12’),axis.title.y = element_text(face = ’bold’, size=’12’))

intervals_b1 = data.frame(Método = c(”Usual”,”Resíduos”,”Pares”,”Ponderado”),
LI = c(2.337,2.423,2.331,2.376),LS = c(3.098,3.028,3.001,3.056))
plot_b1 = ggplot(intervals_b1, aes(x=intervals_b1[,1],y = runif(4,1,3.2)))+geom_errorbar
(aes(ymin = LI, ymax = LS))+xlab(”Métodos”)+ylab(”Limites”)+ggtitle(”IC’s para o coeficiente angular”)+theme(plot.title = element_text(hjust = 0.5,face = ’bold’, size=’15’),axis.title.x = element_text(face = ’bold’, size=’12’),axis.title.y = element_text(face = ’bold’, size=’12’))

grid.arrange(plot_bo,plot_b1,ncol=2)