Revista brasileira de iniciação científica
ISSN: 2359-232X

MÉTODOS BOOTSTRAP: UMA APLICAÇÃO EM DADOS OLÍMPICOS

BOOTSTRAP METHOD: AN APPLICATION IN OLYMPIC DATA

MÉTODO BOOTSTRAP: UNA APLICACIÓN A DATOS OLÍMPICOS

Luciano S. dos Santos[1]
Jalmar M. F. Carrasco[2]
Lizandra C. Fabio[3]

Resumo: Os métodos de reamostragem *bootstrap* têm ocupado um lugar importante no mundo da estatística pela simplicidade e poderio computacional. Ao longo da última década, extensões do método como *bootstrap* bayesiano, duplo, *bootknife,* entre outros, foram propostos. Sabe-se que, o método de estimação por máxima verossimilhança é amplamente utilizado para encontrar estimadores com boas propriedades, no entanto, em algumas situações práticas os estimadores obtidos são viciados, em particular, quando o tamanho da amostra é pequeno. Para contornar este problema, técnicas de reamostragem tipo *bootstrap* podem ser utilizadas. Deste contexto, este artigo tem como objetivo estudar as diferentes extensões do método de *bootstrap* e avaliar a metodologia estudada com o conjunto de dados referente aos jogos olímpicos que o Brasil tem participado; pretendendo conhecer a relação entre o número de medalhas olímpicas obtidas pelo Brasil com as variáveis: número de atletas participantes e o Produto Interno Bruto per capita nas diferentes edições.
Palavras-chave: Método *bootstrap.* Jogos olímpicos. Modelos lineares generalizados.

Abstract: The *bootstrap* resampling method has occupied an important place in the statistical world for its simplicity and computational power. Over the last decade, extensions of the method such as: Bayesian *bootstrap*, double *bootstrap*, bootknife, among others, has been proposed. It is known that the maximum likelihood estimation method is widely used to find estimators with good properties, however, in some practical situations the estimators obtained are biased, particularly when the sample size is small. To get around this problem, *bootstrap* resampling techniques can be used. Thus, this article aims to study the different extensions of the *bootstrap* method and evaluate the methodology studied with the data set referring to the Olympic Games in which Brazil has participated, aiming to know the relationship between the number of Olympic medals obtained by Brazil with the variables: number of participating athletes and Gross Domestic Product per capita in different editions.
Keywords: Bootstrap resampling methods. Olympic games. Generalized linear model.

[1] Graduando. Departamento de Estatística, Instituto de Matemática e Estatística. Universidade Federal da Bahia. E-mail: lucianno0800@gmail.com . ORCID: http://orcid.org/0000-0000-0002-9455
[2] Doutor. Departamento de Estatística, Instituto de Matemática e Estatística. Universidade Federal da Bahia. E-mail:carrasco.jalmar@ufba.br . ORCID: http://orcid.org/0000-0002-0983-1316
[3] Doutor. Departamento de Estatística, Instituto de Matemática e Estatística. Universidade Federal da Bahia. E-mail: Lizandra.fabio@ufba.br . ORCID: http://orcid.org/0000-0003-2910-5634

Revista brasileira de iniciação científica
ISSN: 2359-232X

Resumen: Los métodos de remuestreo *bootstrap* viene ocupando un lugar importante en el mundo de la estadística por su simplicidad y poder computacional. A lo largo de la última década, extensiones del método como *bootstrap* bayesiano, doble, *bootknife*, entre otros, fueron propuestos. Es sabido que, el método de estimación de máxima verosimilitud es ampliamente utilizado para encontrar estimadores con buenas propiedades, sin embargo, en algunas situaciones prácticas los estimadores obtenidos son sesgados, en particular, cuando el tamaño de muestra es pequeño. Para resolver este problema, técnicas de remuestreo tipo *bootstrap* pueden ser utilizados. De esta forma, este artículo tiene como objetivo estudiar las diferentes extensiones del método de *bootstrap* y validar la metodología estudiada con el conjunto de datos referente a los juegos olímpicos que Brasil participó; deseando conocer la relación entre el número de medallas olímpicas conseguidas por Brasil con las variables: número de atletas participantes y el Producto Bruto Interno per cápita en las diferentes ediciones.

Palabras-claves: Método b*ootstrap*. Juegos olímpicos. Modelos lineales generalizados.

Submetido 14/12/2020 **Aceito 01/08/2022** **Publicado 15/08/2022**

Revista brasileira de iniciação científica
ISSN: 2359-232X

Introdução

Os jogos olímpicos têm em sua essência a disputa entre as nações em diversos esportes, sendo um evento multiesportivo. Os jogos olímpicos contêm mais de 60 modalidades, caracterizadas como jogos olímpicos de inverno e verão. Algumas modalidades são disputadas apenas em uma das versões (verão ou inverno). Os Estados Unidos da América (EUA) é atualmente a nação com mais medalhas em toda a história dos jogos olímpicos, seguido da antiga União Soviética e Alemanha. Segundo a Empresa Brasil de Comunicação (EBC) o Brasil, até os jogos olímpicos de 2016 ocupava a 29ª posição no ranking a nível mundial e a 2º posição a nível América Latina (https://agenciabrasil.ebc.com.br/rio-2016/noticia/2016-08/brasil-sobe-de-37o-para-35o-no-quadro-de-medalhas-com-19-conquistadas-no). Além do grande espetáculo apresentado nos jogos olímpicos ao longo dos anos, o alto nível de competitividade dos atletas é apreciado. O fato das disputas esportivas serem um grande atrativo motivou o interesse dos atletas e das nações em entender como os fatores físicos, sociais, econômicos e outros influenciam na respectiva conquista da medalha. Estes fatores são importantes, pois podem influenciar nos resultados obtidos por cada atleta em curta escala (isto é, disputa de uma medalha olímpica), ou em grande escala (medalhas conquistadas por nação).

Neste trabalho utilizaremos um conjunto de dados com o objetivo de utilizar técnicas de reamostragem *bootstrap* e conhecer a relação existente entre o número de medalhas obtidas pelo Brasil em função ao número de participantes e do Produto Interno Bruto per capita (PIBp). Em estudos similares, como por exemplo, em Tang e Li (2015), fatores socioeconômicos foram considerados para explicar tal relação. Comumente, modelos de regressão são utilizados para explicar a relação existente entre uma variável de interesse (variável resposta) e um grupo de variáveis independentes conhecidas como covariáveis. Em particular, um modelo de regressão linear simples define-se como: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, para todo, $i = 1,..,n,$ onde β_0 e $\beta_1$parâmetros desconhecidos e ε_i erros aleatórios, n representa o tamanho da amostra. Os modelos de regressão linear simples tendem a atender condições relacionadas à distribuição da variável aleatória ε_i, sendo elas independente e normalmente distribuídas com média zero e variância comum σ^2. Quando o tamanho da amostra é pequeno algumas suposições relacionadas à variável ε_i podem não estar satisfeitas.

Revista brasileira de iniciação científica
ISSN: 2359-232X

A teoria dos modelos lineares generalizados (MLG) (NELDER E WEDDERBURN, 1972; MCCULLAGH E NELDER, 1989) também pode ser aplicada quando a variável resposta é caracterizada por uma distribuição que pertence à família exponencial; distribuições tais como: Normal, Poisson, Binomial, Gama e Normal Inversa, etc. Contudo, pressupostos para essa classe de modelos podem também ser violados quando ajustamos um tamanho de amostra pequeno. Uma alternativa para contornar este problema são métodos de reamostragem tipo *bootstrap*.

O método de *bootstrap* (EFRON, 1979) é utilizado como uma alternativa para realizar inferências a partir da amostra observada; assume-se que a amostra observada representa à população (ou pseudo-população), permitindo encontrar estimativas adequadas dos parâmetros, como por exemplo, o desvio padrão, viés, intervalos de confiança e realizar teste de hipóteses, entre outros. Lima, F. P. (2017), por exemplo, utilizou métodos de *bootstrap* para realizar inferências nos modelos de regressão beta**;** Dogan, C. D. (2017) aplica métodos de *bootstrap* para encontrar intervalos de confiança utilizando a linguagem de programação R (R CORE TEAM, 2021) e Cirillo, M. A. (2009) avaliou métodos de estimação intervalar para funções lineares binomiais via *bootstrap* finito.

Técnicas computacionais são necessárias para atingir os objetivos do trabalho, desta forma a linguagem de programação Python (ROSSUM E DRAKE, 2012) será utilizada.

Matérias e métodos

Nesta parte do trabalho iremos apresentar o conjunto de dados e os métodos estatísticos necessários para atingir o objetivo do trabalho.

Materiais

O conjunto de dados utilizado neste trabalho foi extraído do site *Kaggle* (https://www.kaggle.com/). Os dados encontram-se disponíveis com o nome "*120 years of Olympic history: athletes and results*". Os dados contém informações de 271.116 atletas olímpicos entre os anos 1896 e 2016, correspondentes a 230 países. Similar a Tang e Li (2015), estudaremos a relação existente entre o número de medalhas olímpicas obtidas pelo Brasil nas edições do verão entre os anos 1896 e 2016 com o número de atletas e o Produto Interno Bruto

Revista brasileira de iniciação científica
ISSN: 2359-232X

per capita (PIB_p) de cada edição; onde estas duas últimas são consideradas como variáveis socioeconômicas. Os dados para a análise são apresentados na Tabela 1.

Edição (Ano)	Medalhas	Nº de participantes	PIB per capita ($)
1960	2	72	210,11
1964	1	61	261,67
1968	3	76	374,79
1972	2	81	586,21
1976	2	79	1.390,62
1980	4	106	1.947,28
1984	8	147	1.578,93
1988	6	160	2.300,38
1992	3	182	2.596,92
1996	15	221	5.166,16
2000	12	198	3.749,75
2004	10	243	3.637,46
2008	16	268	8.831,02
2012	17	248	12.370,02
2016	19	462	8.710,10

Tabela 1: *Ano, número de medalhas, número de participantes, PIB per capita durante do Brasil nos jogos olímpicos de verão.*

Revista brasileira de iniciação científica
ISSN: 2359-232X

Métodos

Modelos lineares generalizados

Durante muitos anos os modelos normais lineares foram utilizados na tentativa de descrever a maioria dos fenômenos aleatórios. Mesmo quando o fenômeno sob estudo não apresentava uma resposta para a qual fosse razoável a suposição de normalidade (PAULA, 2013). Nelder e Wedderburn (1972) abrem um leque de opções para a distribuição da variável resposta e que a mesma pertence à família exponencial. Assim, podemos supor que dados de contagem seguem a distribuição Poisson, dados binários (sucesso e fracasso) podem ser ajustados sob a suposição da distribuição binomial e dados assimétricos com suporte no conjunto dos números reais positivos, uma distribuição gama pode ser adequada. O conjunto de distribuições que Nelder e Wedderburn (1972) definem uma classe de modelos conhecida como modelos lineares generalizados (MLGs).

Nelder e Wedderburn (1972) propuseram também um processo iterativo para a estimação dos parâmetros e introduziram o conceito de função deviance que tem sido largamente utilizado na avaliação da qualidade do ajuste dos modelos MLG's, bem como no desenvolvimento de resíduos e medidas de diagnóstico.

A variável resposta (número de medalhas) é uma variável discreta (contagem). Assim, seguindo a classe dos modelos lineares generalizados (MLGs) (NELDER E WEDDERBURN ,1972; MCCULLAGH E NELDER, 1989) as distribuições Poisson e binomial negativa são adequados para realizar a análise. É comum, na prática, considerar a distribuição normal para analisar dados de contagem, no entanto, esta prática é inadequada devido ao fato da distribuição normal ser recomendada para modelar variáveis contínuas contidas em todo o espaço do conjunto dos números reais. Mesmo assim, neste trabalho iremos considerar a distribuição normal (erroneamente) para mostrar como a má especificação paramétrica pode levar a interpretações inadequadas ou enganosas.

No contexto dos dados olímpicos definimos a seguir 3 modelos a serem utilizados posteriormente na análise.

Revista brasileira de iniciação científica
ISSN: 2359-232X

Modelo 1	Modelo 2	Modelo 3
$(i)\ y_i \sim Poisson(\mu_i)$ $(ii)\ \mu_i = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$	$(i)\ y_i \sim BN(\mu_i, \phi)$ $(ii)\ \mu_i = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$	$(i)\ y_i \sim Normal(\mu_i, \sigma^2)$ $(ii)\ \mu_i = exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})$

Tabela 2: *Modelos quando assumido que a variável resposta segue uma distribuição Poisson, binomial negativo e Normal.*

Método de *bootstrap*

O método de *bootstrap* tem sido utilizado amplamente desde 1979 como a proposta de Efron. *Bootstrap* é um método de reamostragem computacional poderoso para realizar inferência estatística. A ideia básica do método é realizar procedimentos inferências sobre o parâmetro populacional (tal como a média, desvio padrão, etc.), digamos θ, sobre uma amostra aleatória, por meio de reamostragem. Geralmente, os métodos *bootstrap* envolvem os passos a seguir: *i)* Uma amostra aleatória de tamanho *n* extraída de uma população é considerada como tal; *ii)* Extraímos *B* pseudo-amostra de tamanho *n* da amostra original com reposição; *iii)* Para cada pseudo-amostra encontramos as estatísticas θ, totalizando *B* estimativas de θ; *iv)* Definimos a distribuição amostral com as *B* estatísticas *bootstrap* e utilizamos para realizar inferência como: Estimar o erro padrão, viés, assim como encontrar intervalos de confiança para θ.

De maneira técnica, seguindo Efron (1979), considera-se uma amostra $x = (x_1, \ldots, x_n)$ de uma variável aleatória *X*, cuja distribuição está completamente determinada por sua função de distribuição acumulada *F*. Sejam $\theta = t(F)$, uma função de *F* denominada parâmetro e $\hat{\theta} = s(x)$ um estimador de θ. O método consiste em obter, de uma amostra original, um grande número de subamostras (denominada também pseudo-amostras) $x^* = (x_1^*, \ldots, x_n^*)$, em que, para cada pseudo-amostra, encontrar-se suas respectivas estimativas $\hat{\theta}^* = s(x^*)$, com base na distribuição empírica de $\hat{\theta}^*$. Na literatura há duas abordagens para utilizar o método *bootstrap*, uma paramétrica e outra não paramétrica. Na abordagem não paramétrica, as pseudo-amostras são obtidas a partir da estimativa não paramétrica $\hat{F}$ de F, que é uma função de distribuição empírica amostral, definida por $\hat{F}(x) = \#\{x_i \leq x\}/n$, atribuindo probabilidades de $1/n$ a cada

Revista brasileira de iniciação científica
ISSN: 2359-232X

$x_i, i = 1, \ldots, n$, onde $\#\{.\}$ representa a frequência absoluta. Portanto, pode-se afirmar que o estimador de $\theta = t(F)$ é $\hat{\theta} = t(\hat{F})$. Na abordagem paramétrica, F pertence a uma família de modelos.

O método *bootstrap* é útil, na prática, para encontrar, quando não for possível analiticamente, estimativas do viés, erro padrão, erro quadrático médio, etc. de um determinado estimador via reamostragem (EFRON, 1979). Denotemos, por exemplo, o viés de $\hat{\theta} = t(\hat{F})$ por $V_F(\hat{\theta}, \theta) = E_F[s(x)] - t(F)$, então os estimadores *bootstrap* do viés, nas abordagem não paramétrica e paramétrica são dados por $V_{\hat{F}}(\hat{\theta}, \theta) = E_{\hat{F}}[s(x)] - t(\hat{F})$ e $V_{F_{\hat{\xi}}}(\hat{\theta}, \theta) = E_{F_{\hat{\xi}}}[s(x)] - t(F_{\hat{\xi}})$, respectivamente.

Gerar-se B pseudo-amostras *bootstrap*, $(x^{*1}, \ldots, x^{*B})$ de x, calcula-se para cada pseudo-amostra, $(\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B})$, onde $s(x^{*i})$ para $i = 1, \ldots, B$ e pode-se aproximar $E_{\hat{F}}[s(x)]$ e $E_{F_{\hat{\xi}}}[s(x)]$ pela média.

$$\hat{\theta}^{*(.)} = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}^{*i},$$

obtendo, a estimativa *bootstrap* do viés de $V_{\hat{F}}(\hat{\theta}, \theta)$ e $V_{F_{\hat{\xi}}}(\hat{\theta}, \theta)$, que são dadas por $V_{\hat{F}}(\hat{\theta}, \theta) = \hat{\theta}^{*(.)} - s(x)$ e $V_{F_{\hat{\xi}}}(\hat{\theta}, \theta) = \hat{\theta}^{*(.)} - s(x)$, respectivamente. A partir da estimativa *bootstrap* do viés, é possível definir um estimador corrigido até segunda ordem como

$$\underline{\theta}_1 = s(x) - V_{\hat{F}}(\hat{\theta}, \theta) = 2s(x) - \theta^{*(.)}, \underline{\theta}_2 = s(x) - V_{F_{\hat{\xi}}}(\hat{\theta}, \theta) = 2s(x) - \theta^{*(.)},$$

onde $\underline{\theta}_1$ e $\underline{\theta}_2$ são denominadas estimativas CBC (Constant-Bias-Correcting), ver por exemplo, Efron (1979) e Giles e Mentch (2015).

Podemos encontrar também, intervalos de confiança *bootstrap* BCa. O intervalo *bootstrap* BCa (*Bias Corrected and Accelerated*) utiliza os percentis da distribuição *bootstrap* para sua construção. Os percentis utilizados dependem de duas quantidades $\hat{a}$ e $\hat{z}_0$, denominados como correção da tendência e de aceleração, respectivamente. De acordo com Efron e Tibshirani (1994), um intervalo de confiança BCa $(1 - \alpha)$ e definida por $(\hat{\theta}^{\alpha 1}, \hat{\theta}^{\alpha 2})$ onde

$$\alpha_1 = \Phi\left(\hat{z}_0 \frac{\hat{z}_0 + z_{\frac{\alpha}{2}}}{1 - \hat{a}\left(\hat{z}_0 + z_{\frac{\alpha}{2}}\right)}\right), \alpha_2 = \Phi\left(\hat{z}_0 \frac{\hat{z}_0 + z_{\frac{1-\alpha}{2}}}{1 - \hat{a}\left(\hat{z}_0 + z_{1-\frac{\alpha}{2}}\right)}\right),$$

Revista brasileira de iniciação científica
ISSN: 2359-232X

com $\Phi(\cdot)$ função distribuição acumulada da distribuição normal padrão e $z_{(\alpha)}$ com 100α-ésimo percentil de uma distribuição normal padrão. A constante de correção da tendência, $\hat{z}_0$, é obtida da proporção de réplicas *bootstrap* cujas estimativas θ^{*b} são menores que a estimativa original $\hat{\theta}$, que é dada por $\hat{z}^0 = \Phi^{-1}(\#\{\hat{\theta}^{*b} < \hat{\theta}\}/B)$, sendo $\hat{z}_0$ a magnitude da tendenciosidade mediana de $\hat{\theta}^*$, ou seja, a distância entre a mediana de $\hat{\theta}^*$ e $\hat{\theta}$, em uma escala normalizada. Tem-se $\hat{z}_0$ igual a zero se exatamente a metade das medidas de θ^{*b} forem menores ou iguais a $\hat{\theta}$. Dentre as várias possibilidades de obter $\hat{a}$, e dada em termos dos valores *Jackknife* de $\hat{\theta} = s(x)$

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}\cdots - \hat{\theta}_i)^3}{6\left(\sum_{i=1}^{n}(\hat{\theta}\cdots - \hat{\theta}_i)^2\right)^{3/2}},$$

onde $\hat{\theta}\cdots = \sum_{i=1}^{n}\hat{\theta}_i/n \quad e\ \hat{\theta}_{(i)} = s(x_i), i = 1,\ldots,n.$

Ao longo dos anos, extensões do método *bootstrap* foram propostas na literatura. Desta forma, neste trabalho iremos descrever de maneira detalhada, como descrita em Lima (2017), as extensões a seguir:

1. O *Bootstrap* Bayesiano (RUBIN, 1981) surgiu como método análogo ao *bootstrap* proposto por Efron em 1979. Ambas as técnicas de *bootstrap* apresentam resultados semelhantes na parte inferencial. Os métodos divergem na forma de realizar a reamostragem. Dada uma amostra aleatória $(x_1, \ldots, x_n)$, os elementos da amostra original tem probabilidade $1/n$ de ser selecionado no método *bootstrap* clássico, enquanto no método *bootstrap* bayesiano cada replicação gera uma probabilidade posterior para todo $x_i,\ i = 1,\ldots,n$. No quadro abaixo, lista-se os passos que descreve o método *bootstrap* Bayesiano:

Para $i = 1, \ldots, B$

Fazer

1. Considere a amostra aleatória $(x_1,\ldots,x_n)$;
2. Gerar uma amostra aleatória $(u_1,\ldots,u_{n-1})$ de uma distribuição uniforme no intervalo aberto (0, 1);
3. Ordenar a amostra $u_{(0)} = 0 \leq u_{(1)} \leq \ldots \leq u_{(n-1)} \leq u_{(n)} = 1$;

Revista brasileira de iniciação científica
ISSN: 2359-232X

4. Encontramos o vetor $(d_1, \ldots, d_n)$ onde $d_i = u_i - u_{i-1}$;
5. Uma pseudo amostra com reposição de $(x_1, \ldots, x_n)$ é encontrada. Associar a cada y_i a probabilidade g_i de ser selecionado.
6. Calculamos a média da estatística, $s(x)$, das pseudo-amostra geradas.

Fim

2. O *Bootstrap* suavizado (EFRON, 1979) é uma versão do *bootstrap* não paramétrico, essa versão pressupõe que a distribuição dos dados é contínua. Gera-se uma pseudo-amostra $y* = (y_1^*, \ldots, y_n^*)$, a partir da amostra original. A amostra suavizada é gerada, $y = y_i^* + \varepsilon_i \ , i = 1, \ldots, n$, onde ε_i representa um ruído que segue uma distribuição normal $(0, h^2)$, em que $h = s/\sqrt{n}$ sendo $s^2 = \sum_{i=1}^{n}(y_i - \hat{y})^2/(n-1)$. No quadro abaixo, lista-se os passos que descreve o método *bootstrap* suavizado:

Para $i = 1, \ldots, B$

Fazer

1. Considere a amostra aleatória $(x_1, , \ldots, x_n)$;
2. Gerar uma pseudo amostra $(x_1^*, \ldots, x_n^*)$ de $(x_1, \ldots, x_n)$;
3. Uma pseudo amostra suavizada é obtida cujos elementos são $y_i = y_i^* + \varepsilon_i \, com \, i = 1, \ldots, n$.
4. Calculamos a média da estatística, $s(x)$, das pseudo-amostra geradas.

Fim

3. O *Bootstrap* paramétrico (EFRON, 1979) pressupõe que a amostra aleatória em questão, segue uma distribuição F_θ em que θ e o parâmetro de interesse. As pseudo-amostras $x^* = (x_1^*, \ldots, x_n^*)$ são geradas a partir da estimativa de θ. No quadro abaixo, lista-se os passos que descreve o método *bootstrap* paramétrico:

Para $i = 1, \ldots, B$

Fazer

1. Considere a amostra aleatória $(x_1, \ldots, x_n)$;

Revista brasileira de iniciação científica
ISSN: 2359-232X

2. Gere uma pseudo amostra $(x_1^*, \ldots x_n^*)$ em que $x_i^* \sim F_{\hat{\theta}}$;
3. Para cada pseudo-amostra encontre a estatística de interesse, $s(x)$;
4. Calculamos a média das estatísticas geradas.

Fim

4. O *Bootstrap* duplo (EFRON, 1983) é um método que se mostrou bem eficiente quando o interesse é encontrar intervalos de confiança. Esta forma de reamostragem consiste em utilizar um segundo nível de reamostragem, ou seja, reamostrar a partir da primeira amostragem. E gerado B_1 pseudo-amostras no primeiro nível, para cada pseudo-amostra é gerada uma segunda reamostragem de tamanho B_2. Este método acaba tendo um custo computacional alto, devido às que são geradas $B_1 \times B_2$ pseudo-amostras. No quadro abaixo, lista-se os passos que descreve o método *bootstrap* duplo:

Para $i = 1, \ldots, B$

Fazer

1. Considere a amostra aleatória $(x_1, \ldots, x_n)$;
2. Fixar B_1 como o número de réplicas *bootstrap* do primeiro nível;
3. Gere uma pseudo-amostra, $(x_1^*, \ldots, x_n^*)$ a partir de $(x_1, \ldots, x_n)$ com reposição;

Para $j = 1, \ldots$

Fazer

1. Fixar B_2 como o número de réplicas *bootstrap* do segundo nível;
2. Gere uma pseudo-amostra $(x_1^{**}, \ldots, x_n^{**})$ de $(x_1^*, \ldots, x_n^*)$ com reposição;
3. Calcular a estatística de interesse para cada pseudo-amostra de segundo nível;
4. Calcular a média das estimativas de segundo nível para cada interação;

Fim

4. Calcular então a média das médias das estimativas;

Fim

5. O *Bootstrap* duplo rápido (DAVIDSON E MACKINNON, 2007) ao contrário do *bootstrap* duplo, em que em seu processo de reamostragem gera $B_1 \times B_2$ pseudo-

Revista brasileira de iniciação científica
ISSN: 2359-232X

amostras, gera então um total de 2B pseudo-amostras. Nesta variação do método *bootstrap* duplo, o processo tem um ganho computacional processando um número menor de pseudo-amostras. No quadro abaixo, lista-se os passos que descreve o método *bootstrap* duplo rápido:

Para $i = 1, \ldots, B$

Fazer

1. Considere a amostra aleatória $(x_1, \ldots, x_n)$;
2. Fixar B como o número de réplicas *bootstrap*.
3. Gere uma pseudo amostra $(x_1^*, \ldots, x_n^*)$ de $(x_1, \ldots, x_n)$ com reposição;
4. Em seguida gere uma pseudo-amostra $(x_1^{**}, \ldots, x_n^{**})$ de $(x_1^*, \ldots, x_n^*)$ com reposição;
5. Calcular a estatística de interesse para cada pseudo-amostra do segundo nível;
6. Calcular a média das estatísticas geradas.

Fim

6. O *Bootknife* (HESTERBERG E HESTERBERG, 1999) é uma junção dos métodos *Bootstrap* e *Jackknife*. O conceito se baseia em omitir uma informação da amostra aleatória *bootstrap*. No quadro abaixo, lista-se os passos que descreve o método *Bootknife*:

Para $i = 1, \ldots, B$

Fazer

1. Considere a amostra aleatória $(x_1, \ldots, x_n)$;
2. Fixar B como o número de réplicas *bootstrap*.
3. Seja $k = B/n$. Gere k pseudo-amostras com reposição a partir da amostra $(x_1, \ldots, x_n)$ onde foi removida a i-ésima observação (amostra *Jackknife*).
4. Para cada amostra *Jackknife* encontre a estatística de interesse, $s(x)$;
5. Calculamos a média das estatísticas geradas.

Fim

Revista brasileira de iniciação científica
ISSN: 2359-232X

Finalmente, os códigos implementados na linguagem de programação Python, disponíveis no repositório do primeiro autor (https://github.com/santos-luciano/bootstrap) , descrevem os diferentes métodos abordados nesta seção.

Coeficiente de correlação de Spearman

O coeficiente de correlação de Spearman é uma medida não paramétrica que descreve a relação entre duas variáveis. Similar ao coeficiente de correlação de Pearson, este varia ao longo do intervalo [-1,1]. Quanto mais próximo dos extremos (-1 ou 1), maior é a força da correlação. Já valores próximos de 0 implica em correlação fraca ou inexistente. O método para encontrar o coeficiente utiliza simplesmente os postos, sem fazer nenhuma suposição. Essencialmente, o que é encontrado é o coeficiente de correlação de Pearson nos postos. A representação matemática é dada da forma

$$r = 1 - \frac{6\sum_i d_i^2}{(n^3 - n)},$$

em que n é o número de pares (x_i, y_i), $i = 1, \dots, n$ e $d_i = (posto\ de\ x_i\ dentre\ os\ valores\ de\ x) - (posto\ de\ y_i\ nos\ valores\ de\ y)$. Note que se os postos de x são exatamente iguais aos postos de y, então todos os d_i serão zero e r será 1.

Índice de desempenho

Definimos o índice de desempenho de cada nação da forma

$$Desempenho_j = número\ de\ medalhas_j / número\ de\ participantes_j,$$

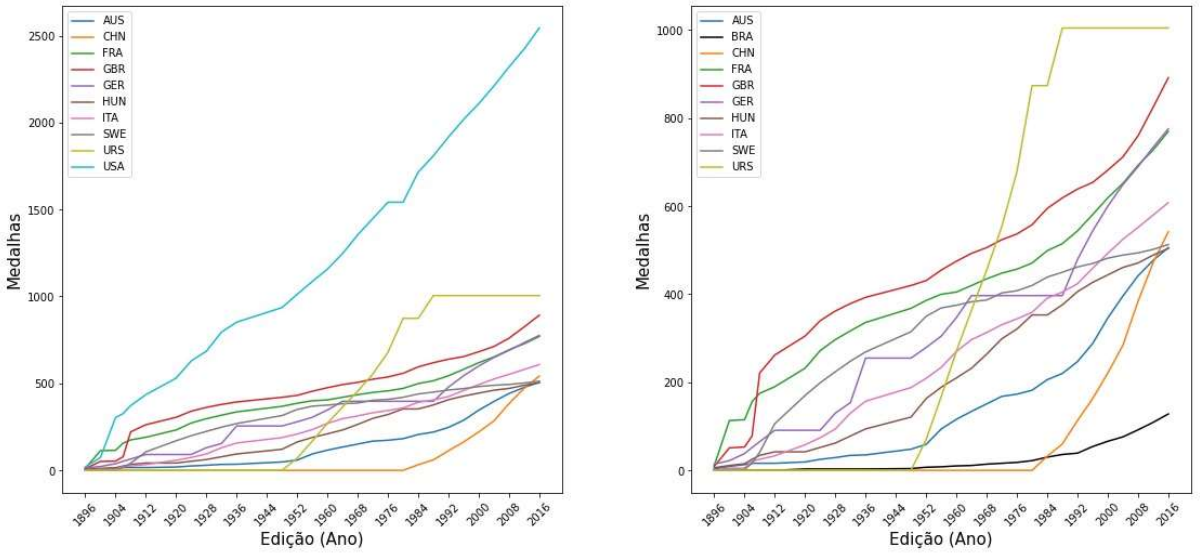
Em que j representa a uma j −ésima nação.

Resultados e Discussão

Alguns resultados obtidos nos permitirá entender como o Brasil tem-se desempenhado ao longo das edições olímpicas. Por exemplo, na Figura 1(a) observa-se o número de medalhas olímpicas ao longo de 1986 a 2016 das 10 nações com maior número de medalhas. Pode-se observar claramente que os Estados Unidos de Norte-americana apresenta uma considerável superioridade em relação aos outros países com mais de 2500 medalhas no total. A URS que disputou os jogos olímpicos entre os anos 1952 e 1988, mesmo participando em nove edições

Revista brasileira de iniciação científica
ISSN: 2359-232X

das 29, ocupa a segunda posição no total de medalhas conquistadas. O Brasil ao longo de todas as edições, como observado na Figura 1(b), encontra-se muito inferior quando comparado com os 10 melhores países, no entanto observamos um leve crescimento a partir de 1992. Para melhorar a visualização da Figura 1(b) foi excluída a informação referente aos Estados Unidos de Norte-americana.



(a) *(b)*

Figura 1: *(a) Número de medalhas acumuladas das 10 nações com melhores resultados e (b) Número de medalhas acumuladas incluindo o Brasil, retirando Estados Unidos de Norte América.*

A Tabela 3 mostra algumas medidas descritivas do número total de atletas das 10 melhores nações, assim como para o Brasil. Observa-se na tabela que países com mais medalhas conquistadas são países com maior número de atletas na competição ao longo dos anos. Podemos observar também que oito dos 10 países que têm mais medalhas olímpicas estão entre os que têm mais participações na competição. Ainda, podemos observar que aproximadamente o número de atletas que participaram do Brasil, representa em média quase a terceira parte dos três primeiros colocados. Observa-se também que em pelo menos 50% das edições os USA apresentou o maior número de participantes.

Revista brasileira de iniciação científica
ISSN: 2359-232X

País	Média	Desvio Padrão	Mínimo	Primeiro quartil	Mediana	Terceiro quartil	Máximo
USA	370,07	176,27	14,00	287,50	358,00	527,50	648,00
URS	359,78	82,32	281,00	295,00	317,00	410,00	489,00
FRG	345,00	63,45	275,00	290,00	347,00	389,00	424,00
GDR	278,80	45,48	225,00	259,00	267,00	297,00	346,00
GER	265,90	162,26	19,00	129,25	293,50	420,50	464,00
GBR	263,24	145,95	6,00	208,00	257,00	310,00	735,00
RUS	261,00	198,66	3,00	44,25	337,00	433,50	454,00
FRA	242,17	138,96	1,00	145,00	238,00	309,00	716,00
CHN	237,00	175,66	1,00	54,00	266,00	365,00	583,00
UKR	229,50	13,61	204,00	230,00	230,50	237,00	243,00
BRA	129,65	107,05	1,00	65,50	81,00	190,00	462,00

Tabela 3: *Medidas descritivas do número de atletas no período de 1896 a 2016 das 10 melhores nações e Brasil*

Com o intuito de mensurar a relação existente entre o número de medalhas conquistadas e número de participantes em cada edição, encontra-se o coeficiente de correlação de Spearman (CHEN E POPOVICH, 2002) igual a 0,8926, indicando uma alta relação entre as duas variáveis. A Tabela 4 mostra o índice de desempenho de cada nação, quanto maior a proporção, menor número de atletas é necessário para a conquista de medalhas. Por exemplo, a URS apresenta o maior aproveitamento com respeito à GBR e USA. Com respeito ao Brasil, foi encontrado um desempenho equivalente a 4,29%, desempenho relativamente pequeno quando comparado com as 10 nações com maior número de medalhas.

Nação	Desempenho (%)	Nação	Desempenho (%)
URS	31,04	GDR	29,34

USA	24,55	RUS	17,62
CHN	17,59	GER	14,59
FIN	13,70	ROU	13,65
HUN	13,20	SWE	12,73

Tabela 4: *Desempenho das 10 melhores nações na conquista medalhas com respeito ao total de atletas participantes,*

Ao longo da história brasileira foi escutado no mundo que o Brasil é o país do futebol, entretanto no contexto dos jogos olímpicos o Brasil tem se destacado mais em outras modalidades, como o judô, navegação e natação como pode ser observado na Tabela 5.

No quadro de medalhas mostrado na Tabela 5, o esporte de maior desempenho do Brasil é o judô, com um total de 22 medalhas conquistadas ao longo da história nos jogos olímpicos, sendo a maioria delas de bronze. O esporte no qual o Brasil tem mais medalhas de ouro é a Vela, sendo um total de 07 medalhas.

Bronze	**Prata**	**Ouro**	**Total**	**Esporte**
8	3	5	16	Atletismos
4	1	0	5	Basquetebol
3	7	3	13	Vôlei praia
3	1	1	5	Box
1	2	0	3	Canoagem
2	0	1	3	Hipismo
2	5	1	8	Futebol
1	2	1	4	Ginástica
15	3	4	22	Judô
1	0	0	1	Pentatlo moderno
8	3	7	18	Navegação
1	2	1	4	Tiroteio

Revista brasileira de iniciação científica
ISSN: 2359-232X

9	4	1	14	Natação
2	0	0	2	Tae-kwon-do
2	3	5	10	Voleibol
62	36	30	128	Total

Tabela 5: *Quadro de medalhas olímpicas do Brasil ao longo de todas as edições olímpicas que o Brasil participou,*

Identificamos nos dados correspondentes ao Brasil um crescimento considerável do número de medalhas a partir dos jogos olímpicos de 1980, época em que começou a surgir os maiores vencedores da história olímpica brasileira, incluindo nomes como Robert Scheidt e Torben Grael, ambos disputaram os jogos olímpicos no esporte do iatismo, Acreditamos que este crescimento pode estar relacionado a alguns fatores, como investimento no esporte, preparação dos atletas, fatores econômicos, etc. Seguindo Tang e Li (2015), consideramos a variável Produto Interno Bruto per capita (PBI_p) junto ao número de medalhas conquistadas ao longo das edições. Podemos observar na Figura 2 um crescimento simultâneo de ambas variáveis ao longo do tempo.

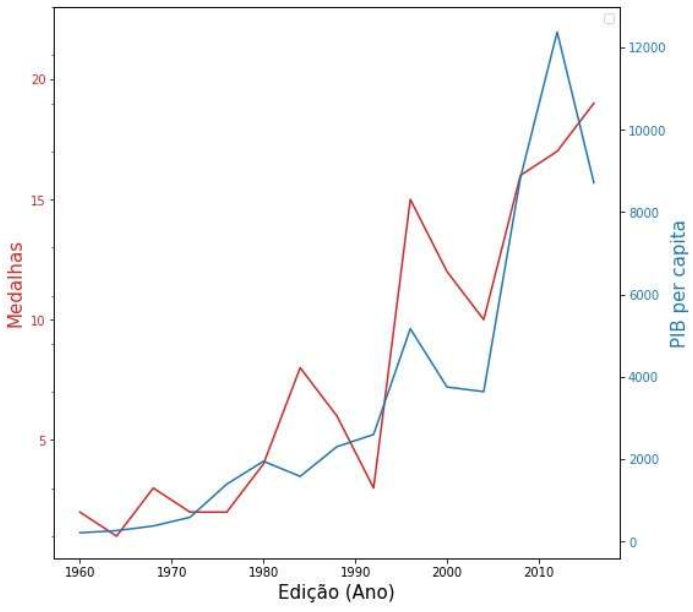


Figura 2: *Medalhas do Brasil desde 1900,*

Revista brasileira de iniciação científica
ISSN: 2359-232X

Para explicar melhor a relação existente entre o número de medalhas conquistadas (variável resposta) com o número de participantes e PIB per capita do Brasil, os modelos lineares generalizados são considerados. Em particular, as distribuições Poisson e Binomial Negativa são suposições adequadas para a variável resposta.

Consideremos $(y_1, x), (y_2, x), , , , , (y_n, x)$ pares de observações de tamanho $n = 15$, sendo y_i, $i = 1, 2, , , , , n$ uma variável resposta (número de medalhas) e x matriz de planejamento contendo as variáveis independentes (x_1 = número de atletas participantes e x_2 = PIB per capita do Brasil, em dólares), Mostram-se na Tabela 6 medidas descritivas para as variáveis em estudo.

	Média	**Desvio Padrão**	**Mínimo**	**1° quartil**	**Mediana**	**3° quartil**	**Máximo**	**CV(%)**
y	8,00	6,34	1,00	2,50	6,00	13,50	19,00	79
x_1	173,60	107,19	61,00	80,00	160,00	232,00	462,00	62
x_2	3580,76	3677,88	210,12	988,42	230,38	4457,96	12370,02	103

Tabela 6: *Medidas descritivas da variável y (número de medalhas), x_1 (número de participantes) e x_2 (PIB per capita do Brasil.*

Pode-se observar na Tabela 6, que em média o número de medalhas obtidas por edição no Brasil é oito; em média o número de participantes por edição é 173,6 e o PIB per capita representa 3.580,76 dólares por habitante. O coeficiente de variação (CV) indica que a variabilidade em torno da sua média para de x_2 é quase duas vezes a variabilidade existente em x_1, Na Figura 3 observa-se a relação linear existente entre as variáveis em estudo.

Revista brasileira de iniciação científica
ISSN: 2359-232X

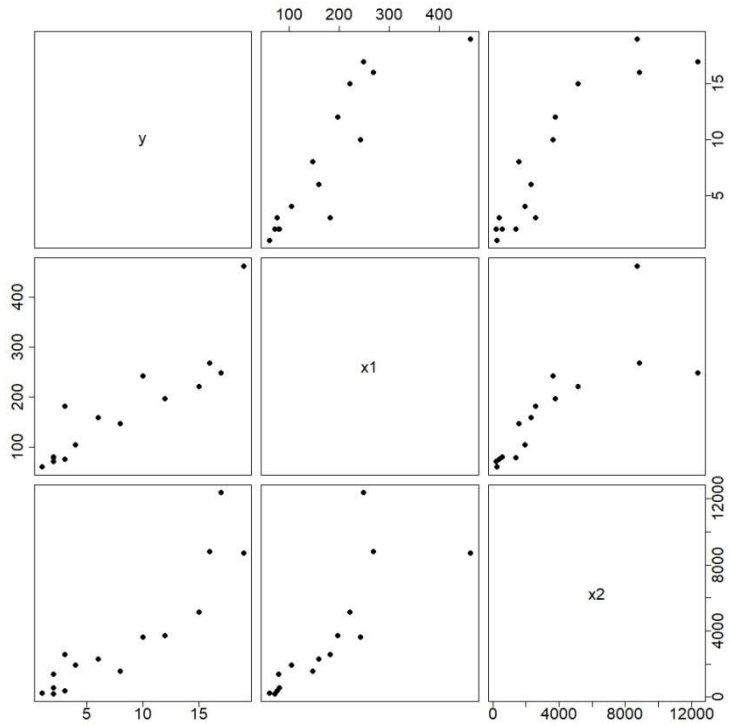
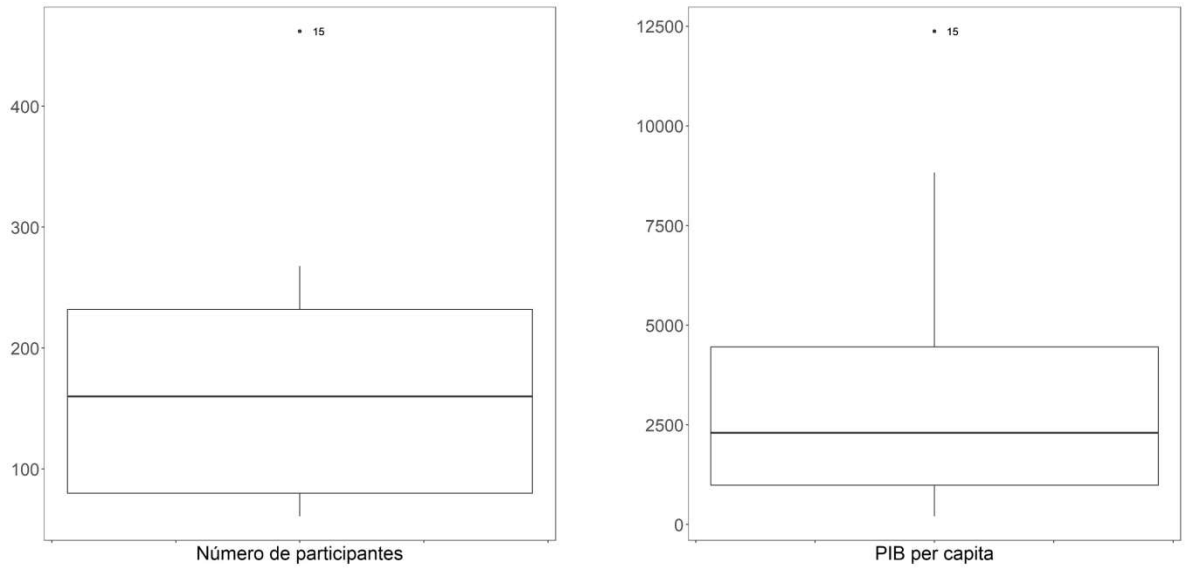


Figura 3: *Gráficos de dispersão entre as variáveis: número de medalhas, número de participantes e PIB per capita.*

Gráficos de caixas (*boxplot*) são construídos para as variáveis x_1 e x_2 como mostra Figura 4. Claramente, observa-se a presença de uma observação atípica ou *outlier*. A edição considerada atípica corresponde ao ano 2016, em que o Brasil teve seu melhor desempenho com o maior número de atletas (462).



Revista brasileira de iniciação científica
ISSN: 2359-232X

Figura 4: *Gráficos de boxplot para as variáveis: número de participante (esquerda) e PIB per capita (direita).*

Métodos de *bootstrap* são utilizados para encontrar estimativas corrigidas dos parâmetros, viés, erros padrão e limite inferior e superior do intervalo de confiança com 95% de confiança (intervalos de credibilidade quando o método bayesiano é utilizado) para os modelos definidos na seção métodos. As Tabelas 7-9 mostram os resultados para os 3 modelos considerados, a saber: quando a variável resposta assume seguir uma distribuição Poisson, binomial negativa, normal, respectivamente. Observar-se que a variável número de participantes e PIB per capita são significativas ao explicar a variável resposta para os três modelos. Isto é, os limites de cobertura aos 95% de confiança não contém zero. O intercepto apresenta um viés considerável, no entanto β_1 e β_2 apresentam viés próximo de zero. Diferente das Tabelas 7 e 8, na Tabela 9 a estimativa associada ao intercepto apresenta sinal contrário para todos os métodos *bootstrap*. Considerou-se para todos os métodos de *bootstrap* 1000 réplicas. Nesta análise assume-se que a variável resposta y_1, y_2,...,y_{15}, número de medalhas obtidas nas 15 edições que o Brasil participou dos jogos olímpicos de verão, são variáveis aleatórias independentes, considera-se que os intervalos de tempo entre as edições olímpicas podem ser consideradas relativamente grandes. Métodos que consideram auto correlação temporal podem também ser utilizados como alternativa de análise dos dados, no entanto, essa análise não é o foco deste trabalho e deixe-se em aberto como pesquisa futura a ser realizada.

Método	Parâmetro	Estimativa	Viés	Erro padrão	LI(95%)	LS(95%)
Paramétrico	β_0	0.76653	-0.27595	0.50487	0.73092	0.79775
	β_1	0.00433	0.00162	0.00472	0.00407	0.00469
	β_2	0.00010	0.00000	0.00014	0.00010	0.00011
Bayesiano	β_0	0.63127	-0.41121	0.63321	0.59220	0.67447
	β_1	0.00502	0.00231	0.00707	0.00456	0.00547
	β_2	0.00011	0.00001	0.00025	0.00010	0.00013
Suavizado	β_0	0.74666	-0.29582	0.51324	0.71316	0.77649
	β_1	0.00446	0.00175	0.00493	0.00417	0.00479
	β_2	0.00011	0.00001	0.00014	0.00010	0.00012

Revista brasileira de iniciação científica
ISSN: 2359-232X

Duplo	β_0	0.57326	-0.46922	0.02024	0.57201	0.57450
	β_1	0.00562	0.00291	0.00026	0.00560	0.00564
	β_2	0.00010	-0.00000	0.00001	0.00010	0.00010
Duplo rápido	β_0	0.56478	-0.47770	0.69656	0.52312	0.60850
	β_1	0.00598	0.00327	0.00817	0.00553	0.00651
	β_2	0.00008	-0.00002	0.00032	0.00006	0.00010
Bootknife	β_0	0.70762	-0.33486	0.53589	0.67620	0.74236
	β_1	0.00466	0.00194	0.00528	0.00430	0.00498
	β_2	0.00010	0.00000	0.00017	0.00009	0.00012

Tabela 7: *Estimativas dos parâmetros, viés, erro padrão, limite inferior e superior do intervalo para o modelo 1.*

Método	**Parâmetro**	**Estimativa**	**Viés**	**Erro padrão**	**LI(95%)**	**LS(95%)**
Paramétrico	β_0	0.93668	-0.06181	0.23954	0.92074	0.95083
	β_1	0.00262	-0.00030	0.00206	0.00250	0.00276
	β_1	0.00013	0.00003	0.00008	0.00013	0.00014
	ϕ	27.03552	-21.05058	13.44951	26.20757	27.84577
Bayesiano	β_0	0.93421	-0.06428	0.24501	0.91768	0.94830
	β_1	0.00267	-0.00026	0.00205	0.00255	0.00280
	β_1	0.00013	0.00003	0.00007	0.00013	0.00014
	ϕ	27.68255	-20.40355	13.76877	26.86686	28.57005
Suavizado	β_0	0.94176	-0.05673	0.21624	0.92931	0.95577
	β_1	0.00255	-0.00037	0.00194	0.00242	0.00266
	β_1	0.00013	0.00003	0.00008	0.00013	0.00014
	ϕ	26.90309	-21.18301	12.94828	26.07562	27.62172
Duplo	β_0	0.81643	-0.18206	0.01720	0.81535	0.81744
	β_1	0.00314	0.00022	0.00014	0.00314	0.00315
	β_1	0.00014	0.00004	0.00000	0.00014	0.00014
	ϕ	24.56498	-23.52112	0.12833	24.55765	24.57316
Duplo rápido	β_0	0.81938	-0.17911	0.40563	0.79317	0.84276

Revista brasileira de iniciação científica
ISSN: 2359-232X

	β_1	0.00296	0.00003	0.00383	0.00271	0.00318
	β_1	0.00015	0.00005	0.00014	0.00014	0.00016
	ϕ	25.08091	-23.00519	13.02549	24.27553	25.84002
Bootknife	β_0	0.89532	-0.10317	0.30641	0.87587	0.91380
	β_1	0.00291	-0.00002	0.00259	0.00275	0.00306
	β_1	0.00013	0.00003	0.00009	0.00013	0.00014
	ϕ	27.75391	-20.33219	13.54910	26.95791	28.58802

Tabela 8: *Estimativas dos parâmetros, viés, erro padrão, limite inferior e superior do intervalo para o modelo 2.*

Método	Parâmetro	Estimativa	Viés	Erro padrão	LI(95%)	LS(95%)
Paramétrico	β_0	-0.25566	-0.22995	1.30323	-0.34047	-0.17705
	β_1	0.02692	-0.00036	0.01880	0.02575	0.02803
	β_2	0.00106	0.00014	0.00070	0.00102	0.00111
	σ^2	4.13431	-1.01865	1.82803	4.02361	4.23854
Bayesiano	β_0	-0.38141	-0.35570	2.12871	-0.51325	-0.24427
	β_1	0.02883	0.00156	0.03238	0.02686	0.03079
	β_2	0.00103	0.00011	0.00131	0.00095	0.00111
	σ^2	3.46337	-1.68959	2.30286	3.30691	3.60080
Suavizado	β_0	-0.19844	-0.17273	1.25517	-0.27970	-0.12844
	β_1	0.02633	-0.00094	0.01775	0.02529	0.02744
	β_2	0.00108	0.00016	0.00064	0.00104	0.00113
	σ^2	4.31643	-0.83652	1.85913	4.20957	4.43614
Duplo	β_0	-0.42121	-0.39549	0.08179	-0.42690	-0.41643
	β_1	0.02718	-0.00009	0.00100	0.02712	0.02725
	β_2	0.00113	0.00021	0.00003	0.00113	0.00113
	σ^2	3.33667	-1.81628	0.14659	3.32821	3.34780
Duplo rápido	β_0	-0.45652	-0.43081	2.23034	-0.58916	-0.30948
	β_1	0.02850	0.00123	0.03354	0.02640	0.03056
	β_2	0.00108	0.00016	0.00128	0.00101	0.00116
	σ^2	3.31448	-1.83848	2.27910	3.17175	3.46014

Revista brasileira de iniciação científica
ISSN: 2359-232X

Bootknife	β_0	-0.32941	-0.30370	1.38009	-0.41810	-0.23932
	β_1	0.02742	0.00015	0.02029	0.02612	0.02871
	β_2	0.00107	0.00015	0.00073	0.00103	0.00112
	σ^2	4.07509	-1.07786	1.95430	3.95389	4.20148

Tabela 9: *Estimativas dos parâmetros, viés, erro padrão, limite inferior e superior do intervalo para o modelo 3.*

Continuando com a análise, sabe-se que métodos *bootstrap* permitem encontrar estimativas corrigidas de forma fácil. Sejam $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ e $\hat{\beta}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^*)$ estimativas obtidas por algum método usual (por exemplo, máxima verossimilhança) e estimativas *bootstrap*, respectivamente. Uma estimativa corrigida $\hat{\beta} = 2\beta - \beta^*$. Desta forma, as Tabelas 10 e 11 mostram as estimativas corrigidas de máxima verossimilhança para os modelos 1 e 2, respectivamente.

Método	**Parâmetro**	**Estimativa Original**	**Estimativa Corrigida**
Paramétrico	β_0	1.04248	1.31843
	β_1	0.00271	0.00109
	β_2	0.00010	0.00010
Bayesiano	β_0	1.04248	1.45369
	β_1	0.00271	0.00040
	β_2	0.00010	0.00009
Suavizado	β_0	1.04248	1.33830
	β_1	0.00271	0.00096
	β_2	0.00010	0.00009
Duplo	β_0	1.04248	1.51170
	β_1	0.00271	-0.00020
	β_2	0.00010	0.00010
Duplo rápido	β_0	1.04248	1.52018
	β_1	0.00271	-0.00056
	β_2	0.00010	0.00012
Bootknife	β_0	1.04248	1.37734

Revista brasileira de iniciação científica
ISSN: 2359-232X

	β_1	0.00271	0.00077
	β_2	0.00010	0.00010

Tabela 10: *Estimativas de máximo verossimilhança e estimativas corrigidas do modelo 1*

Método	Parâmetro	Estimativa Original	Estimativa corrigida
Paramétrico	β_0	0.99849	1.06031
	β_1	0.00292	0.00322
	β_2	0.00010	0.00007
	ϕ	48.08610	69.13668
Bayesiano	β_0	0.99849	1.06278
	β_1	0.00292	0.00318
	β_2	0.00010	0.00007
	ϕ	48.08610	68.48964
Suavizado	β_0	0.99849	1.05522
	β_1	0.00292	0.00329
	β_2	0.00010	0.00007
	ϕ	48.08610	69.26911
Duplo	β_0	0.99849	1.18056
	β_1	0.00292	0.00270
	β_2	0.00010	0.00006
	ϕ	48.08610	71.60722
Duplo rápido	β_0	0.99849	1.17760
	β_1	0.00292	0.00289
	β_2	0.00010	0.00005
	ϕ	48.08610	71.09129
Bootknife	β_0	0.99849	1.10166
	β_1	0.00292	0.00294
	β_2	0.00010	0.00007
	ϕ	48.08610	68.41829

Revista brasileira de iniciação científica
ISSN: 2359-232X

Tabela 11: *Estimativas de máximo verossimilhança e estimativas corrigidas do modelo 2.*

Verificamos se os pressupostos dos modelos são satisfeitos, utilizamos o teste Anderson-Darling (Anderson e Darling, 1952) para verificar se os resíduos de Pearson são normalmente distribuídos. Observa-se que a um nível de significância de 5%, não há evidência suficiente para rejeitar a hipótese nula, isto é, os resíduos são normalmente distribuídos. Finalmente, as Tabelas 12 e 13 apresentam-se os valores preditos junto aos valores reais para os diferentes métodos de estimação.

Edição	Paramétrico	Bayesiano	Suavizado	Duplo	Duplo Rápido	Bootknife	Real
1960	3.01	2.76	2.98	2.71	2.75	2.90	2
1964	2.88	2.63	2.85	2.56	2.59	2.77	1
1968	3.11	2.87	3.08	2.82	2.86	3.01	3
1972	3.25	3.01	3.22	2.96	2.99	3.15	2
1976	3.51	3.26	3.48	3.17	3.15	3.39	2
1980	4.18	3.97	4.16	3.90	3.88	4.07	4
1984	4.80	4.68	4.81	4.74	4.81	4.74	8
1988	5.48	5.42	5.50	5.47	5.51	5.43	6
1992	6.21	6.25	6.26	6.37	6.43	6.21	3
1996	9.62	10.11	9.80	10.22	9.98	9.73	15
2000	7.51	7.70	7.60	7.81	7.77	7.54	12
2004	9.02	9.53	9.18	9.95	10.07	9.19	10
2008	17.30	19.21	17.83	19.11	17.75	17.74	16
2012	22.97	25.72	23.76	24.21	20.93	23.38	17
2016	39.61	50.18	41.84	56.18	56.06	43.24	19

Tabela 12: *Valores preditos para o número médio de medalhas utilizado no ajuste do modelo 1.*

Revista brasileira de iniciação científica
ISSN: 2359-232X

Edição	Paramétrico	Bayesiano	Suavizado	Duplo	Duplo Rápido	Bootknife	Real
1960	3.17	3.17	3.17	2.92	2.90	3.10	2
1964	3.10	3.10	3.10	2.84	2.83	3.02	1
1968	3.27	3.28	3.28	3.03	3.01	3.21	3
1972	3.41	3.42	3.41	3.17	3.15	3.35	2
1976	3.78	3.78	3.78	3.53	3.54	3.70	2
1980	4.37	4.38	4.37	4.15	4.17	4.30	4
1984	4.64	4.65	4.62	4.49	4.45	4.62	8
1988	5.28	5.30	5.26	5.17	5.16	5.27	6
1992	5.82	5.84	5.79	5.78	5.76	5.84	3
1996	9.10	9.13	9.05	9.38	9.56	9.16	15
2000	7.09	7.11	7.05	7.15	7.20	7.11	12
2004	7.86	7.90	7.79	8.11	8.08	7.99	10
2008	16.82	16.84	16.72	18.23	19.15	16.97	16
2012	25.64	25.57	25.61	28.18	30.90	25.45	17
2016	27.53	27.81	27.00	32.98	33.37	29.35	19

Tabela 13: *Valores preditos para o número médio de medalhas utilizado no ajuste do modelo 2.*

Modelo	Métodos					
	Paramétrico	Bayesiano	Suavizado	Duplo	Duplo Rápido	Bootknife
Modelo1	11.41 [9,54-13,28]	9,74 [8,01-11,47]	10,34 [8,56-12,13]	9,24 [7,55-10,92]	9,71 [7,99-11,44]	10,95 [9,12-12,78]
Modelo 2	13,64 [11,41-15,88]	13,51 [11,28-15,73]	13,88 [11,62-16,14]	12,10 [10.01-14.18]	11,85 [9,79-13,91]	13,02 [10,84-15,20]

Tabela 14: *Previsão pontual e intervalar de 95% de confiança do número de medalhas do Brasil nos jogos olímpicos de 2021.*

Segundo o comitê olímpico do Brasil, participaram 314 atletas na olimpíada em Tóquio 2021 e de acordo ao Fundo Monetário Internacional (https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWOR

Revista brasileira de iniciação científica
ISSN: 2359-232X

LD/BRA) o Brasil obteve um PIB per capita de 7.741,153 dólares. A Tabela 14 mostra as estimativas de predição pontual e intervalar de 95% de confiança para o número de medalhas que Brasil conquistaria em Tóquio 2021. Na edição dos jogos olímpicos de 2021, Brasil conquistou 21 medalhas olimpíadas (07 de ouro, 06 de prata e 08 de bronze), quantidade ligeiramente superior aos valores apresentados na Tabela 4. Acreditamos que esta ligeira superioridade seja causada por diversos fatores dificilmente controláveis, por exemplo, os efeitos causados pelo vírus SARS-CoV-2, em que muitos atletas não têm demostrado seu máximo desempenho. Também, podem-se considerar outras variáveis socioeconômicas no modelo com o intuito a obter uma melhor precisão nos valores preditos.

Conclusões

Neste trabalho estudamos o método de *bootstrap* e algumas extensões para explicar a relação existente entre o número de medalhas obtidas pelos atletas Brasileiro ao longo da história olímpica em função ao número de participantes e o Produto Interno Bruto per capita. Os resultados obtidos mostraram que os métodos *bootstrap* utilizados para predizer o número de medalhas conquistadas nas próximas olimpíadas no Japão em 2021 é uma boa alternativa como método de estimação quando o tamanho da amostra é pequeno. Observamos que as estimativas do viés para o modelo binomial negativa é menor comparado com aos obtidos para o modelo normal e Poisson. O método *bootstrap* duplo apresentou um custo computacional superior aos outros métodos. Concluímos também, que ao assumir erroneamente que a variável resposta segue uma distribuição normal o sinal associado ao parâmetro β_0 é contrário comparada aos resultados obtidos assumindo distribuição Poisson ou binomial negativa para a variável resposta. Essa divergência pode levar a interpretações enganosas. O Brasil na edição dos jogos olímpicos de 2021 conquistou 21 medalhas olimpíadas (07 de ouro, 06 de prata e 08 de bronze), quantidade superior aos valores preditos obtidos pelo modelo Poisson e binomial negativa. Sugere-se a incorporação de outras variáveis socioeconômicas no modelo para melhorar a estimação dos valores de predição.

Agradecimentos

Os autores agradecem ao editor e os avaliadores pelos comentários; à Universidade Federal da Bahia (UFBA) pela bolsa de Iniciação Cientifica associado ao projeto N°16907, concedida ao primeiro autor.

Referências

AKAIKE, H. A new look at the statistical model identification**. IEEE Transactions on Automatic Control**. v.19, p. 716-723, 1974.

ANDERSON, T. W. e DARLING, D. A. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. **Annals of Mathematical Statistics**. v.23, p. 193–212, 1952.

CIRILLO, M. A. *et al.* Avaliação de métodos de estimação intervalar para funções lineares binomiais via Bootstrap finito. **Ciência e Agrotecnologia.** v, 33, p, 1741 -1746, 2009.

CHEN, P.Y. e POPOVICH, P.M. **Correlation: parametric and nonparametric measures**. Thousand Oaks: Sage Publication, 2002.

CLÍMACO, G. N. **Otimização da extração de compostos bioativos da beterraba por metodologia de superfície de resposta e método de Bootstrap.** 2019. Dissertação (mestrado em Engenharia de Alimentos)-Universidade Estadual de Maringá.

DAVIDSON, R. e MACKINNON, J. G. Improving the reliability of bootstrap tests with the fast double bootstrap**. Computational Statistics Data Analysis.** v.51, p. 3259-3281, 2007.

DOGAN, C. D. Applying bootstrap resampling to compute confidence intervals for various statistics with R. **Eurasian Journal of Educational Research**. v.68, p. 1-17, 2017.

EDGARD MATSUKI. Brasil sobe de 37º para 35º no ranking histórico das Olimpíadas. **Agência Brasil**, c2016. Disponível em: https://agenciabrasil,ebc,com,br/rio-2016/noticia/2016-08/brasil-sobe-de-37o-para-35o-no-quadro-de-medalhas-com-19-conquistadas-no. Acesso em: 08 de jul. de 2021

EFRON, B. Bootstrap methods: another look at the jackknife**. The annals of Statistics**. v.7, p. 1–26, 1979.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation, **Journal of the American Statistical Association**. v.78, p. 316–331, 1983.

EFRON, B. e TIBSHIRANI, R. J. **An introduction to the bootstrap.** Chapman and Hall: New York, 2014.

GILES, H. e MENTCH, L. Bootstrap bias corrections for ensemble methods. **Statistics and Computing**. V. 28, p. 77-86, 2018.

HESTERBERG, T. C. **Smoothed bootstrap and jackboot sampling**. MathSoft. Inc. Seattle, 1999.

KAGGLE. Your Machine Learning and Data Science Community, Disponível em: https://kaggle,com. Acesso em: 08 de jul. de 2021

LIMA, F. P. **Inferência bootstrap em modelos de regressão beta.** 2017. Tese (doutorado Estatística) - Universidade Federal de Pernambuco.

MCCULLAGH, P. e NELDER, J. A. **Generalized Linear Models.** Chapman and Hall: London, 1989.

MOREIRA, G. R. F. **Aplicação de método estocástico no cálculo das provisões de sinistros.** 2020. Trabalho de conclusão de curso de graduação - Universidade Federal de São Paulo.

NELDER, J. A. e WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society**, v.135, p. 370-384, 1972.

PAULA, G. A. **Modelos de regressão com apoio computacional.** Instituto de Matemática e Estatística, Universidade de São Paulo, 1993.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna: Austria, 2021.

ROSSUM, G. V. e DRAKE, F. L. **Python Tutorial.** Python Software Foundation, 2012.

RUBIN, D. J. The bayesian bootstrap. **The annals of Statistics.** v.9, p. 130–134, 1981.

TANG, G. e LI, J. Regression analysis-based chinese olympic games competitive sports strength evaluation model research. **The Open Cybernetics Systemics Journal.** v.9, p. 2729–2735, 2015.

THE WORLD BANK. Is a unique global partnership fighting poverty worldwide through sustainable solutions. Disponível em: https://data,worldbank,org/. Acessado em: 10 de jul, de 2021.

WIT, E. *et al*. 'all models are wrong,,,': an introduction to model uncertainty**. Statistica Neerlandica**, v.66, p. 217–236, 2012.