Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 86 (2016) 228 – 231

Procedia Computer Science

2016 International Electrical Engineering Congress, iEECON2016, 2-4 March 2016, Chiang Mai, Thailand

Bootstrapping with R to make generalized inference for regression model

Jutatip Sillabutra[a], Prasong Kitidamrongsuk[a,*], Chukiat Viwatwongkasem[a], Chareena Ujeh[b], Siam Sae-tang[b], Khanokporn Donjdee[b]

[a]Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, 10400, Thailand
[b] Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok, 10400, Thailand

Abstract

Bootstrap is a resampling procedure drawn from an original sample data with replacement allocation method to build a sampling distribution of a statistic for statistical inference. This paper focuses to validate the generalized linear regression model by using the bootstrap method in order to make generalization of statistical inference to the different settings outside the original. The first application involved the bootstrap regression coefficients of predictors in the classical regression model while the others emphasized the bootstrap responses for binary outcomes in the logistic regression and for count data in the Poisson regression. The results on the bootstrap regression coefficients perform well even if the original data were restricted with small sample sizes and/or non-normal errors. The confidence intervals based upon the normal theory is quite narrower than the percentile interval and the bootstrap t interval. For the results of the bootstrap responses along a single predictor, both percentile confidence intervals of logistic and Poisson regression models perform well with a nice bandwidth of bootstrap responses for generalization.

Keywords: Bootstrapping for Regression; Generalized Inference; Model Validation.

1. Introduction

In statistical modeling, generalized linear model (GLM) allows the response variable (Y) to be related to the predictors $X_1,...,X_k$ via a canonical link function $g(\mu)$:

$$g(\mu)=\beta_0+\beta_1X_1...+\beta_kX_k+offset$$

where $\mu=E(Y)$ is the response mean depending on linear predictors $X_1,...,X_k$, $\beta_0,\beta_1,...,\beta_k$ are regression coefficients, and $offset$ refers to a specific term used in Poisson regression for rate data. Some observed link functions for a random sample of size n are:

doi:10.1016/j.procs.2016.05.103

$\hat{\mu} = \hat{Y} = b_0 + b_1 X_1 + \ldots + b_k X_k$ where $g(\hat{\mu}) = \hat{\mu} = \hat{Y}$ for normally continuous variable Y

$\ln\left(\dfrac{\hat{p}}{1-\hat{p}}\right) = b_0 + \sum_{i=1}^{k} b_i X_i$ where $g(\hat{\mu}) = \ln\left(\dfrac{\hat{\mu}}{1-\hat{\mu}}\right) = \ln\left(\dfrac{\hat{p}}{1-\hat{p}}\right)$ for binary outcomes Y

$\ln \hat{\mu} = b_0 + \sum_{i=1}^{k} b_i X_i$ where $g(\hat{\mu}) = \ln \hat{\mu}$ for Poisson counting variable Y

$\ln \hat{\mu} = b_0 + \ln N + \sum_{i=1}^{k} b_i X_i$ where $\ln N = offset$, N = exposure time (person-time),

for Poisson counting variable Y in the rate (Y / N) of interest.

After obtaining the regression results, model validation is the next important step to assess whether the results of analysis are likely to hold outside of the original research setting. In other words, we attempt to determine the generalization of a regression model to the same target population but the different setting. There are many statistical tools for model validation, such as cross-validation, resampling technique with split-half data, Jackknife and bootstrap resampling methods. The bootstrap method introduced by Efron in 1979 [1] is a resampling procedure with replacement to estimate the sample statistics, such as medians, variances, percentiles, to perform significance tests, to construct confidence intervals, to validate the regression models, and to approximate the sampling distribution of some statistics. With bootstrapping inference, the population is the original sample, as the sample is the bootstrap samples. A gap of study is the use of bootstrap to make generalization of regression model to another setting outside the original and the R program is implemented for these.

2. Methodology and Design

Suppose that a GLM function based on an original sample of size n is obtained as

$$g(\hat{\mu}_i) = b_0 + b_1 X_{i1} + \ldots + b_k X_{ik}, \qquad \text{for } i = 1, 2, \ldots, n$$

The straight forward approach is to collect the response and predictor values for each i observation $\mathbf{z}_i' = [Y_i, X_{i1}, \ldots, X_{ik}]$. The original n observations $\mathbf{z}_1', \mathbf{z}_2', \ldots, \mathbf{z}_n'$ can be resampled for each observation with replacement allocation under equal probability $1/n$, leading to the resulting bootstrap samples, $\mathbf{z}_{b1}'^{*}, \mathbf{z}_{b2}'^{*}, \ldots, \mathbf{z}_{bn}'^{*}$, assuming producing m sets (replicates) of bootstrap samples. For each b replicate $(b = 1, 2, \ldots, m)$, the response mean estimate $\hat{\mu}_{bi}^{*}$ and the regression coefficient estimates $\mathbf{b}_{bj}^{*} = [b_{bo}^{*}, b_{b1}^{*}, \ldots, b_{bk}^{*}]'$ can be computed. Then the bootstrap average under m bootstrap replicates for the response mean and the regression coefficients are $\hat{\mu}_b^{*} = \frac{1}{m}\sum_{b=1}^{m} \hat{\mu}_{bi}^{*}$ and $\mathbf{b}_b^{*} = \frac{1}{m}\sum_{b=1}^{m} \mathbf{b}_{bj}^{*}$, respectively. Also, their standard errors of a series of bootstrap replicates can be approximated as

$$SE(\hat{\mu}_b^{*}) = \sqrt{\frac{1}{m-1}\sum_{b=1}^{m} (\hat{\mu}_{bi}^{*} - \hat{\mu}_b^{*})^2} \text{ and } SE(\mathbf{b}_b^{*}) = \sqrt{\frac{1}{m-1}\sum_{b=1}^{m} (\mathbf{b}_{bj}^{*} - \mathbf{b}_b^{*})(\mathbf{b}_{bj}^{*} - \mathbf{b}_b^{*})'}, \text{ respectively.}$$

The first bootstrap confidence intervals are based on normal approach as $\hat{\mu}_b^{*} \pm t_{n-(k+1),\alpha/2}\, SE(\hat{\mu}_b^{*})$ and $\mathbf{b}_b^{*} \pm t_{n-(k+1),\alpha/2}\, SE(\mathbf{b}_b^{*})$ where $t_{n-(k+1),\alpha/2}$ is the critical value of t with the right probability $\alpha/2$ for $n-(k+1)$ degrees of freedom; if sample size n is large, then Z distribution values are used instead of t confidence intervals. Secondly, the 95% percentile bootstrap confidence interval as a nonparametric approach can be constructed from $\mathbf{b}_{b(lower)}^{*} < \mathbf{b} < \mathbf{b}_{b(upper)}^{*}$ where $\mathbf{b}_{b(lower)}^{*} = \mathbf{b}_b^{*}$ at $0.025\,m$ and $\mathbf{b}_{b(upper)}^{*} = \mathbf{b}_b^{*}$ at $0.975\,m$ of the m ordered bootstrap replicates. Lastly, the bootstrap t confidence interval can be applied by using bootstrap