STAPRO: 8064 Model 3G pp. 1–9 (col. fig: NIL)

ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bootstrapping for multivariate linear regression models

Daniel J. Eck

Department of Biostatistics, Yale School of Public Health, 60 College St. LEPH PO Box 208034, New Haven, CT., 06510, USA

A R T I C L E I N F O

Article history:
Received 24 April 2017
Received in revised form 2 November 2017
Accepted 6 November 2017
Available online xxxx

Keywords:
Multivariate bootstrap
Multivariate linear regression model
Residual bootstrap

A B S T R A C T

The multivariate linear regression model is an important tool for investigating relationships between several response variables and several predictor variables. The primary interest is in inference about the unknown regression coefficient matrix. We propose multivariate bootstrap techniques as a means for making inferences about the unknown regression coefficient matrix. These bootstrapping techniques are extensions of those developed in Freedman (1981), which are only appropriate for univariate responses. Extensions to the multivariate linear regression model are made without proof. We formalize this extension and prove its validity. A real data example and two simulated data examples which offer some finite sample verification of our theoretical results are provided.

1. Introduction

The linear regression model is an important and useful tool in many statistical analyses for studying the relationship among variables. Regression analysis is primarily used for predicting values of the response variable at interesting values of the predictor variables, discovering the predictors that are associated with the response variable, and estimating how changes in the predictor variables affects the response variable (Weisberg, 2005). The standard linear regression methodology assumes that the response variable is a scalar. However, it may be the case that one is interested in investigating multiple response variables simultaneously. One could perform a regression analysis on each response separately in this setting. Such an analysis would fail to detect associations between responses. Regression settings where associations of multiple responses are of interest require a multivariate linear regression model for analysis.

Bootstrapping techniques are well understood for the linear regression model with a univariate response (Bickel and Freedman, 1981; Freedman, 1981). In particular, theoretical justification for the residual bootstrap as a way to estimate the variability of the ordinary least squares (OLS) estimator of the regression coefficient vector in this model has been developed (Freedman, 1981). Theoretical extensions of residual bootstrap techniques appropriate for the multivariate linear regression model have not been formally introduced. The existence of such an extension is stated without proof and rather implicitly in subsequent works (Freedman and Peters, 1984; Diaconis and Efron, 1983). In this article we show that the bootstrap procedures in Freedman (1981) provide consistent estimates of the variability of the OLS estimator of the regression coefficient matrix in the multivariate linear regression model. Our proof technique follows similar logic as Freedman (1981). The generality of the bootstrap theory developed in Bickel and Freedman (1981) provide the tools required for our extension to the multivariate linear regression model.

2. Bootstrap for the multivariate linear regression model

The multivariate linear regression is

$$Y_i = \beta X_i + \varepsilon_i, \quad (i = 1, \dots, n), \tag{1}$$

E-mail address: daniel.eck@yale.edu.

https://doi.org/10.1016/j.spl.2017.11.001

Please cite this article in press as: Eck D.J., Bootstrapping for multivariate linear regression models. Statistics and Probability Letters (2017), https://doi.org/10.1016/j.spl.2017.11.001.

where $Y_i \in \mathbb{R}^r$ and $r > 1$ in order to have an interesting problem, $\beta \in \mathbb{R}^{r\times p}$, $X_i \in \mathbb{R}^p$, and the $\varepsilon_i's \in \mathbb{R}^r$ are errors having mean zero and variance–covariance matrix Σ where $\Sigma > 0$. It is assumed that separate realizations from the model (1) are independent and that $n > p$. We further define $\mathbb{X} \in \mathbb{R}^{n\times p}$ as the design matrix with rows X_i^T, $\mathbb{Y} \in \mathbb{R}^{n\times r}$ is the matrix of responses with rows Y_i^T, and $\varepsilon \in \mathbb{R}^{n\times r}$ is the matrix of all errors with rows ε_i^T. The OLS estimator of β in model (1) is $\hat{\beta} = \mathbb{Y}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$. We let $\widehat{\varepsilon} \in \mathbb{R}^{n\times r}$ denote the matrix of residuals consisting of rows $\widehat{\varepsilon}_i^T = (Y_i - \hat{\beta}X_i)^T$. The multivariate linear regression model assumed here is slightly different than the traditional multivariate linear regression model. The traditional model makes the additional assumptions that the errors are normally distributed and the design matrix $\mathbb{X}$ is fixed.

We consider two bootstrap procedures that consistently estimate the asymptotic variability of $\mathrm{vec}(\hat{\beta})$ under different assumptions placed upon the model (1), where the vec operator stacks the columns of a matrix so that $\mathrm{vec}(\hat{\beta}) \in \mathbb{R}^{rp\times 1}$. The first bootstrap procedure is appropriate when the design matrix $\mathbb{X}$ is assumed to be fixed and the errors are constant. In this setup, residuals are resampled. The second bootstrap procedure is appropriate when $(X_i^T, \varepsilon_i^T)^T$ are realizations from a joint distribution. In this setup, cases $(X_i^T, Y_i^T)^T$ are resampled. It is known that bootstrapping under these setups provides a consistent estimator of the variability of $\mathrm{var}(\hat{\beta})$ in model (1) when $r = 1$ (Freedman, 1981). Convergence theorems are stated in terms of the Mallows metric for two probability measures μ, ν in $\mathbb{R}^k$. The Mallows metric is

$$d_l^p(\mu,\nu) = \inf_{U\sim\mu, V\sim\nu} E^{1/l}\left(\|U - V\|^l\right). \tag{2}$$

A brief description of useful properties of (2) is stated in the beginning of Section 4. We now provide the needed multivariate bootstrap extensions.

2.1. Fixed design

We first establish the residual bootstrap of Freedman (1981) when $\mathbb{X}$ is assumed to be a fixed design matrix. Resampled, starred, data is generated by the model

$$\mathbb{Y}^* = \mathbb{X}\hat{\beta}^T + \varepsilon^*, \tag{3}$$

where $\varepsilon^* \in \mathbb{R}^{n\times r}$ is the matrix of errors with rows being independent. The rows in ε^* have common distribution $\widehat{F}_n$ which is the empirical distribution of the residuals from the original dataset, centered at their mean. Now $\hat{\beta}^* = \mathbb{Y}^{*T}\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$ is the OLS estimator of β from the starred data. This process is performed a total of B times with a new estimator $\hat{\beta}^*$ computed from (3) at each iteration. We then estimate the variability of $\mathrm{vec}(\hat{\beta})$ with

$$\mathrm{var}^*\left\{\mathrm{vec}(\hat{\beta})\right\} = (B-1)^{-1}\sum_{b=1}^{B}\left\{\mathrm{vec}(\hat{\beta}_b^*) - \mathrm{vec}(\bar{\beta}^*)\right\}\left\{\mathrm{vec}(\hat{\beta}_b^*) - \mathrm{vec}(\bar{\beta}^*)\right\}^T$$

where $\hat{\beta}_b^*$ is the residual bootstrap estimator of β at iteration b and $\bar{\beta}^* = B^{-1}\sum_{b=1}^{B}\hat{\beta}_b^*$. We summarize this bootstrap procedure in Algorithm 1.

Algorithm 1. Bootstrap procedure with fixed design matrix.

Step 1. Set B and initialize $b = 1$.
Step 2. Sample residuals from $\widehat{F}_n$, with replacement, and compute $\mathbb{Y}^*$ as in (3).
Step 3. Compute $\hat{\beta}_b^* = \mathbb{Y}^{*T}\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$, store $\mathrm{vec}(\hat{\beta}_b^*)$, and let $b = b + 1$.
Step 4. Repeat Steps 2–3 $B - 1$ times.

Before the theoretical justification of the residual bootstrap is formally given, some important quantities are stated. The residuals from the regression (3) are $\widehat{\varepsilon}^* = \mathbb{Y}^* - \mathbb{X}\hat{\beta}^{*T}$. The variance–covariance matrix Σ in model (1) is then estimated by

$$\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i\widehat{\varepsilon}_i^T - \hat{\mu}^2, \qquad \hat{\mu}^2 = \left(n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i\right)\left(n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i\right)^T.$$

Likewise, the variance–covariance estimate from the starred data is

$$\widehat{\Sigma}^* = n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i^*\widehat{\varepsilon}_i^{*T} - \hat{\mu}^{*2}, \qquad \hat{\mu}^{*2} = \left(n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i^*\right)\left(n^{-1}\sum_{i=1}^{n}\widehat{\varepsilon}_i^*\right)^T.$$

Let I_k denote the $k \times k$ identity matrix. Theorem 1 provides bootstrap asymptotics for the regression model (1). It extends Theorem 2.2 of Freedman (1981) to the multivariate setting.

Theorem 1. *Assume the regression model (1) where the errors have finite fourth moments. Suppose that $n^{-1}\mathbb{X}^T\mathbb{X} \to \Sigma_X > 0$. Then, conditional on almost all sample paths $Y_1, \ldots, Y_n$, as $n \to \infty$,*