A Bootstrap Approach to Robust Regression

Dr. Hamadu Dallah
Department of Actuarial Science and Insurance
University of Lagos
Akoka, Lagos, Nigeria.

Abstract

We focus on the derivation of consistent estimates of the standard deviations of estimates of the parameters of a multiple regression model fitted via a robust procedure, namely, the so-called M (M for maximum likelihood) regression fitting method. M-regression is mostly actualized by way of weighted least squares (WLS). It is common knowledge that most commonly used statistical packages offering WLS assume that the weights are fixed. In this scenario M-regression yields standard errors that are inconsistent and unstable, moreso if the underlying sample is small. The alternative approach on offer in this article is the bootstrap. Using the re-sampling mechanism inherent in bootstrapping, it is demonstrated empirically that bootstrap standard errors are smaller than their M-regression counterparts.

Key words: M-Regression, WLS, Standard Errors, Bootstrap Methods, and Bootstrap Standard Errors.

1. Introduction

Bootstrapping was first introduced into regression by Efron (1979). Since then much research has gone into investigating the performance of the bootstrap method in regression. Freedman (1981) offers an early theoretical analysis of the asymptotic theory of the bootstrap for regression and correlation models. Specifically, the author has shown that the bootstrap approximation to the distribution of least squares parameters estimates is valid. Freedman's work was extended by Wu (1986) whose intervention itself was extensively discussed by Efron and Tibshirani (1993) and Wilcox (2001). Freedman and Peters (1984) present the bootstrap in the context of an econometric regression model, describing the demand for energy by industry. The main finding is that for generalized least squares with estimated covariance matrix, the asymptotic formula for standard errors can be too optimistic, sometimes by quite large factors. Thus, the bootstrap procedure is appreciably better than the conventional asymptotic approach when applied to the finite – sample situation. Stine (1985) uses the bootstrap to set prediction intervals in regression. These intervals approximate the nominal coverage probability in small samples without requiring specific assumptions about the sampling distribution. The asymptotic properties of the intervals do not depend upon the sampling distribution and Monte Carlo results suggest that invariance approximately holds for relatively small samples. Furthermore, Stine states that the use of the bootstrap does however require certain assumptions; for example, assumptions such as that the specified model be the correct model. In the same vein Efron (1983, 1986) extended the problem of prediction rule to general exponential families with emphasis on logistic regression. After establishing a general theory for prediction rule, Efron uses the bootstrap to estimate error rate of a prediction rule and also determine how biased the apparent error rate is. Breiman (1996) demonstrates the use of the bootstrap for the more primary purpose of producing efficient estimates of regression parameters. Tibshirani and Knight (1999) have proposed a bootstrap – based method for enhancing a search through a space of models, including applications to regression models. Finally, Hamadu (2003) has extensively studied the use of bootstrapping under a variety of regression settings.

This article reports yet another contribution to the kinds of research efforts described above; that is, research efforts directed towards the study of the performance of the bootstrap in regression. Specifically, we demonstrate empirically that the bootstrap is a veritable instrument to enhance the efficiency of robust (M) regression.

We briefly review M regression in Section 2. Section 3 describes the critical steps of the bootstrap in regression. We show an empirical example in Section 4. The article is concluded with a summary and some comments in Section 5.

2. Review of M-Regression

The usual multiple regression model, in matrix notation is

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad (2.1)$$

where,

$\underline{Y}$ is an $n \times 1$ vector of observations of the response variable Y

X is an $n \times p$ (design) matrix of known constants

$\underline{\beta}$ is an $p \times 1$ vector of unknown regression coefficients and

$\underline{\varepsilon}$ is an $n \times 1$ vector of random errors.

It is assumed that elements of $\underline{\varepsilon}$ are independent and identically distributed and $V(\underline{\varepsilon}) = \sigma^2 I_n$ where I_n is an $n \times n$ identity matrix and σ^2 (>0) is a constant. For the estimation of $\underline{\beta}$ by ordinary least squares (OLS) it is further required that the data at hand be well – behaved, that is, that data are devoid of outliers.

Robust or specifically M-regression is a good alternative to OLS in the event that there are outliers in the data. M-regression is described as follows.

Consider the function

$$\sum \rho\left(\frac{Y_i - \underline{X}_i\underline{\beta}}{\hat{\sigma}}\right) \qquad (2.2)$$

where Y_i is the ith element of $\underline{Y}$ (2.2)

$\underline{X}_i$ is the ith row of X and $\hat{\sigma}$ is a robust estimate of σ. The function is to be maximized with respect to the elements of $\underline{\beta}$. Thus, differentiating (2.2) partially with respect to the elements of $\underline{\beta}$, say β_j, and equating the derivatives equal to zero, we have

$$\sum x_{ij}\psi\left(\frac{Y_i - \underline{X}_i\underline{\hat{\beta}}}{\hat{\sigma}}\right) = 0 \qquad (2.3)$$

where $\psi(.)$ represent $\rho'(.)$, that is, derivative of ρ, and x_{ij} is the (ij)th element of X. The maximizing values $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ associated with the p equations are called the M estimators of the elements, $\beta_1, \beta_2, ..., \beta_p$ of $\underline{\beta}$, or we can just say that $\underline{\hat{\beta}}$ is M estimator of $\underline{\beta}$. Hogg (1979) gives a detailed account of how $\underline{\hat{\beta}}$ can be improved upon using weighted least squares (WLS). This is summarized in the following steps:

1. Begin with initial estimates $\underline{\hat{\beta}}_0$ and $\hat{\sigma}_0$. (Note that it is convenient to take OLS estimate of $\underline{\beta}$ to be initial estimate $\underline{\hat{\beta}}_0$ and following this $\hat{\sigma}_0 = median\left\{\left|Y_i - \underline{X}_i\underline{\hat{\beta}}_0\right|\right\}$
2. Calculate residuals $r_i = Y_i - \underline{X}_i^t\underline{\hat{\beta}}_0$, $i = 1,2,...,n$
3. Calculate weights $w_i = \psi(r_i)/r_i$
   Hence form $n \times n$ diagonal matrix of weights W whose diagonal elements are w_i
4. Carry out weighted least squares (WLS) to yield new
   $\underline{\hat{\beta}}_{(1)} = (X^tWX)^{-1}X^tW\underline{Y}$
5. Iterate between Step 2 through Step 4 until convergence.

A few pertinent remarks are in order:

(i) an approach that is slightly different from the above is to estimate $\underline{\beta}$ and σ simultaneously . Dutter (1977) has described how this can be done.

(ii) The choice of an appropriate weighting function ψ is critical in M-regression. Hogg (1979) has given some tips to guide selection of an appropriate ψ function from among those that are commonly used in practice, Huber's, and Tukey's biweight function s are used in the present article.

3. The Bootstrap in Robust Regression

Robust estimators such as $\underline{\hat{\beta}}$ of $\underline{\beta}$ in model (2.1) are not maximum likelihood estimators in the classical sense. This is because the form of the distribution of $\underline{\varepsilon}$ is not known. Specifically, the distribution function $F(\underline{\varepsilon})$ is not specified. By the same token $F(\underline{\hat{\beta}})$) is unknown. Which goes to show that M estimators are essentially non parametric. We venture to say that this non-parametric environment provides a proper setting for the bootstrap methodology to be applied.

Let $\underline{\hat{\varepsilon}} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n)^t = \underline{Y} - X\,\underline{\hat{\beta}}$ denote the residuals from the fitted (robust) regression. The bootstrap sample $\hat{\varepsilon}_1^*, \hat{\varepsilon}_2^*, ..., \hat{\varepsilon}_n^*$ is generated by sampling $\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_n$ with replacement. Thus, the bootstrap sample leaves out some elements from $\hat{\varepsilon}_i$ $(i = 1,2,...,n)$ but could include other elements two, three, four or more times. Now defining the bootstrap observations as

$$\underline{Y}_i = \underline{X}_i \underline{\hat{\beta}} + \underline{\hat{\varepsilon}}_i^* \ , \quad \text{i=1,2,…,n}$$

we can obtain $\underline{\hat{\beta}}^*$ as the solution to

$$\sum x_{ij}\left(\frac{Y_i^* - \underline{X}_i \underline{\hat{\beta}}^*}{\hat{\sigma}}\right) = 0 \ ; \quad j = 1,2,..., p \qquad (2.4)$$

Notice the similarity of (2.3) and (2.4); the similarity simply shows that applying the robust estimator to the original sample ($\underline{Y}$, X) yields $\underline{\hat{\beta}}$, and applying the same estimator to ($\underline{Y}^*$, X) yields $\underline{\hat{\beta}}^*$, namely, the bootstrap estimate. As indicted earlier $F(\underline{\hat{\beta}})$ represents the true but unknown distribution function of $\underline{\hat{\beta}}$, and $\hat{F}(\underline{\hat{\beta}}^*)$ denotes the observed distribution function of $\underline{\hat{\beta}}^*$, which is known by virtue of the fact that it is obtained via many Monte Carlo repetitions of the bootstrap sampling process described earlier . That is, if we draw bootstrap samples a large number of times, B times say, then the B values of $\underline{\hat{\beta}}^*$ will yield $\hat{F}(\underline{\hat{\beta}}^*)$ which approximates a maximum likelihood estimate of $F(\underline{\hat{\beta}})$. The bootstrap variance estimates the true but unknown variance of $\underline{\hat{\beta}}$.

In practical terms, there are two ways to carry out bootstrapping in regression analysis where one has data ($\underline{Y}$, X) following the model in (2.1).

One way is to resample the residuals from the fitted model and the other is to resample the data ($\underline{Y}$, X).

3.1 Bootstrapping Regression Via Residual Resampling

Residual bootstrapping proceeds using the following steps:

i. Perform regression with the original sample ($\underline{Y}$, X) to calculate predicted values $\underline{\hat{Y}}$ and residuals $\underline{r}$

ii. Randomly resample the residuals with replacement, but leave X and $\underline{\hat{Y}}$ values unchanged. Let the bootstrap residuals be denoted by r*.

iii. Construct new $\underline{Y}^*$ values by adding $\underline{r}^*$ to the original predicted values to yield $\underline{Y}^* = \underline{\hat{Y}} + \underline{r}^*$.

iv. Regress $\underline{Y}^*$ on the original X variable(s).

v. Repeat steps (ii) to (iv) B times.

We then study the distribution of the bootstrap estimate $\underline{\hat{\beta}}^*$ across the B bootstrap samples.

3.2 Data Bootstrapping

Data resampling, otherwise called model – free bootstrap, bootstraps regression without assuming fixed X or identically distributed errors. It proceeds as follows:

i. Randomly choose samples of size n, sampling complete cases ($\underline{Y}$, X) from the original data with replacement
ii. Within each bootstrap sample regress $\underline{Y}$ * on the X* variable(s) as usual

Unlike residual resampling, data resampling, as noted above, does not assume independent and identically distributed errors. Since it allows for other possibilities, and also admits random X values as a new source of sample-to-sample variation, data resampling often yields results quite different from those expected under the usual regression assumptions. Stine (1990) recommends basing the choice of residual versus data sampling on how the data were collected. Residual resampling would be preferred if the fixed X assumption is realistic. Otherwise, if X varies as randomly as $\underline{Y}$ then data resamping should be the choice. In either case we want the process of bootstrap resampling to mimic the way in which the sample was originally selected from the population.

4. Application

4.1 Description of Data

When oil prices rose during the 1970s, wood stoves came back into fashion for heating in parts of the country. Although it is often cheaper than other sources, wood burning pollutes both outdoor and indoor air. The following table gives measures of the peak carbon monoxide (CO) levels during 11 tests of wood-burning stoves. Robust methods are particularly appropriate here due to two unusual tests (9&10): the stove F overheated, possibly due to overfilling with wood, and experimenters reduced airflow by using a damper that caused the house to fill with smoke. Such incidents are common with no airtight stoves, especially with inexperienced operators (see Hamilton 1992).

Table 6.1 Data on Indoor carbon monoxide pollution from wood-burning stoves

Test	Stove Type	Burning Time (hours)	Amount of Wood Burned (kg)	Peak House CO (ppm)
1	Airtight	14.8	37.3	2.8
2	Airtight	8.8	38.4	1.2
3	Airtight	13.0	21.2	1.6
4	Airtight	13.7	27.2	2.0
5	Airtight	18.5	40.6	1.2
6	Airtight	18.0	43.2	1.4
7	Airtight	16.1	24.2	3.8
8	Non airtight	8.724.4	7.7	
9	Non airtight	10.4	32.4	35.0
10	Non airtight	5.4	23.2	43.0
11	Non airtight	9.5	38.6	3.5

X = Burning Time, X_2 = Amount of Wood Burned and as mention above Y = CO

4.2 Regression Model for the Data

The following regression model is proposed for the data:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1,2,...,11 \qquad (4.1)$$

where,

$\beta_0, \beta_1, \beta_2$ are regression parameters, and ε is random error in Y assumed to have constant variance, that $V(\varepsilon) = \sigma^2$. Furthermore, for inference purposes, it is necessary to assume that $\varepsilon \sim N(0, \sigma)$.

For fitting the model in (4.1) to the data, we used three robust regression techniques, namely, robust biweight regression on one hand and two bootstrap-based robust procedures on the other.
The results of the regression fits are presented in Table 4.1.

Table 4.1 Regression Estimates and their Corresponding Standard Errors in Brackets.

	Estimates		
Methods of Estimation	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
WLS based on Huber's weight with *c=1.345*	20.497 (10.779)	-0.977 (0.681)	-0.066 (0.311)
WLS based on Biweight weight with *c = 4.875*	10.153 (1.913)	-0.347 (0.121)	-0.065 *(0.052)*
Robust M-regression via Model Bootstrap with *B=500*	53.462 (0.08)	-0.650 (0.081)	-0.280 (0.006)
Robust M-regression via Data Bootstrap with *B=500*	35.262 *(1.150)*	-0.456 (0.153)	-0.101 (0.04)

Relevant entries in the table show that estimates of regression from the bootstrap robust regression fitting methods have uniformly smaller standard errors than those from the biweight regression. This result is an indication that bootstrapping can serve as an instrument for boosting the efficiency of robust regression, which in essence is the main aim of this research. However, we are surprised at the large differences in magnitudes of the estimated coefficients although each estimate maintains the same sign across the three methods under consideration. As for the two bootstrap robust regression models, it is hardly surprising that their results differ. It is not a surprise because, as noted earlier in section 3.2 above, data resampling, does not necessarily assume that the design X is fixed; instead, it can admit random X which occasions greater variability in the estimation data. Consequently, data resampling often yields results that are quite different from those of residual resampling; the latter depending on the usual least squares assumptions for its validity.

References

Andrews et al (1972), Robust Estimates of Location. Survey and Advances, Princeton, F *Princeton, University Press*

Breiman, P. (1996) On Robust Estimation. *Annals of Statistics,* **9.** 1196 - 1217.

Freedman, D.A. (1981), Bootstrapping Regression Models. *Annals of Statistics*, 9, 1218-1228.

Freedman, D.A and Peters S.C. (1984), Bootstrapping Regression Equation: some Empirical Results. *Journal of American Statistical Association*, **79**, 97, 106.

Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife, Annuals of Statistics, 7, 1- 26.

Efron, B. (1983), Estimating the Error Rate of a Prediction Rule: Improvement on Cross-validation. *Journal of American Statistical Association*, 78, 316-330.

Efron, B (1987), Better Bootstrap Confidence Intervals (with discussion). *Journal of American Statistical Association*, 82, 171-185.

Efron, B. and Tibshirani R. (1993), introduction to the Bootstrap, Chapman and Hall International t Thomson Publishing, New York**.**

Hamadu, D (2003), Bootstrapping Heteroscedastic Regression Models . Unpublished PhD Thesis. Department of Mathematics, University of Lagos, Lagos, Nigeria.

Hamilton, L.C. (1992), regression with Graphics: A second Course in Applied Statistics, Duxbury Press. California.

Hampel, F.R. (1974), The Influence Curve and its role in Robust Estimation. Journal of American Statistical Association, **69**, 383 – 394.

Huber, P. J. (1967), The Behaviour of Maximum Likelihood Estimates under Nonstandard Conditions. Proceedings Fift Berkeley Symposium Mathematic – Statistics and Probability **1**, 221 -233.

Huber, P. J. (1981), Robust Regression. John Wiley and Sons – New York.

Stine, R. A (1985) Bootstrap Prediction Intervals for Regression. Journal of American Statistical Association, **80**, 1026-1031.

Tibshirani , R. and Knight, RK (1999), Model Search by the Bootstrap "Bumping." *Journal of Computational and Graphical Statistics*, 8, 671-686.

Wilcox, R. R. (2001). Fundamentals of Modern Statistical Methods. Springer-Verlag New York.

Wu, C. J. F. (1986), Jackknife Bootstrap and other Sampling Methods in Regression Analysis. *The Annals of Statistics*, 14, 1261-1294.