Boot-Strapping in Simple Regression

Project 2
Regression Analysis

Amal Agarwal
Roll. no.: 09D11001

Sagar Chordia
Roll. no.: 09005013

Tuhin Sarkar
Roll. no.: 09007030

Advisor: Prof. Siuli Mukhopadhyay

Department of Mathematics
Indian Institute of Technology, Bombay
2013

Abstract

This project focusses on the technique of Bootstrapping in Simple Regression. It combines the least squaress estimate (OLS), least absolute deviation estimate (LAD), least median estimate (LMS) methods of estimation of parameters with the Bootstrap technique, when the overall error distribution is unknown or is not the normal distribution. The primary aim is to improve stability of regression coefficient and reduce the length of confidence interval.

Chapter 1

Review of Bootstrapping in Simple Regression

1.1 Introduction

Bootstrap re-sampling methods can be divided into two types- Model based bootstrap regression and Cases based bootstrap regression. Three estimation techniques with different algorithm have been combined with the Bootstrap method in this project- Least squaress estimate (OLS), Least absolute deviation estimate (LAD) and Least median estimate (LMS).

1.2 Model based bootstrap regression

This method primarily include following steps:

- Firstly, establish regression model with all samples and estimate the regression coefficients $\widehat{\beta_0}, \widehat{\beta_1}$.
- Then re-sample the random residual and calculate the dependent variables, that is (X_i^*, Y_i^*) in $Y_i^* = \beta_0 + \beta_1 X_i^* + \varepsilon\epsilon_i$, (i = 1, 2, ..., n) is a model-based Bootstrap sample.

1.3 Cases based bootstrap regression

This method primarily include following steps:

- The independent variable x and dependent variable y in correlation model regression are random variables and accord with joint distribution F(x,y). $E(y_i|x_i) = f(x_i) = \beta_0 + \beta_1 x_i$, (i = 1, 2, ..., n), where β_0, β_1 are determined constants independent with X and n is the sample number.
- Randomly select (x_i^*, y_i^*) in original samples.

1.4 Confidence intervals of Regression Coefficients β

The Bootstrap t-method to estimate the Confidence Interval for the Regression coefficient β include the following steps:

- Re-sample a group of Bootstrap samples $x_1^*, x_2^*, ..., x_n^* \equiv X^*$ from the samples. Bootstrap samples are used to calculate $\theta^* = (\widehat{\beta_1}^* - \widehat{\beta}_1)/\{\widehat{\sigma}^*/\sqrt{\sum_{i=1}^n [x_i - \overline{x}]^2}\}$ where $\widehat{\sigma} = \sqrt{(n-2)^{-1}\sum_{i=1}^n [y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]}$
- Repeat the above step for all Bootstrap samples B and sort the obtained values of θ^*
- Approximating $\theta_{1-\alpha/2}$ with $\widehat{\omega}_{\alpha/2} = \theta^*_{[(1-\alpha/2)B]}$, we can get $1 - \alpha/2$ confidence interval for β_1 as
$(\widehat{\beta}_1 - \widehat{\omega}_{\alpha/2}\widehat{\sigma}/\sqrt{\sum_{i=1}^n [x_i - \overline{x}]^2}, \widehat{\beta}_1 + \widehat{\omega}_{\alpha/2}\widehat{\sigma}/\sqrt{\sum_{i=1}^n [x_i - \overline{x}]^2})$

1.5 Estimation methods for Regression Coefficients β

1.5.1 Least Squares Method estimate (OLS)

According to the basic principles of least squares method, the best fit line should make the distance between those points and the straight line minimum. This is same as the least square square minimization that we do normally. Fortuneately this is same for both model based and cases bootstrap regression.

1.5.2 Least Absolute Deviation Regression (LAD)

In this method we minimize the sum of absolute deviation. Thus objective is to choose $\beta 0, \beta 1$ such that $min_{\beta 0, \beta 1} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]$

- Take any two points $a(x_i, y_i), b(x_i, y_i)$ $(1 \leq \text{i} \leq \text{j} \leq \text{n})$ in the n sample points and the coefficients of straight line equation passing a and b points are $\beta_0 = y_j, \beta_1 = (y_j - y_i)/(x_j - x_i)$.
- Let $\beta_{i,j} = (\beta_0, \beta_1), d_{i,j} = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]$
- Take d = $min_{i,j}(d_{i,j})$.
- $\beta_{i,j} = (\beta_0, \beta_1)$ corresponding to d is the LAD regression coefficient.

1.5.3 Least Mean squares regression (LMS)

This method minimizes residual squares median.

- Let $\widehat{d} = \infty, (\widehat{\beta}_0, \widehat{\beta}_1) = (\infty, \infty), (i, j, k) = (1, 1, 1)$
- Reorder $x_i, x_j, x_k, (i, j, k = 1, 2, ..., n) satisfying x_i < x_j < x_k$.
- Calculate$\beta_1 = y_i - y_k / x_i - x_k, \beta_0 = (y_j + y_k - \beta_1(x_j + x_k)$.
- If $d_{i,j,k} = med(y_i - (\beta_0 + \beta_1 x_i))^2, (i = 1, 2, ..., n)$.
- If $d_{i,j,k} < \widehat{d}, let \widehat{d} = d, (\widehat{\beta}_0, \widehat{\beta}_1) = (\beta_0, \beta_1)$
- Repeat 4-7 until (i, j, k) takes all (m, n, q)—(m, n, q = 1, 2, ..., n),
- $(\widehat{\beta}_0, \widehat{\beta}_1)$ is the LAD median regression coefficient.

Chapter 2

Simulation Results

2.1 Figures And Tables (generated from Matlab Code)

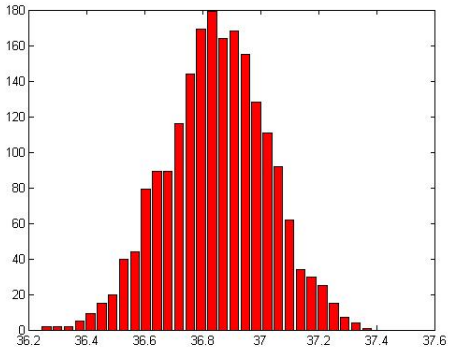

Figure 2.1: Histogram of coefficient estimation by OLS method for model-based Bootstrap model (error is normal distribution)

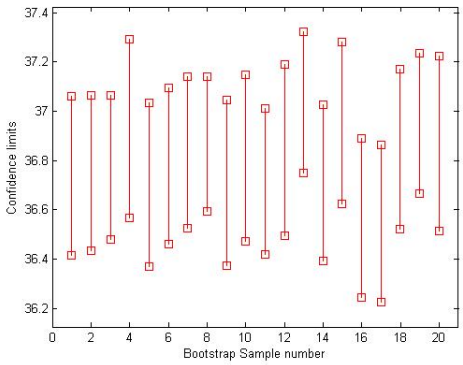

Figure 2.2: Confidence interval of coefficient β_1 estimation by OLS method for model-based Bootstrap model (error is normal distribution and B = 2000)

β_1	statistic	OLS	LAD	LMS
Confidence Interval	number containing β_{10}	825	2000	1683
Lower Limit of Confidence Interval	minimum	36.82132	36.91626	36.62037
	maximum	37.11288	37.29047	37.49834
	mean	36.99008	37.09256	37.03786
	median	36.99151	37.09249	37.03595
	standard deviation	0.045735	0.057817	0.08292
Upper Limit of Confidence Interval	minimum	37.03052	36.85465	36.62037
	maximum	37.32048	37.26026	37.57
	mean	37.1553	37.0529	37.11343
	median	37.15331	37.05355	37.11116
	standard deviation	0.044622	0.062112	0.081627
$\hat{\beta}_1$	minimum	36.92759	36.88545	36.66743
	maximum	37.21668	37.27537	37.53417
	mean	37.07269	37.07273	37.07564
	median	37.07169	37.07306	37.07306
	standard deviation	0.043174	0.059905	0.082197

Figure 2.3: Simulation regression results of Model-based error uniform distribution ($\varepsilon = N(0,1)$, B = 2000, $\alpha = 0.05$)

β_1	Statistic	OLS	LAD	LMS
Confidence Interval	Number containing β_{10}	511	1771	1833
Lower Limit of Confidence Interval	minimum	-1.33659	-1.80676	-3.86731
	maximum	1.121802	1.472863	1.48619
	mean	0.472889	0.308815	-1.07647
	median	0.485913	0.383328	-0.98229
	standard deviation	0.226764	0.463216	0.661598
Upper Limit of Confidence Interval	minimum	0.36316	0.367907	-3.86731
	maximum	2.832343	3.414443	6.435961
	mean	1.503245	1.63573	2.751256
	median	1.481236	1.558748	2.70095
	standard deviation	0.298448	0.451254	0.979869
$\hat{\beta}_1$	minimum	-0.4577	-0.71943	-1.66747
	maximum	1.920027	2.274507	3.337119
	mean	0.988067	0.972273	0.837392
	median	0.978469	0.98243	0.952451
	standard deviation	0.243123	0.431192	0.733461

Figure 2.4: Cases simulation regression results (B = 2000, $\alpha = 0.05$, $\rho = 0.7$)

2.2 Analysis of the Simulation Results

For the Model Based Bootstrap regression:

- Normal Error: Comparing the upper and lower limits for standard deviation of OLS, LAD and LMS confidence intervals, OLS estimate is slightly larger than that of LAD and LMS, and the upper and lower limits for standard deviation of LAD is slightly larger than that of LMS, but the mean and median degree close to original β_1 of LAD and LMS has no absolute relationship; that is, when the error is not large, the robustness advantage of LAD and LMS methods can't be reflected. The mean and median of $\widehat{\beta}_1$ are close, which should take OLS estimate with simple calculation.

- Uniform Error: Comparing the standard deviation of upper and lower limits for confidence interval of OLS, LAD and LMS , estimate of OLS is larger than that of LMS which is larger than that of LAD, but $\widehat{\beta}_1$ standard deviation of LMS is smaller than that of LAD, peak of LMS is more obvious and confidence interval range is the smallest. Thus, when error $\epsilon_i \sim U(-\sqrt{12}, \sqrt{12})$, re-sampling Bootstrap and LMS method can obtain more stable and accurate values.

For the Cases Bootstrap, if we compare the three methods OLS, LAD and LMS $\widehat{\beta}_1$ value of LMS estimate is the most accurate and its standard deviation is the smallest.

So in order to get most accurate linear regression coefficient β_1 estimation, we should adopt the combination of cases Bootstrap re-sampling and LMS estimation to apply cases linear regression analysis.

2.3 Conclusions

This project gives an idea that Bootstrap Method can be employed to estimate the regression coefficients when the error terms are not normal.

Bibliography

[1] Jiehan Zhu (Corresponding author) and Ping Jing, The Analysis of Bootstrap Method in Linear Regression Effect, Journal of Mathematics Research, Vol. 2, No. 4, November 2010

[2] Kesar Singh and Minge Xie, Bootstrap: A Statistical Method

[3] MATLAB, version 7.10.0.499 (R2010a)