The Analysis of Bootstrap Method in Linear Regression Effect

Jiehan Zhu (Corresponding author) & Ping Jing

China University of Mining and Technology, Beijing 100083, China

E-mail: zjiehan@gmail.com

Funded by collage st. innovative projects

Abstract

This paper combines the least squaress estimate, least absolute deviation estimate, least median estimate with Bootstrap method. When the overall error distribution is unknown or it is not the normal distribution, we estimate the regression coefficient and confidence interval of coefficient, and through data simulation, obtain Bootstrap method, which can improve stability of regression coefficient and reduce the length of confidence interval.

Keywords: Bootstrap method, Linear regression, Confidence interval, Least squares estimate, Least absolute deviation estimate, Least median estimate

1. Bootstrap Regression

This paper focuses on linear regression. The traditional regression analysis method assumes that the regression equation is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, (i = 1, 2, \cdots, n)$, where random error $\varepsilon_i \sim N(0, \sigma^2)$. Under the assumption of error normality, the coefficient β can be estimated, while in the significant test of regression, the corresponding distribution of test statistics is obtained.

However, when the error ε is not normal or its distribution is unknown, how to estimate the regression coefficient ,how to estimate the confidence interval of coefficient and how to significantly test the regression equation? The following uses Bootstrap method to solve the above problems. According to the different regression relationships, Bootstrap re-sampling methods can be divided into two types.

1.1 Model-based Bootstrap regression

The independent variable x in correlation model regression is a controllable variable(general variable) and only y is a random variable. Random sampling error is $\varepsilon_i, (i = 1, 2, \cdots, n)$. ε_i in regression equation accords with Gauss-Markov assumption:

$$E(\varepsilon_i) = 0; Var(\varepsilon_i) = \sigma^2; Cov(\varepsilon_i, \varepsilon_j) = 0, (i \neq j) \tag{1}$$

But ε_i is not always a normal distribution. It is be noted that σ^2 is not the variance of residual $e_i = y_i - \hat{y}$. Normalize the residual e_i to obtain the revised residual $r_i = e_i - E(e_i)/\sqrt{Var(e_i)}, (i = 1, 2, \cdots, n)$. In order to better model the actual distribution of residual with the experience distribution, the revised residual can be centralized. Denoted the revised residual after centralizing $\bar{r} = \sum_{i=1}^{n} r_i$ by $\widetilde{r_i} = r_i - \bar{r}$.

Based on model-based re-sampling in linear regression: $x_1, x_2, \cdots, x_n$ are unchanged, i.e. $x_i^* = x_i, (i = 1, 2, \cdots, n)$ and re-sample the regression residuals. Firstly, establish regression model with all samples and estimate the regression coefficients $\hat{\beta_0}, \hat{\beta_1}$.Then re-sample the random residual r_i^* and calculate the dependent variables, that is (X_i^*, Y_i^*) in $y_i^* = \hat{\beta_0} + \hat{\beta_1} x_i + \varepsilon_i^*, (i = 1, 2, \cdots, n)$ is a model-based Bootstrap sample.

1.2 Cases Bootstrap Regression

The independent variable x and dependent variable y in correlation model regression are random variables and accord with joint distribution $F(x, y)$.

$$E(y_i|x_i) = f(x_i) = \beta_0 + \beta_1 x_i, (i = 1, 2, \cdots, n), \tag{2}$$

where β_0, β_1 are determined constants independent with X and n is the sample number. Assume

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, (i = 1, 2, \cdots, n), \tag{3}$$

where ε_i matches Gauss-Markov assumption, that is , matches equations (1), (2) and (3).

Based on cases re-sampling in linear regression: Random select (x_i^*, y_i^*) in original samples, i.e. cases Bootstrap sample.

1.3 Confidence Interval for Bootstrap Regression Coefficient β

Since the error ε_i and the distributions of β_0 and β_1 are unable to determine, it is difficult to get a statistic $\theta(x)$ whose distribution is known. Bootstrap-t method can avoid this difficulty well. The following gives its specific steps, where takes β_1 for example.

1. Re-sample a group of Bootstrap samples $\{x_1^*, x_2^*, \cdots, x_n^*\} \equiv X^*$ from the samples. Bootstrap samples are used to calculate $\theta^* = (\hat{\beta}_1^* - \hat{\beta}_1)/\{\hat{\sigma}^* / \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\}$, where $\hat{\sigma} = \sqrt{(n-2)^{-1}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}$.

2. $\hat{\sigma}^* = \sqrt{(n-2)^{-1}\sum_{i=1}^{n}(y_i^* - (\hat{\beta}_0 + \hat{\beta}_1 x_i^*))^2}$.

3. Repeat step 1 B times and the last statistic values form a data set $\{\theta_j^* | j = 1, 2, \cdots, B\}$ and sort this statistics $\{\theta_j^* | j = 1, 2, \cdots, B\}$ satisfying $\theta_{(1)}^* \le \theta_{(2)}^* \le \cdots \le \theta_{(B)}^*$.

4. Approximating $\theta_{1-\alpha/2}$ with $\hat{\omega}_{\alpha/2} = \theta_{[(1-\alpha/2)B]}^*$, we can get $1-\alpha$ confidence interval of β_1

$$(\hat{\beta}_1 - \hat{\omega}_{\alpha/2}\hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \hat{\beta}_1 + \hat{\omega}_{\alpha/2}\hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}).^{(HiroshiKon,2003)}$$

Thus, even if the overall distribution is uncertain, we can also estimate some statistics and their confidence intervals to solve parameters interval estimation and hypothesis testing problems which are difficult by conventional methods.

2. Estimation Method of Regression Coefficient β

The following describes three common estimation methods of regression coefficient β, and combines them with Bootstrap method.

2.1 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, which can also be expressed as the squares sum of distance minimum, even if the error squares sum $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ is minimal. In the model-based least squares method, assume $\varepsilon_i \sim N(0, \sigma^2)$, thus $y_i \sim N(\beta_0 + \beta_1 x, \sigma^2)$. The regression coefficient

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}, \hat{\beta}_1 = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) / \sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{4}$$

is gotten.

For the cases regression function, from equation (2) we can prove $E_X[E(Y|X)] = E(Y)$ holds. Then

$$E[\{Y - E(Y)\}\{X - E(X)\}] = E_X[X\{E(Y|X) - E(Y)\}]. \tag{5}$$

From equation (5), we have

$$\begin{aligned} \beta_0 &= E(Y) - (Cov(X,Y)/Var(X))E(X) = \overline{y} - \hat{\beta}_1\overline{x}^{(HiroshiKon,2003)}, \\ \beta_1 &= Cov(X,Y)/Var(X) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) / \sum_{i=1}^{n}(x_i - \overline{x})^{2(HiroshiKon,2003)}. \end{aligned} \tag{6}$$

Although coefficients of (6) and (4), obtained by least squares method, are the same, the model-based least squares method is derived under the assumption that $\varepsilon \sim N(0, \sigma^2)$. However, the cases least squares method doesn't assume that $\varepsilon \sim N(0, \sigma^2)$. When the samples are consistent with Gauss-Markov assumption (1), adopting classic OLS can obtain satisfactory results, but if existing extraordinary points or heavy-tailed (Gauss-Markov assumption doesn't hold), the results obtained by OLS is difficult to accept.

Here are two kinds of robust regression estimation methods of the regression coefficient β .

2.2 Least Absolute Deviation Regression (LAD)

The first robust estimation method of regression coefficient β was the least absolute deviation estimation regression presented by the Edgeworth in 1887, whose principle is that the least absolute deviation of regression equation minimizes, that is $\min_{\beta_o,\beta_1}\{\sum_{i=1}^{n}|y_i - (\beta_0 + \beta_1 x_i)|\}$. This paper uses the simplest LAD estimation regression algorithm.

1. Take any two points $a(x_i, y_i), b(x_i, y_i)(1 \le i \le j \le n)$ in the n sample points and the coefficients of straight line equation passing a and b points are $\hat{\beta}_0 = y_j, \hat{\beta}_1 = (y_j - y_i)/(x_j - x_i)$. Let

$$\beta_{i,j} = (\hat{\beta}_0, \hat{\beta}_1), d_{i,j} = \sum_{i=1}^{n}|y_i - (\beta_0 + \beta_1 x_i)|.$$

2. Take $d = \min_{i,j}\{d_{i,j}\}$.

3. $\beta_{i,j} = (\hat{\beta_0}, \hat{\beta_1})$ corresponding to d is the LAD regression coefficient.(*YadolahDodge*,2008)

That LAD regression takes Bootstrap re-sampling method in any case will like sample median number, so the distribution between the coefficient difference by Bootstrap re-sampling and ordinary least absolute deviation regression and difference by regression coefficient obtained by LAD regression and real coefficient isn't relevant, and they are similar only in the case of large samples. Application of smooth Bootstrap can improve estimation accuracy.(*A.C.Davison*,1997)

2.3 Least Median Squares Regression(LMS)

The least median squares regression makes the residual squares median minimum. For linear regression situation, $median(y_i - (\beta_0 + \beta_1 x_i))^2$ minimizes, where using original LMS cured Algorithm. Its specific proof can be found in the book: Algorithms and Complexity for Least Median of Squaress Regression.

1. Let $\hat{d} = \infty, (\hat{\beta_0}, \hat{\beta_1}) = (\infty, \infty), (i, j, k) = (1, 1, 1)$,

2. Reorder $x_i, x_j, x_k, (i, j, k = 1, 2, \cdots, n)$ satisfying $x_i < x_j < x_k$,

3. Calculate $\beta_1 = y_i - y_k / x_i - x_k, \beta_0 = (y_j + y_k - \beta_1(x_j + x_k))$,

4. $d_{i,j,k} = med(y_i - (\beta_0 + \beta_1 x_i))^2, (i = 1, 2, \cdots, n)$,

5. If $d_{i,j,k} < \hat{d}$, let $\hat{d} = d, (\hat{\beta_0}, \hat{\beta_1}) = (\beta_0, \beta_1)$,

6. Repeat 4∼7 until (i, j, k) takes all $\{(m, n, q)|(m, n, q = 1, 2, \cdots, n)\}$,

7. $(\hat{\beta_0}, \hat{\beta_1})$ is the LAD median regression coefficient.(*J.MSteele*,1956)

Table 1 briefly summarizes a few basic properties of three regression coefficient estimate methods, where the estimate variance of $\hat{\beta_1}$ is $Var\hat{\beta_1} = \hat{\sigma}^2 / \sum_{i=1}^{n}(x_i - \overline{x})^2, \hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 / n - 2$.

3. Data Simulation

In this section, data simulation uses computer-generated pseudo-random number technique, where draw out data from known distributions to regression analyze by Bootstrap method, and then according to regression results, test fit situation of Bootstrap method results and assumption distribution.

3.1 Model-based Bootstrap Regression Analysis

According to the definition of model-based regression model, we establish a known distribution

$$y_i = \beta_{00} + \beta_{10} x_i + \varepsilon_i, \tag{7}$$

where ε_i matches Gauss-Markov assumptions. Without loss of generality, we assume $\beta_{00} = 0, \beta_{10} = 37, Var(\varepsilon_i) = 1, x \sim U(0, 10)$ and samples number $n = 14$ in this experiment.

In order to explain the effects of Bootstrap method in condition that error is normally distributed and is not normal distribution. There particularly takes distributions of two common errors: normal distribution and uniform distribution, i.e. do model-based Bootstrap regression analysis under $\varepsilon_i \sim N(0, 1)$ and $\varepsilon_i \sim U(-\sqrt{12}, \sqrt{12})$.

For a more intuitive analysis for Bootstrap method estimate results of model-based regression model, the distribution of regression coefficients $\hat{\beta_1}$ and $1-\alpha$ confidence interval for the former 20 groups of $\hat{\beta_1}$ of 2000 groups of Bootstrap samples by OLS estimate method, respectively.

Since $y \sim N(\beta_0 + \beta_1 x, \sigma^2), T \sim t(n-2)$, graphic 2 is close to the normal distribution. 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 β_{10} 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 $\hat{\beta_1}$ are close, which should take OLS estimate with simple calculation.

Uniform distribution is more discrete than normal distribution. So the standard deviations of all the statistics in Table 3 increase than that of Table 2, the maximum and minimum values of upper and lower limit for confidence interval in Table 2correspondingly increase, mean and median decrease; that is, confidence interval fluctuates larger and the size of interval decrease. Mean of $\hat{\beta}$ is close to 37 and the median deviation is relatively large. It is seen that using LAD and LMS can obtain more accurate value.

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 $\hat{\beta_1}$ 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 $\varepsilon_i \sim U(-\sqrt{12}, \sqrt{12})$,

re-sampling Bootstrap and LMS method can obtain more stable and accurate values.

3.2 Case Bootstrap Regression Analysis

By computer we randomly select original samples from two-dimensional normal distribution $N(0, 1, 1, 1, \rho)$ to do cases Bootstrap regression analysis simulation. According to the properties of two-dimensional normal distribution, we have $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$,

$$E(Y_i|X_i) = \mu_Y + \rho X_i. \tag{8}$$

Under the condition that error $\mu_Y = \beta_{00}, \rho = \beta_{10}$, extract a set of original samples from distribution $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ complying with cases regression equation (3), where $\beta = (\beta_{00}, \beta_{10})^{(HiroshiKon,2003)}$.

For the confidence interval is almost symmetrical about 0, only the confidence interval limit statistics are given here. Table 9 shows that the tandard deviation for absolute value of $\hat{\beta}_1$ s in cases Bootstrap regression is not larger than that of model-based Bootstrap regression, but relative $\hat{\beta}_1$ value is larger than that of model-based Bootstrap regression. The reason is that x, y values in $N(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ is small and rounding error in calculation relatively increases, leading that the regression results are not satisfactory, which is the cause that LAD estimation results in Table 4 are not accurate than that of OLS. In the process of LAD regression iteration, every step has some rounding errors, but OLS need not iteration and rounding error should not be influenced greatly.

In the three methods, $\hat{\beta}_1$ value of LMS estimate is most accuration and its standard deviation is the smallest. So in order to get most accurate linear regression coefficient $\hat{\beta}_1$ estimation, we should adopt the combination of cases Bootstrap re-sampling and LMS estimation to apply cases linear regression analysis.

Acknowledgements

Dear Prof. Ping Jing, Words could hardly express my most sincere and heartfelt thanks for your never-ending support, encouragement, guidance, wisdom, experience, and insight into my essay from start to finish. Thank you so so much!

References

A. C. Davison & D. V. Hinkely. (1997). *Bootstrap methos and their application*. Cambridge: Cambridge University Press(312).

George Casella & Roger L. Berger. (2002). *Statistical Inference*. Belmont: Duxbury Press(448).

Hiroshi Kon & Masaaki Hitoshi. (2003). Statistics Calculator - a new method of probability calculation, *statistics Frontiers of Science 11*, Tokyo: Iwanami Shoten, 41-59.

J.M Steele & W. L. Steiger. (1956). Algorithms And Complexity for Least Median of Squaress Regression . *Discrete Applied Mathematics*, 14:93-100.

Yadolah Dodge. (2008). *The concise encyclopedia of statistics*. Springer(299-301).

Table 1. Comparsion between regression and least squares regression

	OLS	LAD	LMS
Estimation principle	$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$	$\sum_{i=1}^{n}\|y_i - (a + bx_i)\|$	$median(y_i - (\beta_0 + \beta_1 x_i))^2$
Contamination data	irresistible pollution	resistible light pollution	resistible grave pollution
Breakdown point	0%	0%	50%
Regression coefficient	a unique solution	multiple solutions possibly	a unique solution

Table 2. Simulation regression results of Model-based error normal distribution

($\varepsilon_i \sim N(0,1), B = 2000, \alpha = 0.05$)

β_1	statistic	OLS	LAD	LMS
confidence interval	number containing β_{10}	1964	1939	1976
lower limit of confidence interval	minimum	36.94	35.23	35.26
	maximum	37.79	36.86	37.24
	mean	37.29	36.35	36.61
	median	37.29	36.37	36.61
	standard deviation	0.007844	0.1870	0.1624
limit of confidence interval	minimum	36.08	36.71	36.75
	maximum	37.13	38.38	38.63
	mean	36.80	37.25	37.44
	median	336.80	37.23	37.43
	standard deviation	0.1017	0.1640	0.1406
$\hat{\beta}_1$	minimum	36.77	36.46	36.53
	maximum	37.43	37.11.	37.47
	mean	37.04	36.80	37.03
	median	37.04	36.79	37.01
	standard deviation	0.07261	0.06818	0.1062

Table 3. Simulation regression results of Model-based error uniform distribution

($\varepsilon_i \sim U(-\sqrt{12}, \sqrt{1}12), B = 2000, \alpha = 0.05$)

β_1	statistic	OLS	LAD	LMS
confidence interval	number containing β_{10}	1990	1977	1995
lower limit of confidence interval	minimum	36.993	36.98	36.87
	maximum	37.085	37.21	37.04
	mean	37.029	37.07	36.95
	median	37.029	37.07	36.95
	standard deviation	0.01088	0.02321	0.01605
higher limit of confidence interval	minimum	36.900	37.78	36,96
	maximum	37.015	37.03	37.12
	mean	36.961	36.92	37.04
	median	36.961	36.92	37.04
	standard deviation	0.012646	0.02437	0.01411
$\hat{\beta}_1$	minimum	36.97	36.94	36.94
	maximum	37.04	37.06	37.06
	mean	36.99	36.99	36.99
	median	37.00	36.99	36.99
	standard deviation	0.008011	0.01203	0.009643

Table 4. Case simulation regression results

($B = 2000, \alpha = 0.05, \rho = 0.7$)

β_1	statistic	OLS	LAD	LMS
confidence interval	number containing β_{10}	1982	1934	1977
limit of confidence interval	minimum	0.0000	0.0000	0.0000
	maximum	0.2772	0.7562	0.3789
	mean	0.0001386	0.001302	0.0001894
	median	0.0000	0.0000	0.0000
	standard deviation	0.006198	0.02932	0.008471
$\hat{\beta}_1$	minimum	0.4059	0.9689	0.9368
	maximum	0.9517	0.2201	0.3942
	mean	0.7190	0.6693	0.6926
	median	0.7194	0.6492	0.6937
	standard deviation	0.0737	0.7488	0.07223

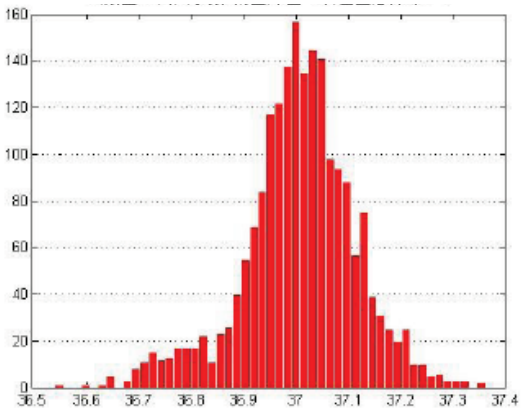

Figure 1. Histogram of coefficient β_1 estimation by OLS method for model-based Bootstrap model (error is normal distribution)

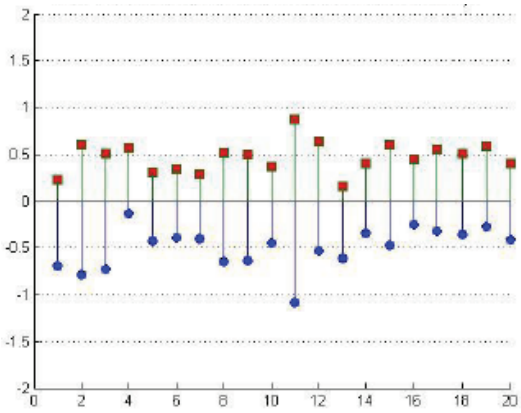

Figure 2. Confidence interval of coefficient β_1 estimation by OLS method for model-based Bootstrap model(error is normal distribution and $B = 2000$)