INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Bootstrapping Two-step Least Squares Approach for Solving Heteroscedasticity Problem in Linear Regression Analysis

Abueme Onyinye Maureen
Student, Department of Statistics
Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
Obiora-Ilouno Happiness
Associate Professor, Department of Statistics,
Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
Ogbonna Uchenna Austine
Student, Department of Statistics,
Nnamdi Azikiwe University, Awka, Anambra State, Nigeria

Abstract:
In this study, the problem of heteroscedasticity which is referred to as unequal variability in linear regression analysis is considered. Bootstrapping Two Step Least Squares (BWLS) for controlling heteroscedasticity in linear regression analysis was proposed. This method was compared with other existing methods (Ordinary Least Squares, Weighted Least Squares, Two Step Least Squares, Bootstrap Weighted Least) based on the Root Mean Square Error (RMSE) of the regression coefficient and the Euclidean norm was used to measure the accuracy across all coefficients. Simulated data was used in the comparison. The results show that the proposed bootstrap method performed better than the other methods in performing regression analysis in the presence of heteroscedasticity across all the sample sizes and at various levels of heteroscedasticity considered from the analysis. Real life data was used to demonstrate the results, which corresponds with result obtain from the simulation studies.

Keywords: *Heteroscedasticity, weighted least squares, two step least squares, bootstrap weighted least, bootstrapping two step least squares, regression analysis*

1. Introduction

Regression analysis is a statistical method that takes a set of data points and creates a model that can then be used to better understand the relationship between the variables or make predictions about future results. Regression was published by Legendre in 1805 and by Guass in 1809. The term 'regression' was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. Regression analysis is a statistical technique for investigating and modelling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

The most popular method of fitting a regression model is least square method. The least squares method is a form of mathematical regression analysis that finds the line of best fit for a dataset, providing a visual demonstration of the relationship between the data points. Each point of data is representative of the relationship between a known independent variable and an unknown dependent variable, Douglas et al (2012).

Suppose the data consists of n observations $\{Y_i, X_i\}_{i=1}^{n}$. Each observation includes a scalar response yi and a vector of p predictors (or regressors) X_i.

$y_i = X_i^T \beta + \varepsilon_i$,

where β is a p×1 vector of unknown parameters;ε_i's are unobserved scalar random variables (errors) which account for the discrepancy between the actually observed responses y_i and the 'predicted outcomes'$X_i^T\beta$; and T denotes matrix transpose, so that $X^T\beta$ is the dot product between the vectors x and β.

One of the basic assumptions in the application of the ordinary least squares method is that the error terms have constant variance (homogeneity of variance). Several cases arise in practice which violates this assumption, that is, a situation of heteroscedasticity which has serious consequence for the ordinary least squares estimator, thus the amount and reliability of the information about the value of the dependent variable for each level of the independent variables may differ. As a result, the Ordinary Least Squares estimator can no longer be Best Linear Unbiased Estimate (BLUE).

This work examines the effect of bootstrapping in working with heteroscedasticity. Bootstrapping is a re-sampling method that uses the original data to estimate a parameter or estimate the standard error of an estimate (Chernick 1999).

The bootstrap was introduced in 1979 by Bradley Efron at Stanford University. Efron first gave an extended account of the term in An Introduction to the Bootstrap with Efron and R.J.Tibshirani (1993) where theyexplained, that the use of the term'bootstrap'was derived from the phase to pull oneself up by one's own bootstrap. To bootstrap means to make use of existing resources to raise oneself toa new situationby makinguseofwhatisalready present. The simplest bootstrap method involves taking the original data set of n size, and obtaining a sampling from the sample to form a new sample (called a 'resample' or bootstrap sample) that is also of size n. The bootstrap sample is taken from the original by sampling with replacement. This process is repeated a large number of times (typically 1,000 to 10,000 times), and for each of these bootstrap samples, the mean is computed which is called bootstrap estimates.

2. Literature Review

Guan (2003) proved that bootstrapping is an alternative to obtain standard errors forestimated parameters, (when he compared results from Monte Carlo simulations with thosefrom parametric models) bootstrapped standard errors tend to be more conservatives than theparametric estimated coefficients. He concluded by stating that the number of repetitions andsample size both play important roles in the bootstrap method.

Obiora-Iluonu et al (2016) compared the linear discriminant method with proposed bootstrapmethod to identify which of the method performs better based on their error rate, reportedthat the bootstrap methods produced smaller error rate indicating that the proposed bootstrapalgorithm yield a better reduced error rate.

Wang et al (2006) discussed how bootstrapping is used to approximate the standard error ofcertain estimators that were then used to create a particular linear regression model.

According to Karlis (2004), the bootstrap standard errors of the TL and OL coefficients aresubstantially larger than the estimated asymptotic OLS standard errors, because of theinadequacy of the bootstrap in small samples. The confidence intervals based on the bootstrapstandard errors are very similar to the percentile intervals of the TL and OL coefficients.However, the confidence intervals based on the OLS standard errors are quite different frompercentiles and confidence intervals based on the bootstrap standard errors.

3. Materials and Methods

The Bootstrap Two-Step Least Square (BTSLS) is proposed. This procedure involves bootstrapping the already existing Two-Step Least Square. This method was compared with other existing methods; Ordinary Least Squares (OLS), Weighted Least Squares (WLS), Two Step Least Squares (TSLS), Bootstrap Weighted Least Squares (BWLS)

3.1. Bootstrap Two-Step Least Square (BTSLS)

This method involves adding a bootstrap smoothing to the two-step least squares.

Let $Z_i = (Y_i, X\,ji)'$ be the n sample size for the resampling, Where $Y_i = (y_1, y_2, ..., yn)'$ is a column vector of dependent variable and $X_{ji} = (x_{ji}, x_{j2}, ..., xjn)'$ is the matrix of dimension n x p for the independent variable where $j = 1, 2, ..., k.\ \ i = 1, 2, ..., n$

1. Draw a sample $(z_1^{(b)} z_2^{(b)}, \dots, z_n^{(b)})$ with replacement from the original sample, with $\frac{1}{n}$ probability of sampling eachZ_i, and label each element as $Z_i^{(b)} = (Y_i^{(b)}, X_{ji}^{(b)})$
2. On each bootstrap sample, perform two-step least square use in section 3.4 and record your regression coefficients.
3. Find the average of the all the bootstrap estimates of each coefficient, which is bootstrap estimate for the BTSLS

$$\hat{\beta}^{(BTSLS)} = \sum_{b=1}^{1000} \hat{\beta}^{(BTSLS)} / 1000 = \overline{\hat{\beta}}^{(BTSLS)}$$

Data will be simulated using the R statistical package. The sample size will be varied as 25, 50, 100. The severity of heteroscedasticity will be varied as low, high and very high. The Root Mean Square Error (RMSE) of the regression coefficient and the Euclidean norm will be used to measure the accuracy across all coefficients.

3.2. Root Mean Square Error (RMSE)

The root mean square error measures the accuracy of the methods in estimating the slope parameter. It is a function of both the bias and the variance. The RMSE is given by;

$$RMSE = (var(b_1) - bias^2)^{\frac{1}{2}}$$

3.3. Euclidean Norm (Enorm)

Enorm is the distance between coordinators

$x = (x_1, x_2, \dots\dots\dots, x_n) \in R^n$ by

$\|x\| = \sqrt{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}$

x is called a point on a vector and $x_1 x_2 \dots\dots x_n$ are called the coordinates of x. n is called the dimension of space, R (real numbers).

Methods	Coefficients			RMSE			ENORM
OLS	$b_0 = 1.18$	$b_1 = 5.81$	$b_2 = 7.01$	0.87	0.94	0.01	1.29
	$b_0 = 2.26$	$b_1 = 4.88$	$b_2 = 7.00$	0.89	0.86	0.01	1.23
	$b_0 = 2.33$	$b_1 = 4.85$	$b_2 = 7.00$	0.48	0.46	0.00	0.66
WLS	$b_0 = 1.23$	$b_1 = 5.81$	$b_2 = 7.01$	0.90	0.97	0.01	1.32
	$b_0 = 2.26$	$b_1 = 4.88$	$b_2 = 7.00$	0.89	0.86	0.01	1.24
	$b_0 = 2.31$	$b_1 = 4.84$	$b_2 = 7.00$	0.47	0.45	0.00	0.66
BWLS	$b_0 = 1.23$	$b_1 = 5.78$	$b_2 = 7.01$	0.83	0.87	0.01	1.20
	$b_0 = 2.17$	$b_1 = 4.99$	$b_2 = 7.00$	0.78	0.78	0.01	1.10
	$b_0 = 2.38$	$b_1 = 4.84$	$b_2 = 6.99$	0.42	0.42	0.00	0.60
TSWLS	$b_0 = 0.86$	$b_1 = 5.93$	$b_2 = 7.01$	1.09	0.91	0.04	1.42
	$b_0 = 2.00$	$b_1 = 5.19$	$b_2 = 7.00$	0.38	0.51	0.00	0.64
	$b_0 = 1.99$	$b_1 = 5.05$	$b_2 = 7.00$	0.17	0.24	0.00	0.31
BTSLS	$b_0 = 1.04$	$b_1 = 6.02$	$b_2 = 7.00$	0.53	0.60	0.01	0.80
	$b_0 = 2.03$	$b_1 = 5.13$	$b_2 = 6.99$	0.32	0.40	0.00	0.51
	$b_0 = 2.01$	$b_1 = 4.97$	$b_2 = 7.00$	0.16	0.21	0.00	0.27

Table 1: Results for Sample Size 25, 50, 100 (Level of Heteroscedasticity Is Low)

When the sample size is 25, 50 and 100 with low heteroscedasticity, the BTSLS outperformed the other methods with the least ENORM and smallest RMSE.

Methods	Coefficients			RMSE			ENORM
OLS	$b_0 = 4.21$	$b_1 = 3.70$	$b_2 = 6.99$	1.25	1.51	1.40	2.41
	$b_0 = 3.76$	$b_1 = 3.66$	$b_2 = 7.00$	1.23	0.97	0.94	1.83
	$b_0 = 3.35$	$b_1 = 4.10$	$b_2 = 4.10$	1.23	0.97	0.94	1.83
WLS	$b_0 = 3.51$	$b_1 = 3.95$	$b_2 = 6.99$	1.22	1.55	1.40	2.42
	$b_0 = 3.78$	$b_1 = 3.78$	$b_2 = 7.00$	1.84	1.48	0.91	2.53
	$b_0 = 3.44$	$b_1 = 4.04$	$b_2 = 6.99$	1.84	1.48	0.91	2.53
BWLS	$b_0 = 3.72$	$b_1 = 4.39$	$b_2 = 6.99$	1.17	1.44	1.27	2.25
	$b_0 = 3.59$	$b_1 = 3.83$	$b_2 = 7.00$	1.11	0.90	0.85	1.67
	$b_0 = 3.26$	$b_1 = 4.14$	$b_2 = 6.99$	1.11	0.90	0.85	1.67
TSWLS	$b_0 = 2.15$	$b_1 = 5.26$	$b_2 = 6.99$	1.48	1.69	1.55	2.73
	$b_0 = 1.99$	$b_1 = 4.68$	$b_2 = 6.99$	0.46	1.63	0.93	1.93
	$b_0 = 2.20$	$b_1 = 4.51$	$b_2 = 6.99$	0.46	1.63	0.93	1.93
BTSLS	$b_0 = 1.19$	$b_1 = 6.15$	$b_2 = 7.09$	0.47	0.91	0.73	1.26
	$b_0 = 1.43$	$b_1 = 5.16$	$b_2 = 7.00$	0.28	0.47	0.40	0.69
	$b_0 = 2.17$	$b_1 = 4.55$	$b_2 = 7.00$	0.28	0.47	0.40	0.69

Table2: Results for Sample Size 25, 50, and 100 (Level of Heteroscedasticity Is High)

When the sample size is 25, 50 and 100 with high heteroscedasticity, the BTSLS outperformed the other methods with the least ENORM and smallest RMSE.

Methods	Coefficients			RMSE			ENORM
OLS	$b_0 = -1520.61$	$b_1 = 870.26$	$b_2 = 32.47$	79.5 4	128.48	63.68	163.99
	$b_0 = 296.84$	$b_1 = -267.49$	$b_2 = 6.88$	79.54	128.48	63.68	163.99
	$b_0 = 221.31$	$b_1 = -160.28$	$b_2 = 6.11$	15.40	34.06	17.66	41.34
WLS	$b_0 = -1511.41$	$b_1 = 866.59$	$b_2 = 32.28$	29.78	24.47	65.35	75.87
	$b_0 = 233.42$	$b_1 = -229.16$	$b_2 = 6.79$	29.78	24.47	65.35	75.87
	$b_0 = 208.39$	$b_1 = -164.18$	$b_2 = 6.43$	24.15	33.98	19.55	46.05
BWLS	$b_0 = -1419.47$	$b_1 = 792.94$	$b_2 = 31.90$	64.45	106.57	58.12	137.44
	$b_0 = 260.92$	$b_1 = -240.46$	$b_2 = 6.88$	64.45	106.57	58.12	137.44
	$b_0 = 202.80$	$b_1 = -145.23$	$b_2 = 5.99$	10.43	24.51	14.07	30.13
TSWLS	$b_0 = -12.50$	$b_1 = 14.61$	$b_2 = 7.22$	6.96	11.09	3.84	13.65
	$b_0 = 3.06$	$b_1 = 3.58$	$b_2 = 6.99$	6.96	11.09	3.84	13.65
	$b_0 = 3.61$	$b_1 = 2.90$	$b_2 = 6.99$	8.40	43.10	31.05	53.78
BTSLS	$b_0 = -246.02$	$b_1 = 162.93$	$b_2 = 386.49$	2.66	5.58	4.77	7.8
	$b_0 = 4.56$	$b_1 = 1.70$	$b_2 = 7.00$	2.66	5.58	4.77	7.8
	$b_0 = 4.36$	$b_1 = 1.84$	$b_2 = 7.00$	0.63	3.08	1.37	3.43

Table3: Results for Sample Size 25, 50, and 100 (Level of Heteroscedasticity Is Very High)

When the sample size is 25, 50 and 100 with low heteroscedasticity, the BTSLS outperformed the other methods with the least ENORM and smallest RMSE.

Methods	Models	RMSE
OLS	$Y = 4364.726 - 4.296046X_1 + 2.599252\ X_2$	493.3117
WLS	$Y = 4647.777 - 5.251522\ X_1 + 2.641275\ X_2$	493.7551
BWLS	$Y = 4332.571 - 4.235595\ X_1 + 2.61899\ X_2$	493.3741
TSLS	$Y = 1207.564 + 4.864112\ X_1 + 3.023242\ X_2$	532.0568
BTSLS	$Y = 2279.356 + 1.183883X_1 + 3.188121X_2$	467.6002

Table4: Results for the Application of the Methods on the Real Data Set

After the application of the methods to the real data, the results correspond to the results from the simulations in the sense that the bootstrap two- step produced the least RMSE than any other method in the presence of heteroscedasticity.

4. Conclusion

Based on the summary of findings from the simulations and the real-life data analysis, it can be concluded that the BTSLS method is an improved method for performing regression analysis in the presence of heteroscedasticity when the sample size is relatively small and large.

5. References

i. ChernickM. R. (1999) Bootstrap methods, A practitioner`sGuide, Wiley, New York.
ii. Douglas C. Montgomery, Elizabeth A, Peck and Geoffery G. Vining. 'Linear Regression Analysis', 5th Edition. Wiley, New York.
iii. EfronB.(1983)'Estimatingtheerrorrateofapredictionrule:Improvementoncross-validation'.JournalofAmericanstatisticalassociation78,316-330.
iv. Efron B and Tibshirani R. J. (1993). 'An introduction to the Bootstrap'. Chapman and Hall, now to K N 4.
v. Guan, W. (2003). 'From the help desk: Bootstrapped standard errors'. Stata Journal 3(1):71–80.
vi. Guass C. F. (1809) 'Theoria Motus Corporum Coelesticim. Perthes, Hamburg. Translation reprinted as Theory of the Motions of the Heavenly Bodies Moving about the sun in sections Dover', New York.
vii. Karlis, D. (2004). 'An introduction to bootstrap methods'. 171h conference of Greek Statistical Society, Greece.
viii. Obiora-Ilouno, H.O., B.N. Chidinma and C.A. Uzuke, (2016). 'Bootstrap method for estimating error rate in linear discriminant analysis (LDA)'. .J. Nat. Sci. Res., 6: 80-85.
ix. Wang, J., Carpenter, J.R., &Kepler, M.A. (2006). 'Using SAS to conduct nonparametric residual bootstrap multilevel modelling with a small number of groups'. Computer Methods and Programs in Biomedicine, 82, 130-143