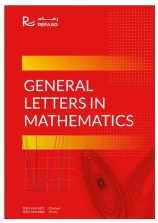
Available online at www.refaad.com
Gen.Lett. Math., 12 (2) (2022), 96-104
Research Article

General Letters in Mathematics (GLM)
Journal Homepage: https://www.refaad.com/Journal/Index/1
ISSN: 2519-9277 (Online) 2519-9269 (Print)



Estimating Regression Coefficients using Bootstrap with application to Covid-19 Data

Rojeen Taha Ahmad[a,*], Shelan Saied Ismaeel[a]

[a]*Department of Mathematics, Faculty of Science, University of Zakho, Zakho, Kurdistan Region, Iraq*

Abstract

The linear regression model is often used by researchers and data analysts for predictive, descriptive, and inferential purposes. When working with empirical data, this model is based on a set of assumptions that are not always satisfied. In this situation, using more complicated regression algorithms that do not strictly rely on the same assumptions might be one answer. Nevertheless, transformations provide a simpler technique for improving the validity of model assumptions and allow the user to continue using the well-known model of linear regression. The main objective of this project is to provide a transformation for the linear model's response and predictor variables, as well as parameter estimation methods before the transformation and after the transformation. The bootstrap approach has been effectively used for many statistical estimates and inference issues, according to the paper.

Keywords: Power transformations, Bootstrap technique, MM-estimator, GM6-Estimator, trafo package.
2010 MSC: 62F40, 62J02, 62J05.

1. Introduction

The linear regression is one of the most often used statistical approaches for studying the relationship between two or more random variables. A number of assumptions must be fulfilled in order for this model to be used properly. These assumptions include linearity, homoscedasticity, and normality, which are all connected to the functional form and error terms. Nevertheless, in practice, these assumptions are not always achieved. As a result, the practitioner must consider how to proceed with the analysis in such a situation. One option is to do the study while disregarding the model assumption violations, however this is not encouraged because it would most likely result in misleading findings. Another option is to employ more complicated approaches like generalized linear regression or methods of non-parametric, which may better match the data and the issue. The use of suitable transformations is a third method which is also the topic of this study. The main goal of this project is to create a transformation that transforms the response and predictor variables of a linear model, as well as parameter estimate methods before the transformation and after the transformation.

*Corresponding author
Email addresses: rojen.ahmad@staff.uoz.edu.krd. (Rojeen Taha Ahmad), Shelan.ismaeel@uoz.edu.krd. (Shelan Saied Ismaeel)

doi:10.31559/glm2022.12.2.6

Received 27 Apr 2022 : Revised : 21 Jun 2022 Accepted: 5 Jul 2022

The trafo package can be used to estimate, choose, and compare different transformation families. The following are the transformation families contained in the package: Bickel-Doksum [3] , Box-Cox [24] Dual, Glog [7] , Gpower, Log, Log-shift opt, Manly, Modulus, Neglog [22] , Reciprocal and Yeo-Johnson. The package makes it easier to compare linear models with transformed and untransformed dependent variables, as well as linear models with different transformations applied to the dependent variable. Moreover, to estimate the optimal transformation parameter, the package uses maximum likelihood methods, skewness and divergence minimization.
The method of Ordinary Least Square (OLS) is the most often used strategy for estimating the parameters of linear regression because of its general acceptance, neat statistical features, and computing simplicity. Under assumptions normality of regression errors, the OLS estimates offer a lot of appealing qualities. When there are outliers in a dataset, however, OLS estimations become ineffective. Robust methods that are known to be resistant to outliers can be used to solve the problem of outliers on parameter estimations. In the literature [12],[14],[23] and [25], there are several resilient estimating methods such as M, MM, LMS, and LTS. According to [11] , Schweppes proposed a new robust method known as a bounded influence. (GM-estimator) Generalized M-estimator as a remedy for M-estimator sensitivity to high leverage points [1] and [11]. In the literature [1] and [23], several different forms of GM-estimators have been suggested. These methods, on the other hand, have achieved a modest breakdown point of $1/k$, where k is the number of regression coefficients including the intercept [20]. Multi-stage GM estimators were created as a remedial approach. GM6, which was introduced by [6], is one of the most common varieties of multi-stage GM-estimator. In the GM6 algorithm [15] , the least trimmed of squares is used as an initial estimate. The GM6 estimator's initial d–weight function is stated in terms of robust Mahalanobis distance (RMD), which was calculated using robust location and scatter estimators derived from minimal volume ellipsoid (MVE) [17]. For the linear model, we used transformed data for the response variable and the predictor variable, and then used robust MM and robust GM6 to estimate parameters in simple and multiple regression.

Because the bootstrap re-sampling technique is with replacement, the bootstrap samples may contain more outliers than the original sample [16]. As a result, the confidence intervals and variance estimates are impacted, resulting in bootstrap distribution breakdown. To deal with possible outliers, we might use a robust estimator, however, this may not be enough because robust estimation is only supposed to perform well up to a specific number of outliers.
The remainder of the paper is structured as follows. Section 2 contains the transformations and parameter estimation methods. Section 3 mentions the application of using real data part 4 includes some concluding remarks.

2. Main body

2.1. Transformations method:

The equation summarizes and describes the relationship between a continuous dependent variable y and several covariates x (either continuous or discrete) defined by $y_i = \beta_i x_i^T + e_i$,with $i = 1,\ldots,n$, This is also known as the model of linear regression, and it is made up of a deterministic and random components that are based on several assumptions such as, homoscedasticity, linearity and normality. For fulfilling the model assumptions, many ways have been presented. We focus on parameter estimation and transformation, such as the use of non-linear transformations of the dependent and independent variables. The transformations for predictor variable (Square Root, Inverse, Logarithmic) and the transformations for response variable implemented in the trafo package primarily help in the achievement of normality. Nevertheless, the majority of them correct other assumptions at the same time.
If the response variable contains negative values, the values are moved by a deterministic shift a by default in trafo package, so that $y + a > 0$. In one case, a square root transformation with deterministic shift is given [2] . For this estimation, the trafo package includes several methodologies. Each estimating

method's advantage is determined by the research analysis and underlying data. We showed how the trafo package makes it simple for users to determine which transformations are suitable for satisfying the model assumptions concerning linearity, normality and homoscedasticity. trafo is the only R package, to the best of our knowledge that supports this decision process.

Table 1: Diagnostic checks provided in the package trafo

Assumption	Diagnostic check
1) Normality	Skewness and kurtosis Shapiro-Wilk test Quantile-quantile plot Histograms
2) Homoscedasticity	Breusch-Pagan test Residuals vs. fitted plot Scale-location
3) Linearity	Scatter plots between y and x Observed vs. fitted plot

The maximum likelihood estimate approach discovers the set of transformation parameter values that maximize the likelihood function of the dataset under the given transformation [4] . This is a standard approach that is used in several of the R packages described above [10] and [21]. Because the user can only determine whether the transformation is beneficial by checking the above-mentioned assumptions, the trafo package includes a wide range of the diagnostic tests [5] and [19]. The fast check, which determines if a transformation is helpful, uses a smaller selection. Table 1 shows the implemented diagnostic checks for the untransformed and transformed models, as well as two different transformed models, and shows which diagnostics are conducted in the fast check. Additionally, graphs such as the Cook's distance plot by the residuals vs leverage plot are provided to help in the detection of outliers.

2.2. *Ordinary Least Square estimator*

In terms of the observations, the model of multiple linear regression can be written as matrix notation $y = X\beta + e$, where y is response values, X is predictor variables, β is the parameters, and e is the error terms. The goal of regression analysis is to find unknown parameter estimates. The least squares criteria is used to obtain the best estimate of β's using the OLS, which minimizes the sum of squared distances of all points from the actual observation to the regression surface [9].

2.3. *The MM-estimator*

Yohai (1987) developed a special type of M-estimation called MM estimation [25]. High breakdown value estimation and efficient estimation are combined in MM-estimation. Under normal error, Yohai's MM estimator was the first with a high breakdown point and high efficiency [18] . The MM-estimators are divided into three stage procedures. Calculating an S-estimate with influence function is the first stage $\rho(x) = 3\left(\frac{x}{c}\right)^2 - 3\left(\frac{x}{c}\right)^4 + 3\left(\frac{x}{c}\right)^6$ if $|x| \leqslant c$, otherwise $\rho(x) = 1$. The tuning constant, c, has been set at 1.548. The MM parameters that produce the minimum value of $\sum_{i=1}^{n} \rho(\frac{y_i - x_i'\widehat{\beta}_{MM}}{\widehat{\sigma}_0})$ are calculated in the second stage where $\rho(x)$ is the function of influence used in the first stage, with a tuning constant of 4.687, and $\widehat{\sigma}_0$ is the standard deviation of the residuals (estimate of scale form the first step). The final stage computes MM estimate of the scale as the answer to

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho\left(\frac{y_i - x_i'\widehat{\beta}}{s}\right) = 0.5$$

2.4. The GM6-Estimator

The GM6 estimator was proposed by [6] and is defined as a solution of normal equations as follows:

$$\sum_{i=1}^{n} \pi_i \psi \left\{ \frac{y_i - x_i' \widehat{\beta}}{\widehat{\sigma} \pi_i} \right\} x_i = 0$$

Where
$\psi = \acute{\rho}$: weight function, π_i, $i = 1, 2, \ldots, n$: the i^{th} initial weight element of the diagonal matrix W, $\hat{\sigma}$: the scale estimate, $\widehat{\beta}$: the vector of parameters estimates. **Stage 1:** Using the Least Trimmed Squares (LTS) estimator, calculate the residuals (r_i). **Stage2:** Calculate the estimated scale (σ) of residuals,

$$s = (1.4826) * \left(1 + \frac{5}{(n-p-1)}\right) * (median \ (\ |r_i|),$$

where r_i is obtained from Stage 1.
Stage 3: Calculate standardized residuals(e_i), *where*, $e_i = \frac{r_i}{s}$.
Stage 4: Compute initial weight, denoted as π_i, *where*, $\pi_i = min \left\{ 1, \frac{\chi^2_{0.95,k}}{RMD(MV\ E)} \right\}$, where RMD denoted as Robust Mahalanobis Distance, MVE is the Minimum Volume Ellipsoid and k is the number of predictors in the model.
Stage 5: Calculate the function of bounded influence, $t_i = \frac{e_i}{\pi_i}$.
Stage 6: To get the GM6 estimations, calculate one-step Newton Raphson.

2.5. Bootstrap Techniques

By using the resampling technique, the Bootstrap is a general nonparametric approach developed by Efron (1979) that can be used to build a statistical inference and sampling distribution for estimators. Bootstrapping technique uses sample data as population from which repeated samples are drawn. Furthermore, because of difficulties to find distribution of the some complex estimators of robust, bootstrapping can be used as an alternative for the method for computing standard errors of desired estimates. Because bootstrap does not need distributional assumptions, it can yield more accurate inferences when the data is not well behaved or sample size is small see, for more information [8] and [13]. **Fixed-X Bootstrapping** The fixed-X bootstrapping procedure can be used to compute the bootstrapping standard errors for regression coefficients if an independent variable is fixed. The procedure is given as follows [23]:
Stage 1: Let $(\widehat{\beta})$ be an estimate of regression coefficients based on a random sample of observations $(x_1, x_2, \ldots, x_n)$. Create a robust regression model and calculate the fitted values $(\widehat{y})$ and residuals $(r_i = y - \widehat{y})$.
Stage 2: Find the resampled residuals $(\widehat{r}_i^b)$ by randomly selecting (J) samples of size n and replacing them with residuals (r_i). These J samples are sometimes referred to as bootstrap samples $(j = 1, 2, \ldots, J)$. J=1000 is the application, (Andersen, 2008). **Stage 3:** Identify J sets of bootstrap fitted values $(\widehat{y}_i^b = \widehat{y}_i + \widehat{r}_i^b)$ and add the residuals $(\widehat{r}_i^b)$ to the fitted values $(\widehat{y})$ **Stage 4:** Get J sets of regression coefficients $(\widehat{\beta}_i^b)$ by regressing each set of J bootstrap fitted values $(\widehat{y}_i^b)$ on the fixed model matrix X. As a result, the regression coefficients are estimated as follows:

$$\overline{\widehat{\beta}}_i^{\,b} = \frac{1}{J} \sum_{j=1}^{J} \widehat{\beta}_j^J, \quad j = 1, 2, \ldots, J$$

Stage 5: For bootstrapping regression coefficients $(\widehat{\sigma}^2(\widehat{\beta}_i^b))$, the bootstrapping variance is defined:

$$\widehat{\sigma}^2(\widehat{\beta}_i^b) = \frac{1}{J-1}\sum_{j=1}^{J}(\widehat{\beta}_{ij}^b - \overline{\widehat{\beta}}_i^b)^2, i = 1,2,\ldots,p \quad \ldots(*)$$

2.6. APPLICATION

The R program was used to analyze the data in order to obtain the transformation function and parameter estimate methods for these variables. In both simple and multiple regression, two types of data are used. The total number of observations is 130 for simple and 105 for multiple. It can be found at (https://gov.krd/coronavirus-en/dashboard/) for the daily new cases, active cases, and new recovered from October 1, 2021, to January 13, 2022 and from October 1, 2021, to February 7, 2022. The daily data of COVID- 19 cases of Duhok and the active cases were represented by (Y), the predictors were new recovered (X_1) and new cases (X_2). Consider the dependent variable(Y) as active cases and independent variable(X) as new cases for a simple data analysis.

Table 2: Summary table for untransformation and transformation of simple data

	OLS				
	RMSE	Skewness	Kurtosis	P_Shapiro	P_BreuschPagan
Untransformed	1480.081	1.22	5.59	1.342230e-08	0.075
Transformed(y)	0.35	0.24	3.2	0.0703	0.99

Table(2) displays the original (untransformed) and transformed models that were evaluated on the data using OLS for original and transformation data (the (logshiftopt) model to transform the response variable). Since the RMSE of converted data is smaller than original data, and the Shapiro-Wilk test for normality of the residuals for the transformed model is larger than a 5% level of significance, the best performance was achieved. Moreover, the skewness indicates that the residuals in the transformed model are more symmetric, and the kurtosis is nearer to 3, the normal distribution's kurtosis value. The Breusch-Pagan test shows that the transformed model has homoscedasticity. Figure 1 shows diagnostic charts that support these two findings.

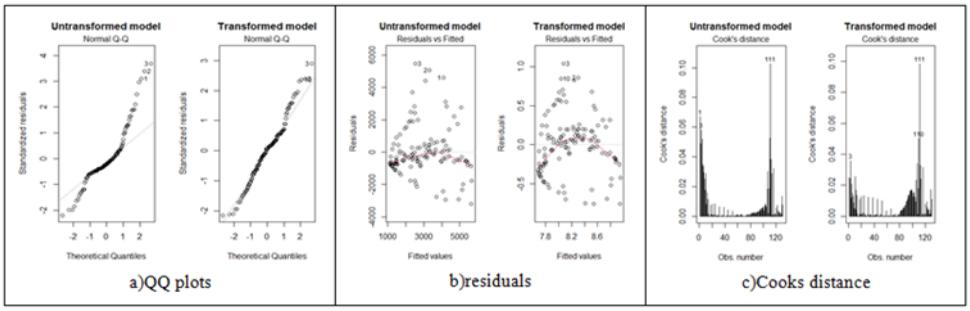


Figure 1: Simple model by using plot (f-trafo). (a) shows QQ plots error terms of the untransformed and the transformed model. (b) shows the residuals against the fitted values of the untransformed and the transformed model. (c) Cooks distance of the untransformed and the transformed model.

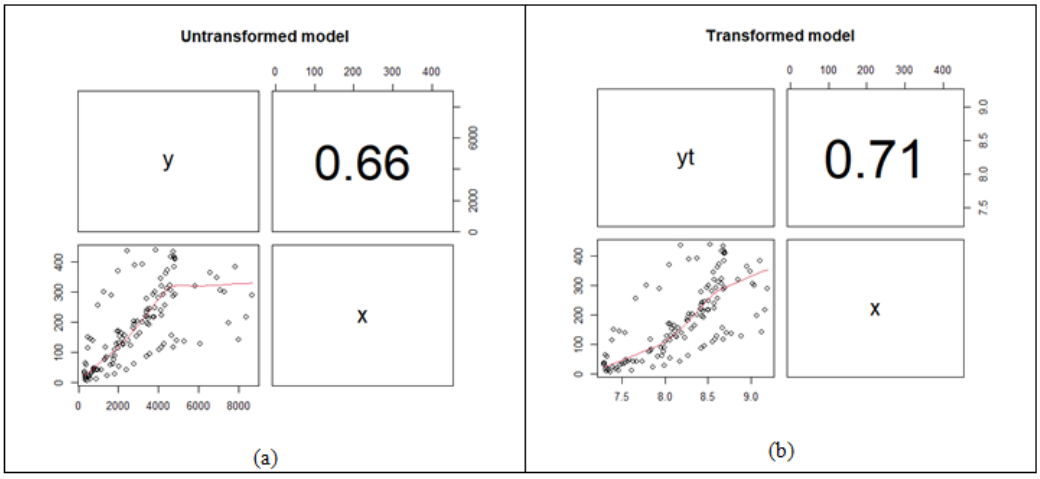


Figure 2: simple model by using plot(f-trafo). (a) shows the scatter plot of the untransformed (b) shows scatter plot of the transformed. The numbers specify the correlation coefficient between the dependent and independent variable.

To evaluate the linearity assumption, scatter plots of the response variable against the independent variable can help. The assumption of linearity is violated in the untransformed model, as seen in Figure 2. The original and transformed data, on the other hand, have a linear relationship. Also, the residual plot and Normal Q-Q Plot in Figures 1 (a) and 1 (b) shows residuals based on predicted raw scores from the transformation regression model, suggesting that the transformation to achieve linear regression model was successful. Moreover, as shown in Figure 1 (c) cooks distance before and after the transformation, the data contains some outliers.
The dependent variable in the second data analysis was chosen with (y) representing active cases and (x_1) representing New Recovered and (x_2) representing New Cases.

Table 3: Summary table for untransformation and transformation of multiple data

	OLS				
	RMSE	Skewness	Kurtosis	P_Shapiro	P_BreuschPagan
Untransformed (original)	811	0.867706	4.19627	0.0003936	2.915674e-07
Transformed (x,y)	1.033	0.081509	3.42260	0.3666	0.11902005

Table (3), the best results were obtained by transforming both x and y using a dual model for the outcome variable (y) and a logarithmic model for the explanatory variable (X_1). It shows that the transformed data has a lower RMSE than the original data, and that the p-values of the Breusch-Pagan and the Shapiro-Wilk tests are greater than 0.05. Furthermore, the skewness shows that the residuals in the transformed data are more symmetric, and the kurtosis is closer to 3, implying that the violation of the homoscedasticity assumption can be fixed by transformation normality and homoscedasticity in data.

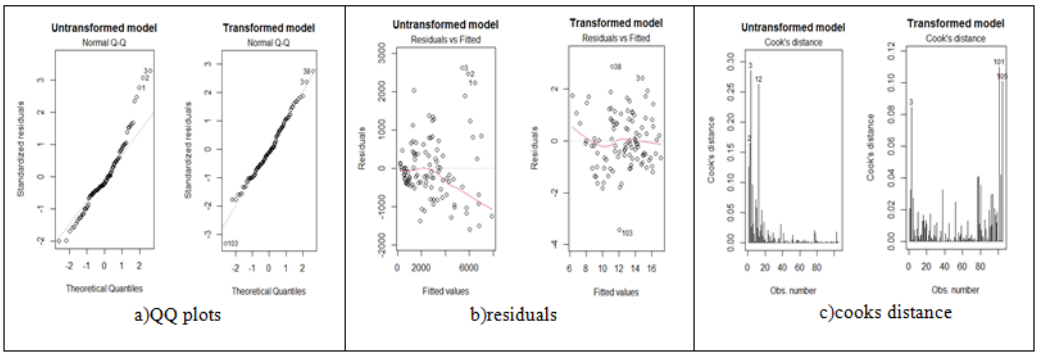


Figure 3: multiple model by using plot(f-trafo). (a) shows QQ plots error terms of the untransformed and the transformed model. (b) shows the residuals against the fitted values of the untransformed and the transformed model. (c) Cooks distance of the untransformed and the transformed model.

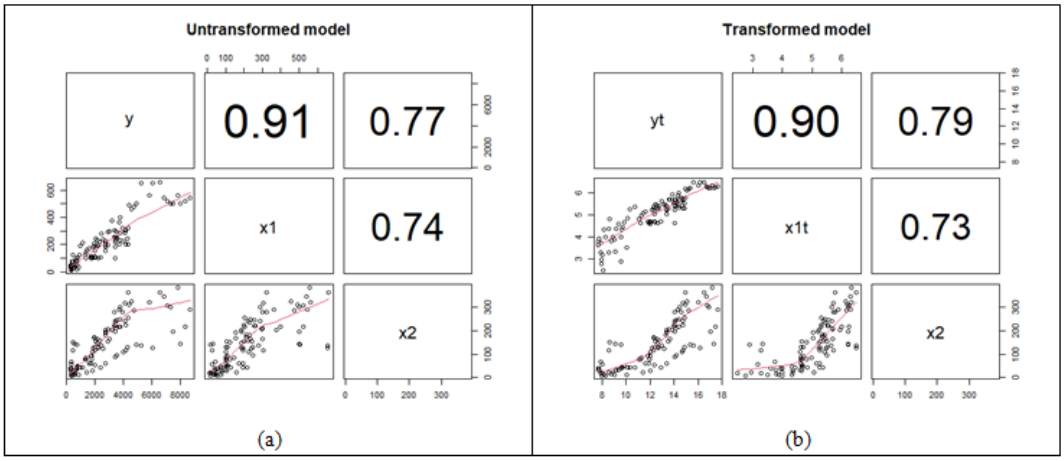


Figure 4: multiple model by using plot(f-trafo). (a) shows the scatter plot of the untransformed (b) shows scatter plot of the transformed. The numbers specify the correlation coefficient between the dependent and independent variable.

The residual plot and Normal Q-Q plot in Figures 3 (a) and 3 (b) look to be residuals built on predicted raw scores from transformation regression equation while Figure 4 (b) appears to linearity. The transformation data is used to achieve linear regression. Also, as seen in Figure 3 (c) for Cooks distance, the data contains some outliers before transformation and after transformation.

Table 4: The summary results based on different regression methods and their Corresponding Standard Errors for simple real data

	Boot-OLS		Boot-MM		Boot-GM6	
	b0	b1	b0	b1	b0	b1
	S.E.(b0)	S.E.(b1)	S.E.(b0)	S.E.(b1)	S.E.(b0)	S.E.(b1)
Untransformed	1029.9	10.5	672.9	10.9	541.16	11.78
	(229.1)	(1.05)	(146.6)	(0.67)	(140.20)	(0.62)
Transformed(y)	7.66	0.002	7.62	0.003	7.5	0.003
	(0.056)	(0.00025)	(0.055)	(0.00025)	(0.055)	(0.00026)

For both original and transformed data, Table (4) shows how to estimate parameter and standard error using boot strapping. (Boot-OLS) method is the worst in both cases since it has the highest standard

error. However, the MM estimator's standard error is lower than the GM6, and their results are consistently better than the GM6 estimator. This finding suggests that bootstrapping may be used to improve the efficiency of robust regression, which is the major goal of this study.

Table 5: Regression Estimates and their Corresponding Standard Errors for multiple data

	Boot-OLS			Boot-MM			Boot-GM6		
	b0	b1	b2	b0	b1	b2	b0	b1	b2
	S.E.(b0)	S.E.(b1)	S.E.(b2)	S.E.(b0)	S.E.(b1)	S.E.(b2)	S.E.(b0)	S.E.(b1)	S.E.(b2)
Untransformed	49.62	9.42	4.37	101.12	7.13	6.21	239.6	5.49	7.33
	(144.62)	(0.71)	(1.20)	(118.5)	(0.53)	(0.87)	(128.48)	(0.81)	(1.08)
Transformed (x,y)	1.08	2.03	0.007	1.31	1.97	0.008	-0.008	2.27	0.006
	(0.68)	(0.16)	(0.0015)	(0.67)	(0.15)	(0.001)	(0.96)	(0.22)	(0.0017)

Table (5) shows the coefficient parameter estimates and standard errors from the original and transformed in multiple regression using the bootstrap technique using the OLS estimator, MM-estimator and GM6-estimator. In this case, one thousand bootstrapped samples are used. The results show that the OLS estimates of the original and transformed data are large when compared to the other estimators. Table (5) shows that regression estimates using the bootstrap robust (MM-estimator) regression fitting techniques have uniformly smaller standard errors. This finding suggests that bootstrapping can be used to improve the efficiency of robust regression, which is the main goal of this study.

3. Conclusion

This paper examines the performance of transformation data and three bootstrap regression methods. We illustrated how the trafo package makes it simple for the user to decide which transformation is suitable for fulfilling the assumptions of linearity, homoscedasticity, and normality. Trafo is the only R package that we are aware of that supports this choice procedure. The numerical results show that the proposed Boot-MM outperforms the Boot-OLS and Boot-GM6.

Acknowledgment

The authors would take the opportunity to thank those who helped them complete the work.

References

[1] R. Andersen *Modern methods for robust regression*,Sage, 152 (2008). https://doi.org/10.4135/9781412985109

[2] M. S. Bartlett *The use of transformations*, Biometrics , **3** 1 (1947),39–52. 1 2.1

[3] P. J. Bickel and K. A. Doksum *An analysis of transformations revisited*, J. Am. Stat. Assoc., **76** 374 (1981),296–311. 1

[4] G. E. P. Box and D. R. Cox *An analysis of transformations*,Journal of the Royal Statistical Society: Series B (Methodological), **26** 2 (1964),211–243. 2.1

[5] T. S. Breusch and A. R. Pagan *A simple test for heteroscedasticity and random coefficient variation*, Econometrica: Journal of the econometric society, (1979),1287–1294. https://doi.org/10.2307/1911963 2.1

[6] C. W. Coakley and T. P. Hettmansperger *A bounded influence, high breakdown, efficient regression estimator*, J. Am. Stat. Assoc., **88** 423 (1993),872–880. https://doi.org/10.1080/01621459.1993.10476352 1, 2.4

[7] B. P. Durbin, J. S. Hardin, D. M. Hawkins, and D. M. Rocke, *A variance-stabilizing transformation for gene-expression microarray data*, Bioinformatics, **18** suppl-1 (2002),S105–S110. https://doi.org/10.1093/bioinformatics/18.suppl$_1$.s1051

[8] B. Efron and R. Tibshirani *Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy*, Stat. Sci., (1986),54–75. https://doi.org/10.1214/ss/1177013816 2.5

[9] C. Flexeder *Generalized lasso regularization for regression models*,Institut für Statistik, (2010). 2.2

[10] J. Fox and S. Weisberg *An R companion to applied regression. Sage*, Thousand Oaks, (2011). 2.1

[11] R. W. Hill and P. W. Holland *Two robust alternatives to least-squares regression*, J. Am. Stat. Assoc., **72** 360a (1977),828–833. https://doi.org/10.2307/2286469

[12] P. J. Huber, *Robust statistics*,John Wiley and Sons, (2004). 1
1

[13] S. S. ISMAEEL *Robust Diagnostic and Robust Estimation methods for Fixed Effect Panel Data Model in Presence of high Leverage Points and Multicollinearity*, (2017). 2.5

[14] A. M. Leroy and P. J. Rousseeuw *Robust regression and outlier detection*, Wiley Ser. Probab. Math. Stat., (1987). 1

[15] H. Midi, S. S. Ismaeel, J. Arasan, and M. AMohammed *Simple and Fast Generalized-M (GM) Estimator and Its Application to Real Data Set*, Sains Malaysiana, **50** 3 (2021), 859–867. https://doi.org/10.17576/jsm-2021-5003-26 1

[16] M. R. Norazan, H. Midi, and A. Imon *Estimating regression coefficients using weighted bootstrap with probability*, WSEAS Trans. Math., **8** 7 (2009), 362–371. 1

[17] P. J. Rousseeuw *Multivariate estimation with high breakdown point*, Math. Stat. Appl., **8** 37 (1985),283–297. https://doi.org/10.1007/978-94-009-5438-$0_2$01

[18] P. J. Rousseeuw and A. M. Leroy *Robust regression and outlier detection*, John wiley and sons, (2005). 2.3

[19] S. S. Shapiro and M. B. Wilk *An analysis of variance test for normality (complete samples)*, Biometrika, **52** 3/4 (1965),591–611. https://doi.org/10.2307/2333709 2.1

[20] D. G. Simpson, D. Ruppert, and R. J. Carroll *On one-step GM estimates and stability of inferences in linear regression*, J. Am. Stat. Assoc., **87** 418 (1992),439–450. https://doi.org/10.1080/01621459.1992.10475224 1

[21] W. N. Venables and B. D. Ripley *Modern applied statistics with S. 4th Springer*, New York, 118 (2002). 2.1

[22] J. Whittaker, C. Whitehead, and M. Somers *The neglog transformation and quantile regression for the analysis of a large credit scoring database*, J. R. Stat. Soc. Ser. C (Applied Stat., **54** 5 (2005),863–878. https://doi.org/10.1111/j.1467-9876.2005.00520.x 1

[23] R. R. Wilcox *Introduction to robust estimation and hypothesis testing (Statistical Modeling and Decision Science)*, Academic press, (2005). 1, 2.5

[24] I. Yeo and R. A. Johnson, Kjell A *A new family of power transformations to improve normality or symmetry*,Biometrika, **87** 4 (2000),954–959. https://doi.org/10.1093/biomet/87.4.954 1

[25] V. J. Yohai *High breakdown-point and high efficiency robust estimates for regression*, Ann. Stat., (1987),642–656. https://doi.org/10.1214/aos/1176350366 1, 2.3