Science World Journal Vol 14(No 2) 2019
www.scienceworldjournal.org
ISSN 1597-6343
Published by Faculty of Science, Kaduna State University

Full Length Research Article

RESIDUAL BOOTSTRAPS FOR REGRESSION MODEL VALIDATION

[1]Adewole Ayoade I., [2]Loyinmi A.C.

[1,2]Department of Mathematics, College of Science and Information Technology, Tai Solarin University of Education, Ijagun, Ogun State, Nigeria

*Corresponding Author Email Address: hayorhade2005@gmail.com Phone: +2348055124368

ABSTRACT
Validation is a useful and necessary part of the model-building process, identification of one or several "good" regression models is not the end of the model-building process and these models must be evaluated by various diagnostic procedures before the final regression model is determined. Residual Bootstrap method in regression model validation accomplish the goal of constructing appropriate sampling distributions empirically using the data at hand instead of statistician relying on theoretical sampling distributions like the normal, t and f where appropriateness for any given problem always rest on untestable assumptions. Validation statistics of interest such as standard error (SE), mean square error (MSE) and coefficient of determination (R^2) were used as criteria for selecting the best model suitable for predictive purposes.The research work concluded that to reduce the problem of overffited models in regression analysis, residual bootstrap approach should be employed in checking the validation of regression model as it gives a better estimates and stable value of coefficient of determination.

Keywords: Bootstrapping, Validation, Residual and Regression models

1. INTRODUCTION

Fitting a regression model is not the end of a regression analysis on a set of data. The question as how valid the fitted regression comes to mind almost immediately. Model Validation requires checking the model against independent data to see how well it predicts (Predictive accuracy of the model is how the model validate a new data set). Model selection and validation are critical in predicting a dependent variable given the independent variable. The correct selection of variables minimizes the model mismatch error while the selection of suitable model reduces the model estimation errors. Models are validated to minimize the model prediction error.

Model validity refers to the stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function, and the ability to generalize inferences drawn from the regression analysis. Model validation is an important step in the modeling process and helps in assessing the reliability of models before they can be used in decision making (Jannath and Tsuchido, 2003). Validation typically involves the comparison of model predictions with analogous observation of the calculated pack experiment not used to develop the model or with literature values.

Snee (1977) researched extensively by listing four methods of validation, Dempster *et al.*(1977) worked on the use of stimulation studies to assess the mean square error as a validating technique supported Snee's suggestion on comparing prediction of these regression procedures for validation. Furthermore, Renduer and Pun (1980), kleijnen and Deflandre (2006) and recently Oredein *et al. (*2011), Charles *et al.* (2016) and Lan and Phoebe (2018) also worked on model validation.
Bootstrap was introduced by Efron (1979) as a general approach to statistical inference based on building a sampling distribution for a statistic by re-sampling from the data at hand. Also Efron and Gong (1983) depicted several bootstrapping techniques in obtaining nearly unbiased estimates of future model without holding back data when making the final estimates of model parameters. A fewer recent publication on bootstrapping in simulations are available such as; McCullagh (2000), Fox (2002), Davison *et al.*(2003) and Lendasse *et al.*(2005) discussed the application of fast bootstrap methodology for regression model selection likewise, Olatayo (2010) showed how bootstrap method can be applicable to complicated data structure, such as time series data.

This research work describes how residual bootstrap techniques can be applied in checking the validity of a regression models. Residual bootstraps analysis on two data sets pertaining telecommunication and market stocks was carried out with a view of studying model validation process. Statistic such as standard error (SE), mean square error (MSE) and coefficient of determination (R^2) were used as criteria for selecting the best model suitable for predictive purposes.

2. MATERIALS AND METHODS

In this research work, focus is on the implementation of validating a regression model by residual bootstrapping approach, it was achieved by resampling the residuals from the fitted models. Residuals were estimated as the difference between the calculated value of the dependent variable against the predicted variable. The data involved two data set; first data set was a stock exchange data obtained from Nigeria stock exchange office in Marina, Lagos, Nigeria. The stock exchange data consist of number of deals, quantity traded and value of shares as the independent variables predicting the all share index per week. The observations were selected for 40 weeks.

The second data set pertains to different hourly readings of bytes received in Airtel telecommunications in Victoria Island Lagos, Nigeria. The telecommunication data consists of Bytes transmitted, Link utilization received, link utilization transmitted real time and Best effort, which were used as the independent variables, predicting the Bytes received. The observations were recorded for 300 hours. The model was fitted to the sample and calculate the measure of predictive accuracy that is the standard error, coefficient of determination (R^2) and Mean square error

(MSE). The validation statistics was estimated using the following estimations:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 \quad (2.1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2} \quad (2.2)$$

where $Y_i, \hat{Y}_i \text{ and } \bar{Y}$ are observed values, predicted values and mean value of dependent variable respectively.
The bootstrap method estimates typically assumes that data are independent and identically distributed (i.i.d.) and this was assumed in this work.

The bootstrap standard error estimates reflect the variability between the estimates obtained if the sample is repeatedly taken from the population. Estimate the standard error $S_e(\hat{\theta})$ by the sample standard deviation of the *N* replication where

$$\hat{S}_e\left(\hat{\theta}\right) = \left(\frac{\sum_{n=1}^{N}\left\{\hat{\theta}^{"}(n) - \hat{\theta}^{"}(.)\right\}^2}{N-1}\right)^{\frac{1}{2}} \quad (2.3)$$

where $\hat{\theta}^{*}(.) = \frac{\sum_{n=1}^{N}\hat{\theta}^{"}(n)}{N}$ (2.4)

The approach of resampling the residuals in bootstrapping the regression models was achieved through the following algorithm;

Considering the simple linear regression model
$Y = \beta x + \varepsilon$ (2.5)

(i) Fit the model to the original data, obtain the estimates $\hat{\beta}$, retain the fitted values $\hat{Y}_i$ and the residuals $\hat{\varepsilon}_i$, $i = 1, \ldots, n$.
(ii) Compute a bootstrap sample ε_i^* with the recursive formula $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ by sampling with replacement.
(iii) construct new data Y* with the resampled residuals ε_t^*
(iv) Create the bootstrap values for the response variables by adding the original predicted values to the bootstrap residuals i.e. $Y^* = \hat{Y}_i + \varepsilon_i^*$
(v) Regress Y* on the original X variable(s).
(vi) Fit the model using $Y^* = \hat{Y}_i + \varepsilon_i^*$
(vii) Estimate parameter of interest in validation of regression models such as standard error, coefficient of determination (R^2) and Mean square error from the fitted model.
(viii) Repeat the procedure in N times.

3. RESULTS AND DISCUSSION

The study examined approaches of validating regression models using residual bootstrap approach. The bootstrap models were obtained from 1000 bootstrap replication. The bootstrap Y values was computed by adding resample residuals onto the ordinary least squares regression fit. The N =1000 bootstrap samples were generated randomly to reflect the exact behaviour of bootstrap estimations.
The summary of the research findings reveals the performance of residual bootstrapping in validating regression model based on the execution of the outlined algorithm listed in section 2 above is illustrated as follows;
The validating model obtained for stock exchange data using bootstrap residual resampling is

$Y = 1.3261 + 0.2571x_1 - 0.0196x_2 + 0.3359x_3$ (3.1)

Also

$Y = 7.0732 - 0.00294x_1 - 0.1842x_2 + 0.0552x_3 - 0.3670x_4 - 0.9261x_5$ (3.2)

was obtained as the validating model for telecommunication data set using bootstrap residual resampling procedures. The above models were chosen because they generated highest and lowest value of R^2 and standard error, mean square error respectively, as a criterion of model validation.
The criterion statistic used for the choice of validity models are presented in the table below:

Table 1. Summary of validated model using residual bootstrap.

VALIDATION STATISTIC	STOCK EXCHANGE DATA	TELECOMMUNICATION DATA
Standard error	0.2216	0.3759
Mean square error	0.1335	0.4617
R^2	0.9774	0.9926
Adj R^2	0.9205	0.9510

Two set of data samples were considered to check how residuals bootstrap approach works on small and larger data set in validating regression models. The number of bootstrap replications N depends on the application and size of sample and computer availability.
From above results, it was discovered that the larger the bootstrap replicate, the higher and stable coefficient of determination is, that is; it gives a better validity model. The residuals bootstrapping techniques yields minimum and temperate values of standard error (SE), mean square error (MSE) and high values of coefficient of determination (R^2) which are good and reliable for predictive purposes. It was observed from the above, comparing the data set in Telecommunication and stock exchange data, bootstrap gives a better coefficient of determination both in small and large sample data sets as it conforms to the work of Oredein *et al.,* (2011) that compared the use of data splitting techniques and sample bootstrapping in validating regression models

4. Conclusion and Recommendation

Validation is a useful and necessary part of the model-building process, identification of one or several "good" regression models is not the end of the model-building process. These models must be evaluated by various diagnostic procedures before the final

regression model is determined. This model then needs to be validated before it is considered to be satisfactory.

A good use of the residual bootstrap for validation purpose is in estimating the optimism of an index of predictive accuracy. Bootstrap method is preferable in validating linear regression because of some theoretical properties like having any distributional assumptions on the residuals and hence allows for inference even if the errors do not follow normal distributions. Residual resampling assures fixed X values and independently and identically distributed errors (but not necessarily normal) that is, it assures that the residual found for the i^{th} case could equally well have occurred with the j^{th} case instead, residual resampling randomly reassigns the original-sample residuals to new case. The n sets of X values from the original sample remain unchanged in each bootstrap sample. The study recommend that statisticians should persuade researchers to validate their models and correct for variable selection effects in their models. The outlined procedures in this work for selection and validation of predictive models may be valuable for similar applications in the analysis and modelling of complex manufacturing processes and systems.

REFERENCES

Charles Laurin, Dorret B. and Giotta L. (2016). The use of vector bootstrapping to improve variable selection precision in lasso models. Statistical applications in genetics andmolecular Biology. 2016(15$). 305 -320.

Davison, A.C., Hinkley, D.V. and Young, G.A. (2003), *Recent Developments in bootstrap methodology*. *Statistical Science*, Vol. 18, No. 2, 141-157.

Dempster, A.P., Schatzoff, M., & Wermuth, N.(1977) "A Simulation Study of Alternatives to Ordinary Least Squares," *Journal of the American Statistical Association*, 72, 77-91.

Efron, B. (1979a) Bootstrap methods: another look at the jackknife. *Ann. Statist*. 7, 1-26.

Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife and cross validation. *Amer. Statistician* 37, 36-48.

Fox, J. (1997). *Applied Regression Analysis*, Linear Models and Related Methods.

Jack P.C Kleijnen and David Deflandre (2006). Validation of regression metamodels in simulation: Bootstrap approach. European Journal of operational research 170 (2006) 120 – 131.

Jannath, A. and Tsuchido, T.(2003) Predictive microbiology: A review,*Biocontrol Science* 8,1-7.

Lan T. Tran and Phoebe Tran (2018). Validating Geospatial Regression models with bootstrapping. International journal of Geospatial and Environmental research vol 5; n0:1 Article 1.

Lendasse, A., Simon, G., Wertz, V. and Verleysen, M. (2005) Fast bootstrap methodology for regression model selection. Neurocomputing, 64, 161-181.

McCullagh, P. (2000). Resampling and exchangeable arrays. *Bernoulli* 6,285-301.

Olatayo T.O. (2010). On Truncated Geometric Bootstrapping Method for Stochastic Time Series Process. Unpublished Ph.D Thesis.

Oredein A.I, Olatayo T.O. and Loyinmi A.C. (2011). On validating regression models with Bootstraps and Data splitting techniques. Global Journal of Science Frontier Research.11(6) 1.0, 1-6.

Rencher, A.C., and Pun, Fu Ceayong, (1980), Inflation of R^2 in best subset regression, *Technometrics* 22 (1), 49-53.

Snee, R.O, (1977), Validation of Regression Models, Methods and Examples, Techniques 19, 415 – 428.

Tejedor, W., Rodrigo, M. and Martinez, A. (2001). *Journal of Food Protection* 10, 1631-1635.

Winkelman R. and Mehmud S. (2007) "A Comparative Analysis of Claims Based Tools for Health Risk Assessment, *Society of Actuaries*.