Global Journal of Science Frontier Research
Volume 11 Issue 6 Version 1.0 September 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN : 2249-4626 &Print ISSN : 0975-5896

On Validating Regression Models with Bootstraps and Data Splitting Techniques

By A.I Oredein , T.O Olatayo , A.C Loyinmi
Tai Solarin University of education, Ijebu-Ode, Nigeria.

Abstract - Model validity is the stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function and ability to generalize inference 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. This research work therefore seeks to study regression model validation process by bootstrapping approach and data splitting techniques. We review regression model validation by comparing predictive index accuracy of data splitting techniques and residual resampling bootstraps. Various validation statistic such as the mean square error (MSE), Mallow's cp and R2 were used as criteria for selecting the best model and the best selection procedure for each data set. The study shows that bootstrap provides the most precise estimate of R2 which reduce the risk over fitted models than in data splitting techniques..

Keywords : *Validation, bootstrap, Data splitting techniques, coefficient of determination, and stepwise regression .*

GJSFR Classification: *FOR Code: 170202*

Strictly as per the compliance and regulations of:

On Validating Regression Models with Bootstraps and Data Splitting Techniques

A.I Oredein[α], T.O Olatayo[Ω], A.C Loyinmi[β]

Abstract - Model validity is the stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function and ability to generalize inference 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. This research work therefore seeks to study regression model validation process by bootstrapping approach and data splitting techniques.

We review regression model validation by comparing predictive index accuracy of data splitting techniques and residual resampling bootstraps. Various validation statistic such as the mean square error (MSE), Mallow's cp and R^2 were used as criteria for selecting the best model and the best selection procedure for each data set. The study shows that bootstrap provides the most precise estimate of R^2 which reduce the risk over fitted models than in data splitting techniques.

Keywords : Validation, bootstrap, Data splitting techniques, coefficient of determination, and stepwise regression.

I. Introduction

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 error. Models are validated to minimize the model prediction error. A more flexible model can better represent the data may also more easily lead the user astray by noise in the data. Determining the right form of the model in order to reduce model mismatch error is accomplished during model construction phase, whereas determining the correct model parameter can be achieved at the model selection and validation.

Once a regression model has been constructed, it is important to confirm the goodness of fit of the model and the statistic significance of the estimated parameters, commonly used are check of goodness of fit include analysis of the pattern of residuals and hypothesis testing, statistically significance checked by an f-test of the overall fit , followed by t-test of individual parameters interpretation of these diagnostic tests.

Validation is an essential part of model building, its application and levels of confidence in usage are highly important. It entails checking the R^2 statistic from the regression fit, carrying out a diagnostic of the residual either through exploratory statistic, checking the mean confirmatory statistics, checking the mean square error and also the mallow C_p statistic.

Model validity refers to stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function and ability to generalize inferences drawn from the regression analysis. Validation is a useful and necessary part of the model building process. A good fit of a model to the data set is not an only goal of model validation but also to get a perfect fit by n-1 parameter to a data set with n cases. i.e. its predictive accuracy of the model is how the model validates a new dataset.

Model validation requires checking the model against independent data to see how well is predicts. Several researchers have work extensively on model validation using Jackniffing, Data splitting techniques, data resampling bootstraps regression without assuring fixed X or identically distributed errors. A drawn back of cross validation is the choice of the number of observation to hold out each fit. Also cross validation may not fully represent the variability of variable selection. The major disadvantages of data splitting techniques in model validation is that different investigators using the same data could split the data differently generate different models, hence obtain different validating result. Snee (1997) researched extensively on method of validation, Neumann et al (1977) and Shapiro (1984) have employed Monte Carlo testing to estimate artificial predictability in tropical Cyclane prediction models. Renduer and Run (1980) and Lanzante (1984) carried out set of Monte Carlo test to examine false predictability and the inflation of R^2 as a function of sample size, size of the predictor and number of predictors selected.

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 1988). Hall and Wilson(1991),Davison and Hinkely(1997),Efron(1998)

Author α : Department of Mathematics, Tai Solarin University of education, Ijebu-Ode, Nigeria.

Author Ω : Department of Mathematical Science, Olabisi Onabanjo University, Ago-Iwoye, Nigeria.

considered the application of bootstrap method to regression models from model based resampling approach.

These research works examine the validation of regression by comparing the predictive accuracy of data splitting techniques and the newly introduced bootstrapping approach by Efron (1993), to check the significance of each method in regression model validation. This work proposes a procedure for construction, selection and validation of regression models.

However, in regression model validation analysis, fewer reports have shown how bootstrap can be used in estimating the distribution of any validation statistic in random simulation with replicated runs. Unfortunately, this simplicity and versatile techniques of bootstrapping approach in validation seems not to be well known among simulation users and researchers. A fewer recent publication on bootstrapping in simulations are Cheng (2004), Deflandre and Kleijnen etal (2001) and Willemain etal (2003. This research work will extensively shows how bootstrap technique can be applied in checking the validity of a regression models using residual resampling. This work will rely less on theoretical sampling distribution like the normal, χ^2, t and F, whose appropriateness for any given always rest on untestable assumptions. Instead we will construct appropriate sampling distribution empirically through bootstrap method using the data at hand.

II. Material and Methodology

Validating regression model was implemented in this work by bootstrapping approach and using the technique of data splitting. In data splitting, three different regression procedures were used to fit regression model to two different data sets. The data sets have many variables predicting the response variable. In data splitting, we split the data sets into two separate samples using one part for modeling and the other for testing the model. We also hope to see if the peculiarities of the original set will be seen in the split modeling set.

The first data set is a stock exchange data using Number of deals; 'Quality traded' and 'values of shares' as the independent variables predicting the 'All share index' per week. The observations were selected over 50 weeks.

The second data set pertains to different hourly readings of bytes received in telecommunication industry. 'Bytes transmitted', link utilization received, link utilization transmitted, 'Real time' and 'Best effort' were used as the independent variables predicting, 'Bytes received'. The observations were recorded over 130 hours. In data splitting techniques we employed the approach of stepwise regression procedures in selecting variables into a regression model. These include Forward Selection, Backwards Elimination and Best subset Regression. They add or remove variables one at a time until some stopping rule is satisfied.

The forward selection regression procedure sequentially adds variables to the model one at a time. It starts with an empty model and adds the variable that has the smallest p value usually less than 0.05 or 0.1 to the model.

Aside the p-value criterion, at any stage in the selection process, forward selection adds the variable that has the highest partial correlation, increases R^2 the most, and gives the largest absolute t of F statistic to the model. This procedure is a model reduction method. The Backward Elimination regression procedure starts with all the predictors in the model and sequentially deletes variables from the model. At any stage, in the selection process, it deletes the variables with the smallest absolute t or F-statistic, largest p-value and smallest R^2 ,backward Elimination procedure gives an adequate model since the procedure involves starting the model building with all the variables and deleting the variables that add nothing to the model.

Best subset regression examines all possible models and chooses the one with the most favorable value of some summary measure such as large adjusted, R^2 smallest Mallow's C_p and smallest standard error. All possible regression has a large advantage over stepwise procedures in that it can let the analyst see competing models, models that are almost as good as the best.

Data splitting has the advantage of allowing hypothesis tests to be confirmed in the test sample, however, the major disadvantages it has is that different investigators using the same data could split the data differently and generate different models, hence obtaining different validating results.

a) Bootstrap Estimate Of Standard Error

The bootstrap was introduced in 1979 as a computer based method for estimating the standard error of $\hat{\theta}$ .The bootstrap estimate of standard error requires no theoretical calculations, and is available no matter how mathematically complicated the estimator $\hat{\theta} = s(x)$. Bootstrap methods depend on the notion of a bootstrap sample . A bootstrap sample is defined to be a random sample of size n drawn from F , X^* is defined as

$$X^* = (x_1^*, x_2^* \ldots\ldots\ldots\ldots\ldots, x_n^*) \quad)$$

And

$$\hat{F} \rightarrow (x_1^*, x_2^* \ldots\ldots\ldots, x_n^*)$$

The star notation indicates that is not the actual data set *x*, but rather a randomized or resample version of X, in other word. Bootstrap sample can be

defined as bootstrap data points x_1^*, x_2^* ....., x_n^* that are random sample of size *n* drawn with replacement from the population of *n* objects(x_1, x_2……, x_n).The bootstrap data set(x_1^*, x_2^*,…….. x_n^*) consists of the original data set (x_1, x_2……, x_n) some appearing zero times, once, twice etc.

Corresponding to a bootstrap data set $X^*$is a bootstrap replication of $\hat{\theta}$,

$$\hat{\theta}^* = s(x^*)$$

s(x^*) is the mean of the bootstrap data set.

$$\bar{x}^* = \frac{\sum_{i=1}^{n} x_i^*}{n}$$

The bootstrap estimate of $S_{ef}(\hat{\theta})$i.e.the standard error of a statistical $\hat{\theta}$, is a plug-in estimate that uses the empirical distribution function $\hat{F}$ in place of the unknown distribution *F*. Specifically the bootstrap estimate of $S_{ef}(\hat{\theta}^*)$. is defined by

$$S_{ef}(\hat{\theta}^*).$$

In other words, the bootstrap estimate of $S_{ef}(\hat{\theta})$ is the standard error of $\hat{\theta}$ for the data sets of size n randomly sampled from F.

b) The Bootstrap Algorithm For Estimating Standard Errors

1. Select *B* independent bootstrap samples (X^{*1}, X^{*2}. ……..,X^{*B}) each consisting of n data values drawn with replacement from $x = (x_1, x_2 \ldots, x_n)$. (*B* is the number of bootstrap samples used).
2. Evaluate the bootstrap replication corresponding to each bootstrap sample:

$$\hat{\theta}(b) = S(X^{*b}) \text{ where } b = 1, 2\ldots, B.$$

3. Estimate the standard error $S_{ef}(\hat{\theta})$ by the sample standard deviation of the *B* replicatication where

$$\widehat{se}_B = \frac{[\sum \theta^*(b) - \theta^*(.)^2]}{(B-1)^{1/2}}$$

where

$$\hat{\theta}^*(.) = \sum_{b=1}^{B} \frac{\theta^*(b)}{B}$$

$$\widehat{bias}_B = \hat{\theta}(.) - t(\hat{F})$$

The bootstrap algorithm above works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications and estimating the standard error of $\hat{\theta}$ by the empirical standard deviation of the replications.

c) Bootstrap Estimate of Bias

F is an unknown probability distribution, given data $x = (x_1, x_2 \ldots, x_n)$ by random sampling

$$F \rightarrow x$$

To estimate a real value parameter

$$\theta = (F)$$

Let statistic $\hat{\theta}$ = s(*x*) to be an estimator, using plug-in-estimate

$$\hat{\theta} = t(\hat{F})$$

The bias of $\hat{\theta} = s(x)$ as an estimate of θ is defined to be the difference between the expectation of $\hat{\theta}$ and the value of the parameter $\hat{\theta}$.

$$bias_f = bias_f(\hat{\theta}, \theta) = E_F[s(x)] - t(F)$$

The bootstrap estimate of bias is defined to be the estimate $bias_F$

$$bias_{\hat{F}} = E_{\hat{F}}[s(X^*)] - t(\hat{F})$$

d) Validation Using Bootstrap

Efron and Gong, Efron and Tibshirani,(1993) describe several bootstrapping procedures for obtaining nearly unbiased estimates of future model performance without holding back data when making the final state of model parameters. With the "simple bootstrap", one repeatedly fits the model in a bootstrap sample and evaluates the performance of the model on the original data.

A simple regression bootstrap called residual resampling was achieved through the following algorithm with the aid of computer

(i) Perform regression with the original sample; calculate predicted values ($\hat{Y}$) and residuals(e)
(ii) Randomly resample the residuals, but leave X and ($\hat{Y}$) unchanged.
(iii) Construct new Y* values by adding the original predicted values to the bootstrap residuals i.e Y* = $\hat{Y}$ + e*
(iv) Regress Y* on the original X variable(s).
(v) Repeat step (ii) – (iv) several times.
(vi) Estimate parameter of interest in validation of regression models such as R^2 and MSE.

The ability to study the arbitrariness of how a stepwise variable selection algorithm selects "important" factors is a major benefit of bootstrapping.

III. Results

Summary of result obtained in data splitting techniques using validating set of stock exchange data

	R^2	ADJ R^2	MSE
LSE	0.617	0.625	0.036
FORWARD	0.3955	0.3952	0.008
BACKWARD	0.3955	0.3922	0.008
BEST SUBSET	0.624	0.630	0.007

We went further in comparing the MSE obtained from the modeling set and validation of stock exchange data.

The table below shows the summary of the MSE obtained

Set	No. of Obs.	LSQ.	FWD	BKWD	Best Subset
Modeling	30	0.036	0.036	0.0362	0.0356
Validating	20	0.007	0.008	0.008	0.007

Hence the validated model is

$$Y \quad -1.417 - 0.023x_1 - 0.001x_2 + 0.176x_3$$
$$R^2 \quad 0.4985$$
$$R^2 Adj \quad 0.4960$$
$$MSE \quad 0.007$$
$$SE. \quad 0.08026$$

a) Analysis of Telecommunication Data

Summary of the result obtained in data splitting techniques using validating set of telecommunication data

	R^2	ADJ R^2	MSE
LSE	0.3124	0.2846	24.42
FORWARD	0.403	0.486	29.28
BACKWARD	0.402	0.482	29.28
BEST SUBSET	0.426	0.415	24.4

Summary of the MSE for Telecommunication Data

Set	No. of Obs.	LSQ.	FWD	BKWD	Best Subset
Modeling	110	163.18	157.5	157.52	163.07
Validating	20	24.42	29.28	29.28	24.4

MSE's from the validating set are smaller than those from the modeling set. This is not far from our expectation as this can be attributed to the distance of the observation of the validating set from the modeling Hence the validated model of telecommunication data is

$$Y \quad 3.67 + 0.87x_1 + 0.002x_4$$
$$R^2 \quad 0.426$$
$$Adj\ R^2 \quad 0.415$$
$$Cp \quad 4.78$$
$$S\,E. \quad 4.93$$
$$MSE \quad 29.816$$

b) Statistical Analysis of Bootsrap Approach In Validating Regression Models

The validating model obtained for stock exchange data using bootstrap residual resampling is

$Y = -34.4188 - 0.0793X_1 + 0,2684X_2 + 0.6504X_3$

$R^2 = 0.9854$, Adj $R^2 = 0.9848$, S.E = 0.6424, MSE $= 0.8015$, $N = 50$

Also s

$Y = 4.5094 + 0.6077X_1 - 0.0380X_2 + 0.0169X_3 + 0.0829X_4 - 0.0168X_5$ with $R^2 = 0.9899$, Adj $R^2 = 0.9895$, S.E $= 0.8826$, MSE $= 0.7790$, $N = 130$ 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 MSE respectively, as a criterion of model validation. Summary of validating statistics in validated models using data splitting and bootstraps

VALIDATING TECHNIQUES		R^2	Adj R^2	MSE
Data Splitting	Stock	0.4985	0.4960	0.007
	Telecommu-nication	0.426	0.4150	29.816
Bootstra-pping				
	Stock	0.9854	0.9848	0.8015
	Telecommu-nication	0.9899	0.9895	0.7790

IV. Discussion of Results

The bootstrap models were obtained from 100 bootstrap replication. The bootstrap Y values were computed by adding resample residuals onto the ordinary least squares regression fit. The B=100 bootstrap samples were generated randomly to reflect the exact behaviour of bootstrap estimations.

Residual resampling assumes fixed X values and independently and identically distributed errors (but not necessarily normal) that i.e. it assumes 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. Using a Monte Carlo algorithm, B bootstrap sample are generated by drawing from the empirical distribution with replacement. For each boot sample, the statistics of interest such as R^2, MSE, and standard error of estimate was calculated.

Two set of data samples were considered to check how bootstrap approach and data splitting techniques work on small and larger data set in validating regression models. The number of bootstrap replications B depends on the application and size of sample and computer availability.

From above results, it was discover that the larger the bootstrap replicate, the higher and stable R^2 is i.e. it gives a better validity model. It was also observed that in validating regression model bootstrap approach gives a better and higher R^2 in each replicates than the value of R^2 in the validated models of data splitting techniques.

advantages of allowing hypothesis tests to be confirmed in the test samples, but the major disadvantages of this method compared to bootstrap approach in model validation is that different investigators using the same data could split the data differently and senate difference models, hence obtain different validating results. It was observed from the above, comparing the data set in Telecommunication and stock exchange data, bootstrap give a better R^2 both in small and large sample data sets compared to stepwise regression i.e. the risk of over fitting was reduced. Also R^2 varies inversely with SSE, and it also increases if and only if MSE decreases, R^2 does not take account of number of parameters in the regression model.

This research work has demonstrated the use of validating regression models by data splitting techniques and bootstrap. In data splitting techniques, the data were split according to time into two samples with a view of using the second samples to validate the predictions made by the first sample.

Three regression procedures were used to build models on each data set for comparison purposes and to test their predictive abilities on each unique data set. Our criterion for test was the validation MSE(s) which was obtained by predicting each dependent variable value in the validation sample and averting the squared errors.

In bootstrapping approach, validation of the regression models was achieved by adding the resample residuals unto the least square regression fit, holding the regression design fixed. The least squares estimate from each bootstrap samples was obtained and the validation statistic of interest such as R^2 than stepwise regression in data splitting techniques.

Bootstrapping seems to work better than stepwise regression in validating regression models. In the simplest form of bootstrapping, instead of repeatedly analyzing subset of the data, analyst repeatedly analyse subsamples of the data and each subsamples is a random sample with replacement from the full sample. Bootstrapping allows us to gather many alternative version of the single statistic that would ordinarily be calculated from one sample and compute the statistical interest for each of the data sets.

V. Conclusion

Bootstrapping the model fitting process is a much better way to get unbiased estimates of model performance without sacrificing sample size, here we are validating the full n- subject model.

The most important advantage of bootstrapping in validating regression models over data splitting techniques are to need smaller sample than data splitting techniques and its practical performance is frequently much better in the sense that the risk of over fitted models are reduced as it gives a better and stable value of R^2.

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.

In a nutshell in validating regression models, bootstrapping procedures are useful than data splitting in the following situation:

(i) When the theoretical distribution of a statistic is complicated or unknown.
(ii) When the sample size is insufficient for straightforward statistical inference.
(iii) When power calculations have to be performed and a small pilot sample is available.

References Références Referencias

1. Cheng, R.C.H., 2004. *Resampling methods*. Working Paper, Southampton University, Southampton, UK,Chapter 16 in Elsevier Handbooks in Operations Research and Management Sciences.
2. Davison , A.C and Hinkely, D.V. 1997, Bootstrap methods and their application. Cambridge University Press.
3. Deflandre, D., Kleijnen, J.P.C., 2003. Statistical analysis of random simulations: *Bootstrap tutorial*. Simulation News Europe (38/39), 29-34.
4. Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife and cross-validation. Amer. Statistician 37, 36-48.
5. Efron, B., Tibshirani. R.,(1993). Introduction to the Bootstrap. Chapman and Hall, London.
6. Efron, B., (1983), Estimating the error rate of a Prediction rule improvement on improvement on crossvalidation" Journal of the American Statistical Association.
7. Feng, C.-X., Yu, Z. and Kusiak, A., Selection and validation of predictive regression and neural networks modeling data from designed experiment. IIE Trans., 2004 (under third review).
8. Friedman, L.W., Friedman, H.H., 1995. Analyzing simulation output using the bootstrap method. Simulation 64 (2), 95-100.
9. Gershenfeld, N., The Nature of Mathematical modeling, 1999 (Cambridge University Press: Cambridge, UK).
10. Hall , P. and Wilson S.R. 1991. Two guidelines for bootstrap hypothesis testing. Biometrics, 20, 231-246.
11. Kleijnen, J.P.C., Cheng, R.C.H., Bettonvil, B., 2001. Validation of trace-driven simulation models: Bootstrapped tests. Management Science 47 (11), 1533-1538.te.
12. Lazante, J.R., 1984: Strategies for assessing skill and significance of screening regression models with

emphasis on Monte Carlo techniques. J. Climate Appl. Meteor., 23, 1454-1458.
13. Mallows, C.L. (1997) C_P and prediction with many regressors: comments on Mallows (1995). Technometric, 39(1), 115-116.
14. Neuman,C.J., M.B.Lawrence and E.L, Caso, 1977; Monte Carlo significance testing as applied to statistical tropical cyclone prediction models. J.App., Meteor., 16,1165-1174.
15. Olatayo T.O (2010). On Truncated Geometric Bootstrapping Method for Stochastic Time Series Process. Unpublished Ph.D Thesis. University of Ibadan, Nigeria.
16. Oredein A.I (2011) On Validating regression procedures with bootstraps and data splitting techniques. Unpublished M.Sc.Thesis. Olabisi Onabanjo University, Ago-Iwoye, Nigeria.
17. Rencher,A.C., and F.C.Pun 1980; Inflation of $R sup 2$ in best subset regression. Technometrics, 22, 49-53.
18. Shapiro, L.J.,1984: Sampling errors in statistical models of tropical cyclone motion: A comparison of predictor screening and EOF techniques. Mon, Wea. Rev., 112, 1378-1388.
19. Snee, R.O, (1977), Validation of Regression Models, Methods and Examples, Techniques 19. 415 – 428.
20. Willemain, T.R., Bress, R.A., Halleck, L.S., 2003. Enhanced simulation inference using bootstraps of historical inputs. IIE Transactions 35(9), 851-862.