Multiple Regression Model Fitted for Rice Production Forecasting in Nepal: A Case of Time Series Data

Chuda Prasad Dhakal

Submitted: 15 August 2018; Accepted: 12 September 2018

Published online: 26 September 2018 **DOI: http://dx.doi.org/10.3126/njs.v2i0.21157**

ABSTRACT

Background: Fitting a multiple regression model is always challenging and the level of difficulty varies according to the purpose for which it is fitted. Two major difficulties that arise while fitting a multiple regression model for forecasting are selecting 'potential predictors' from numerous possible variables to influence on the forecast variable and investigating the most appropriate model with a subset of the potential predictors.

Objective: Purpose of this paper is to demonstrate a procedure adopted while fitting multiple regression model (with an attempt to optimize) for rice production forecasting in Nepal.

Materials and Methods: This study has used fifty years (1961-2010) of time series data. A list of twenty-one predictors thought to impact on rice production was scanned based upon past literature, expert's hunches, availability of the data and the researcher's insight which left eleven possible predictors. Further, these possible predictors were subjected to family of automated stepwise methods which left five ‘potential predictors’ namely harvested area, rural population, farm harvest price, male agricultural labor force and, female agricultural labor force. Afterwards, best subset regression was performed in Minitab Version 16 which finally left three 'appropriate predictors' that best fit the model namely harvested area, rural population and farm harvest price.

Results: The model fit was significant with $p < .001$. Also, all the three predictors were found highly significant with $p < 0.001$. The model was parsimonious which explained 93% variation in rice production with 54% overlapping predictive work done. Forecast error was less than 5%.

Conclusion: Multiple regression model can be used in rice production forecasting in the country for the enhanced ease and efficiency.

Keywords: Cross-validation, forecast error, model selection, predictive work done, predictors, variable selection.

Address correspondence to the authors: Institute of Agriculture and Animal Sciences, Rampur Campus, Chitwan, Nepal.
Email: chudadhakal@iaas.edu.np

Copyright & License @ Central Department of Statistics, TU, 2018

INTRODUCTION

Fitting a multiple regression model is always challenging and the level of difficulty varies according to the purpose for which it is fitted. Two major difficulties that arise while fitting a multiple regression model for forecasting are: i) selecting 'potential predictors' from numerous possible variables to influence on the forecast variable and, ii) investigating the most appropriate model with a subset of the potential predictors. And, scientists have mentioned that the methods used for variable selection are not free from the debates. For this reason, careful use of them is an essential factor.

Karim (n.d.) reveals that where there is no clear-cut theory, the problem of selecting predictors for a regression equation becomes quite important. To this milieu, guidelines given by Makridakis, Wheelwritht, and Hyndman(1998, p.276) are: i) experts and other knowledgeable people in the general area counsel about which predictors should be incorporated and which should not be of greater importance and ii) the data on the respective predictors are available. Other major principle of variable selection is, model should be parsimonious. According to Stephanie (2018), parsimonious models are simple models with great explanatory predictive power. They explain data with a minimum number of predictor variables. No more variables than necessary are used. Accordingly, the author tells that, parsimonious models have optimal parsimony, or just the right amount of predictors needed to explain the model well.

According to Draper and Smith (1998), commonly in a model, one likes to include as many predictors as are necessary or, as few predictors as possible to make less cost, easy monitor and to keep the variance of the predictors reasonably small. Fearfully, both of which careless about fitting a best regression equation. To obtain a best regression equation, Draper and smith (1998, p.327) has mentioned i) all possible regression using three criteria: R^2, s^2 and the Mallows C_p; ii) best subset regression using R^2 ˒ R^2 (adjusted), and $C_{p;}$ iii) Stepwise regression; iv) backward elimination; and v) some variations on previous methods. Also, in the same context, authors have mentioned that a great deal of personal judgment will be a necessary part of any of the methods discussed. After 'potential predictors' are fixed, there are numerous methods almost with the same amount of controversies and the debates to locate a best model. As 'appropriate predictors' are the ultimate predictors that locate the best model out of all possible models that could be fit with the 'potential predictors', this is always a challenging task. Sometimes the model which consists of all investigated potential predictors becomes the optimum model, whereas in many other times this might not be the case. This issue of selecting the best combination of the potential predictors to fit a model comes under the theory of model selection criteria. In this line, Taylor and Cobb (2004) reveals this to be an “unsolved” problem in statistics where there are no magic procedures to get the “best model” as is said chiefly.

Other methods we can find in common are, Akaike's Information Criteria (AIC), and Bayesian Information Criteria (BIC). These procedures are valuable for quickly producing regression equations worth for further consideration. Bowerman, O'Connell, and Koehler(2005) have mentioned that different computer packages carry out regression with these methods with slight variation. However, no matter what effort is done to select a best single model, at the end with best subset regression method, there sometimes is the danger that different criteria may lead to different "best" models. In such situation according to Draper and Smith (1998, p.328) researcher's common sense, basic knowledge of the data being analyzed; and, according to Makridakis et al. (1998, p.276) certain amount of creativity are applied throughout the variable selection procedure. Besides these, to locate the best model, Mundry and Nunn (2009); Hyndman and Athanasopolous (2014) on stepwise methods, and Pardoe(2015) on best subset regression, are also considered. Last but not the least, to its end, this paper explains the various characteristics of the model fitted with all the attempts mentioned above.

MATERIALS AND METHODS

For forecasting Methods and Principles, Makridakis et al. (1998); Hyndman and Athanasopulos (2014); Bowerman et al. (2005) have comprehensively described how a multiple regression model is fitted (making an attempt to optimize) for forecasting. This study is primarily based upon these principles. Fifty years (1961-2010) time series data of eleven 'possible predictors' suggested by the experts and the past research experience, were acquired collectively from the office of the Department of Hydrology and Metrology (DHM) and, the respective websites of International Rice Research Institute (IRRI), Ministry of Agriculture Development (MOAD) and, Food and Agriculture Organization (FAO).

Out of the fifty years (1961-2010) time series data considered for the study; first thirty-five years (1961-1995) data were used to make the model whereas, the last 15 years (1996-2010) 30% of the total data, called the test data were put aside and used later for cross validating the model. The test sample according to Hyndman and Athanasopulos (2014) should not be less than the 20% of the total sample size. Appropriate predictors were selected in various distinct stages. First hand collection was a list of 21 'prompted predictors'. Makridakis et al. (1998, p.276) have suggested long list of the predictor variables that impact on the forecast variable are prepared based on i) hunches of experts and other knowledgeable people, ii) availability of the data, and iii) practical time and constraints. With this foundation and some other creativity and brainstorming, this list of 'prompted predictors' was better organized and refreshed founding upon researcher's insight, past literature, experts' hunches, and availability of the data. This led to have a list of 11 'possible predictors'. Thirdly, from this list of 'possible predictors', unimportant ones were eliminated using the family of stepwise methods: forward selection, backward elimination and stepwise (forward and backward) selection.

Forward selection starts picking up the predictors one by one, given that alpha to enter (in this case 0.25) is pre-specified. It picks up the strongest predictor first and at the last the weakest predictor which has at least the specified amount of significance to influence the forecast variable. Backward elimination, first enters all predictors in the model. It then starts sorting out with the weakest predictor eliminated one by one in every step. Regression is run until there would be any predictor which won't meet the pre-specified criterion of alpha to remove (in this case 0.1). Stepwise method is combination of the forward and the backward approach. This moves ahead with forward approach at the same time having a backward look at it. It therefore has both levels of significance (in this case, alpha to enter = 0.05, alpha to remove = 0.1) are specified earlier. For instance, having a check on, if the first variable included with forward selection is quite unnecessary in the presence of other predictors or that the first variable deleted in backward stepwise could be the first variable included in forward selection. This approach (forward and backward) of variable selection takes care about which and performs accordingly. And once, potential predictors are identified they are subjected to the model selection procedure (in this case the best subset regression in Minitab 16) which comes out with the final subset of the predictors called the 'appropriate predictors' to best fit the model.

Software used were SPSS 20, Minitab 16 and STATA 12. As far as it was practicable the variable selection procedure were carried out in the software considered to minimize the error in selection due to slight variations among the software.

RESULTS

The model

The followings were the prompted predictors for the model.

harvested area, rice yield, rice consumption per capita, total consumption milled rice, stock exchange-milled rice, seed consumption, number of released and registered varieties, export quantity, import quantity, export price, farm harvest price, irrigated rice area, fertilizer consumption, number of tractors, numbers of harvesters, number of threshers, rural population, male labor force in agriculture, female labor force in agriculture, annual mean rain fall and annual mean temperature

A careful scan reduced this into a list of eleven 'possible predictors' (Table 1).

Table1. List of the possible predictors.

SN	Predictors
1	Harvested area
2	Farm price at harvest
3	Fertilizer consumption
4	Number of tractors
5	Seed consumption
6	Annual mean rainfall
7	Annual mean temperature
8	Number of registered and released varieties
9	Rural population
10	Male labor force in agriculture
11	Female labor force in agriculture

This list of 'possible predictors' was then set for stepwise methods of variable selection. For better outcome, to eliminate any variation between software to software, every step of variable selection was conducted in both SPSS and Minitab, and commonly eliminated variables were removed from the selection procedure. In the meantime forward selection [with 0.25 alpha to enter (deliberate that no possible predictor was eliminated)], and backward selection [with 0.1 alpha to remove (deliberate that no unimportant variable was included)] were conducted one after another. Remaining seven predictors were then subjected to stepwise (forward and backward) selection with alpha to enter = 0.05, and alpha to remove = 0.1. For this, commonly eliminated predictors both by SPSS and Minitab were *Fertilizer consumption* and *Seed consumption* (Table 2).

Table 2. SPSS/Minitab stepwise selection (alpha to enter =0.05, alpha to remove = 0.1).

SN	Predictors
1	Harvested area
2	Farm price at harvest
3	~~Fertilizer consumption~~
4	~~Seed consumption~~
5	Rural population
6	Male labor force in agriculture
7	Female labor force in agriculture

Table 3 shows the list of potential predictors.

Table 3. List of the 'potential predictors'**.**

SN	Predictors
1	Harvested area
2	Rural population
3	Farm harvest price
4	Male labor force in agriculture
5	Female labor force in agriculture

Best subset regression was then conducted in Minitab [*Stat > Regression > Best Subsets*]. The output (Table 4) locates the 'appropriate predictors' to fit the ultimate 'best model' that could be fit from the available set of 'potential predictors'.

Table 4. Best subset regression.

No. of Variables	R^2	Adj.R^2	Mellows C_P	SE	H Area	R Pop^n	FHP	MLF	FLF
1	81.1	80.5	55.3	227.14	X				
1	73.5	72.7	89.8	268.70			X		
2	84.9	83.9	40.0	206.30	X		X		
2	82.2	81.1	52.3	223.84	X				X
3	93.3	92.6	3.6	139.60	X	X	X		
3	90.2	89.2	17.9	169.13	X			X	X
4	93.4	92.5	5.1	140.77	X		X	X	X
4	93.3	92.4	5.6	141.76	X	X	X	X	
5	93.6	92.6	6.0	140.47	X	X	X	X	X

Adj. = Adjusted; SE = Standard error; H area = Harvested area; Pop^n = Population; FHP= Farm harvest price; MLF = Male labor force; FLF = Female labor force

The fourth column in (Table 4) shows Mallows Cp statistic. In this column, the Cp value 3.6 for the model consisting of the predictors *harvested area*, *price at harvest* and *rural population* is minimum among all other Cp values. Both criteria for model to be the best model; smallest Cp value and the Cp value (3.6) is less than the number of parameters (4) in the model are satisfied. This reveals that the model is 'the best' model among all other possible models that could be fit by the available set of 'potential predictors'.

Also, the model has minimum standard error $s(= 139.60)$ and the highest value for adjusted R-square (= 92.6) among all other models. These are the other additional criteria for a model to be a best model. This signified the model with the specified predictors (i.e. *harvested area*, *price at harvest* and *rural population*) was the best model. Finally, the 'appropriate predictors' therefore were investigated (Table 5).

Table 5. List of appropriate predictors**.**

SN	Predictor
1	Harvested area
2	Rural population
3	Price at harvest

The final model

Model investigated was

Production = -1619 + 5.26 × (**Harvested area)** – 0.239 × (**Rural population**) + 0.321 × (**Price at harvest**)

Yet, this was a crude model which needed multiple regression assumption testing including multiollinearity issue, outliers and influential points etc. and cross validation of the model. In the form of case studies each of these issues has already been attempted, accomplished and published. Accordingly as, assumption testing (Dhakal, 2017); issues of outliers and influential points (Dhakal, 2017); cross validation of the model (Dhakal, 2017) and forecast accuracy measures (Dhakal, Sthapit, & Devkota, 2014). The model was over all significant, F (3, 34) = 147.70, p < 0.001, and all the predictors in the model too, were found to be highly significant, p < 0.001.

Variance explained in the forecast variable was, (R^2) = 93%. Unique contributions due to the predictors were (19%, 12% and 8% totaling 39%). This justified the combination of the predictors to be highly effective with 54% overlapping predictive work. Adjusted R-squared (92%) and Predicted R-squared (90%); signify that the model was neither susceptible with sampling fluctuation nor was it over fitting. Forecast error below 5% is an indicator to validate the model for forecasting. Also, the model was found superior to Naïve forecast method and the value of Tracking Signal (-6.97) for the model, was not much contradictory to the general rule of thumb -4<TS<4. However, this signified the model was not perfect and can be improved.

Physical significance of the investigated model

Increased inputs for the predictors *harvested area* and *price at harvest* have positive significance and therefore expected to increase rice production. For each 000' hectare increase in *harvested area*, 5.26 thousand (t) and for every per unit (NRs/t) increase in *price at harvest*, 0.321 thousand (t) of rice is expected to be increased. However, *rural population* has negative coefficient. This conveys negative significance in rice production. For every additional (000) *rural population* in the production process, 0.239 thousand (t) rice is expected to be diminished.

Inclusion of the predictor *area harvested* with its respective coefficient, in the model indicates low level technology in agriculture system. It signifies, production is increased if area harvested is increased. Had there been other predictor /s instead, for instance *improved variety* etc. we could have said, technology enhanced production. Likewise, as compared to other seemingly prominent predictors such as *fertilizer consumption*, *annual rainfall,* etc. *price at harvest* included in the model has a significant implication. This indicates, farmers encouraged at their farm giving them good price for their produce, have a positive effect in producing more rice in the country. However, the case of negative coefficient of the third predictor, *rural population* included in the model, appears to contradict the common belief, 'more labor is employed more will be the production.' This agrees with the theory of 'law of marginal diminishing return.' Labor force results in decreased production when it continues to go further from its peak, keeping other factors constant. This suggests that country would have to better plan its labor force in agriculture. A possible shifting is needed perhaps in the industrial sector to rise country's revenue.

CONCLUSION

This paper demonstrated a fit of multiple regression model with time series data (1961-2010) for rice production forecasting Nepal. Variable selection, which by its nature is a complicated process, was the prime challenge to come to its end. However, different facilities have now been invented and therefore was made possible. For this case stepwise methods and the best subset regression were used. But, because these methods were no more debate less, cautions were taken to cope with the variations between such facilities. For this reason, also the Cp criterion was employed with enough creativity and the research experience. The model so fitted was found highly significant, showing its potential to be used in real life by the concerned planners and the policy makers. It therefore, for this fit, was concluded that multiple regression model could be scientifically used in forecasting, and the related stakeholders could be benefited from this model especially for the enhanced ease, and efficiency for rice production forecasting in the country. In addition to this, this study to its limit, has spread the relevant factors and the idea about them to contribute in rice production at national level in the present scenario.

Last but not the least, this study was carried out through the educational perspective, hence forth, future work might consider increasing the precision of the model in any aspects like increasing its reliability, validity etc. A forecast model using the same data set but with different approach could be developed and the model's efficiency could be compared. Further the various results obtained throughout the study, has created an opportunity to replicate the whole study or part/s of it and to move forward for improved confidence of the model.

CONFLICT OF INTEREST

The author confirms there is no conflict of interest.

ACKNOWLEDGEMENTS

The writers of all papers and books who are cited are all indebted. In addition, thanks also go to all writers of books, papers, etc. which were helpful in any means while preparing this research article.

REFERENCES

Bowerman, B. L., O'Connell, R. T., & Koehler, A.B. (2005). Forecasting, time series, and regression: An applied approach. (4th ed.). 10 Davis Drive Belmont, CA 94002, USA: Thomson Brooks/Cole.

Dhakal, C. P. (2017). A naïve approach for comparing a forecast model. International Journal of Thesis Projects and Dissertations, *5*(1), 1-3. Retrieved from www.researchpublish.com

Dhakal, C. P. (2017). Dealing with outliers and influential points while fitting regression. Journal of Institute of Science and Technology, *22*(1), 61-65.

Dhakal, C. P. (2017). Testing assumptions in linear regression models: A case study. International Journal of Basic and Applied Agricultural Research, *14*(3), 40-44.

Dhakal, C. P., Sthapit, A. B., & Devkota, N. R. (2014). Forecast accuracy measure: An overview. Samajiki Sandarsh, *2*(4), 121-127.

Draper, N. R., & Smith, H. (1998). Applied regression analysis. (3rd ed.). USA: John Wiley & Sons, Inc.

Hyndman, R. J., & Athanasopoulos, G. (2014). Forecasting: Principles and practice. Retrieved from http://otexts.com/fpp/ Accessed on 02.05.2014

Karim, E. M. (n.d.). Selection of best regression equation by sorting out variables. Institute of statistical research and training: University of Dhaka Bangladesh. Retrieved from http://www.angelfire.com/ab5/get5/selreg.pdf

Makridakis, S., Wheelwritht, S. C., & Hyndman, R. J. (1998). Forecasting methods and application. (3rd ed.). John Wiley & Sons, Inc.

Mundry, R., & Nunn, C. L. (2009). Stepwise model fitting and statistical inference: Turning noise into signal pollution. American Naturalist, *173*(1), 119-123. doi:10.1086/593303

Pardoe, I. (2015). Best subset regression. PennSate, Stat 501, Regression Methods. Retrieved from 3.11.2015 from https://onlinecourses.science.psu.edu/stat501/node/330

Stephanie. (2018). Parsimonious model: definition, ways to compare models. Statistics How To. Retrieved from http://www.statisticshowto.com/parsimonious-model/

Taylor, J., & Cobb, K. (2004). Model selection. Statistics 262: Intermediate biostatistics. Retrieved from http://statweb.stanford.edu/~jtaylo/courses/stats262/spring.2004/notes/week9.pdf

Reference to this paper should be made as follows:

Dhakal, C. P. (2018). Multiple regression model fitted for rice production forecasting in Nepal: A case of time series data. *Nep. J. Stat., 2*, 89-98.