INT. J. BIOL. BIOTECH., 9 (4): 399-404, 2012.

FORECASTING OF AGRICULTURE PRODUCTION THROUGH BIASED ESTIMATORS

Mansoor Ahmed, Muhammad Hanif* and Nasir Jamal

Department of Mathematics and Statistics, Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi, Pakistan, E-mail: mhpuno@hotmail.com

ABSTRACT

The aim of this paper is to study the biased estimators in forecasting the agriculture production. The main idea relies on using the ridge regression estimators to forecast the groundnut production. The motivation behind the study is to use the ridge regression estimators is that it overcomes the problem of Multicolinearity that often occurs in the time series data. Our simulation study reveals that the forecasting through ridge estimators found much better than the forecasting of groundnut production by using time series econometric model.

Key words: Autoregressive integrated moving average, Biased estimator, Biasing constant, Ridge estimator, Variance inflation factor.

INTRODUCTION

Ridge regression is one of the popular methods for the solutions of problem related to Multicolinearity. This method is the modification of the least squares method that allows biased estimators of the regression coefficients. Therefore, these biased estimators are preferred as they have a larger probability of being close to the true parameters. In presence of multicollinearity, selection of ridge parameter plays an important role, because the idea of that adding a small constant to the diagonal elements of the matrix $X X'$ will improve the conditioning of a matrix has been recognized by numerical analysis, because this would dramatically decrease its 'condition number (Jamal and Muhammad, 2007). Complete elimination of multicollinearity is not possible but its degree can be reduced by adopting ridge regression, principal components regression. For further study about the Ridge regression see, e.g., Batah and Gore (2008), Batah, Gore and Özkale (2009), Fuller (2002), Rao and Singh (1997), Özkale and Kaciranlar (2007) and the reference therein.

Multicollinearity is a case of multiple regression in which the predictor variables are themselves highly correlated. If the goal is to know how different explanatory variables X related with explained variable Y then multicollinearity is a big problem. Multicollinearity is a matter of degree, not a matter of presence or absence. In presence of multicollinearity the ordinary least squares estimators are imprecisely estimated (Madala and Kajal, 2007). The the presence of multicollinearity among the predicted variables can be observed through correlation matrix, variance influence factor (VIF), eigenvalues of the correlation matrix and auxiliary regression. Koutsoyiannis (2007) studied that the degree of the multicollinearity becomes more severe as $\mathbf{X}\square/$ approaches zero. If the goal is to understand how the different predicted variables impact on the dependent variable, then multicollinearity causes a serious problems like P values may mislead and the confidence intervals on the regression coefficients may very wide. Usually the ordinary least squares estimator is unbiased estimator but in the presence of multicollinearity ordinary least squares estimators could becomes unstable due to their large variance, which leads to poor prediction.

MATERIAL AND METHODS

All important factors which may cause effect the production of groundnut such as rainfall, temperature, chemical fertilizers, number of ploughs and area sown are considered for the development of the forecasting model of groundnut production. The parameters which determine the climate of locality are rainfall, temperature, humidity, air pressure, snowfall, winds, light, clouds and storms. Out of these factors, the rainfall and temperature are the most critical factors for modeling area and production of any crop. Three times period of rainfall and temperature maximum and minimum data from the month of April to September is considered for groundnut production model. Temperature affects crop from seed germination to harvesting and even further during storage. In general, photosynthesis and respiration take place slowly at low temperature. Mean Maximum temperature for the month of April and May is taken as independent variable for the development of groundnut production forecasting model.

The data on the number of ploughs of groundnut fields is selected for groundnut yield estimation survey is taken as an independent variable in groundnut production model. The data on this variable is taken from groundnut yield estimation survey forms of Attock district, which is selected for this study. Filled in groundnut yield estimation survey forms for last twenty one years from 1990-2010 were collected from District office of crop reporting service, Agriculture Department Attock. These forms were inspected of 25 sample villages of district Attock. The data on use of Urea, DAP and number of ploughs was compiled from yield estimation survey forms. The variables used in the study are defined as under:

Y = Production

X1 = Area
X2 = Ploughs
X3 = Urea
X5 = Rainfall in April and May
X6 = Rainfall in June
X7 = Minimum Temperature in July
X8 = Minimum Temperature in August

Detection of Multicollinearity

Multicollinearity usually arises when one or more of the regressors are exact or approximately linear combinations of the regressors. It is often detected by observing the value of R2 and low value of t-ratios. Another way to find the multicollinearity is simply the inspection of the off-diagonal elements in $X \square X$ **.**

Auxiliary Regression

One way of finding out which *X* variable is related to other *X* variables is to regress each *Xi* on the remaining *X* variables and compute the corresponding *R2*which is designate as *Ri2* each one of these regressions is called an auxiliary regression (Griffiths *et al*., 2001). In general linear regression model, if $E(\varepsilon i\ 2) \neq \sigma 2$ then the heterosedasticity is said to be present and the assumption of constant variance of error term is violated. The Spearman Rank Correlation Test is used to detect the heteroscedasticity. This is very simple test for heteroscedasticity which can be applied to both small and large data (Green, 2012). Autocorrelation correlation is defined as correlation between numbers of the same series of observations ordered in time in case of time series data or space in case of cross section data. In regression context, autocorrelation means correlation between εi across observations (Butt, 1999).

Ridge Trace

A number of procedures have been developed for obtaining biased estimators of regression coefficients. One of these procedures is Ridge Trace (Hoerl and Kennard, 1970). A commonly used method of determining the biasing constant is based on ridge trace and variance inflation factor. The ridge trace is simultaneous plot of the values of the *p-1* estimated ridge standardized regression coefficients for the estimated regression coefficients for the different values of C, usually between 0 and 1 (Kutner *et al*., 2005*)*. Extensive experience has indicated that the estimated regression coefficients may fluctuate widely as biasing constant is changed slightly from 0, and some may change Signs. Gradually, however, these wide fluctuations cease and the magnitudes of the regression coefficients tend to move slowly towards zero as biasing constant is increased further. At the same time, the values of variance inflation factor for regression coefficients on different biasing constants biasing constant tends to fall rapidly as biasing constant is changed from zero, and gradually the variance inflation factor values also tend to change only moderately and as biasing constant increases further. Usually the plot of the estimated ridge standardized regression coefficients becomes stabilize and the variance inflation factor value for each *X* variable becomes approximately equal to one for the same value of biasing constant (Marquardt, 1970).

ARIMA Procedure

The ARIMA procedure analyzes and forecasts equally spaced univariate time series data, transfer function data, and intervention data using the Autoregressive Integrated Moving-Average or autoregressive moving-average model. An model predicts a value in a response time series as a linear combination of its own past values, past errors which is also called shocks or innovations, and current and past values of other time series (Ansley and Newbold, 1980).

Forecasting

The forecasting performance of an econometric model is tested on the basis of the difference between the predicted values of the dependent variable and the actual values of the dependent variable. The smaller the difference between them, batter the forecasting performance of the model. Various measures have been proposed for evaluating the forecasting performance of econometric models. These measures are Root Mean Square Error, Mean

Absolute Error, Mean Absolute Percent Error, Theil's Inequality Coefficient and The Janus Quotient (Raymond, 1975).

RESULTS AND DISCUSSION

Tests for deduction of multicollinearity, heteroscedasticity, and autocorrelation are conducted on the data. Results of these tests, their clarifications and remedial measures are as follow:

Deduction of Multicolinearity

Auxiliary regressions are fitted to see which explanatory variable is linearly related to other explanatory variables. The value of Fi are calculated and having the F distribution. The values of Fi show that all the explanatory variables are linearly related to other Xi's. All the results are given in Table 1.

Table 1. Results of auxiliary regressions

Variables	X1	X2	X3	X4	X5	X6	X7	X8
R^2	0.97	0.79	0.99	0.99	0.98	0.99	0.97	0.96
F_i	86.78	9.51	49.50	24.50	18.80	14.16	16.54	8.92

On the other hand off-diagonal elements of matrix show that all regressors *Xi* are linearly related with each other.

$$XX = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ 1 & & & & & & & \\ 0.421 & 1 & & & & & & \\ 0.819 & 0.693 & 1 & & & & & \\ 0.824 & 0.672 & 0.997 & 1 & & & & \\ 0.887 & 0.698 & 0.941 & 0.928 & 1 & & & \\ 0.759 & 0.740 & 0.880 & 0.858 & 0.965 & 1 & & \\ 0.945 & 0.973 & 0.940 & 0.928 & 0.892 & 0.927 & 1 & \\ 0.906 & 0.985 & 0.923 & 0.907 & 0.988 & 0.902 & 0.829 & 1 \end{bmatrix}$$

Detection autocorrelation and hetroscedasticity

Von Neumann Ratio is calculated to check the presence of autocorrelation in the data which is given Vc = 3.28, whereas Vt = 1.38, which shows absence of autocorrelation in the data. By using Spearman Rank Correlation the computed value of t = 0.8754 which shows that there is no evidence of systematic relationship between the explanatory variables and hence there is no heteroscedasticity.

Remedial Measures of Multicolinearity

Ridge Regression is used to remove Multicollinearity from the data. The ridge regression coefficient for $0 < C < 1$ are given below:

From the standardized Ridge Regression coefficients which are shown in Table 2 are stabilized between C = 0.4 and C = 0.5. We take its average value i.e C = 0.4500 as a biasing constant (Pasha and Shah, 2004). The standardized Ridge Regression coefficients at C = 0.4500 are {0.2305 , 0.0652, 0.1134, 0.1170, 0.0712, 0.1611, -0.0661, - 0.1577} and retransformed Ridge Regression coefficients are {0.11, 17.18, 0.99, 1.97, 0.28, 0.70, -13.18, -104.72}. Therefore the groundnut production model based on agricultural data for the period 1990 to 2005 as under:

$$Y = 4186.22 + 0.11\ X_1 + 17.18X_2 + 0.99\ X_3 + 1.97X_4 + 0.28X_5 + 0.7X_6 - 13.18X_7 - 104.72X_8$$

The increase in area of sowing of groundnut crop increases the production of the crop. The use fertilizers and number of Ploughs play a significant role to enhance the production of groundnut crop. Good rains in the months of April, May and June contributed to increase the production of groundnut crop. Mean Minimum temperature in the months of July and August turn out as a significant but negative effect on groundnut production in the District because groundnut crop required good rain at that time.

Table 2. Ridge estimators for different values of biasing constant

C	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
0.1	0.3679	0.0831	0.0893	0.1038	0.0185	0.2038	-0.0015	-0.1735
0.2	0.3022	0.0712	0.1039	0.1118	0.0426	0.1836	-0.0335	-0.1729
0.3	0.2649	0.0672	0.1099	0.1153	0.0576	0.1721	-0.0511	-0.1667
0.4	0.2401	0.0656	0.1127	0.1167	0.0675	0.1643	-0.0620	-0.1605
0.5	0.2221	0.0649	0.1139	0.1171	0.0742	0.1582	-0.0694	-0.1550
0.6	0.2083	0.0646	0.1142	0.1168	0.0789	0.1533	-0.0744	-0.1502
0.7	0.1971	0.0644	0.1140	0.1162	0.0822	0.1491	-0.0780	-0.1459
0.8	0.1879	0.0642	0.1135	0.1154	0.0846	0.1454	-0.0806	-0.1421
0.9	0.1801	0.0640	0.1128	0.1144	0.0863	0.1421	-0.0825	-0.1387

Fitting of Time Series Model

For the purpose of comparison with the ridge estimates, time series analysis has been made on the agricultural data. The results are shown in Table 3 and Table 4.

Table 3. Estimated values of coefficients and their standard errors for stationarity data

Variable	Coefficient	Std. error	t-Statistics	Prob.
Yt-1	-1.17	0.16	-7.18	0.00
C	127.48	1007.69	0.12	0.90

Table 4. Estimated values of coefficients and their standard errors of Time series model

Variable	Coefficient	Std. error	t-Statistics	Prob.
AR(1)	0.67	0.15	4.31	0.0001
MA(1)	-0.96	0.06	-16.58	0.0000

Table 3 shows that t value of Yt-1 co efficient is -7.1897 and this value in absolute terms is much higher than even 1 percent critical τ value of -3.6155 which suggesting that data is stationary at its 1st difference. Table 4 Shows that time series econometric model is significant for both AR (1) and MA (1) , also it is we defined earlier that data is stationary at 1st difference so ARIMA model is applicable.

Forecasting Performance of Model

Forecasting production of groundnut crop for the period 2006 to 2010 from Ridge Regression model is calculated and the percentage difference between the actual and forecasted values of groundnut production is given in Table 5.

Table 5. Forecasted groundnut production from Ridge Regression

Sr. No	Year	Production		Percentage
		Actual	Forecasted	Increase/Decrease
1.	2006	1511.1	1477	2 2.24
2.	2007	1329.2	1238.2	6.81
3.	2008	1641.1	1593.2	2.91
4.	2009	559.6	863.9	-4.36
5.	2010	441.0	668.7	-1.63

Forecasted production of groundnut crop for the period 2006 to 2010 from ARIMA model is calculated and the percentage difference between the actual and forecasted values of groundnut production is given in Table 6.

Table 6. Forecasted groundnut production from ARIMA model

Sr. No	**Year**	**Production**		**Percentage**
		Actual	**Forecasted**	**Increase/Decrease**
1.	2006	1511.1	1347.00	10.8596
2.	2007	1329.	2 829.29	37.6098
3.	2008	1641.1	1450.90	11.5897
4.	2009	559.6	1056.245	-88.7500
5.	2010	441.0	760.15	-72.3696

Table 5 and 6 show the percentage difference of forecasted production from actual production by using Ridge Regression model is significantly low as compared to difference in actual and forecasted production by using ARIMA model.

Various measures like Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percent Error (MAPE), Theils Inequality co-efficient and Janus Quotient are computed for evaluating the performance of econometric model. Computed values of these measures are given in Table 7, which show that the performance of model fitted by Ridge Regression is far better than the time series model. Therefore Ridge Regression model is preferred over ARIMA model to forecast the groundnut production.

Table 7. The forecasting performance of an econometric model

S. No	**Method**	**Ridge**	**ARIMA**
1.	Root Mean Square Error	47.33	5693.19
2.	Mean Absolute Error	37.86	4190.55
3.	Mean Absolute Percentage Error	0.03	28.50
4.	Theil's Inequality co-efficient	0.019	0.16
5.	Janus	Quotient	18.89 89.78

REFERENCES

Ansley, C. and P. Newbold (1980). Finite Sample Properties of Estimators for Autoregressive Moving Average Models, *Journal of Econometrics*, 13: 159-183.

Batah, F. and S. Gore (2008). Improving precision for jackknifed Ridge type estimation. *Far East Journal of Theoretical Statistics*, 24: 157–174.

Batah, F., D. Gore and M. Özkale (2009). Combining Unbiased Ridge and Principal Component Regression Estimators. *Communication in Statistics - Theory and Methods*, 38: 2201–2209.

Butt, A. R. (1999). *Least Squares Estimation of Econometric Models, Department of Economic*, (1st Ed.), National Book Foundation, pp 63-123.

Fuller, W. A. (2002). Regression estimation for survey samples. *Survey Methodology*. 28: 5–23.

Green, W. H. (2007). *Econometric Analysis*, (7th Ed.) New Jersey, Prentice Hall, pp 212-227.

Griffiths, W. E., R. C. Hill and G. G. Judge (2001). *Learning and Practicing Econometrics*, John Wiley & Sons, Inc., New York, 1st Edition, pp 431-39 and pp 483-500.

Hoerl, A. E. and R. W. Kennard (1970). Ridge Regression: Biased Estimation for Non-orthogonal Problems, *Technometrics*, 12: 55-67.

Jamal, N. and Q. R. Muhammad (2007). Ridge Regression; A tool to Forecast wheat area and production, *Pakistan Journal of Statistics and Operation Research*, 3: 125-134.

Kibria, B. M. (2003). Performance of Some Ridge Regression Estimators. *Commun. Statist – Simulation and computation*, 32: 419–435.

Koutsoyiannis, A. (2007). *Theory of Econometrics*, (2nd Ed.), Oxford University press, Ontario, pp 233-253.

Kutner, M. H., C. J. Nachtsheim, J. Neter and W. Li (2005). *Applied Linear Statistical Models*, (5th Ed.), McGraw-Hill, Inc., New York, pp 411-414.

Maddala, G. S. and L. Kajal (2007). *Introduction to Econometrics*, (4th Ed.), John Wiley & Sons, Inc., New York, pp 456-493.

Marquardt, D.W. (1970). Generalized Inverse, Ridge Regression, Biased Linear Estimation, and Non- Linear Estimation. *Technometrics*, 12: 591-612.

Özkale, M and S. Kaciranlar (2007). Comparisons of the Unbiased Ridge Estimation to the Other Estimations. *Communication in Statistics - Theory and Methods*, 36: 707–723.

Pasha, G. R. and A. A. S Muhammad (2004). Application of Ridge Regression to multicolinear data. *Journal of Research* (Science), 15(1): 97-106.

Rao, J. N. K. and A. C. Singh (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. *Proceedings of the section on survey research methods*. American Statistical Association. Pp.57–64.

Raymond M. L. (1975). Reviewed work, *American Journal of Agricultural Economics*, 57: 344.

(*Accepted for publication September 2012*)