



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

An analysis of Indian agricultural workers: a ridge regression approach

Banti Kumar¹, Manish Sharma^{2*}, Anil Bhat², and Pawan Kumar²

¹Department of Physical Sciences and Languages, COBS, CSK HPKV Palampur, Himachal Pradesh

²Sher-e-Kashmir University of Agricultural Sciences & Technology-Jammu (J), Chatha 180 009, J&K

*Corresponding author: manshstat@gmail.com

Abstract The agriculture sector in India cannot productively employ the growing rural labour force, and farmers are forced to look for other employment for their livelihood. This paper attempts to understand the increasing marginalization of the Indian agricultural workforce through different parameters which were influenced with multicollinearity. To handle multicollinearity in the data, this paper uses the ridge regression technique. The results shows that the value of the ridge constant K was found to be 0.02, at which the ridge regression model estimated the number of Indian agricultural workers more precisely than the ordinary least squares model.

Keywords Agricultural workers, multicollinearity, variance inflation factor, ridge regression, R square

JEL codes C1

Half the world's labour force, about 1.3 billion, work in agricultural production. The remuneration of agricultural workers is stumpy, work conditions put an undue burden on them, and employment is extremely irregular. About 170,000 agricultural workers shift occupations every year, maybe because their risk of dying on the job is at least twice that of workers in other sectors (ILO 2015). The number of agricultural labourers, 258.93 million in 2004–05, fell 30.59% to 228.34 million in 2011–12. Agricultural workers have been observed over the past few decades to be inclined towards other job avenues.

In India, agricultural landholdings are small, and these become even smaller with each passing generation. These land holdings do not meet the needs of growing families, and farmers are forced to shift from agriculture. The surveys of the National Sample Survey Office (NSSO) have been indicating the growing marginalization of the workforce, and that marginalization has been highlighted by the 2011 census. The share of marginal workers in the workforce grew from 22.2% in 2001 to 24.8% in 2011. In the

Indian economy, 25% of workers worked for less than six months in a year. The proportion of marginal workers in all workers grew from 13% in 1991 to 16% in 2011. The agriculture sector is not able to productively employ the growing rural labour force. Farmers are forced to look for other employment, where the prospects are transitory and unintended.

The economic data, including data related to farmers and agriculture, that is collected for analysis has repetitive variables or faulty dummy variables, and multicollinearity is a common problem. Multicollinearity may cause the coefficient estimates to swing violently, based on the other independent variables in the model, and the coefficients may become very sensitive to trifling changes in the model. Multicollinearity may lead to misleading results when a researcher tries to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

Multicollinearity influences the estimations of the coefficients used in economic decisions and strategic

planning; it also influences the estimations of monetary supplies and demands. All such parameters are basic and essential in economic planning, and these should be applied not only in research but in ground applications of specialized authorities such as finance ministries, central banks, and pricing units.

The ordinary least squares (OLS) estimator becomes inconsistent in the presence of multicollinearity in the data (Gujarati 1995), and the method of least squares breaks down because the estimates calculated with less confidence and regression coefficients tend to produce estimates that are too large in absolute value. Over the years, economists and statisticians have studied the problem of multicollinearity in economic data and suggested remedial measures.

Scott (1976) combined regression and factor analysis in agricultural economics research and suggested that the chi-square test be used to detect multicollinearity. Mittelhammer et al. (1980) used exact and stochastic restrictions to mitigate the effects of multicollinearity in an aggregate agricultural production function of Thailand.

The partial least squares path modelling was used by Enaami et al. (2011) to overcome the problem in agricultural production data for the least squares estimation of the Cobb–Douglas production functions. Principal component analysis can solve the collinearity among variables (Gwelo 2019).

The ridge regression technique is used to obtain consistent estimates of predictor variables in the presence of multicollinearity (Hoerl and Kennard 1970). Mahajan et al. (1977) applied the ridge regression technique to estimate parameters in marketing models in the presence of multicollinearity.

This study attempts to understand the factors responsible for the number of Indian agricultural workers; it uses the ridge regression technique to solve the problem of multicollinearity in macroeconomic data. The study uses the classical linear regression model and the OLS method, and it takes Indian agricultural workers as the endogenous variable. To estimate the unknown parameters, this study takes as exogenous variables the literacy rate, average size of holding, number of establishments, gross cropped area, net sown area, population density, and the inflation rate.

Materials and methods

The cross-sectional secondary data for Indian agricultural workers includes agricultural labourers and cultivators. The data for the year 2011–12 was collected from the Economic Survey of India and the Census Survey of India, the websites of the Ministry of Statistics and Programme Implementation and the Reserve Bank of India, and the data book published by the Indian Agricultural Statistics Research Institute for all the states and union territories of India.

The number of Indian agricultural workers (IAW) is the endogenous variable, the literacy rate (LR) was the explanatory variable, average size of holding (ASH) in hectares, number of establishments (EST) in numbers, gross cropped area (GCA) in hectares, net sown area (NSA) in hectares, population density (DEN) in per square and inflation rate (IR) in percentage.

We used Pearson's coefficient (Pearson 1920) to study the relationship among the endogenous variable and the exogenous variables. To detect multicollinearity, we used the condition number (C^*), variance inflation factor, and Q value. The C^* is the square root of maximum eigen value divided by the minimum eigenvalue (Vinod and Ullah 1981; Reuben and Emenonye 2013). If the condition number is above 30, the regression is said to have significant multicollinearity:

$$C^* = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

The condition number is $C^{**} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$ (Montgomery and Peck 1992).

If C^{**} is less than 100, multicollinearity does not pose a serious problem. If C^{**} lies between 100 and 1,000, the multicollinearity is moderate to strong. If C^{**} exceeds 1,000, the multicollinearity is severe.

The variance inflation factor measures the multicollinearity in a set of multiple regression variables. O'Brien (2007) examines the various rules of thumb for the variance inflation factor; they find its threshold values in the context of several other factors that influence the variance of regression coefficients,

and they suggest that ridge regression be used or the independent variables be combined into a single index.

If the variance inflation factor of a variable exceeds 10, the multicollinearity may cause a problem in estimating the parameters. Sufian (2010) suggests that if there are p explanatory variables, the expected squared distance of the OLS estimators from their true values (Chatterjee and Price 1977) is given by $L^2 = \sigma^2 \sum_{i=1}^p VIF_i$. The smaller the distance, the more precise the OLS estimates. If the explanatory variables are orthogonal, each variance inflation factor will be equal to 1 and $L^2 = p\sigma^2$.

$$\text{Therefore, the ratio } Q = \frac{\sigma^2 \sum_{i=1}^n VIF_i}{p\sigma^2} = \frac{\sum_{i=1}^n VIF_i}{p}$$

can also be used as a measure of multicollinearity. If the value of Q is large, multicollinearity may be present. Multicollinearity may be detected by several approaches—variable deletion, restrictions on the parameters, Bayesian estimation, and the ridge regression technique—and this paper uses the ridge regression technique to detect multicollinearity.

Ridge regression technique

The coefficient estimates of the linear regression rely on the independence of the model terms. In the presence of multicollinearity, the columns of the design matrix X have an appropriate linear dependence, and the matrix $(X'X)^{-1}$ becomes close to singular. As a result, the least squares estimate $\hat{\beta} = (X'X)^{-1}X'y$ becomes highly sensitive to random errors in the observed response y , producing a large variance and effect on

the size, sign, and significance of the parameters. The ridge regression technique helps in solving the problem by estimating regression coefficients using $\hat{\beta} = (X'X + KI)^{-1}X'y$, where, K is the ridge parameter and (I) is the identity matrix.

If $K=0$, the ridge estimates are also the OLS estimates. By sacrificing a small amount of bias in the estimates, more reasonable coefficients may be obtained and can lead to dramatic reductions in the variance of the estimated model coefficients. The value of the biasing constant K is to be determined by the minimization of the variance inflation factors and by the maximization of the determination coefficient R^2 of the model (Mardikyan and Cetin 2008). Marquardt (1970) proposed that the value of K should be found such that the variance inflation factor is between 1 and 10, probably closer to 1.

Results and discussion

To study the Indian agricultural workers, we calculated for the exogenous variables their minimum, maximum, mean, standard deviation, and coefficient of variation (CV) (Table 1). The maximum variability was due to population density (207.59%), followed by GCA (136.69%), NSA (134.70%), EST (123.26%), ASH (108.60%), IR (15.44%), and LR (10.31%). The overall average literacy rate during the study period was 78.65%; the standard deviation was 8.11%.

The average (standard deviation) of ASH was 1.86 (2.02) hectares. The EST was 1195.08 (14,714.58). The GCA was 4039.00 (5,521.22) hectares. The NSA was 5,574.37 (7,508.75) hectares. The DEN was 1,035.94

Table 1 Summary statistics of exogenous variables affecting Indian agricultural workers

Variables (in units)	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation (%)
LR (%)	63.8	93.9	78.65	8.11	10.31
ASH (hectares)	0.27	11.00	1.86	2.02	108.60
EST ('000)	3	4,433	1,195.08	14714.58	123.26
GCA ('000hectares)	1	17,551	4,039	5521.22	136.69
NSA ('000 hectares)	2	25,540	5574.37	7,508.75	134.70
DEN (per sq km)	17.00	9,340.00	1,035.94	2,150.59	207.59
IR (%)	7.62	18.32	14.24	2.20	15.44

Indian agricultural workers (IAW) literacy rate (LR), average size of holding (ASH, number of establishments (EST, gross cropped area (GCA), net sown area (NSA), population density (DEN) and inflation rate (IR)

Table 2 Correlation matrix between Indian agricultural workers and exogenous variables

Variables	IAW	LR	ASH	EST	GCA	NSA	DEN	IR
IAW	1							
LR	-0.510**	1						
ASH	-0.126	-0.200	1					
EST	0.796**	-0.246	-0.142	1				
GCA	0.879**	-0.462**	0.083	0.747**	1			
NSA	0.894**	-0.476**	0.084	0.752**	0.990**	1		
DEN	-0.179	0.314	-0.157	-0.130	-0.209	-0.205	1	
IR	0.246	-0.264	0.339*	0.236	0.417*	0.419*	-0.004	1

Indian agricultural workers (IAW) literacy rate (LR), average size of holding (ASH, number of establishments (EST, gross cropped area (GCA), net sown area (NSA), population density (DEN) and inflation rate (IR)

* = Significant at 5%, ** = Significant at 1%

(2,150.59) per sq km. The IR was 14.24 (2.20)%. The standard deviation exceeded the average value for all the exogenous variables except LR and IR, indicating the huge variation in the data on the states and union territories of India and that a robust model is required to handle the problem.

The relations between the variables—LR (-0.510**), EST(0.796**), GCA (0.879**), and NSA (0.894**)—were significantly correlated with the number of Indian agricultural workers (Table 2). The literacy rate is negatively related with the response variable. The number of establishments, gross cropped area, and net sown area were positively related with the response variable. The relationship between the exogenous variables—gross cropped area and net sown area—were significantly and highly correlated ($r=0.99^{**}$), which might be an indicator of multicollinearity. So, we used the C^* , VIF, and the Q value to detect multicollinearity in the data (Table 3).

The condition index value of exogenous variables—GCA, DEN, and IR, greater than 1000—showed the presence of multicollinearity. The VIFs for the coefficients of the EST and the GCA exceeded 10, indicating the presence of multicollinearity among the exogenous variables. The Q value was found to be 15.89, implying that the distance of the OLS estimators from their true values, as measured by Q, is over 15.89 times greater than if the explanatory variables were orthogonal, further indicating the presence of multicollinearity.

Table 3 Condition index and variance inflation factor of exogenous variables

Exogenous variables	Condition index		Variance inflation factor
	$C^* = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$	$C = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$	
LR	1	1	1.46
ASH	24.26	4.92	1.28
EST	140.86	11.86	2.60
GCA	42,930.97	207.19	50.60
NSA	347.63	18.64	52.71
DEN	20,508,050.18	4,528.58	1.16
IR	50,100,364.84	7,078.16	1.40

Indian agricultural workers (IAW) literacy rate (LR), average size of holding (ASH, number of establishments (EST, gross cropped area (GCA), net sown area (NSA), population density (DEN) and inflation rate (IR)

Ridge regression estimates

Figure 1 represents the ridge trace of the standardized betas of exogenous variables of Indian agricultural workers against the ridge constant (K). The standardized beta estimates started stabilizing in the interval of ridge constant values (K) in the 0.01–0.1 range. The ridge trace of the variance inflation factor values of exogenous variables of Indian agricultural workers against the ridge constant (K) also started stabilizing in the same interval of ridge constant values (K) (Figure 2). The variance inflation factor values of all the exogenous variables of Indian agricultural

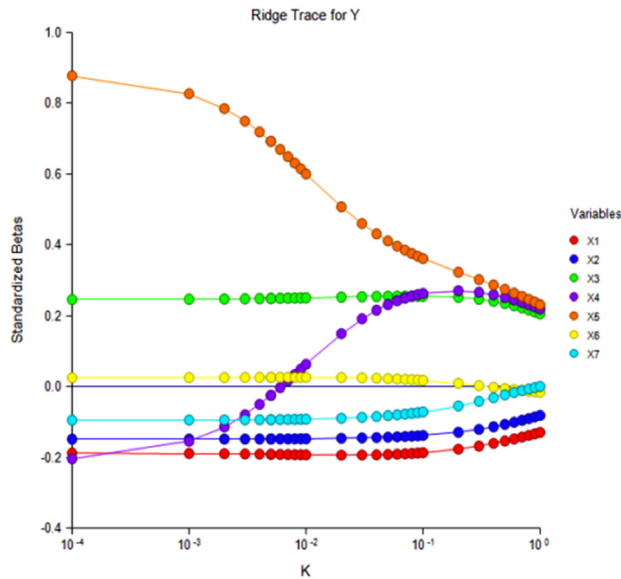


Figure 1 Ridge trace of standardized betas for exogenous variables of IAW

workers became less than 1, in line with Marquardt (1970). From Figures 1 and 2, it was concluded that the values of the ridge constant (K) lie in the interval of 0.01 to 0.1.

Then, the optimum value of the ridge constant (K) is obtained through several iterations using R^2 . The values of the coefficient of determination were found to be increasing at a faster rate between the range of ridge constant $K=1$ to $K=0.02$, and it became somewhat constant or increased at a slower rate (Table 4). The value of the variance inflation factor after $K=0.02$ was

Table 4 Ridge constant (K) analysis through coefficient of determination and variance inflation factor

K	R square	Max. VIF
0	0.8881	52.7122
0.001	0.8872	43.5577
0.003	0.8857	31.2464
0.007	0.8832	18.4563
0.01	0.8816	13.4868
0.02	0.8772	6.3209
0.06	0.8625	1.7705
0.20	0.8184	0.9298
0.40	0.7648	0.5350
0.60	0.7188	0.3869
1	0.6430	0.2356

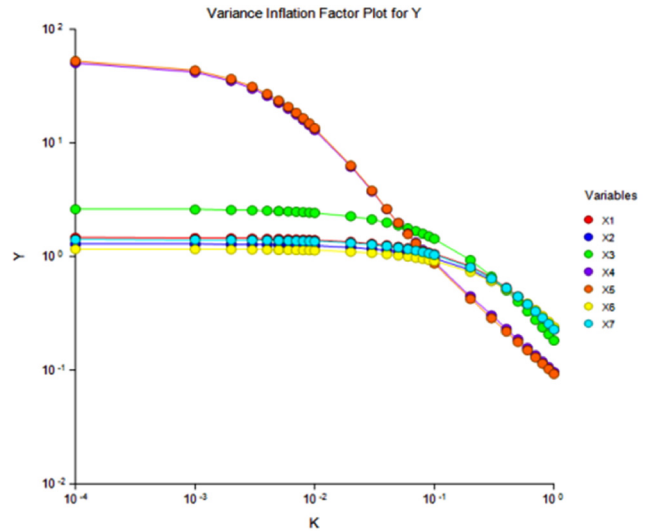


Figure 2 Plot of VIF of exogenous variables of IAW

more than 10, and the optimum value of the ridge constant was 0.02 for the problem under study. Mardikyan and Cetin (2008) use a similar methodology.

The ridge regression model for Indian agricultural workers was statistically significant; the model is adequate with respect to the exogenous variables (Table 5). The R^2 (0.884) indicates that 88.4% of the variation in the dependent variable is due to the explanatory variables in the study. The functional analysis of Indian agricultural workers revealed that the literacy rate was negatively significant whereas the EST and NSA were positively significant. The values of regression coefficient were $-208,302$ for LR, $1,502.06$ for EST and 591.15 for NSA; the results in terms of the sign are similar as in Table 2.

The parameters estimated using through the ridge regression technique and the OLS method were compared, and found that the ANOVA of both models were significant at 5% level of significance (Table 6). When using the ridge regression technique, the exogenous variables LR, EST, and NSA were found to be significant, whereas in the OLS method only the LR and EST significantly affected the variables under study. The value of the coefficient of determination when using the ridge regression technique (0.877) was greater than when using the OLS method (0.872). The mean square error value was smaller in the case of the ridge regression technique, in line with Malthouse (1999). The regression coefficient of the gross cropped

Table 5 The estimates of regression coefficients through ridge regression model with ridge constant K=0.02

Variable	Intercept	LR	ASH	EST	GCA	NSA	DEN	IR
Regression coefficient	2.37E+07	-208,302*	-1,005,285	1,502.06*	237.61	591.15*	106.06	-353,857
Mean square error (MSE) (In E+13)						1.18		
Coefficient of determination (In percent)						87.70*		

Indian agricultural workers (IAW) literacy rate (LR), average size of holding (ASH, number of establishments (EST, gross cropped area (GCA), net sown area (NSA), population density (DEN) and inflation rate (IR)

*Significant at 5%.

Table 6 Comparison among ridge regression coefficients with respect to ordinary least square estimates

Exogenous variable	Intercept	LR	ASH	EST	GCA	NSA	DEN	IR	R ²	MSE (E+13)
Least regression coefficient	24,713,127.87	-207,079.18*	-193,652	1,748.36*	-291.63	963.52	177.18	-536,973	0.872*	1.22
Ridge regression coefficient	2.37E+07	-208,302*	-1E+06	1,502.06*	237.61	591.15*	106.06	-353,857	0.877*	1.18

Indian agricultural workers (IAW) literacy rate (LR), average size of holding (ASH, number of establishments (EST, gross cropped area (GCA), net sown area (NSA), population density (DEN) and inflation rate (IR)

*Significant at 5%.

area is estimated using both the ridge regression technique and the OLS method, and the sign is different in each case; and the NSA is found to be positive and significant in the ridge regression technique but non-significant in the OLS model.

The number of agricultural labourers has fallen in India because the wage has fallen and the literacy rate has risen. Farmers may return to practising agriculture if they adopt modern, efficient, technology-driven practices and the net sown area increases, and if more of agriculture-related infrastructure (cold storage facilities) and products (implements) are made available.

Summary and conclusions

The economic variables are usually correlated; therefore, the model exhibited the problem of multicollinearity, and it affected the sign, size, and significance of the estimates of exogenous variables with respect to endogenous variables while modelling. Policy planners have to tackle the problem of multicollinearity before modelling, or they may use ridge regression techniques.

The optimum value of the ridge constant K was found to be 0.02. The ridge regression estimates were biased, but theoretically they were more stable and reliable, while the OLS method yielded estimates of the wrong sign and offered no significant meaning when interpreted. The ridge regression technique yielded statistically significant estimates of the right sign if the possible bias is ignored. The variances of the coefficients obtained by ridge regression were smaller than the OLS method; and R² was more at K=0.02.

The exogenous variables— the literacy rate, number of establishments, and net sown area—significantly affected the number of Indian agricultural workers in which the literacy rate is negatively related to the number whereas the number of establishments and the net sown area were positively related to the number.

Acknowledgements

The authors are very thankful to the referees for suggestions and bringing the paper in this format.

References

Chatterjee, S, and B Price. 1977. *Regression analysis by example*. John Wiley and Sons.

- Enaami, M, S A Ghani, and Z Mohamed. 2011. Multicollinearity problem in Cobb-Douglas production function. *Journal of Applied Sciences* 11 (16): 3015–021. <https://dx.doi.org/10.3923/jas.2011.3015.3021>
- Gujarati, D N. 1995. *Basic econometrics*. McGraw-Hill, New York.
- Gwelo, A S. 2019. Principal components to overcome multicollinearity problem. *Oradea Journal of Business and Economics* 4 (1): 79–91. <https://dx.doi.org/10.47535/1991ojbe062>
- Hoerl, A E, and R W Kennard. 1970. Ridge regression: biased estimation for non-orthogonal problem. *Technometrics* 12 (1): 55–67. <https://dx.doi.org/10.1080/00401706.1970.10488634>
- Mahajan, V, A K Jain, and M Bergier. 1977. Parameter estimation in marketing models the presence of multicollinearity: an application of ridge regression. *Journal of Marketing Research* 14 (4): 586–91. <https://dx.doi.org/10.1177/002224377701400419>
- Malthouse, E C. 1999. Ridge regression and direct marketing scoring models. *Journal of Interactive Marketing* 13 (4): 10–23. [https://dx.doi.org/10.1002/\(SICI\)1520-6653\(199923\)13:43.0.CO;2-3](https://dx.doi.org/10.1002/(SICI)1520-6653(199923)13:43.0.CO;2-3)
- Mardikyan, S, and E Cetin. 2008. Efficient choice of biasing constant for ridge regression. *International Journal of Contemporary Mathematical Sciences* 3 (11): 527–36. <https://www.m-hikari.com/ijcms-password2008/9-12-2008/cetinIJCMS9-12-2008.pdf>
- Marquardt, D W. 1970. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* 12 (3): 591–612. <https://dx.doi.org/10.2307/1267205>
- Mittelhammer, R C, D L Young, D Tasanasanta, and J T Donnelly. 1980. Mitigating the effects of multicollinearity using exact and stochastic restrictions: the case of an aggregate agricultural production function in Thailand. *American Journal of Agricultural Economics* 62 (2): 199–210. <https://dx.doi.org/10.2307/1239685>
- Montgomery, D C, and E A Peck. 1992. *Introduction to linear regression analysis*. John Wiley and Sons.
- O'Brien, R M. 2007. A caution regarding rules of thumb for variance inflation factor. *Quality & Quantity* 41 (5): 673–90. <https://dx.doi.org/10.1007/s11135-006-9018-6>
- Pearson, K. 1920. Notes on the history of correlation. *Biometrika* 13 (1): 25–45. <https://dx.doi.org/10.1093/biomet/13.1.25>
- Reuben, F O, and C Emenonye. 2013. Solutions to multicollinearity diagnostics. *Discrete Mathematics* 54A: 12860–863. [https://www.elixirpublishers.com/articles/1363783430_54A%20\(2013\)%2012860-12863.pdf](https://www.elixirpublishers.com/articles/1363783430_54A%20(2013)%2012860-12863.pdf)
- Scott, J T, Jr. 1976. Combining regression and factor analysis for use in agricultural economics research. *Journal of Agricultural and Applied Economics* 8 (2): 145–49. <https://dx.doi.org/10.1017/S0081305200013376>
- Sufian, A J M. 2010. *An analysis of poverty—a ridge regression approach*. Paper presented at the Proceedings of 4th International multi-conference on Society, Cybernetics, and Informatics (IMSCI 2010), 29 June to 2 July, Orlando, Florida, USA. https://www.iiis.org/CDs2010/CD2010SCI/SOIC_2010/PapersPdf/WA355CW.pdf
- Vinod, H D, and A Ullah. 1981. *Recent advances in regression models*. Marcel Dekker, New York.

Received: 6 February 2020 Accepted: 18 July 2020