Biological Forum – An International Journal **17(1): 08-13(2025)**

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

Addressing Multicollinearity in Path Analysis: Insights from Ridge Regression in Rice Yield Components

Kotadiya Trunal, Singh S.K., S. Jayasudha*, Kumar Ankit and Futane Aachal
Department of Genetics and Plant Breeding, Institute of Agricultural Sciences,
Banaras Hindu University, Varanasi (Uttar Pradesh), India.

(Corresponding author: S. Jayasudha)*
(Received: 03 October 2024; Revised: 18 November 2024; Accepted: 04 December 2024; Published online: 03 January 2025)
(Published by Research Trend)

ABSTRACT: Multicollinearity among predictor traits is a major challenge in path coefficient analysis, distorting the estimation of direct and indirect effects on grain yield in rice. This study addresses this issue by applying ridge regression to an M2 population of rice (*Oryza sativa* L.), evaluating its effectiveness in stabilizing coefficients and improving model interpretability. Severe multicollinearity was identified in the conventional path analysis model, with traits like plant height (PH), Culm height (CH), days to maturity (DM), panicle weight (PW), and grain number per panicle (GNP) exhibiting inflated Variance Inflation Factor (VIF) values (>10). Ridge regression significantly reduced multicollinearity, stabilizing path coefficients while retaining key traits. In the modified Model 6, ridge regression combined with the exclusion of highly collinear traits (e.g., plant height (PH) and days to flowering (DF)) achieved superior model fit (R2=0.870R^2 = 0.870) and reduced residual effects. Total tillers (TT) and biomass yield (BY) consistently emerged as primary yield determinants, with realistic direct effects (0.340 and 0.428, respectively). Additionally, ridge regression preserved essential interdependencies among traits like PW and GNP, which were distorted in the conventional model. This study provides a robust framework for handling multicollinearity in path coefficient analysis, emphasizing the practical utility of ridge regression in breeding programs.

Keywords: Multicollinearity, Path coefficient analysis, Ridge regression, Correlation analysis, Predictor variables, Direct effect, Variance Inflation Factor (VIF), Rice.

INTRODUCTION

Rice (*Oryza sativa* L.) is one of the most widely consumed staple crops globally and plays a vital role in ensuring food security, particularly in many developing countries where it is a primary source of nutrition for millions of people. Improving rice yields is crucial to meet the increasing global demand, and understanding relationships between morphological traits can assist plant breeders in identifying key yield-contributing traits to optimize selection criteria, and gaining a deeper understanding of the genetic architecture of yield to improve breeding efficiency. To get insights into these complex relationships between yield and its component traits, path coefficient analysis plays a crucial role. An appropriate utilising path coefficient analysis as an aid to the selection of genotype can make significant strides in developing high-yielding, resilient rice varieties to meet the growing global demand for food. The technique was proposed by Dewey & Lu (1959) to separate the simple correlation coefficient between the seed yield (as the dependent variable) and its components (as predictor variables) into direct and indirect effects (that affect grain yield through the other variables). However, the path coefficient is a standardized partial regression coefficient, therefore, no or weak associations among predictor variables are a necessary assumption needed to satisfy the goodness of fit for the path analysis model (Neter *et al*., 1996). Correlation and path analysis in rice was done by various scientists before like Singh *et al*. (2018); Zahid *et al*. (2006); Patil *et al*. (2005); Mahto *et al*. (2003); Sarawgi *et al*. (1997), etc.

Multicollinearity, a condition where two or more predictor variables are highly correlated, poses a significant challenge in path analysis, as it can lead to inaccurate estimates of the relationships between yield component traits, thereby hindering the identification of the most critical yield component traits contributing to grain yield. In fact, independence among yield components is rarely found under field conditions. This case is usually called a multicollinearity problem (EL-Taweel *et al*., 2012). Multicollinearity can lead to an overestimation of direct effects, producing effect values greater than 1, which are not meaningful for interpreting path analysis (Satyanarayana *et al*., 2023; Muthuramu *et al*. 2023). Similarly, Gravois & Helms (1992) mentioned that when the ordinary path analysis model is used in the presence of multicollinearity, some path coefficients may exceed one, which is considered a negative effect of multicollinearity. Williams *et al*. (1979) stated that the negative effects of

multicollinearity may be enough to reject the ordinary path analysis model. Various statistical methods and techniques have been employed to handle multicollinearity in path analysis, including ridge regression, principal component regression, and partial least squares regression. These methods can be effective in addressing multicollinearity, but the choice of method depends on the specific characteristics of the data and the research question at hand. Path coefficient analysis relies on the accurate estimation of regression weights of the predictor variables (yield attributing traits) to interpret causal relationships among them. Ridge regression improves the reliability of these coefficients when predictors are highly correlated, avoiding exaggerated effects of multicollinearity (Xu *et al.*, 2014; Pelzer *et al*., 2009). It retains all original predictors in the model while principal component regression and partial least squares regression methods reduce predictors to latent components and may lose some granular details (Zou & Hastie 2005). Thus, no information is completely discarded in the ridge regression method. The ridge regression method adds a penalty term proportional to the squared magnitude of coefficients that effectively shrinks coefficients towards zero without fully eliminating variables. This reduces the variance of coefficient estimates in the presence of multicollinearity, improving their stability (Hoerl & Kennard 1970). Thus, ridge regression preferred method for resolving multicollinearity in path coefficient analysis which improves model interpretability and predictive accuracy. In this study, we have made a framework for dealing with multicollinearity in path coefficient analysis on the mutant M_2 population of rice using the strategies suggested by Olivoto *et al*. (2017).

MATERIAL AND METHODS

The experimental material used for this study was an M_2 population of the rice cultivar Gobindabhog. The 910 mutant seedlings were transplanted at the spacing of 20 × 20 cm as one seedling per hill and each individual was phenotyped for 14 traits as mentioned in Table 1.

Table 1: List of the traits evaluated.

Sr. No.	Traits	Denotation	Method of evaluation
1	Plant height	**PH** (cm)	Height from base to the tip of the panicle on the main culm
2	Culm height	**CH** (cm)	Height from base to the basal joint of the panicle on the main culm
3	Total tillers	**TT** (No.)	Number of reproductive tillers (that bear the panicle) in the individual plant
4	Days to flowering initiation	**DF** (days)	Number of Days from the nursery sowing to the emergence of the first panicle
5	Days to maturity	**DM** (days)	Number of Days from the nursery sowing to the maturity (80% mature grain/panicle)
6	Panicle length	**PL** (cm)	Length from the basal joint of the panicle (neck) to the tip
7	Panicle weight	**PW** (g)	Average weight of the whole panicle
8	Primary panicle branches	**PPB** (No.)	Number of primary branches in the panicle on the main culm
9	Grain number per panicle	**GNP** (No.)	Average number of fertile spikelets/panicles
10	Spikelet fertility	**SF** (%)	Spikelet fertility % = $\frac{\text{Total number of fertile spikelets}}{\text{Total number of spikelets(fertile and sterile)}} \times 100$
11	Test weight of seeds	**TW** (g)	Weight of the 1000 filled grains
12	Biomass yield per plant	**BY** (g)	Weight of the whole above ground plant parts after harvesting and drying
13	Grain yield per plant	**GY** (g)	Weight of all fertile spikelets (grains) threshed from all the panicles of each plant
14	Harvest index	**HI** (%)	Harvest Index (%) = $\frac{\text{Grain yield per plant (only grain weight)}}{\text{Biomass yield per plant (straw weight + grain weight)}} \times 100$

Statistical analysis. The simple correlation matrix was computed among traits as outlined by Snedecor & Cochran (1989). The conventional and modified models of path analysis were formulated by considering the grain yield per plant (GY) as a dependent variable and the other 13 traits as independent variables. The pairs. panels() and path_coeff() functions of the "psych" and "metan" R packages were used for the analysis. A common measure called the Variance Inflation Factor (VIF) was used to test the presence of multicollinearity among the predictor variables (yield attributing traits) in different models (Hair *et al.*, 1992). If the VIF value is above 10 for any trait, this indicates the presence of multicollinearity, then such models could not estimate the correct contribution of predictor variables. Here we have applied two important strategies for the ideal model construction, **(i)** exclude the traits from the model if VIF>10 and **(ii)** apply the modified path analysis model (ridge regression-based model).

The modified path analysis model was proposed by Carvalho *et al.* (1999) to correct the negative effects of the multicollinearity problem. The proposed model adds a small bias constant value (k) to the diagonal elements (unity value) of the correlation matrix of yield components (predictor variables). The k value would range from 0 to 1. When k=0, the modified model would behave like the normal model. Two methods have been used to determine the optimum value of k,

(a) ridge trace exam (Hoerl & Kennard 1970 a and b) uses a 2D chart showing how the path coefficients vary as a function of k ($0 < k < 1$) to select the smallest value of k which is capable of stabilising most path coefficients (<1) and **(b)** VIF based approach select the smallest value of k at which VIF values become less than 10.

A total of 5 different models of path coefficients were estimated as mentioned in Table 2. The impact of different models was evaluated by analysing the direct effects of traits with multicollinearity, residual effect square and coefficient of determinants (R^2).

Table 2: Models of Path Coefficient Analysis.

Model	Method used	Traits excluded
Model 1	Normal path analysis (Conventional path analysis)	-
Model 2	Normal path analysis (Modification)	PH
Model 3	Normal path analysis (Modification)	PH, DF
Model 4	Normal path analysis (Modification)	PH. DF, PW
Model 5	Modified path analysis (Modification)	PH
Model 6	Modified path analysis (Modification)	PH, DF

RESULTS AND DISCUSSION

Correlations among traits. The strongest correlation was observed between culm height (CH) and plant height (PH) (r = 0.98), followed by days to flowering (DF) and days to maturity (DM) (r = 0.96), and panicle weight (PW) and grain number per panicle (GNP) (r = 0.94). High correlations were also noted for PW and spikelet fertility (SF) (r = 0.85), total tillers (TT) and biomass yield (BY) (r = 0.77), and GNP and SF (r = 0.75). These inter-correlations indicated significant multicollinearity among predictor traits.

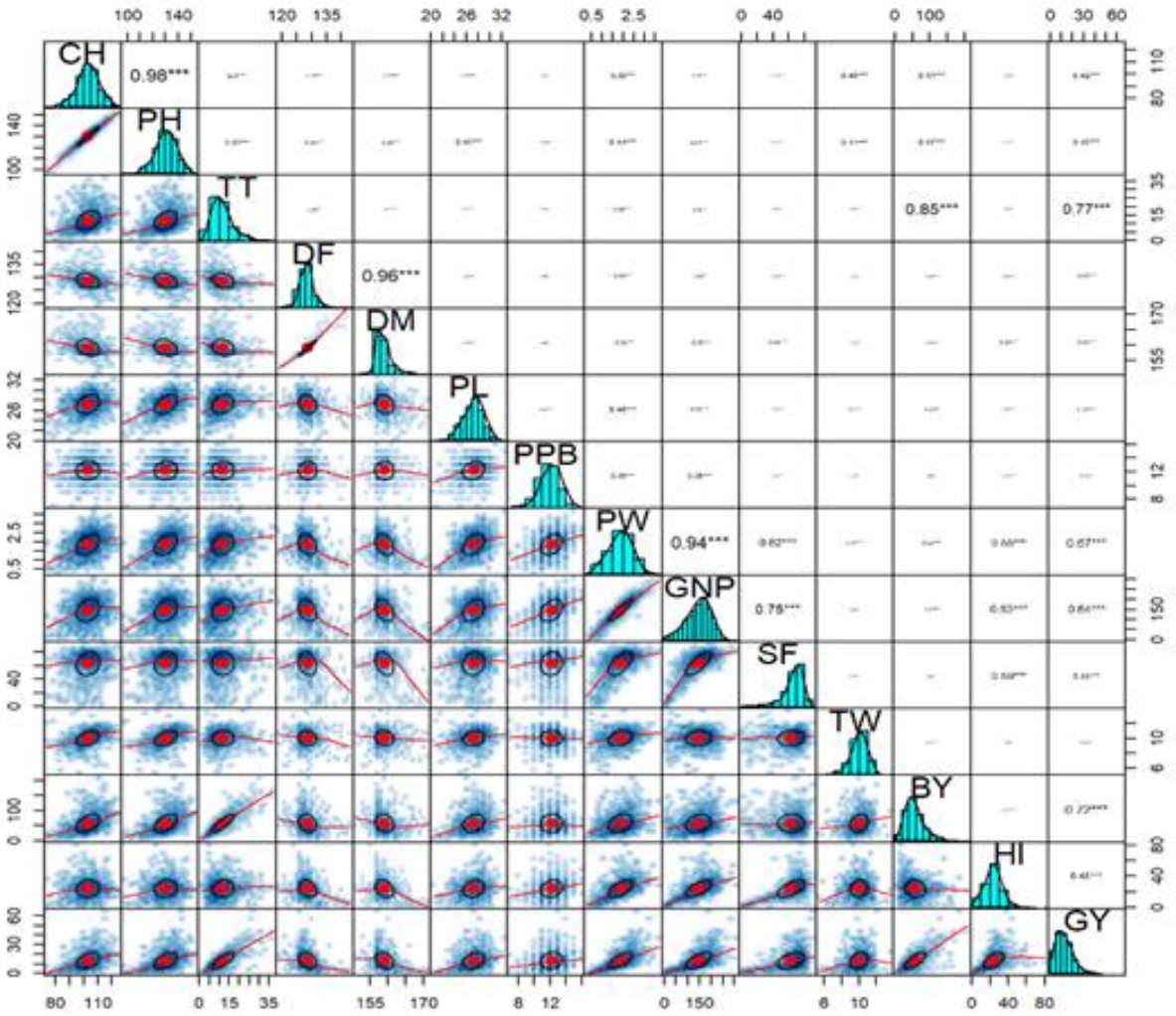

Fig. 1. Correlation Plot.

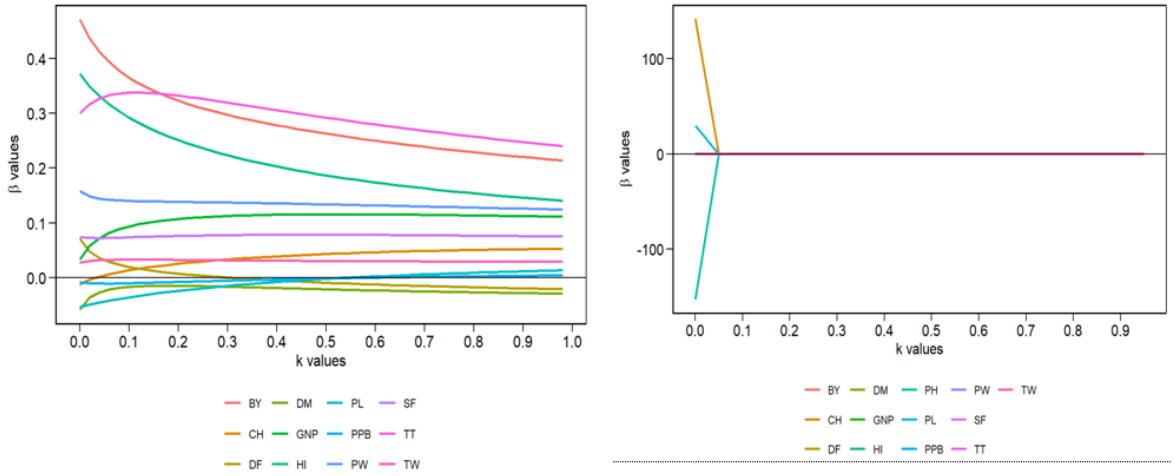

Fig. 2. Ridge trace examination plot 1 (left) and plot 2 (right).

Table 3: The VIF of different traits in the Model.

	Traits	Model1	Model2	Model3	Model4	Model5	Model6
VIF	**CH**	**-613840.633**	1.772	1.771	1.646	1.575	1.574
	PH	**-711761.075**	-	-	-	-	-
	TT	4.601	4.601	4.601	4.436	3.334	3.332
	DF	**11.328**	**13.094**	-	-	6.534	-
	DM	**11.954**	**13.810**	1.292	1.287	6.871	1.217
	PL	**-26939.525**	1.272	1.266	1.234	1.193	1.189
	PPB	1.362	1.366	1.366	1.329	1.224	1.224
	PW	**12.750**	**13.064**	**13.025**	-	6.532	6.525
	GNP	**16.400**	**16.413**	**16.404**	4.060	7.880	7.879
	SF	3.266	3.345	3.274	2.833	2.480	2.459
	TW	0.916	1.505	1.505	1.240	1.282	1.282
	BY	5.947	5.983	5.973	5.661	4.210	4.208
	HI	2.580	2.581	2.580	2.493	2.137	2.136
Condition Number (CN)		-6209314.515	124.034	113.941	33.369	57.236	52.955
Multicollinearity		Severe	Moderate	Moderate	Weak	Weak	Weak

Path Coefficient analysis: The Model 1 exhibited negative VIF values for CH, PH and PL while excessive VIF values for traits like DF (11.328), DM (11.954), PW(12.750) and GNP(16.400) representing the presence of severe multicollinearity among predictor variables. As well as higher condition number (CN) also confirms the severe multicollinearity in data. This led to exaggerated and opposite direct effects for traits such as CH (142.022) and PH (-152.944). Other traits like PL (29.702) and PW (11.354) showed inflated contributions, undermining their reliability in explaining grain yield. The ridge trace examination plot 2 demonstrated perfect complementarity between CH and PL with PH which leads to over-estimated effects with opposite directions. Looking to the perfect complementarity PH with CH it has been removed permanently in all other models. Its removal in model 2 led to an improvement in the coefficient of determinants (0.888) compared to model 1(0.856) as well as the multicollinearity issue of CH (VIF = 1.772) and PL(VIF = 1.272) resolved. The exclusion of one more trait (DM) in model 3 resolves the VIF of the DM (1.292) and the direct effects of traits were not much affected. The removal of PW in model 4 resolves the multicollinearity issues at all levels (CN<100) but the exclusion of PW leads to an overestimation of GNP direct effects (0.190) that had shown the requirement of the integration of ridge regression-based model.

The optimum k value for the ridge regression in models 5 and 6 has been fixed at 0.04 by analysing the ridge trace examination plot 1. Model 5 has been evaluated with all traits except PH. The ridge regression was found to deal with multicollinearity without excluding the correlated traits. Still, the direct effects of DF(0.038) and DM(-0.027) were observed in opposite directions. Ridge regression combined with the exclusion of PH and DF in model 6 achieved the highest stability and interpretability, with a coefficient of determination (R2=0.870) and residual effect square of 0.130. Ridge regression effectively resolved multicollinearity, stabilizing direct and indirect effects. For instance, CH (0.002) and DM (0.008) showed realistic and interpretable direct effects and TT (0.340) retained its strong direct effect on grain yield, emphasizing its role as a key yield determinant. Ridge regression significantly stabilizes VIF values, reducing them to below the critical threshold of 10, as seen for GNP (Model 1: 16.4; Model 6: 7.879). In Model 6, traits like TT and BY emerged as the most consistent predictors of grain yield, with direct effects of 0.340 and 0.428, respectively. Indirect effects between traits like PW and GNP were more consistent, highlighting their interdependence without overestimation.

Table 4: Direct effects of predictor variables.

	Traits	Model1	Model2	Model3	Model4	Model5	Model6
Direct Effects	**CH**	**142.022**	-0.012	-0.012	0.004	0.002	0.002
	PH	**-152.944**	-	-	-	-	-
	TT	0.303	0.300	0.299	0.281	0.340	0.340
	DF	0.312	**0.071**	-	-	**0.038**	-
	DM	-0.305	**-0.058**	0.012	0.015	**-0.027**	0.008
	PL	**29.702**	-0.053	-0.055	-0.047	-0.047	-0.048
	PPB	-0.020	-0.008	-0.009	-0.017	-0.011	-0.011
	PW	0.260	0.158	0.162	-	0.150	0.151
	GNP	0.055	0.034	0.032	**0.190**	0.076	0.075
	SF	0.023	0.074	0.079	0.049	0.076	0.078
	TW	0.167	0.028	0.028	0.051	0.034	0.033
	BY	0.505	0.471	0.469	0.494	0.429	0.428
	HI	0.367	0.372	0.371	0.384	0.344	0.344
Coefficient of Determinants (R^2)		0.856	0.888	0.888	0.886	0.870	0.870
Residual effect square		0.144	0.112	0.112	0.114	0.130	0.130

Table 5: Path Coefficient Matrix (Model 6) of direct and indirect effects of yield components.

Variable	CH	TT	DM	PL	PPB	PW	GNP	SF	TW	BY	HI
CH	0.002	0.108	-0.002	-0.012	0.001	0.058	0.017	0.014	0.013	0.194	0.023
TT	0.000	0.340	-0.002	-0.009	0.000	0.041	0.016	0.005	0.003	0.344	0.030
DM	0.000	-0.077	0.008	0.007	0.000	-0.051	-0.027	-0.026	-0.002	-0.072	-0.105
PL	0.000	0.068	-0.001	-0.048	-0.002	0.056	0.021	0.007	0.005	0.108	0.060
PPB	0.000	0.000	0.000	-0.008	-0.011	0.043	0.026	0.005	0.000	0.005	0.058
PW	0.001	0.092	-0.003	-0.017	-0.003	0.151	0.066	0.046	0.009	0.137	0.188
GNP	0.000	0.073	-0.003	-0.013	-0.004	0.134	0.075	0.056	0.003	0.087	0.209
SF	0.000	0.022	-0.003	-0.005	-0.001	0.090	0.054	0.078	0.003	0.012	0.194
TW	0.001	0.026	-0.001	-0.007	0.000	0.041	0.006	0.007	0.033	0.079	0.022
BY	0.001	0.273	-0.001	-0.012	0.000	0.049	0.015	0.002	0.006	0.428	-0.044
HI	0.000	0.029	-0.003	-0.008	-0.002	0.083	0.046	0.044	0.002	-0.054	0.344

This study highlights the limitations of conventional path analysis in the presence of severe multicollinearity and demonstrates the advantages of different modifications in path analysis models for resolving these issues. The VIF is an effective measure of multicollinearity in path analysis that explains well the abnormal estimation of the direct effects. The 2 strategies followed for ideal model construction show that according to trait behaviour in population, selection of strategies varies as for PH and DM exclusion was effective while for PW ridge regression-based strategy suits well. PH, DM, and GNP have also shown similar associations with GY in the mutant M2 population as observed by Chandar *et al.* (2023). The PH and DF are highly colinear traits with CH and DM respectively and the interpretation of yield based on one of them can be appropriate so their exclusion was a suitable approach. While excluding traits residual effect should not increase drastically and this consideration was taken care of in this study. On other hand ridge regression significantly reduced VIF values, stabilizing coefficients without the need to exclude critical traits like PW and GNP. This aligns with previous findings that emphasize ridge regression's utility in retaining all predictors while controlling for multicollinearity (Hoerl & Kennard 1970). In Model 6, traits like TT and BY emerged as the most consistent predictors of grain yield which provides reliable insights into trait selection for breeding programs. The ridge regression penalty shrank the overestimated coefficients, such as those for CH and PH in Model 1, making the analysis robust. Thus, the combination of ridge regression with selective exclusion of highly collinear traits (e.g., PH and DF) in Model 6 achieved superior model fit ($R2=0.870R^2$ = 0.870) and resolved multicollinearity issues (CN < 100) (Hair *et al.*, 1992).

CONCLUSIONS

This study demonstrates that ridge regression is an effective method for addressing multicollinearity in path coefficient analysis. By stabilizing coefficients and retaining critical predictors, ridge regression provides robust insights into the relationships among yield-related traits. Traits such as TT and BY consistently emerged as key contributors to grain yield, highlighting their potential as selection criteria in rice breeding. The proposed approach offers a valuable framework for addressing multicollinearity in other crop improvement studies. The stability of coefficients in Models 5 and 6 underscores the importance of integrating ridge regression into path coefficient analysis. This approach allows for the inclusion of interrelated traits like GNP and PW, which are vital for understanding complex yield dynamics in rice (Hoerl & Kennard 1970; Xu *et al*., 2014).

Acknowledgement. We are thankful to Banaras Hindu University (BHU) for funding this research and to the faculty of the Department of Genetics and Plant Breeding, Institute of Agricultural Sciences, BHU, for providing academic and technical assistance.

REFERENCES

Carvalho, S. P., Cruz, C. D. & Carvalho, C. G. P. (1999). Estimating gain by use of a classic selection under multicollinearity in wheat (*Triticum aestivum*). *Genetics and Molecular Biology*, *22*, 109-113.

Chandar, S. R. H., Susmitha, P., Ganesh, P., Mounika, A. & Anand, S. M. K. (2023). Evaluation of EMS induced mutant population using character associations and principle component analysis in rice (*Oryza sativa* L). *Biological Forum – An International Journal, 15*(1), 616-622.

Dewey, D. R. & Lu, K. H. (1959). A correlation and path coefficient analysis of components of crested wheatgrass seed production. *Agronomy Journal*, *51*, 515-518.

El-Taweel, A. M. S. A., El-Koomy, M. B. A. & Fares, W. M. (2012). Path analysis to control the multicollinearity among yield components in maize (*Zea mays* L.). *Egypt. J. Agron.*, *34*, 213-226.

Gravois, K. A. & Helms, R. S. (1992). Path analysis of rice yield and yield components as affected by seeding rate. *Agronomy Journal*, *84*(1), 1-4.

Hair, J. F., Anderson, J. R., Tatham, R. L. & Black, W. C. (1992). Multivariate data analysis. MacMillan Pub. Comp., A Division of MacMillan, Inc., USA.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69-82.

Hoerl, A. E. & Kennard, B. G. (1970b). Ridge regression: Biased estimation for monorthogonal problems. *Technometrics*, *12*, 55-68.

Mahto, R. N., Yadava, M. S. & Mohan, K. S. (2003). Genetic variation, character association and path analysis in rainfed upland rice. *Indian Journal of Dryland Agriculture Research and Development*, *18*(2), 196-198.

Muthuramu, S. & Thangaraj, K. (2023). Genetic variability, heritability, genetic advance and trait association studies in rainfed rice genotypes. *Biological Forum – An International Journal*, *15*(11), 108-112.

Neter, J**.,** Kutner, M. H., Wasserman, W. & Nachtsheim, C. J. (1996). *Applied linear statistical models* (4th ed.). McGraw-Hill/Irwin.

Olivoto, T., de Souza, V. Q., Nardino, M., Carvalho, I. R., Ferrari, M., de Pelegrin, A. J. & Schmidt, D. (2017). Multicollinearity in path analysis: a simple method to reduce its effects. *Agronomy Journal*, *109*(1), 131-142.

Patil, P. V. & Sarawgi, A. K. (2005). Studies on genetic variability, correlation and path analysis in traditional aromatic rice accessions. *Annals of Plant Physiology*, *19*(1), 92-95.

Pelzer, G. M. M., Feelders, A. J. & Groenen, P. J. F. (2009). Comparison of ridge regression and partial least squares regression in multicollinear data. *Journal of Applied Statistics*, *36*(4), 429-446.

Sarawgi, A. K., Rastogi, N. K. & Soni, D. K. (1997). Correlation and path analysis in rice accessions from Madhya Pradesh. *Field Crops Research*, *52*(1-2), 161-167.

Satyanarayana, P. V., Kumar, K. M., Babu, P. U., Srinivas, T. & Manojkumar, D. (2023). Association studies for identifying the selection criteria among early varieties of rice in North Coastal Zone of Andhra Pradesh. *Biological Forum – An International Journal*, *15*(11), 329-335.

Singh, R., Yadav, V., Mishra, D. N. & Yadav, A. (2018). Correlation and path analysis studies in rice (*Oryza sativa* L.). *Journal of Pharmacognosy and Phytochemistry*, *7*(1), 2084-2090.

Snedecor, G. W. & Cochran, W. G. (1989). Statistical methods, 8th Edn. *Ames: Iowa State Univ. Press Iowa*, *54*, 71-82.

Williams, W. A., Qualset, C. O. & Geng, S. (1979). Ridge regression for extracting soybean yield factors 1. *Crop Science*, *19*(6), 869-873.

Xu, Q., Liang, H. & Li, G. (2014). Adaptive ridge penalized model estimation in high-dimensional data. *Journal of Multivariate Analysis*, *123*, 160-171.

Zahid, M. A., Akhter, M., Sabar, M., Manzoor, Z. & Awan, T. (2006). Correlation and path analysis studies of yield and economic traits in Basmati rice (*Oryza sativa* L.). *Asian J. Plant Sci*, *5*(4), 643-645.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67* (2), 301-320.

How to cite this article: Kotadiya Trunal, Singh, S.K., S. Jayasudha, Kumar Ankit and Futane Aachal (2025). Addressing Multicollinearity in Path Analysis: Insights from Ridge Regression in Rice Yield Components. *Biological Forum – An International Journal, 17*(1): 08-13.