Int. J. Agr. Ext. 12 (03) 2024. 379-388 DOI:-10.33687/ijae.012.003.5418

Available Online at EScience Press
International Journal of Agricultural Extension
ISSN: 2311-6110 (Online), 2311-8547 (Print)
https://esciencepress.net/journals/IJAE

EVALUATING AGRONOMIC FACTORS INFLUENCING COTTON YIELD USING MULTIVARIATE REGRESSION AND MODEL VALIDATION

[a,b]Iftakhar Ahmed*, [a]Mohammad H. Lakho, [a]Riaz A. Buriro, [c]Aijaz A. Khooharo

[a] *Department of Statistics, Faculty of Agricultural Social Sciences, Sindh Agriculture University, Pakistan.*

[b] *Department of Economics, Faculty of Management and Social Sciences, Lasbela University of Agriculture, Water and Marine Sciences, Uthal, Balochistan, Pakistan.*

[c] *Department of Agricultural Education Extension and Short Courses, Faculty of Agricultural Social Sciences, Sindh Agriculture University, Pakistan.*

ARTICLE INFO

Article History
Received: August 10, 2024
Revised: November 12, 2024
Accepted: December 01, 2024

Keywords
Agronomic parameters
Regression model's assumption
Mahalanobis distance (MD)
Model Validation
Standardized beta coefficient
Variance Inflation factors (VIF)
Cotton

ABSTRACT

Regression analysis is a statistical technique used to estimate connections between variables that exhibit a cause-and-effect relationship. Despite its widespread use for identifying correlations and predicting outcomes, it is crucial to validate the assumptions and reliability of multivariate regression models. This study focuses on multivariate regression analysis, which utilizes multiple independent variables to predict cotton yield in kilogram per hectare (YldPhec). The secondary dataset was sourced from the Cotton Research Station (CRS) in Uthal, Lasbela, Balochistan, Pakistan. The analysis included eight predictors, revealing that the intercept and PlntPl had marginally significant positive effects (Beta = 4425.26, 0.04; $p < 0.1$), while Grmn and FbrSnt demonstrated highly significant positive effects (Beta = 48.76, 177.97; $p < 0.001$). Conversely, BolWt and StpLt exhibited significant negative effects (Beta = -741.49, -246.77; $p < 0.001$), with Lnt also showing a significant negative effect (Beta = -145.57; $p < 0.01$). Additionally, MikV and BolPp were not significant. Zero-order correlation analyses indicated strong positive relationships for Grmn (0.56) and PlntPl (0.50), while BolWt (-0.41) and Lnt (-0.42) showed strong negative correlations with the dependent variable. Tolerance values exceeding 0.39 and Variance Inflation Factor values less than 3 indicate that multicollinearity is not a significant concern among the predictors. The Shapiro-Wilk, Rainbow, and Studentized Breusch-Pagan test statistics confirmed the normality of residuals, the linearity of the model, and the absence of significant heteroscedasticity, with p-values of 0.74, 0.84, and 0.32, respectively. Confirming these assumptions enhances the model's validation and underscores the significance of the identified predictors in explaining the variability of YldPhec.

Corresponding Author: Iftakhar Ahmed
Email: iftkhanzada@gmail.com

INTRODUCTION

The agricultural sector is the cornerstone of Pakistan's economy, which is predominantly agrarian (Rehman et al., 2015). A significant proportion of Pakistan's population relies on agriculture for their livelihood and food security (Naseer et al., 2020). The cropping sector contributes 6.8% to the overall GDP of the country (Shakeel et al., 2023). Approximately 1.3 million out of 5 million farmers cultivate cotton (*Gossypium hirsutum L.*) on 6.0 million acres, accounting for 15% of the country's

total cultivated area (Abubakar et al., 2023). The cotton crop constitutes 0.8% of the GDP and contributes 4.5% to the agricultural value addition (Naveed et al., 2024).

Regression analysis is a statistical technique employed to identify correlations and predict outcomes based on cause-and-effect relationships among variables (Brillinger et al., 1967; Graybill, 1961; Sestelo et al., 2016; Wen et al., 2017; Zyskind, 1964). Despite its widespread application, it is essential to validate the assumptions and reliability of multivariate regression models, particularly for predicting cotton yield. This study examines the assumptions and the influences of various predictors on cotton yield.

The research questions focus on understanding the relationships between the dependent variable (cotton yield) and various independent variables. The study seeks to determine the causal connections, assess the strength of these connections, and evaluate their influence on cotton yield prediction. Additionally, it investigates the impact of specific independent variables or groups of variables under controlled conditions and develops methods for systematic evaluation.

The objectives are to explore the causal connections among variables, assess their influence on cotton yield prediction, validate the assumptions of the multivariate regression model—including normality, linearity, and homogeneity of variance—and evaluate the impact of specific variables under controlled conditions.

METHODOLOGY

Linear regression is useful for predicting quantitative response (Hastie et al., 2021). Univariate regression analysis involves using a single independent variable (Wang et al., 2020), while multivariate regression analysis involves using more than one independent variable (Izenman, 2013; Poon and Feng, 2023). In both cases, equations showing linear relationships between dependent and independent variables are formulated (Jakešová, 2014; Tabachnick and Fidel, 2019). Multivariate regression analysis simultaneously considers the variations in multiple independent factors that affect the dependent variable. (Kawano et al., 2023; Ünver and Gamgam, 1999). This descriptive study focuses on agronomic parameters to determine the significance of independent variables to the response variable in regression analysis. Multiple linear regression presumes that the model's response variable with p predictors is a linear function of the model parameters (Dhamodharavadhani S. and Rathipriya R., 2021). This is expressed compactly and in matrix notation, respectively, as follows:

$$\boldsymbol{y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i} \quad (1)$$

$$\underline{\boldsymbol{Y}}_{\ n\times 1} = \underline{\boldsymbol{X}}_{\ n\times p}\,\underline{\boldsymbol{\beta}}_{p\times 1} + \underline{\boldsymbol{\varepsilon}}_{\ n\times 1} \quad (2)$$

Where $\underline{Y}$ is the vector of responses, $\underline{X}$ is a matrix of known regressors, $\underline{\beta}$ is the vector of unknown parameters and $\underline{\varepsilon}$ is the vector of random error (Rawlings et al., 1998). Where the errors (ε_i) are assumed to be independently and identically normal random variables with a zero mean and a constant variance (Chatterjee and Hadi, 2012; Zhang, 2022). The vector of regression coefficients is estimated as:

$$\boldsymbol{\widehat{\beta} = (X'X)^{-1}X'Y} \quad (3)$$

The following considerations must be made to secure them:

The coefficient of determination (R_p^2) indicates the percentage variation explained by predictors to the response variable (Faraway, 2002; Finch et al., 2016; Frost, 2019) and the strength of the relationship (Frederick et al., 2019) is shown by

$$\boldsymbol{R_p^2 = \frac{SSR}{SST}} \quad (4)$$

A large value of R_p^2 will always result in a large model (Haldar and Miller, 1992). However, the adjusted coefficient of multiple determination (R_{adj}^2) is a variation of the ordinary R_p^2 statistic that reflects the number of factors in the model (Derksen and Keselman, 1992) and is computed by

$$\boldsymbol{R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}} \quad (5)$$

A predictor's relatively high variance inflation factor (VIF) value suggests that it might be collinear with other predictors Use the 0.95 quantiles of a chi-square distribution with p degrees of freedom as the cut value if the multivariate distribution is normal in the model (Denis, 2021). The VIF of the coefficient (β_p) estimator for p explanatory variables is given as

$$\boldsymbol{VIF_p = \frac{1}{(1-R_p^2)}} \quad (6)$$

While the tolerance is the reciprocal of the VIF, can also be used to identify collinearity in the regression model (Goldengorin et al., 2015; Park and Klabjan, 2020; Young, 2016) produced by

$$\boldsymbol{1 - R_p^2} \quad (7)$$

A key component of statistical analysis is the accurate detection of outliers. In multivariate settings, the masking effect also occurs, making the supplementary

visual inspection of the dataset much more difficult (Becker and Gather, 1999). While, Mahalanobis distances relying on the sample mean and covariance matrix struggle to identify all multivariate outliers in a dataset (Filzmoser et al., 2008; Mayrhofer and Filzmoser, 2023)

$$\boldsymbol{MD^2_{\mu,\Sigma}(x) = (x-\mu)'\Sigma'(x-\mu)} \quad (8)$$

Use the 0.95 quantile of a chi-square distribution with p degrees of freedom as the cut value if the multivariate distribution is normal (Rousseeuw and Zomeren, 2012).

Research Data

The secondary dataset was sourced from the Cotton Research Station (CRS) in Uthal, Lasbela, Balochistan. The dataset includes variables such as yield in kilograms per hectare (YldPhec) to regress on eight independent variables: germination (Grmn) in percentage, plant population (PlntPl), bolls per plant (BolPp), boll weight (BolWt) in grams, lint (Lnt) in percentage, staple length (StpLt) in millimetres, Mike value (MikV) in μ $inch^{-1}$, and fiber strength (FbrSnt) in G tex^{-1}.

Analytical procedure

Using RStudio, descriptive statistics were used to assess each quantitative variable's univariate normality, outliers, skewness, and kurtosis before the analytical process began. Variance inflation factors (VIFs), tolerance values, and correlation matrices were used to address potential multicollinearity. The assumptions of multiple regression were then investigated, including multivariate normality, linearity, heteroscedasticity, and constant variances. Outliers were also evaluated to improve the validity of the model.

Empirical Model

The empirical model used a multivariate regression approach to predict cotton yield based on independent variables. The model's assumptions were tested and validated to ensure the reliability of the predictions.

RESULTS AND DISCUSSION

To ensure the dataset's suitability for regression analysis, a descriptive analysis was conducted on 255 observations for each variable presented in Table 1. It offers key insights into their central tendencies, variability, and distribution shapes. Grmn exhibits moderate variability with a slight left skew and a relatively flat distribution. PlntPl demonstrates near symmetry and a distribution approximating normality. YldPhec shows high variability with a slight right skew and a flatter distribution. BolPp is characterized by a left skew and a more peaked distribution. BolWt displays a slight right skew and a flatter distribution. Lnt has a slight left skew and a flatter distribution. StpLt is nearly symmetrical with a flatter distribution. MikV shows a slight right skew and a distribution close to normal. FbrSnt exhibits a slight right skew and a flatter distribution.

Table 1. Descriptive statistics.

Variables	Mean	SD	Median	Min	Max	Range	Skew	Kurtosis	SE
Grmn	89.32	6.17	90.00	75.00	100.00	25.00	-0.41	-0.63	0.39
PlntPl	38109.70	2150.63	38587.50	32309.00	43078.70	10769.70	0.09	-0.15	134.68
YldPhec	2524.87	802.41	2512.90	1077.00	4307.90	3230.90	0.33	-0.82	50.25
BolPp	36.08	2.99	36.40	26.10	41.20	15.10	-0.93	1.27	0.19
BolWt	3.52	0.34	3.50	3.00	4.20	1.20	0.39	-0.95	0.02
Lnt	37.33	1.17	37.40	34.40	39.30	4.90	-0.27	-0.72	0.07
StpLt	26.82	0.78	26.70	25.20	28.40	3.20	-0.07	-0.69	0.05
MikV	4.32	0.39	4.30	3.30	5.20	1.90	0.50	-0.19	0.02
FbrSnt	26.96	1.08	26.80	24.60	29.30	4.70	0.25	-0.68	0.07

Table 2 shows the correlation matrices, including the correlation coefficients and corresponding p-values, indicating significant correlations among the variables, both positive and negative. The top section of the diagonal displays correlation coefficient values, ranging from -1 to 1, which indicate the strength and direction of the relationships between variable pairs. A value close to 1 signifies a strong positive correlation, while a value approaching -1 denotes a strong negative correlation (Izenman, 2013). A value around 0 indicates minimal to

no linear correlation between the variables. Its lower section (below the diagonal) shows the p-values corresponding to these correlations, where lower p-values indicate greater statistical significance. Significant positive correlations were observed between Grmn and PlntPl, YldPhec, and MikV. Conversely, Grmn exhibited negative correlations with BolPp, BolWt, Lnt, StpLt, and FbrSnt. PlntPl demonstrated significant positive correlations with YldPhec and MikV, while showing negative correlations with BolPp, BolWt, Lnt, StpLt, and FbrSnt. YldPhec was positively correlated with MikV and negatively correlated with BolPp, BolWt, Lnt, StpLt, and FbrSnt. BolPp had positive correlations with BolWt, Lnt, StpLt, MikV, and FbrSnt. BolWt was positively correlated with Lnt, StpLt, and FbrSnt, but negatively correlated with MikV. Lnt showed positive correlations with StpLt and FbrSnt, and a negative correlation with MikV. StpLt was positively correlated with FbrSnt and negatively correlated with MikV. MikV exhibited a negative correlation with FbrSnt.

Table 2. Correlation Matrices with P-values.

Predictor	Grmn	PlntPl	YldPhec	BolPp	BolWt	Lnt	StpLt	MikV	FbrSnt
Grmn		0.64	0.53	-0.28	-0.13	-0.42	-0.07	0.1	-0.13
PlntPl	0		0.47	-0.18	-0.23	-0.27	-0.13	0.22	-0.2
YldPhec	0	0		-0.24	-0.35	-0.4	-0.24	0.28	-0.11
BolPp	0	0	0		0.58	0.49	0.31	0.1	0.33
BolWt	0.04	0	0	0		0.16	0.27	-0.27	0.26
Lnt	0	0	0	0.01	0		0.41	-0.33	0.49
StpLt	0.25	0.04	0	0	0	0		-0.19	0.59
MikV	0.12	0	0	0.12	0	0	0		-0.33
FbrSnt	0	0	0.08	0	0	0	0	0	

The left section of the scatterplot matrix in Figure 1 visualizes the significant relationships between multiple variables at once and is useful for spotting correlations, patterns, and potential outliers in the data. Outliers in these scatterplots are identified as data points that significantly deviate from the main cluster of points. In addition, it also shows linear connections and is almost elliptic in shape. In the upper and right sections of the diagonal in Figure 1, positive correlations are typically represented by blue shades, while red shades depict negative correlations. Darker colors signify stronger positive or negative correlations, whereas lighter colors indicate weaker correlations. The data was analysed using the Mahalanobis distances (for df=8, quantile =0.10, Chi-square=14.68) to confirm the multidirectional extreme value. The following 25 observations were excluded from the analysis: 24, 28, 29, 33, 34, 35, 46, 49, 65, 68, 73, 75, 76, 77, 78, 79, 84, 92, 121, 123, 124, 138, 174 and 187. Before delving into multiple linear regression analyses, univariate normality was assessed for all quantitative variables. To scrutinize the univariate normality assumptions, the skewness and kurtosis coefficients of the variables were examined and presented in Table 3. Skewness values carry greater significance when assessing the normality assumption, whereas if the kurtosis coefficient does not deviate significantly from the normal distribution, it can be considered that this variable follows a normal distribution (Büyüköztürk, 2018). The skewness and kurtosis values are considered reliable indicators of distribution characteristics due to their small standard errors. The skewness values for all variables fall within the acceptance range of -0.38 and 0.70 and cannot be considered skewed. However, kurtosis analysis shows that all the variable's platykurtic values lie between 2.11 and 2.96.

The findings from the simple correlations, VIFs, and tolerance values were examined to determine whether there were any multiple relations between the variables. Collinearity in the explanatory variables is considered significant when the VIF exceeds 10 and tolerance values are less than 0.10 (Ott and Longnecker, 2010). Hence, higher VIF values indicate a more serious impact of collinearity on the accuracy of slope estimation. Upon examining the simple correlations, it is noted that all correlation coefficients have absolute values below 0.57. Furthermore, Table 4 indicates that no two predictor variables exhibit a perfect relationship. Additionally, all tolerance values are greater than 0.38, while the VIF values for all variables are less than 3

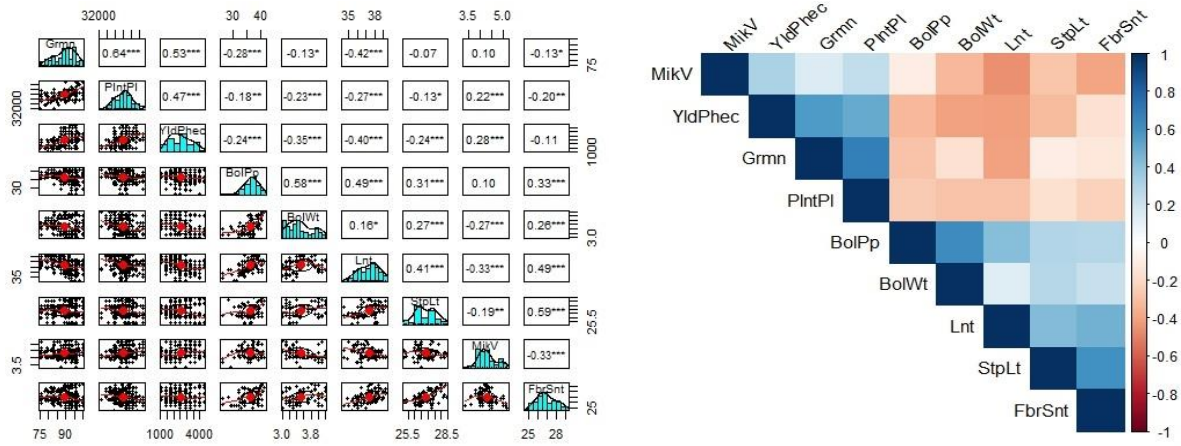

Figure 1. Scatterplot Matrix and Heatmap

Table 3. Univariate normality.

Variables	N	Skewness		Kurtosis	
		Statistic	Std. Error	Statistic	Std. Error
YldPhec	230	0.38	0.16	2.24	0.32
Grmn	230	-0.36	0.16	2.30	0.32
PlntPl	230	0.07	0.16	2.80	0.32
Lnt	230	-0.31	0.16	2.28	0.32
BolWt	230	0.39	0.16	2.10	0.32
BolPp	230	-0.38	0.16	2.96	0.32
StpLt	230	-0.18	0.16	2.35	0.32
MikV	230	0.70	0.16	2.69	0.32
FbrSnt	230	0.23	0.16	2.33	0.32

Table 4. Multiple Relations Correlations

Variables	Zero Order	Partial	Part	Tolerance	VIF
Grmn	0.56	0.33	0.25	0.44	2.27
PlntPl	0.50	0.11	0.08	0.47	2.13
BolPp	-0.32	0.10	0.07	0.39	2.55
BolWt	-0.41	-0.27	-0.20	0.41	2.44
Lnt	-0.42	-0.19	-0.13	0.41	2.45
StpLt	-0.32	-0.25	-0.18	0.57	1.76
MikV	0.32	0.08	0.06	0.54	1.84
FbrSnt	-0.16	0.23	0.17	0.53	1.88

After scrutinizing the data to confirm adherence to assumptions and subsequently addressing any issues, a multiple linear regression analysis was conducted. The results, which examine the significance of the eight independent variables concerning the dependent variable, YldPhec (Table 5). The ANOVA reveals the impact of eight predictors on the dependent variable YldPhec significantly across 230 observations. The model explains 51% of the variance in YldPhec, as evidenced by the R^2 of 0.51 and an adjusted R^2 of 0.49, indicating a moderate fit. The value of the standardized beta coefficient (std. Beta) in Table 6 reflects the relative importance, standardized strength, and direction of the relationships between each predictor and the dependent variable in the regression model (Imdadullah, 2017). Its largest absolute value for Grmn is 0.37, which indicates its strongest relationship with YldPhec compared to other predictors. Following Grmn, there are BolWt, StpLt, FbrSnt, Lnt, PlntPl, BolPp and MikV. This order signifies the decreasing strength of their relationships with the YldPhec. The lowest absolute value for MikV is 0.08, revealing weakest relationship with the YldPhec after controlling for other predictors in the model.

Table 5. Analysis of variance.

SV	SS	DF	MS	F	Sig.
Regression	74395176.11	8	9299397.01	28.68	0.00
Residual	71652210.84	221	324218.14		
Total	146047386.9	229			

Residual standard error: 569.40 on 221; R-squared: 0.51 Adjusted R-squared: 0.49

Table 6. Parameter estimation and analysis.

Model	Beta	Std.Error	Std.Beta	T	95% CI
Intercept	4425.26^{0}	2488.71		1.78	-479.38 — 9329.89
Grmn	48.76***	9.28	0.37	5.26	30.48 — 67.04
PlntPl	0.04^{0}	0.03	0.12	1.69	-0.01 — 0.09
BolPp	33.97	23.92	0.11	1.42	-13.16 — 81.10
BolWt	-741.49***	175.83	-0.31	-4.22	-1088.01 — -394.98
Lnt	-145.57**	51.78	-0.21	-2.81	-247.61 — -43.53
StpLt	-246.77***	64.85	-0.24	-3.81	-374.58 — -118.96
MikV	168.02	141.28	0.08	1.19	-110.41 — 446.45
FbrSnt	177.97***	49.88	0.23	3.57	79.67 — 276.27

Significance: 0 '***' 0.001 '**' 0.01 '*' 0.05 '^{0}' 0.1 ''

However, in terms of individual predictors, Grmn and FbrSnt positively influence YldPhec significantly, suggesting that their increases result in a proportional increase in YldPhec. Similarly, PlntPl and Intercept show substantial positive effects on YldPhec, being statistically significant. Conversely, BolWt, Lnt, and StpLt negatively affect YldPhec, implying that their higher values correspond to lower YldPhec. Lastly, BolPp and MikV have no statistical significance, and their coefficients are not significantly different from zero. The multiple linear regression equation is based on the analysis's findings and is written as below:

$$\begin{aligned} \text{YldPhec} = {} & 4425.26 + 48.76\,\text{Grmn} + 0.04\,\text{PlntPl} \\ & + 33.97\,\text{BolPp} - 741.49\,\text{BolWt} \\ & - 145.57\,\text{Lnt} - 246.77\,\text{StpLt} + 168.02\,\text{MikV} + \\ & 177.97\,\text{FbrSnt} \quad (9) \end{aligned}$$

The test results evaluating a regression model's assumptions are presented in Table 7. The Shapiro-Wilk test yields a p-value of 0.74, indicating no significant deviation from normality in the residuals. Similarly, the Rainbow test result of 0.84, with 0.82 p-value, provides no statistically significant evidence against the null hypothesis of linearity, thereby supporting the model's linearity assumption. The Studentized Breusch-Pagan, and Breusch-Pagan tests for constant, and homogeneity variance produce p-values of 0.32, and 0.08, respectively, suggesting no significant violations of these assumptions. Therefore, the assumptions of normality, linearity, and constant and homogeneous variance in the model's residuals are not significantly violated.

The residual histogram, box, and normal Q-Q plots in Figure 2 reveal that the residuals are generally normally distributed, although there are some deviations in the tails. These deviations suggest that further analysis is needed to assess the extent of non-normality and determine if it is significant. However, this is not considered problematic and is not severe enough to invalidate concerns about the model's performance.

Table 8 presents the outlier test results for the model's residuals. Observation 171, with a studentized residual of 3.0, significantly deviates from the mean, suggesting it could be an outlier based on the unadjusted p-value is 0.003. However, after applying the Bonferroni correction, the adjusted p-value becomes 0.67, indicating that the observation is not statistically significant when accounting for multiple comparisons.

Table 7. Test of Assumption

Test	Shapiro-Wilk	Rainbow	Studentized Breusch-Pagan	Breusch-Pagan
Statistics	0.99	0.84	9.25	3.09
DF	-	115, 106	8	1
P-value	0.74	0.82	0.32	0.08

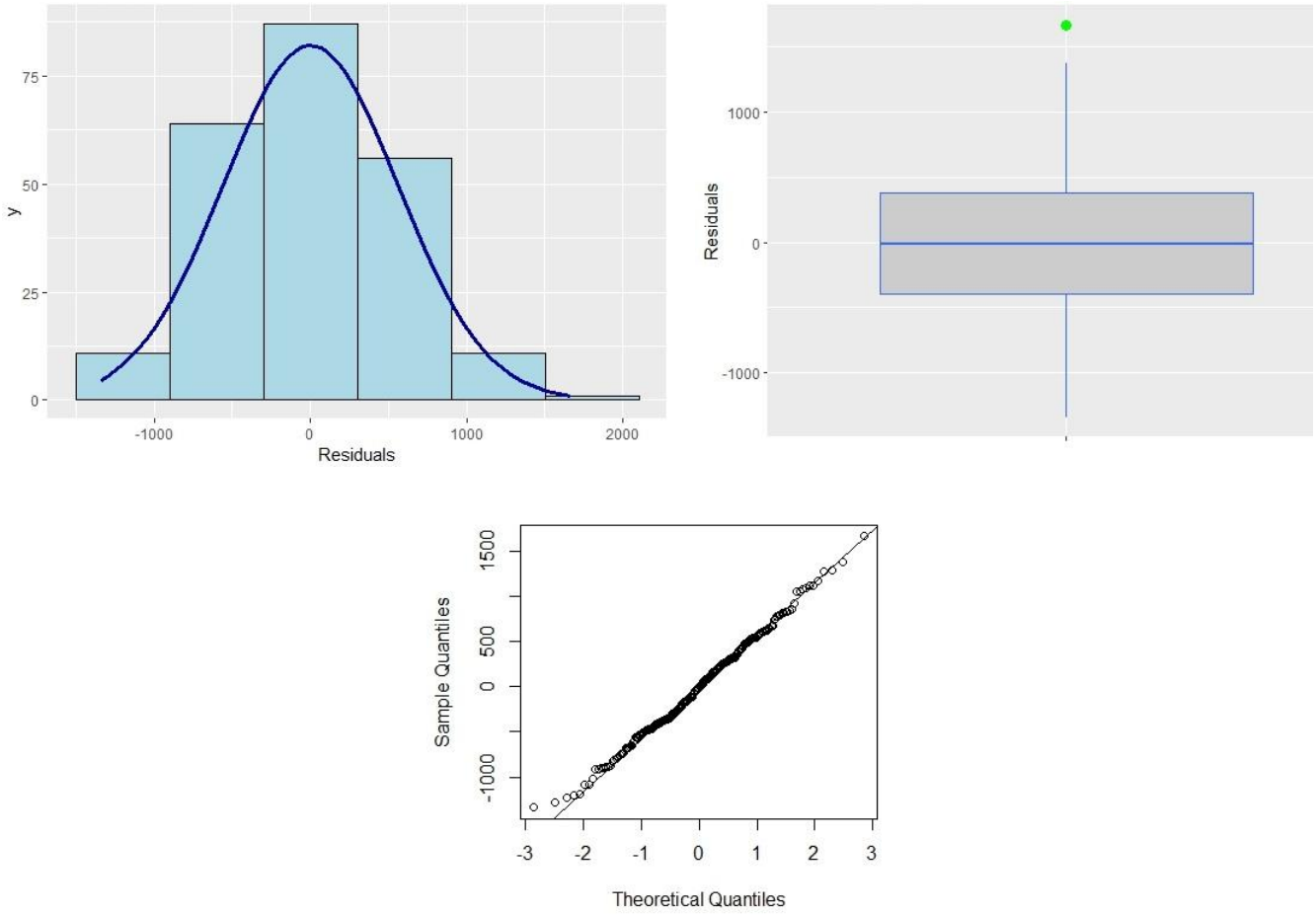

Figure 2. Residual Histogram, Box, and QQ Plot.

Table 8. Test of studentized residuals.

Observation	Studentized Residual	Unadjusted P-value	Bonferroni P-Value
171	3.00	0.003	0.67

Table 9. Test of Skewness and Kurtosis

Test	Statistics	SE	t-value	P-value
Skewness	0.13	0.16	0.79	0.21
Kurtosis	-0.19	0.32	-0.57	0.28

Table 9 demonstrates that the model's residuals approximate normal distribution at the 0.05 significance level. The skewness statistic of 0.13 indicates near symmetry, while the kurtosis statistic of -0.19 suggests the residuals exhibit slightly less peakedness compared to a normal distribution.

CONCLUSION

This study successfully investigated the causal connections among the dependent and independent variables in the context of cotton yield and validated the multivariate assumptions. The descriptive statistics reveal that YldPhec, Grmn, and MikV showed near-normal distributions, while PlntPl, BolPp, BolWt, Lnt, StpLt, and FbrSnt displayed slight skewness and flatter distributions. The correlation analysis highlights a significant positive correlation of Grmn, PlntPl, and MikV with YldPhec and a negative correlation of BolPp, BolWt, Lnt, StpLt, and FbrSnt with YldPhec. The scatterplot matrix effectively visualizes these correlations, patterns, and potential outliers, providing a comprehensive dataset overview. The study provides valuable insights

into the relationships between various agronomic predictors and YldPhec, informing breeding programs and agronomic practices aimed at optimizing yield through the management of key predictors such as germination, boll weight, lint, staple length, and fiber strength.

Outlier analyses identified 25 observations as outliers, subsequently excluded from the study. Univariate normality assessments indicate skewness values fall within acceptable ranges, while kurtosis values suggest flatter distributions. Collinearity assessments indicate no significant multicollinearity, with all VIF values below 10 and tolerance values above 0.10. The study assessed the strength of these connections and their influence on the prediction of cotton yield, with the regression model explaining 51% of the variance in YldPhec. Significant predictors included Grmn and FbrSnt, which had positive effects, and BolWt, Lnt, and StpLt, which had negative effects, underscoring the importance of these parameters in determining yield.

The validation of the multivariate regression model assumptions confirmed the normality of residuals, the model's linearity, and no significant heteroscedasticity. However, there was marginal evidence of some concerns regarding homogeneity variance. Meeting these assumptions is required for a valid multilinear regression analysis. Confirming these assumptions enhances the model's validity and underscores the significance of the identified predictors in explaining the variability of YldPhec. Finally, the impact of specific independent variables or groups of variables on cotton yield was evaluated under controlled conditions. An outlier, Observation 171, in the model residuals, was noted; however, it was not statistically significant after adjustment for multiple comparisons.

To enhance cotton yield, it is recommended to focus on agronomic practices that improve Grmn, PlntPl, MikV, and FbrSnt due to their positive effects. Conversely, it is crucial to address and mitigate the negative impacts of BolPp, BolWt, Lnt, and StpLt. Ensuring the exclusion of outliers and maintaining acceptable skewness and kurtosis values will support the robustness of the analysis, while monitoring collinearity will ensure no significant multicollinearity affects the results.

To ensure the validity of the multivariate regression model for predicting cotton yield, it is recommended to confirm the normality of residuals, linearity, and absence of significant heteroscedasticity while addressing any concerns regarding the homogeneity of variance.

ACKNOWLEDGEMENTS

Mr. Sultan Ahmed Baloch, scientific officer at the Cotton Research Station (CRS), Uthal, Lasbela, Balochistan, Pakistan, has significantly contributed to the advancement of the field of cotton research through his invaluable insights and support. We would also like to express heartfelt gratitude to all Ph.D. scholars, with a special mention to Mr. Lutfullah Rodini, Mr. Ahmed Khan Memon, Mr. Asif Arain, and Dr. Shahmir Ali Kalhoro for their unwavering support and encouragement throughout this journey.

REFERENCES

Abubakar, M., Sheeraz, M., Sajid, M., Mehmood, Y., Jamil, H., Irfan, M., and Shahid, M. 2023. Analysis of Cotton Value Chain in Pakistan: Identifying the Process and Critical Factors in Sustainable Agribusinesses. Journal of Arable Crops and Marketing, 5(2): 63–74.

Becker, C., and Gather, U. 1999. The Masking Breakdown Point of Multivariate Outlier Identification Rules. Journal of the American Statistical Association, 94(447): 947–955.

Brillinger, D. R., Malinvaud, E., and Silvey, A. 1967. Statistical Methods of Econometrics. Economica, 34(136): 451.

Büyüköztürk, Ş. 2018. Sosyal bilimler için veri analizi el kitabı. Sosyal Bilimler Için Veri Analizi El Kitabı, 1–214.

Chatterjee, S., and Hadi, A. S. 2012. Regression Analysis by Example (Fifth). A Johan Wiley and Sons, Inc., Publication.

Denis, D. J. 2021. Simple and Multiple Linear Regression. In Applied Univariate, Bivariate, and Multivariate Statistics Using Python.

Derksen, S., and Keselman, H. J. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 45(2): 265–282.

Dhamodharavadhani S., and Rathipriya R. 2021. Variable Selection Method for Regression Models Using Computational Intelligence Techniques. In Handbook of Research on Machine and Deep

Learning Applications for Cyber Security (pp. 417–236). IGI Global.

Faraway, J. J. 2002. Practical Regression and Anova using R (Third). www.r-project.org.

Filzmoser, P., Maronna, R., and Werner, M. 2008. Outlier identification in high dimensions. Computational Statistics and Data Analysis, 52(3): 1694–1711.

Finch, W. H., Bolin, J. E., and Kelley, K. 2016. Multilevel Modeling Using R. Chapman and Hall/CRC.

Frederick, O., Maxwell, O., Ifunanya, O., Udochukwu, E., Kelechi, O., Ngonadi, L., and Idris, H. K. 2019. Comparison of Some Variable Selection Techniques in Regression Analysis. American Journal of Biomedical Science and Research, 6(4): 281–293.

Frost, J. 2019. Regression Analysis (First Edit).

Goldengorin, B. I., Malyshev, D. S., Pardalos, P. M., and Zamaraev, V. A. 2015. A tolerance-based heuristic approach for the weighted independent set problem. Journal of Combinatorial Optimization, 29(2): 433–450.

Graybill, F. A. 1961. An introduction to linear statistical models.

Haldar, S., and Miller, A. J. 1992. Subset Selection in Regression. In Journal of Marketing Research (Second, Vol. 29, Issue 2). Chapman and Hall/CRC.

Hastie, T., Tibshirani, R., James, G., and Witten, D. 2021. An Introduction to Statistical Learning with Application in R. In Springer Texts (Second Edi, Vol. 102). Springer Science+Business Media.

Imdadullah, M. 2017. Addressing Linear Regression Models with Correlated Regressors: Some Package Development in R. Bahuddin Zakariya University Multan, Pakistan.

Izenman, A. J. 2013. Modern Multivariate Statistical Techniques Regression, Classification, and Manifold Learning (2nd ed.) Springer.

Jakešová, J. 2014. The validity and reliability study of the Czech version of the motivated strategies for learning questionnaire (MSLQ). New Educational Review, 35(1): 54–65.

Kawano, S., Fukushima, T., Nakagawa, J., and Oshiki, M. 2023. Multivariate regression modeling in integrative analysis via sparse regularization.

Mayrhofer, M., and Filzmoser, P. 2023. Multivariate outlier explanations using Shapley values and Mahalanobis distances. Econometrics and Statistics, xxxx, 21.

Naseer, M. A. ur R., Ashfaq, M., Razzaq, A., and Ali, Q. 2020. Comparison of water use efficiency, profitability and consumer preferences of different rice varieties in Punjab, Pakistan. Paddy and Water Environment, 18(1): 273–282.

Naveed, M., Maqsood, M. F., and Cheema, A. R. 2024. Modeling the contribution of district-level cotton production to aggregate cotton production in Punjab (Pakistan): an empirical evidence using correlated component regression approach. Journal of Excellence in Social Sciences, 3(3), 147–164.

Ott, L., and Longnecker, M. 2010. An introduction to statistical methods and data analysis. Brooks/Cole Cengage Learning.

Park, Y. W., and Klabjan, D. 2020. Subset selection for multiple linear regression via optimization. Journal of Global Optimization, 77(3): 543–574.

Poon, E., and Feng, C. 2023. Univariate and Multiple Regression Analyses in Medical Research. Biometrical Letters, 60(1): 65–76.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. 1998. Applied Regression Analysis: A Research Tool (Second). Springer-Verlag New York Berlin Heidelberg.

Rehman, A., Jingdong, L., Shahzad, B., Chandio, A. A., Hussain, I., Nabi, G., and Iqbal, M. S. 2015. Economic perspectives of major field crops of Pakistan: An empirical study. Pacific Science Review B: Humanities and Social Sciences, 1(3): 145–158.

Rousseeuw, P. J., and Zomeren, B. C. Van. 2012. Unmasking Multivariate Outliers and Leverage Points Unmasking Multivariate Outliers and leverage Points. Journal of the American Statistical Association, 85(411):, 633–639.

Sestelo, M., Villanueva, N. M., Meira-Machado, L., and Roca-Pardiñas, J. 2016. FWDselect: An R package for variable selection in regression models. R Journal, 8(1): 132–148.

Shakeel, M., Hassan, ul H., Chaudhry, K. A., and Tahir, M. N. 2023. View of What Affects Crop Production in Pakistan_ The Role of Agriculture Employment, Machinery and Fertilizer Consumption.pdf. Bulletin of Business and Economics, 12(3): 541–546.

Tabachnick, B., and Fidel, L. S. 2019. Using Multivariate Statistics. In Pearson Education, Inc. (Seventth). Pearson.

Ünver, Ö., and Gamgam, H. 1999. Uygulamalı istatistik yöntemler. Siyasal Kitabevi.

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. 2020. A simple new approach to variable selection in regression, with application to genetic fine mapping. Journal of the Royal Statistical Society. Series B: Statistical Methodology, 82(5): 1273–1300.

Wen, C., Zhang, A., Quan, S., and Wang, X. 2017. BeSS: An R Package for Best Subset Selection in Linear, Logistic and CoxPH Models.

Young, D. S. 2016. Normal tolerance interval procedures in the tolerance package. R Journal, 8(2): 200–212.

Zhang, F. 2022. Economic Research on Multiple Linear Regression in Fruit Market inspection and Management. Applied Mathematics and Nonlinear Sciences, 8(1): 1951–1966.