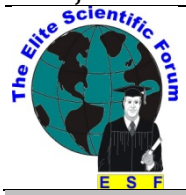
Pak. j. life soc. Sci. (2015), 13(2): 86-90 E-ISSN: 2221-7630;P-ISSN: 1727-4915




**Pakistan Journal of Life and Social Sciences**
www.pjlss.edu.pk

RESEARCH ARTICLE

# Fitting Regression Model with Some Assumptions for the Identification of Cotton Yield Factors

Muhammad Amin*, Atif Akbar and Muhammad Awais Manzoor
Department of Statistics, Bahauddin Zakariya University, Multan

***ARTICLE INFO***
Received: Feb 12, 2015
Accepted: Aug 10, 2015
Online: Aug 20, 2015

***Keywords***
Cotton yield
Cotton yield factors
Influential Observations
Multicollinearity
Regression modeling

***Corresponding Author:**
ma_amin15@yahoo.com

**ABSTRACT**

In agriculture and related fields many relationships exist that need to be identified in quantitative way. Regression modeling plays an important role for the determination of such relationships and also the isolation of factors that greatly affect the target or response variable. For reliable and valid results, one has to check the regression assumptions like influential observations, multicollinearity etc. In this study, we have fitted the regression models with and without satisfying the some regression assumptions for the identification of cotton yield factors. For the analysis purposes, the required data was collected from the district Khanewal. It was observed when regression assumptions were satisfied, model goodness ($R^2$) was improved from 68% to 92%, $R^2$ (adjusted) was improved from 62% to 90%) and standard error of the estimates reduced from 8.298 to 2.348. These better results indicated that the pesticide used for seed, all type of fertilizers (DAP, potash and Ganwara), water frequency, and previously sown crops were the significant factors for cotton yield with p-values as less than 0.05. So cotton yield was 90% explained by these factors.

## INTRODUCTION

Cotton is the main cash crop of national economy of Pakistan and is the second most abundant source of edible oil in the world (Ali et al., 2009). Cotton is important for textile products which are commonly used in daily life. Various factors affect the cotton yield like cotton variety, seed, water, pesticides and fertilizers used etc. The use of modern technologies has led to the development of new hybrid lines of cotton and farmer got the cotton varieties which produce maximum yield with minimum cost. Worldwide, Pakistan is the fourth largest cotton producer. Cotton crop itself and its products contribute 10% in GDP of Pakistan (Muhammad et al., 2009).

Many factors affect the cotton yield and are responsible to decrease the cotton yield. In the literature, different researchers have given the different opinion for the identification of cotton yield factors. Ali (1983) identified the biological factors under socioeconomic constraints. Khan *et al.* (1986) and Hassan (1991) found that unskilled farmers, low finance and marketing facilities, and lack of agricultural inputs were responsible for low crop yield. Nabi (1991) found that farm size, labor, seed, fertilizer, irrigation, number of cultivation and working capital were the important factors in the production process. Anwar (1998) showed that the management of these variables can increase the cotton yield. Bakhsh et al. (2005) found that education, plant protection measures, fertilizer and land preparation are the most contributing factors for obtaining the maximum cotton yield.

Chaudhry et al. (2009) studied the cotton yield and related factors i.e. cultivation, seed, DAP, urea, irrigation, plant protection and hoeing in the Multan. They have used the Cobb-Douglas production function for that purpose and found that all these factors are significant for cotton yield. Muhammad et al. (2009) studied the effect of weed management on cotton productivity and have found that weed management provides better cotton yield. Saleem et al. (2009) studied the effect of row spacing on earliness and cotton yield and have found that cotton is maximum when row spacing was 75 cm. Saleem et al. (2010) studied the effect of nitrogen on seed cotton yield and fiber quality and have concluded that varieties are not significant for boll weight and seed cotton yield. Sui et al. (2013) gave the method to assess the cotton yield using plant height mapping system.

One of the main purpose of the regression model is to identify the factors which contribute maximum to the response variable. Regression models results are only reliable if results met the regression assumptions. These assumptions include no multicollinearity, constant error variance, no autocorrelation, no outlier and influential observation etc. (Pardoe, 2012). In this article, we have paid special attention to the regression model assumptions diagnostics and its impact on the cotton yield model for the identification of factors. Regression diagnostics includes outliers and influential observations analysis, as these are the observations which are not compatible with the remaining data values/observations and these can effect on the model estimates and predicted values. With the presence of these values, the fitted model results may indicate the significance of the factors which are actually playing no role in the response variables and vice versa. In the similar manner, with the exclusion of these data values may be some factors showing significance to explain the response variable. Outliers are tested by studentized residuals while influential observations are tested by Cook's distance method as proposed by Cook (1977).

So under these circumstances, the main theme of this research was to search cotton yield model and associated factors without outlier and influential observation. Also, the comparison of the fitted model with and without outlier and influential observation has been presented. Our interest was not only to identify the outlier and influential observations but also model fitting with and without outlier and influential observations. Some additional factors that were not considered in previous studies of cotton yield model had also been added in this study.

## MATERIALS AND METHODS

Plant experiment evaluation is the basic need for agriculture research for the identification of significant factors which are responsible to change the output of plant experiments. The data was collected from the two tehsils of district Khanewal of Multan division. The questionnaire was given to the farmers and ask the questions regarding cotton yield, farmer education, cottonseed variety, pesticide used for cotton seed and for crop, fertilizer (DAP, Potash, Urea), land type, bushes elimination and previously sown crop. Cotton yield measured in Munds per acre, famer education categorized as illiterate (0), Primary (1), middle (2), Matric (3), Intermediate (4), Graduate (5) and Masters (6). And another factor is the cotton verity with codes; Bt.009 (1), Bt.121(2), Bt.142(3), Bt.456(4), Bt.703(5), Bt.886(6) and SJ1(7). Pesticide used for cottonseed variety is measured dichotomously as yes (1) or No (0), pesticide for a cotton crop is noted and asked in frequency form. Similarly types fertilizer (Urea, DAP, Potash, and Ganwara) counted in frequency per acre. Similarly, bushes elimination is if yes (1) otherwise No (0). Land type is categorized as Kachi (0), Paki (1), and Ratli (2). Sometimes the previously sown crop may also affect the cotton yield here categorized as if the previously sown crop is cotton (1) and if wheat is the previously sown crop then coded as 2. These variables with description and notation are represented in table 1.

Multiple regression analysis with qualitative/categorical and quantitative factors is used for the identification of cotton yield factors. Here we can't use stepwise regression because some factors' effect is hidden due to regression assumptions. Here we use some qualitative factors to study the effect of qualitative factors on cotton yield. The outlier is tested through studentized test while influential observation is tested through Cook's distance method. Other assumptions like multicollinearity are tested through variance inflation factor (VIF), autocorrelation is tested through DW-test and heteroscedasticity is tested through Glejser test. In Glejser test, we regressed the absolute estimated residuals on the squared independent variables and test the significance by F-test. For further details about these assumptions tests, reader are referred to Pardoe (2012) and Gujarati (2004). For the analysis purposes, we use statistical software, Sigma XL V.7.

## RESULTS

This section contains the analysis regarding cotton yield and associated factors. Also, the analysis with and without regression assumptions (no outlier, influential observation, autocorrelation, multicollinearity, and heteroscedasticity) was made for model goodness of fit and also to identify factors which contribute cotton yield.

**Regression model for cotton yield per acre without satisfying the regression assumptions**

The fitted model with violating the regression assumptions and with 12 independent variables which assumed to be responsible for cotton yield was given as;

y = 28.353 + 0.2899 x4 + 0.2452 x5 + 8.157x6 + 3.564x7 + 0.4316x8 + 0.2599x9 -1.553x1_1 - 3.379x1_2 + 0.2742x1_3 + 2.925x1_4 - 4.536x1_5 - 2.817x1_6 + 2.51 x2_2 + 9.554x2_3 - 1.386x2_4 + 8.497x2_5 + 0.8266x2_6 + 4.035x2_7 + 2.147x3_1 - 1.296x10_1 - 5.781x10_2 - 1.648 x11_1 - 10.441x12_2

From Table 2, it was found that all variables contributed in cotton yield as shown by p-value (0.0000) of the F-test. $R^2$ value (68%) indicated that the model as fitted explained 68% of the variability in cotton yield per acre model (Table 3). The adjusted $R^2$, a more suitable parameter for comparing models with different numbers of independent variables, was 62.16%. The standard error of the estimate showed that the

**Table 1: Cotton yield and its associated factors description and notation**

| Sr. No. | Name of variable | Notation | Variables Nature | Unit |
|---|---|---|---|---|
| 1 | Cotton Yield per acre | $\Gamma$ | Continuous | Munds |
| 2 | Farmer Education | $\chi_1$ | Categorical | |
| 3 | Cotton variety | $\chi_2$ | Categorical | |
| 4 | Pesticide used for seed | $\chi_3$ | Dichotomous | |
| 5 | Pesticide frequency | $\chi_4$ | Discrete | |
| 6 | Urea | $\chi_5$ | Discrete (frequency) | 50 kg per acre |
| 7 | DAP | $\chi_6$ | Discrete (frequency) | 50 kg per acre |
| 8 | Potash | $\chi_7$ | Discrete (frequency) | 50 kg per acre |
| 9 | Ganwara | $\chi_8$ | Discrete (frequency) | 50 kg per acre |
| 10 | Water frequency | $\chi_9$ | Discrete (frequency) | |
| 11 | Land type | $\chi_{10}$ | Categorical | |
| 12 | Bushes | $\chi_{11}$ | Dichotomous | |
| 13 | Previously sown crop | $\chi_{12}$ | Categorical | |

**Table 2: Cotton yield model estimates with and without problematic observations**

| Predictor | Full Data | | | | | After excluding outliers and Influential observations | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Term | Coefficient | SE Coefficient | T | P | VIF | Coefficient | SE Coefficient | T | P | VIF |
| Constant | 28.3528 | 6.5857 | 4.3052 | 0.0000 | | 42.5316 | 2.5546 | 16.6490 | 0.0000 | |
| x4 | 0.2900 | 0.4857 | 0.5969 | 0.5516 | 1.8761 | 0.2107 | 0.2599 | 0.8107 | 0.4204 | 2.1953 |
| x5 | 0.2452 | 0.7337 | 0.3342 | 0.7388 | 2.1955 | 0.3711 | 0.3064 | 1.2113 | 0.2300 | 2.2755 |
| x6 | 8.1573 | 1.6331 | 4.9950 | 0.0000 | 2.2059 | 3.0064 | 1.0775 | 2.7902 | 0.0068 | 5.1943 |
| x7 | 3.5644 | 1.7392 | 2.0495 | 0.0425 | 1.1724 | 3.6081 | 0.8280 | 4.3576 | 0.0000 | 1.8456 |
| x8 | 0.4316 | 1.0128 | 0.4261 | 0.6707 | 1.6771 | 1.5784 | 0.5352 | 2.9490 | 0.0044 | 3.1520 |
| x9 | 0.2599 | 0.2617 | 0.9931 | 0.3226 | 2.7136 | -0.6862 | 0.1333 | -5.1493 | 0.0000 | 2.4095 |
| x1_1 | -1.5529 | 2.5459 | -0.6100 | 0.5430 | 2.2017 | -1.1449 | 1.0350 | -1.1062 | 0.2726 | 2.7283 |
| x1_2 | -3.3787 | 2.8174 | -1.1992 | 0.2327 | 1.3685 | -2.3894 | 0.9740 | -2.4531 | 0.0167 | 1.4874 |
| x1_3 | 0.2742 | 2.1530 | 0.1274 | 0.8989 | 2.1970 | 0.0927 | 0.8009 | 0.1158 | 0.9082 | 2.1977 |
| x1_4 | 2.9253 | 3.5119 | 0.8330 | 0.4064 | 1.6716 | 2.5978 | 1.3335 | 1.9481 | 0.0555 | 1.7544 |
| x1_5 | -4.5362 | 2.7510 | -1.6489 | 0.1017 | 1.7407 | -1.6649 | 1.1034 | -1.5089 | 0.1360 | 1.9087 |
| x1_6 | -2.8173 | 5.6226 | -0.5011 | 0.6172 | 1.3497 | -2.4684 | 3.0753 | -0.8027 | 0.4250 | 1.6455 |
| x2_2 | 2.5095 | 3.6598 | 0.6857 | 0.4942 | 2.1473 | 2.4716 | 1.2725 | 1.9422 | 0.0563 | 2.5389 |
| x2_3 | 9.5540 | 3.5626 | 2.6818 | 0.0083 | 2.7780 | 4.7374 | 1.2729 | 3.7217 | 0.0004 | 1.8433 |
| x2_4 | -1.3857 | 5.1719 | -0.2679 | 0.7892 | 1.5123 | -0.6233 | 2.1185 | -0.2942 | 0.7695 | 1.5446 |
| x2_5 | 8.4973 | 3.8179 | 2.2257 | 0.0278 | 2.3368 | 2.3357 | 3.0889 | 0.7562 | 0.4522 | 1.6601 |
| x2_6 | 0.8266 | 2.7772 | 0.2976 | 0.7665 | 4.1096 | 3.5581 | 0.9493 | 3.7483 | 0.0004 | 3.3975 |
| x2_7 | 4.0354 | 5.3593 | 0.7530 | 0.4529 | 1.6238 | 5.6949 | 2.8285 | 2.0134 | 0.0480 | 1.3920 |
| x3_1 | 2.1466 | 1.9837 | 1.0821 | 0.2813 | 1.8000 | 2.8959 | 0.8385 | 3.4535 | 0.0010 | 2.4562 |
| x10_1 | -1.2957 | 2.0012 | -0.6474 | 0.5185 | 2.0409 | 0.3228 | 0.8483 | 0.3806 | 0.7047 | 2.8618 |
| x10_2 | -5.7811 | 4.9353 | -1.1714 | 0.2437 | 8.2741 | -0.0051 | 1.7657 | -0.0029 | 0.9977 | 9.9848 |
| x11_1 | -1.6478 | 4.0256 | -0.4093 | 0.6830 | 5.3590 | -1.1877 | 1.2284 | -0.9669 | 0.3370 | 4.7085 |
| x12_2 | -10.4407 | 1.9525 | -5.3474 | 0.0000 | 1.8069 | -15.0348 | 1.0222 | -14.7087 | 0.0000 | 2.7708 |

**Table 3: Model summary and assumptions tests with and without outliers and Influential observations**

| Tools | | Without satisfying assumptions | With satisfying assumptions |
|---|---|---|---|
| $R^2$ | | 68% | 92.63% |
| $R^2$ (Adjusted) | | 62.16% | 90.13% |
| Standard Error of the estimates | | 8.298 | 2.348 |
| F-test | | 11.64 | 37.14 |
| P-value | | 0.0000 | 0.0000 |
| Durbin Watson (DW) test | *Statistic* | 1.158 | 2.006 |
| | *P-value* | 0.0000 | 0.3370 |
| Glejser test (Heteroscedasticity) | *Statistics* | 2.073 | 1.834 |
| | *P-value* | 0.0058 | 0.0284 |

standard deviation of the residuals was 8.298. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the P-value was less than 0.05, there was an indication of possible serial correlation at the 95.0% confidence level. We had also found from the Glejser test that there was heteroscedasticity in the errors as p-value for this test was 0.0000. To determine which variable was contributing and which one was more

significant as other we considered table 2. We found that seed variety Bt.142 and Bt.703 were the significant cottonseed varieties. Cottonseed variety Bt.142 might be beneficial for the farmers to increase the better yield as indicated by the rate of change. While Bt-703 was also significant but not beneficial for the farmers because this seed users attained low cotton yield as Bt.142. DAP and previously sown crop were the significant factors for the cotton yield. While other are insignificant but they were contributed in cotton yield but hidden due to regression problems as we discussed earlier. This might be due to the fact that there were outliers and influential observations in the data values which may cause for autocorrelation and hetero-scedasticity in the fitted model residuals. So these data values were removed for better results. In the next section, a similar model was fitted after deleting the outliers and influential observations.

**Multiple Regression model for cotton yield per acre with satisfying model assumptions**

The output showed the results of fitting a multiple linear regression model after deleting all influential observations to describe the relationship between cotton yield per acre and 12 independent (qualitative and quantitative) variables. The equation of the fitted model was;

y = 42.532 + 0.2107x4 + 0.3711x5 + 3.006x6 + 3.608x7 + 1.578x8 - 0.6862x9 - 1.145x1_1 - 2.389x1_2 + 0.0927x1_3 + 2.598x1_4 - 1.665x1_5 - 2.468x1_6 + 2.472x2_2 + 4.737x2_3 - 0.6233x2_4 + 2.336x2_5 + 3.558x2_6 + 5.695x2_7 + 2.896x3_1 + 0.3228x10_1 - 0.0051x10_2 - 1.188x11_1 - 15.035x12_2.

Since the P-value of the F-test in the Table 3 is less than 0.05, there was a statistically significant relationship between the cotton yield and associated factors with the exclusion of substantial values observations.

The $\boldsymbol{R}^2$ indicated that the model as fitted explained 92.63% of the variability in cotton yield per acre. The adjusted R-squared statistic, that was more suitable for comparing models with different numbers of independent variables, was 90.13%. The standard error of the estimate showed the standard deviation of the residuals to be 2.348 which was much smaller as model fitted variability with full data. This value could be used to construct prediction limits for new observations. The DW statistic testing the residuals showed that P-value was greater than 0.05, there was no indication of serial autocorrelation in the residuals at the 95.0% confidence level.

After deleting all influential observations, the results of the fitted model as shown in Table 3 were entirely different as compared to the full data model results. From Table 2, it was found that pesticide used for seed, all type of fertilizers (DAP, potash and Ganwara), water frequency, land type and previously sown crops were contributed significantly towards cotton yield. So it could be speculated that all these factors strongly affected the cotton yield. Increase in water frequency reduced the cotton yield as results have shown.

On comparing the fitted models as shown in Table 3, significant results satisfying some of the regression assumptions (no multicollinearity, autocorrelation, influential observations, and outliers) were found. It was found that the $R^2$ and $R^2$ (Adjusted) are increased from (68 % to 92.36%) and (62.16% to 90.13%), respectively (Table 3). The standard error of the estimates decreased from (8.298 to 2.348) and indicated the better results with satisfying regression assumptions. The degree of heteroscedasticity was reduced with omitting influential observation as Glejser test p-value (0.0058) to (0.0284). It was observed that the exclusion of outlier and influential observation may overcome the other problems like autocorrelation and heteroscedasticity and multicollinearity.

## DISCUSSION

All previous studies for the cotton yield model did not give any attention to regression assumptions. So here we have studied the effect of influential observations and outliers on the cotton yield fitted regression model results. Here, we also studied the effect of outliers and influential observations on the regression assumptions. We have seen that outliers and influential observations have strongly affected the model results. We also observed that after deleting all these values, the factors which were insignificant including all these observations actually they were important for the cotton yield model now showing their role in model building. When attention is given to regression assumptions, then water frequency, farmer education, some cottonseed varieties showing their significant role which had not seen without satisfying the regression assumptions. These results coincides with previous literature (Khan et al., 1986; Hassan, 1991; Nabi, 1991; Bakhsh et al., 2005).

The outliers and influential observations were also the cause of multicollinearity, heteroscedasticity, and autocorrelation. It was also observed that after deleting these values, the multicollinearity, heteroscedasticity and autocorrelation were reduced substantially. The model error was reduced dramatically and better results were observed. On the basis of our findings it was found that, types of fertilizer (DAP, Potash and Ganwara), water frequency, farmer education, cottonseed variety and previously sown crops were the significant factors to change the cotton yield. On the basis of these results, it may be suggested that in model fitting in all agriculture related fields, one has to identify and remove the influential observations and outliers. Then refit the model for the better decision, analysis, and interpretation and determine the factors which contributing significant roles in model building.

**REFERENCES**


Ali A, M Tahir, M Ayub, I Ali, A Wasaya and F Khalid, 2009. Studies on the effect of plant spacing on the yield of recently approved varieties of cotton. Pakistan Journal of Life and Social Sciences, 7: 25-30.

Ali MM, 1983. Constraint Research in Pakistan, Some Methodological Issues and Consideration, (Social Science Division), Pakistan. Agricultural Research Council, Islamabad.

Anwar N, 1998. Optimization of quantitative factors contributing to yield variability of wheat. A case study of Rachna Doab. M.Sc. (Honors) Thesis, Department of Agriculture-Economics, University of Agriculture, Faisalabad, Pakistan.

Bakhsh K, I Hassan and A Maqbool, 2005. Factors affecting cotton yield: A case study of Sargodha (Pakistan). Journal of Agriculture and Social Sciences, 4: 332-334.

Chaudhry IS, MB Khan, and M Anwar, 2009. Factors affecting cotton production in Pakistan: empirical evidence from Multan district. Journal of Quality Technology Management, II: 91-100.

Cook RD, 1977. Detection of influential observation in linear regression. Technometrics, 19:15-18.

Gujarati DN, 2004. Basic Econometrics, 4th Edition. The McGraw-Hill.

Hassan I, 1991. Determination of factors inhibiting adoption of improved technology in cotton production. M.Sc. (Agric. Econ.) Thesis, University of Agriculture, Faisalabad, Pakistan.

Khan BR, BM Khan, A Razzaq, M Munir, M Aslam, S Ahmad, NI Hashmi and PR Hobbs, 1986. Effect of different tillage implements on the yield of wheat. Pakistan Journal of Agricultural Research, 7: 141–7.

Muhammad D, MN Afzal, I Raza, and MA Mian, 2009. Growth and development of cotton (Gossypium hirsutum L.) as affected by different methods of pendimethalin application. Pakistan Journal of Weed Sciences and Research, 5: 11-17.

Nabi M, 1991. Relationship between crop productivity and input usage. Journal of International Development, 8: 68–88.

Pardoe I, 2012. Applied Regression Modeling, 2end Edition. John Wiley & Sons, Inc.

Saleem MF, SA Anjum, A Shakeel, MY Ashraf and HZ Khan, 2009. Effect of row spacing on earliness and yield in cotton. Pakistan Journal of Botany, 41: 2179-2188.

Saleem MF, MF Bilal, M Awais, MQ Shahid, and SA Anjum, 2010. Effect of nitrogen on seed cotton yield and fiber qualities of cotton (Gossypium Hirsutum L.) cultivars. Journal of Animal and Plant Sciences, 20: 23-27.

Sui R, DK Fisher, and KN Reddy, 2013. Cotton yield assessment using plant height mapping system. Journal of Agriculture Sciences, 5: 23-31.