Agriculture Association of Textile Chemical and Critical Reviews Journal (2022) 01-05

AATCC Review

26 August 2022: Received
29 October 2022: Revised
04 December 2022: Accepted
20 December 2022: Available Online

http://aatcc.peerjournals.net/

Research Article Open Access

Investigation of Diagnostic checks and finding Outliers in fitting Regression models

R.Vasanthi[1]*, R. Pangayar Selvi[2*], B. Sivasankari[3] and M. Kalpana[4]

[1,2]Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Coimbatore, India
[3]Agricultural College and Research Institute, Tamil Nadu Agricultural University, Madurai, India
[4]Agricultural College and Research Institute, Tamil Nadu Agricultural University, Coimbatore, India

Abstract

The fitting of regression model has problems related to non-linearity, multicollinearity, serial correlation and heteroscedasticity which involves very long and complex procedure of calculations and analysis. This study focuses on an improvement in the model fit based on R2 value. An attempt is made to investigate the outliers in any data set and to increase the R2 square value after the removal of outliers. In this study, a hypothetical data set is considered. The data set indicates consumption as a dependent variable and Income, Food size are considered as independent variables. The regression model for Actual data indicates the R2 value is 0.455. After the removal of outliers using the cook's distance, the revised R2 value is 0.578. This indicates that the outlier in the data set plays a vital role in the model fit. Therefore it is necessary to remove the outlier if any in the data, before proceeding to further analysis.

Keywords:*Diagnostic, Hat matrix, MATLAB, Outlier, Regression, R2value*

Introduction

The main aim of regression modeling and analysis is to develop a good predictive relationship between the dependent (response) and independent (predictor) variables. Regression Diagnostics plays a vital role in finding and validating such a relationship. Once a regression model has been constructed it may be important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include the R-squared, analysis of the pattern of the residuals, and hypothesis testing. Statistical significance can be checked by an F-test of the overall fit, followed by t-tests of individual parameters. Interpretations of these diagnostics tests rest heavily on the model assumptions. Although examination of the residuals can be used to invalidate a model, the results t-test, and F-test are sometimes difficult to interpret if the model assumptions are violated. Regression diagnostics is to identify the influential data. Diagnostics are certain quantities computed from the data with the purpose of pinpointing influential points, after which these outliers can be removed are corrected.

*CorrespondingAuthor: R.Vasanthi and R. Pangayar selvi.
E-mail Address: - vasanthi@tnau.ac.in and pangai@tnau.ac.in
DOI: https://doi.org/10.58321/AATCCReview.2022.10.04.01

Review of Literature

The study on Bauchi Local Government Area determined the costs and returns of rice production among farmers. Primary data were collected with the aid of structured questionnaires which were administered to fifty (50) purposively selected rice farmers. The article showed the input variables Seeds, herbicides, and farm size are significant. With these variables, the R^2 value is 0.931 [1].

The study on Resource use efficiency in Rice production in Kwande Local Government Area of Benue State, Nigeria was examined. From this article the variables Land, and fertilizer are significant with R^2 value is .895 [2].

The resource-use efficiency in sorghum production in the coastal region of Andhra Pradesh was

examined. Data for the study were collected from 100 sorghum producers in seven villages in the study area pertaining to the 2008-09 crop season. In this article, the variables all are significant, except Expenditure on seeds. With these variables, the R^2 value is 0.718 his study examines the resource use efficiency in rice production in Kwande Local Government Area of Benue State Nigeria. The data for the study was collected from 100 rice farmers in the four districts of the study area using a simple random sampling technique [3].

An investigation undertaken in central Gujarat, has estimated the technical efficiency in rice production and has assessed the effect of farm-specific socio-economic factors on this technical efficiency. A stochastic frontier production function has been estimated to determine the technical efficiency of individual farms and variance, as well as regression analyses, have been carried out to find the influence of socioeconomic factors. From this article the variables Operational area, experience in rice cultivation, education level of the farmer, Number of working family members, and Distance of field from canal irrigation structure are significant. With these variables, R^2 value is 0.3174 [4].

According to the above articles, it is seen that the production of rice shows high beta coefficients for highly related variables and given higher R^2 value, whereas the technical efficiency study includes the variables which lead to a less significant R^2 value. Hence it is clear that the value of R^2 depends upon the independent variable selection. Therefore there is a Diagnostics check for justification.

Materials & Methods

Diagnostics:

(i) Hat Matrix:

The Hat matrix comes from the formula for the regression Y.

$$\hat{Y} = X\beta$$

$=X(X'X')^{-1}X'Y$

=HY

Where H = $X(X'X')^{-1}X'$ is the Hat matrix

$\therefore \hat{Y}$ = HY

The Hat matrix transforms Y into the predicted scores. The diagonals of the Hat matrix indicate which values will be Outliers or not.

(ii) Outliers:

An outlier is an observation that is substantially different from all other ones and can make a large difference in the results of regression analysis.(Fox.AJ., 1972) Outliers play an important role in regression. An outlier is a data point that diverges from an overall pattern in a sample. It has a large residual (the distance between the predicted value and the observed value (y). In linear regression, an outlier is an observation with a large residual. In other words, it is an observation whose dependent-variable value is unusual given its value on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem [5]. Outliers lower the significance of the fit of a statistical model because they do not coincide with the model's prediction.

(iii) Cook's distance (Di):

Evaluating large or unusual observations in regression models are the purpose of Cook's Distance. It is a summary measure of the influence of a single case (observation) based on the total changes in all other residuals when the case is deleted from the estimation process.

Cook's Distance can be calculated using the following formula,

$$D_i = \frac{e_i^2}{P\,MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

Where,

h_{ii} is the i-th diagonal element of the hat matrix;
e_i is the crude residual (i.e., the difference between the observed value and the value fitted by the proposed model);

MSE is the mean square error of the regression model;
P is the number of fitted parameters in the model

To identify potential outliers, one Rule of Thumb is to treat point i as an outlier when:

Where

$$D_i \geq \frac{4}{n-(k+1)}$$

n is the number of observations
k is the number of parameters

Results & Discussion

Diagnostic analysis was performed for the hypothetical data set (Annexure I) in which consumption is considered as a dependent variable and Income, Food size are considered as independent variables. The leverage points, the Cooks distance and upper limit for Cooks were worked out using MATLAB program and the regression analyses was done with and without Outliers for the dataset and the results are tabulated below

The upper limit for Cooks distance was worked out as 0.0851 using formula given above. According to the Rule of Thumb observations for which the Cooks distance exceeded the upper limit are classified as outliers. From the Table.1. in our data set 4 observations namely 12, 42, 44 and 49 were outliers their Cooks distance are respectively 0.1882, 0.0999, 0.1018 and 0.1974 which are all greater than the upper limit.

The original data had 50 observations with four outliers. Table 2 shows that the R^2 value of 50 observations was 0.455. Table .3 indicate that after removing the outliers there were 46 observations and the revised value of R^2 is 0.579. The adjusted value R^2 also increased from 0.432 to 0.559. This shows that the removal of outliers improves the explanatory power of the model and also improves the precision of the regression coefficients and R-Square value [7]

Table 1: Cook's distance

S.No	Cooks Distance	S.No	Cooks Distance	S.No.	Cooks Distance	S.No	Cooks Distance	S.No	Cooks Distance
1.	0.0037	11.	0.0068	21.	0.0028	31.	0.0054	41.	0.0113
2.	0.0037	**12.**	**0.1882**	22.	0.0019	32.	0.0093	**42.**	**0.0999**
3.	0.0037	13.	0.0769	23.	0.0087	33.	0.0059	43.	0.0029
4.	0.0037	14.	0.0093	24.	0.0275	34.	0.0042	**44.**	**0.1018**
5.	0.0037	15.	0.0000	25.	0.0053	35.	0.0070	45.	0.0258
6.	0.0037	16.	0.0007	26.	0.0011	36.	0.0028	46.	0.0038
7.	0.0037	17.	0.0011	27.	0.0002	37.	0.0088	47.	0.0187
8.	0.0037	18.	0.0182	28.	0.0093	38.	0.0076	48.	0.0167
9.	0.0037	19.	0.0123	29.	0.0000	39.	0.0453	**49.**	**0.1974**
10.	0.0037	20.	0.0202	30.	0.0010	40	0.0128	50.	0.0026

Summary and conclusion

Diagnostics checks are very important for regression analysis. To arrive at a suitable Multiple Linear Regression model for the data set, the researcher has to carry out the diagnostic checks and out layers, if any have to be removed from the data set, and the updated data set has been used for further analysis. In this regard cook's distance is a very useful statistic to identify outliers. In this study, the diagnostic checks have been illustrated with the data set. This study shows the importance of diagnostic checks in fitting regression models.

Future Scope of the Study

This residual analysis and note outlying cases can lead to valuable insights for strengthening the model to adopt this model. This outlier fixation gives insight to modify or fit the correct model for analysis of data for strengthening the model.. Finally, the outliers can be detected and play any significant influence on the parameter estimate.

Conflict of Interest

The authors declares that there is no Conflict of Interest. The authors had full access to all set of data, with an explanation of the nature and extend of access to all of the data in this study and authors take complete responsibility for the integrity of the data and accuracy of the data analysis.

Acknowledgement

I acknowledge Department of Physical Science and Information and Technilogy, Agricultural Engineering College and Research Institute, Tamil Nadu Agricultural University, Coimbatore for providing the necessary facility to carryout the work.

Table 2- Regression Analysis with Outliers

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.67467256				
R Square	**0.455183063**				
Adjusted R Square	0.431999363				
Standard Error	504.7628577				
Observations	50				
ANOVA					
	Df	*SS*	*MS*	*F*	*Significance F*
Regression	2	10004794	5002397	19.63375	6.34E-07
Residual	47	11974920	254785.5		
Total	49	21979714			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	331.3086797	254.3287	1.302679	0.199031	-180.335
Income	0.056091264	0.011326	4.952342	9.88E-06	0.033306
Fsize	129.5656701	36.13522	3.585578	0.000798	56.87098

Table 3- Regression Analysis without Outliers

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.760803				
R Square	**0.578822**				
Adjusted R Square	0.559232				
Standard Error	414.1317				
Observations	46				
ANOVA					
	Df	***SS***	***MS***	***F***	***Significance F***
Regression	2	10135011	5067505	29.54726	8.43E-09
Residual	43	7374719	171505.1		
Total	45	17509730			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	273.24	211.2963	1.29316	0.20286	-152.88
Income	0.063436	0.010091	6.286159	1.4E-07	0.043085
Fsize	107.0002	32.58815	3.283408	0.002043	41.27993

References

[1.] Sani RM , Malumfashi AI, Daneji MI, Alao OO (2007) Economics of rice production:a case study of Bauchi local government area, Bauchi state, Nigeria. Continental J. Agricultural Economics 1: 7 – 13.

[2.] David Terfa Akighir, Terwase Shabu (2011) Efficiency of Resource use in Rice Farming Enterprise in Kwande Local Government Area of Benue State, Nigeria. International Journal of Humanities and Social Science. 1(3).

[3.] Chapke RR, Biswajit Mondal, Mishra JS (2011)

Resource-use Efficiency of Sorghum (Sorghum bicolor) Production in Rice (Oryza sativa)-fallows in Andhra Pradesh, India. J Hum Ecol 34(2): 87-90

[4.] Anuradha Narala, Zala YC (2010) Technical Efficiency of Rice Farms under Irrigated Conditions in Central Gujarat. Agricultural Economics Research Review 23:375-381.

[5.] Fox AJ (1972) Outliers in time series. Journal of the Royal Statistical Society, Series B 34:350–363.

[6.] Kalman RE, et al. (1960) A new approach to linear filtering and prediction problems. Journal of basic Engineering 82 (1):35–45.

[7.] Stephen Raj S, Senthamarai Kannan K (2017) Detection of Outliers in Regression Model for Medical Data. International Journal of Medical Research & Health Sciences 6(7): 50-56