Research Paper

Volume : 3 | Issue : 2 | February 2014 • ISSN No 2277 - 8179

Outlier Detection in Simple Linear Regression Models and Robust Regression – A Case Study on Wheat Production Data

STATISTICS

KEYWORDS : Cook's Distance, DFFITS, DFBETAS, COVRATIO, Robust Regression

B. Vinoth — Research Scholar, Department of Statistics.

A. Rajarathinam — Professor, Department of Statistics Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.

ABSTRACT

In this study, the presence of outliers in wheat production data based on residuals obtained from the fitted simple linear regression model have been studied. Outliers were detected based on the following: i) residual analyses using standardized residuals, studentised residuals, jackknife residuals and predicted residuals; ii) residuals plots such as the graph of predicted residuals, the Williams graph, and the Rankit Q-Q plot; and iii) scalar measures of influence statistics such as cook's Di (measures the change in the estimates that results from deleting each observation), DFFITSi (measures the change in the predicted value of the dependent variable when the current case is omitted from the calculations), DFBETASj(i) (measures the influence on regression coefficients), Atkinson measures, and the Covariance ratio (measure of model performance). Robust regression was employed to avoid the influence of outliers, and the linear growth rate was calculated based on the fitted robust regression model.

1. INTRODUCTION

In agricultural economics, the linear regression model is an important statistical tool used to study trends and calculate the growth rates of various agricultural productions. A number of researchers (Panse, 1964; Dey, 1975; Reddy, 1978; Narain et al., 1982; Kumar and Rosegrant, 1994; Kumar, 1997; Vaishnav and Dixit, 1998; Joshi and Saxena, 2002; Singh and Srivastava, 2003; Shah et al., 2005; Patil et al., 2009; Parmar, 2010) have used linear or exponential models to estimate growth rates that are currently being used by country planners and policy makers.

The method of ordinary least squares (OLS) has been used to fit the model and estimate the model parameters. There are several assumptions that must be fulfilled for the OLS model to be valid. When the regression model does not meet the fundamental assumptions, predictions and estimations based on the model may be biased. It is usually not possible to detect departures from the underlying assumptions of regression analysis by examination of standard summary statistics such as the t or F statistics or R^2. These are global model properties, and as such, they do not ensure model adequacy.

When fitting the regression model, one can find outliers (observations with large residuals), leverage points (observations with extreme values on a predictor variable) or influential points (observations that substantially change the estimate of the regression coefficients when removed; influence can be thought of as the product of leverage and outliers). Outliers will inflate the error variance as well as the standard errors, causing the confidence interval to become stretched. Furthermore, the estimation cannot become asymptotically consistent. Leverage points can have a dramatic effect on the model summary statistics, R^2 and $\hat{\sigma}^2$, etc. Influential points will have a noticeable impact on the model parameters; they "pull" the regression model in their direction.

Appropriate measures must be taken when the production data are contaminated with unusual data points. If the unusual data points are neither data entry errors nor observations from a differing population, then we have no compelling reason to exclude them from the analysis. In this case, robust regression might be a good strategy because it offers a compromise between excluding these points from the analysis entirely and including all the data points and treating them equally in OLS regression. The idea of robust regression is to weigh the observations differently based on how well behaved they are.

In this study, we investigate trends and calculate the growth rate in wheat production based on a Simple Linear Regression (SLR) model. Furthermore, we investigate the presence of outliers, leverage and influential data points based on different residual plots and scalar measures of influence statistics such as cook's D_i (measures the change in the estimates that results from deleting each observation), $DFFITS_i$ (measures the change in the predicted value of the dependent variable when the current case is omitted from the calculations), $DFBETAS_{j(i)}$ (measures the influence on regression coefficients), Atkinson measures, and the Covariance ratio (measures of model performance). Robust regression techniques are employed to minimise the impact of outliers in the estimation processes, and growth is calculated based on the fitted robust regression model.

2. MATERIALS AND METHODS

This study, based on wheat production time-series data pertaining to the period from 1950-51 to 2009-2010, was retrieved from the Gujarat government's official web site: www.agri.gujarat.gov.in.

2.1 Diagnostics based on residual analysis

The residuals (defined as the difference between the observed values and those predicted by the regression model) are often examined because they not only take the response variable into account but also aid in the investigation of regression assumptions. The residuals in their raw form are often not used because they do not approximate the unobserved errors well; the diagonal values of the hat matrix need to be equal and the off-diagonal values need to be sufficiently small to ensure that the residuals are uncorrelated and have constant variance (Cook and Weisberg, 1982). Chatterjee and Hadi (1988) mention that the ordinary residuals are not appropriate for diagnostic purposes; a transformed version of them is preferable. Thus, transformations of residuals such as standardised residuals, studentised residuals, jackknife residuals and predicted residuals, are often preferred over raw residuals because they overcome some of the limitations of raw residuals.

2.1.1 Standardised residuals (Normalised)

A logical scaling for the residuals is the standardised residuals

$d_i = \frac{e_i}{\sqrt{MS_{RES}}}$, i=1,2,3, …, n where MS_{RES} is the mean square residual. The standardised residuals have mean zero and approximately unit variance. Consequently, large standardised residuals ($d_i > 3$) potentially indicate an outlier.

2.1.2 Internally studentised residuals: Daniel and Wood (1971) introduced a type of residual defined as $r_i = \frac{e_i}{\sqrt{MS_{RES}(1-h_{ii})}}$ where h_{ii} is the leverage for observation i. The studentised residuals are the preferred residuals because they are easily related to the t-distribution and have been standardised to equal variance (Welsch and Kuh, 1977). Belsley et al. (1980) recommend a cutoff value of |2| for r_i.

2.1.3 Jackknife residuals: The studentised residual r_i discussed above is often considered an outlier diagnostic. It is cus-

tomary to use MS_{RES} as an estimate of σ^2 in computing r_i. This is referred to as internal scaling of the residual because MS_{RES} is an internally generated estimate of σ^2 obtained from fitting the model with all n observations. Another approach would be to use an estimate of σ^2 based on a data set with the ith observation removed. Denote the estimate of σ^2 as obtained by .

$$S^2_{(i)} = \frac{(n-p)MS_{RES} - \left(e_i^2/(1-h_{ii})\right)}{(n-p-1)}.$$

The estimate of σ^2 in the above equation is used instead of MS_{RES} to produce jackknife residuals (externally studentised residuals), usually called R-student and given by

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}} \quad i = 1,2,3, \ldots, n$$

In many situations, t_i will differ little from the studentised residual r_i. However, if the ith observation is influential, then $S^2_{(i)}$ can differ significantly from MS_{RES}, and thus the R- student statistic will be more sensitive to this point.

2.1.4 Predicted residuals: The predicted residuals are the ordinary residuals weighted according to the diagonal elements of the hat matrix; they are calculated by

$e_{P,i} = \frac{e_i}{(1-h_i)}$, i=1,2,3, ..., n where e_i is the ith raw residual.

2.2 Diagnostics based on residuals plots

Meloun and Militky (2001) used the following types of plots using residuals to detect influential observations.

2.2.1 The graph of predicted residuals

The graph of predicted residuals (Williams, 1973) has the predicted residuals $\hat{e}_{P,i}$ on the x-axis and the ordinary residuals $\hat{e}_i$ on the y-axis. The high-leverage points are easily detected by their location, as they lie far outside the line $y = x$. The outliers are located on the line $y = x$, but far from its central pattern.

2.2.2 The Williams graph

The Williams graph (Williams, 1973) has the diagonal elements H_{ii} on the *x*-axis and the jackknife residuals $\hat{e}_{J,i}$ on the *y*-axis. Two boundary lines are drawn: the first for outliers, $y = t_{0.95}(n-p-1)$ and the second for high-leverages, $x = 2p/n$. Note that $t_{0.95}(n-p-1)$ is the 95% quantile of the Student distribution with (n-p-1) degrees of freedom.

2.2.3 The Rankit Q-Q plot

The rankit (Q-Q plot) has the quantile of the standardised normal distribution $u\,p_i$ for $P_i = i/(n+1)$ on the x-axis and the ordered residuals $\hat{e}_{S,i}, \hat{e}_{P,i}, \hat{e}_{J,i}$ (increasingly ordered values of various types of residuals) on the y-axis.

2.3 Diagnostics based on scalar measures of influence

The general procedure for assessing the influence of a point in a regression analysis is to determine the changes that occur when that observation is deleted. Several measures of influence have been developed using concepts cited in Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985), and Chatterjee and Hadi (1988). Several influence measures are discussed below, each of which measures the effect of deleting the ith observation. These measures are important to the study because they communicate the influence of a specified observation on all parameters.

2.3.1 Cook's D measure

Cook (1977) has suggested an influence measure, Cook's D that uses the squared distance between the least-squares estimate based on all n points $\hat{\beta}$ and the estimate obtained by deleting the ith point $\hat{\beta}_{(i)}$ or, equivalently, the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the ith observation. This distance measure can be expressed in general form by

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})X'X(\hat{\beta}_{(i)} - \hat{\beta})}{p\,MS_{RES}} = \frac{(\hat{y}_{(i)} - \hat{y})^T(\hat{y}_{(i)} - \hat{y})}{p\,MS_{RES}} = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$$

Thus Cook's distance is a multiplicative function of two basic quantities. The first is the square of the standardised residual (r_i^2) and the potential function $h_i\,(1-h_i)$, where h_i is the ith diagonal element of the hat matrix, and p is the parameter. If a point is influential, its deletion causes large changes, and the value of D_i will be large. There are different options regarding what cut-off values to use for spotting highly influential points. A simple operational guidance of $D_i > 1$ has been suggested (Kenneth and Robert, 1990), (Cook and Weisberg, 1982). Others have suggested $D_i > 4/n$, where n is the number of observations. Plots of D_i against i are often useful as influential points will stand out. If all values of D_i are similar, then there is likely no influential point, and so no further action needs to be taken.

2.3.2 Welsch and Kuh Measure

Differences in FITS (DFFITS), proposed by Welsch and Kuh (1977), is a measure similar to Cook's distance. It is used to investigate the influence of the ith observation on the predicted value. It is defined as

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_i}}, \qquad i = 1,2,3,\ldots,n$$

Thus, DFITS is the scaled difference between the ith fitted value obtained from the full data and the ith fitted value obtained by deleting the ith observation. The difference is scaled by $\sqrt{MS_{Res}h_{ii}}$. It can be written as

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} r_i, \quad i = 1,2,3,\ldots,n, \quad \text{where } r_i$$

is the standardised residual. Thus, $DFFITS_i$ is the R-student value multiplied by the leverage of the ith observation $\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$. Belsley, Kuh, and Welsch (1990) have suggested that any observation for which $|DFFITS_i| > 2\sqrt{p/n}$ warrants attention.

2.3.3 DFBETAS (Difference in beta coefficients)

Cook's D_i reveals the impact of the ith observation on the entire vector of the estimated regression coefficients. The influential observations for the individual regression coefficients are identified by $DFBETAS_{j(i)}$, j=0,1,2,…, p, where each $DFBETAS_{j(i)}$ is the standardised change in $\hat{\beta}_j$ when the ith observation is deleted from the analysis. Thus, $DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{s_{i}\sqrt{c_{jj}}}$, where c_{jj} is the (j+1)[st] diagonal element from $(X^TX)^{-1}$. The cutoff point of $2/\sqrt{n}$ has been suggested by Belsley, Kuh, and Welsch as the point that will tend to highlight the same proportion of influential points across data sets.

Volume : 3 | Issue : 2 | February 2014 • ISSN No 2277 - 8179

2.3.4 Atkinson Measure

This measure is the modified version of Cook's measure D_i suggested by Atkinson (1985) that is even more closely related to Belsley's $DFFITS_i$ and has the form

$$A_i = \left|\hat{e}_{j,i}\right| \times \sqrt{\frac{n-p}{p} \times \frac{h_i}{1-h_i}}$$

where $\hat{e}_{j,i}$ is the jackknife residual. The Atkinson measure A_i enhances the sensitivity of distance measure to high-leverage points. This measure is also convenient for graphical interpretation; Atkinson recommends that signed values of A_i be plotted in any of the customary methods for residuals.

2.3.5 A Measure of Model Performance (COVRATIO)

Belsley et al. (1980) have proposed the following numerical statistic to identify outliers in linear regression models based on the determinant ratio of observations on the covariance matrix of the parameter estimates given by

$$COVRATIO_i = \frac{\left|(X_{(i)}^T X_{(i)})^{-1} S_{(i)}^2\right|}{\left|(X^T X)^{-1} MS_{RES}\right|}, \quad i = 1, 2, \ldots, n,$$

, where $X_{(i)}$ is the X matrix without the ith observation. Clearly, if $COVRATIO_i > 1$, the ith observation improves the precision of estimation, whereas if $COVRATIO_i < 1$, inclusion of the ith point degrades the precision. Computationally,

$$COVRATIO_i = \frac{(S_{(i)}^2)^p}{MS_{RES}^p}\left(\frac{1}{1-h_{ii}}\right), \quad \text{where } (1/1-h_{ii}) \text{ is}$$

the ratio of $\left|(X_{(i)}^T X_{(i)})^{-1}\right|$ to $\left|(X^T X)^{-1}\right|$, so that a high-leverage point will make $COVRATIO_i$ large. If the ith observation is an outlier, $S_{(i)}^2 / MS_{RES}$ will be much less than unity. If $COVRATIO_i > 1 + 3p/n$ or if $COVRATIO_i < 1 - 3p/n$, then the ith point should be considered influential.

3. RESULTS AND DISCUSSIONS

To study the trends in wheat crop production, a Simple Linear Regression (SLR) model is employed. The fitted model is further analysed for the presence of influential points as well as outliers using residuals produced by the fitted model. The findings are discussed below.

3.1 Simple Linear Regression Model

A SLR model is fitted to the wheat production time-series data pertaining to the period from 1950-51 to 2009-10 (n=60) using the following SAS procedure.

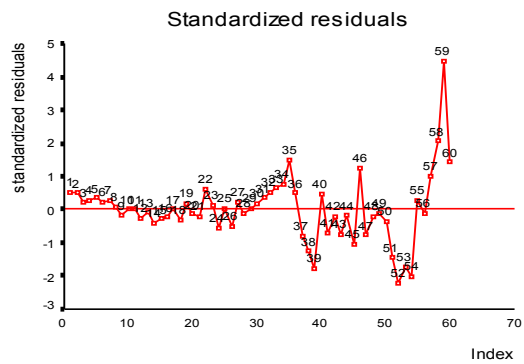
```
PROC REG DATA=DELIVERY;
MODEL Y=X/INFLUENCE;
OUTPUT OUT = SASdata1
PREDICTED=pre1
RESIDUAL = res1
L95M =l95m1
U95M = u95m1
L95=L951
U95=u951
STDP=stdp1
STDR=stdr1
STUDENT=student1
COOKD=cookd1
PRESS=press1
RSTUDENT=rstudent1
DFFITS=defits1
COVRATIO=covratio1;
RUN;
```

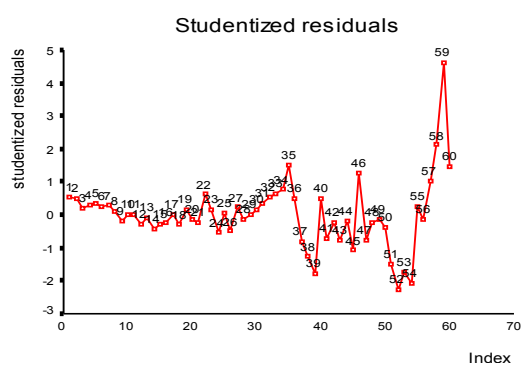
The following SLR model

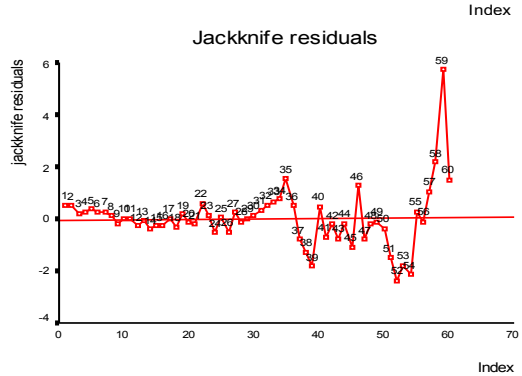
$Y = -215.09 + 335.73^{**} X$ with $R^2 = 66^{**}$ % (where Y is the wheat production and X is the time (year)) gives the trends in wheat production with Root MSE = 4208.54, Coefficient of Variation=42 % and the Durbin-Watson D statistic = 0.863, which is non-significant, indicating that the residuals from the fitted model are correlated. Other types of residuals from the fitted model include standardised residuals, studentised residuals, jackknife residuals and predicted residuals; and different regression diagnostic statistics such as Cook's D, DFFITS, DFBETA and COVRATIO have been calculated and given in Appendix-I and Appendix-II, respectively.

3.2 Diagnostics based on residuals

The standardised residuals d_{52} = 2.2295, d_{54}= 2.0369, d_{58}=2.0572 and d_{59}=4.4633; internally studentised residuals r_{52} = 2.2782, r_{54}= 2.0869, r_{58}=2.1204 and r_{59}=4.6080; and jackknife residuals (R-student) t_{52} = 2.3669, t_{54}= 2.1511, t_{58}=2.1886 and t_{59}=5.7375 exceeded the cutoff value of ± 2; therefore, the observations at the data points 52, 54, 58 and 59 are considered outliers. The predicted residuals are also very high at these points.







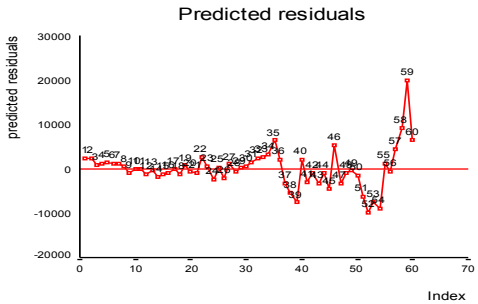


Fig.1: Diagnostics based on standardized, studentized, Jackknife and predicted residuals plots

3.3 Diagnostics based on residuals plots

Using the residuals produced by the fitted SLR model, the graph of predicted residuals, Williams graph and Rankit Q-Q plot have been drawn to detect outliers and other influential points. The findings are discussed in sequence as follows.

3.3.1 The graph of predicted Residuals

The following graph of predicted residuals has the predicted residuals $\hat{e}_{P,i}$ on the *x*-axis and the ordinary residuals $\hat{e}_i$ on the *y*-axis. The observation at the 59th data point is far from its central pattern; hence, the observation at this point may be considered an outlier.

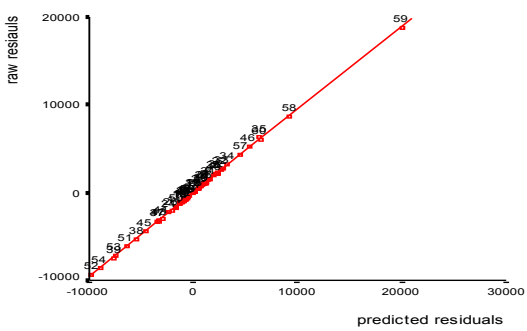


Fig.2: Predicted residuals plot

3.3.2 The Williams graph

The Williams graph is drawn by plotting the diagonal elements H_{ii} on the *x*-axis and the jackknife residuals $\hat{e}_{J,i}$ on the *y*-axis. Two boundary lines are drawn, the first for outliers,

$y = t_{0.9}\ (n-p-1)$ 1.671 and the second for high-leverages, $x = 2p/n$ = 4/60= 0.0667. The points 35, 58, 59 and 60 were above the boundary lines, indicating that the observations at these points are outliers.

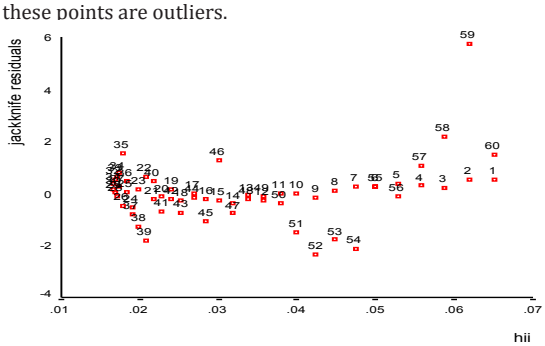


Fig.3: Williams graph

3.3.3 The Rankit Q-Q plot

The rankit (Q-Q plot) has the quantile of the standardised normal distribution $u\,p_i$ for $P_i = i/(n+1)$ on the *x*-axis and the ordered residuals $\hat{e}_{S,i}, \hat{e}_{P,i}, \hat{e}_{J,i}$ (increasingly ordered values of various types of residuals) on the *y*-axis. The observation at the 59th data point is far from the rest of the observation; therefore, the observation at the 59th data point may be considered an outlier.

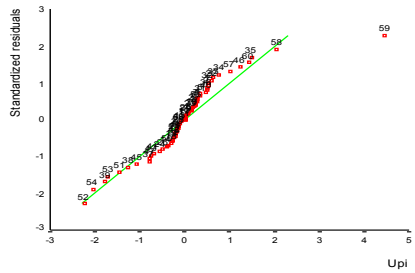


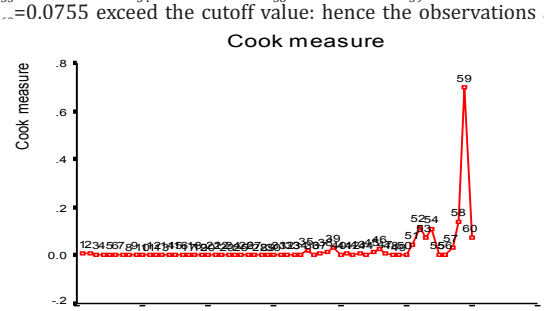
Fig.4: Q-Q plot for Standardized residuals

3.4 Diagnostics based on scalar measures of influence

Based on the fitted SLR model, various measures of influence statistics such as Cook's D_i (measures the change to the estimates that results from deleting each observation), $DFFITS_i$ (measures the change in the predicted value of the dependent variable if the current case is omitted from the calculations), $DFBETAS_{j(i)}$ (measures the influence on regression coefficients), Atkinson measures, and COVRATIO (measures of model performance) have been calculated using SAS PROC REG procedures and are presented in Appendix – II. The findings are discussed in sequence as follows.

3.4.1 Cook distance

The cutoff value for cook distance measure is calculated as $D_i > 4/n$ = 0.0667. It is observed that D_{52}=0.1148, D_{53}=0.0720, D_{54}=0.1082, D_{58}=0.1402, D_{59}=0.6994 and D_{60}=0.0755 exceed the cutoff value: hence the observations at



3.4.2 DFFITS

It is observed that $|DFFITS_{52}|$ = 0.4978, $|DFFITS_{53}|$ =0.3866, $|DFFITS_{54}|$= 0.4796, $|DFFITS_{58}|$=0.5465, $|DFFITS_{59}|$=1.4726, and $|DFFITS_{60}|$=0.3927 exceeded the cutoff value $2\sqrt{(p/n)} = 2\sqrt{(2/60)}$ =0.3652; therefore, the observations at points 52, 53, 54, 58, 59 and 60 may be considered outliers.

3.4.3 DFBETAS

Recalling that the cutoff value of DFBETA= $2/\sqrt{n}$ = $2/\sqrt{60}$ =0.2582, it is observed that $|DFBETAS_{0,58}|$ = 0.2583, $|DFBETAS_{0,59}|$=0.7168, $|DFBETAS_{1,52}|$=0.3877, $|DFBETAS_{1,53}|$=0.3064, $|DFBETAS_{1,58}|$=0.4624, $|DFBETAS_{1,59}|$= 1.2585 and $|DFBETAS_{1,60}|$ = 0.3387 exceeded the cutoff value, which implies that the observations at the data points 58 and 59 are affecting the intercept and the observations at the data points 52, 53, 58, 59 and 60 are affecting the slope. The observations at these data points are considered to have an influence on the parameter estimates.

Volume : 3 | Issue : 2 | February 2014 • ISSN No 2277 - 8179

3.4.4 Atkinson Measure
The index plot of Atkinson measure clearly indicates that the observation at the 59th data point is an outlier that lies far away from the rest of the observations. Furthermore, the ordinary residual e_{59} =18784.17, the standardised residual d_{59}=4.4633, the studentised residual r_{59}=4.6080, the jackknife residual t_{59}=5.7375 and the predicted residuals p_{59}=20021.58 are also above the cutoff values, clearly indicating that the observation at the 59th data point is an outlier.

3.4.5 A Measure of Model Performance
The $COVRATIO_{59}$ exceeds the cutoff 1±3p/n = 1± 3(2)/60 = (0.9, 1.1), indicating that the observation at the 59th data point is an influential point and degrades the precision of the estimation, while the rest of the observations are within the limits of the Covariance ratio.

3.5 Robust regression
Robust regression is applied to the wheat production data using the following SAS procedure PROC ROBUSTREG.

```
PROC ROBUSTREG DATA=DELIVERY;
MODEL Y=X /DIAGNOSTICS LEVERAGE;
OUTPUT OUT = ROBOUT R=RESID
SR=STDRES;

TITLE "Example Multiple linear regression using PROC REG";
RUN;
```

By default, the procedure does M estimation with the bisquare weight function, and it uses the median for estimating the scale parameter. The MODEL statement specifies the covariate effects. The DIAGNOSTICS option requests a table for outlier diagnostics, and the LEVERAGE option adds leverage-point diagnostic results.

Table-1 displays the model fitting information and summary statistics for the response and continuous covariates. The lower quantile (Q_1), median, upper quantile (Q_3), mean and standard deviation are calculated to be 3967.5, 9107.5, 13674.5, 10024.6 and 7196.6, respectively. The robust estimate of the univariate scale, which is computed as the standardised median absolute deviation (MAD) is 7033.5. Because the difference between the standard deviation and the MAD is very low, the presence of extreme values is ruled out in the wheat production data.

Table-1: The ROBUSTREG Procedure model fitting information and Summary Statistics

Method			M Estimation		
Summary Statistics					
Q1	Median	Q3	Mean	S.D	MAD
3967.5	9107.5	13674.5	10024.6	7196.5	7033.5

Table-2 displays the robust parameter estimates, standard errors, and confidence limits. The point estimate of the scale parameter in the linear regression model, which is obtained by the median method, is 2041.336. All parameter estimates lie in the 95% confidence interval. For the wheat production data, M estimation yields the linear model

$$\hat{Y} = 709.34 + 289.06^{**} X \text{ with } R^2 = 55^{**}\%$$

Table-2: The ROBUSTREG Procedure Parameter Estimates

Parameter	DF	Estimate	S.E	95% C.L		χ^2	Pr > χ^2
Intercept	1	709.3	724.5	-710.8	2129.5	.96	0.32
X	1	289.1	20.6	248.5	329.55	95.7	<.0001
Scale	1	2041.336					

The standardised robust residuals, calculated using the ROBUSTREG procedure, that exceed the corresponding cutoff value are given in the Table-5. Observations at data points 35, 39, 46, 52, 54, 58, 59 and 60 are outliers because their standardised robust residuals exceed the cutoff value in absolute value.

Table-3: The ROBUSTREG Procedure Outlier diagnostic statistics

Observation	Standardised Robust Residual	Outlier
35	3.4161	*
39	-3.2100	*
46	3.1547	*
52	-3.8605	*
54	-3.4177	*
58	5.1143	*
59	10.0978	*
60	3.8568	*

The outliers detected by the method of residuals analyses and diagnostics based on scalar measures of influence are similar to those detected by ROBUSTREG procedures.

The following output gives the robust version of goodness-of-fit statistics for the model. The Robust Simple Linear Regression model explains 55 % of variation in wheat production data, which is found to be highly significant.

Goodness-of-Fit

Statistic	Value
R-Square	0.5471
AICR	106.1827
BICR	110.5805
Deviance	4.2667E8

The linear growth rate of wheat production, based on the Robust Simple Linear Regression model, is (289.06/10024.6)*100=2.88, which is found to be highly significant.

4. CONCLUSION
Outliers deserve special attention as they often provide important clues about the model and the process under study. Outliers are commonly present in agricultural production data due to climatic factors or other uncontrolled situations. In this case robust regression is a suitable tool in comparison to simple linear regression models.

ACKNOWLEDGEMENT
The financial assistance received by the first author in the form of Junior Research Fellowship under the Major Research Project scheme of University Grant Commision (UGC), India is highly acknowledged.

REFERENCE

1. Atkinson, A.C., (1985). Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. Claredon Press, Oxford. | 2. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York: John Wiley & Sons. | 3. Chatterjee, S., and Hadi, A.S. (1988). Sensitivity Analysis in Linear Regression. Wiley, New York. | 4. Cook, R.D., Weisberg, S. (1982). Residuals and Influence in Regression. New York: Chapman and Hall, London. | 5. Daniel, C., and Wood, F.S. (1971). Fitting Equations to Data. New York, Wiley. | 6. Dey, A.K. (1975). Rates of growth of agriculture and industry. Econ.Political Weekly, 10:A26-A30. | 7. Joshi, P.K. and Saxena, R. (2002). A profile of pulses production in India. Facts, trends and opportunities. Ind.J.Agric.Econ., 57 : 326-339. | 8. Kenneth, A., and Robert, W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases in Fox. John; and Long, J. Scott (eds.); Modern Methods of Data Analysis, 257-91. Newbury Park, CA: Sage. | 9. Kumar, P. and Rosegrant, M.W. (1994). Productivity and sources of growth for rice in India. Eco. Political Weekly, 29: 183-188. | 10. Kumar, P., (1997). Food security: Supply and demand perspective. Indian Farming. 12: 4-9. | 11. Meloun, M., and Militky, J. (2001). Detection of single influential points in OLS regression model building. Analytica Chimica Acta 439, 161-191. | 12. Narain, P., Pandey, R.K., and Sarup, S. (1982). Perspective plan for foodgrains. Commerce, 145: 184-191. | 13. Parmar, R.S. (2010). Statistical modeling on area, production and productivity of major crops of Middle Gujarat – A case study. Ph.D. thesis submitted to Anand Agricultural University, Anand, Gujarat, India. | 14. Panse, V.G., (1964). Yield trends of rice and wheat in first two five-year plans in India. J.Ind.Soc. Agric.Statistics, 16 : 1-50. | 15. Patil, B.N., Bhonde, S.R. and Khandikar, D.N. (2009). Trends in area, production and productivity of groundnut in Maharashtra. Financing Agric. pp.35-39. | 16. Reddy, V.N. (1978). Growth Rates. Econ.Political Weekly, 13: 806-812. | 17. Shah, A.N., H.Shah and N.Akmal, (2005). Sunflower area and production variability in Pakistan: Opportunities and constraints.HELIA. 28: 165-178. | 18. Singh,A. and Strivastava, R.S.L. (2003). Growth and instability of sugarcane production in Uttar Pradesh: A regional study. Indian J.Agric. Econ., 58: 279-282. | 19. Vaishnav, P.R. and S.K.Dixit (1998). Trends of groundnut productivity in long-term experiment. GAU Res.J., 24: 73-82. | 20. Welsch, R. E. and Kuh, E. (1977). Linear Regression Diagnostics. Technical Report 923-77, Sloan School of Management, Massachusetts Institute of Technology. | 21. Williams, D.X., (1973). Letter to the Editor, Appl. Statist. 22, 407-408. |