DEALING WITH OUTLIERS AND INFLUENTIAL POINTS WHILE FITTING REGRESSION

Chuda Prasad Dhakal

Journal of Institute of Science and Technology

Volume 22, Issue 1, July 2017

ISSN: 2469-9062 (print), 2467-9240 (e)

Editors:

Prof. Dr. Kumar Sapkota

Prof. Dr. Armila Rajbhandari

Assoc. Prof. Dr. Gopi Chandra Kaphle

JIST, **22** (1): 61-65 (2017)

Published by:

Institute of Science and Technology

Tribhuvan University

Kirtipur, Kathmandu, Nepal

JIST 2017, 22 (1): 61-65
ISSN: 2469-9062 (p), 2467-9240 (e)

Research Article

DEALING WITH OUTLIERS AND INFLUENTIAL POINTS WHILE FITTING REGRESSION

Chuda Prasad Dhakal
Tribhuwan University, Institute of Agriculture and Animal Sciences (IAAS), Rampur, Chitwan, Nepal
Corresponding E-mail: chuda.studies@gmail.com

ABSTRACT

Dealing with outliers and influential points while fitting regression is recognizing them, identifying the reasons to their existence in the process and employing the best alternatives to lessen their effect to the fitted regression model. In this paper, before considering elimination of outliers and the influential points while fitting a regression, as they contain important information, issues why unusual observations (possible outliers) appear in the process and how to analyze them to detect if they were real outliers, have been discussed thoroughly. And, when detected as outliers and influential points, to investigate and eliminate their effect in the fitted model, analytic procedures; leverage value, studentized residuals and cook's distance were carefully employed to optimize a multiple regression model for rice production forecasting in Nepal. This model was fitted with 35 years (1961-1995) time series data, collected from Ministry of Agriculture and Cooperatives, Food and Agriculture Organization Statistics Database, International Rice Research Institute and Department of Hydrology and Metrology which to its end was consisted of the three predictors, price at harvest, rural population and area harvested.

Keywords: Outliers, Influential points, Studentized residual, Leverage value, Cook's distance.

INTRODUCTION

In statistics, outlier is an observation that is numerically distant from the rest of the sample in which it occurs. Rahaman *et al.* (2012), Stevens (1983) and Sweet & Martin (2012) have compatibly defined outliers to be the points in a data set which are very different from the other points. Also, Jarrell (1994), Rasmussen (1988) and Stevens (1984), (as cited in Osborne & Overbay, 2004) have mentioned that an outlier is generally considered to be a data point that is far outside the norm for a variable. However, a little skewed definition of outlier given by Dixon (1950) and Waine (1976) is, "values that are dubious in the eyes of the researcher."

Outliers and influential points are sensitive to regression analysis. According to Osborne & Overbay (2004) possible deadly effects can cause increase in error variance to reduce the power of statistical test, help violet the assumptions in the model which ultimately result a biased estimate. To identify the nature of outliers and how to detect them, some discovering of the methods for dealing with them correctly are essential to fit an unbiased regression model. Despite such severity of outliers in regression model fitting, there is a great deal of debate as to what to do with identified outliers. Meanwhile though, the above-mentioned authors congruently reveal different approaches are in practice to deal with outliers and several indicators are used for identifying and analyzing them. According to Osborne & Overbay (2004) a thorough review of the various arguments about outliers and influential points is almost impossible in a single write and there come situations where researchers must use their training, intuition, reasoned argument, and thoughtful consideration in making decisions about outliers and the influential points.

Sweet & Martin (2012) reveal that, the outliers are deleted if a) they are the wrong entry or b) they are some special cases isolated from a common phenomenon in an analysis. Otherwise, run the analysis first without and second with the outlier excluded. If the outlier is exerting an undue influence on the outcomes, both models should reasonably coincide. Otherwise report the both. Stevens (1983) have reported four diagnostics that are useful in identifying outliers. Namely: studentized residuals [standardizing the deleted residuals produces studentized residuals, (studentized residuals, 2017)], the hat elements,

Cook's distance [a combination of each observation's leverage and residual values; the higher the leverage and residuals, the higher the Cook's distance (Andale, 2016), where leverages are defined as a measure of how far away the independent variable values of an observation are from those of the other observations (Leverage statistics, 2016)], and Mahalanobis distance. Rahaman *et al.* (2012) claims different computer-based approaches (distribution-based, distance-based, density-based and deviation-based) have been proposed for detecting outlying data.

If there is an outlier in the data, rather omit it, the preference would be its effect is removed. Possible ways that any data point can be outlier are: it could have, an extreme x value, an extreme y value, an extreme x and y value and it might be distant from the rest of the data, even without x and y values. According to Bowerman *et al.* (2005) an observation may be an outlier with respect to its y value and /or its x value, but an outlier may or may not be influential.

Influence of outliers is identified by computing regression coefficients with and without outliers. Observations that have a large influence on the estimation results of a regression model are called influential observations. When data set includes influential point, things to consider are: the influential point may be bad data viz. the measurement error, and one needs to check the validity of the data point. Andersen (2012) and, Jacoby (2005) have described outlying observation can cause to misinterpret patterns in plots. More importantly, according to the author, separated points can have strong influence on statistical models viz. unusual cases can substantially influence on the fit of the Ordinary Least Square (OLS) model. And therefore, deleting outliers from a regression model can sometimes give completely different results. Cases that are both outliers and high leverage exert influence on both the slopes and intercept of the model, outliers may also indicate that model fails to capture important characteristics of the data. Bowerman *et al.* (2005) have recommended; first dealing with outliers with respect to their y values and explaining that they could affect the overall fit of the model. According to whom if this was done first, other problems become much less important or disappear.

MATERIALS AND METHODS

Data (Appendix A) for the study were collected from(MOAC, 2010), (FAOSTAT, 2013), (IRRI, 2014) and (DHM, 2014).

Osborne & Overbay (2004) claim, if the studentized residual of an observation is 2 (irrespective of any sign) then the observation is an outlier with respect to its y value. If the leverage value for an observation is greater than $2(k+1)/n$, [where k = number of independent variables and n = number of observation], the observation is outlying with respect it's x value. Moreover, a rough rule of thumb, to determine if an observation is influential is, to calculate Cook's distance measure written as $Cook's D$. If Cook's Distances for the outliers are > 1 , then these outliers are the influential points. In addition to this,

> **If none of these appears to be the case, two analyses—one with the influential cases in and one with these cases deleted—could be reported to emphasize the impact of these few points on the analysis. This is a case where researchers must use their training, intuition, reasoned argument, and thoughtful consideration in making decisions.** (Osborne & Overbay, 2004).

Once identified and if there is a reason to believe that these cases arise from a process different from that for the rest of the data, then the cases should be deleted. For example, the failure of a measuring instrument etc. otherwise, two analyses—one with the influential cases in and one with these cases deleted—could be reported to emphasize the impact of these few points on the analysis. During analysis in Minitab following were observed to be unusual observations (Table 1).

Table 1. Unusual observations.

Obs	harv_area	Prodn_paddy	Fit	SE Fit	Residual	St	Resid
21	1265	1832.6	2083.1	66.7	-250.5	-2.04	R
31	1262	2584.9	2421.6	100.4	163.3	1.68	X

21: R denotes an observation with a large standardized residual (possibly *outlying with respect to y value)*

31: X denotes an observation whose X value gives it large leverage (possibly *outlying with respect to x value*).

These observations therefore were subjected to test for outlying and the influential points. Discussion on these is made in the results and discussion section based on the criterions mentioned above. Following (Table 2) are the essential statistics and the related threshold values to locate either or not the suspected observations were outliers and the influential points,

Table 2. Detecting influential observations.

Observation	Studentized residual		Leverage value		Cook's Distance	
	Observed	Thresh hold	Observed	Thresh hold	Observed	Thresh hold
21	2.04	|2|	0.23	0.23	0.31	1
31	1.68	|2|	0.52	0.23	0.76	1

RESULTS AND DISCUSSION

Observations suspected to be outliers or influential points were checked for their possible wrong recordings. But this was not the case, they were found correctly recorded. Therefore, the other criterions were sought to identify if they were outliers or the influential points.

Table 2. shows that observation 21 is not outlying with respect to its *y* value [studentized residual (2.04) > 2, not significantly greater)] nor was it outlying with respect to its *x* value [leverage value (0.23), not greater than the threshold value].And, for the same (observation 21) Cook's Distance (0.31) < 1 proved that this observation was not influential.

Similarly, from the figures in the same (Table1) show that observation 31 was not an outlying due to its *y* value [studentized residual (1.68) < 2], but clearly was an outlier due to its *x* value [leverage value (0.517) > 0.23]. However again, [Cook's Distance (0.76) < 1] showed that this (observation 31) was also, no more any influential point.

As above we came to the conclusion that neither of the suspected observations were the influential outliers. However, keeping in view that the mentioned criterions many a times (as they have been discussed in the theoretical section) could give malicious results, models with and without the suspected observations were computed (Table 3) and cross checked.

Table 3. Model with and without the suspected outliers.

	Description							
Model	S	R^2	$R^2_{(adj)}$	PRESS	$R^2_{(pred)}$	D-W statistic	Overall lack of fit test is significant at	VIF
Original	139.60	93.3%	92.6%	897983	90.03%	1.96 (1.58)	P = .052	HA(8.66) RP(22.11) PH(9.40)
Obs. 21 deleted	132.02	93.8%	93.2%	741806	91.19%	1.38 (1.58)	(P > = .1)	HA(9.527) RP(27.314) PH(11.460)
Obs.31 deleted	135.26	93.9%	93.3%	756545	91.60%	1.96 (1.58)	(P > = .1)	HA(16.04) RP(30.24) PH(8.78)
Obs. 21 and 31 deleted	130.69	94.1%	93.5%	702891	91.65%	1.37 (1.58)	P = .060	HA(18.95) RP(40.02) PH(11.07)

Figures in the parenthesis for the DW statistics column are the thresh hold value. If $DW >$ upper bound (1.58) at 5% level of significance, no correlation exists between the predictor variables (Makridakis *et al.*, 1998).

Table 3. shows that, neither options either 21 or 31 deleted turn by turn or both 21 and 31, deleted together did give better model. For instance, autocorrelation, lack of fit condition and the multicollinear situation were rather degraded in the newer model as compared to the corresponding values of the models for which the suspected observations were deleted. And hence the model was kept as it was before starting to have a check on the suspected values.

CONCLUSION AND RECOMMENDATIONS

Outliers are the values that lie very far from the middle of the distribution in either direction. They take extreme values compared to most of the observations in a data set. Outliers and influential points could have significant impact in the results of any analysis. Outliers not necessarily be influential in affecting the regression coefficients. They are to be dealt with utmost care. If these influential points are to be removed, it may lead to a different model. Outliers are associated each other. In the presence of the first outlier the second might not act as an outlier. Occurrence of outliers may be by chance. If the occurrence is by chance, they are discarded.

In this paper, outliers, with the tools, studentized residual, leverage value and Cook's distance are checked through the deletion approach of the outliers. And, we have demonstrated through examples that outliers and influential points can be carefully dealt with. This procedure can be applied in the cases when similar situation arises.

REFERENCES

Andersen, R. 2012. SOC6078 Advanced Statistics: Outliers and Influential Cases. Department of Sociology, University of Toronto. Retrieved from http://individual.utoronto.ca/andersen/soc6708/6. DiagnosticsII.pdf Accessed on 11.01.2014.

Andale. (2016). Cook's Distance/Cook's D: Definition, Interpretation. (2016). Statistics how to. Retrieved from http://www.statisticshowto.com/cooks-distance/ Accessed on 7.12.2017.

Bowerman, B. L.; O'Connell, R. T., and Koehler, A. B. (2005). Forecasting, Time Series, and Regression: An Applied Approach. (4th ed.), pp. 258-260. Thomson Brooks/Cole, 10 Davis Drive Belmont, CA 94002, USA.

DHM (2010). Government of Nepal, Department of Hydrology and Meteorology.

FAO (2013). FAOSTAT. Retrieved from http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor Accessed on 14.03.2013.

IRRI (2014). *World Rice Statistics Online Query Facility*. International Rice Research Institute. Retrieved from http://ricestat.irri.org:8080/wrs2/entrypoint.htm Accessed on 06.09.2014.

Jacoby, W. G. (2005). Regression III: Advanced methods. Michigan State University. Retrievedfrom http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week3/11.Outliers.pdf Accessed on 03.20.2014.

Leverage statistics. (2016). Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/wiki/Leverage_(statistics)Accessed on 3.29.2017.

MOAC (2010). Statistical information on Nepalese agriculture. Ministry of Agriculture and Co-operatives, Agri-Business Promotion and Statistics Division [ABPSD].

Osborne, J. W. and Overbay, A. (2004). The Power of Outliers (and why researchers shouldalways check for them). *Practical Assessment, Research & Evaluation,* **9** (6). North Carolina State University. Retrieved from pareonline.net/getvn.asp?v=9&n=6 Accessed on 10.7.2013

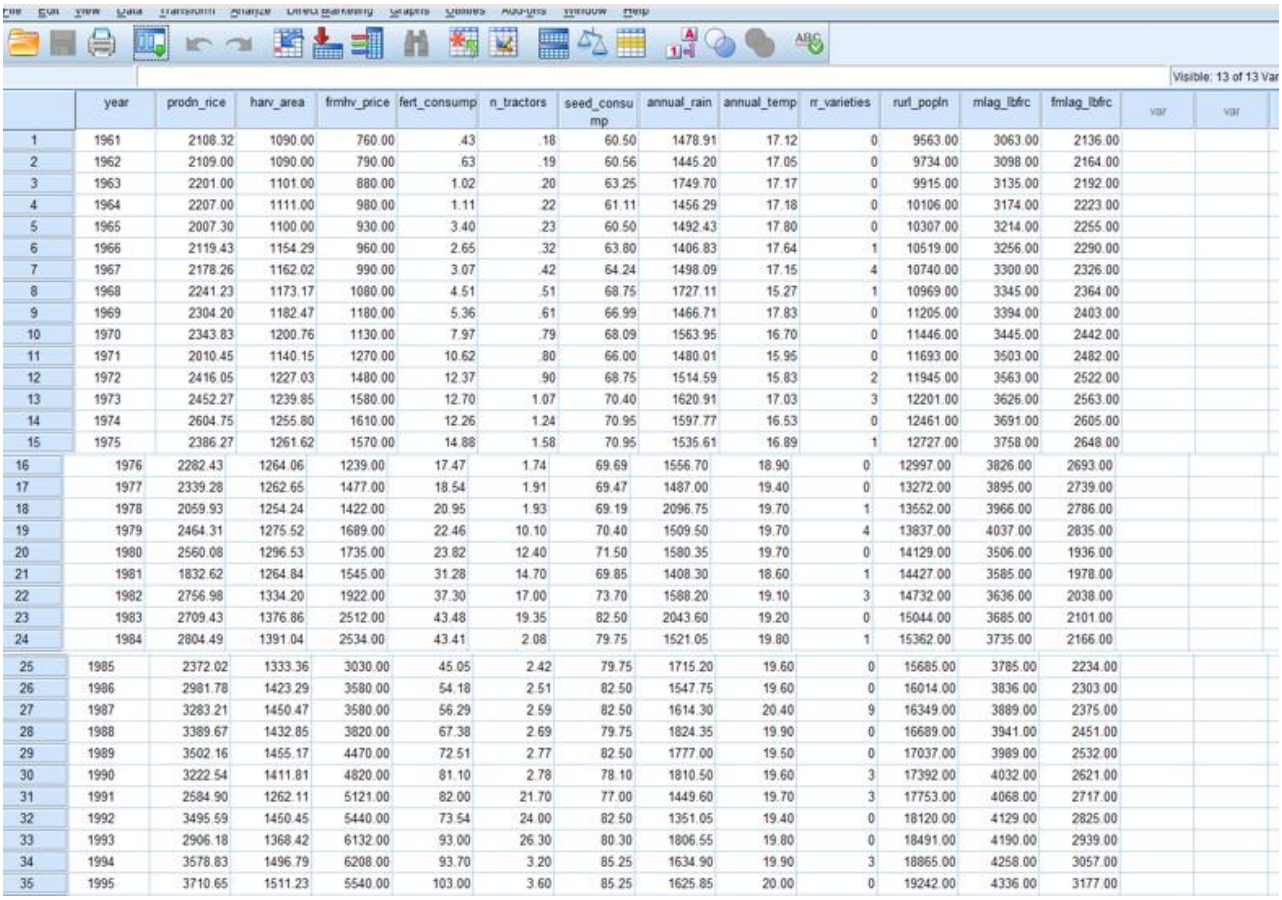
Rahman, S. M. A.; Khaleelur, S.; Mohamed, M. and Kannan, K. S. (2012). Multiple Linear Regression Models in Outlier Detection, *International Journal of Research in Computer Science*, **2** (2).

Stevens. J. (1983). Outliers and Influential Data Points in Regression Analysis. University of Cincinnati. Retrieved from https://pdfs.semanticscholar.org/.../0a48e3f... Accessed on 1.29.2017.

Studentized Residuals. (2017). Stat 426: Applied regression analysis. Penn State Eberly College of Science. Retrieved from https://onlinecourses.science.psu.edu/stat462/node/247 Accessed on 3.29.2017.

Sweet. S. and Martin. K. (2012). Data Analysis with SPSS: A First Course in Applied Statistics (4th ed.), p 181. Pearson Education Inc., Publishing as Allyn and Bacon, USA.

Appendix (A): Data used for fitting the regression model.



	year	prodn_rice	harv_area	frmhv_price	fert_consump	n_tractors	seed_consump	annual_rain	annual_temp	nr_varieties	rurl_popln	mlag_lbfrc	fmlag_lbfrc	var	var
1	1961	2108.32	1090.00	760.00	.43	.18	60.50	1478.91	17.12	0	9563.00	3063.00	2136.00		
2	1962	2109.00	1090.00	790.00	.63	.19	60.56	1445.20	17.05	0	9734.00	3098.00	2164.00		
3	1963	2201.00	1101.00	880.00	1.02	.20	63.25	1749.70	17.17	0	9915.00	3135.00	2192.00		
4	1964	2207.00	1111.00	980.00	1.11	.22	61.11	1456.29	17.18	0	10106.00	3174.00	2223.00		
5	1965	2007.30	1100.00	930.00	3.40	.23	60.50	1492.43	17.80	0	10307.00	3214.00	2255.00		
6	1966	2119.43	1154.29	960.00	2.65	.32	63.80	1406.83	17.64	1	10519.00	3256.00	2290.00		
7	1967	2178.26	1162.02	990.00	3.07	.42	64.24	1498.09	17.15	4	10740.00	3300.00	2326.00		
8	1968	2241.23	1173.17	1080.00	4.51	.51	68.75	1727.11	15.27	1	10969.00	3345.00	2364.00		
9	1969	2304.20	1182.47	1180.00	5.36	.61	66.99	1466.71	17.83	0	11205.00	3394.00	2403.00		
10	1970	2343.83	1200.76	1130.00	7.97	.79	68.09	1563.95	16.70	0	11446.00	3445.00	2442.00		
11	1971	2010.45	1140.15	1270.00	10.62	.80	66.00	1480.01	15.95	0	11693.00	3503.00	2482.00		
12	1972	2416.05	1227.03	1480.00	12.37	.90	68.75	1514.59	15.83	2	11945.00	3563.00	2522.00		
13	1973	2452.27	1239.85	1580.00	12.70	1.07	70.40	1620.91	17.03	3	12201.00	3626.00	2563.00		
14	1974	2604.75	1255.80	1610.00	12.26	1.24	70.95	1597.77	16.53	0	12461.00	3691.00	2605.00		
15	1975	2386.27	1261.62	1570.00	14.88	1.58	70.95	1535.61	16.89	1	12727.00	3758.00	2648.00		
16	1976	2282.43	1264.06	1239.00	17.47	1.74	69.69	1556.70	18.90	0	12997.00	3826.00	2693.00		
17	1977	2339.28	1262.65	1477.00	18.54	1.91	69.47	1487.00	19.40	0	13272.00	3895.00	2739.00		
18	1978	2059.93	1254.24	1422.00	20.95	1.93	69.19	2096.75	19.70	1	13552.00	3966.00	2786.00		
19	1979	2464.31	1275.52	1689.00	22.46	10.10	70.40	1509.50	19.70	4	13837.00	4037.00	2835.00		
20	1980	2560.08	1296.53	1735.00	23.82	12.40	71.50	1580.35	19.70	0	14129.00	3506.00	1936.00		
21	1981	1832.62	1264.84	1545.00	31.28	14.70	69.85	1408.30	18.60	1	14427.00	3585.00	1978.00		
22	1982	2756.98	1334.20	1922.00	37.30	17.00	73.70	1588.20	19.10	3	14732.00	3636.00	2038.00		
23	1983	2709.43	1376.86	2512.00	43.48	19.35	82.50	2043.60	19.20	0	15044.00	3685.00	2101.00		
24	1984	2804.49	1391.04	2534.00	43.41	2.08	79.75	1521.05	19.80	1	15362.00	3735.00	2166.00		
25	1985	2372.02	1333.36	3030.00	45.05	2.42	79.75	1715.20	19.60	0	15685.00	3785.00	2234.00		
26	1986	2981.78	1423.29	3580.00	54.18	2.51	82.50	1547.75	19.60	0	16014.00	3836.00	2303.00		
27	1987	3283.21	1450.47	3580.00	56.29	2.59	82.50	1614.30	20.40	9	16349.00	3889.00	2375.00		
28	1988	3389.67	1432.85	3820.00	67.38	2.69	79.75	1824.35	19.90	0	16689.00	3941.00	2451.00		
29	1989	3502.16	1455.17	4470.00	72.51	2.77	82.50	1777.00	19.50	0	17037.00	3989.00	2532.00		
30	1990	3222.54	1411.81	4820.00	81.10	2.78	78.10	1810.50	19.60	3	17392.00	4032.00	2621.00		
31	1991	2584.90	1262.11	5121.00	82.00	21.70	77.00	1449.60	19.70	3	17753.00	4068.00	2717.00		
32	1992	3495.59	1450.45	5440.00	73.54	24.00	82.50	1351.05	19.40	0	18120.00	4129.00	2825.00		
33	1993	2906.18	1368.42	6132.00	93.00	26.30	80.30	1806.55	19.80	0	18491.00	4190.00	2939.00		
34	1994	3578.83	1496.79	6208.00	93.70	3.20	85.25	1634.90	19.90	3	18865.00	4258.00	3057.00		
35	1995	3710.65	1511.23	5540.00	103.00	3.60	85.25	1625.85	20.00	0	19242.00	4336.00	3177.00		