Visit us - www.researchjournal.co.in ■ DOI : 10.15740/HAS/IRJAES/9.1/25-30
International Research Journal of Agricultural Economics and Statistics
Volume 9 | Issue 1 | March, 2018 | 25-30 ■ e ISSN-2231-6434

Research Paper

Robust modeling in the presence of outliers for food grain production in India

■ **Sunali Mahajan, Manish Sharma and Banti Kumar**

See end of the paper for authors' affiliations

Correspondence to :
Sunali Mahajan
Division of Statistics and Computer Science, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu,
Jammu (J&K) India
Email : sunali12mahajan@gmail.com

ABSTRACT : The traditional ordinary least squares procedure (OLS) is the most frequently used method for analyzing food grain production data (1983-2014), but ignore the presence of outliers or influential data points which may distort the regression estimates obtained from OLS. These data points may remain unnoticed and can have a strong adverse affect on the regression estimates. In this paper, two approaches *i.e.*, robust M-regression and quantile regression to linear robust regression analysis are presented, as these methods provide formal procedure to overcome from the situation of outliers and influential observations and to reduce their influence on the final estimates of the regression co-efficients by using Cobb-Douglas production function. Moreover, 0.90^{th} quantile regression model comes out to be best on the basis of AIC (-47.17), SBIC (-36.91), elasticity of production, marginal value productivity, sign, size and the variables significant effect on foodgrain production than OLS and robust M-regression. Also, the variables NSA and AC were best in order to increase the food grain production on the basis of quantile 0.90^{th} regression, elasticity of production and MVP at 0.90^{th} quantile.

KEY WORDS : Ordinary least square, Outliers, Robust regression, Quantile regression, M-estimator, Food grain production

Paper History :
Received : *28.04.2017;*
Revised : *04.01.2018;*
Accepted : *18.01.2018*

HOW TO CITE THIS PAPER : Mahajan, Sunali, Sharma, Manish and Kumar, Banti (2018). Robust modeling in the presence of outliers for food grain production in India. *Internat. Res. J. Agric. Eco. & Stat.*, **9** (1) : 25-30, **DOI : 10.15740/HAS/IRJAES/9.1/25-30.**

INTRODUCTION :

The term regression was first coined by Galton (1885) in the title of the first paper on the subject "regression mediocrity in heredity stature". In regression analysis, ordinary least square estimators are sensitive to the presence of observations that lie outside the norm for the regression model of interest. The sensitivity of conventional regression methods to these outlier and influential observations can result in co-efficient estimates that do not accurately reflect the underlying statistical relationship and the results are not resistant because of undue influence on estimate of slope as well as intercept (Meintanis and Donatos, 1997). An outlier may arise for many different reasons such as sampling, human, instrument errors etc. and each different reason may require different treatments. To overcome the limitations of the standard Least squares diagnostics, OLS method is directly compared against robust M-estimation and quantile regression methods. Koenker and Bassett (1978); Powell (1984 and 1986); Koenker and Portnoy (1987); Portnoy (1991); Gutenbrunner and Jureckova (1992); Chaudhuri *et al.* (1997); Portnoy and Koenker (1997); Knight (1998); Koenker and Machado (1999); Portnoy

(2003) and He and Zhu (2003) have used conditional quantile models for obtaining consistent estimates of conditional quantiles. Also, Firpo *et al.* (2009) proposed a new regression method to study the impact of changes in the distribution of the exogenous variables on quantiles of the unconditional (marginal) distribution of an outcome variable.

India holds the second largest agricultural land in the world with approximately 179.9 million hectares under cultivation. India today is facing a critical situation in relation to food grain sector. The area under food grain cultivation was 97.32 Mha (1950-51), the productivity stood at 522 kg/ha and production around 51 MT. Population at that point of time was 361.1 million and growing at a modest rate of 1.25 per cent the population by 1961 touched 439.2 million at a growth rate of 1.96 per cent, whereas food grain production increased to about 82 MT (Rai, 2006). In 2013-14, total food grain production in India reached an all-time high of 265.57 MT but in 2014-15 it was 257.07 MT which is lowered by 8.50 MT (Ministry of Consumer Affairs, Department of Food and Public Distribution, 2014-15). In India the estimated projection for the year 2050 will be 2.6 MT in rice, 2.2 MT in wheat, 1.6 MT in pulses (Rai, 2006). So, in order to take the above points in consideration, the primary aim of this study is to compare the parameter estimates for food grain at National level through secondary data over decades by means of OLS, robust M-regression and quantile regression methods, to evaluate the exogenous variables in order to maximize the production of food grain through OLS, robust M-regression and quantile regression methods and drawing conclusions in the presence of outliers and influential observations under the situation where assumptions of LS estimation are untenable.

MATERIALS AND METHODS :

In this study, time series data (1983 to 2014) of food grain have been procured from various online data portals like Ministry of Agriculture, Govt. of India, Directorate of Economics and Statistics, Govt. of India, RBI etc. The exogenous variables used to study the foodgrain production (FP) are net sown area (NSA), net irrigated area (NIA), area under cultivation (AC), consumption of fertilizer (CF), consumption of pesticide (CP) and electricity consumption in agriculture (EC). For outliers and influential observations, the studentized deleted residual and Cook's Distance (Cook, 1979) have been used, respectively. Estimation of parameters has been done through OLS, robust M-regression and quantile regression methods by using Cobb-Douglas production function. The Cobb-Douglas functional form of production functions (multiplicative) is used to represent the relationship of an output to inputs as:

$$y_t = f(x_{k,t}|\beta) \text{ or } y_t = A\sum_{k=1}^{K}(x_{k,t}^{\beta_k})$$

where, k = 1…K is the number of inputs, cross-section i =1 … N, time-series t =1 … T and $\beta_1,\ldots,\beta_k$ are the input elasticities.

The OLS estimator is obtained by $\hat{\beta} = (X'X)-1X'Y$ and is now being criticized more and more for its dramatic lack of robustness (Rousseeuw and Leroy, 1987). More importantly, median regression does not require classical assumptions about the distribution of the regression error terms (Cameron and Trivedi, 2009). To overcome the situation of outliers and influential observations, the robust M-regression and quantile regression methods are used which was first introduced by Huber (1973) and Koenker and Bassett (1978) as a result of making the least square approach robust. Huber's estimator is an extension of the maximum likelihood estimate method which possessing the characteristics of robustness and efficiency (Pol *et al.*, 2006). Instead of minimizing a sum of squares of the residuals, a Huber-type M estimator $\hat{\beta}_M$ of β minimizes a sum of less rapidly increasing function ρ of the residuals:

$$\hat{\beta}_M = {}^{\min}{}_{\beta}\,\rho\left(y_i - \sum_{j=0}^{k} x_{ij}\beta_j\right)$$

The linear conditional quantile function can be estimated by $\hat{\beta}(^2) = \arg\min_{\beta\varepsilon R^p}\sum_{i=1}^{n}\rho_\tau(y_i - x_i'\beta)$ for any quantile τε (0,1). Here, as opposed to OLS, the minimization is done for each subsection defined by ρτ and the quantity $\hat{\beta}(^2)$ is known as the τ^{th} regression quantile.

Elasticity of production and Marginal productivity have also been obtained which is defined as the ratio of proportionate change in output to the proportionate change in a variable input and is expressed as:

$$E_p = \frac{\Delta y/y}{\Delta x_i/x_i}$$

where, Δ is change, y is output and x_i's are inputs. And the Marginal productivity is expressed as:

$$MVP = co-efficient\ of\ x_i * \frac{\text{Geometric mean of } y}{\text{Geometric mean of } x_i}$$

where, y is output and x_i's are inputs.

RESULTS AND DATA ANALYSIS :

Table 1 reveals the summary statistics for endogenous and exogenous variables used in the estimation of production function of foodgrain. The variability can be seen maximum in fertilizer consumption (36.82 %) followed by electricity consumption (35.64 %) whereas, minimum in net sown area (1.66%). Here, the results were consistent for the variables NSA, NIA and AC with co-efficient of variations 1.66, 12.23 and 2.65 per cent as compared to other variables.

The estimated mean for aggregate output stood at 196.43 MT with a minimum value of 140.35MT in 1983 and a maximum of 264.77 MT in 2014.

Table 2 showed that variables NSA, NIA, CF and EC are positively whereas AC and CP are negatively correlated with foodgrain production. Further, regression co-efficients through the traditional OLS present a detailed representation of sign, size and significance of exogenous variables on foodgrain production. The traditional OLS reveal a significant effect of only one variable *i.e.*, consumption of fertilizer on foodgrain production.

Table 3 showed that by studentized deleted residual the observation 32 (*i.e.*, the production for the year 2013-14) was an outlier as their standardized robust residuals exceed the cutoff value -2 to +2 (Meloun and Militky, 2001) and by Cook's distance, observations 31 and 32 (*i.e.*, the production for the years 2012-13 and 2013-14) are influential observations as these values crossed the cut-off line and showed sudden jump according to Fig. 1.

But observation 32 is both outlier and influential observation which has adverse affect on both intercept and slope of the regression line. There are strong reasons to remove outliers and influential but decide to keep them in the analysis and use alternative models to OLS *i.e.*, robust M-regression and Quantile regression models. Thus, outliers and influentials may be the cause of

Table 1: Summary statistics of exogenous variables affecting food grain production in India

Variables (in units)	Mean	Minimum	Maximum	Standard deviation	Co-efficient of variation (%)
FP (million tonnes)	196.43	140.35	264.77	34.36	17.49
NSA (million hectares)	140.92	131.94	143.00	2.35	1.66
NIA (million hectares)	53.46	41.87	63.64	6.54	12.23
AC (million hectares)	123.73	113.87	131.16	3.29	2.65
CF (lakh tonnes)	162.58	77.10	281.22	59.87	36.82
CP (million tonnes)	55531.16	39773.00	75418.00	11549.89	20.79
EC (GWh)	66119.56	18234.00	99023.00	23565.38	35.64

Table 2 : Correlation and estimation of regression co-efficients through OLS of the Cobb-Douglas model

Exogenous variable	Correlation co-efficients with dependent variable	Regression co-efficients (standard error)
NSA	0.1937	0.2883 (1.2949)
NIA	0.6881**	-0.2508 (0.2787)
AC	-0.1936	1.5405 (1.0107)
CF	0.9066**	0.4093* (0.0614)
CP	-0.5498**	-0.0410 (0.1104)
EC	0.7058**	0.1087 (0.0691)

R^2=0.87**, Adj R^2=0.84** and F =27.55** * and ** indicate significance of values at P=0.05 and 0.01, respectively

Table 3: Detection of outliers and influential observations through Studentized deleted residual and Cook's distance of food grain data

Outliers (observations)	Studentized deleted residual value	Influential observations	Cook's distance
32	9.39018	29	0.14880
		31	0.29768
		32	0.31590

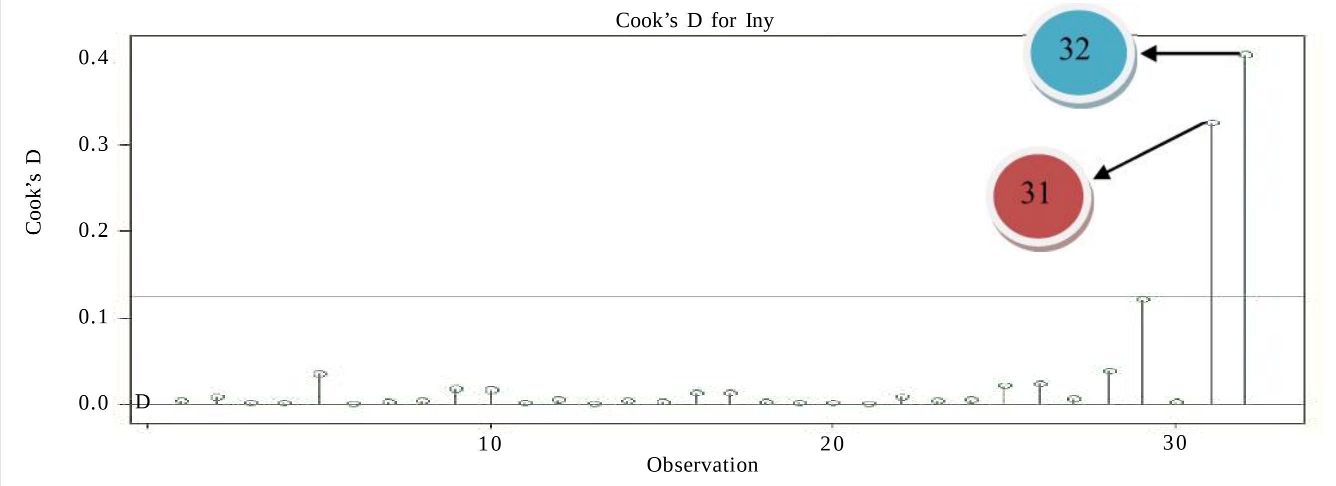

Fig. 1 : Cook's distance for food grain production in India

insignificance of exogenous variable in the regression model.

Table 4 presents a detailed representation of sign, size and significance of exogenous variables on food grain production. The traditional OLS reveal a significant effect of the consumption of fertilizer on food grain production. Unlike traditional OLS, robust M-regression and quantile regression illustrated a positive and statistically significant effect of net sown area (NSA) and consumption of fertilizer (CF) on the production but showed a negative and statistically significant effect of net irrigated area (NIA). These results reveal considerable differences between OLS, robust M-estimation and specifically. 90[th] quantile estimates. The major distinction between the traditional OLS, robust M-regression and quantile regression is the disparity presented by the 0.90[th] quantile regression depicting a significant effect between the agricultural inputs (NSA, NIA, AC, CF, CP and EC) and total production. In addition, quantile regression reveals a clearer representation by depicting 0.90[th] quantile as best wherein each variable maintain a significant effect on food grain production. Also, in OLS, marginal value product (MVP) of resource AC is greater than one, in robust M-regression MVP of NSA is greater than one but when one looks at MVP of 0.90[th] quantile, the resources NSA and AC both are greater than one which

Table 4 : The estimation of regression co-efficients of Cobb-Douglas model as well as marginal value product through robust M-regression and quantile regression at τ=0.90

Variable	Regression co-efficients		Marginal value productivity		
	Robust M-regression (standard error)	τ=0.90 (standard error e-07)	OLS	Robust M-regression	0.90[th] quantile
Constant	-4.7759* (2.2260)	-7.0084* (4.3981)			
NSA	1.6829* (0.6973)	0.8083* (1.3778)	0.39108	2.28290	1.09648
NIA	-0.3324* (0.1501)	-0.4300* (0.2966)	-0.90995	-1.20602	-1.56013
AC	0.2108 (0.5443)	1.4427* (1.0754)	2.37932	0.32558	2.22826
CF	0.4317* (0.0331)	0.4771* (0.0065)	0.52840	0.55732	0.61593
CP	-0.0761 (0.0595)	-0.0555* (0.1175)	-0.00014	-0.00027	-0.00019
EC	0.0606 (0.0373)	0.1164* (0.0736)	0.00035	0.00019	0.00038
SBIC	50.7460	-36.9124			
AIC	35.2366	-47.1726			

* indicate significance of value at P=0.05

Table 5: Elasticity of production for food grain production data

	NSA	NIA	AC	CF	CP	EC
Elasticity	104.7921	2.4325	67.7724	0.9279	6.4964	0.7053

cover both the results of OLS and Robust M. So, from the above findings it is depicted that the use of NSA and AC may be expanded on the basis of MVP (.90th quantile) whereas the use of NIA, FC, PC and EC are curtailed. Moreover, on comparison, the values of AIC and SBIC (-47.17 and -36.91) are minimum in case of 0.90th quantile regression than the robust M-regression. So, on the basis of above discussion 0.90th quantile model comes out to be best in order to increase the food grain production.

Table 5 illustrated a strong and positive effect of NSA and AC on food grain production with 104.79 and 67.77 per cent. Moreover, NIA, CF, CP, and EC showed a small but positive effect on food grain production.

Conclusion:

With reference to our findings, Quantile regression method at .90th quantile comes out to be best for researchers who are estimating the regression parameters in the presence of outliers and influential observations. Our results also indicated that outliers and influential observations should not be automatically rejected but rather should receive special attention and careful examination to determine the cause of their peculiarities. Quantile regression method at .90th quantile allows the researcher's to accommodate data with outliers and influential data points rather than to ignore or delete it.

On the basis of elasticity of production and quantile 0.90th regression, all the exogenous variables are statistically significant in order to increase the food grain production. So, by the results of elasticity of production and MVP (.90th quantile) it is recommended to the farmers that they will use NSA and AC variables more to increase the food grain production.

Proposed model for quantile regression at 0.90th quantile to study the food grain production with respect to exogenous variables is as:

$$Q_{0.90}[\ln(y/x_i)]=-7.0084+0.8083NSA^*-0.4300NIA^*+1.4427AC^*+0.4771CF^*-0.0555CP^*+0.1164EC^*$$

Authors' affiliations:
Manish Sharma and Banti Kumar, Division of Statistics and Computer Science, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, **Jammu (J&K) India**

LITERATURE CITED :

Cameron, A. C. and Trivedi, P. K. (2009). *Microeconometrics using stata*, Stata Corp LP, Texas.

Chaudhuri, P., Doksum, K. and Samarov A. (1997). On average derivative quantile regression. *Annl. Stat.,* **25**(2): 715–744.

Cobb, C.W. and Douglas, P.H. (1928). A theory of production. *American Econ. Rev.,* **18** : 139-65.

Cook, R.D. (1979). Influential observations in linear regression. *J. American Stat. Assoc.,* **19**(1): 15-18.

Firpo, S., Fortin, N.M. and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica,* **77**(3): 953-973.

Galton, F. (1885). Regression towards mediocrity in heredity stature. *J. Anthropological Institute*, **15**: 246-263.

Gutenbrunner, C. and Jureckova, J. (1992). Regression rank scores and regression quantiles. *Annl. Stat.,* **20**(1): 305–330.

He, X. and Zhu, L.X. (2003). A lack-of-fit test for quantile regression. *J. American Stat. Assoc.,* **98** (464):1013-1022.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *Annl Stat.*, **1**(5): 799-821.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Annl Stat.,* **26**(2): 755–770.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica,* **46** (1): 33-50.

Koenker, R. and Portnoy, S. (1987). L-estimation for linear models. *J. American Stat. Assoc.,* **82**(399): 851–857.

Koenker, R. and Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *J. American Stat. Assoc.,* **94**(448): 1296–1310.

Meintanis, S.G. and Donatos, G.S. (1997). A comparative study of some robust methods for coefficient-estimation in linear regression. *Computational Stat. & Data Anal,* **23**: 525–540.

Meloun, M. and Militky, J. (2001). Detection of single influential points in OLS regression model building. *Analytica Chimica Acta*, **439** (2):169-191.

Pol, A.P., Pascual, M.B. and Vazquez, P.C. (2006). Robust estimators and bootstrap confidence intervals applied to tourism spending. *Tourism Mgmt.,* **27**: 42–50.

Portnoy, S. (1991). Asymptotic behavior of regression quantiles in nonstationary, dependent cases. *J. Multivariate Analysis,* **38** (1): 100–113.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise. *Stat. Sci.,* **12**: 279–300.

Portnoy, S. (2003). Censored regression quantiles. *J. American Stat. Assoc.,* **98** (464): 1001-1012.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *J. Econometrics,* **25**: 303–325.

Powell, J.L. (1986). Censored regression quantiles. *J. Econometrics,* **32**:143-55.

Rai, M. (2006). Green revolution II. A lecture note by Director General, Indian Council of Agricultural Research, Ministry of Agriculture, Govt. of India in ASSOCHAM (Association Chambers of Commerce and Industry of India) Summit at New Delhi, India.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection.* Hoboken: Wiley. SAS Users Group International Conference.

WEBLIOGRAPHY

http://dfpd.nic.in/writereaddata/images/pdf/ann-2014-15.pdf.