DOI https://doi.org/10.18551/rjoas.2018-06.06

RANK-BASED ESTIMATION FOR COBB-DOUGLAS MODELLING IN THE PRESENCE OF OUTLIERS

Acquah De-Graft Henry, Associate Professor
Department of Agricultural Economics and Extension, University of Cape Coast,
Cape Coast, Ghana
E-mail: henrydegraftacquah@yahoo.com

ABSTRACT
Ordinary least square (OLS) has been widely used in estimating the Cobb-Douglas production function when analysing the empirical linkage between inputs and outputs. However, the estimates based on OLS technique may be biased by the presence of outliers. Rank-based regression estimation is resistant to outliers and may result in unbiased estimates. The objective of this study is therefore to investigate by use of Monte Carlo methods, the performance of the Rank-based regression and OLS methods in estimating the Cobb-Douglas regression model using data with and without outliers. Monte Carlo simulation results indicate that the estimates of the coefficients of the Cobb-Douglas regression model derived from the Least Squares and the Rank-based estimation methods are accurate and equivalent or close to their true values for normal data regardless of variability in sample size. For data with outliers, Least Squares method is affected by outliers and yields inaccurate estimates of the coefficients of the Cobb-Douglas model across various sample sizes. Rank-based estimation remains robust to outliers in large samples and provides estimates of the coefficients of the Cobb-Douglas Regression model that are accurate and nearly equivalent to their true values. The evidence from Monte Carlo experimentation suggests that the proposed Rank-based estimation is likely to do no worse than the OLS with normal dataset and promise to do better when the dataset has outliers within the Cobb-Douglas production function modelling context. The presence of outliers can bias the results of the OLS estimation of the Cobb-Douglas model and it is recommended that the use of Rank-based regression can be an appropriate method to avoid such biased estimates.

KEY WORDS
Monte Carlo simulation, rank-based regression estimation, Cobb-Douglas production function, ordinary least squares estimation, outlier.

Agricultural economics literature has witnessed extensive use of the Cobb-Douglas production function as a major technique for applied production economics analysis. For example large number of research articles in agricultural economics (Anupama et al., 2005; Mandal et al., 2005; Mruthyunjaya et al., 2005; Pouchepparadjou et al., 2005; Shaheen and Shiyani, 2005; Srinivas and Ramanathan, 2005) have emphasized the use of the Cobb-Douglas production function. Notably, the application of the Cobb-Douglas modelling in these studies uses the OLS estimation methodology (Prajneshu, 2008).

However, the OLS technique used in estimating the Cobb-Douglas production function can lead to misleading results if its fundamental assumptions are not met. In the presence of a small amount of data that behaves differently from the vast majority of the observations (i.e. outliers), the fundamental assumptions of the OLS will be violated and may not be met.

In order words, the presence of outlier observations might bias the results of the OLS method. In an empirical application, Enaami, Mohamed and Ghani (2013) note that outliers could bias the results of the Cobb-Douglas estimation and they solve the problem by developing a new Cobb-Douglas model based on robust methods and the Partial Least Squares paths modelling for parameter estimation. In support of Enaami, Mohamed and Ghani (2013)'s assertion that OLS could bias the results of the Cobb-Douglas estimation in the presence of outliers, this study demonstrates that the Rank-based regression is an

alternative robust method to deal with the problem of bias estimates in Cobb-Douglas modelling context.

Thus an alternative approach to estimate the Cobb-Douglas model while concurrently dealing with the problem of outliers in the data is to employ Rank-based regression. Rank-based regression remains robust to the presence of outliers and has been successfully applied in the estimation of linear models in the presence of outliers as detailed in Jureckova (1971); Jaeckel (1972); McKean and Hettmansperger (1978) and Kloke and McKean (2015). However, very little research has been undertaken to explore the robust and successful Rank-based regression for the Cobb-Douglas production function.

When the production data is contaminated with outliers, Rank-based regression estimation will provide realistic estimates. However, less effort has been made to compare the Rank-based estimation and OLS methods for the estimation of the Cobb-Douglas model in the presence of outliers. The purpose of this research is therefore to investigate by use of Monte Carlo methods, the performance of the Rank-based regression and OLS methods in estimating the Cobb-Douglas production function using data with and without outliers.

The paper is organized as follows. The introduction is followed by the methodology which discusses the Cobb-Douglas model, Ordinary Least Squares (OLS) and Rank-based regression. The results and discussion present Monte Carlo simulations of Cobb-Douglas model and demonstrate the ability of Ordinary Least Squares and Rank-based regression to estimate true values of Cobb-Douglas data generating process. Finally, the paper ends with a conclusion.

METHODS OF RESEARCH

Cobb-Douglas Production Function. The Cobb-Douglas production function is a particular form of production function widely used to represent the technical relationship between the amount of two or more inputs and the amount of output. The Cobb-Douglas production function was developed and tested against statistical development by Charles Cobb and Paul Douglas. In its most standard form for production of a single output with two inputs, the function is:

$$y = AL^{\beta}K^{\alpha}$$

Where: y = output, L and K are inputs (e.g. labour and capital). A is a scale parameter, α and β are elasticities of produce (y) with respect to the input variables.

Linearizing the Cobb-Douglas production function becomes:

$$\ln(y) = A + \beta \ln L + \alpha \ln K + u$$

For constant returns to scale: $\alpha + \beta = 1$

Decreasing returns to scale: $\alpha + \beta < 1$

Increasing returns to scale: $\alpha + \beta > 1$

Ordinary Least Squares (OLS) Estimation. Regression analysis is one of the most widely employed statistical techniques (Takeaki and Hiroshi 2004: xi). The purpose is to illuminate any underlying association between variables by fitting equations to the observed variables, according to some model (Rousseeuw and Leroy 1987:1). The classic linear model relates the dependent or 'response' variable y_i to independent or 'explanatory' variables $x_{i1},...,x_{ip}$ for $i = 1,...,n$, such that:

$$y_i = x_i^T \beta + \epsilon_i \;\; i = 1,...,n$$

Where: $x_i^T = (x_{i1}, x_{i2}, ..., x_{ip})$, $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ and ϵ_i, the 'error' term, is a random variable with expectation 0.

Define the design matrix **X,** and the vectors **Y** and $\boldsymbol{\epsilon}$:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \text{ and } \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Now the classic linear model is $y = x\beta + \epsilon$. The least squares estimator aims to minimize:

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon \quad (3)$$

$$= (y - x\beta)^T (y - x\beta)$$

$$= y^T y - y^T x\beta - \beta^T x^T y + \beta^T x^T x\beta$$

At the minimum:

$$\frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n} \epsilon_i^2\right) = \frac{\partial}{\partial \beta}\left(y^T y - y^T x\beta - \beta^T x^T y + \beta^T x^T x\beta\right)$$

So the least squares estimator $\hat{\beta}$ is the solution to:

$$x^T x\hat{\beta} = x^T y$$

As this minimises $\hat{\epsilon}^T \hat{\epsilon} = \sum_{i=1}^{n} r_i^2$. Thus when $x^T x$ is non-singular the least squares estimator can be evaluated directly from the data:

$$\hat{\beta} = (x^T x)^{-1} x^T y$$

Rank-Based Regression. The aim of Rank-based regression as pertains to least squares, is to estimate the vector of coefficients, β, of a general linear model of the form:

$$y_i = \alpha + x_i^T \beta + e_i \text{ for } i = 1, \ldots, n$$

Where: y_i is the response variable, x_i is the vector of explanatory variables, α is the intercept parameter and e_i is the error term which is assumed as iid with probability density function (pdf), f(t). It can be written in matrix notation as:

$$y = \alpha 1 + X\beta + e$$

Where: $y = [y_1, \dots, y_n]^T$ is the n x 1 vector of response variable, $X = [x_1, \dots, x_n]^T$ is the n x p design matrix, and $e = [e_1, \dots, e_n]^T$ is the n x 1 vector of error terms.

The only assumption on the distribution of the errors is that it is continuous, making the model general. The least squares estimator minimizes the Euclidean distance between **y** and $\widehat{\boldsymbol{y}}_{LS} = \boldsymbol{x}\widehat{\boldsymbol{\beta}}_{LS}$. A different measure of distance which is based on the dispersion function of Jaeckel (1972) is used to obtain the R estimator and the function is given by:

$$D(\beta) = ||Y - X\beta||_{\varphi}$$

Where $||\,.\,||_{\varphi}$ is a pseudo-norm defined as:

$$||\boldsymbol{u}||_{\varphi} = \sum_{i=1}^{n} a(R(u_i))u_i$$

Where: R denotes rank, $a(t) = \varphi\left(\frac{t}{n+1}\right)$, and φ is a non-decreasing, square-integrable score function defined on the interval (0, 1). Assume without loss of generality that it is standardized, so that $\int \varphi$ (u) du = 0 and $\int \varphi^2$ (u) du = 1.

The R estimator of β is defined as:

$$\hat{\beta}_{\varphi} = Argmin\, ||y - X\beta||_{\varphi}. \quad (9)$$

This estimator is a highly efficient one which is robust in the Y-space. A weighted version can attain 50% breakdown in the X-space at the expense of a loss in efficiency (Chang et al., 1999).

RESULTS AND DISCUSSION

Estimation of Cobb-Douglas Model using Rank-based Regression and OLS. The Cobb-Douglas data generating process can be specified as follows:

$$\ln\ y\ =\ 1\ +\ 0\,.7\ \ln\ x_1\ +\ 0\,.3\ \ln\ x_2\ +\ u \quad (10)$$

The independent variables x_1 and x_2 are obtained by taking exponential of uniform distributed random variables. For normal data, the errors u are normally distributed with a mean of 0 and a variance of 0.2 [$u \sim N(0, 0.2)$]. For data with outliers, five observations of the errors generated for the normal data with values generated from the normal distribution with a mean of 0 and a variance of 0.2 is replaced with five observations from the normal distribution with a mean of 20 and variance of 0.2 ($\varepsilon \sim N(20, 0.2)$).

The performance of the Rank-based regression and OLS in estimating the true values of the Cobb-Douglas model are investigated using 1000 regressions based on the Cobb-Douglas Model specified in Eq. (10). The Monte Carlo experiments are conducted under conditions of different sample sizes (50, 150 and 500) and coefficients of $(\beta_0, \beta_1, \beta_2) \in (1.0, 0.7, 0.3)$ are assigned to the Cobb-Douglas model with normal data as well as the data with outliers. The parameters of the Cobb-Douglas model are assigned in the spirit of Behr (2015).

Monte Carlo simulation results obtained for the normal data are reported in Table 1. Results of 1000 Monte Carlo simulations indicate that the estimates of the coefficients of the Cobb- Douglas model obtained from the Rank-based regression analysis are accurate and close to their true parameter values for the normal data (data without outliers) with small and moderate sample sizes (50 and 150). The estimates of the coefficients of the Cobb-Douglas model obtained from the least squares methods are accurate and close to their true parameter values for the data without outliers (Normal data) with small and moderate sample sizes (50 and 150). Additionally, the estimates of the coefficients of the Cobb-Douglas model derived from the Least squares method and the Rank-based regression approach are accurate and equivalent to their true parameter values for normal data with large sample size (500).

Fundamentally, Table 1 demonstrates that in the absence of outliers, the OLS and Rank-based regression analysis performed well, with the averaged estimates all nearly equal or close to their true values of $\beta_0 = 1.0, \beta_1 = 0.7, \beta_2 = 0.3$ regardless of the different sample sizes. These results are consistent with Ryan (1997) who asserts that robust regression estimation technique performs almost as well as OLS when the data has no outliers. Similarly, Chen, Tang, Lu and Tu (2014) noted that in the absence of outliers, OLS and Rank-based methods performed well, with the averaged estimates all nearly identical to the true values in linear regression analysis.

Table 1 – Normal Data (Without Outliers)

Sample Size	*Properties of Data*		Estimates		
		Method	β_0	β_1	β_2
		OLS	1.05	0.64	0.31
N=50	Normal	Rank Regression	1.04	0.70	0.34
Sample Size	*Properties of Data*		Estimates		
		Method	β_0	β_1	β_2
		OLS	1.03	0.70	0.30
N=150	Normal	Rank Regression	1.02	0.70	0.30
Sample Size	*Properties of Data*		Estimates		
		Method	β_0	β_1	β_2
		OLS	1.00	0.70	0.30
N=500	Normal	Rank Regression	1.00	0.70	0.30

Based on 1000 Monte Carlo Simulations.

Monte Carlo studies results derived for the data with outliers are reported in Table 2. Results of 1000 Monte Carlo simulations indicate that the estimates of the coefficients of the Cobb- Douglas model derived from the Rank-based regression analysis are accurate and close to their true parameter values for the data with outliers in large sample (500). Generally, as sample size increases from small through moderate to large sample, estimated coefficients of the Cobb- Douglas model move closer to their true parameter values in the Rank-based regression analysis.

The Ordinary Least Squares method performed poorly in the presence of outliers as illustrated in Table 2. In small, moderate and large samples of 50, 150 and 500 respectively, the Ordinary Least Squares (OLS) estimator performs poorly with its parameter estimates entirely different from the true parameter values of $\beta_0 = 1.0, \beta_1 = 0.7, \beta_2 = 0.3$ as specified in the Cobb-Douglas data generating process.

Noticeably, the results of the Rank-based regression analysis are similar to that of the Ordinary Least Squares and close to their true values when the data contains no outliers. However, when the data contains outliers, Rank-based regression analysis remains robust to outliers in large samples whilst the least squares is influenced by outliers in small, moderate and large samples.

The results are consistent with previous studies. For example, Chen, Tang, Lu and Tu (2014) noted that in the presence of outliers, classic linear models yield extremely large estimates that are un-interpretable, whilst in contrast, the Rank-based regression model generated estimates close to their true values. Furthermore, Ryan (1997) notes that robust methods such as Rank-based estimation methods perform much better than OLS when the data has outliers. Similarly, the results is consistent with Jureckova (1971); Jaeckel (1972); McKean and Hettmansperger (1978) and Kloke and McKean (2015)'s assertion that Rank-based regression remains robust in the presence of outliers in the data. Additionally, Rousseeuw and Leroy (2003) asserts that OLS estimator is extremely sensitive to multiple outliers in linear regression analysis and it can even be easily biased by just a single outlier because of its low breakdown point. Chatterjee and Hadi (2006) also note that unlike OLS estimator, robust regression estimators provide robust estimates even in the presence of multiple outliers. The impact of outliers when using robust regression is minimized by giving smaller weight for outliers in the estimation procedure. These results also confirm the claims of Enaami, Mohamed and Ghani (2013) that outliers in production data bias the results of OLS, while robust regression methods provide unbiased estimates of the Cobb-Douglas model as noted in the Rank-based regression.

Table 2 – Data with Outliers

Sample Size	*Properties of Data*		Estimates		
		Method	β_0	β_1	β_2
		OLS	5.31	-2.25	-1.30
N=50	With Outliers	Rank Regression	1.08	0.62	0.30
Sample Size	*Properties of Data*		Estimates		
		Method	β_0	β_1	β_2
		OLS	2.30	-0.90	0.70
N=150	With Outliers	Rank Regression	1.05	0.70	0.24
Sample Size	*Properties of Data*		Estimates		
		Method	β_0	β_1	β_2
		OLS	1.12	0.82	0.40
N=500	With Outliers	Rank Regression	1.02	0.70	0.30

Based on 1000 Monte Carlo Simulations.

CONCLUSION

The performance of Rank-based regression analysis has been explored in the Cobb-Douglas production function regression modelling. The results suggest that with normal data, the Rank-based regression approach yield comparable results to the OLS. However, when outliers are present in the data, the least squares provides inaccurate estimates of the coefficients of the true Cobb-Douglas model in small, moderate and large samples of data. Rank-based regression estimation, on the other hand, is resistant to outliers and provides exact estimates of the coefficients of the true Cobb-Douglas model in large samples. The Monte Carlo simulation results indicate that the Rank-based regression estimation can be considered an alternative to the OLS technique in estimating the Cobb Douglas production function and may yield accurate results in large samples when the data contains outliers. The evidence from the study suggests that the presence of outliers can bias the results of the OLS estimation of the Cobb-Douglas model while the Rank-based regression can be an appropriate method to produce unbiased estimates in the presence of outliers.

REFERENCES

1. Anupama, J., Singh, R. P. & Kumar, R. (2005). Technical Efficiency in Maize Production in Madhya Pradesh: Estimation and Implications. Agricultural Economics Research Review, 18: 305-15.
2. Behr, A. (2015). Production and Efficiency Analysis with R. Springer International.
3. Chang, W. H., McKean, J. W., Naranjo, J. D. & Sheather, S. J. (1999). 'High-Breakdown Rank Regression,' Journal of the American Statistical Association, 205-219
4. Chatterjee, S. & Hadi, A. S. (2006). Regression Analysis by Example. New Jersey: John Wiley & Sons.
5. Chen, T., Tang, W., Lu, Y. & Tu, X. (2014). Rank regression: An Alternative Regression Approach for Data with Outliers. Shanghai Archives of Psychiatry, *26*(5), 310.
6. Enaami, M., Mohamed, Z. & Ghani, S. (2013). Model Development for Wheat Production: Outliers and Multicollinearity Problem in Cobb-Douglas Production Function. Emirate Journal of Food and Agriculture. 25(1):81-88.
7. Jaeckel, L. A. (1972). Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals. The Annals of Mathematical Statistics*,* 43,1449-1458.
8. Jureckova, J. (1971). Non-parametric Estimate of Regression Coefficients. The Annals of Mathematical Statistics, 42,1328-1338.
9. Kloke, J. & McKean, J. W. (2015). Non-parametric Statistical Methods Using R. New York: CRC Press. ISBN-13:978-1-4398-7343-4.
10. Mandal, S., Datta, K. K., Dayal, Bhu, Minhas, P. S. & Chauhan, C. P. S. (2005). Resource Use Efficiency in Saline Irrigated Environment. Indian Journal of Agricultural Economics, 60: 494-509.
11. McKean, J. W. & Hettmansperger, T. A. (1978). Robust Analysis of the General Linear Model Based on One Step R-Estimates. Biometrika, 65(3):571.
12. Mruthyunjaya, Kumar, Sant, Rajashekarappa, M. T., Pandey, L. M., Ramanarao, S. V. & Narayan, P. (2005). Efficiency in Indian Edible Oilseed Sector: Analysis and Implications. Agricultural Economics Research Review, 18: 153-66.
13. Pouchepparadjou, A., Kumaravelu, P. & Achoth, L. (2005). An Econometric Analysis of Green Technology Adoption in Irrigated Rice in Pondicherry Union Territory. Indian Journal of Agricultural Economics, 60: 660-76.
14. Prajneshu (2008). Fitting of Cobb-Douglas Production Functions: Revisited. Agric. Econ. Res. Rev. 21:289-292
15. Rousseeuw, P. J. & Leroy, A. M. (1987). Robust Regression and Outlier Detection. Hoboken: Wiley.
16. Rousseeuw, P. J. & Leroy, A. M. (2003). Robust Regression and Outlier Detection. New York: John Wiley & Sons.
17. Ryan, T. P. (1997). Modern Regression Methods. New York, NY: John Wiley & Sons, Inc.
18. Shaheen, F. A. & Shiyani, R. L. (2005). Water-use Efficiency and Externality in the Groundwater Exploited and Energy Subsidised Regime. Indian Journal of Agricultural Economics, 60: 445-57.
19. Srinivas, T. & Ramanathan, S. (2005). A Study on Economic Analysis of Elephant Foot Yam Production in India. Agricultural Economics Research Review, 18: 241-52
20. Takeaki, K. & Hiroshi, K. (2004). Generalized Least Squares. Chichester: Wiley.