International Journal of Mathematics and Statistics Studies

Vol.5, No.4, pp.38-44, August 2017

Published by European Centre for Research Training and Development UK (www.eajournals.org)

ROBUSTNESS OF TWO-PART FRACTIONAL REGRESSION MODELS IN MODELLING FRACTIONAL OUTCOMES

Ayiah –Mensah Francis

Department of Mathematics and Statistics. Takoradi Polytechnic

ABSTRACT: *The bounded nature of the fractional dependent variables, for instance in corporate finance leverage ratio clustering with a substantial number of observations at unit interval raises some important issues in estimation and inference. Ordinary Least Square (OLS) regression with Gaussian distributional assumption has been the main choice to model fractional outcomes in many business problems. Nevertheless, it is conceptually flawed to assume Gaussian distribution for a response variable in the interval [0,1]. Tobit model which is a Single-component method for modelling proportional outcome also share properties with OLS. Two-part Fractional regression models have been shown as the most natural way of modelling bounded, proportional response variables. Beta regression method has been used to achieve the objective in this paper.*

KEYWORDS: Fractional outcomes, Ordinary Least Square, Gaussian distribution, Tobit model, Beta regression

INTRODUCTION

The normal distribution is one of the mainly used distributions in statistical analysis. It is sometimes called the Gaussian distribution and has the basis of much parametric statistical analysis. An assumption is often made that the sample under test is from a population with normal distribution. Making this assumption about the data requires the use of parametric tests which are more powerful than their equivalent non-parametric methods.

Many variables by their nature naturally follow the normal distribution, for example, biological variables such as blood pressure, serum cholesterol, height and weight. You could choose to skip the normality check in these cases, though it's always wise to check the sample distribution.

The dependent variable in many economic models is often proportion, percentage or fraction. Cook et al (2008) indicated that many studies in corporate finance ignored the fact that proportion, percentage or fractional data are not normally distributed since data in these form are not observed and defined over the range of the normal distribution. The bounded nature of the fractional dependent variables raises some important issues in estimation and inference. To assume a linear functional form in estimating these models is conceptually flawed.

Until the discovery of fractional regression models (FRM), the simple OLS regression with Gaussian distributional assumption remained the most popular method to model fractional outcomes due to its simplicity. According to Kennedy (2003) some of the desirable properties of the OLS estimate may no longer hold in case of a continuous fractional or proportional dependent variable. It is therefore required to use an appropriate estimation technique to analyze bounded nature data.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

Over the past years researchers who have interest in statistical models for fractional outcomes have developed models for data analysis in the financial, education and other sectors of the economy. The distinctive statistical nature of fractional outcomes is that the variance is not independent of the mean. For example a problem of heteroscedasticity is seen in regression models where the variance shrinks as the mean approaches boundary points [0,1]. It is also important to note that fractional outcomes in a unit interval are not defined on the whole real line hence they should not be considered normally distributed. It is important to note that data measured in a continuous scale and restricted to the unit interval, i.e $0 < y < 1$ as in the case of percentages, proportions, fractions and rates.

An empirical application of the fractional regression models is in the area of financial service industry. In business, there is a necessity to model fractional outcomes in a unit interval [0,1]. The variable of interest in many economic settings is often a fraction or a proportion, being defined only on the unit interval. Such variables are bounded by their nature. In some cases, there is a possibility of nontrivial probability mass accumulating at one or both boundaries raise some interesting estimation and inference issues.

METHODOLOGY

Models for response variables defined on the unit interval was first proposed by Papke and Wooldridge(1996). The Statistical models for fractional outcomes that are considered in this study are the single component model and the two part model. The Tobit model is an example of a single component while the Beta model is used to explain the two-parts model. In the two parts modeling approach, for instance a Logit model separating between boundary points and the open interval of (0,1) will be used initially and the other governing all values in the (0,1) interval by a Beta model.

A logit transformation can be used to solve the problem for response variable values within the unit interval. Given that $y = \frac{1}{1+exp(-X\beta)}$ it will yields the transformed response variable y^* such that $y^* = log\left(\frac{y}{1-y}\right) = X\beta + e$

Tobit Model

The Tobit model which is being used to model outcomes with unit interval [0,1] can be seen in regression when the dependent variable is incompletely observed and the regression when the dependent variable is completely observed but is observed in a selected sample that is not representative of the population. This shares the feature of the OLS regression and leads to inconsistent parameter estimate because the sample is not representative of the population.

The Tobit model which is being used to model outcomes with unit interval [0,1] is assumed that there is a latent variable Y^* such that

$$Y = \begin{cases} 0 & for \quad Y^* \leq 0 \\ \grave{X}\beta + \varepsilon & for \quad 1 > Y^* > 0, \quad where\ the\ error\ term\ \varepsilon \sim Normal\ (0, \sigma^2) \\ 1 & for \quad Y^* \geq 1 \end{cases}$$

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

The response Y is bounded by [0,1] and it might be considered the observable part of a normally distributed variable $Y^* \sim Normal(\grave{X}\beta, \sigma^2)$ being defined on the real line.

The model assumes that the observed dependent variables Y_j for observation $j = 1, \dots, n$

$Y_j = max(Y_j^*, 0)$ where $Y_j^{*\prime}s$ are latent variables generated by the classical linear regression model $Y_j^* = \acute{\beta} X_j + \varepsilon_j$ with X_j a vector of regressors, possibly including 1 for the intercept, and β the corresponding vector of parameters.

The conditional c.d.f. of Y_j given $Y_j > 0$ and X_j is

$$H(y|Y_j > 0, X_j, \beta, \sigma) = P(y_j \le y|Y_j > 0, X_j)$$

$$= \frac{P(0 < Y_j^* \le y|x_j)}{P(Y_j^* > 0|X_j)} = \frac{P(-\acute{\beta} X_j < \varepsilon_j \le y - \acute{\beta} X_j|X_j)}{P(\varepsilon_j > -\acute{\beta} X_j|X_j)}$$

$$= \frac{F\left((y - \acute{\beta} X_j)/\sigma\right) - F(-\acute{\beta} X_j/\sigma)}{F(\acute{\beta} X_j/\sigma)}$$

The maximum likelihood estimator for Tobit model assumes that errors are normal and homoscedastic and would be otherwise inconsistent. As a result the simultaneous estimation of a variance model might be needed to account for the heteroscedasticity as follows

$$E(\varepsilon^2) = \sigma^2 \times \left(1 + EXP(\grave{Z} G)\right)$$

Beta Model

The beta model is based upon the two-parameter Beta distribution and can be employed to model any continuous variable bounded by two known endpoints, for example zero (0) and one (1). With the assumption that Y follows a standard beta distribution defined in the interval (0,1) with two shape parameters ω and τ. While ω is pulling the density toward 0, τ is pushing density toward 1. The density function can be specified as;

$$F(Y) = \frac{\Gamma(\omega + \tau)}{\Gamma(\omega)\Gamma(\tau)} \times Y^{\omega - 1} \times (1 - Y)^{\tau - 1}$$

Ferrari and Cribari-Neto (2004) proposed the reparametrization of the beta distribution. With their proposal, a random variable Y follows a Beta distribution if its probability function (pdf) is given by

$$b(y|\mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma(1-\mu)\phi} y^{\mu\phi - 1}(1 - y)^{(1-u)\phi - 1}, 0 < y < 1$$

Where $0 < \mu < 1$ and $\phi > 0$. Without the loss of generality, ω and τ can be translated into two other parameters, location parameter μ and dispersion parameter ϕ with $\omega = \mu\phi$ and

$\tau = \phi(1 - \mu)$.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

The notation $Y \sim Beta(\mu, \phi)$ with the mean and the variance expressed as $E(y|\mu, \phi) = \mu$ and $Var(y|\mu, \phi) = \frac{V(\mu)}{1+\phi}$ where $V(\mu) = \mu(1-\mu)$, μ is the mean and ϕ can be interpreted as a precision parameter.

Although the Beta distribution is initially considered as flexible since the pdf can have different shapes by considering different values of μ and ϕ. Hahn (2008) and Garcia et al. (2011) noted that it does not take into consideration, the greater flexibility in the variance specification and the tail-area events. Bayes et al. (2012) considered that this fact could limit its application for modeling proportions. In solving the problem by getting some additional flexibility, they provided a regression model which permits varying amounts of dispersion and greater likelihood of more extreme tail-area events by considering Beta rectangular distribution proposed by Hahn(2008).

$$P(y; \mu, \sigma^2) = [2\pi\sigma^2\{y(1-y)\}^3]^{-1/2} exp\left\{-\frac{1}{2\sigma^2} d(y; \mu)\right\}$$

$$y \in (0,1), \quad \mu \in (0,1)$$

$$d(y, \mu) = \frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}$$

$$S^-(\mu; \sigma^2)$$

$$\mu = E(Y)$$

$$V(\mu) = \mu^3(1-\mu)^3$$

This model avoids the problems associated with using linear models in the Data Envelopment Analysis (DEA) framework. The first problem was that when a DEA analyses use a linear conditional mean model given by $E(y|x) = x\theta$, to explain efficiency scores, the linearity assumption is unlikely to hold because the conceptual requirement that the predicted values of y lie in the unit interval will not be satisfied. The second problem is that the marginal effect on the DEA score of a unitary change in covariate x_j, is constant over the entire range of y. This is not compatible with either the bounded nature of DEA scores or the existence of a mass point at unity in their distribution.

The Papke and Wooldridge(1996) approach is to solve the problem where the dependent variable is defined on a unit interval whether or not the boundary values are observed. Ramalho et al. (2010) indicated that the fractional regression models only requires the assumption of a functional form of y that imposes the desired constraints on the conditional mean of the dependent variable such that $E(y|x) = G(x\theta)$, where $G(\cdot)$ is some nonlinear function satisfying $0 \leq G(\cdot) \leq 1$. While Papke and Wooldridge(1996) suggested that the model defined by $E(y|x) = G(x\theta)$ may be consistently estimated by QML, Ramalho et al. (2010) suggested the use of nonlinear least square or maximum likelihood estimation.

The estimation of fractional regression models by QLM is based on the Bernoulli log-likelihood function given by $L_i(\theta) = y_i log[G(x_i\theta)] + (1-y_i) log[1 - G(x_i\theta)]$. According to Gourieroux et al (1984), given that the Bernoulli distribution is a member of the linear exponential family, the QLM estimator of θ defined by $\overset{\wedge}{\theta} \equiv \underset{\theta}{argmax} \sum_{i=1}^{N} LL_i(\theta)$ is consistent and asymptotically normal. Regardless of the distribution of y conditional on x, provided that

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

$E(y|x)$ in $E(y|x) = G(x\theta)$ is indeed correctly specified. Given the bounded nature of a response variable that appears as a proportion fraction, both Hoff (2007) and McDonald (2009) considered the use of Papke and Wooldridge's (1996) logit fractional regression model. Using a proportion in a linear regression model will generally yield nonsensical predictions for extreme values of the regressors.

RESULTS AND DISCUSSIONS

The limitation of the OLS regression necessitated the need for an alternative method in modeling fractions. Tobit encompasses single-component modeling approach to analyze fractional outcome in the close interval [0,1]. Beta covers two part modeling with one model. For instance a Logit model, separating between boundary points and the open interval of (0,1) and the other governing all values in the (0,1) interval by a Beta.

Bayes et al.(2012) did a simulation studies on the influence of outliers and confirm that the Beta rectangular regression model is seems to be a new robust alternative for modeling proportional data. They also revealed that Beta regression model offers sensitivity in the estimation of regression coefficients, sensitivity on the posterior distribution of all parameters.

Vuong(1989) proposed a test which is a likelihood-based measures to compare multiple models with different distributional assumptions called Vuong statistics. Statisticians often prefer to use this test because it is considered as a better model with the individual log likelihoods which is significantly higher than the ones of its rival and is calculated as

$$Vuong \text{ Statistics} = \frac{LR(Model1, Model2) - C}{\sqrt{N \times V}} \sim Normal(0,1)$$

where

$LR(Model1, Model2)$ is the summation of individual log likelihood ratios between the two models.

C is a correction term for the difference in parameter numbers between the two models.

N is the number of records.

V is the variance of individual log likelihood ratio between two models.

Vuong statistics is distributed as a standard Normal (0,1). The model 1 is better with Vuong statistics >1.96 and the model 2 is better with Vuong statistics <-1.96. Liu(2014) analysis compared Tobit (model 1) and Zero inflated Beta (model 2) regressions. The result from Vuong statistics in modeling the financial leverage ratios of business was -5.84 which implied that the the zero-inflated Beta is significantly better than the Tobit.

CONCLUSION

Generally the OLS on the whole sample or just the uncensored sample will provide inconsistent estimate of β. If we consider OLS on the uncensored sample, the expected value of the latent

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

variable of $y|y > 0$ therefore $E[y|y > 0] = X_i\beta + \sigma\lambda(\alpha)$ where $\lambda(\alpha) = \frac{\phi\left(\frac{X_i\beta}{\sigma}\right)}{\Phi\left(\frac{X_i\beta}{\sigma}\right)}$ is the inverse Mills ratio. We have $y_i = X_i\beta + \sigma\lambda\left(\frac{X_i\beta}{\sigma}\right) + e_i$ therefore $E(e_i|X_i, y_i > 0) = 0$. We mistakenly omit $\sigma\lambda\left(\frac{X_i\beta}{\sigma}\right)$ in our OLS regression. The effect of this omitted term will appear in the disturbance term, which means that the Xs will be correlated with the disturbance term, leading to inconsistent estimates.

In the Tobit model, the response Y bounded by [0,1] could be considered the observable part of a normally distributed variable defined on a real line. The unobservable values in the interval is not as a result of the censorship but any value out of the interval is not theoretically defined. As a result the censored normal distribution assumption will not be the best for fractional outcome. Tobit is based on the normal distribution and the probability function of any value in the (0,1) is identical to OLS regression. This share the feature of the OLS regression and leads to inconsistent parameter estimate because the sample is not representative of the population.

It has been shown that the linear or Tobit regression model are not appropriate in modeling fractional data. Beta fractional regression model have flexibility for working with data where there are outlying observations.

REFERENCES

Bayes, C. L., Bazan, J. L. and Garcia, C. (2012). A new robust regression model for proportions. International Society for Bayesian Analysis. pp 841-866.

Choi, Y. M. (2013). Estimation of Fractional Dependent Variables Observed on [0, 1) with an Application to Firm Capital Structure. International Journal of Digital Content Technology and its Applications(JDCTA). Volume 7, Number 13.

Cook, D. Kieschnick R. McCllough,(2008). Regression analysis of proportions in finance with self selection, Journal of Empirical Finance 15, pp 860-867.

Ferrari, S. and Cribari-Neto, F. (2004). "Beta regression for modelling rates and pro-Portions" *Journal of Applied Statistics*, 31: 799-815.

Garcia, C., Garcia, J., and van Dorp, J. R. (2011). Modeling heavy-tailed, skewed and peaked uncertainty phenomena with bounded support." *Statistical Methods and Applications*, 20(4): 463-486.

Gouriéroux, C., Monfort, A. and Trognon, A. (1984), "Pseudo maximum likelihood methods: applications to Poisson models", Econometrica, 52(3), 701-720.

Hoff, A. (2007), "Second stage DEA: comparison of approaches for modelling the DEA score", European Journal of Operational Research, 181, 425-435.

Kennedy, P. (2003). A Guide to Econometrics, The MIT Press, 5th edition.

Hahn, E. D. (2008). "Mixture densities for project management activity times: A robust approach to PERT." *European Journal of Operational Research*, 188: 450-459.

Liu, W. (2014). Modeling fractional outcome with SAS. http://support.sas.com/resources/papers/proceedings14/1304-2014.pdf. Accessed March 2015.

McDonald, J. (2009), "Using least squares and tobit in second stage DEA efficiency

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)

analyses“, European Journal of Operational Research, 197(2), 792-798.
Papke, L. E. and Wooldridge, J. M. (1996) Econometric methods for fractional response variables with application to test pass rates. Journal of Econometrics 11. Pp 619-632.
Ramalho, E.A., Ramalho, J.J.S. and Murteira, J. (2010),“Alternative estimating and testing empirical strategies for fractional regression models“, Journal of Economic Surveys.
Vuong, Q. (1989), Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57, 307 – 333.

ISSN 2053-2229 (Print), ISSN 2053-2210 (Online)