Statistical Methodology 23 (2015) 50–58

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

An approximation of logarithmic functions in the regression setting

Tao Chen [a,b,*], Kenneth A. Couch [c]

[a] Department of Economics, University of Waterloo, Waterloo, ON N2L 3G1, Canada

[b] School of Business, Jianghan University, Hubei, 430056, China

[c] Department of Economics, University of Connecticut, Storrs, CT 06269, USA

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form
5 September 2014
Accepted 5 September 2014
Available online 8 October 2014

Keywords:
Box–Cox transformation
Censoring
Method of moments

ABSTRACT

We consider a method of moments approach for dealing with censoring at zero for data expressed in levels when researchers would like to take logarithms. A Box–Cox transformation is employed. We explore this approach in the context of linear regression where both dependent and independent variables are censored. We contrast this method to two others, (1) dropping records of data containing censored values and (2) assuming normality for censored observations and the residuals in the model. Across the methods considered, where researchers are interested primarily in the slope parameter, estimation bias is consistently reduced using the method of moments approach.

1. Introduction

In the economics literature, there are many examples where variables are expressed in logarithmic terms in a regression setting. Perhaps the best known case is in the estimation of human capital earnings functions [1,2]. In some contexts such models may not be appropriate if measures of variables have censoring at the origin, outside the domain of the logarithmic function.

In this paper, we consider a numerical method for approximating logarithmic transformations when both the dependent variable and some of the regressors have censoring at the origin. This

* Corresponding author at: Department of Economics, University of Waterloo, Waterloo, ON N2L 3G1, Canada. Tel.: +1 5198884567.

E-mail addresses: t66chen@uwaterloo.ca (T. Chen), kenneth.couch@uconn.edu (K.A. Couch).

http://dx.doi.org/10.1016/j.stamet.2014.09.004
1572-3127/

type of problem is common in behavioral social sciences such as economics. A behavioral approach imposes structure on the underlying data generating process that leads to the censored variables. One common example we investigate in this paper is when a researcher assumes that due to mis-recording that is independent of the outcome of interest, all the negative data realizations are replaced by zero's. This assumption allows the researcher to proceed in taking the logarithm of relevant measures by removing all observations of zero. Another popular behavioral assumption we consider is the existence of an unobserved latent variable from a certain parametric family whose negative realizations are set to zero for an observed variable. For this second assumption, a maximum likelihood type of estimator can be justified. We demonstrate through simulations that the method proposed here reduces the extent of parameter bias from these two behavioral approaches when model misspecification occurs and there is substantial censoring (20% or more) of the data. A somewhat related paper by Rigobon and Stoker [4] gives some theoretical results to characterize the bias caused by censored regressors. They do not provide an algorithm to mitigate the bias and their dependent variable is not censored. If only the dependent variable is censored but not the regressors, and one is willing to assume the residual is conditionally symmetric, a seminal paper by Powell [3] solves this problem.

We consider the following model

$$\log(Y) = W'\alpha_0 + \beta_0 \log(X) + \epsilon, \tag{1.1}$$

where $\mathbb{E}[\epsilon|X, W] = 0$, $Y \geq 0$ and $X \geq 0$. We assume $\mathbb{P}\{Y = 0\} > 0$ and $\mathbb{P}\{X = 0\} > 0$ and that the researcher is interested in accurately estimating β_0. A prominent example of an equation that takes this structure is a commonly used expression to estimate the impact of schooling on earnings developed by Becker [1] and Mincer [2]. In that context, Y represents earnings and X measures years of schooling. In this case, it is observations of Y that may be censored. Another literature where such expressions can be found, for example Zimmerman [5], is in the measurement of intergenerational mobility, where Y represents earnings of sons and X represents earnings of fathers. Here, both observations of Y and X may be censored.

A common practice in many empirical studies is to simply throw away the observations if either X or Y takes the value 0 and then run the regression on the log-transformed values. A different approach is to assume the censored observations which are associated with a known parametric distribution such as the normal where the censored analytical expressions of the moments of the observations are available. Then maximum likelihood estimation can be used. However, one concern about this approach is the strong parametric assumption of normality.

Here, we compare these two procedures for dealing with censoring using simulation methods to the use of a Box–Cox transformation of the data while choosing the relevant tuning parameters to minimize a squared-loss function. We demonstrate through the simulation, that in models where α_0 is not of interest, that the method proposed in this paper reduces bias in the estimation of the slope parameter, β_0, relative to the alternatives considered when model misspecification and substantial censoring of the data are present, which is the problem the proposed method is intended to address.

2. A method of moments Box–Cox transformation

Suppose Z is a random variable with finite variance such that $Z \geq 0$ and $\mathbb{P}\{Z = 0\} > 0$. Let $Z^{(\lambda)} := \frac{Z^{\lambda}-1}{\lambda}$. If $Z > 0$, then it is well known that

$$\lim_{\lambda \to 0^+} Z^{(\lambda)} = \log(Z).$$

For $Z = 0$, we have

$$\lim_{\lambda \to 0^+} Z^{(\lambda)} = \lim_{\lambda \to 0^+} -\lambda^{-1} = -\infty.$$

Now define $\mathbb{Z}^{(\lambda)} = \exp(Z^{(\lambda)})$. Now irrespective of whether $Z > 0$ or $Z = 0$,

$$\lim_{\lambda \to 0^+} \mathbb{Z}^{(\lambda)} = \lim_{\lambda \to 0^+} \exp\left(Z^{(\lambda)}\right) = \exp\left(\lim_{\lambda \to 0^+} Z^{(\lambda)}\right) = Z.$$