# A potential solution to problems in ordered choice models involving endogenous ordinal variables for self-reported questions

HASAN, HAMID and Rehman, Attiqur

9 March 2013

# A potential solution to problems in ordered choice models involving endogenous ordinal variables for self-reported questions

Hamid Hasan[1] and Atiq-ur-Rehman[2]

## Abstract

Most of the surveys in social sciences generally consist of ordinal variables. Sometimes researchers need to model behaviour of ordinal variables in simultaneous equation system involving many endogenous ordinal variables. This situation leads to a very complex likelihood function which is extremely hard to solve. The solutions suggested in the literature are even harder to understand by applied researchers. The present study suggests a simulation method to avoid this problem altogether by converting ordinal variables into continuous variables and use standard simultaneous regression models. The proposed method involves generating random numbers from continuous probability distributions (uniform and truncated normal distributions) within a discrete probability distribution. This method can be fruitfully be used in ordered logit and probit models. The limitations of this method are also discussed.

*Keywords:* Endogenous Ordinal variables, Simultaneous Equation System, Ordered Logit, Ordered Probit.

*JEL Codes: C1, C2, C3*

[1] Assistant Professor, International Institute of Islamic Economics, International Islamic University, Islamabad, E-mail: hamidiiiephd@yahoo.co.uk

[2] Assistant Professor, International Institute of Islamic Economics, International Islamic University, Islamabad, E-mail: ateeqmzd@gmail.com Contact No:+92-51-9019439

## 1. Introduction

The present paper provides a discrete-to-continuous conversion method to deal with ordered choice models containing many endogenous ordinal variables. There is no solution exist for simultaneous equation model containing two or more ordinal endogenous variables. Ordered choice models have serious limitations in case of ordinal endogenous variables. Estimation of a simultaneous equation models in such a situation leads to a complex and a high dimensional joint probability distribution and hence the likelihood function (Heckman, 1977). Problems arise because these models assume latent continuous probability distribution underlying discrete manifestation of the distribution. More than two ordinal endogenous variables involve more than two such distributions which lead to intractable models. More often than not, such models do not converge. These types of models are particularly common in economics and other social sciences where ordinal survey data are ubiquitous. Researchers, in this situation, are compelled to use single equation ordered choice models with no ordinal endogenous variables. This would lead to naïve models not capitalizing full information.

Even these single-equation models have problems because of the restrictive assumptions like parallel regression lines, homoscedasticity, constant thresholds, and single-crossing property. Models can not be solved or have no unique solution unless identifying restrictions like zero intercept or zero first-threshold and given variance of underlying distribution are imposed. These restrictions result in rescaled coefficients. Models relaxing some of these assumptions become very complicated and not readily available for practitioners. Baltagi (2010) notes:

> "The estimation of ordered probit models with a panel structure, however, becomes exceedingly difficult with increasing number of categories of the dependent variable and increasing time periods."

## 2. Methodology

The above problems can be avoided altogether if we change discrete random variable into continuous random variable by a suitable method. The present study suggests a simulation method for converting ordinal variables into continuous variables by generating random numbers from continuous probability distributions within a discrete probability distribution. Ordered choice models assume that an unobserved (latent) continuous variable y* is underlying an observed discrete (ordinal) variable y. The y* is divided into various unknown thresholds within which discrete values fall. These thresholds are estimated along with other parameters.

For the sake of understanding, I take example of self-reported happiness or satisfaction. Ordinal choice models assume that self-reported satisfaction can take only integer values within an interval. If a satisfaction scale is from 1 to 4, then self-reported satisfaction level can take on discrete values 1, 2, 3, or 4, but this assumption is not feasible since a person choosing 1 might have a level around 1 but he is forced to choose exactly 1 since he has no other given option close to 1. This fact is also true for other categories. The chosen category, therefore, did not likely to reveal his true satisfaction level. Hence the self-reported satisfaction level is measured with error. Kuklys (2005, p.34) notes that:

> "[…] if we define the functioning to be measured as objectively achieved health status, subjective measures are likely to measure this functioning with error".

Self-reported satisfaction level ($y_R$) is different from true satisfaction level ($y_T$) and it is likely to be reported above, below, or at the level of true satisfaction because

each individual judge his/her own welfare with respect to a reference group and/or individuals understand the ordinal scales differently (anchoring problem), i.e.,

$$y_R > y_T \quad \text{or} \quad y_R < y_T \quad \text{or} \quad y_R = y_T$$
$$\Rightarrow \quad y_R = y_T + \varepsilon$$
$$y_T = y_R - \varepsilon$$
$$y_T - y_R = \varepsilon \qquad (1)$$

where $\varepsilon$ is assumed to be generated by a continuous probability distribution

If:

$$\varepsilon = 0$$

then this is the case of ordered choice models and

$$y_R = y_T \qquad (2)$$

Since $y_R$ is self-reported on an ordinal scale, $y_R$ is a discrete random variable taking integer values between α and β.

And if:

$$\varepsilon \neq 0$$

then $\varepsilon$ is assumed to follow a continuous distribution and $y_R$ becomes a continuous random variable taking any value between two cut-points α and β. i.e., $\alpha \leq y_R \leq \beta$

In what follows, we consider random number generation from uniform and truncated normal distributions.

(i) Random number generation from a uniform probability distribution:

$\varepsilon \sim U(\alpha, \beta)$ where $\varepsilon$ lies between a closed interval $[\alpha, \beta]$ consisting of all reals between $\alpha$ and $\beta$ : $[\alpha, \beta] = \{\varepsilon \in R : \alpha \leq \varepsilon \leq \beta\}$

In this case all outcomes (responses) are equally likely between α and β. We assume that the mean of the distribution is the self-reported response in case of ordinal data. i.e.,

$y_i^R = (\alpha_i + \beta_i)/2 \qquad \text{for i} = 1 \ldots \text{J} \quad \text{categories}$

(ii) Random number generation from a truncated normal probability distribution:

$\varepsilon \sim N(\mu, \sigma^2)$ has a truncated normal distribution if $\varepsilon$ lies between $\alpha$ and $\beta$ such that $(-\infty \leq \alpha < \varepsilon < \beta \leq \infty)$.

In this case outcomes (responses) close to mean are highly probable. Since mean is a location parameter, we assume that the mean of the distribution is the self-reported response for the ordinal data. i.e.,

$y_i^R = \mu_i \qquad \text{for i} = 1 \ldots \text{J} \quad \text{categories}$

The cut points (α and β) can be determined in one of the following two methods:

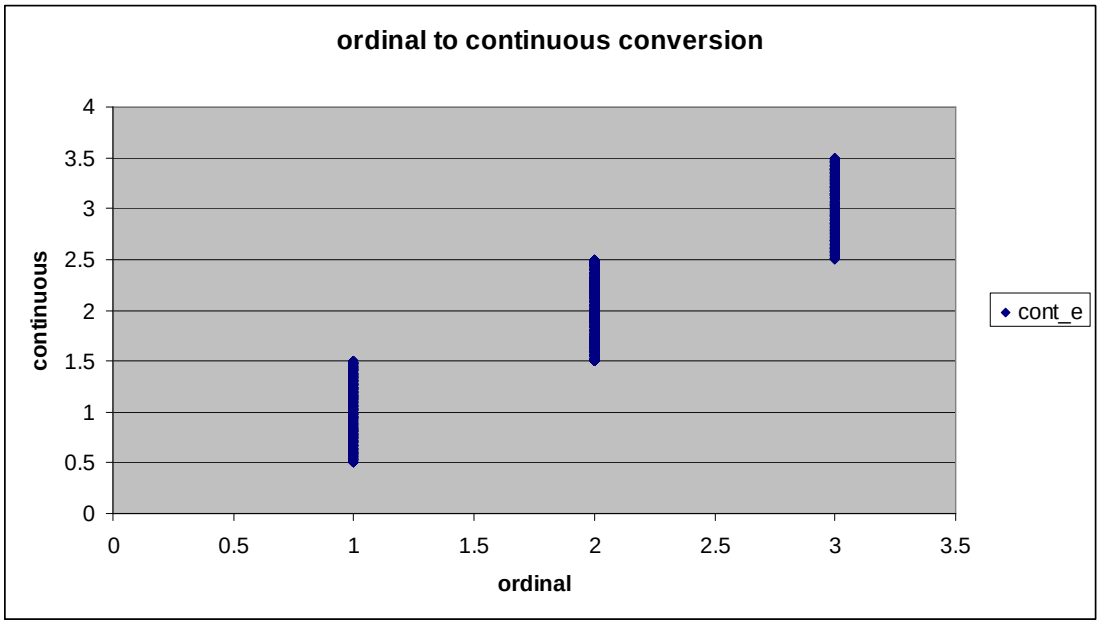
1) Arbitrary cut points are generated assuming equal distances between categories.
2) Data-based cut-points: using the data distribution around each category, the cut points are determined by inverse cumulative distribution function.

The cut-points are useful for our purpose if the tails of distributions (histograms) of a converted variable are connected or overlapped so that data become continuous and the distance between cut-points are equal. Since if cut-points are equally apart then OLS is roughly corresponds to ordered choice models (Long and Freese, 2006).

In case of data-based cut points, the tails of histograms are not connected because of the narrow range of cut points around a category whereas in case of arbitrary cut-points, the tails are connected. The distance between categories can be unequal for the data-based cut points but not for the arbitrary cut-points. Therefore, I will use arbitrary cut-points in ensuing analysis since they satisfy the required conditions.

The following graph shows conversion of an ordinal variable on x-axis into continuous variable on y-axis by a continuous probability distribution. Ordinal variable can take integer values between 1 and 3, that is, 1, 2, and 3, and continuous variable, on the other hand, can take any value between 0.5 and 3.5.

Figure 1




## 2. Comparison of histograms: ordinal versus uniform and truncated normal

The first histogram shows self-reported categories in actual data which are discrete. The second and third histograms show continuous versions of the first histogram. These are drawn by generating random numbers around each category from uniform and truncated normal distributions respectively by the first method. The last histogram is drawn by generating truncated normal distributions by the second method.

Figure 2 Histogram of an ordinal variable

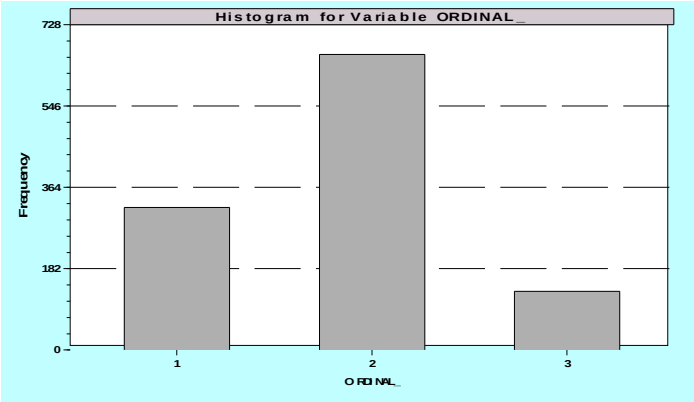



Figure 3 Histogram of a uniform continuous variable around ordinal categories (method 1)

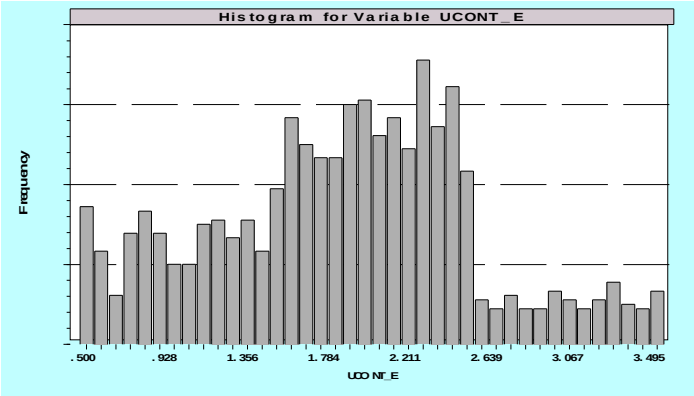



Figure 4 Histogram of a truncated normal variable around ordinal categories (method 1)

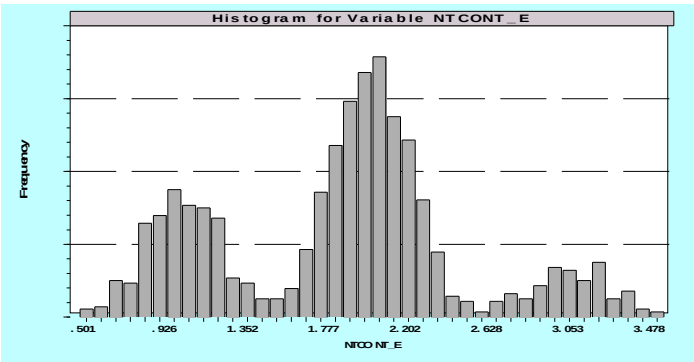
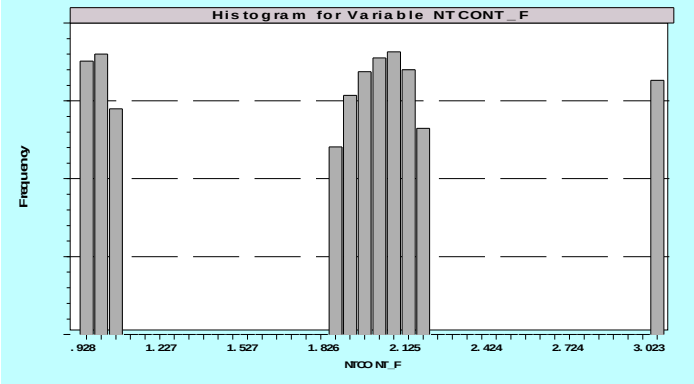


Figure 5 Histogram of a truncated normal variable around ordinal categories (method 2)




**3. Behaviour of ordinal versus continuous variables across individuals:**

The following graphs show behaviour of ordinal variable across individuals and its continuous counterpart generated by uniform random numbers and truncated normal random numbers respectively. The last graph shows behaviour across individuals when second method is used.

Figure 6

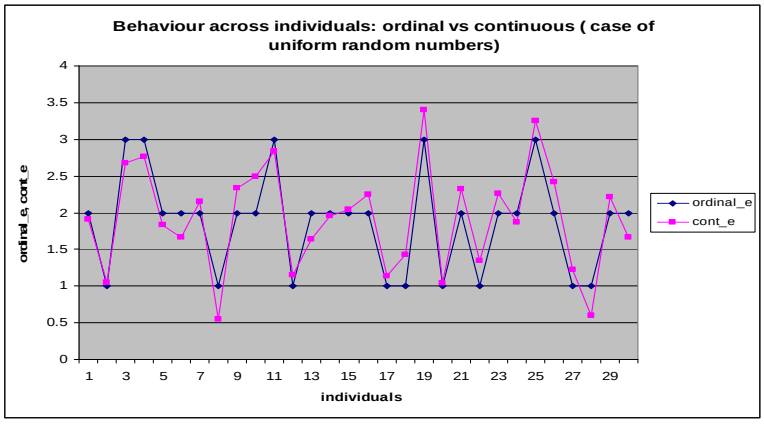


Figure 7

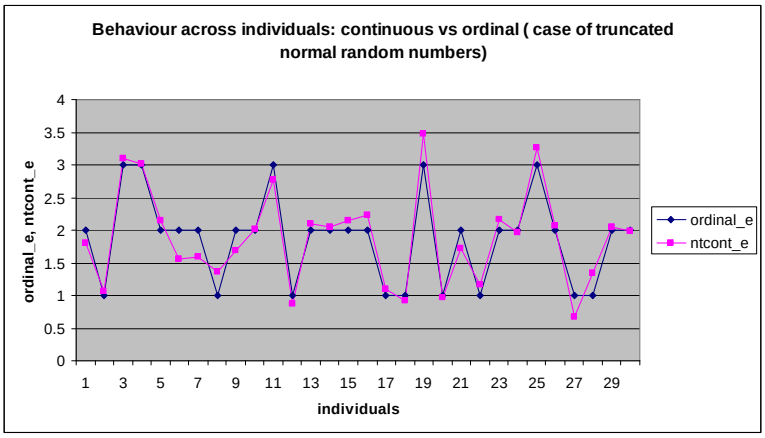
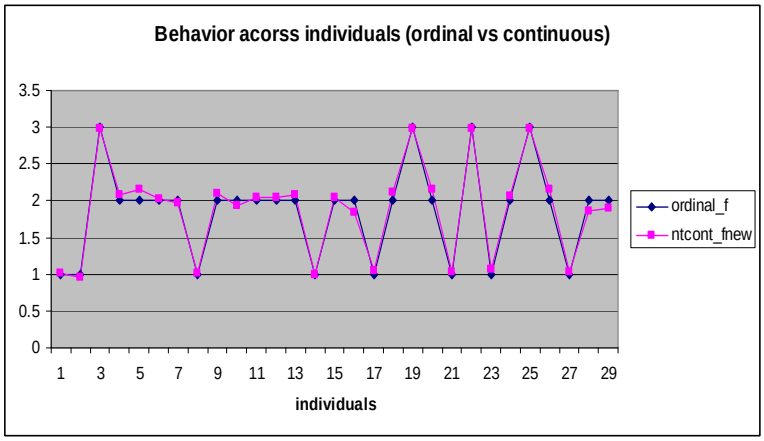



Figure 8




## 4. The Pearson and Polyserial correlations between ordinal and continuous variables

The Pearson coefficient of linear correlation between ordinal variable and its uniform counterpart is 0.901 whereas between ordinal variable and its truncated normal counterpart is 0.951. The Polyserial correlation (a correlation between ordinal and continuous variables) between ordinal and uniform as well as between ordinal and truncated normal is 0.999.

## 5. Predictions of ordinal variable from continuous variables

The following tables show prediction of ordinal variable from uniformly generated random variable and from truncated normally generated random variable by ordered Probit models. The both tables show very high level of predictions.

Table 1

Predictions for ordinal variable from uniformly distributed variable

| Cross tabulation of predictions. Row is actual, column is predicted. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model = Probit . Prediction is number of the most probable cell. | | | | | | | | | | |
| Actual | Row Sum | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 319 | 319 | 0 | 0 | | | | | | | |
| 1 | 661 | 0 | 661 | 0 | | | | | | | |
| 2 | 131 | 0 | 0 | 131 | | | | | | | |
| Col Sum | 1111 | 319 | 661 | 131 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2

Predictions for ordinal variable from truncated normally distributed variable

| Cross tabulation of predictions. Row is actual, column is predicted. | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model = Probit . Prediction is number of the most probable cell. | | | | | | | | | | |
| Actual | Row Sum | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 319 | 319 | 0 | 0 | | | | | | | |
| 1 | 661 | 0 | 661 | 0 | | | | | | | |
| 2 | 131 | 0 | 0 | 131 | | | | | | | |
| Col Sum | 1111 | 319 | 661 | 131 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## 6. The unbaisedness and consistency

Finally, we analyse the consistency and unbiasedness of the method introduced above. The analytical computations are very difficult because of very complex likelihood; therefore we design a simulation experiment to see if the method introduced has the two desirable properties. The simulation design is as follows:

6.1 Data Generating Process

Taking the data generating process $y = \alpha + \beta x + \varepsilon$ where $x$ is ordinal variable between taking values 1,2,3 and 4. Y is the continuous random variable which depends on the ordinal variable $x$ and $\alpha, \beta$ are the parameters. We set $\alpha = \beta = 1$ and generate $x$ using discrete uniform distribution and $\varepsilon$ is standard normal random variable.

6.2 Estimation

For estimation we convert the ordinal variable $x$ to continuous variable $x^*$ by generating continuous random variable centred at $x$. The variable $y$ was than regressed on $x$ using OLS and the difference between parameter and its estimate i.e. $d = \hat{\beta} - \beta$ was calculated.

The process was repeated 100,000 times for various sample sizes. The results are reported below:

Table 2: Properties of Estimates after converting

| Parameter $\beta$ | Sample size | Average of estimates | Average of the difference d | Standard deviation of d |
|---|---|---|---|---|
| 1 | 40 | 0.994 | 0.006 | 0.231496 |
| 1 | 100 | 0.996 | 0.004 | 0.117411 |
| 1 | 200 | 1.002 | -0.002 | 0.079831 |
| 1 | 400 | 0.999 | 0.001 | 0.054283 |

It can be observed from the table that the estimator is unbiased since the average of estimates is equal to the true value of the parameter. Furthermore, the Monte Carlo

standard deviation of estimates decreases with the sample size which indicates that the estimates are consistent and converge to the true value as the sample size increase. The experiment was repeated for multiple endogeneous case and the distribution of estimates revealed same pattern.

## 7. Conclusion

The above results show that continuous data can reliably be used instead of discrete data since most of the econometric and statistical methods are best suited for continuous (interval or ratio levels) data. Moreover, the statistical measures like means, variances, covariances, and correlations are meaningful for continuous data. Another advantage of using converted data is that data become independently and identically distributed. This facilitates the application of central limit theorem and hence valid inference.

There is at least one limitation of the proposed method. The converted variable does not extend from minus infinity to plus infinity and hence does not have any relationship with an underlying latent probability distribution usually assumed in ordered choice models though that assumption creates single-crossing property. This and other potential limitations can be removed if we can tackle this problem using Gibbs sampling technique. But this technique is not user-friendly and applied researchers generally have difficulty using it.

## References

Baltaqi, B. H. (2010). Econometrics. Springer.

Heckman, James. J. (19**77)**. Dummy Endogenous Variables in a Simultaneous Equation System, Econometrica, Vol. 46, No. 4, pp. 931-959.

Kuklys, W., 2005. Amartya Sen's Capability Approach – Theoretical Insights and Empirical Applications. Springer, Berlin et al.

Long, J. Scott and Jeremy Freese. (2006). *Regression Models for Categorical Dependent Variables Using Stata, Second Edition.* College Station, Texas: Stata Press.