Measurement Error Bias from Social Network Data used in Discrete Choice Models

Michael Maness
Graduate Research Fellow
University of Maryland
Department of Civil and Environmental Engineering
1173 Glenn Martin Hall
College Park, MD 20742
Phone: 301-405-6864
Fax: 301-405-2585
Email: mmaness@umd.edu

Over the past two decades, travel behavior analysis has begun shifting focus from the individual to the social. Accordingly, discrete choice models, a common modeling technique in travel behavior analysis, have begun to integrate social context into its framework (Dugundji and Walker 2005, Paez and Scott 2007). Social influence, defined as the tendency to engage in behavior similar to others, has been one approach to this integration. Thus far these efforts have resulted in work that showing that social influences may be relevant in travel decision making. But, these models are hampered by strong assumptions on social network structure, limited availability of data, and difficulties in model applicability. This project has theoretical aims to improve the understanding of the endogeneity problem in discrete choice models of social influence caused by measurement error and missing data in social network data collection. The practical aims are to provide practical guidance for practitioners in the use of social influence models. This doctoral dissertation research seeks to answer the question of: What bias does measurement errors in the specification of social networks induce in parameter estimation for social influence models of discrete choice? This paper expands on work by Paez et al. (2008) from analyzing measurement errors in regression models of social influence to discrete choice models of social influence.

Currently results are preliminary as this is ongoing dissertation work. For this extended abstract, the methodology and results section will be combined to provide a walk-through of a particular run of the experiment. We begin with a population of 256 agents with individual specific characteristics distributed normally with mean -1 and standard deviation 2. Then an initial choice was simulated using a logit model and the individual specific characteristic parameter equal to 1.2. For this run, this resulted in 173 individuals choosing not to bike.

These agents are socially connected to one another through a social network from the Bernoulli graph distribution. In a Bernoulli graph distribution, the probability of a link between two individuals has a set probability (in this case, a 5% chance) and it is independent from the presence of other links in the network. For this run, a draw from this graph distribution was taken and set as the social network for the individuals in this run. With a social network and individual

specific characteristics, choice decisions were simulated using equation (1) and a social influence parameter equal to 5.0.

$$U_n(+1) - U_n(-1) = \alpha + \beta x_n + \delta \bar{y}_n + \varepsilon_i \qquad (1)$$

This resulted in 216 individuals now choosing not to bike. This social network and choice configuration is used as the base or "true"data.

The next step is to generate graphs (i.e. social networks with the same individuals, but different links) similar to the base data. For this run, we used other graphs from the same Bernoulli graph distribution and generated 4000 graphs. We then estimated binary logit models on these new choice and social network configurations. Figure 1, 2, and 3 show the distribution of the constant parameter, individual specific characteristics parameter, and social influence parameter estimates, respectively.

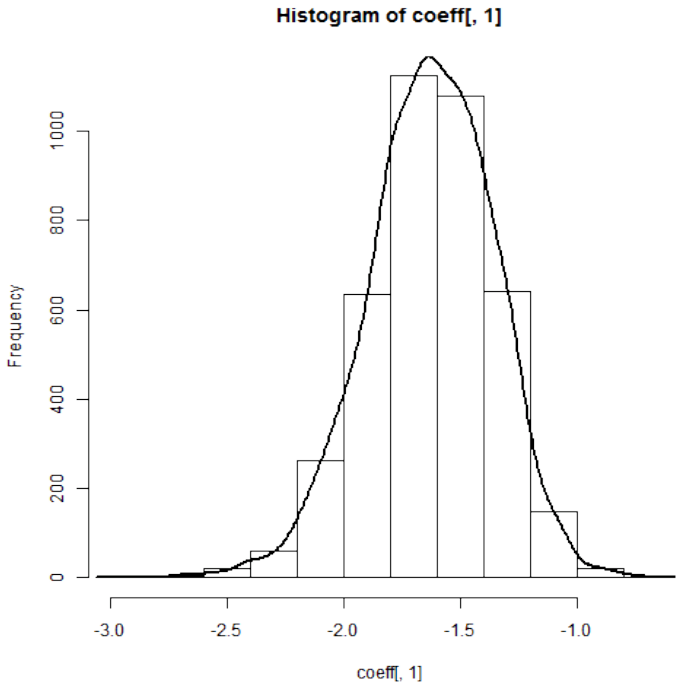

Figure 1. Constant Parameter Distribution

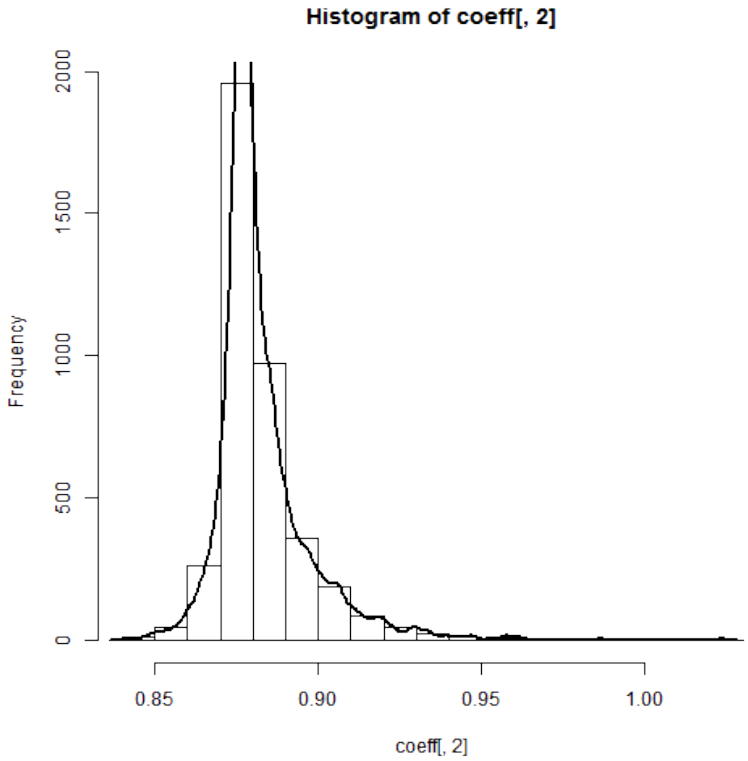

Figure 2. Individual specific characteristic parameter distribution

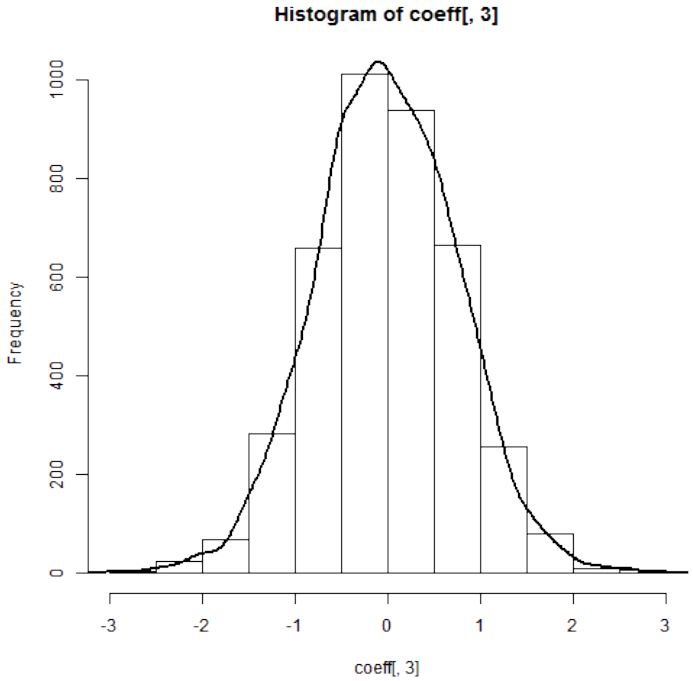

Figure 3. Social influence parameter distribution

The individual specific characteristic parameter is underestimated with mean of 0.883 and a range of 0.840 to 1.020. The social influence parameter is greatly underestimated and even has an incorrect sign an approximate half of the model estimations. These parameter estimates have a mean is -0.011 in range between 3.000 and 2.840. These preliminary results show that (at least for this run) measurement errors on the order of only knowing the likely distribution of social networks may lead to underestimation and even issues with the direction of influence. The constant term compensates for this uncertainty in graph structure and the multitude of paths for the diffusion of influence. This work will be expanded upon by choosing graphs that are closer to the true graph and not just in the same distribution.

The expected results will generalize this to include other distributions of individual specific characteristics as well as other graph distributions. More sparse and more dense Bernoulli graphs will be generated as well as Markov graphs, which are networks that include triangle and star configurations. Markov graphs are more realistic representations of true social networks than Bernoulli graphs due to the possibility of transitivity and varying popularity of individuals. The final paper will have an extensive discussion of the patterns between different graph structures and specific guidelines on the level of bias generated.

References

Dugundji, E. R., & Walker, J. L. (2005). Discrete choice with social and spatial network interdependencies: an empirical example using mixed generalized extreme value models with field and panel effects. *Transportation Research Record: Journal of the Transportation Research Board*, 1921, 70-78.

Páez, A., & Scott, D. M. (2007). Social influence on travel behaviour: a simulation example of the decision to telecommute. Environment and Planning A, 39(3), 647-665.

Páez, A., Scott, D. M., & Volz, E. (2008). Weight matrices for social influence analysis: An investigation of measurement errors and their effect on model identification and estimation quality. Social Networks, 30(4), 309-317.