A MONTE CARLO STUDY OF THE EFFECT OF SAMPLE BIAS ON THE MULTINOMIAL LOGIT CHOICE MODEL COEFFICIENTS.

A Thesis presented in partial fulfilment
of the requirements of the
degree of Master of Business Studies at Massey University.

Grant K. Bell

1996

ABSTRACT

This thesis reports the findings of a Monte Carlo simulation into the effect of sample bias on the parameters of the multinomial logit (MNL) choice model. At issue is the generalisability of parameter estimates obtained from biased samples to the balance of the population. An actual data set of 164 respondents was used to estimate an aggregate model. Using these parameters as the true coefficients of choice behaviour, an unbiased sampling distribution of the MNL parameters was derived by repeatedly fitting aggregate models to artificially generated sets of individual responses. Subsequently, the biased sampling distribution was derived by selectively eliminating those individuals at the tails of the sample distribution based on their correlation with one of the independent variables.

The expected values of the biased and unbiased sampling distributions were compared to assess the sensitivity of the model to sample bias. The research found the biased coefficients changed by significantly more than the proportion of individuals removed. However, this sensitivity was predictable as the percentage change in the value of the coefficients was related to the size of the coefficient. It was also found that the coefficients of the unbiased variables were not significantly influenced by bias on another variable. The ratio between the unbiased variables was also maintained. It was concluded that although sensitive to bias, the estimates produced by the MNL model could be modified to reflect the different effect of the bias on the coefficients. Additionally, there was no evidence to suggest that the MNL estimates were not reflecting the effects of interest when calibrated on unbiased samples.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisors, Michael Fox and Malcolm Wright, for their insight, professionalism, and patience.

The support, financial and otherwise, provided by the Marketing Department was also greatly appreciated.

Special thanks to my family whose support made this possible.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1.0 INTRODUCTION

The multinomial logit (MNL) model is used by marketers to predict consumer product or brand choice behaviour as a function of that brand's or product's attributes and the characteristics of the consumer. The MNL is a random utility model which assumes that choice consists of both a systematic (or explainable) component and a random component. It is widely used in industry, to guide strategy, and in academia, where it furnishes results which test specific knowledge claims about consumer behaviour.

However, a number of potential sources of error exist in the estimation of the coefficients of the model. These include specification error, breaches of the assumptions underlying the model (namely, independence from irrelevant alternatives, IIA, and independently and identically distributed errors, IID), measurement error, aggregation bias, random sampling error, and systematic sampling error (or bias). The first five of these have already been investigated by a number of authors including Gordon, Lin, Osberg, and Phipps (1994), Ben-Akiva and Lerman (1989), Jones and Landwehr (1988), Batsell and Polking (1985), Lee (1982), Horowitz (1981), and Chamberlain (1980). Systematic sampling error or bias has received the least attention. This is surprising given the fact that the MNL is nearly always calibrated on samples drawn from the population of interest.

In marketing, three of the more common sources of data used to estimate the MNL include retail scanner data, consumer panels, and experimental data. This data either describes actual market behaviour with associated product attributes and consumer characteristics or hypothetical choice behaviour with the alternative attributes and actual consumer characteristics of a selected sample of individuals.

The main repercussion of calibrating the MNL on samples is the requirement that at the very least the model functions as expected in conditions of sampling error. It would appear that research

specifically directed at examining the stability or otherwise of the MNL model in situations where sample bias is prevalent has not been attempted. This study investigates the behaviour of the MNL coefficients when estimated on biased samples. In particular, the Monte Carlo method is employed to generate biased sampling distributions which are compared with a benchmark unbiased sampling distribution. The bias is simulated by removing individuals who are most (or least) highly correlated with one particular independent variable.

Three main consequences of this simulated sample bias are explored. The effect on the biased variable is examined to determine if the coefficients vary by the same proportion as the bias. Ideally, the MNL's biased coefficients would change by a proportionately less amount. This would mean that the MNL is producing estimates that reflect the effect being modelled more than the sampling error. If the coefficients change by proportionately more than the bias, then the stability of the model would be questioned.

Another effect of interest is the change in the unbiased variables caused by the simulated bias. It would be desirable (and expected) for the unbiased variables to remain unchanged. As the Monte Carlo method used here generates individuals with no interactions between the independent variables, we would not expect any significant change to occur. However, if this does occur, then it would be a major handicap to the MNL.

Finally, any change in the coefficients of the unbiased variables should not significantly impact on the ratio between them. If this does occur, then the difference in the coefficient sizes are both a reflection of the effect and the error. The MNL would therefore be unjustifiably sensitive to bias.

If the MNL is found to be sensitive to sample bias, then not only is the collection of an unbiased sample important, but the assumptions underlying the model may be dubious.