Robustness of Latent Class Measurement Error Models

Brian Meekins, Daniell Toth
Bureau of Labor Statistics

Abstract

The technique of latent class analysis relies on a number of model assumptions which might be violated by the underlying process being investigated. This study is to determine the reliability of the analysis done on four stage Markov Latent Class models containing the classification of individuals in one of two indicator categories. The estimation is done using the EM algorithm on simulated data under specified model assumptions where those assumptions are violated to varying degrees.

KEY WORDS: Markov Latent Class, Measurement Error, Simulation

1. Introduction

In recent research, Tucker et al (2003) employed Markov Latent Class Analysis (MLCA) in order to estimate the extent of underreporting on the Consumer Expenditure Interview Survey (CEIS). Using second order Markov Models the authors used a single indicator at four different waves of an interview to estimate a complimentary latent or true variable for the same waves. The indicator was the report of a purchase in that interview wave, while the latent variable was hypothesized to be the "true" or actual purchase of a given commodity. Many different commodities were evaluated and results indicated that routine monthly purchases such as utilities had a much higher accuracy rate (probability that purchase was reported given an actual purchase) than commodities such as shoes or household furnishings, that are purchased less routinely. While these results are reassuring because they are in concert with the author's intuition, the models, in order to be estimated, make a number of assumptions that are likely violated. The purpose of this work is to use simulated data to evaluate the impact of violating two key assumptions of the model used in Tucker et al.

2. The Consumer Expenditure Survey

The CEIS is a rotating panel survey, where consumer units (CUs, usually households) are interviewed in five separate quarters about there purchases in the previous quarter. Data regarding purchases in the first interview are typically not used in analysis as this interview is treated as a "bounding" interview to avoid telescoping. Consistent with this practice Tucker et al. estimated models using the last four waves of the survey. The interview iteself is quite long and burdensome so that reporting error is likely. Approximately 8,000 CUs are interviewed annually (Tucker et al. combined six years of data).

3. Second Order Markov Model and Model Assumptions

Let $\{L_i\}_{i=1}^{n}$ be a set of independent random vectors

$$L_i = [w_i, x_i, y_i, z_i]$$

satisfying

Equation 1

$$P(x_i = j) \,|\, (w_i = k) = P(y_i = j \,|\, x_i = k) = (z_i = j \,|\, y_i = k) = p_{jk}$$

For $j,k \in \{1,2\}$ with initial probabilities

$$P(w_i = 1) = p_1 \quad \text{and} \quad P(w_i = 2) = p_2$$

Let

Equation 2

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}$$

and

Equation 3

$$\mathbf{P}_0 = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$$

The probabilities in (equation 2) and (equation 3) define the dynamics of the underlying latent variables as they propagate through time. We assume independent measurement (and/or reporting) error in each of the manifest variables, given by the components of,

$$M_i = [a_i, b_i, c_i, d_i],$$

corresponding to an observation of each component of L_i. These errors are defined by

Equation 4

$$\begin{aligned} P(a_i = j) \mid (w_i = k) &= P(b_i = j) \mid (x_i = k) \\ &= P(c_i = j) \mid (y_i = k) \\ &= P(d_i = j) \mid (z_i = k) = q_{jk} \end{aligned}$$

Let

Equation 5

$$\mathbf{Q} = \begin{pmatrix} q11 & q12 \\ q21 & q22 \end{pmatrix}$$

3.1 Likelihood Function

Equation 6

$$\begin{aligned} &L(\theta; M) = P(a, b, c, d \mid \theta) \\ &= P_\theta(a \mid w) P_\theta(b \mid x) P_\theta(c \mid y) P_\theta(d \mid z) P_\theta(w, x, y, z) \end{aligned}$$

where $P_\theta(\cdot) = P(\cdot \mid \theta)$ and $M = (M_1, \ldots, M_n)^T$

Equation 7

$$\begin{aligned} P_\theta(a \mid w) &= P_\theta(a \mid w = 0) + P_\theta(a \mid w = 1) \\ &= (q_{a.10} + q_{a.11})^a (1 - (q_{a.10} + q_{a.11}))^{1-a} \end{aligned}$$

The probabilities of b, c, and d are similar. Next we find

Equation 8

$$P_\theta(w, x, y, z) = P_\theta(w) P_\theta(x \mid w) P_\theta(y \mid w, x) P_\theta(z \mid x, y)$$

In addition, to those just described, two assumptions were made in order to estimate the second order Markov model in Tucker et al. First it was assumed that the relationship of the indicator to the latent variable was stationary or did not change over time, that is:

Test Assumption 1

$$P(a \mid w) = P(b \mid x) = P(c \mid y) = P(d \mid z)$$

This assumption is likely to be violated. Analysis of patterns of expenditure reports over the four panel waves show differences in the total amount of expenditure reported, especially in the third quarter where expenditure reports are typically low. However, a second order Markov model with one indicator per latent construct cannot be estimated without assuming some stationarity. The second assumption made by Tucker et al. to be evaluated in this work is the assumption that no false positive reports of expenditures are given by the respondent or,

Test Assumption 2

$$\begin{aligned} P(a = 1 \mid w = 0) &= P(b = 1 \mid x = 0) \\ &= P(c = 1 \mid y = 0) \\ &= P(d = 1 \mid z = 0) = 0 \end{aligned}$$

While it is certainly uncommon for CUs to report purchases they did not make, it probably does happen. For Tucker et al. fixing the probability of false positives to zero greatly simplified the estimation task by creating "partially latent" variables - the advantages gained outweighed the possibility of bias in the estimates.

4. Design of Simulations

The current work uses simulated data to examine the assumptions made by Tucker et al. However, in order to make the results more general, a second order Markov latent class model with a single indicator of a latent construct at each of four time points without covariates is estimated. This can be viewed as a simplified version of the model used in Tucker et al.

The data on which this model is estimated is varied in three ways. Firstly, the sample size is varied (n = 250; 500; 1,000; 2,000; 4,000). Sample sizes in Tucker et al. were on the order of 3,000 to 5,000. Initially, the authors of this work believed that in order to observe large biases in estimates due to violations of the assumptions, smaller sample sizes would be needed. This proved not to be the case.

Secondly, data is generated that violates the stationarity assumption of the model. The third quarter reporting error was increased in proportion to the reporting error of the other three quarters. The accuracy rate, $P(a=1|w=1)$ for the first, second, and fourth quarters is set to 0.75 (based on work by Tucker et al.), the third quarter probability is adjusted such that $P(c=1|y=1) = .75a$ where a=(0.8, 0.85, 0.9, 0.95, 1.0). Therefore, the measurement error for quarters one, two, and four id defined as:

$P(a=0|w=1) = P(b=0|x=1) = P(d=0|z=1) = 0.25$

while the measurement error for quarter three varies:

$P(c=0|y=1) = \{0.25, 0.2875, 0.325, 0.3625, 0.4\}$

Finally, data are generated that violate the model's false positive assumption. The degree that the data violate this assumption is based on the probability of a report being given when no purchase was actually made *P(a=1|w=0)*. This probability, which we call 1-*q* was varied from 0 to .2, such that q={0.8, 0.85, 0.9, 0.95, 1.0). Therefore, the data varied from 20 percent of positive reports being false (very unlikely in reality) to no positive reports being false.

The model is estimating using ℓ_{EM} under *R*2.4.1. Consistent with Tucker et al. starting values were supplied for both the transition probabilities and the measurement error probabilities to aid with convergence. These starting values were either 0.6 or 0.4 depending on if the value in the data was above or below 0.5. The convergence criteria was set to $1x10^{-6}$.

For each of the three conditions that are varied (n, a, and q), 1,000 iterations were conducted where simulated data were generated. However, in order to obtain more stable estimates (and avoid the some of the pitfalls of local maxima) each of the 1,000 iterations was estimated ten times and the best fitting model, as determined by the value of the BIC L-square was selected. Thus, a total of 1,250,000 estimates were produced, 125,000 iterations were conducted, for 125 different combinations of a, q, and n.

A number of measures were recorded, including fit statistics (both the BIC and AIC based on L-square or Loglikelihood and the dissimilarity index) and distance measures derived from parameters set in the simulated data compared to those that were estimated (estimated – true). Both the square root of the squared differences between the estimated and true and the unsquared (signed) distance measures were used in analysis. In addition, these distances were calculated for the entire model combined, where all estimated probabilities versus their corresponding true values were calculated and combined into a single measure. and the distance of the probability of most interest *P(a=1|w=1)* from the estimated to the true.

5. Results

For the sake of brevity the analysis of the data is confined to the bias measures. Future research will examine model selection procedures and will therefore examine the model fit statistics more rigorously. In addition, we found that measures that estimate the amount of bias in all model parameters combined tends to obscure the bias in the measure of interest the bias in the reporting accuracy, or it's compliment reporting error. In addition, bias measures based on the squared deviations of the estimated minus the true, while helpful, will be ignored in the results, because these too obscures the underlying distribution of the bias. Results will then focus on the bias in *P(a=1|w=1) = q.11*.

The authors quickly realized that the variance in the bias was larger than expected for n <= 1,000. Figure 1 shows the distribution of the bias of *q.11* (or *P(a=1|w=1)*) by a and q for n = 1,000. Note that for very large departures from model assumptions the variance of the distribution of the bias is quite large. Because the value of the bias is bounded at a maximum of 0.25 (the true value .75 – the maximum probability of 1.0), we see that the mode of the histogram is located at this boundary. Even with data where no violations of the assumptions are made (a=1, q=1) we can note a relatively large number of points at this boundary level. Because of this problem we choose to conduct most of our analysis where n >=2,000.

Figure 2 presents the same histogram for n=4,000. For mild departures in a, and even severe departures in q, the distribution of the bias appears to be somewhat normal, but positively skewed. Where the data does not depart from the model assumptions at all we can still discern the positive skewness of this distribution. The authors were somewhat surprised at this result, expecting a normally distributed, symmetrical result. Figure 3, presents the histogram for the bias in *q.11* for n=40,000 and no departures of the data from model assumptions. Note that, while the variance is quite small and the skewness is greatly reduced, the positive bias remains and is actually more pronounced. The mean of the bias is 0.041.

Figure 4 shows the mean of the bias by *a* for each level of n (each plot is a particular level of n with the bottom left n=250 and top right n=4,000). While the mean bias does not vary significantly by a for low n, for higher values of n, the bias decreases sharply over *a*, only to increase where a is close to one (where there is no difference in the third quarter reporting error. Figure 5 shows a relatively steady decrease in the variance of the bias as *a* increases.

Firgure 6 shows the mean of the bias by *q* for each level of n. There are relatively modest decreases in the mean bias by *q* indicating that violations in *a* are perhaps more serious. Figure 7 shows the variance of the bias by *q* for each level of n. Note the slight but steady improvement in the variance of the bias.

In order to explore the relationship of *a* and *q* further, the mean of the bias and the mean of the variance were calculated for each combination or *a*, *q*, and *n*. Figure 8 plots the mean of the mean bias by *n* for all combinations of *a* and *q*. Likewise, Figure 9 plots the mean of the variance of the bias by *n* for all combinations of *a* and *q*. Note the rapid decrease in the variance of the bias over *n* regardless of *a* and *q*. For mean bias, the most striking result is the increase in bias with *n* as *a* and *q* approach 1.

5. Conlusion

From our findings we conclude the following. Estimation of simple second order Markov models require a sample greater than or equal to 2,000. Sample sizes smaller than 2,000 appear to be sensitive to even the smallest violations in model assumptions, where the estimate of measurement error can be hardly assumed to be unbiased.

In addition, it appears that even with no violations in model assumptions the estimate for the measurement error in second order Markov latent class models will be biased. This is true, even for sample sizes larger than 40,000. Therefore, the estimate produced for the reporting error appears to be neither unbiased nor consistent.

Simple second order Markov models appear to be robust to violations in the false positive rate. Even when 20 percent of the positive reports were false, the mean bias and variance are not significantly larger than those when the assumption is not violated.

The models are sensitive to violations in the stationarity of measurement error assumption. Even modest departures in the reporting error for the third quarter, as compared to other quarters, led to large increases in both the variance and mean of the bias. However, the exact nature of this relationship is difficult to obtain because of the result that bias exists for no departures from the model.

References

Tucker, Clyde, Paul Biemer, and Brian Meekins. 2003. "Latent Class Models and Estimating of Errors in Consumer Expenditure Reports." *Proceedings of the 2003 Joint Statistical Meetings*. San Francisco, CA. ASA.

Biemer, Paul. 2004. "An analysis of classification error for the revised current population survey employment questions." *Survey Methodology*. 30(22):127-140.

Biemer, Paul and Bushery, J.M. 2000. "On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data." *Survey Methodology*.26(2):139-152.

Vermunt, Joeren. 1996. *Log-linear Event History Analysis*. Tilberg University Press. The Netherlands.

Tables and Figures

Figure 1

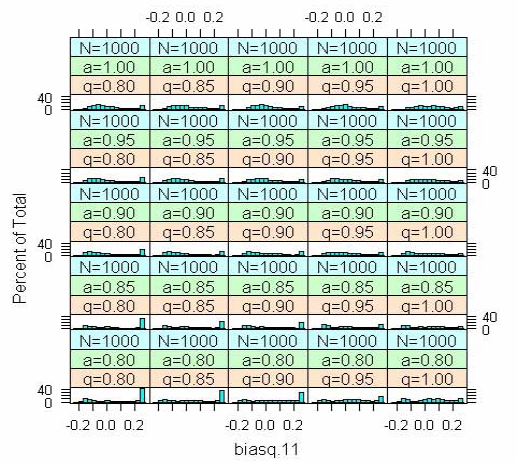


Figure 2

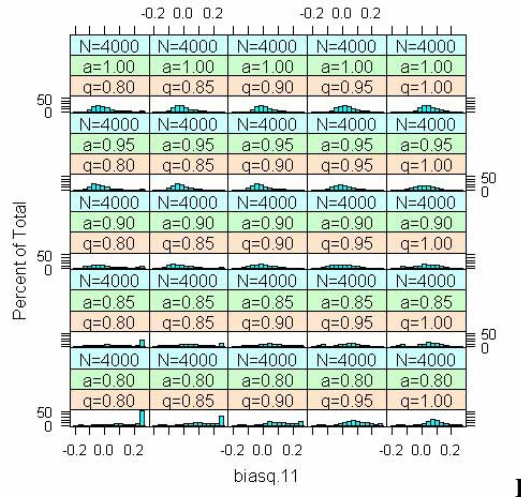


Figure 3

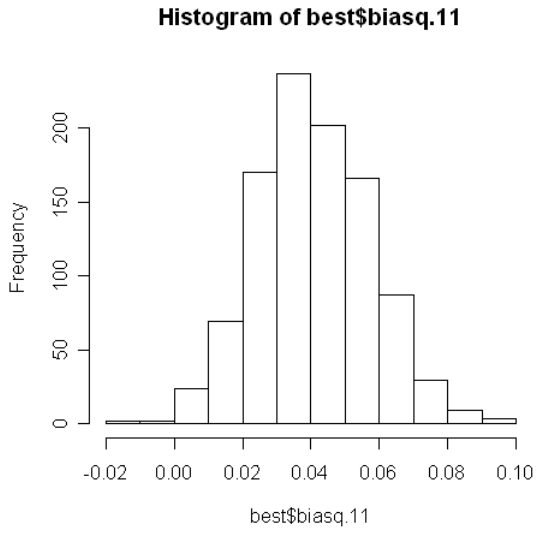


Figure 4

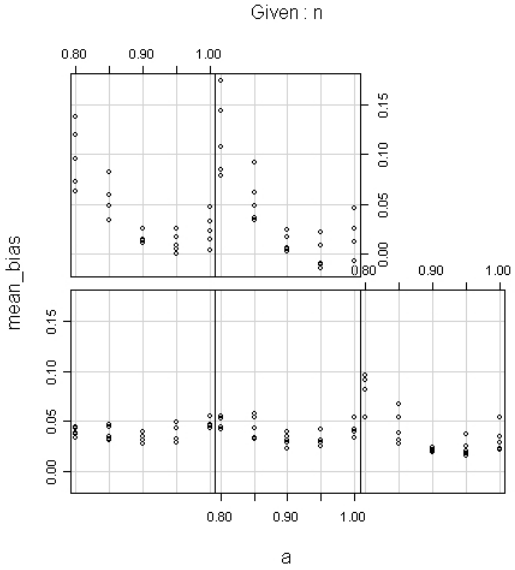


Figure 5

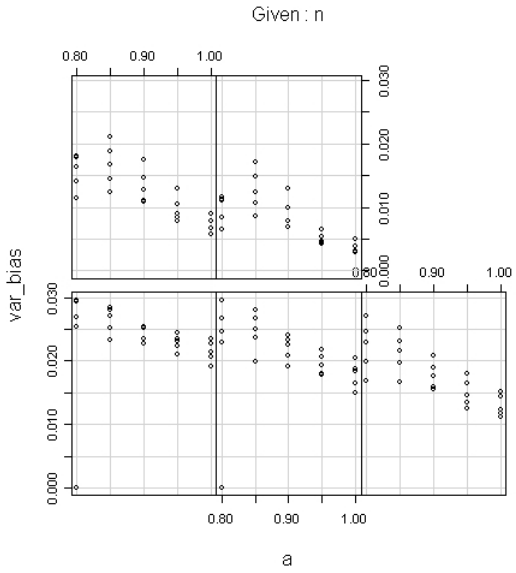


Figure 6

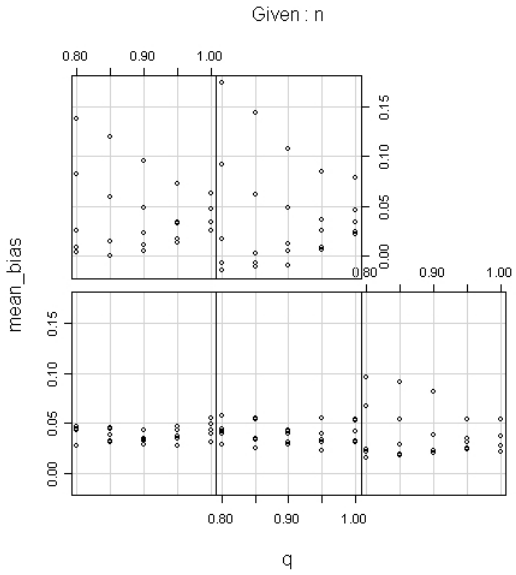


Figure 7

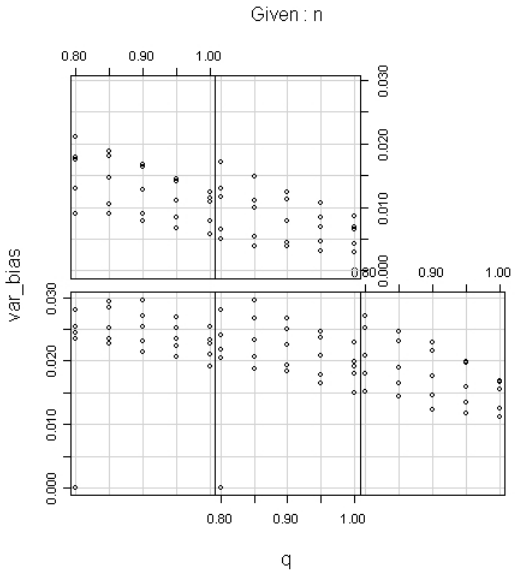


Figure 8

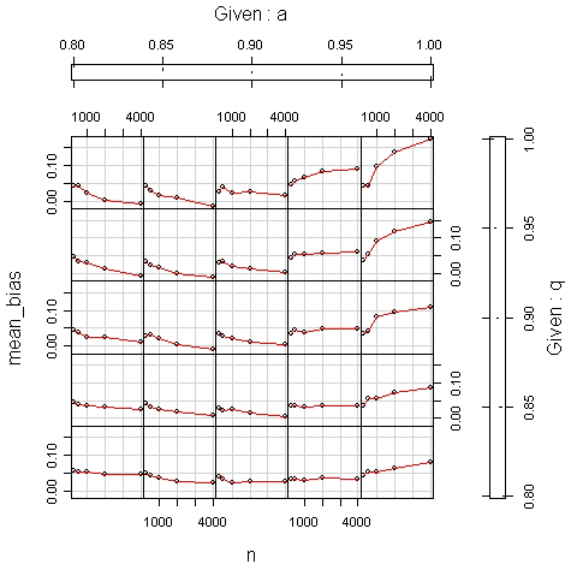
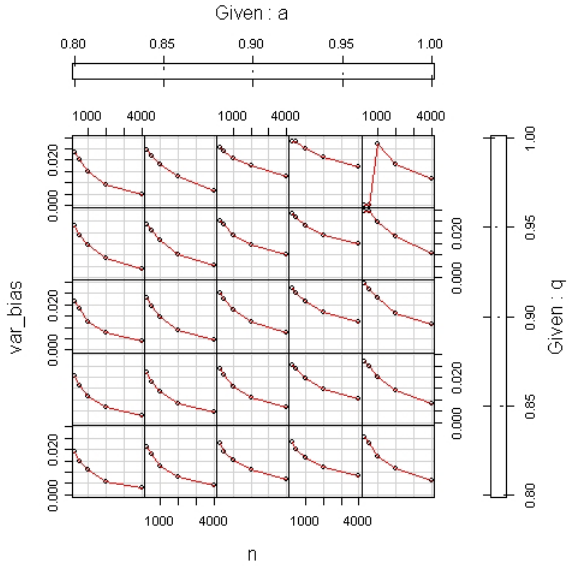


Figure 9


Given : a
0.80
0.85
0.90
0.95
1.00
1000
4000
0.000
0.020
var_bias
n
Given : q