DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 7

5-1-2010

Assessing Classification Bias in Latent Class Analysis: Comparing Resubstitution and Leave-One-Out Methods

Marc H. Kroopnick
Association of American Medical Colleges, mkroopnick@aamc.org

Jinsong Chen
The George Washington University, cjs@gwu.edu

Jaehwa Choi
The George Washington University, jaechoi@gwu.edu

C. Mitchell Dayton
University of Maryland, dayton@umd.edu

Recommended Citation

Kroopnick, Marc H.; Chen, Jinsong; Choi, Jaehwa; and Dayton, C. Mitchell (2010) "Assessing Classification Bias in Latent Class Analysis: Comparing Resubstitution and Leave-One-Out Methods," *Journal of Modern Applied Statistical Methods*: Vol. 9 : Iss. 1 , Article 7.
DOI: 10.22237/jmasm/1272686760
Available at: http://digitalcommons.wayne.edu/jmasm/vol9/iss1/7

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 52-63

1538 – 9472/10/$95.00

Assessing Classification Bias in Latent Class Analysis: Comparing Resubstitution and Leave-One-Out Methods

Marc H. Kroopnick
Association of American Medical Colleges

Jinsong Chen Jaehwa Choi
The George Washington University

C. Mitchell Dayton
University of Maryland, College Park

This Monte Carlo simulation study assessed the degree of classification success associated with resubstitution methods in latent class analysis (LCA) and compared those results to those of the leave-one-out (L-O-O) method for computing classification success. Specifically, this study considered a latent class model with two classes, dichotomous manifest variables, restricted conditional probabilities for each latent class and relatively small sample sizes. The performance of resubstitution and L-O-O methods on the lambda classification index was assessed by examining the degree of bias.

Key words: Resubstitution methods, multivariate classification, latent class analysis, leave-one-out, lambda classification index.

Introduction

Classifying individuals into groups is a popular multivariate technique, methods for which include: logistic regression analysis and discriminant function analysis with manifest group membership and cluster analysis and latent class analysis (LCA) with latent group membership (Everitt, Landau & Leese, 2001). Measures of classification success, however, can be biased in the positive direction because the data used for model estimation are also used to evaluate the success of classification (Hand, 1986). Measures of classification success based on the same data used to fit the model are referred to as resubstitution measures (Huberty, 1994; Clancy, 1997). The leave-one-out method (L-O-O), initially proposed by Lachenbruch (1967) to obtain approximately unbiased classification success measures, may be a viable alternative to the resubstitution method. Huberty (1994) also provides an illustration of the L-O-O method compared to other methods in the context of discriminant function analysis.

Two common measures for classification success in LCA are proportion correctly classified, P_c, and the statistic, λ (lambda), which adjusts P_c for chance level classification into the largest latent class (Goodman & Kruskall, 1954). Investigation of this bias in small samples sizes was suggested in Dayton (1998) but has yet to be widely addressed in the latent class literature. In order to assess the degree of bias, the traditional resubstitution computation of λ and the λ computed using the L-O-O method were compared to a theoretical value for λ.

Latent Class Analysis

Latent Class Analysis (LCA) is a statistical technique for multivariate categorical data that is used to discover subtypes of

Marc H. Kroopnick is a Senior Measurement Research Analyst. He recently earned a Ph.D. from the Department of Measurement, Statistics and Evaluation at the University of Maryland, College Park. Email: mkroopnick@aamc.org. Jinsong Chen is an Ed.D. candidate in the Graduate School of Education and Human Development. Email: cjs@.gwu.edu. Jaehwa Choi is an Assistant Professor of Educational Research in the Graduate School of Education and Human Development. Email: jaechoi@gwu.edu. C. Mitchell Dayton is a Professor Emeritus and past Chair in the Department of Measurement, Statistics, and Evaluation. Email: cdayton@umd.edu.

individuals or to confirm hypothesized subtypes of individuals (see Dayton, 1998, for more latent class model details). LCA is useful for: (1) estimating latent class proportions (class sizes) for two or more latent classes and conditional probabilities for the manifest variables; and (2) assigning individuals to the latent classes using Bayes' theorem. An example of LCA is locating distinctive cognitive diagnostic categories from examinees' answers to achievement test items in an educational context. Subsequently, Bayes' theorem can be used to assign examinees to the diagnostic categories that are most likely based on their observed responses.

Theoretical Framework

Successful classification of individuals into latent classes is a fundamental component to LCA. Following Dayton (1998), Bayes' theorem is used to determine the posterior probability of membership in each latent class, *t,* given a specific response vector, **y**$_s$:

$$P(t \mid \mathbf{y}_s) = \frac{P(t \mid \mathrm{y}_s) \times \pi_t^X}{\Sigma[P(t \mid \mathrm{y}_s) \times \pi_t^X]} \tag{1}$$

where π_t is the latent class proportion, *X* is the latent variable with levels (classes) *t* in *T*, and $\mathbf{\Sigma}[P(t \mid \mathrm{y}_s) \times \pi_t^X]$ is the unconditional (across all latent classes) probability for the response vector **y**$_s$. All individuals with the same response pattern are classified into the latent class, *t,* with the largest posterior probability corresponding to its response vector, **y**$_s$. The following formula expresses the proportion correctly classified, *P*$_c$:

$$P_c = \frac{\Sigma[n_s \times \max P(t \mid \mathbf{y}_s)]}{N} \tag{2}$$

where max P(*t*| **y**$_s$) is the largest posterior probability for response **y**$_s$ across all latent classes *T* , *n*$_s$ is the number of cases corresponding to the response vector **y**$_s$, and *N* is the total number of cases. Note that the number of possible response vectors is 2^v, where *v* is the number of manifest variables; thus, 2^v elements would be in the summation at the population level and, for sample based analyses, up to 2^v elements.

Chance level of correct classification, which is maximized by classifying all cases into the largest latent class, is not accounted for in *P*$_c$. Goodman and Kruskall (1954) developed the λ (lambda) statistic as an adjusted value of *P*$_c$.

$$\lambda = \frac{(P_c - \pi_M^X)}{(1 - \pi_M^X)}, \tag{3}$$

where π_M^X represents the largest latent class proportion.

Considering that the parameter estimation and classification success for the latent class model are based on the same data (i.e., resubstitution), Dayton (1998) noted that values for *P*$_c$ and λ tend to be biased upward (more so with small sample sizes) and that research investigating the magnitude and methods to correct for this have yet to be studied in great detail; thus, this provided the motivation for this study. Work by Dias and Vermut (2006), however, used bootstrapping techniques to assess classification uncertainty in LCA. Their research brought to light the risk of using traditional resubstitution methods, especially at the individual response vector level.

The Leave-One-Out Method

A so-called jackknife method for determining an unbiased estimate for classification accuracy was developed by Lachenbruch (1967). His study focused on discriminant analysis and his method has been named the leave-one-out (L-O-O) method (Huberty, 1994). This method involves two basic steps. First, the model is estimated in the sample with one observation deleted, and then the resulting parameter estimates are used to classify the single deleted observation. This process was carried out *N* times so that each observation was deleted and classified. Consequently, the measure of successful classification is the proportion of times that the deleted observation was correctly classified (Huberty, 1994).

In order to investigate the bias reduction property of the L-O-O method, Lachenbruch

(1967) conducted a small Monte Carlo simulation study with 300 replications for a two group discriminant analysis The proportions of correct classifications according to both the resubstitution and L-O-O methods were calculated and empirical 95% confidence intervals (CIs) were obtained for those proportions. The CIs for the L-O-O method contained the true population value 93.3% of the time and the resubstitution method contained the true value 84.7% of the time. These results suggested the appropriateness and usefulness of Lachenbruch's L-O-O technique. Lachenbruch's procedure, with modifications, was employed in this LCA study, which involved a greater number of replications.

Methodology

Simulation Conditions

This study considered a latent class model with two classes, dichotomous manifest variables, restricted conditional probabilities for each latent class and relatively small sample sizes. The number of manifest variables considered was 4 and 6; this was purposefully small due to the small sample size focus of the study and the computation complexity associated with additional variables. Sample size varied in three ways based on the number of manifest variables.

Simulation sample sizes were 3, 5, or 7 times the number of possible response vectors. For example, applying the first weight, 3, to the four variable case yields a sample size of $3\times2^4 =$ 48. The latent class proportions and conditional probabilities for responses to the manifest variables followed a structure similar to that used in Holt and Macready (1989). The first set of latent class proportions had no discrepancy (.5, .5), and the second set had a large discrepancy, (.8, .2).

Three sets of conditional probabilities were tested; the first set had a small disparity (.7, .4), the second set had a moderate disparity (.8, .3) and the last had larger disparity (.9, .05). The first number in the set corresponded to the conditional probability of a positive response to all items for the larger latent class (if there was one) and the second number applied to the smaller latent class (if there was one). Thus, the conditional probabilities were homogeneous across manifest variables within each latent class. In sum, this simulation included the following number of cells: 2 (number of variables)*3 (sample size cases)*2 (latent class proportions)*3 (conditional probability sets) for a total of 36 simulation conditions.

Data Generation and LCA Parameter Estimation

Monte Carlo simulation methods were used to generate data consistent with the parameters described above. MATLAB (The MathWorks Inc., 2007) was used to conduct the simulation. Following guidelines in Holt & Macready (1989), there were 500 replications per cell. The flexible Expectation-Maximization (EM) (Dempster, Laird & Rubin, 1977; McLachlan & Krishnan, 1997) algorithm was programmed in MATALB to provide the maximum-likelihood estimates (MLE) of the parameters in the latent class model. The iterative EM algorithm is a popular parameter estimation technique in LCA because there is no closed form formulation for their MLE computation (Dayton, 1998). It is the default estimation method in LEM (Vermut, 1997) or Mplus (Muthén & Muthén, 2004) and, typically, LEM or MPlus would be the program of choice, but MATLAB offers more advanced and useful data manipulation options. The accuracy of the costume MATALB code was compared the estimates obtained in Mplus.

Resubstitution and L-O-O Methods for Lambda Computation

The performance of resubstitution and L-O-O methods on the lambda (λ) classification index was assessed by examining the degree of bias. Thus, for each replication in each simulation cell, the L-O-O and resubstitution lambda was computed and compared to the theoretical λ value. The calculation of the sample based resubstitution $P_c^{'}$ and λ', followed equations (2) and (3), respectively, but used the MLE parameter estimates obtained from the LCA estimation from the sample data associated with each replication in each cell.

The L-O-O method calculation was conducted in a similar fashion to that of the Lachenbruch (1967) simulation study, but was modified for LCA. A description of this procedure is: For each response vector from the generated sample data, each unique response vector was deleted and the parameters re-estimated. The max $P'(t|\ \mathbf{y}_s)$ for the deleted response vector, $\mathbf{y}_s$, was determined according to equation (1), but based on the re-estimated parameters from the $N - 1$ cases. The deleted response vector was placed back in the data set and the process was repeated for the next unique response vector.

After this process, each of the (up to) 2^v max $P'(t|\ \mathbf{y}_s)$ values was weighted by the appropriate $n_{s,}$ summed, and divided by N (equation 2); essentially this is a jackknifed $P_c^{'}$, which will be called P_c^*. Alternately the equivalent procedure (described above) could be conducted by deleting each case instead of each unique response vector and equally weighting the max $P'(t|\ \mathbf{y}_s)$ associated with each deleted case. The latter was performed for this study. Note that the L-O-O method based estimate for this index requires N estimations and the possibility exists for not getting a converged solution during each of the N estimations. If the estimation associated with a given deleted case failed to converge, the case was eliminated from the analysis and N was adjusted accordingly.

This value appeared in the numerator of the L-O-O method lambda, which will be called λ^*. The maximum latent class proportion estimate used to compute λ' was also used to compute λ^*. This provided a means by which to be able to directly compare the degree of classification success above the chance success of classifying all simulees in the largest estimated latent class proportion based on the entire dataset, $\pi_M^{'X}$. The formula for λ^* is:

$$\lambda^* = \frac{(P_c^* - \pi_M^{'X})}{(1 - \pi_M^{'X})}. \tag{4}$$

Simulation Study Outcomes

The two outcome measures evaluated were the degree of bias and the performance of 95% confidence intervals based on λ' and λ^* in capturing the true value, λ. The true value, λ, was computed by applying the true population generating parameters to equations (1), (2) and (3). First, to evaluate the bias of λ' and λ^*, the mean of the estimates, *M,* was computed and compared to the theoretical value for lambda. The percent difference between each mean and corresponding λ was reported.

Second, within each cell, up to 500 (depending on the number of converged solutions) 95% CIs were computed for each λ' and λ^*. As noted, for the L-O-O method, an estimate of λ^* is treated as a converged solution unless the N estimations do not converge while there is only one estimation required to obtain λ', the resubstitution value. The method for CI construction was based on the method for computing proportion CIs developed by Wilson (1927) and further described by Newcombe (1998). The computation of the interval is as follows:

$$\frac{2np + z^2 \pm \sqrt{z^2 + 4npq}}{2(n + z^2)}, \tag{6}$$

where p is the lambda value, q is 1- p, n is the sample size for the given cell, and z is 1.96. The degree of bias was measured by subtracting the proportion of times the two types (resubstitution and L-O-O) of CIs contained the theoretical λ from 95%. Note that both of these measures are reasonable methods, but not necessarily the only ways, to assess the performance of the two methods in terms of bias (i.e., comparing the observed to statistic to truth).

Results

The simulation outcome measures described above are summarized in Tables 1 and 2 for the 4 and 6 variables cases, respectively. Note that, except for the confidence interval coverage for one cell of the study, the difference between both simulation outcome measures associated with resubstitution and L-O-O methods was very small; i.e., less than .02 in absolute value. Figures 1 and 3 provide a graphical display of the outcome measures for the 4 variable case and Figures 2 and 4 provide a graphical display for the six variable case. While the results for the resubstitution and L-O-O methods mirrored each other, trends emerged from the various factors manipulated.

Table 1: Simulation Results when ν = 4

N	LC Max	Cond. Prob.	%RE	%LOO	.95-%RE	.95-%LOO	M_{RE} - λ	M_{LOO} - λ
48	0.500	(.7,.4)	0.330	0.274	0.620	0.676	0.204	0.200
48	0.500	(.8,.3)	0.682	0.680	0.268	0.270	0.066	0.068
48	0.500	(.9,.05)	0.992	0.994	-0.042	-0.044	0.009	0.009
48	0.800	(.7,.4)	0.010	0.018	0.940	0.932	0.531	0.526
48	0.800	(.8,.3)	0.164	0.154	0.786	0.796	0.242	0.233
48	0.800	(.9,.05)	0.964	0.962	-0.014	-0.012	0.014	0.014
80	0.500	(.7,.4)	0.356	0.344	0.594	0.606	0.091	0.094
80	0.500	(.8,.3)	0.728	0.736	0.222	0.214	0.027	0.028
80	0.500	(.9,.05)	0.986	0.986	-0.036	-0.036	0.004	0.004
80	0.800	(.7,.4)	0.032	0.036	0.918	0.914	0.419	0.413
80	0.800	(.8,.3)	0.262	0.248	0.688	0.702	0.164	0.163
80	0.800	(.9,.05)	0.930	0.930	0.020	0.020	0.007	0.007
112	0.500	(.7,.4)	0.360	0.344	0.590	0.606	0.016	0.016
112	0.500	(.8,.3)	0.756	0.758	0.194	0.192	0.006	0.006
112	0.500	(.9,.05)	0.982	0.982	-0.032	-0.032	0.002	0.002
112	0.800	(.7,.4)	0.042	0.050	0.908	0.900	0.317	0.312
112	0.800	(.8,.3)	0.356	0.356	0.594	0.594	0.117	0.116
112	0.800	(.9,.05)	0.928	0.926	0.022	0.024	0.003	0.003

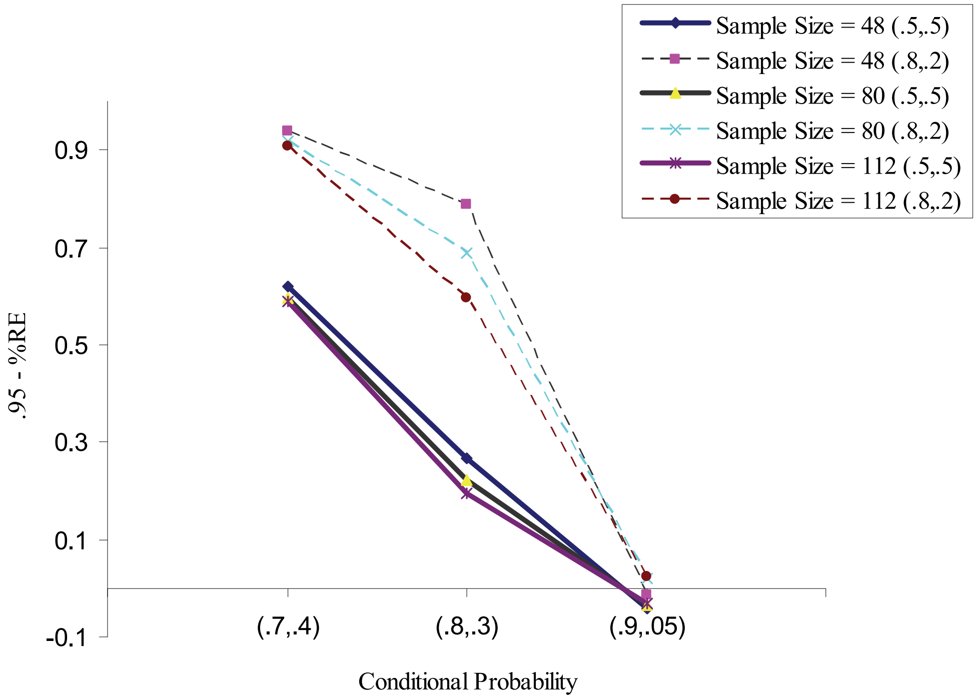
Note: LC MAX is the first latent class population proportion; Cond. Prob. is the population conditional probability for all responses; %RE is the percentage of the resubstitution method CIs containing λ; %RE is the percentage of the resubstitution method CIs containing λ; M_{RE} is the mean of the λ estimates based on the resubstitution method; M_{RE} is the mean of the λ estimates based on the L-O-O method.

Table 2: Simulation Results when $\nu = 6$

N	LC Max	Cond. Prob.	%RE	%LOO	.95-%RE	.95-%LOO	M_{RE} - λ	M_{LOO} - λ
192	0.500	(.7,.4)	0.554	0.558	0.396	0.392	-0.021	-0.024
192	0.500	(.8,.3)	0.878	0.878	0.072	0.072	-0.011	-0.011
192	0.500	(.9,.05)	0.990	0.990	-0.040	-0.040	0.000	0.000
192	0.800	(.7,.4)	0.046	0.050	0.904	0.900	0.247	0.240
192	0.800	(.8,.3)	0.538	0.542	0.412	0.408	0.047	0.047
192	0.800	(.9,.05)	0.972	0.972	-0.022	-0.022	0.002	0.002
320	0.500	(.7,.4)	0.494	0.496	0.456	0.454	-0.054	-0.054
320	0.500	(.8,.3)	0.836	0.836	0.114	0.114	-0.011	-0.011
320	0.500	(.9,.05)	0.986	0.986	-0.036	-0.036	0.000	0.000
320	0.800	(.7,.4)	0.092	0.086	0.858	0.864	0.130	0.128
320	0.800	(.8,.3)	0.636	0.636	0.314	0.314	0.020	0.020
320	0.800	(.9,.05)	0.968	0.968	-0.018	-0.018	0.000	0.000
448	0.500	(.7,.4)	0.446	0.446	0.504	0.504	-0.053	-0.054
448	0.500	(.8,.3)	0.850	0.850	0.100	0.100	-0.010	-0.010
448	0.500	(.9,.05)	0.994	0.994	-0.044	-0.044	0.000	0.000
448	0.800	(.7,.4)	0.176	0.178	0.774	0.772	0.099	0.096
448	0.800	(.8,.3)	0.588	0.588	0.362	0.362	0.009	0.010
448	0.800	(.9,.05)	0.974	0.974	-0.024	-0.024	0.000	0.000

Note: LC MAX is the first latent class population proportion; Cond. Prob. is the population conditional probability for all responses; %RE is the percentage of the resubstitution method CIs containing λ; %RE is the percentage of the resubstitution method CIs containing λ; M_{RE} is the mean of the λ estimates based on the resubstitution method; M_{RE} is the mean of the λ estimates based on the L-O-O method.

Figure 1: .95 - %RE and 95 - %LOO over Conditional Probabilities when ν = 4



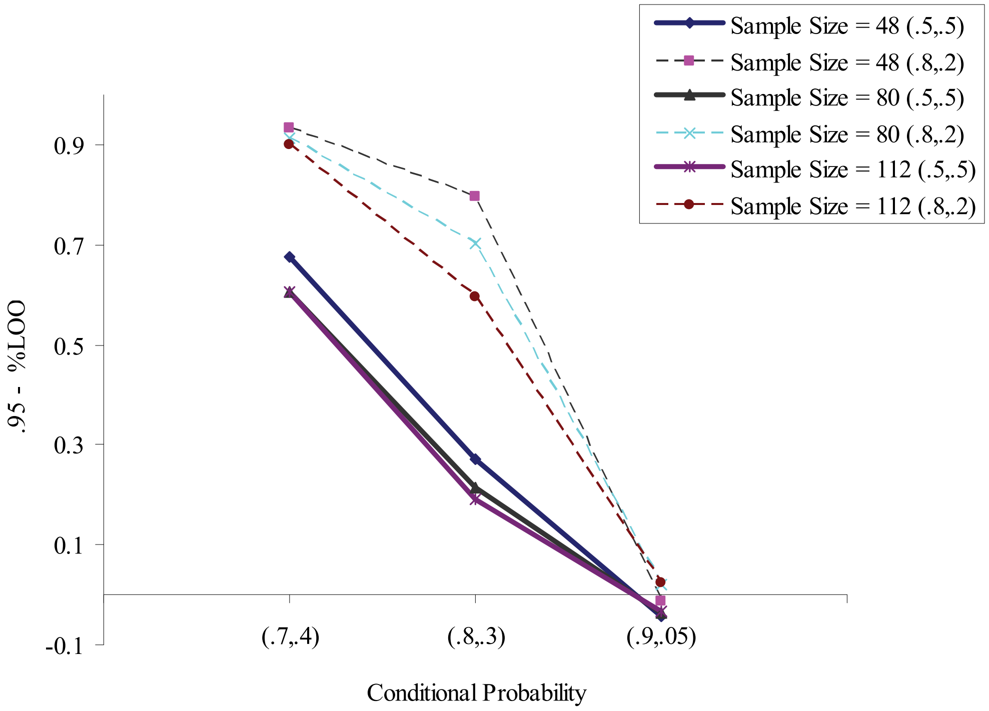
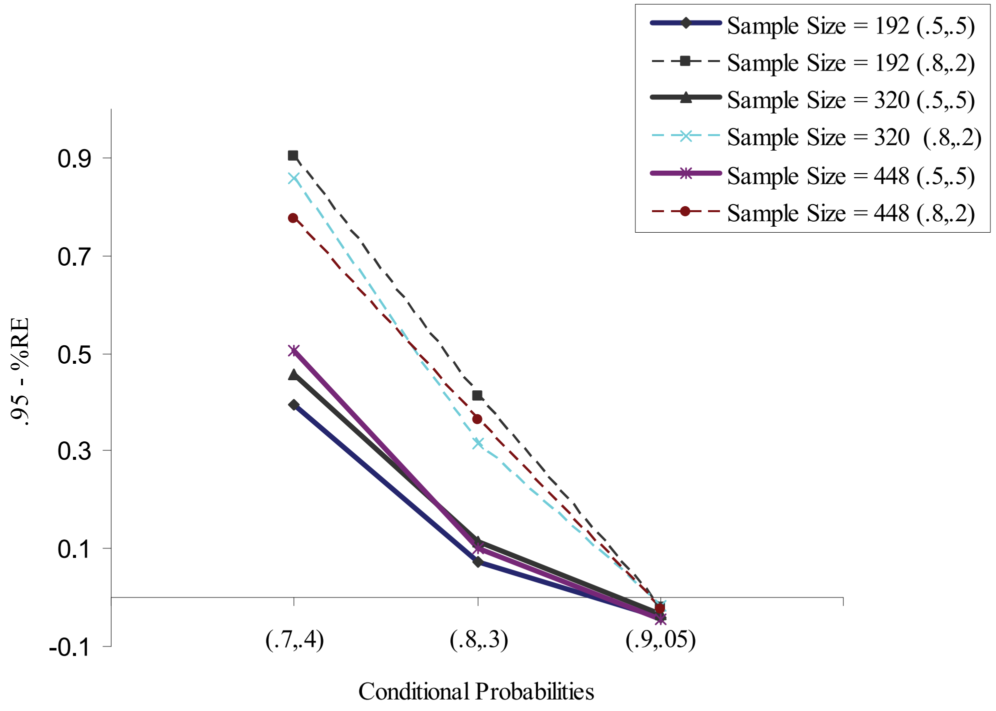


Figure 2: .95 - %RE and 95 - %LOO over Conditional Probabilities when ν = 6



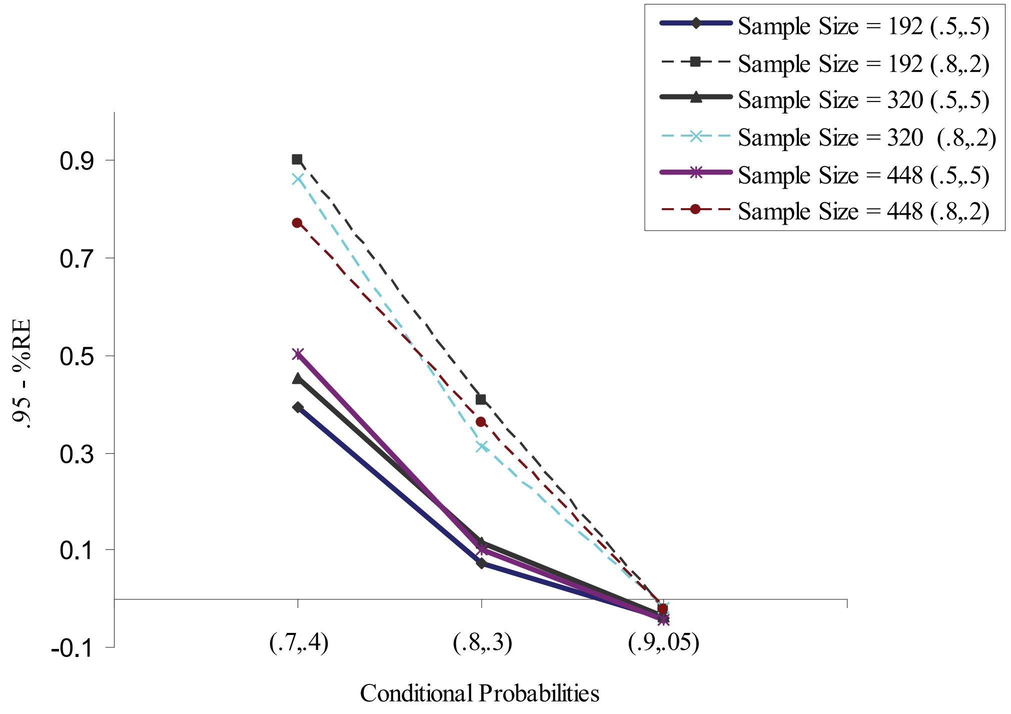


Figure 3: RE and LOO BIAS over Conditional Probabilities when ν = 4

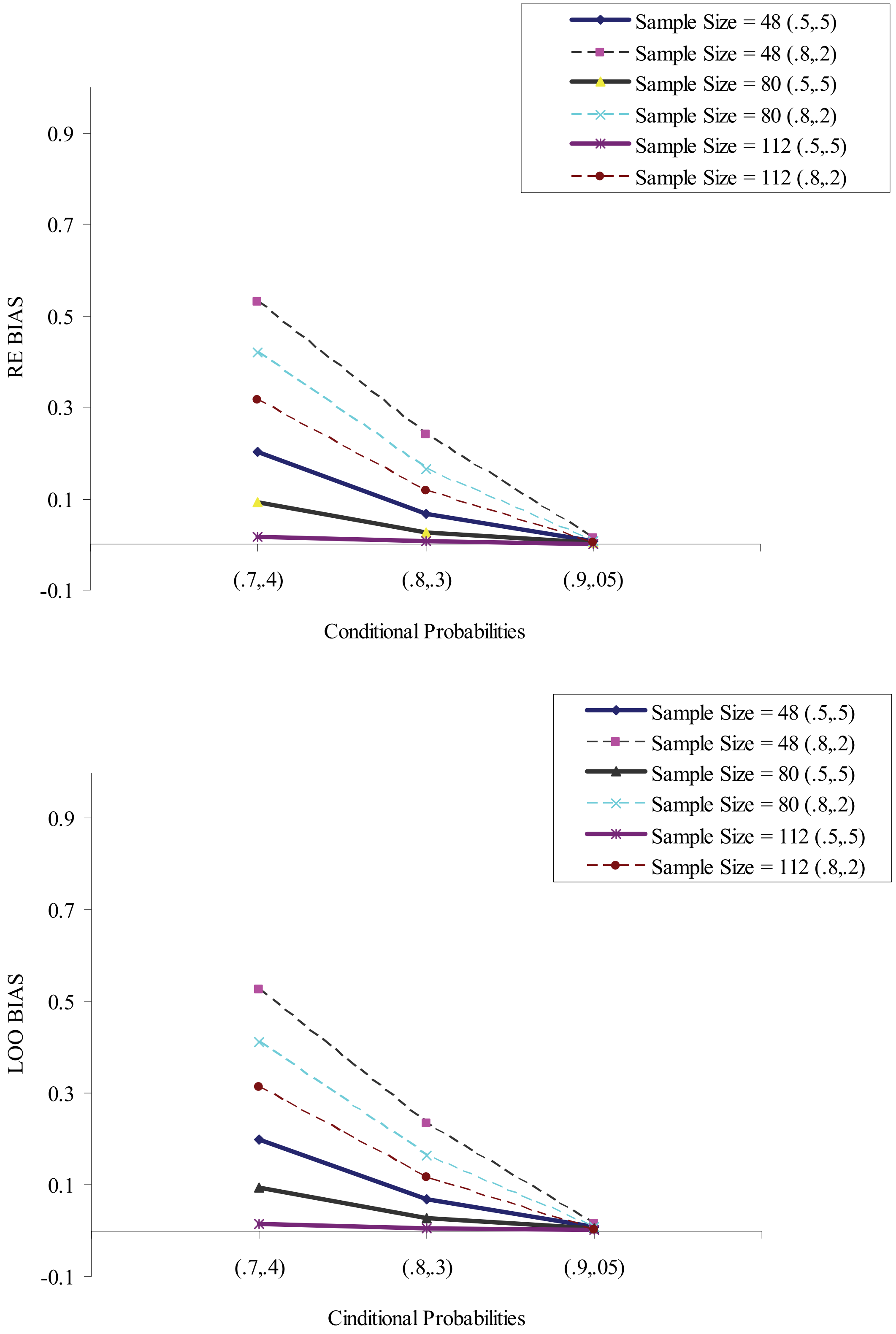
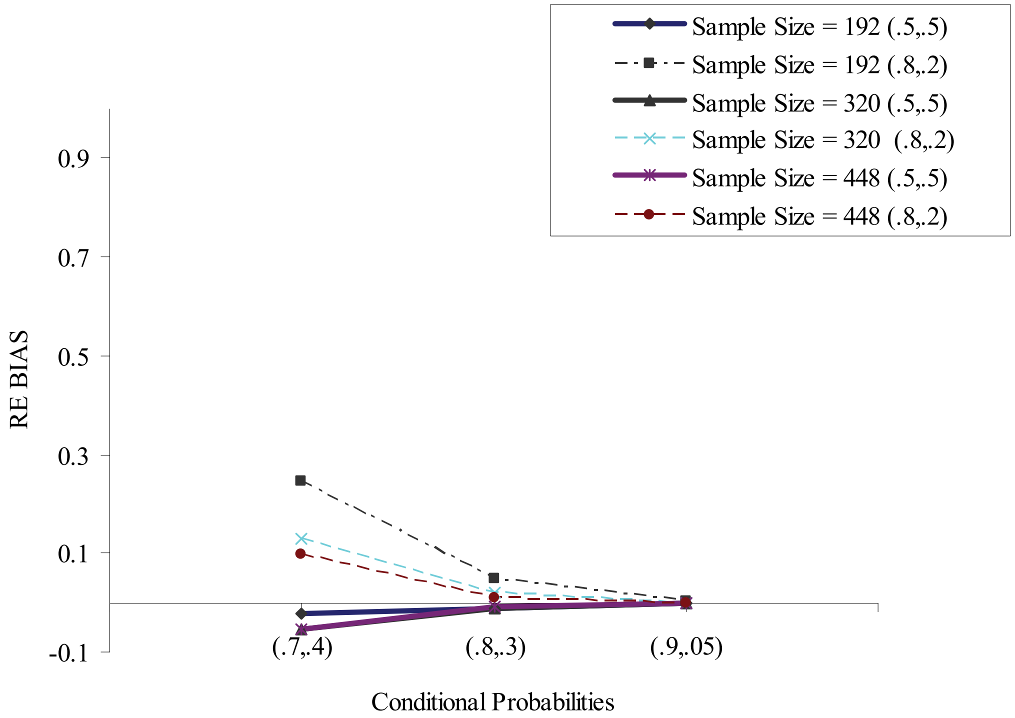
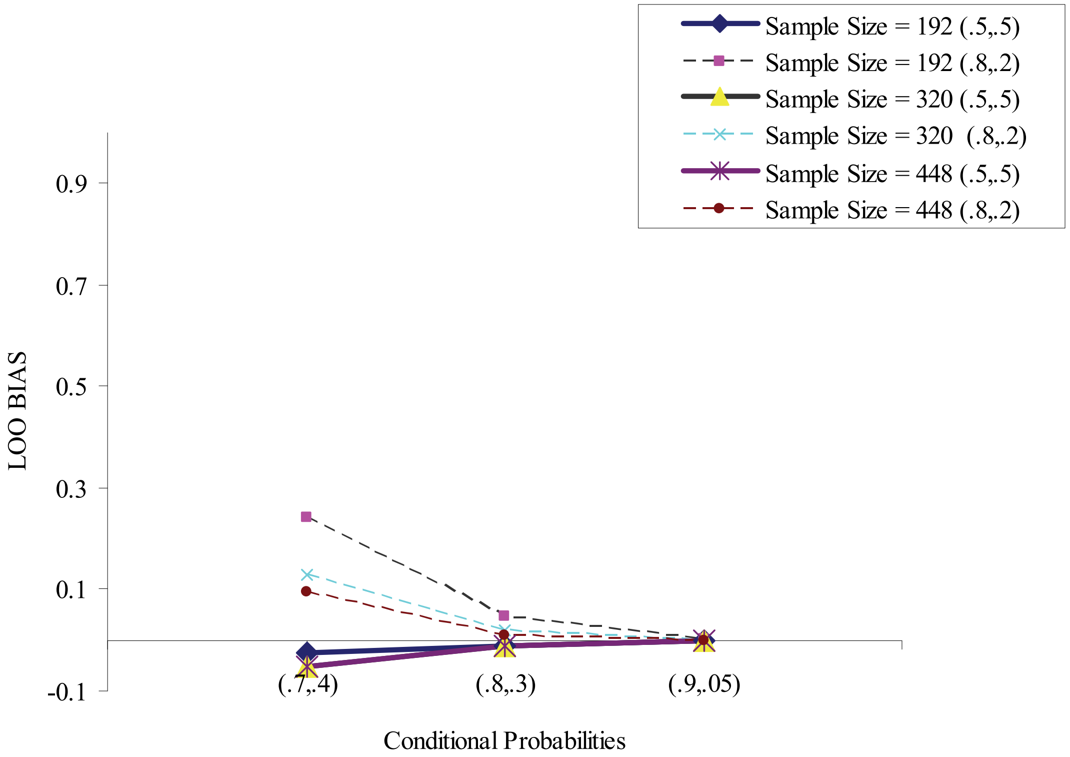


Figure 4: RE and LOO BIAS over Conditional Probabilities when ν = 6





Overall, results were largely consistent with expectations: Assessing classification accuracy improves with increasing samples size, larger numbers of variables, more discrepant conditional probabilities, and equal (i.e., less discrepant) latent class proportions. In terms of absolute numbers, the outcome measures from the simulation strongly suggested the best results across all other conditions occurred when the conditional probabilities were the most discrepant. In sum:

- Overall, more bias and less confidence interval coverage for the (.8, .2) latent class proportions resulted compared to the (.5, .5) latent class proportions.
- Overall, more bias and less confidence interval coverage for the 4 variable case was observed compared to the 6 variable case.
- For any given pair of latent class proportions, bias decreased and confidence interval coverage increased as sample size increased.
- For any given pair of latent class proportions, the variability of bias across sample sizes decreased as the discrepancy of conditional probabilities increased.
- For any given pair of latent class proportions, as the discrepancy of the conditional probabilities increased, the bias decreased and the confidence interval coverage increased.

Conclusion

The primary purpose of the study was to illustrate differences between L-O-O and resubstitution methods for assessing classification accuracy in latent class analysis. Simulation results indicated very little difference in the methods based on outcome measures. However, the accuracy measures did vary over the factors manipulated in this study and should provide researchers with a guide regarding what to expect in their studies. It is important to note that when the conditional probabilities were very discrepant, other factors had little influence and accuracy was high.

Generalizing beyond the factors and the scope of this study should be approached cautiously. As noted earlier, only a two class latent class model with restricted conditional probabilities and relatively small sample sizes was considered. Research comparing and evaluating these classification accuracy measures applied to more complicated latent class models, larger sample sizes and an increased number of variables is warranted. This research provides a baseline of possible outcomes when those future studies are conducted.

References

Clancy, E. A. (1997). Factors influencing the resubstitution accuracy in Multivariate classification analysis: implications of study design in ergonomics. *Ergonomics*, *40*(*4*), 417-427.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13,* 195- 212.

Dayton, C. M. (1998). *Latent class scaling analysis: Quantitative applications in the social sciences series, no. 126*. Thousand Oaks, CA: Sage Publications.

Dempster, A, Laird, N., & Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the royal statistical society*, *Series B, 39*(*1*), 1-38.

Dias, J. G., & Vermunt, J. K. (2006). Bootstrap methods for measuring classification uncertainty in latent class analysis. In A. Rizzi & M. Vichi (Eds.), *Proceedings in computational statistics*, 31-41). Heidelberg, Germany: Springer.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster Analysis*, *4th Ed.* New York: Oxford University Press.

Goodman, L. A., & Kruskall, W. H. (1954). Measures of association for cross-classification. *Journal of the American statistical association*, *49*, 732-764.

Hand, D. J. (1986). Recent advances in error rate estimation. *Pattern Recognition Letters*, *4*, 335-346.

Holt, J. A., & Macready, G. B. (1989). A simulation study of the difference Chi-square statistic for comparing latent class models under violation of regularity conditions. *Applied Psychological Measurement*, *13*, 221-231.

Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.

Kolb, R. & Dayton, C. M. (1996). Correcting for nonresponse in latent class analysis. *Mutivariate Behavorial Research*, *31*, 7-32.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence Intervals for the probability of misclassification in discriminant analysis. *Biometrics*, *23*, 639-645.

McLachlan, G. & Krishnan, T. (1997). *The EM algorithm and extensions: Wiley series in probability and statistics*. New York: Wiley.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (*3rd Ed.*). Los Angeles, CA: Muthén & Muthén.

Newcombe, R. (1998). Two-sided confidence intervals for the single proportion: A comparative evaluation of seven methods. *Statistics in Medicine*, *17*, 857-872.

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science*, *12*(*1*), 103-124.

The Math Works Inc. (2007). Documentation for Math Works products, R2007b [Electronic version]. Retrieved 10/07/2007 from http://www.mathworks.com/access/helpdesk/help/helpdesk.html.

Vermunt, J. K. (1997). *LEM 1.0: A general program for categorical ldata*, Tilburg: Tilburg University.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209-212.