Using Latent Class Models to Better Understand Reliability in Measures of Labor Force Status[1]

Bac Tran, Clyde Tucker

***Bac.Tran@census.gov*, U.S. Census Bureau, 4700 Silver Hill Road, Washington, D.C. 20233-1912**
***Tucker.Clyde@bls.gov* , Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Washington, DC 20212**

Key Words: Latent Class Models, Reinterview, Panel Data, Unemployment, First-order Markov, Second-order Markov, Mover-Stayer, Rotation Group, Month-in-Sample, and Classification Probability.

I. Introduction

The Current Population Survey (CPS) is a U.S. national household survey conducted by the U.S. Census Bureau for the Bureau of Labor Statistics (BLS). It is designed to generate national and state-level estimates of labor force characteristics such as: employed (E), unemployed (UE), and not in the labor force (NILF); demographic characteristics; and other characteristics of the 16^+ non-institutionalized civilian population. Therefore, measurement error is an important non-sampling study for the CPS survey. The CPS uses addresses from the most current U.S. Census, adding new construction, as the frame. The total sample size is about 72,000 assigned households per month. The CPS uses a 4-8-4 rotating panel design, i.e. 4 months in, 8 months out, and 4 months in. In the CPS, the same respondents are interviewed at several points following the pattern 4-8-4 (Current Population Survey Tech paper 66). For any given month, the CPS sample is grouped into eight sub-samples corresponding to the eight rotation groups.

Latent Class Analysis (LCA) has been used to estimate the response bias in survey data (Van De Pol and De Leeuw 1986; Biemer 2004; Tucker et al. 2002; Tucker et al. 2008). It has the further advantage of being able to not only estimate the simple response variance, or unreliability, but also the sources and causes of unreliability that may help explain bias as well (Bassi et al. 2000; Biemer and Bushery 2000). Currently, the Census Bureau is using a second interview; or reinterview, in parallel with LCA to estimate unreliability. The results showed that LCA agreed with reinterview results (Tran 2003, 2007). The LCA method can save some of the cost of reinterview, substantially increases the sample size for estimating unreliability, and avoids the added respondent burden.

LCA uses data collected over several waves of a survey. Previous research (Biemer and Bushery 2000, Tran and Winters 2003, Tran and Mansur 2004, and Tran and Nguyen 2007) applied one of the traditional LCA models in panel data, first-order latent Markov. However, that model could not deal with the unobserved heterogeneity in the sense that there were groups of sample persons having different transition and error probabilities. The Mover-Stayer (MS) LCA models models (Langeheine & Van De Pol, 1990; Goodman, 1961; Hagenaar et al., 2002; Vermunt et al, 1999) were used to deal with this unobserved heterogeneity by making assumptions about which respondents are "movers" and "stayers." The MS LCA model contains two first-order Markov chains. One is for unobserved 'stayers' such as those employed, who usually want to keep their jobs. The other is for the unobserved "mover" group consisting of the unemployed, who are searching for jobs. The results showed that MS outperformed just the first-order Markov model by itself in estimating the CPS unreliability.

Both the first-order and the MS models employ the Markov assumption, which assumes the current state depends only upon the previous state. A natural question is if the dependency goes beyond the previous state (i.e., second-order), is the MS model still missing some additional information about the dynamics in the measurement of labor force status? This paper is seeking an answer for that question by going through past LC models and comparing their results in terms of classification probabilities. We will present five different models in this paper: first-order, second-order, MS with first-order, MS with second-order, and MS with second-order where mover and stayer are defined alternatively by observing whether a reported change in labor force status occurred over a four-month period. We use those five models to see which one best fits the CPS data. The data used in the analysis was from January 2008 to December 2009, andLatentGold4.5, software developed by Statistical Innovations, was used to implement the model estimation.

II. Latent Class Models

Latent Class Analysis (LCA) treats the true classification of the labor force status as an unobserved variable. The observed variables (A_1, A_2, A_3, and A_4) are labor force status of a respondent in four consecutive months (see Figure 1 and 2) obtained from the CPS survey. They are fallible indicators of the latent variable Xs. LCA suggests a relationship between observed variables and latent variables through a mathematical equation. Under the equation the table of observed data is viewed as a partial table from a full table of observed and unobserved data. The Markov assumption is employed, hence the so-called Markov Latent Class.

A. Model 1: First-Order Latent Markov Model

Let y_{it} denote the observed value of the dependent variable at time point t for a person with response pattern i. In this paper y_{it} is a categorical variable with M = 3 categories (E, UE, and NILF). The total number of time points is T + 1, where in this analysis four time points were used (T= 0, 1, 2, 3). The response vector of length T + 1 containing all the responses for respondent i is denoted by $\mathbf{y_i}$, and the associated model probability by $P(\mathbf{y_i})$. A first-order Markov LC model is a LC model with T + 1 latent variables, each having K categories (Van de Pol and Langeheine, 1990). In this paper we will only work with models in which K = M = 3, the number of latent labor force states which is equal to the number of observed labor force states. Let x_t denote a possible

[1] *This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau, or the Bureau of Labor Statistics.*

value of the latent variable at time point t, where $x_t = 0, 1, 2$ (E, UE, NILF). The first-order Markov LC model has the following form:

$$P(\mathbf{y_i}) = \sum_{x_0=1}^{3}\sum_{x_1=1}^{3}\sum_{x_2=1}^{3}\cdots\sum_{x_4=1}^{3} P(x_0)\prod_{t=1}^{4}P(x_t|x_{t-1})\prod_{t=0}^{3}P(y_{it}|x_t) \quad (1)$$

The First-Order Markov assumption states that:

$$P(x_t|x_{t-1}) = P(x_t|x_{t-1}, x_{t-2}, \ldots, x_0).$$

The assumptions of the first-order Markov Latent Class in the equation (1) are:

1. x_t is independent of $x_{t-2}, x_{t-3}, \ldots, x_0$;
2. There is no unobserved heterogeneity; and
3. Classification errors are independent across time points.

$P(x_0)$: initial latent state probabilities;
$P(x_t \mid x_{t-1})$: transition probability; and
$P(y_{it}|x_t)$: classification error probabilities.

The unknown model probabilities to be estimated are the initial latent state probabilities P(x0), the latent transition probabilities $P(x_t|x_{t-1})$, and the classification error probabilities $P(y_{it}|x_t)$.

B. Model 2: Second-Order Latent Markov Model

The second-order model differs from first-order models in which the current state not only depends on previous state but also on the state before that, in other words:

$$P(\mathbf{y_i}) = \sum\sum\sum\sum P(x_0)\,P(x_1|x_0)\prod_{t=2}^{4}P(x_t|x_{t-1},x_{t-2})\prod_{t=0}^{4}P(y_{it}|x_t) \quad (2)$$

C. Model 3: Mover-Stayer Latent Markov Model

The MS model, a two-class mixed Markov Latent Class model, assumes that there are two unobserved subgroups with different transition probabilities. These subgroups are theoretically defined as described in the introduction. The MS model has the form:

$$P(y_i) = \sum_{w=1}^{2}\sum_{x_0=1}^{3}\sum_{x_1=1}^{3}\cdots\sum_{x_T=1}^{3} P(w)P(x_0 \mid w)\prod P(x_t \mid x_{t-1}, w)\prod P(y_{it} \mid x_t, w) \quad (3)$$

where w=1 or 2 denotes two classes of latent variables, e.g., w=1 (the Stayers class) is for those Employed or NILF, and w=2 represents the Mover class and includes only the Unemployed.

In our study we assume classification error is time homogeneous. The path diagrams for the first-order model and MS model, applied to the CPS labor force with four consecutive time periods, are given below:

Figure 1: First-Order

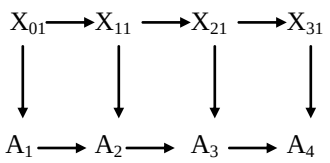
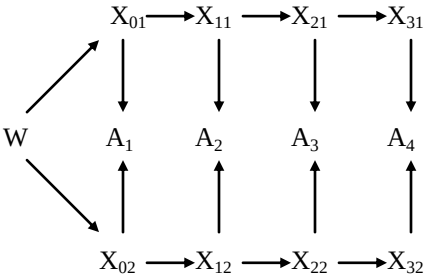


Figure 2: Mover-Stayer



The latent variable W represents the two subgroups, movers and stayers. The Xs are latent variables that represent the true labor force status at four time points. A_1, A_2, A_3, and A_4 are the observed labor force variables used as the indicators of the Xs.

D. Model 4: Mover-Stayer Latent Markov Model with Second-Order

In analogy to **II.B** additional probabilities $P(x_1 \mid x_0)$ and $P(x_2 \mid x_1, x_0)$, $P(x_3 \mid x_2, x_1)$, $P(x_4 \mid x_3, x_2)$ were introduced to the MS model to take into account the influence of the state before the previous state. In other words:

$$P(y_i) = \sum_{w=1}^{2}\sum_{x_0=1}^{3}\sum_{x_1=1}^{3}\cdots\sum_{x_4=1}^{3} P(w)P(x_0 \mid w)P(x_1 \mid x_0, w)\prod P(x_t \mid x_{t-1}x_{t-2}, w)\prod P(y_{it} \mid x_t, w) \quad (4)$$

E. Model 5: Model II.D with Redefined Stayer Group

Model 5 is obtained from Model 4 by redefining the mover and stayer classes based on the observed transitions from month-to-month, also described in the introduction. The theoretically defined movers and stayer provide more explanatory power, but the models associated with them may not fit as well as when the observed transitions are used instead.

III. Validation Process

Previous research (Tran & Nguyen, 2007) provided a proof of validation process. Below are brief results in the simulation study.

The four questions we want to answer with the simulation study are:

1. Is it is possible to detect whether model assumptions are violated?
2. Are the estimated misclassification probabilities unbiased when the correct model is specified?
3. Are the estimated misclassification probabilities biased when model assumptions are violated when an incorrect model is specified?
4. Are the estimated class sizes biased when model assumptions are violated when an incorrect model is specified?

The results from simulations provide the answers as follows:

1. Yes, it is possible to detect that model assumptions are violated, but only for large violations.
2. Yes, estimates of the misclassification probabilities are unbiased when the right model is specified.
3. Yes, there is an upward bias in the estimates of the misclassification probabilities, but it is surprisingly small. Only with a very extreme (and unrealistic) second-order process do we see substantial bias in the estimated misclassification probabilities obtained with an incorrect first-order model.
4. Yes, estimates of the class sizes are biased downwards. With weak violations, this bias is negligible.

IV. Results

We focus on one set of parameters (see II.A): $P(y_{it}|x_t)$ which is called classification error probabilities. In our case they are:

P(Observed Labor Force = i | Latent Variable =j).

When i = j that probability is called a correct classification probability, and when $i \neq j$ the probability indicates the error classification probability. In this analysis i, j = E, UE, and NILF.

We compared our estimates of the CPS classification probabilities for five different models. The results are summarized in Attachment A. This Attachment also shows the log likelihood, BIC, AIC3, number of parameters, and dissimilarity index for model selection purposes. As usual the largest log likelihood, the smallest BIC and AIC3, and dissimilarity index smaller than 0.05 were used to do the model selection. The number of parameters follows the parsimony principle.

We can see that Model 4, the MS model with the second-order Markov improved the correct classification for the Unemployed class, which had a historical inconsistency, where the probability describing the fit of the observed labor force using the unobserved latent variable was lowest. Thus, the second-order effect is present and the even earlier labor force status is related to the most recent one. Furthermore, the results with Model 5, while consistent with Model 4, adds nothing new and is less theoretically pleasing than Model 4.

V. Limitation

There are limitations when using maximum likelihood procedure with missing values. The procedure can deal with missing values on response variables, but not with missing values on covariates, and it assumes that the missing data are missing at random (MAR). Local independence among the indicators is required to make LCA work. This assumption is hard to verify in practice (Vacek, 1985). This study applied a simple MS model in which there were two classes, mover and stayer. There could be more than two classes. The weight we used for the analysis was the average of the second stage weights from four-month CPS data. We need to develop a weighting scheme that is better than the averaging. LCA uses panel data for analysis, and combines all the same pattern into groups (marginal). It cannot look at the data at the person level like the reinterview.

VI. Conclusions

LCA has been used to estimate the measurement error efficiently by various researchers. It is also used in parallel with reinterview at the Census Bureau to estimate measurement error in CPS labor force status. This paper enriches the MS models by introducing the second-order term which is naturally realistic in survey data. This was clarified by comparing the classification probabilities in the second-order MS models which improved the correct classification probability for UE group to slightly over 90 percent. This figure used to be at most in the low 80's range for the first-order Markov model. However, as mentioned in **V**, LCA cannot work at the person level; therefore, the reinterview is needed whenever there are changes in design for which we need to investigate the effects on the behaviors of individual respondents. Furthermore, dropping the reinterview for measuring reliability will not save that much money, given that most of the reinterview program centers around detecting curbstoning. At the same time, we can use the individual observed transitions to validate the theoretical assertions about movers and stayers in the MS Markov LCA Model.

References

Bassi, Franesca, Jacques A. Hagenaars, Marcel A. Croon and Jeroen Vermunt. (2000). "Estimating True Changes When Categorical Panel Data Are Affected by Uncorrelated and Correlated Classification Errors: An Application to Unemployment Data." *Sociological Methods & Research* 29: 230-268.

Biemer, P.P. (2004). "The Twelfth Morris Hansen Lecture Simple Response Variance: Then and Now", *Journal of Official Statistics*, 20 (3): 417-439

Biemer, P. and Wiesen, C. (2002). "Measurement Error Evaluation of Self-Reported Drug Use: A Latent Class Analysis of the US National Household Survey on Drug Abuse," *Journal of Royal Statistical Society*, Part 1, 165, pp. 97-119.

Biemer, P. and Bushery, J. (2000). "On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data," *Survey Methodology*, Vol. 26, No. 2, pp. 139-152.

Clyde Tucker, Brian Meekins, and Biemer, P. (2008). "A Microlevel Latent Class Model for Measurement Error in the Consumer Expenditure Interview Survey"f

Goodman, L.A. (1961). "Statistical Methods for the Mover-Stayer Model," *Journal of the American Statistical Association*, 81, 354-365

Hagenaars, J.A, and McCutcheon A.L. (2002). Applied Latent Class Analysis. Cambridge University Press.

McCutcheon, A.L. (1987). Latent Class Analysis. Newbury Park, CA: Sage.

Tran, Bac and Nguyen, J. (2007). "Estimating the Measurement Error in the Current Population Survey Labor Force–A Latent Class Analysis Approach with Sample Design," *2007 Federal Committee on Statistical Methodology*. SESSION IV-A

Tran, Bac and Mansur, K. (2004). "Analysis of the Unemployment Rate in the Current Population Survey- A Latent Class Approach," *2004 Proceedings of the American Statistical Association*, Statistical Computing Section [CD-ROM], Alexandria, VA: American Statistical Association.

Tran, Bac and Winters, F. (2003). "Markov Latent Class Analysis and Its Application to the Current Population Survey in Estimating the Response Error," *2003 Proceedings of the American Statistical Association*, Statistical Computing Section [CD-ROM], Alexandria, VA: American Statistical Association.

Tucker, C., Biemer, P., and Vermunt, J. (2002). "Estimation Error in Reports of Consumer Expenditures," Proceedings of the ASA, Survey Research Methods Section, New York, NY.

U.S. Census Bureau (2006). *Current Population Survey: Design and Methodology*, Technical Paper 66. October 2006.

Vacek, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics, 41*, 959-968.

Van De Pol, F., and De Leeuw, J. (1986). "A Latent Markov Model to Correct for Measurement Error," *Sociology Method and Research*, 15, 118-141

Van De Pol, F., and Langeheine, R. (1990). "Mixed Markov Latent Class Models," *Sociology Methodology*, 213-247

Vermunt, J.K., Langeheine, R., and Bockenholt, U. (1999). "Latent Markov Models With Time-Constant and Time-Varying Covariates," *Journal Education and Behavioral Statistics*, 24, 178-205

Vermunt, J.K. and Magidson, J. (2007) Technical Guide to Latent Gold 4.5. Belmont Massachusetts: Statistical Innovations Inc.

Correct Classification

P(Observed = I | Unobserved = i) where i= E, UE, NILF

	Employed	Unemployed	NILF	E	U	NILF
Jan08-Apr08						
First-order Markov	0.9870 (0.0011)	**0.7311 (0.0217)**	0.9673 (0.0026)	0.6176	0.0357	0.3466
Second-order Markov	0.9921 (0.0011)	**0.7974 (0.0198)**	0.9903 (0.0032)	0.6139	0.0374	0.3487
Mover-Stayer	0.9915 (0.0010)	**0.7945 (0.0142)**	0.9825 (0.0012)	0.6135	0.0361	0.3504
Mover-Stayer with 2nd-order	0.9944 (0.0005)	**0.9104 (0.0283)**	0.9906 (0.0018)	0.6124	0.0329	0.3546
Observed MS with 2nd-order	0.9957 (0.0022)	**0.9310 (0.0258)**	0.8918 (0.0069)	0.6118	0.0796	0.3087
May08-Aug08						
First-order Markov	0.9890 (0.0011)	**0.7596 (0.0214)**	0.9653 (0.0031)	0.6190	0.0450	0.3360
Second-order Markov	0.9927 (0.0030)	**0.7973 (0.0161)**	0.9831 (0.0018)	0.6187	0.0428	0.3385
Mover-Stayer	0.9933 (0.0012)	**0.8118 (0.0140)**	0.9794 (0.0014)	0.6181	0.0418	0.3400
Mover-Stayer with 2nd-order	0.9944 (0.0038)	**0.8958 (0.0296)**	0.9896 (0.0054)	0.6198	0.3400	0.0402
Observed MS with 2nd-order	0.9891 (0.0015)	**0.9088 (0.0237)**	0.9934 (0.0029)	0.6249	0.0413	0.3338
Sep08-Dec08						
First-order Markov	0.9893 (0.0015)	**0.7806 (0.0184)**	0.9694 (0.0024)	0.6080	0.0487	0.3433
Second-order Markov	0.9915 (0.0012)	**0.8047 (0.01220**	0.9885 (0.0019)	0.6086	0.0470	0.3444
Mover-Stayer	0.9910 (0.0010)	**0.8434 (0.0141)**	0.9840 (0.0010)	0.6092	0.0455	0.3453
Mover-Stayer with 2nd-order	0.9996 (0.0005)	**0.8940 (0.0219)**	0.9427 (0.0028)	0.6060	0.3463	0.0477
Observed MS with 2nd-order	0.9943 (0.0023)	**0.8851 (0.0298)**	0.8883 (0.0074)	0.6104	0.0811	0.3086
Jan09-Apr09						
First-order Markov	0.9856 (0.0011)	**0.8034 (0.0141)**	0.9666 (0.0022)	0.5875	0.0648	0.3478
Second-order Markov	0.9977 (0.0010)	**0.8622 (0.0104)**	0.9883 (0.0016)	0.5866	0.0638	0.3497
Mover-Stayer	0.9887 (0.0013)	**0.8307 (0.0144)**	0.9705 (0.0022)	0.5875	0.0624	0.3501
Mover-Stayer with 2nd-order	0.9996 (0.0003)	**0.9215 (0.0227)**	0.9912 (0.0025)	0.5860	0.0611	0.3529
Observed MS with 2nd-order	0.9939 (0.0030)	**0.8916 (0.0157)**	0.9917 (0.0034)	0.5902	0.0614	0.3484
May09-Aug09						
First-order Markov	0.9847 (0.0013)	**0.8118 (0.0130)**	0.9594 (0.0029)	0.5882	0.0691	0.3428
Second-order Markov	0.9871 (0.0013)	**0.8422 (0.0109)**	0.9880 (0.0044)	0.5889	0.0685	0.3426
Mover-Stayer	0.9890 (0.0009)	**0.8500 (0.0097)**	0.9779 (0.0010)	0.5884	0.0662	0.3454
Mover-Stayer with 2nd-order	0.9897 (0.0041)	**0.8935 (0.0134)**	0.9919 (0.0080)	0.5909	0.0663	0.3428
Observed MS with 2nd-order	0.9988 (0.0017)	**0.9106 (0.0233)**	0.9631 (0.0052)	0.5806	0.0616	0.3578
Sep09-Dec09						
First-order Markov	0.9889 (0.0016)	**0.8317 (0.0111)**	0.9653 (0.0019)	0.5779	0.0682	0.3538
Second-order Markov	0.9963 (0.0016)	**0.8750 (0.0185)**	0.9867 (0.0032)	0.5775	0.0666	0.3560
Mover-Stayer	0.9915 (0.0010)	**0.8508 (0.0080)**	0.9820 (0.0011)	0.5783	0.0652	0.3565
Mover-Stayer with 2nd-order	0.9973 (0.0016)	**0.9127 (0.0201)**	0.9890 (0.0032)	0.5782	0.0641	0.3578
Observed MS with 2nd-order	0.9947 (0.0025)	**0.8707 (0.0187)**	0.9917 (0.0041)	0.5794	0.0694	0.3512

LCA Model Selections

Periods

		Log likelihood	BIC	AIC3	NParms	Diss. Index
Jan08-Apr08	First-order Markov	-222422.3983	445332.063	444964.7965	40	0.0389
	Second-order Markov	-224573.6914	449610.2859	449261.3828	38	0.0119
	Mover-Stayer	-204050.503	408478.5271	408194.006	31	0.0138
	Mover-Stayer with 2nd-order	-205870.3711	412240.0443	411863.7422	41	0.0115
	observed MS with 2nd-order	-205871.9084	412243.1189	411866.8168	41	0.0115
May08-Aug08						
	First-order Markov	-212460.1654	425407.6641	425040.3307	40	0.0346
	Second-order Markov	-214767.6341	429998.235	429649.2683	38	0.0134
	Mover-Stayer	-212633.7135	425645.1103	425360.4269	31	0.016
	Mover-Stayer with 2nd-order	-214601.7773	429703.0715	429326.5547	41	0.0126
	observed MS with 2nd-order	-214594.5995	429688.7159	429312.1991	41	0.0126
Sep08-Dec08						
	First-order Markov	-208616.3009	417719.3405	417352.6017	40	0.0374
	Second-order Markov	-210649.9762	421762.3543	421413.9525	38	0.0114
	Mover-Stayer	-208772.6155	417922.4535	417638.231	31	0.0141
	Mover-Stayer with 2nd-order	-210515.6966	421530.3004	421154.3931	41	0.0106
	observed MS with 2nd-order	-210508.6156	421516.1385	421140.2312	41	0.0106
Jan09-Apr09						
	First-order Markov	-222422.3983	445332.063	444964.7965	40	0.0389
	Second-order Markov	-224573.6914	449610.2859	449261.3828	38	0.0108
	Mover-Stayer	-222630.7255	445639.0826	445354.451	31	0.0139
	Mover-Stayer with 2nd-order	-214599.7469	429699.0106	429322.4939	41	0.0126
	Observed MS with 2nd-order	-224465.3998	449430.2478	449053.7996	41	0.01
May09-Aug09						
	First-order Markov	-232728.9394	465945.6297	465577.8789	40	0.0349
	Second-order Markov	-222411.8363	445384.0292	444961.6727	38	0.0392
	Mover-Stayer	-233008.8365	466395.6799	466110.6731	31	0.0146
	Mover-Stayer with 2nd-order	-234891.5763	470283.0971	469906.1525	41	0.0117
	Observed MS with 2nd-order	-234893.7269	470287.3985	469910.4539	41	0.0116
Sep09-Dec09						
	First-order Markov	-225247.8764	450983.1100	450615.7528	40	0.0369
	Second-order Markov	-227239.9918	454942.9728	454593.9835	38	0.0098
	Mover-Stayer	-225445.3145	451268.3308	450983.629	31	0.0143
	Mover-Stayer with 2nd-order	-227118.8688	454737.2787	454360.7376	41	0.0092
	Observed MS with 2nd-order	-227105.0654	454709.672	454292.1309	41	0.0091