Latent Class Analysis of Worker Knowledge of Their Employment Status

Stanislav Kolenikov[1], Kelly Daley[2]
[1] Abt Associates, 8405 Colesville Rd #300, Silver Spring, MD 20910
[2] 10420 Little Patuxent Parkway, Suite 300, Columbia, MD 21044

Abstract
The Worker Classification Survey conducted by Abt Associates for the Department of Labor in 2014 measured several concepts that are difficult for respondents to report accurately. These concepts include their own employment status as well as details about their employer, benefits, and tax forms. The survey also asked questions that some respondents may purposefully answer inaccurately. For example, certain response patterns may indicate that respondents are "working under the table" and not paying taxes on their income. We analyzed measurement error in this survey using latent class analysis (LCA), a statistical technique that links discrete latent variables, or discrete classes, with discrete responses. We have been able to identify respondents who clearly are employees of their companies; those who clearly are non-employees (e.g., contractors or consultants), as well as those whose status is unclear despite detailed collected self-reports on their documented status with the company, as well as their economic relations and job responsibilities.

Key Words: employment statistics; latent class analysis; measurement error; survey methodology

1. Basics of Latent Class Analysis

The Worker Classification Survey conducted by Abt Associates for the Department of Labor in 2014 measured several concepts that are difficult for respondents to report accurately. These concepts include their own employment status (Biemer 2004) as well as details about their employer, benefits, and tax forms. The Worker Classification Survey also asks questions that some respondents may purposefully answer inaccurately. For example, certain response patterns would indicate that respondents are "working under the table" and not paying taxes on their income. Some of the responses in the survey, thus, are likely to contain measurement error. The primary purpose of measurement error analysis is to identify questions that are flawed and which elicit unreliable or biasing data.

One tool for investigating measurement error in surveys is latent class analysis (LCA) which can be used to estimate the error in survey items without relying on external validation data. LCA is a confirmatory latent variable model with categorical observed variables (e.g., closed categories of survey questions) and categorical latent variables (e.g., a small number of population subgroups). The model coefficients are unconditional class probabilities (that can be thought of as population prevalences) and probabilities of binary or multinomial responses given the class membership (that can be thought of as the probabilities of accurate response, assuming that all members of the class are expected to answer in a certain way). Survey respondents in the same class would be expected to have

similar patterns of responses to the questions of interest, while respondents in difference classes tend to have different patterns.
LCA finds occasional use in survey statistics and methodology. Statistical foundations of the latent class models were developed in 1970s by Leo Goodman and Shelby Haberman. These models were introduced into survey statisticians' arsenal by Allan McCutcheon (1987). A team of the Bureau of Labor Statistics statisticians, including Clyde Tucker and Brian Meekins, along with Paul Biemer (RTI), produced a series of papers in early 2000s addressing measurement issues in BLS surveys, including unemployment measurement in the Current Population Survey and underrerporting of expenditures in Consumer Expenditure Interview Survey (see summary and references in Tucker, Meekins and Biemer 2010). Biemer and Wiesen (2002) and Kreuter et. al. (2008) give examples of how LCA identifies poorly performing items. Monographs by Hagenaars and McCutcheon (2009) and Biemer (2011) provide comprehensive treatment of the current state of the technique.

1.1 Estimation and testing of latent class models

A latent class model analyzes a high dimensional contingency table to assess whether a model posited by the researcher provides adequate fit. Suppose K multinomial outcome variables, $y_1 \in \{1, \dots, m_1\}, \dots, y_K \in \{1, \dots, m_K\}$ are observed in the survey. The K-way contingency table then represents a saturated model with no structure imposed:

$$p_{i_1 i_2 \dots i_K} = \mathrm{Prob}\left[y_1 = i_1,\ \dots, y_K = i_K\right]$$

This saturated model states that a probability to give a specific response on a specific item cannot be teased out, as everything is interdependent. A simple latent class model can postulate that the probability to give a particular response is dependent on one variable only, which is the latent class that a person belongs to:

$$\mathrm{Prob}[y_1 = i_1] = p_{i_1} = \sum_{c=1}^{C} \pi_c \times \mathrm{Prob}\left[y_j = i_j \middle| \text{class } c\right],$$

where the marginal class probability π_c shows how many people there are of class c; and the probability associated with a cell in the K-dimensional table is

$$p_{i_1 i_2 \dots i_K} = \sum_{c=1}^{C} \pi_c \times \prod_{j=1}^{K} \mathrm{Prob}[y_j = i_j | \text{class } c]$$

This assumption is frequently referred to as "local independence assumption": given the class membership, responses to the distinct variables $y_1, \dots, y_K$ are independent of each other. An efficient estimation procedure can be constructed using the principles of the EM algorithm (Fienberg et al 2007), which in this case represents rewriting Bayes theorem several times in and out. The expectation step computes class probabilities π_c by aggregating conditional probabilities of a class given the response patterns. The maximization step updates the estimated probabilities $\mathrm{Prob}\left[y_j = i_j \middle| \text{class } c\right]$ as a function of the observed cell frequencies and the modeled frequencies in $(K + 1)$-dimensional table of the observed and latent variables. Once the EM algorithm is declared convergent, goodness of fit test and standard errors can be obtained. Pearson χ^2 test comapres expected vs. observed frequencies, while the likelihood ratio χ^2 test is based on log-likelihood of fitted vs. saturated models. While the EM algorithm finds the local optima of the ML surface, it does not produce the derivatives in the way that Newton-Raphson maximization procedures do, and hence does not naturally produce the standard errors for the coefficients. (Also, since some of the probabilities may be estimated to be exactly 0 or exactly 1 for perfectly measured responses, the regularity conditions associated with the gradient-based methods may be violated.) In this application, we used the bootstrap standard errors to assess the uncertainty around the parameter estimates.

2. Measurement Error Analysis of the Worker Classification Survey

2.1. Data

To better document workers' understanding of issues related to job classification, Abt Associates, under contract to the Department of Labor (DOL), designed and fielded a survey to collect information about workers' knowledge about their current job classification and the rights and benefits associated with their job status. The survey featured a national, overlapping dual frame landline and cell phone RDD design. The survey design included a sample allocation of 70 percent cell phone and 30 percent landline. Abt Associates interviewers completed 2,554 interviews with respondents sampled through the landline frame and 5,949 interviews with respondents sampled through the cellular frame, for a total of 8,503 interviews. The target population was individuals age 18 and over who worked for pay or profit in the 30 days before survey administration. The survey uses weights that combine the probability of selection, sampling frame integration composite factors, non-response adjustments for people who are included on the household roster but do not complete the interview, and calibration adjustments to population controls by age, gender, education, race, Hispanic ethnicity, region, and employment status, based on ACS and CPS data. The survey included previously used questions as well as new questions developed specifically for this study.

2.2. Analysis Variables

We use LCA in the Worker Classification Survey to estimate the rate of false negative and false positive responses to survey questions addressing the following constructs of The Worker Classification questionnaire:

1) SELF: self-report of the employment status, a composite variable that determines whether or not a survey respondent received the detailed questions in the ERT; see Section X). Examples of items included in this construct:
    - At your main job, are you employed by government, by a private company, a nonprofit organization, or are you self-employed?
    - At your main job, are you an employee?
    - How certain are you that you are an EMPLOYEE on your main job and not another type of worker such as an independent contractor, or temporary worker? ("very certain" to "not at all certain")
    - Do you usually refer to your work at your main job as…? (your business, your practice, your client, your job)
2) TAX: treatment by the firm based on the tax documents the worker receives and submits (forms W-2, W-4, W-9, 1099, schedule K-1);
3) Behavioral control, based on the functions that the worker performs at the work place, and the degree of control over these functions that they exercise. CTRL1 composite is based on 10 items, and a somewhat relaxed CTRL2 is based on a subset of 7 of these items. Some of the items in the latter construct include:
    - On your main job, do you report directly to a manager, supervisor, foreman or someone else who regularly directs and controls HOW you do your work?
    - Do you determine your own schedule or the hours that you work?

- Do you need approval for your schedule or hours worked?
- Do you need permission to leave your place of work, or can you come and go at will?
- Does the nature of your work require that you provide your duties at a specific location?

4) NONCTRL: financial and employment relationships that the worker forms with the firm. Some items in the NONCTRL construct include:
    - Are you working only until a specific project is completed?
    - Were you hired for a fixed period of time?
    - On your main job, have you ever invested your own money in the company where you work to support the day-to-day operations?
    - In the event that the company where you work loses money, would you continue to earn your wage for the work you perform?
    - Besides your main job, do you perform similar paid work for others?
    - Are you required to get approval from your main job in order to provide these duties and activities for others?

The latter two groups of variables form the Economic Realities Test (ERT).

The estimated error rates can be used in a qualitative way to inform the interpretation of the survey results. E.g., there is a one-to-one relation between being an employee and receiving the form W-2, so the members of the latent class "employee" will nearly always report receiving W-2. If the model estimates the probability of choosing the "Yes" answer at 100%, then this question is free of false negatives. If, on the other hand, the model estimate is that only 98% of the members of the class report receiving a W-2, and the remaining 2% report not receiving a W-2 (but all other characteristics, determined by the patterns of other variables, point out to these respondents being proper employees), then we can conclude that this question suffers from a 2% false negative error rate. Similarly, the members of the latent class "nonemployee" are expected to never or almost never report receiving W-2, so a non-zero estimate for that item is the measure of the false positive rate.

Unlike in many other analysis, missing data in the form of "Don't know" answer is an informative response for LCA, as it is just one of the multinomial options along with "Yes" and "No". Moreover, given the topic of the survey as that of employee *knowledge* of their status, the "Don't know" response is an important insight into that knowledge.

Each latent class analysis model requires specification of the variables to be analyzed and specification of the expected number of classes. Interactions of the survey variables can also be used. For instance, if a worker is required to wear a uniform and be present onsite, this increases the likelihood that they are an employee on top of the main effects of the uniform variable and the site presence variables by themselves. As latent class models are estimated by maximum likelihood, likelihood ratio tests against the null model will be used to test model fit, and select the appropriate number of classes. Various model simplifications (e.g., assuming perfect indicators of employment that have no error, at least in some classes – e.g., a person who has never been self-employed would never report having formed their own business or LLS, which is one of the financial relations variable) can also be entertained. Additional diagnostics suggested by Biemer (2011) can also be checked.

2.3. Fitting the latent class model

Since the ERT block was only administered to those who were considered non-employees or underdetermined status, the subsample for the analysis was restricted to the 3,158 cases that were administered the ERT component. The remaining 5,345 cases were clearly classified as employees, mostly based on the TAX construct of how the firm treats them. As the analysis considers measurement error within the existing sample, our analysis uses unweighted data.

Models with two to four classes were considered. For each number of classes, the model was run 20 times with random starting values, to increase the likelihood that a globally optimal solution is found. A two-class solution was found to be stable (the same solution was reproduced in all 20 runs), however it demonstrated critical lack of fit (likelihood ratio test T=115.3, Prob[χ^2 (48)>115.3] <10^{-4}), and therefore we did not consider this model. Among the three class solutions, the best fitting one had fair fit (likelihood ratio test χ^2 p-value = 0.015), and it was found in 15 runs out of 20. The best fitting four class solution had acceptable fit (p>0.12), however it was only identified in 3 out of 20 runs, which cannot be deemed stable.

A complementary test of fit is provided by the residuals (difference between the empirical values in the non-zero cells of the saturated null model vs. the model-implied values). If the model fits accurately, the normalized residuals (i.e., residuals divided by the appropriate standard error) should demonstrate a close fit to the standard normal distribution. For the three-class model, Kolmogorov-Smirnov test based on n=65 residuals of the contingency table produced p-value p=0.2850, which was deemed satisfactory.

The analysis was performed in Stata 14, using custom coded programs to estimate latent class models via EM algorithm.

2.4. The preferred latent class model

The best fitting model contained three classes. The likelihood ratio test (against a saturated null model in which no structure on the joint five-way table of the outcome variables was imposed) was T=60.56, Prob[$\chi^2(39)$>60.56]=0.0150. Estimated coefficients are reported in Table 1.

Based on the results reported in the above table, the sample of the respondents who received the ERT section of the instrument (as they were not classified by the TAX status), we identify three groups that share common patterns of responses.

1. Class A (59% of the subsample) consists of workers who clearly identify themselves as employees (with 98.4% probability of being classified as an employee by SELF measure). However, treatment by employer is less clearcut (only 58.0% are classified as employees by TAX measure in that class). Ultimately, only 12% of the members of this class are classified as employees by the ERT Ctrl 1, and only 37.6% are classified as employees by other factors.
2. Class B (24% of the sample) clearly identifies themselves as non-employees (84.7% probability of being classified as a non-employee by SELF measure), which is largely confirmed by employer treatment (57.3% probability of being classified as non-employee by TAX measure). This class is clearly classified as non-employees by the ERT control factor variables (98.5% probability of being classified as a non-employee) and the ERT factors other than control (81.1% probability of being classified as a non-employee). For all ERT variables, the probability of being classified as an employee, while reported as positive, is insignificantly different from zero.

Table 1: Three class solution of latent class analysis.

	Class A	Class B	Class C
Class probability	0.5876	0.2426	0.1698
	(0.0216)	(0.0089)	(0.0198)
SELF == Employee	0.9842	0.0100	0.6097
	(0.0054)	(0.0074)	(0.0608)
SELF == Non-employee	0.0058	0.8473	0.0003
	(0.0029)	(0.0290)	(0.0005)
SELF == undetermined	0.0100	0.1427	0.3899
	(0.0051)	(0.0306)	(0.0607)
TAX == Employee	0.5799	0.1288	0.1701
	(0.0181)	(0.0153)	(0.0352)
TAX == Nonemployee	0.0411	0.5731	0.2195
	(0.0068)	(0.0206)	(0.0352)
TAX == undetermined	0.3790	0.2981	0.6104
	(0.0182)	(0.0168)	(0.0365)
ERT CTRL1 == NE,	0.7972	0.9849	0.9326
ERT CTRL2 == NE	(0.0102)	(0.0054)	(0.0174)
ERT CTRL1 == NE	0.0795	0.0118	0.0473
ERT CTRL2 == EE	(0.0068)	(0.0050)	(0.0141)
ERT CTRL1 == EE	0.1232	0.0033	0.0200
ERT CTRL2 == EE	(0.0086)	(0.0024)	(0.0097)
ERT NONCTRL ==	0.5244	0.8113	0.7894
Nonemployee	(0.0177)	(0.0174)	(0.0348)
ERT NONCTRL ==	0.3759	0.0031	0.0261
Employee	(0.0182)	(0.0031)	(0.0190)
ERT NONCTRL ==	0.0996	0.1856	0.1846
Undetermined	(0.0092)	(0.0176)	(0.0292)

The reported values are class probabilities (top row), and conditional probabilities of response given the class (other rows). Bootstrap standard errors in parenthesis.

3. Class C (17% of the sample) tends to think of themselves as employees (61.0% probability of being classified as employees by SELF report), but their treatment by employers is mostly unclear (61.0% probability of an undetermined TAX, i.e., the respondent does not know what tax document(s) are that they receive from the principal firm). ERT analysis tends to classify them as non-employees (93.3% probability of being classified as non-employee by CTRL variables; 78.9% of being classified as non-employee by NONCTRL factors other than control).

Thus while Class B appears to consist of nonemployees, Classes A and C are the workers in difficult or confusing situations. A fraction of those in Class A are employees, but others are not despite their thinking of themselves as such. For the members of class 1, the probabilities of being classified as an employee by the control factors are in low double digits, while probabilities of being classified as employees by TAX and NONCTRL factors other than control are in 30% (NONCTRL=employee) to 60% (TAX=employee) range. Likewise problematic are the members of Class C: they do not think of themselves as non-employees (the probability of SELF=nonemployee answer is virtually zero), yet the ERT CTRL and NONCTRL measures point that they are non-employees.

The predicted class probabilities for the 3,158 cases are shown in Figure 1. While some cases fall into class 1 or class 2 with near certainty, Class C is not as clearly defined. Respectively, 1374, 651 and no cases have predicted probabilities higher than 90% of being in classes A through C. Prediction by the highest probability/modal class gives 2067 cases (65.5%) in Class A, 759 cases (24.0%) in Class B, and 332 cases (10.5%) in Class C. Only one case was largely unresolved, with all three predicted class probabilities below 0.5.
A complementary view of the stability and interpretation of the classes and the various solutions is presented on Figure 2 based on prediction from the best fitting models. For the three classes solution described above:

- 3A (58%): self-identify as employees, treatment by principal is less clear-cut;
- 3B (24%): are non-employees, and they know it, and they are treated as such;
- 3C (17%): tend to think of themselves as employees, treatment by principal unclear, ERT points to them not being employees.

For the four class solution:

- 4A (39%): self-identify as employees, treatment by principal is less clear
- 4B (27%) self-identify as employees, treatment by principal is less clear, and they are less likely than 4A to be identified as employees by NONCTRL
- 4C (27%): mostly think of themselves as non-employees, ERT confirms
- 4D (7%): don’t know their status, treatment by principal is unclear, ERT points to them not being employees

The riverplot shows that the distinctions between classes 3B, 3C, 4C and 4D are somewhat blurred. Also, forcibly adding a fourth class to the model leads to an immaterial split of the class 3A.

3. Conclusions

Latent class analysis helped identify some of the fine structure of employment classification of the more difficult cases that were given the ERT section of the survey. About a quarter (unweighted) of respondents appear to be nonemployees by most measures. However, among the other respondents, there appears to be a conflict between their self-reporting as employees and the Economic Reality Test results. Quantification of response propensities produced by this latent class analysis demonstrated a certain degree of measurements error. This in turn calls for caution is designing, implementing, and interpreting surveys that rely on self-reports of work and employment. More accurate definitions and instruments may need to be adopted to more clearly define concepts as fluid and as confusing as employment (and especially self-employment).

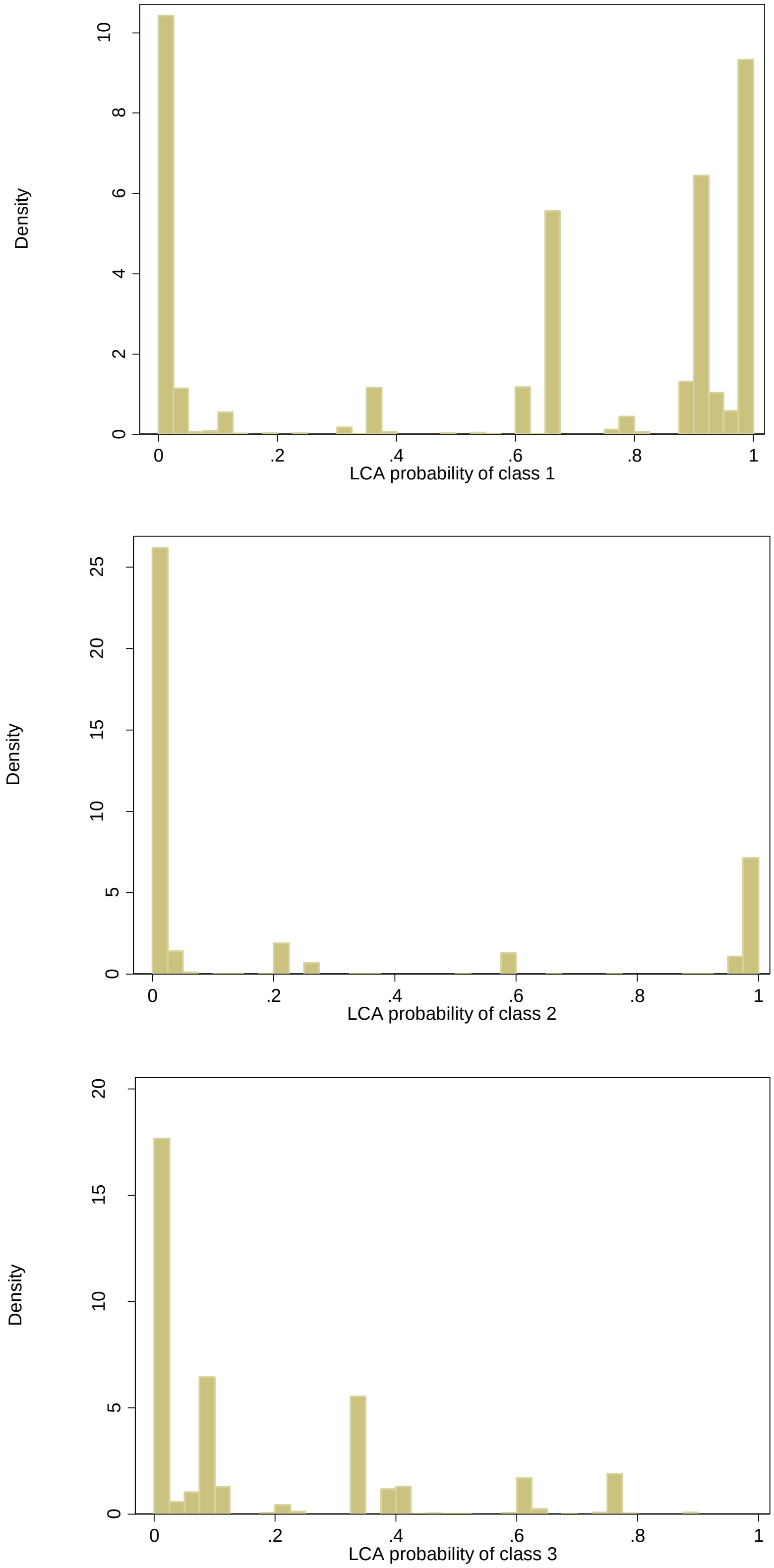

Figure 1: Predicted class probabilities for the three-class LCA solution.

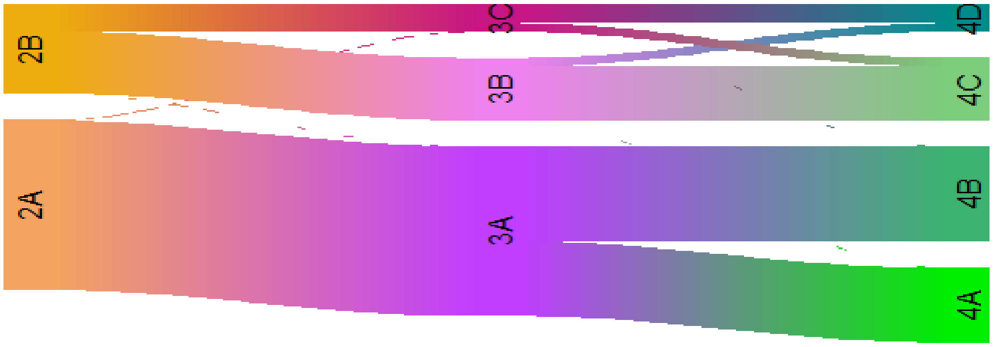

Figure 2: Relations between classes for the models with different number of classes.

References

Biemer, P. P. (2011). Latent Class Analysis of Survey Error. Wiley Series in Survey Methodology, Hoboken, NJ.

Biemer, P. P. 2004. “Modeling Measurement Error to Identify Flawed Questions.” In Methods for testing and Evaluating Survey Questions (Eds S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, and E. Singer) pp. 225-246, New York: Wiley.

Biemer , P. , and Wiesen , C. (2002), Latent class analysis of embedded repeated measurements: An application to the National Household Survey on Drug Abuse, Journal of the Royal Statistical Society, Series A , 165 (1), 97 – 119 .

Fienberg, S.E., Hersh, P., Rinaldo, A., and Zhou, Y. (2007). Maximum Likelihood Estimation in Latent Class Models for Contigency Table Data. Technical report #9-2007, Department of Statistics, Carnegie Mellon University. Available from http://www.stat.cmu.edu/tr/tr857/tr857.html.

Hagenaars J.A. & McCutcheon, A.L. (2009). Applied Latent Class Analysis. Cambridge University Press.

Hui, S.L. and Walter, S.D. 1980. “Estimating the Error Rates of Diagnostic Tests.” Biometrics, 36, 167-171.

Kreuter, F., Yan, T., Tourangeau, R. 2008. “Good Item or Bad - Can Latent Class Analysis Tell?: The Utility of Latent Class Analysis for the Evaluation of Survey Questions.” Journal of the Royal Statistical Society, Series A, 171: 1-16.

McCutcheon, A.L. 1987. Latent Class Analysis. Beverly Hills: Sage.

Tucker, C., Meekins, B., and Biemer, P. (2010). Latent Class Analysis of Consumer Expenditure Reports, in: Proceedings of the Survey Research Methods Section of the American Statistical Association, Alexandria, VA. Available from https://www.bls.gov/osmr/pdf/st100250.pdf.