A Brief History of Classification Error Models

Marcus Berzofsky[1], Paul Biemer[1], and William Kalsbeek[2]
[1]RTI International, 3040 Cornwallis Rd, Research Triangle Park, NC 27709
[2]UNC-Chapel Hill, 730 MLK Jr Blvd, CB 2400, Chapel Hill, NC 27599

Abstract

Classification error analysis aims to measure misclassification rates and to correct estimates of proportions for misclassification. We will trace development of the key models developed for this type of analysis and compare and contrast them. Of particular interest are the latent class models developed by Lazarasfeld and Henry (1968), the Census Bureau Model developed by Hansen, Hurwitz, and Pritzker (1964), and the finite mixture probability models developed by Bross (1954), Tenenbein (1972), Hui and Walter (1980) and others. This paper will show how these three primary methods can be viewed from a broader log linear model with latent variables perspective and how under that framework classification error can be better specified through greater model flexibility.

Key Words: latent class analysis, classification error models, measurement error, reliability analysis, Census Bureau Model, finite fixture models, Hui-Walter Method

1. Introduction

Measurement error is defined as the difference between the true value of a measurement and the value obtained during the measurement process (Lessler & Kalsbeek, 1992). Classification error is a type of measurement error by which the respondent does not provide a true response to a survey item. For nominal categorical data this can occur in one of two ways: a false negative response or a false positive response. A false negative response occurs when a respondent indicates an event did not occur when it did. A false positive response occurs when a respondent indicates an event did occur when it really did not.

Sudman and Bradburn (1974) write that the conceptual framework of measurement error comes from one of three sources: the task to be accomplished, the interviewer, or the respondent. The task to be accomplished includes the mode of the interview (e.g., in person or via telephone) and the questionnaire itself. In terms of the mode, Sudman and Bradburn (1974) point out that the nature of the interview (e.g., sensitive subject matter) may impact how a respondent answers when talking directly to an interviewer as opposed to answering an unseen person or, in the case of audio computer assisted self interviews (ACASI), a computer. The questionnaire impacts measurement error by how it is structured. Sudman and Bradburn (1974) indicate that measurement error will be smaller if the questionnaire has greater structure and is laid out clearly. The second source of measurement error is the interviewer. Whether the interviewer is required to follow a strict script or is allowed leeway in obtaining responses impacts the level of measurement error. The third source of measurement error is the respondent. Sudman and Bradburn (1974) state that the motivation of the respondent plays a major role in the quality of the data provided.

As a type of non-sampling error, measurement error cannot be corrected for during the design portion of the study. Therefore, it is important that a survey analyst quantifies the impact of measurement error on survey estimates prior to drawing any conclusions. Lohr (1999, pp. 9–10) outlines eight ways in which a classification error may be induced that fit into the Sudman and Bradburn (1974) framework:

- People sometimes do not tell the truth
- People do not always understand the questions
- People forget (e.g., telescoping or recall bias)
- People give different answers to different interviewers

- People may say what they think an interviewer wants to hear or what they think will impress the interviewer
- A particular interviewer may affect the accuracy of the response by misreading questions, recording responses inaccurately, or antagonizing the respondent
- Certain words mean different things to different people
- Question wording and order have a large effect on responses obtained

Recognizing the importance of needing to quantify the impact that measurement error, and more specifically classification error, can have on survey estimates, three separate groups independently developed methods to address classification error. Psychometricians developed latent class analysis (LCA) for latent variables, government survey statisticians developed the Census Bureau Model which uses reliability analysis, and biostatisticians and epidemiologists developed methods using finite fixture probability models. This paper will discuss each of these approaches and then discuss latent class analysis for measuring classification error, a more recent approach that combines elements of all three approaches. The paper then presents some examples of how LCA for measuring classification error has been used in practice.

2. Latent class analysis for latent variable analysis

2.1 Background

LCA for latent variable analysis was developed by two psychometricians, Lazarasfeld and Henry (1968). Their goal was to use manifest variables found in a survey to develop a definition for an unobserved latent variable and determine how respondents should be categorized. For example, an analyst may want to create categories for a socioeconomic status indicator. While this status was not directly measured, the analyst may use a combination of manifest variables such as income level, age, and education status. Using these manifest variables, LCA helps the analyst determine the number of categories that socioeconomic status should be split into and how respondents should be organized into each category.

2.2 How it works

LCA is a modeling technique that consists of two components: a structural component and a measurement component (Bassi, Hagenaars, Croon, and Vermunt, 2000). The structural component consists of the latent variables and how they are defined and the measurement component consists of the manifest variables conditioned on the latent variables. Based on this structure, a model with three manifest variables (indicated by A, B and C) representing one latent variable (indicated by X) is written as

$$\pi_{abc}^{ABC} = \sum_{x} \pi_{x}^{X} \pi_{a|x}^{A|X} \pi_{b|x}^{B|X} \pi_{c|x}^{C|X} \quad (1)$$

In this model, π_{x}^{X} is the structural component where π_{x}^{X} represents the probability of being classified into category X=x and $\pi_{a|x}^{A|X}\pi_{b|x}^{B|X}\pi_{c|x}^{C|X}$ is the measurement component of the model where $\pi_{a|x}^{A|X}$ is the classification probability for manifest variable A. $\pi_{b|x}^{B|X}$ and $\pi_{c|x}^{C|X}$ have similar definitions. Under a dichotomous definition for A and X, $\pi_{2|1}^{A|X}$ represents the false negative rate for manifest variable A and $\pi_{1|2}^{A|X}$ represents the false positive rate for manifest variable A.

When Lazarasfeld and Henry (1968) developed this method they were primarily interested in the structural component of the model since it is the structural component that indicates how the latent variable of interest should be defined. Thus, while recognizing the contribution of the measurement component and the resulting error rates, Lazarasfeld and Henry did not focus on their model estimates. However, as we will show in Section 5, the measurement component of this model is a very powerful tool in quantifying the classification error rates.

3. Census Bureau Model

3.1 Background

The Census Bureau Model, also called reliability analysis, was developed by Hansen, Hurwitz, and Pritzker (1964) with the goal of determining how reliable an estimate is. Instead of looking at error rates, the Census Bureau Model decomposes the variance of an estimate into simple random variance (SRV), the variation due to measurement error, and sampling variance (SV), the variation due to random sampling. The Census Bureau Model is most commonly used in an interview-reinterview setting where an initial interview is followed-up by a second interview after a reasonably amount of time (several weeks to a few months) where several of the key questions asked in the first interview are repeated.

3.2 How it works

The Census Bureau model is based on two stage cluster sampling where the first stage is the selection of individuals and the second stage is repeated measurements on an item from each individual (Biemer, 2004). This approach measures the number of individuals that would be classified in the affirmative by the survey process by determining the probability, for each individual, of responding affirmatively over several repeated measurements of an item. This process assumes that each repeated measurement is independent of all prior measurements.

Through the repeated measurements, the Census Bureau Model is able to determine both SRV and SV where the sum of these two variance components equals the total variance. Using these variance components the Census Bureau Model defines the index of inconsistency (I) as

$$I = \frac{SRV}{SV + SRV} \tag{2}$$

In other words, the index of inconsistency is the proportion of variance due to simple random variance and, therefore, measurement error. The reliability ratio, R, defined as $R = 1 - I$, represents the proportion of variance due solely to sampling variation. The Census Bureau (1985) published guidelines for interpreting the index of inconsistency. They advised the following rule of thumb: $0 \le I \le 0.2$ is good, $0.2 < I \le 0.5$ is moderate, and $I > 0.5$ is poor.

4. Finite fixture models

4.1 Background

A frequent goal for epidemiologists and biostatisticians is to determine if a new method for testing a disease, presumably faster and/or less expensive to implement, is as effective as the existing method where the existing method is considered a gold standard. Bross (1954) developed a probability model for two binomial tests and Tenenbein (1972) extended this model for comparing methods that follow a multinomial distribution. Hui and Walter (1980) extended the finite fixture model approach further so that neither method being compared needed to be a gold standard.

4.2 How it works

The finite fixture model approach treats the method deemed the gold standard as truth and, therefore, without error and the other method as fallible and with error. Using the joint distribution of the two methods, Bross (1954) showed that the false negative rate, θ, and the false positive rate, ϕ, can be calculated directly. Figure 1 illustrates the case where the two methods each have dichotomous outcomes where p_{11} is the probability of both methods having an affirmative agreement, p_{22} is the probability of both methods having a negative agreement, and p_{12} and p_{21} being the probabilities of disagreement.

		Item B (Gold Standard)	
Item A (Fallible)		Yes = 1	No = 2
	Yes = 1	p_{11}	p_{12}
	No = 2	p_{21}	p_{22}

Figure 1. Finite Fixture Model Method

Under this design, Bross (1954) showed that if π is the probability of an affirmative response under the gold standard that θ and ϕ are defined as

$$\theta = \Pr\left[A = 2 \mid B = 1\right] = \frac{p_{21}}{p_{11} + p_{21}} \tag{3}$$

and

$$\phi = \Pr[A = 1 \mid B = 2] = \frac{p_{12}}{p_{12} + p_{22}} \tag{4}$$

Given these, Bross (1954) proved that probability of an affirmative response to the fallible method, p, was a function of π, the true prevalence rate from the gold standard method, θ, and ϕ. Namely,

$$p = (1-\theta)\pi + \phi(1-\pi) \tag{5}$$

Thus, the bias between the gold standard method and the fallible method is a function of the joint probabilities where

$$\text{Bias} = p - \pi = -\theta\pi + \phi(1-\pi) = p_{12} - p_{21} \tag{6}$$

Hui and Walter (1980) derived a closed form solution for the maximum likelihood estimates of π, θ, and ϕ. In doing this, the Hui-Walter method does not require the assumption that one of the measurements is a gold standard without error. However, Hui and Walter note that, in general, for R tests applied to S populations, there are $(2^R - 1)S$ degrees of freedom for estimating $(2R+1)S$ parameters. Therefore, in the important case of two indicators from a single population, there are 5 parameters to be estimated (i.e., the true prevalence, the false negative rate from each item, and the false positive rate from each item) with only three degrees of freedom from which to make the estimates. Therefore, the model is deemed not identifiable (Goodman, 1974), and estimates cannot be calculated for the parameters of interest.

Thus, in order to derive the maximum likelihood estimates for two measurements, additional constraints must be placed on the model parameters for it to be identifiable. To achieve this, an analyst has several possible options available. For example, in some circumstances, it may make sense to set the false positive rate for each measurement equal to 0. This constraint may make sense when analyzing the probability of using certain illegal drugs. It this situation, one might be willing to assume that it is highly unlikely that a respondent would indicate using these drugs when they are not. Another possibility is to set the false negative rate equal to the false positive rate for each measurement. This constraint may make sense in a study in which, for each item, it is just as likely for a respondent to misclassify themselves in the affirmative as it is to misclassify themselves in the negative. The Hui-Walter (1980) method does not require constraints on the false negative or false positive rates. Instead, Hui and Walter determined that if the population is split by a two level grouping variable (e.g., sex, age, etc.) and assumes 1) the true prevalence estimate is different across groups, and 2) the classification error rates for a measurement are the same across groups then a saturated model estimating six parameters (i.e., $\pi_{x|1}^{X|G}, \pi_{x|2}^{X|G}, \theta_A, \theta_B, \phi_A,$ and ϕ_B) with six degrees of freedom is obtained.

5. Latent class analysis for measuring classification error

5.1 Bringing the three methods together

Over the past twenty years, statisticians have worked to combine elements from each of the three methods described into a single comprehensive method for measuring classification error. This method is primarily based on LCA for latent variable analysis (Lazarasfeld and Henry, 1968), but has a couple of key differences (see, for example, Hagenaars, 1993; Biemer, 2004). Specifically, like finite fixture models the outcome being modeled has a clear predetermined definition and the focus of the model is on the measurement component and not the structural component. Thus, in Latent class analysis for measuring classification error there is a predefined latent variable, which is often dichotomous, measured by indicators in the survey. An indicator is a highly correlated manifest variable defined as follows: Let X and Y be dichotomous latent variables and A be a manifest variable. Then, A is an indicator for X if for any other dichotomous latent variable Y

$$P(A = c \mid X = C) \geq P(A = C \mid Y = c) \text{ for c=1,2} \tag{7}$$

For instance, if the latent variable being measured is whether one used marijuana in the past 12 months then an example of a good indicator is a survey item asking if one used marijuana in the past 12 months.

Furthermore, LCA for measuring classification error can be viewed through a log-linear framework (see Vermunt, 1997). This allows greater flexibility and ease in model selection and determining model fit because it allows an analyst the ability to use the same model selection techniques used for log-linear models without latent variables. Under this framework, several software packages such as LEM (Vermunt, 1997), Latent Gold (Vermunt and Magidson, 2005), and M-plus (Munthen and Munthen, 1998 – 2005) have been developed to conduct LCA.

LCA for measuring classification error has four key assumptions when three or more indicators are used. Specifically,

- Univocality: Indicators measure the same latent variables. In other words, two dichotomous latent variables X and Y are univocal if:
$$\text{Corr}(X, Y) = \rho_{XY} = 1 \tag{8}$$
- Local independence: Error rates between indicators are independent of one another. In other words, two indicators A and B for the dichotomous latent variable X are locally independent if
$$\text{Corr}(A, B \mid X) = \rho_{AB|X} = 0 \tag{9}$$
- Homogeneity within group: Error rates for an indicator are the same for all persons in a group. So, if A is an indicator for latent variable X and H is a grouping variable define $\lambda_i(a, x \mid H) = P(A_i = a \mid X_i = x, H = g)$. There is group homogeneity when $\lambda_i(a, x \mid H) = \lambda(a, x \mid H)$ for all *i* subjects in group H. Therefore, if a hidden grouping variable H creates group homogeneity then an observed or manifest grouping variable G creates group homogeneity if
$$\text{Corr}(G, H) = \rho_{GH} = 1 \tag{10}$$
- Simple random sample: Respondents were selected by a simple random sample

When only two indicators are available one can use the Hui-Walter method and its two additional assumptions. The cell probabilities for an LCA model with three indicators (A, B, and C) for one dichotomous latent variable (X) and a grouping variable (G), such as gender or age, that meets these assumptions can be defined as

$$\pi_{gabc}^{GABC} = \pi_g^G \sum_x \pi_{x|g}^{X|G} \pi_{a|gx}^{A|GX} \pi_{b|gx}^{B|GX} \pi_{c|gx}^{C|GX} \tag{11}$$

While the Census Bureau Model focuses on the variance components of an estimate it too can be viewed through the LCA for measurement error framework via a latent agreement model (Guggenmoos-Holzmann and Vonk, 1998). Like the Census Bureau Model, the latent agreement model looks at intra-observer agreement. To accomplish this it splits

respondents into two groups: those easily classified (conclusive) and those difficult to classify (inconclusive). The classification status is considered a latent variable and modeled using the LCA framework.

5.2 Examples of LCA for measuring classification error

LCA for measuring classification error has been used to determine the quality of estimates on several surveys. In this paper we will illustrate three such examples.

- Sinclair and Gastwirth (1996) and Biemer and Bushery (2001) used the Hui-Walter method to measure error rates in employment status reported in the Current Population Survey. They both found that the error rates for the categories employed and not in the labor force were relatively small, but the error rates for unemployed were high indicating that respondents are least comfortable admitting to being unemployed.
- Biemer and Wiesen (2002) looked at measurement error in the National Survey for Drug Use in Households for the estimation of marijuana use in the past 12 months. They found that the wording of sensitive questions can make a dramatic difference in how one responds. For example, they found that respondents would admit to the last time they used marijuana in a frequency question, but when asked if they were a 'marijuana user' they reported that they were not. Biemer and Wiesen surmised that respondents who were occasional users of marijuana objected to the term 'marijuana user' because it implied a higher and more frequent level of usage.
- Kruter, Yan, and Torangeau (2008) looked to see if LCA could be used to determine the quality of survey instrument items to correctly measure the latent variable of interest. The authors imbedded an item they knew to be inferior for measuring the latent variable they were interested in and they wanted to see if LCA correctly determined that this item was inferior. Their findings found that LCA will correctly identify the inferior indicator when all of the LCA assumptions are fully met. When some of the assumptions are not met then the LCA results may not be as expected.

References

Biemer, P. (2004). Simple response variance: then and now. *Journal of Official Statistics, 20*, 417–439.

Biemer, P., & Bushery, J. (2001). Application of Markov latent class analysis to the CPS. *Survey Methodology, 26*(2), 136–152.

Biemer, P., & 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.

Bross, I. (1954). Misclassification in 2 X 2 tables. *Biometrics, 10*, 478–486.

Guggenmoos-Holzmann, I., & Vonk, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Statistics in Medicine, 17*, 797–812.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Newbury Park, CA: Sage Publications.

Hansen, M., Hurwitz, W. N., & Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. In C. R. Rao (Ed.), *Contributions to statistics* (pp. 111–136). Calcutta: Pergamon Press, Ltd.

Hui, S. L., & Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics, 36*, 167–171.

Kreuter, F., Yan, T., and 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*(3), 723–738.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys*. New York: Wiley.

Lohr, S. L. (1999). *Sampling: Design and analysis*. New York: Duxbury Press.

Muthen, L. K. and B.O. Muthen (1998-2005). *Mplus*. Los Angeles, CA: Muthen & Muthen.

Sinclair, M., & Gastwirth, J. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: Application to labor force data. *Journal of the American Statistical Association, 91*, 961–969.

Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis*. Chicago: Aldine.

Tenenbein, A. (1972). A double sampling scheme for estimating from misclassified multinomial data with application to sampling inspection. *Technometrics, 14*(1), 187–202.

U.S. Census Bureau. (1985). *Evaluating censuses of population and housing* (STD-ISP-TR-5). Washington, DC: U.S. Government Printing Office.

Vermunt, J. (1997). *Log-linear models for event histories*. Newbury Park, CA: SAGE Publications

Vermunt, J.K (1997). LEM: A General Program for the Analysis of Categorical Data. Tilburg Netherlands: Department of Methodology and Statistics, Tilburg, Netherlands.

Vermunt, J.K. and Magidson, J (2005). Technical Guide to Latent Gold 4.0: Basic and Advanced. Belmont, MA: Statistical Innovations.