Published in final edited form as:

Biom J. 2013 July ; 55(4): 541–553. doi:10.1002/bimj.201200020.

Latent class regression: inference and estimation with two-stage multiple imputation

Ofer Harel^{1,*}, Hwan Chung², and Diana Migliorette³

¹Department of Statistics, 215 Glenbrook Rd., Unit 4120, University of Connecticut, Storrs, CT, USA

²Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul 136-701, Korea

³Department of Public Health Sciences, UC Davis School of Medicine, One Shields Ave., Med Sci 1C, Room 144, Davis, CA 95616, USA

Abstract

Latent class regression (LCR) is a popular method for analyzing multiple categorical outcomes. While non-response to the manifest items is a common complication, inferences of LCR can be evaluated using maximum likelihood, multiple imputation, and two-stage multiple imputation. Under similar missing data assumptions, the estimates and variances from all three procedures are quite close. However, multiple imputation and two-stage multiple imputation can provide additional information: estimates for the rates of missing information. The methodology is illustrated using an example from a study on racial and ethnic disparities in breast cancer severity.

Keywords

Latent class regression; Missing data; Missing information; Multiple imputation

1 Introduction

Latent class analysis (LCA) (Goodman, 1974; Clogg & Goodman, 1984) explains the relationships among categorical variables in a cross-classified contingency table by positing the existence of an unobserved classifier. Extensions to the traditional LCA allow covariates to predict class membership (Dayton & Macready, 1988; Bandeen-Roche *et al.*, 1997). Latent class regression (LCR) allows the latent class prevalences to vary with covariates, but the meaning of the latent class is still determined by the manifest items. Using this framework, LCR can examine influences of covariates on the marginal distribution of latent class memberships through a binary or polytomous regression.

In many cases, data sets include missing values, leading to several complications. Exclusion of individuals with missing data (case deletion) makes specific assumptions about the missing data mechanism and can result in bias and/or loss of efficiency. In LCA, the latent class membership of each individual is always missing. However, there may also be missing values for the manifest variables that measure the latent classes. Most procedures do not distinguish between these two types of missing values. By ignoring the distinction between

Conflict of Interest

The authors have declared no conflict of interest.

the missing value types, LCR parameters can be estimated by maximum likelihood (ML) using the EM algorithm (Dempster *et al.*, 1977; Agresti & Lang, 1993). Alternatively, one can use a simulation-based technique such as Markov chain Monte Carlo (MCMC) (Gilks *et al.*, 1998; Brémaud & Bremaud, 1999; Liu, 2001) or carry out inference by multiple imputation (MI) (Rubin, 1987; Schafer, 1997; Harel & Zhou, 2007). Harel & Miglioretti (2007) demonstrated the use of MI when there are no missing manifest variables.

We propose a two-stage MI procedure for LCA/LCR to separate the missing information into two distinct types: missing information due to (i) the unobserved latent class variable (measurement error) and (2) non-response on manifest variables. This separation will render further understanding of the missing information of the model, and how the uncertainty in the model is separated between measurement error and non-response. Two-stage MI consists of two imputation steps: we first impute m sets of complete manifest variables, then we impute k sets of latent class memberships, given the imputed items. These two steps take into account the missing information caused by missing items and measurement error, respectively.

In Section 2 we discuss missing data and their assumptions, while in section 3, we describe LCR and present the standard estimation algorithm including EM and MI. As an alternative to EM and MI, two-stage MI is described in Section 4. We compare the results from three types of estimation methods (ML, MI, two-stage MI) through a breast cancer study in Section 5. Concluding discussion and directions for future research are presented in Section 6.

2 Missing data and missing data assumptions

Missing data is a common complication in applied research. While most statistical software would use complete case analysis as a default, it has been shown that ignoring the missing data (or treating it improperly) can lead to biased results, inefficient parameter estimates and misleading inferences (e.g. Little & Rubin, 2002; Imbens & Manski, 2004; Harel & Zhou, 2007; Harel *et al.*, 2012). There is considerable amount of research which led to better procedures to deal with incomplete data such as maximum likelihood (Little & Rubin, 2002), Bayesian analysis (Gelman *et al.*, 2003) and multiple imputation (Rubin, 1987) to mention a few. All missing data procedures rely on some missing data assumptions which relate the observed and unobserved data. In general, one can separate the data Y into observed part Y_{obs} and missing part Y_{mis} .

2.1 Missing data assumptions

Missing data techniques depend on the nature of the missing data mechanism. Little & Rubin (2002) defined the terms missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR refers to scenarios in which the outcomes are independent of the mechanism governing the missingness; under MAR the outcomes may depend on the mechanism governing the missingness but only through the observed measurements. If MAR does not hold, there is dependence of the mechanism governing the missingness with unobserved outcomes. MAR is the weakest, most general condition which allows ignoring the modeling of the mechanism governing the missingness. Harel & Schafer (2009) extended these notions for cases with two types of missing values.

In this set-up, the missing values $Y_{mis} = (Y_{mis}^A, Y_{mis}^{B|A})$ are further separated into two parts.

In particular, case deletion analysis will be appropriate only under the MCAR assumption. In this manuscript, MAR is assumed for ML and MI procedures. While for two-stage MI, we will assume extended ignorability (Harel & Schafer, 2009; Harel, 2009b) in our models,

which imply that both mechanisms (governing Y_{mis}^A and $Y_{mis}^{B|A}$) can be ignored. Extensions for other missing data assumptions are discussed in Harel & Schafer (2009).

3 Latent class regression and estimation algorithms

3.1 Latent class regression model

The basic idea of LCA is that associations among manifest items can be explained by the class membership. In other words, manifest items are conditionally independent given class membership. This assumption is referred to as local independence by Lazarsfeld & Henry (1968), and is a crucial feature that allows one to draw inferences about the latent classes. Dayton & Macready (1988) extended LCA to incorporate covariates, allowing them to influence the marginal distribution of class membership through a generalized-logit link function. They computed ML estimates via computationally intensive simplex method, whereas Bandeen-Roche *et al.* (1997) applied an EM algorithm.

Let $\mathbf{Y} = (Y_1, \dots, Y_D)$ be D discrete manifest items measuring latent classes, where variable Y_d takes possible values from 1 to r_d . We denote the latent class variable by L with membership ranging from 1 to C . The joint probability that an individual belongs to class l and provides responses $\mathbf{y} = (y_1, \dots, y_D)$ would be

$$\begin{aligned} \Pr(\mathbf{Y}=\mathbf{y}, L=l) &= \Pr(L=l) \prod_{d=1}^D \Pr(Y_d=y_d|L=l) \\ &= \gamma_l \prod_{d=1}^D \prod_{q=1}^{r_d} \rho_{dq|l}^{I(y_d=q)}, \end{aligned} \quad (1)$$

where $I(y_d=q)$ is the indicator function which has the value 1 if y_d is equal to q and 0 otherwise. In (1), the following two sets of parameters are estimated: (1) $\rho_{dq|l} = \Pr(Y_d=q|L=l)$ represents the probability of the response q to the d th item for a given class l , and (2) γ_l represents the probability of belonging to the class l . As the variable L is latent, the marginal probability of a particular response pattern $\mathbf{y} = (y_1, \dots, y_D)$ without regard for the unobserved class membership is

$$\Pr(\mathbf{Y}=\mathbf{y}) = \sum_{l=1}^C \gamma_l \prod_{d=1}^D \prod_{q=1}^{r_d} \rho_{dq|l}^{I(y_d=q)}.$$

Here, we have assumed local independence—that is, the items $Y_1, \dots, Y_d$ are assumed to be unrelated within each class. This assumption is the crucial feature of LCA that allows us to draw inferences about the unobserved class variable.

To specify an LCR, we incorporate the p vector of covariates, $\mathbf{x} = (x_1, \dots, x_p)$, allowing them to influence the prevalence of class membership γ_l through a logistic regression. However, measurement parameters (i.e., ρ -parameters) are not allowed to vary along with these covariates. Then, the likelihood of LCR can be presented as

$$\Pr(\mathbf{Y}=\mathbf{y}|\mathbf{x}) = \sum_{l=1}^C \gamma_l(\mathbf{x}) \prod_{d=1}^D \prod_{q=1}^{r_d} \rho_{dq|l}^{I(y_d=q)}. \quad (2)$$

The probability that the individual has class membership l is specified by the generalized-logit link function given as

$$\gamma_l(\mathbf{x}) = \Pr(L=l|\mathbf{x}) = \frac{\exp\{\mathbf{x}'\beta_l\}}{\sum_{j=1}^C \exp\{\mathbf{x}'\beta_j\}},$$

where $\beta_l = (\beta_{l1}, \dots, \beta_{lp})'$ for $l = 1, \dots, C-1$ is a $p \times 1$ vector of regression coefficients influencing log-odds that an individual falls into class l relative to the baseline class C (i.e., $\beta_C = \mathbf{0}$).

3.2 EM algorithm

The most common approach to estimate unknown parameters in an LCR is ML estimation using the EM algorithm (Bandein-Roche *et al.*, 1997; Dempster *et al.*, 1977). EM is an iterative procedure in which each iteration consists of two steps, the E-step (expectation) and M-step (maximization). Iterating between these two steps produces a sequence of parameter estimates that converges reliably to a local or global maximum of the likelihood function. The E-step computes the conditional probability that the i th individual belongs to class l given $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ and $\mathbf{y}_i = (y_{i1}, \dots, y_{iD})$,

$$\begin{aligned} \theta_{il} &= \Pr(L=l|\mathbf{X}=\mathbf{x}_i, \mathbf{Y}=\mathbf{y}_i) \\ &= \frac{\gamma_l(\mathbf{x}_i) \prod_{d=1}^D \prod_{q=1}^{r_d} \rho_{dq|l}^{I(y_{id}=q)}}{\sum_{j=1}^C \gamma_j(\mathbf{x}_i) \prod_{d=1}^D \prod_{q=1}^{r_d} \rho_{dq|j}^{I(y_{id}=q)}}. \end{aligned} \quad (3)$$

This conditional probability can be obtained from (1) and (2) using provisional estimates provided by the previous iteration. The M-step maximizes the complete data likelihood (i.e. the likelihood for the cross-classification by l and $\mathbf{y}_i$) with respect to θ_{il} and $\rho_{dq|l}$ -parameters. In particular, when θ_{il} is known, updates for $\rho_{dq|l}$ -parameters can be calculated by the standard Newton-Raphson method for multinomial regression, provided that the computational routines allow fractional responses rather than integer counts. The item-response probabilities can be interpreted as parameters in a multinomial distribution when θ_{il} is known, so we have

$$\hat{\rho}_{dq|l} = \frac{\sum_{i=1}^n \theta_{il} I(y_{id}=q)}{\sum_{i=1}^n \theta_{il}}.$$

3.3 Multiple imputation via MCMC sampling

A Bayesian method for LCR has been implemented as an iterative two-step (I-step and P-step) procedure which can be regarded as an extended form of a data augmentation (Tanner & Wong, 1987), where steps of a Metropolis algorithm are embedded for $\rho_{dq|l}$ -parameters into the Gibbs sampler (Chung *et al.*, 2006; Robert & Casella, 2004). In the I-step, given current simulated parameter values, we calculate the posterior probabilities of class membership given in (3). Then we draw the class membership for the i th individual from a multinomial(1, $\theta_{i1}, \dots, \theta_{iC}$). For a sample of n observations, class membership should be drawn independently for $i = 1, \dots, n$. Once class memberships have been imputed, the augmented data likelihood factors into independent likelihood functions for the β_l and $\rho_{dq|l}$ -parameters. In the P-step, we draw new random values for β_l 's and $\rho_{dq|l}$'s independently. Applying a Jeffreys prior to β_l -parameter, new random values for $\rho_{dq|l}$'s are drawn from a Dirichlet distribution.

Using an improper uniform prior for β -parameters, we generate β 's by the Metropolis algorithm. Computational details for LCR parameter simulation via MCMC are described in Chung *et al.* (2006).

In a typical application of MCMC, it is common to summarize the values of simulated parameters by computing long-run averages: after the burn-in period, averaging the output stream of simulated parameters produces estimates for the posterior means and variance (Tierney, 1994). As an alternative to averaging over the simulated parameters, one could also carry out inference about parameters of interest through MI (Rubin, 1987). In MI for LCR, we retain the m sets of class variables $I_1, \dots, I_m$ and missing values for the items $y_1, \dots, y_m$ from the I-steps, spacing them enough cycles apart to ensure that they are essentially independent. Treating these imputed values as known, we compute point and variance estimates for the β 's and γ 's by standard complete-data methods for proportions and regression coefficients. The results are then combined by using Rubin's rules (Rubin, 1987, p.76). There are several advantages for the use of MI over ML. First, using MI, it is very simple to get the standard errors of the parameter estimates, while it is more complex using ML procedures. Second, the rates of missing information are useful diagnostic tool for LCR and are produced automatically using MI (Harel & Miglioretti, 2007). Lastly, using MI, one can use different models for the imputation and analysis stages, allowing adding important variables to the imputation model even though they are not of interest in the analysis model. Therefore, we chose to use these advantages and introduce the use of two-stage MI to LCR. Computational methods for generating MI under some useful multivariate models are described by Schafer (1997).

4 Two-stage multiple imputation

4.1 Methods

Two-stage MI (Harel, 2009b; Kinney & Reiter, 2009) is an extension to the nested MI proposed by Shen (2000) and used by Rubin (2003). Similarly to any missing data procedure, the main purpose of the procedure is to estimate and make inference regarding a population quantity Q . Shen (2000) developed this procedure for computation convenience. Harel (2009b) demonstrated that separating the missing values into two types allows us to assess the different contribution of each missing data type to the overall uncertainty of the inference. Harel & Schafer (2009) also introduced extended ignorability conditions which allow both ignorable and non-ignorable missing data processes. Different missing data types, their factorizations, and the extended ignorability conditions are described in more detail in Harel (2009b).

When applying two-stage MI, we are interested in estimating the population quantity Q , while separating the missing values (Y_{mis}) into two types, Y_{mis}^A and Y_{mis}^B . In our case, we are interested in estimating class probabilities and regression coefficients (of the LC regression) while the two types of missing values are the missing class identifiers and missing values for the manifest items. Since the class membership is missing with probability 1, it can be considered as MCAR. Additionally assuming that the missing manifest items are MAR leads to an ignorable model (Harel & Schafer, 2009). Under this ignorability condition, the imputation stage of two-stage MI consists of two steps. In the first stage, we draw m independent values of Y_{mis}^A from their predictive distribution,

$$Y_{mis}^{A(i)} \sim \Pr(Y_{mis}^A | Y_{obs}) \text{ for } i=1, 2, \dots, m.$$

We then draw k conditionally independent values of Y_{mis}^B for each imputation of Y_{mis}^A ,

$$Y_{mis}^{B(i,j)} \sim \Pr(Y_{mis}^B | Y_{obs}, Y_{mis}^{A(i)}) \text{ for } j=1, 2, \dots, k.$$

Applying two-stage MI to LCR, we partition the missing information into two types: one caused by latent class variable and the other caused by non-response to the manifest items. We also can estimate the same population parameters as in Section 3 using the two-stage MI for the LCR. In the first step, like conventional MI, we first draw m sets of the class variable $I_1, \dots, I_n$ and missing values for items $y_1, \dots, y_n$ from the I-steps. We then impute the class variable and retain another $k-1$ sets of $I_1, \dots, I_n$ using each of the complete items from the first step. A total of mk imputations of the class variable will be generated, which will reduce to the conventional MI when $k=1$. Similarly to the conventional MI, analyzing each data set by a complete-data method will result in $N=mk$ sets of point estimates and their standard errors. In addition, this method also provides the rates of two different types of missing information, and the sum of those two is equivalent to the rate of missing information provided by conventional MI.

The two-stage MI combination rule, an extension to Rubin's rules (Rubin, 1987) for the conventional MI, may follow Shen's rules (Shen, 2000) to calculate final parameter estimates and missing information taking into account the variability due to the two kinds of missing data. Given $N=mk$ imputations, we can calculate $N=mk$ different sets estimates of model parameters and their variances. In our case, the parameters are the β 's and γ 's as specified in (2) in Section 3. Let us refer to the parameter estimates as Q and their respective variance estimates as $\hat{U}$. The assumption involved is that, with complete data, p intervals and tests would be based on a normal approximation. That is, $(\hat{Q}-Q)/\sqrt{\hat{U}} \sim N(0, 1)$. The final point estimates based on two-stage MI is the average of the $N=mk$ complete-data estimates,

$\bar{Q}_{..} = \frac{1}{mk} \sum_{i=1}^m \sum_{j=1}^k \hat{Q}_{ij}$, where i and j index the complete-data sets. The uncertainty in Q arises from three components. The estimated complete-data variance of the average of the

mk complete-data variances calculated within each data set, $\bar{U}_{..} = \frac{1}{mk} \sum_{i=1}^m \sum_{j=1}^k \hat{U}_{ij}$. The between-nested-imputation variance, B , is the variance across the nested imputations,

$B = \frac{1}{m-1} \sum_{i=1}^m (\bar{Q}_{i.} - \bar{Q}_{..})^2$, where $\bar{Q}_{i.} = \frac{1}{k} \sum_{j=1}^k \hat{Q}_{ij}$. The within-nested-imputation variance, W , is the variance within the nested imputations,

$W = \frac{1}{m(k-1)} \sum_{i=1}^m \sum_{j=1}^k (\hat{Q}_{ij} - \bar{Q}_{i.})^2$. Summing those three components and adjusting for correction factors derived by Shen (2000), provides the total variance, T , defined as $T = \bar{U}_{..} + (1-k^{-1})W + (1+m^{-1})B$. We perform inferences about Q by assuming that $T^{-1/2}(Q - \bar{Q}_{..})$ is approximately t -distributed with the estimated degrees of freedom ν , where

$\hat{\nu}^{-1} = \frac{1}{m(k-1)} \left(\frac{(1-1/k)W}{T} \right)^2 + \frac{1}{m-1} \left(\frac{(1+1/m)B}{T} \right)^2$ (Shen, 2000). The estimated overall population rate

of missing information is $\hat{\lambda} = \frac{B + (1-k^{-1})W}{\bar{U}_{..} + B + (1-k^{-1})W}$. If non-response in the manifest items (i.e., Y_{mis}^A) were observed, then the between-nested-imputation variance would vanish. Therefore, the estimated rate of missing information due to the latent class variable (i.e., Y_{mis}^B) given the complete manifest items is $\hat{\lambda}^{B|A} = \frac{W}{\bar{U}_{..} + W}$. The difference between these rates, $A = \hat{\lambda} - \hat{\lambda}^{B|A}$, represents the amount by which the rate of missing information would drop if the manifest items were complete.

The rates of missing information can be unstable at times (Rubin, 1987; Schafer, 1997). To keep these estimates stable, Harel (2007) developed the method for the choice of the number of imputations using the asymptotic distributions of the rates of missing information and control for the amount of variability wanted in the estimation procedure. When using MI or two-stage MI in our study, we chose the number of imputations based on this method in Harel (2007). In addition, a number of methodologies have been proposed to measure the impact of a missing value or a group of missing values by Harel (2008), Harel & Stratton (2009), and Harel (2009a). Generally speaking, it is important to note that there is an increased push to use more imputations as indicated by Graham *et al.* (2007), Bodner (2008), and White *et al.* (2011), to name a few.

5 Latent class regression analysis for breast cancer severity

5.1 Data motivation

To demonstrate the methodology, we analyzed data from a previously reported study on racial and ethnic disparities in breast cancer severity (Smith-Bindman *et al.*, 2006). These data come from a prospective cohort obtained by pooling data from seven mammography registries that constitute the National Cancer Institute-funded Breast Cancer Surveillance Consortium (<http://breastscreening.cancer.gov/>).

The scientific question of interest was whether observed racial and ethnic differences in breast cancer severity were due to differences in screening utilization. Smith-Bindman *et al.* (2006) separately modeled each cancer severity measure excluding subjects with missing data for each response. Here, we model cancer severity as a latent class with four binary manifest variables: tumor size (*Size*: $> 15\text{mm}$ vs. $\leq 15\text{mm}$), American Joint Committee on Cancer (AJCC) 5th edition stage (*Stage*: $II+$ vs. I), grade (*Grade*: III/IV vs. I/II), and Estrogen Receptor (ER) status (*ER*: $-$ vs. $+$), and we keep all subjects including those with incomplete data. Using these four items, three cancer severity classes were identified from a LCA (Bandeem-Roche *et al.*, 1997): early stage cancer, late stage/ER positive, and high grade plus ER negative cancer. These latent classes are described in section 5.2. Table 1 indicated that the goodness-of-fit for the two-class LCA was significantly worse than those for the three-class LCA and models with more than three classes were not identifiable with four binary items. Therefore, we chose the three-class LCA for the following analysis. In addition, from conversation with the cancer experts, the three class model has a meaningful scientific interpretation.

In the first analysis, we include race/ethnicity (*Race*: Non-Hispanic White, Non-Hispanic African-American/Black, and Hispanic) to examine disparities in cancer severity. We then additionally included screening history (*History*: 1–2 years since last mammogram, 3–4 years since last mammogram, first screening mammogram, no prior mammogram) to determine if this accounts for any observed racial disparities.

We included 14,060 women who were diagnosed with invasive breast cancer between 1996 and 2003. Out of 14,060 women, only 8,550 (61%) have complete information for all four manifest variables. Tumor size is missing for 10% of the sample, stage is missing for 9% of the sample, while grade and ER status are missing 18.5% and 28%, respectively. There are 16 different patterns of missing values which are summarized in Table 2. In the first pattern we see that 8,550 subjects have complete information for all four variables. In the second pattern we see that 109 subjects are missing tumor size, but have complete information on all other variables. Many researchers confuse the amount of missing data with the rates of missing information. Under MCAR these two measures will be tied to one another; however, this is not the case in general (Harel, 2007). Since the MCAR assumption is not

realistic in this example, we can not tie the percentages above to the rates of missing information, and instead estimate it using MI.

The present analyses are focused on estimation of LCR parameters and two types of missing information (due to the latent class variable and missing manifest items). Whereas this paper's focus is not on the substantive findings, we provide some speculative interpretation of the observed results. We report and compare the results from EM, MI, and two-stage MI, and discuss the additional information one can get using each procedure.

Following Sections (3.2) and (3.3), ML estimation or MCMC can be used to estimate latent class regression parameters. If some items are ignorably missing for some subjects, the EM and MCMC procedures are readily available to accommodate these missing values (Chung *et al.*, 2006). Routines for fitting latent class models with missing items have been implemented in Mplus (Muthén & Muthén, 1998), Latent GOLD (Vermunt & J., 2000) and WinLTA (Collins *et al.*, 1999). We analyzed the data using EM, MI and two-stage MI using an R code (available as supplemental material). We used the MCMC procedure (Chung *et al.*, 2006) to draw random values for parameters from the posterior distributions for two-stage MI. In the first stage, we imputed the class memberships and missing items $m = 150$ times. Another ($k = 2$) sets of class memberships were imputed conditional on each set of complete items in the first stage. We chose the number of imputations based on the results in Harel (2007), estimating that the rates of missing information will vary between 40 and 60 percent. Since the rates of missing information can be unstable with a small number of imputations, a larger number of imputations than typically suggested for simple missing data problems is needed. After imputation, we combined the results using Shen's rules (Shen, 2000) as described in Section 4.

5.2 Data analysis

We investigate the effect of race/ethnicity on the severity of breast cancer with and without adjustment for prior mammography utilization. We are interested in whether adjustment for mammography history reduces the effects of race/ethnicity on cancer severity. Smith-Bindman *et al.* (2006) examined the effect of race/ethnicity on mammography utilization in the BCSC cohort that included all women, not just those with breast cancer. They found that African-American and Hispanic women were less likely to receive adequate mammographic screening than non-Hispanic White women and that women screened less often were diagnosed with breast cancer with more severe tumor characteristics. This motivates the hypothesis that less frequent screening utilization could be a reason why minorities are more likely to be diagnosed with cancers with worse prognostic characteristics.

ML-based item prevalences and item probabilities are very similar for both models and are shown in Table 3. Women in the *Early Stage* class are more likely to have tumors with good prognostic characteristics; they have a low probability of having large tumors (18%), late stage tumors (12%), high grade tumors (16%), and ER-negative tumors (7%). Women in the *Late Stage* class have poor prognosis due to a high prevalence of large tumors (97%) and late stage disease (95%), but many of these tumors tend to be lower grade (60%) and ER positive (87%), which have better treatment options. The *High grade + ER Negative* class is a severe cancer class of aggressive tumors with poor prognosis, with high prevalence of large tumors (67%) and late stage tumors (61%) of high grade (100%) and ER negative status (85%).

Table 4 provides the regression coefficients from ML, MI, and two-stage MI from Model 1, the LCR model without adjustment for mammography history, and Model 2, the LCR model with adjustment for mammography history. We present the table in such manner to show the additional information each procedure provides. In particular, ML provides the estimates

and their variances; MI adds the rates of missing information, and two-stage MI separates the missing information into two parts. The MI estimates are the same as those of two-stage MI, which are slightly different than those from EM algorithm (ML). The exponentiated coefficients may be interpreted as estimated odds ratios. For instance in two-stage MI, the estimated odds of belonging to the *Late Stage* class versus the *Early Stage* class are $\exp(0.34) = 1.40$ and $\exp(0.40) = 1.49$ times higher for African-Americans and Hispanics than Whites, respectively. This confirms that African-Americans and Hispanics are diagnosed with a later stage of disease compared with Whites. Similarly, when comparing the *High Grade* class versus the *Early Stage* class African-Americans have $\exp(1.42) = 4.14$ times higher risk than Whites while Hispanics have $\exp(0.49) = 1.63$ times higher risk than Whites, which leads to the same conclusion as above. Adjusting for mammography history only slightly attenuated the effects of race/ethnicity. This suggests that differences in cancer severity by race/ethnicity are not accounted for by differences in mammography utilization.

Table 4 shows that the rates of missing information for both Models 1 and 2 are moderate for the African-American and Hispanic indicators (20% – 40%), but are slightly elevated for Model 2 compared with Model 1. In contrast, the rates of missing information for the intercept are reduced in Model 2 compared with Model 1. For both models, most of the missing information is due to measurement error (missing latent class memberships), B^A , and only small amount is due to missing items, A . However, some indicators such as the indicator for African-American race has the same amount of missing information due to missing values and due to measurement error. This suggests that a decrease in the rate of missing tumor characteristics will not affect the results much, except perhaps in African-American women. To decrease the amount of missing information, resources should focus on the better measurement of the latent classes (e.g. add more cancer characteristics that predict cancer severity) as well as decreased missing data in African-Americans. The rates of missing information for the *Late Stage* class are lower than the rates of missing information for the *High grade + ER Negative* class, which can be attributed to the different size of the groups.

6 Conclusions

In this manuscript we introduce a method for estimating the rate of missing information due to incomplete data in latent class regression. Our two-stage MI procedure allows the separation of the missing information into two parts, measurement error (the missing latent variable) and non-response/missing manifest variables. The parameter estimates from the two-stage MI approach are very similar to standard MI and the EM algorithm (ML), but the MI procedures add the rates of missing information compared to ML estimation. Two-stage MI further separates the rates of missing information into two parts: rates of missing information due to non-response and rates of missing information due to measurement error.

The knowledge of the rates of missing information and in particular, the separation of these rates of missing information into two parts can be useful in evaluating the model and also for design of future studies. For example, in Section 5 we found that the missing information is mostly due to measurement error, and although only 61% of subjects have complete data, the impact of these missing values is negligible compared to the missing information due to measurement error except in African-American women. The percentage of African-American women missing a response to items *Size*, *Stage*, *Grade*, and *ER* is 17.2%, 16.0%, 24.1%, and 28.6%, respectively. These missing rates are clearly higher than those of their counterparts, especially for items *Size*, *Stage*, and *Grade*. We believe that a greater frequency of missing items played a role in this exception regarding the African-American women. Measurement items should be added with discretion: adding a measurement item with a strong relationship with cancer severity could produce a positive effect in terms of a

decrease in measurement error. However, adding an item with a higher non-response rate could have a negative effect on missing information. For future studies, resources should be used to explore the criteria on how to select the items to enhance the measurement model by finding the optimal condition to balance these two types of information.

It is important to note that not all missing values are created equally. Subjects with missing data who will be eliminated from a complete case analysis may have 1, 2, 3 or 4 missing items. In Table (2), it is apparent that when using complete case analysis, only 61% of the data is available. There are approximately 85% of subjects with at least one missing value and approximately 93% of subjects with at least two missing values. In this paper we treat all missing values in the same way. Harel (2008) develop the idea of outfluence where different missing values can have different impact on the inference of interest. Using this measure we can find the relative impact of subjects with one missing values to all others, the relative impact of those with two missing values, etc. But this is not the main point of this paper. Our main point was to illustrate that two stage MI can separate the rate of missing information into two parts, one due to the unobserved latent class variable and another due to the nonresponse in manifest variable.

In this manuscript, we assume ignorable missing data mechanisms. Harel & Schafer (2009) have shown that when separating the missing data into two parts, it is possible to use different ignorability assumptions. Harel & Schafer (2009) used a ML approach, but it is very simple to use the two-stage MI approach as well. We defer further discussion of missing data assumptions to Harel & Schafer (2009).

Acknowledgments

The first author efforts were partially supported by Award Number K01MH087219 from the National Institute of Mental Health. The second author efforts were partially supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0009437 and 2012R1A1A2040773). The data collection for this work was supported by the National Cancer Institute-funded Breast Cancer Surveillance Consortium (U01CA63740, U01CA86076, U01CA86082, U01CA63736, U01CA70013, U01CA69976, U01CA63731, U01CA70040, HHSN261201100031C). The collection of cancer data used in this study was supported in part by several state public health departments and cancer registries throughout the U.S. For a full description of these sources, please see: <http://www.breastscreening.cancer.gov/work/acknowledgement.html>. We thank the participating women, mammography facilities, and radiologists for the data they have provided for this study. A list of the BCSC investigators and procedures for requesting BCSC data for research purposes are provided at: <http://breastscreening.cancer.gov/>. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute, the National Institute of Mental Health, the National Institutes of Health or the National Research Foundation of Korea.

References

1. Agresti A, Lang J. Quasi-symmetric latent class models, with application to rater agreement. *Biometrics*. 1993; 49:131–139. [PubMed: 8513099]
2. Bandeen-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*. 1997; 92:1375–1386.
3. Bodner T. What improves with increased missing data imputations? *Structural Equation Modeling*. 2008; 15:651–675.
4. Brémaud, P.; Bremaud, P. Markov chains: Gibbs fields, Monte Carlo simulation, and queues. Springer-Verlag Inc; 1999.
5. Chung H, Flaherty BP, Schafer JL. Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society, Series A*. 2006; 169:723–743.
6. Clogg CC, Goodman LA. Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*. 1984; 79:762–771.

7. Collins, L.; Flaherty, B.; Hyatt, S.; Schafer, J. Winlta user's guide part 1. The Methodology Center, Penn State University; 1999. Software
8. Dayton CM, Macready GB. Concomitant-variable latent-class models. *Journal of the American Statistical Association*. 1988; 83:173–178.
9. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*. 1977; 39:1–38.
10. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian Data Analysis, Second Edition* (Chapman & Hall/CRC Texts in Statistical Science). 2. Chapman and Hall/CRC; 2003.
11. Gilks, WRe; Richardson, Se; Spiegelhalter, DJe. *Markov chain Monte Carlo in practice*. Chapman & Hall Ltd; 1998.
12. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*. 1974; 61:215–231.
13. Graham J, Olchowski A, Gilreath T. How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*. 2007; 8:206–213. [PubMed: 17549635]
14. Harel O. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology*. 2007; 4:75–89.
15. Harel O. Outfluence – The impact of missing values. *Model Assisted Statistics and Applications*. 2008; 3:161–168.
16. Harel, O. The impact of model mis-specification on the outfluence. In: Vanhoof, K.; Ruan, D.; Li, T.; Wets, G., editors. *Intelligence decision making systems*, World Scientific Proceedings Series on Computer Engineering and Information Science. World Scientific; 2009a. p. 221-226.
17. Harel, O. *Strategies for data analysis with two types of missing values: From theory to application*. Lambert Academic Publishing; 2009b.
18. Harel O, Miglioretti D. Missing information as a diagnostic tool for latent class analysis. *Journal of Data Science*. 2007; 5:269–288.
19. Harel O, Pellowski J, Kalichman S. Are we missing the importance of missing values in HIV prevention randomized clinical trials? review and recommendations. *AIDS and Behavior*. 2012:1–12. [PubMed: 21476006]
20. Harel O, Schafer J. Partial and latent ignorability in missing-data problems. *Biometrika*. 2009; 96:37–50.
21. Harel O, Stratton J. Inferences on the outfluence – how do missing values impact your analysis? *Communications in Statistics-Theory & Methods*. 2009; 38:2884–2898.
22. Harel O, Zhou XH. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*. 2007; 26:3057–3077. [PubMed: 17256804]
23. Imbens GW, Manski CF. Confidence intervals for partially identified parameters. *Econometrica*. 2004; 72:1845–1857.
24. Kinney S, Reiter J. Inferences for two stage multiple imputation for nonresponse. *Journal of Statistical Theory and Practice*. 2009; 3:307– 318.
25. Lazarsfeld, PF.; Henry, NW. *Latent structure analysis*. Houghton Mifflin; Boston: 1968.
26. Little, R.; Rubin, D. *Statistical analysis with missing data*. J. Wiley & Sons; New York: 2002.
27. Liu, JS. *Monte Carlo strategies in scientific computing*. Springer-Verlag Inc; 2001.
28. Muthén, L.; Muthén, B. *Mplus user's guide*. Muthén & Muthén; Los Angeles: 1998.
29. Robert, CP.; Casella, G. *Monte carlo statistical methods*. 2. Springer; New York: 2004.
30. Rubin, D. *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons; New York: 1987.
31. Rubin DB. Nested multiple imputation of nmes via partially incompatible mcmc. *Statistica Neerlandica*. 2003; 57:3–18.
32. Schafer, JL. *Analysis of incomplete multivariate data*. Chapman & Hall; London: 1997.
33. Shen, Z. PhD thesis. Department of Statistics, Harvard University; Cambridge, MA: 2000. Nested multiple imputation.
34. Smith-Bindman R, Miglioretti D, Lurie N, Abraham L, Barbash R, Strzelczyk J, Dignan M, Barlow W, Beasley C, Kerlikowske K. Does utilization of screening mammography explain racial

- and ethnic differences in breast cancer? *Annals of Internal Medicine*. 2006; 144:541–554. [PubMed: 16618951]
35. Tanner WA, Wong WH. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*. 1987; 82:528–550.
36. Tierney L. Markov chains for exploring posterior distributions (with discussion). *Annals of Statistics*. 1994; 22:1701–1762.
37. Vermunt, J.; JM. Latent gold user's guide. Statistical Innovations, Inc; Belmont, MA: 2000.
38. White I, Royston P, Wood A. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*. 2011; 30:377–399. [PubMed: 21225900]

Goodness-of-fit statistics for a series of LCA models with different numbers of classes

Table 1

Number of classes	Log-likelihood	Number of parameters	Degrees of freedom	AIC	BIC
2	-26650.8	9	6	53319.6	53387.6
3	-26260.8	14	1	52549.7	52655.4
4	Not identifiable				

Table 2

Missingness patterns and their frequencies in the manifest items (1=observed and 0=missing).

<i>Size</i>	<i>Stage</i>	<i>Grade</i>	<i>ER</i>	Freq.
1	1	1	1	8550
0	1	1	1	109
1	0	1	1	58
0	0	1	1	185
1	1	0	1	1000
0	1	0	1	31
1	0	0	1	55
0	0	0	1	96
1	1	1	0	2171
0	1	1	0	78
1	0	1	0	66
0	0	1	0	244
1	1	0	0	713
0	1	0	0	148
1	0	0	0	98
0	0	0	0	458

Table 3

The probabilities of having poor prognostic factors within each latent class and class prevalences.

Manifest items	Latent classes		
	<i>Early Stage</i>	<i>Late Stage</i>	<i>High grade + ER Negative</i>
Large size	0.18	0.97	0.67
Late stage	0.12	0.98	0.61
High grade	0.16	0.40	1.00
ER negative	0.07	0.13	0.85
Class prevalence	0.58	0.30	0.12

Table 4

Estimated regression coefficients (Est.), their standard deviations (Std.), total rate of missing information (), rate of missing information due to the class variable (B/A), and rate of missing information due to non-response in the items (A). Model 1 represents the LCR with race/ethnicity as a covariate, while Model 2 includes the mammography history as an additional covariate (*Early Stage* class is the baseline).

Model 1	Late Stage			High grade + ER Negative		
	Est.	Std.	B/A	Est.	Std.	B/A
ML	Intercept	-0.73	0.02	-1.73	0.03	
	Black	0.50	0.08	1.43	0.09	
	Hispanic	0.43	0.08	0.50	0.11	
MI	Intercept	-0.73	0.08	-1.73	0.08	0.86
	Black	0.49	0.10	1.45	0.12	0.43
	Hispanic	0.43	0.09	0.30	0.14	0.43
Two-stage MI	Intercept	-0.72	0.08	-1.73	0.08	0.86
	Black	0.50	0.10	1.45	0.12	0.40
	Hispanic	0.43	0.09	0.50	0.15	0.31
Model 2	Late Stage			High grade + ER Negative		
	Est.	Std.	B/A	Est.	Std.	B/A
ML	Intercept	-1.19	0.08	0.44	0.11	
	Black	0.35	0.09	1.42	0.09	
	Hispanic	0.40	0.08	0.50	0.11	
	1-2 yrs	-1.96	0.13	-2.30	0.23	
	3-4 yrs	-1.97	0.09	-1.48	0.11	
	1 st time	-1.47	0.09	-1.25	0.13	
MI	Intercept	1.17	0.12	-0.44	0.17	0.60
	Black	0.34	0.11	1.42	0.12	0.34
	Hispanic	0.40	0.09	0.50	0.14	0.36
	1-2 yrs	-1.95	0.16	-2.36	0.34	0.50
	3-4 yrs	-1.96	0.11	-1.46	0.16	0.52
	1 st time	-1.46	0.12	-1.24	0.18	0.49

Model 2	Late Stage			High grade + ER Negative				
	Est.	Std.	B/A	A	Est.	Std.	B/A	A
Intercept	1.18	0.12	0.51	0.40	0.11	0.17	0.59	0.39
Black	0.34	0.11	0.37	0.19	0.18	1.42	0.37	0.18
Hispanic	0.40	0.09	0.29	0.24	0.05	0.49	0.14	0.36
Two-stage MI								
1-2 yrs	-1.96	0.16	0.32	0.22	0.10	-2.35	0.34	0.51
3-4 yrs	-1.97	0.11	0.37	0.27	0.10	-1.46	0.17	0.54
1 st time	-1.46	0.12	0.36	0.27	0.09	-1.24	0.18	0.49

All estimates are significant with p-value < .05.