Journal of Statistical Distributions and Applications
a SpringerOpen Journal

METHODOLOGY Open Access

Missing data approaches for probability regression models with missing outcomes with applications

Li Qi and Yanqing Sun*

*Correspondence: yasun@uncc.edu
Department of Mathematics and Statistics, The University of North Carolina at Charlotte, 28223 Charlotte, NC, USA

Abstract

In this paper, we investigate several well known approaches for missing data and their relationships for the parametric probability regression model $P_\beta(Y|X)$ when outcome of interest Y is subject to missingness. We explore the relationships between the mean score method, the inverse probability weighting (IPW) method and the augmented inverse probability weighted (AIPW) method with some interesting findings. The asymptotic distributions of the IPW and AIPW estimators are derived and their efficiencies are compared. Our analysis details how efficiency may be gained from the AIPW estimator over the IPW estimator through estimation of validation probability and augmentation. We show that the AIPW estimator that is based on augmentation using the full set of observed variables is more efficient than the AIPW estimator that is based on augmentation using a subset of observed variables. The developed approaches are applied to Poisson regression model with missing outcomes based on auxiliary outcomes and a validated sample for true outcomes. We show that, by stratifying based on a set of discrete variables, the proposed statistical procedure can be formulated to analyze automated records that only contain summarized information at categorical levels. The proposed methods are applied to analyze influenza vaccine efficacy for an influenza vaccine study conducted in Temple-Belton, Texas during the 2000-2001 influenza season.

Mathematics Subject Classification: Primary 62J02; Secondary 62F12

Keywords: Augmented inverse probability weighted estimator; Asymptotic results; Automated records; Auxiliary outcome; Efficiency; Inverse probability weighted estimator; Mean score estimation; Recurrent events; Vaccine efficacy; Validation sample

1 Introduction

Suppose that Y is the outcome of interest and X is a covariate vector. One is often interested in the probability regression model $P_\beta(Y|X)$ that relates Y to X. In many medical and epidemiological studies, the complete observations on Y may not be available for all study subjects because of time, cost, or ethical concerns. In some situations, an easily measured but less accurate outcome named auxiliary outcome variable, A, is supplemented. The relationship between the true outcome Y and the auxiliary outcome A in the available observations can inform about the missing values of Y. Let V be a subsample of

the study subjects, termed the validation sample, for which both true and auxiliary outcomes are available. Thus observations on (X, Y, A) are available for the subjects in V and only (X, A) are observed for those not in V.

It is well known that the complete-case analysis, which uses only subjects who have all variables observed, can be biased and inefficient, cf. Little and Rubin (2002). The issues also rise when substituting auxiliary outcome for true outcome; see Ellenberg and Hamilton (1989), Prentice (1989) and Fleming (1992). Inverse probability weighting (IPW) is a statistical technique developed for surveys by Horvitz and Thompson (1952) to calculate statistics standardized to a population different from that in which the data was collected. This approach has been generalized to many aspects of statistics under various frameworks. In particular, the IPW approach is used to account for missing data through inflating the weight for subjects who are underrepresented due to missingness. The method can potentially reduce the bias of the complete-case estimator when weighting is correctly specified. However, this approach has been shown to be inefficient in several situations, see Clayton et al. (1998) and Scharfstein et al. (1999). Robins et al. (1994) developed an improved augmented inverse probability weighted (AIPW) complete-case estimation procedure. The method is more efficient and possesses double robustness property. The multiple imputation described in Rubin (1987) has been routinely used to handle missing data. Carpenter et al. (2006) compared the multiple imputation with IPW and AIPW, and found AIPW as an attractive alternative in terms of double robustness and efficiency. Using the maximum likelihood estimation (MLE) coupled with the EM-algorithm (Dempster et al. 1977), Pepe et al. (1994) proposed the mean score method for the regression model $P_\beta(Y|X)$ when both X and A are discrete.

In this paper, we investigate several well known approaches for missing data and their relationships for the parametric probability regression model $P_\beta(Y|X)$ when outcome of interest Y is subject to missingness. We explore the relationships between the mean score method, IPW and AIPW with some interesting findings. Our analysis details how efficiency is gained from the AIPW estimator over the IPW estimator through estimation of validation probability and augmentation to the IPW score function. Applying the developed missing data methods, we derive the estimation procedures for Poisson regression model with missing outcomes based on auxiliary outcomes and a validated sample for true outcomes. Further, we show that the proposed statistical procedures can be formulated to analyze automated records that only contain aggregated information at categorical levels, without using observations at individual levels.

The rest of the paper is organized as follows. Section 2 introduces several missing data approaches for the probability regression model $P_\beta(Y|X)$, where the outcome Y may be missing. Section 3 explores the relationships among these estimators. The asymptotic distributions of the IPW and AIPW estimators are derived and their efficiencies are compared. Section 3 investigates efficiency of two AIPW estimators, one is based on the augmentation using a subset of observed variables and the other is based on the augmentation using the full set of observed variables. The procedures for Poisson regression using automated data with missing outcomes are derived in Section 4. The finite-sample performances of the estimators are studied in simulations in Section 5. The proposed method is applied to analyze influenza vaccine efficacy for an influenza vaccine study conducted in Temple-Belton, Texas during the 2000-2001 influenza season. The proofs

of the main results are given in the Appendix A, while the proof of a simplified variance formula in Section 4 is placed in the Appendix B.

2 Missing data approaches

Consider the probability regression model $P_\beta(Y|X)$, where Y is the outcome of interest and X is a covariate vector. Let A be the auxiliary outcome for Y and V be the validation set such that observations on (X, Y, A) are available for the subjects in V and only (X, A) are observed for those in $\bar{V}$, the complement of V. In practice, the validation sample may be selected based on the characteristics of a subset, Z, of the covariates in X. We write $X = (Z, Z^c)$. For example, Z may include exposure indicator and other discrete covariates and Z^c may be the exposure time. Let (Z_i, X_i, Y_i, A_i), $i = 1, \ldots, n$, be independent identically distributed (iid) copies of (Z, X, Y, A). Let $\xi_i = I(i \in V)$ be the selection indicator.

Most statistical methods for missing data require some assumptions on missingness mechanisms. The commonly used ones are missing completely at random (MCAR) and missing at random (MAR). MCAR assumes that the probability of missingness in a variable is independent of any characteristics of the subjects. MAR assumes that the probability that a variable is missing depends only on observed variables. In practice, if missingness is a result by design, it is often convenient to let the missing probability depend on the categorical variables only. There is also simplicity in statistical inference by modeling the missing probability based on the categorical variables. We introduce the following missing at random assumptions.

MAR I: ξ_i is independent of Y_i conditional on (X_i, A_i) and ξ_i is independent of Z_i^c conditional on (Z_i, A_i).

MAR II: ξ_i is independent of (Y_i, Z_i^c) conditional on (Z_i, A_i).

Since the conditional density $f(y, z^c|\xi, z, a) = f(z^c|\xi, z, a) f(y|z^c, \xi, z, a) = f(z^c|z, a)$ $f(y|z^c, z, a) = f(y, z^c|z, a)$, MAR I implies MAR II. It is also easy to show that MAR II implies MAR.

Let $\hat{\pi}_i$ be the estimator of the conditional probability $\pi_i = P(\xi_i = 1|X_i, A_i)$, and $\hat{\pi}_i^z$ the estimator of $\pi_i^z = P(\xi_i = 1|Z_i, A_i)$. Let $S_\beta(Y|X)$ denote the partial derivatives of $\log P_\beta(Y|X)$ with respect to β. Let $\hat{E}\left\{S_\beta(Y|X_i)|X_i, A_i\right\}$ be the estimator of the conditional expectation $E\left\{S_\beta(Y|X_i)|X_i, A_i\right\}$, and $\hat{E}\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}$ the estimator of $E\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}$. We investigate several estimators of β based on the following estimating equations with different choices of W_i:

$$\sum_{i=1}^{n} W_i = 0, \tag{1}$$

where W_i takes one of the following forms:

$$W_i^{I1} = \frac{\xi_i}{\hat{\pi}_i^z} S_\beta(Y_i|X_i) \tag{2}$$

$$W_i^{E1} = \xi_i S_\beta(Y_i|X_i) + (1 - \xi_i)\hat{E}\left\{S_\beta(Y|X_i)|Z_i, A_i\right\} \tag{3}$$

$$W_i^{A1} = \frac{\xi_i}{\hat{\pi}_i^z} S_\beta(Y_i|X_i) + \left(1 - \frac{\xi_i}{\hat{\pi}_i^z}\right)\hat{E}\left\{S_\beta(Y|X_i)|Z_i, A_i\right\} \tag{4}$$

$$W_i^{I2} = \frac{\xi_i}{\hat{\pi}_i} S_\beta(Y_i|X_i) \tag{5}$$

$$W_i^{E2} = \xi_i S_\beta(Y_i|X_i) + (1-\xi_i)\hat{E}\left\{S_\beta(Y|X_i)|X_i, A_i\right\} \tag{6}$$

$$W_i^{A2} = \frac{\xi_i}{\hat{\pi}_i^z} S_\beta(Y_i|X_i) + \left(1 - \frac{\xi_i}{\hat{\pi}_i^z}\right)\hat{E}\left\{S_\beta(Y|X_i)|X_i, A_i\right\}. \tag{7}$$

$$W_i^{A3} = \frac{\xi_i}{\hat{\pi}_i} S_\beta(Y_i|X_i) + \left(1 - \frac{\xi_i}{\hat{\pi}_i}\right)\hat{E}\left\{S_\beta(Y|X_i)|X_i, A_i\right\}. \tag{8}$$

The estimator $\hat{\beta}_{I1}$ obtained by using W_i^{I1} is an IPW estimator where a subject's validation probability π_i^z depends only on the category defined by (Z_i, A_i). Because $E\left\{(\pi_i^z)^{-1}\xi_i S_\beta(Y_i|X_i)\right\} = E\left\{S_\beta(Y_i|X_i)\right\} = 0$, the estimator $\hat{\beta}_{I1}$ is approximately unbiased. The estimator $\hat{\beta}_{I2}$ obtained by using W_i^{I2} is also an IPW estimator but with the validation probability π_i depending on the category defined by (Z_i, A_i) and the additional covariate Z_i^c.

The estimator $\hat{\beta}_{E1}$ obtained by using W_i^{E1} is the mean score estimator where the scores $S_\beta(Y_i|X_i)$ for those with missing outcomes are replaced by the estimated conditional expectations given (Z_i, A_i). The estimator $\hat{\beta}_{E2}$ obtained by using W_i^{E2} is the mean score estimator where the scores $S_\beta(Y_i|X_i)$ for those with missing outcomes are replaced by the estimated conditional expectations given (X_i, A_i). The estimator $\hat{\beta}_{E2}$ is the mean score estimator in Pepe et al. (1994). The mean score estimator is the MLE estimator employing the EM-algorithm (Dempster et al. 1977) under the assumption that the auxiliary outcome is noninformative in the sense that the probability model $P_\theta(A|Y, X)$ is unrelated to β.

The estimator $\hat{\beta}_{A1}$ obtained using W_i^{A1} is the AIPW estimator augmented with the estimated conditional expectation $\hat{E}\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}$. The estimator $\hat{\beta}_{A2}$ obtained using W_i^{A2} is the AIPW estimator augmented with the estimated conditional expectation $\hat{E}\left\{S_\beta(Y|X_i)|X_i, A_i\right\}$. The estimator $\hat{\beta}_{A3}$ is obtained using W_i^{A3}. The W_i^{A3} differs from W_i^{A2} in that the estimated validation probability is $\hat{\pi}_i$ instead of $\hat{\pi}_i^z$.

Suppose that $\hat{\pi}_i^z$ is an asymptotically unbiased estimator of $\bar{\pi}_i^z$ and that $\hat{E}\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}$ is asymptotically unbiased of $\bar{E}\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}$, where both $\bar{\pi}_i^z$ and $\bar{E}\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}$ are functions of (Z_i, A_i). Under MAR II, if one of the equalities, $\bar{\pi}_i^z = \pi_i^z$ and $\bar{E}\left[S_\beta(Y|X_i)|Z_i, A_i\right]\} = E\left[S_\beta(Y|X_i)|Z_i, A_i\right]\}$, holds, then

$$E\left\{\left(\bar{\pi}_i^z\right)^{-1}\xi_i S_\beta\left(Y_i|X_i\right)\right\} + E\left\{\left(1 - \left(\pi_i^z\right)^{-1}\xi_i\right)\bar{E}\left[S_\beta\left(Y|X_i\right)|Z_i, A_i\right]\right\} = E\left\{S_\beta(Y_i|X_i)\right\} = 0,$$

which entails that the estimator $\hat{\beta}_{A1}$ has the double robust property in the sense that it is a consistent estimator of β if either $\hat{\pi}_i^z$ is a consistent estimator of π_i^z or $\hat{E}\left[S_\beta(Y|X_i)|Z_i, A_i\right]$ is a consistent estimator of $E\left[S_\beta(Y|X_i)|Z_i, A_i\right]$. Similarly, under MAR I, the estimator $\hat{\beta}_{A2}$ possesses the double robust property in that $\hat{\beta}_{A2}$ is a consistent estimator of β if either $\hat{\pi}_i^z$ is a consistent estimator of π_i^z or $\hat{E}\left[S_\beta(Y|X_i)|X_i, A_i\right]$ is a consistent estimator of $E\left[S_\beta(Y|X_i)|X_i, A_i\right]$. The estimator $\hat{\beta}_{A3}$ has similar double robust property as $\hat{\beta}_{A2}$.

3 Method comparisons and asymptotic results

Let $V(X_i, A_i)$ denote the subjects in V with values of (X, A) equal to (X_i, A_i), $n^V(X_i, A_i)$ the number of subjects in $V(X_i, A_i)$, and $n(X_i, A_i)$ the number of subjects with values of (X, A) equal to (X_i, A_i). When X and A are discrete and their dimensionality is reasonably small, the probability $\pi_i = P(\xi_i = 1|X_i, A_i)$ can be estimated by $\hat{\pi}_i = n^V(X_i, A_i)/n(X_i, A_i)$. The

conditional expectation $E\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ can be nonparametrically estimated based on the validation sample,

$$\hat{E}\left\{S_\beta(Y|X_i)|X_i,A_i\right\} = \sum_{j\in V(X_i,A_i)} S_\beta\left(Y_j|X_j\right)/n^V(X_i,A_i), \tag{9}$$

Under MAR I, $\hat{E}\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ is an unbiased estimator of $E\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$. Now we let $V(Z_i,A_i)$ denote the subjects in V with values of (Z,A) equal to (Z_i,A_i), $n^V(Z_i,A_i)$ the number of subjects in $V(Z_i,A_i)$, and $n(Z_i,A_i)$ the number of subjects in the sample with values of (Z,A) equal to (Z_i,A_i). A nonparametric estimator of $\pi_i^z = P(\xi_i = 1|Z_i,A_i)$ is given by $\hat{\pi}_i^z = n^V(Z_i,A_i)/n(Z_i,A_i)$. A nonparametric estimator of $E\{S_\beta(Y|X_i)|Z_i,A_i\}$ is given by

$$\hat{E}\left\{S_\beta(Y|X_i)|Z_i,A_i\right\} = \sum_{j\in V(Z_i,A_i)} S_\beta\left(Y_j|X_j\right)/n^V(Z_i,A_i). \tag{10}$$

Under MAR II, (Y_i,X_i) is independent of ξ_i conditional on (Z_i,A_i), then $\hat{E}\left\{S_\beta(Y|X_i)|Z_i,A_i\right\}$ is an unbiased estimator of $E\{S_\beta(Y|X_i)|Z_i,A_i\}$.

Proposition 1. *Suppose that $X = (Z,Z^c)$ and A are discrete and their dimensionality is reasonably small. Under the nonparametric estimators $\hat{\pi}_i^z = n^V(Z_i,A_i)/n(Z_i,A_i)$, $\hat{\pi}_i = n^V(X_i,A_i)/n(X_i,A_i)$ and the estimators for the conditional expectation defined in (9) and (10), the estimators $\hat{\beta}_{I1}$, $\hat{\beta}_{E1}$ and $\hat{\beta}_{A1}$ are equivalent, and the estimators $\hat{\beta}_{I2}$, $\hat{\beta}_{E2}$, $\hat{\beta}_{A2}$ and $\hat{\beta}_{A3}$ are equivalent. However, the estimator $\hat{\beta}_{A2}$ is different from $\hat{\beta}_{A1}$ unless Z_i^c is linearly related to Z_i in which case β is not identifiable.*

The results of Proposition 1 are very intriguing since research has shown that the AIPW and the mean score methods are more efficient than the IPW method. It is also intriguing that the AIPW estimators $\hat{\beta}_{A2}$ and $\hat{\beta}_{A3}$ are actually the same estimators, not affected by the validation probability. To further understand these approaches, we investigate the asymptotic properties of these methods where (X,A) are not necessarily discrete. Through the asymptotic analysis, we gain insights about what matters to the efficiency in terms of the selections of the validation sample and the augmentation function.

Suppose that $\tilde{E}\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ is a consistent parametric/nonparametric estimator of $E_a\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$, where $E_a\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ is $E\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ or $E\{S_\beta(Y|X_i)|Z_i,A_i\}$. Let $\pi(X_i,A_i,\psi)$ be the parametric model for the validation probability π_i, where ψ is a q-dimensional parameter. We show in Corollary 2 that the nonparametric estimator of $\pi(X_i,A_i,\psi)$ can also be expressed in the parametric form when (X_i,A_i) are discrete. Let ψ_0 be the true value of ψ. Under MAR I, the MLE $\hat{\psi} = \left(\hat{\psi}_1,\ldots,\hat{\psi}_q\right)$ of $\psi = (\psi_1,\ldots,\psi_q)$ is obtained by maximizing the observed data likelihood,

$$\prod_{i=1}^{n}\{\pi(X_i,A_i,\psi)\}^{\xi_i}\{1-\pi(X_i,A_i,\psi)\}^{1-\xi_i}.$$

The validation probability π_i is estimated by $\tilde{\pi}_i = \pi\left(X_i,A_i,\hat{\psi}\right)$. Then by the standard likelihood based analysis, we have the approximation

$$\hat{\psi}-\psi_0 = n^{-1}\sum_{i=1}^{n}\left(I^\psi\right)^{-1}S_i^\psi + o_p\left(n^{-1/2}\right), \tag{11}$$

where S_i^{ψ} and I^{ψ} are the score vector and information matrix for $\hat{\psi}$ defined by

$$\begin{aligned} S_i^{\psi} &= \frac{(\xi_i - \pi(X_i, A_i, \psi_0))}{\pi(X_i, A_i, \psi_0)(1 - \pi(X_i, A_i, \psi_0))} \frac{\partial \pi(X_i, A_i, \psi_0)}{\partial \psi}, \\ I^{\psi} &= E\left\{ \frac{1}{\pi(X_i, A_i, \psi_0)(1 - \pi(X_i, A_i, \psi_0))} \left(\frac{\partial \pi(X_i, A_i, \psi_0)}{\partial \psi} \right)^{\otimes 2} \right\}, \end{aligned} \tag{12}$$

where $a^{\otimes 2} = aa'$.

Consider the IPW estimator $\hat{\beta}_I$ obtained by solving the estimating equation

$$U_I = \sum_{i=1}^{n} \frac{\xi_i}{\tilde{\pi}_i} S_{\beta}(Y_i|X_i) \tag{13}$$

and the AIPW estimator $\hat{\beta}_A$ based on solving the estimating equation

$$U_A = \sum_{i=1}^{n} \left[\frac{\xi_i}{\tilde{\pi}_i} S_{\beta}\left(Y_i|X_i\right) + \left(1 - \frac{\xi_i}{\tilde{\pi}_i}\right) \tilde{E}\left\{ S_{\beta}(Y|X_i)|X_i, A_i \right\} \right]. \tag{14}$$

Theorem 1. *Assume that $P_{\beta}(Y|X)$ and $\pi(X, A, \psi)$ have bounded third-order derivatives in a neighborhood of the true parameters and are bounded away from 0 almost surely, both $-E\left\{\left(\partial^2/\partial\beta^2\right)\left(\log P_{\beta}(Y|X)\right)\right\}$ and I^{ψ} are positive definite at the true parameters. Then, under* MAR I*,*

$$n^{1/2}\left(\hat{\beta}_I - \beta\right) = I^{-1}(\beta) n^{-1/2} \sum_{i=1}^{n} Q_i^I + o_p(1),$$

$$n^{1/2}\left(\hat{\beta}_A - \beta\right) = I^{-1}(\beta) n^{-1/2} \sum_{i=1}^{n} Q_i^A + o_p(1),$$

where $I(\beta) = E\left\{-\left(\partial^2/\partial\beta^2\right)\left(\log P_{\beta}(Y|X)\right)\right\} = Var\left(S_{\beta}\left(Y_i|X_i\right)\right)$,

$$Q_i^I = \xi_i/\pi_i S_{\beta}(Y_i|X_i) - E\left\{ {\pi_i}^{-2} \xi_i S_{\beta}(Y_i|X_i) \left(\partial \pi(X_i, A_i, \psi_0)/\partial \psi\right)' \right\} \left(I^{\psi}\right)^{-1} S_i^{\psi}$$

and $Q_i^A = \xi_i/\pi_i S_{\beta}(Y_i|X_i) + (1 - \xi_i/\pi_i) E_a\left\{S_{\beta}(Y|X_i)|X_i, A_i\right\}$.

Both $n^{1/2}\left(\hat{\beta}_I - \beta\right)$ and $n^{1/2}\left(\hat{\beta}_A - \beta\right)$ have asymptotically normal distributions with mean zero and covariances equal to $I^{-1}(\beta) Var\left(Q_i^I\right) I^{-1}(\beta)$ and $I^{-1}(\beta) Var\left(Q_i^A\right) I^{-1}(\beta)$, respectively. Further,

$$Var\left(Q_i^I\right) = Var\left(Q_i^A\right) + Var(B_i + O_i) \tag{15}$$

and

$$Var\left(Q_i^A\right) = I(\beta) + Var\left(\left(1 - \frac{\xi_i}{\pi_i}\right)\left\{S_{\beta}(Y_i|X_i) - E_a\left\{S_{\beta}(Y|X_i)|X_i, A_i\right\}\right\}\right), \tag{16}$$

where $O_i = E\left\{{\pi_i}^{-2} \xi_i S_{\beta}(Y_i|X_i) \left(\partial \pi(X_i, A_i, \psi_0)/\partial \psi\right)'\right\} \left(I^{\psi}\right)^{-1} S_i^{\psi}$ and $B_i = (1 - \xi_i/\pi_i)$ $E_a\left\{S_{\beta}(Y|X_i)|X_i, A_i\right\}$.

Suppose that the validation probability $\pi_i = P(\xi_i = 1|X_i, A_i)$ depends only on (Z_i, A_i). That is, $\pi_i = \pi_i^z = P(\xi_i = 1|Z_i, A_i)$. Suppose that $\tilde{\pi}_i$ is the MLE of π_i^z under the parametric family $\psi(Z_i, A_i, \psi)$. Let $\hat{\beta}_{A1}$ be the estimator obtained by solving (14) where the augmented term, $\tilde{E}\left\{S_{\beta}(Y|X_i)|X_i, A_i\right\}$, is a consistent parametric/nonparametric estimator of $E\left\{S_{\beta}(Y|X_i)|Z_i, A_i\right\}$. Let $\hat{\beta}_{A2}$ be the estimator obtained by solving (14) where $\tilde{E}\left\{S_{\beta}(Y|X_i)|X_i, A_i\right\}$ is a consistent parametric/nonparametric estimator of $E\left\{S_{\beta}(Y|X_i)|X_i, A_i\right\}$. The following corollary presents the asymptotic results for two AIPW estimators of β, one that corresponds to the augmentation based on a subset,

(Z, A), of observed variables and the other that corresponds to the augmentation based on the full set, (X, A), of the observed variables.

Corollary 1. *Suppose that the validation probability* $\pi_i = P(\xi_i = 1|X_i, A_i)$ *depends only on* (Z_i, A_i). *Under the conditions of Theorem 1,*

$$n^{1/2}\left(\hat{\beta}_{A1} - \beta\right) \xrightarrow{\mathcal{D}} N\left(0, I^{-1}(\beta) + I^{-1}(\beta)\Sigma_{A1}(\beta)I^{-1}(\beta)\right), \tag{17}$$

and

$$n^{1/2}\left(\hat{\beta}_{A2} - \beta\right) \xrightarrow{\mathcal{D}} N\left(0, I^{-1}(\beta) + I^{-1}(\beta)\Sigma_{A2}(\beta)I^{-1}(\beta)\right), \tag{18}$$

where $\Sigma_{A1}(\beta) = E\left[((1-\pi_i^z)/\pi_i^z)Var\{S_\beta(Y_i|X_i)|Z_i, A_i\}\right]$ *and* $\Sigma_{A2}(\beta) = E\left[((1-\pi_i^z)/\pi_i^z)\right.$ $\left.Var\left\{S_\beta(Y_i|X_i)|X_i, A_i\right\}\right]$. *The asymptotic variance of* $\hat{\beta}_{A2}$ *is smaller than the asymptotic variance of* $\hat{\beta}_{A1}$ *if the covariates* Z_i *are a proper subset of* X_i.

Suppose that (Z, A) are discrete taking values (z, a) in a set $\mathcal{Z}$ of finite number of values. If the number of parameters in ψ equals the number of values $\psi_{z,a} = P(\xi_i = 1|Z_i = z, A_i = a)$ for all distinct pairs (z, a), then $\psi = \{\psi_{z,a}\}$ and $\pi(z, a, \psi) = \psi_{z,a}$. Further, $\frac{\partial \pi(z,a,\psi_0)}{\partial \psi}$ can be viewed as a column vector with 1 in the position for $\psi_{z,a}$ and 0 elsewhere. The information matrix I^ψ defined in (12) has the expression,

$$I^\psi = \sum_{z,a} \rho(z, a)\left\{\frac{1}{\pi(z, a, \psi_0)(1 - \pi(z, a, \psi_0))}\frac{\partial \pi(z, a, \psi_0)}{\partial \psi}\left(\frac{\partial \pi(z, a, \psi_0)}{\partial \psi}\right)'\right\},$$

where $\rho(z, a) = P(Z_i = z, A_i = a)$. It follows that I^ψ is a diagonal matrix and its inverse matrix is also diagonal. The MLE $\hat{\psi}_{z,a} = n^V(z, a)/n(z, a)$ is in fact the nonparametric estimator for $\psi_{z,a}$ based on the proportion of validated samples in the category specified by (z, a). The equation (11) can be expressed as

$$\hat{\psi}_{z,a} - \pi(z, a, \psi_0) = n^{-1}\frac{n^V(z, a) - n(z, a)\pi(z, a, \psi_0)}{\rho(z, a)} + o_p\left(n^{-1/2}\right),$$

for $(z, a) \in \mathcal{Z}$.

By Threom 1, the possible efficiency gain of the AIPW estimator over the IPW estimator is shown through the equation (15). The AIPW estimator is more efficient unless $\mathrm{Var}(B_i + O_i) = 0$. In particular, from the proof of Theorem 1, we have

$$n^{-1/2}U_A = n^{-1/2}\sum_{i=1}^{n} Q_i^A + o_p(1) \tag{19}$$

$$n^{-1/2}U_I = n^{-1/2}\sum_{i=1}^{n} Q_i^A - n^{-1/2}\sum_{i=1}^{n}(B_i + O_i) + o_p(1), \tag{20}$$

where B_i and O_i are defined following (16). The following corollary presents the analysis of the term $n^{-1/2}\sum_{i=1}^{n}(B_i + O_i)$ when (Z_i, A_i) are discrete to understand how efficiency may be gained from the AIPW estimator over the IPW estimator.

Corollary 2. *Under the conditions of Theorem 1, suppose that* $X = (Z, Z^c)$ *and* (Z, A) *are discrete taking values* (z, a) *in a set* $\mathcal{Z}$ *of finite number of values. Suppose that the*

validation probability $\pi_i = P(\xi_i = 1|X_i, A_i)$ *only depends on* (Z_i, A_i) *and* $\psi_{z,a} = P(\xi_i = 1|Z_i = z, A_i = a)$ *is estimated nonparametrically by* $\hat{\psi}_{z,a} = n^V(z,a)/n(z,a)$. *Then*

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^{n}(O_i + B_i) \\ & = -\sum_{z,a} n^{-1/2}\sum_{j=1}^{n}\frac{\xi_j - \pi(z,a,\psi_0)}{\pi(z,a,\psi_0)} I(Z_j = z, A_j = a) \\ & \quad \Big[E\{S_\beta(Y_j|X_j)|Z_j = z, A_j = a\} - E\left\{S_\beta(Y_j|X_j)|Z_j = z, Z_j^c, A_j = a\right\}\Big]. \end{aligned} \tag{21}$$

By Corollary 2, (19) and (20), $\hat{\beta}_A$ is more efficient than $\hat{\beta}_I$ unless $Var\{S_\beta(Y_j|X_j)|Z_j = z, A_j = a\} = 0$ for all (z,a) for which $P(Z_i = z, A_i = a) \neq 0$. If $X = Z$ and the validation probability $\pi_i = P(\xi_i = 1|X_i, A_i)$ is nonparametrically estimated with the cell frequencies $\hat{\psi}_{z,a} = n^V(z,a)/n(z,a)$, then $\hat{\beta}_A$ and $\hat{\beta}_I$ are asymptotically equivalent.

Remark Consider the estimators of β obtained based on the estimating equation (1) corresponding to different choices of W_i given in (2) to (8). If (Z, A) are discrete and the validation probability $\pi_i^z = P(\xi_i = 1|Z_i, A_i)$ is estimated nonparametrically by the cell frequency, then by Theorem 1 and Corollary 2, $\hat{\beta}_{A1}$ and $\hat{\beta}_{I1}$ have same asymptotic normal distributions as long as $\hat{E}[\,S_\beta(Y|X_i)|Z_i, A_i]$ is a consistent estimator of $E[\,S_\beta(Y|X_i)|Z_i, A_i]$. But $\hat{\beta}_{A2}$ is more efficient than $\hat{\beta}_{I1}$ as long as $\hat{E}[\,S_\beta(Y|X_i)|X_i, A_i]$ is a consistent estimator of $E[\,S_\beta(Y|X_i)|X_i, A_i]$ since $\mathrm{Var}(B_i + O_i)$ is not zero by (21). These results are not affected by whether $E[\,S_\beta(Y|X_i)|Z_i, A_i]$ and $E[\,S_\beta(Y|X_i)|X_i, A_i]$ are estimated nonparametrically or based on some parametric models. In addition, by Theorem 1, Corollary 1 and 2, $\hat{\beta}_{A3}$ and $\hat{\beta}_{I2}$ have the same asymptotic normal distributions as long as $\hat{E}[\,S_\beta(Y|X_i)|X_i, A_i]$ is a consistent estimator of $E[\,S_\beta(Y|X_i)|X_i, A_i]$.

4 Poisson regression using the automated data with missing outcomes

Many medical and public health data are available only in aggregated format, where the variables of interest are aggregated counts without being available at individual levels. Many existing statistical methods for missing data require observations at individual levels. Applying the missing data methods presented in Section 3, we derive some estimation procedures for the Poisson regression model with missing outcomes based on auxiliary outcomes and a validated sample for true outcomes. Further, we show that, by stratifying based on a set of discrete variables, the proposed statistical procedure can be formulated so that it can be used to analyze automated records which do not contain observations at individual levels, only summarized information at categorical levels.

Let Y represent the number of events occurring in the time-exposure interval $[0, T]$ and Z the covariates. We consider the Poisson regression model,

$$P(Y = y|Z, T) = \exp\left\{-T\exp(\beta' Z)\right\}\left\{T\exp(\beta' Z)\right\}^y/y!, \tag{22}$$

where Z is a vector of $k+1$ covariates and β a vector of $k+1$ regression coefficients. In practice, the exact number of true events may not be available for all subjects. We may instead have the number of possible events, namely, the auxiliary events. For example, in the study of vaccine adverse events associated with childhood immunizations, the

number of auxiliary events A for MAARI is collected based on ICD-9 codes through hospital records. Further diagnosis may indicate that some of these events are false events. The number of true vaccine adverse events, Y, can only be confirmed for the subjects in the validation set V. Suppose that Z is the vaccination status, 1 for the vaccinated and 0 for the unvaccinated. Then, under Poisson regression, $\exp(\beta)$ is the relative rate of event occurrence per unit time of the exposed versus unexposed. We assume that the number of automated events A can be expressed as $A = Y + W$, where W is the number of false events independent of Y conditional on (Z, T). Suppose that W follows the Poisson regression model

$$P(W = w|Z, T) = \exp\left\{-T\exp(\gamma' Z)\right\}\left\{T\exp(\gamma' Z)\right\}^{w}/w!, \tag{23}$$

where $\gamma' = \left(a_0, a_1, \gamma_1, \cdots, \gamma_{k-1}\right)$.

We apply the missing data methods introduced in Section 3 on model (22). The variables (Z_i, T_i, Y_i, A_i) are observed for the validation sample V and (Z_i, T_i, A_i) are observed for the nonvalidation sample $\bar{V}$. While the covariate Z can be considered as categorical, it is natural to consider the exposure time T as a continuous variable. We assume that the validation probability depends only on the stratification of (Z, A). That is, the validation sample is a stratified random sample by the categories defined by (Z, A). Of those estimators discussed in Section 2, there are only two different estimators, $\hat{\beta}_{I1}$ and $\hat{\beta}_{A2}$. We show in Section 4.3 that the proposed method can be formulated so that it can be used to analyze the automated records with missing outcomes. First we derive the explicit estimation procedures for $\hat{\beta}_{I1}$ and $\hat{\beta}_{A2}$ and their variance estimators under model (22).

4.1 Inverse probability weighting estimation

We adopt all notations introduced in Section 3. In particular, let $\pi_i^z = P(\xi_i = 1|Z_i, A_i)$ and $\hat{\pi}_i^z = n^V(Z_i, A_i)/n(Z_i, A_i)$. Let $X = (Z, T)$ and $X_i = (Z_i, T_i)$ to be consistent with earlier notations. The score function for subject i under model (22) is $S_\beta(Y_i|X_i) = Z_i'(Y_i - T_i\exp(\beta' Z_i))$. The estimator $\hat{\beta}_{I1}$ is obtained by solving $\sum_{i=1}^n(\xi_i/\hat{\pi}_i^z)S_\beta(Y_i|X_i) = 0$, where $S_\beta(Y_i|X_i) = Z_i'(Y_i - T_i\exp(\gamma' Z_i))$. By Corollary 1, $\sqrt{n}(\hat{\beta}_{I1} - \beta)$ converges in distribution to a normal distribution with mean zero and the variance matrix $I^{-1}(\beta) + I^{-1}(\beta)\Sigma_{A1}(\beta)I^{-1}(\beta)$, where $\Sigma_{A1}(\beta) = E\left[((1 - \pi_i^z)/\pi_i^z)\mathrm{Var}\left\{S_\beta(Y_i|X_i)|Z_i, A_i\right\}\right]$.

The information matrix $I(\beta) = E(Z_i Z_i' T_i\exp(\beta' Z_i)) = \sum_z P(Z_i = z)zz'\exp(\beta' z)E(T_i|Z_i = z)$ can be estimated by $\hat{I}(\hat{\beta})$ which is obtained by replacing β with $\hat{\beta}_{I1}$, $P(Z_i = z)$ by the sample proportion of the event $\{Z_i = z\}$, and $E(T_i|Z_i = z)$ with the sample average exposure time for those with covariates $Z_i = z$. The matrix $\Sigma_{A1}(\beta)$ can be estimated by

$$\hat{\Sigma}_{A1}(\hat{\beta}) = \sum_{a,z}\hat{\rho}(a, z)\frac{1 - \hat{\rho}^v(a, z)}{\hat{\rho}^v(a, z)}\widehat{\mathrm{Var}}\{S_\beta(Y|X)|A = a, Z = z\}, \tag{24}$$

where $\hat{\rho}(a, z)$ is the estimator of $P\{A_i = a, Z_i = z\}$, $\hat{\rho}^v(a, z)$ is the estimator of $P\{i \in V|A_i = a, Z_i = z\}$, and $\widehat{\mathrm{Var}}\left\{S_\beta(Y|X)|A = a, Z = z\right\}$ is an estimator of $\mathrm{Var}\left\{S_\beta(Y_i|X_i)|Z_i, A_i\right\}\Big]$ which is derived in the following.

Since A is observed for all subjects, W can be determined if Y is known, and undetermined otherwise. The IPW estimator, $\hat{\gamma}_{I1}$, of γ can be estimated by solving the equation $\sum_{i=1}^n(\xi_i/\hat{\pi}_i^z)S_\gamma(W_i|X_i) = 0$, where $S_\gamma(W_i|X_i) = Z_i'(W_i - T_i\exp(\gamma' Z_i))$. The conditional distribution of Y given $A = a$, T, and $Z = z$ is Binomial (a, p_z), where $p_z = \exp(\beta' z)/(\exp(\beta' z) + \exp(\gamma' z))$. Since this conditional distribution does

not depend on T, the outcome Y and T are conditionally independent given (A, Z). Therefore, $\text{Var}\left\{S_\beta(Y|X)|A,Z\right\} = ZZ'\left\{\text{Var}(Y|A,Z) + \exp(2\beta'Z)\text{Var}(T|A,Z)\right\}$. The variance $\text{Var}(Y|A = a, Z = z)$ can be estimated by $a\hat{p}_z(1-\hat{p}_z)$, where $\hat{p}_z = \exp(\hat{\beta}'z)/\left\{\exp(\hat{\beta}'z) + \exp(\hat{\gamma}'z)\right\}$, and $\text{Var}(T|A = a, Z = z) = E(T^2|A = a, Z = z) - \{E(T|A = a, Z = z)\}^2$ can be estimated nonparametrically using the first and the second sample moments conditional on each category with $A = a$ and $Z = z$.

4.2 Augmented inverse probability weighted estimation

Under the assumption that W follows the Poisson regression model (23) and is independent of Y conditional on (Z, T), $E\left\{S_\beta(Y|X)|Z,T,A\right\} = AZ'\frac{\exp(\beta'Z)}{\exp(\beta'Z)+\exp(\gamma'Z)} - TZ'\exp(\beta'Z)$. Let $\hat{E}\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ be the estimator of $E\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ for a given β by substituting γ by its estimator $\hat{\gamma}_{I1}$ of Section 4.1. Then the estimator $\hat{\beta}_{A2}$ is obtained by solving

$$\sum_{i=1}^{n}\frac{\xi_i}{\hat{\pi}_i^z}S_\beta(Y_i|X_i) + \left(1-\frac{\xi_i}{\hat{\pi}_i^z}\right)\hat{E}\left\{S_\beta(Y|X_i)|X_i,A_i\right\} = 0. \tag{25}$$

By Corollary 1, $\sqrt{n}(\hat{\beta}_{A2} - \beta)$ converges in distribution to a normal distribution with mean zero and the variance matrix where $I^{-1}(\beta)+I^{-1}(\beta)\Sigma_{A2}(\beta)I^{-1}(\beta)$, where $\Sigma_{A2}(\beta) = E\left[((1-\pi_i^z)/\pi_i^z)\text{Var}\{S_\beta(Y_i|X_i)|X_i,A_i\}\right]$. The information matrix $I(\beta)$ can be estimated by $\hat{I}(\hat{\beta})$ given in Section 4.1. The conditional variance $\text{Var}\left\{S_\beta(Y|X)|Z = z, T, A = a\right\} = ap_z(1-p_z)z^{\otimes 2}$ can be estimated by $a\hat{p}_z(1-\hat{p}_z)$, where $\hat{p}_z = \exp(\hat{\beta}'z)/(\exp(\hat{\beta}'z) + \exp(\hat{\gamma}'z))$. It follows that $\Sigma_{A2}(\beta)$ can be consistently estimated by

$$\hat{\Sigma}_{A2}(\beta) = \sum_{a,z}\hat{\rho}(a,z)\frac{1-\hat{\rho}^v(a,z)}{\hat{\rho}^v(a,z)}a\hat{p}_z(1-\hat{p}_z)z^{\otimes 2},$$

where $\hat{\rho}(a,z)$ is the estimator of $P\{A_i = a, Z_i = z\}$ and $\hat{\rho}^v(a,z)$ is the estimator of $P\{i \in V|A_i = a, Z_i = z\}$.

4.3 Estimation using the automated data

This section formulates the missing data estimation procedure for (22) based on the automated (summarized) information at categorical levels defined by relevant covariates of the model. In particular, we show that $\hat{\beta}_{I1}$ and $\hat{\beta}_{A2}$ and their variance estimators can be formulated using the automated data at categorical levels.

In many applications it is convenient to write $Z = (1, Z_{(1)}, Z_{(2)})$ and $\beta = (b_0, b_1, \theta')'$, where $Z_{(1)}$ is the treatment indicator ($Z_{(1)} = 1$ for the exposed group and $Z_{(1)} = 0$ for the unexposed group) and $Z_{(2)} = (\eta_1, \cdots, \eta_{k-1})'$ as the other covariates, and $\theta = (\theta_1, \cdots, \theta_{k-1})'$. For the applications involving the automated data records, we let $\eta_1, \cdots, \eta_{k-1}$ be $k-1$ dummy variables representing k groups. Without loss of generality, we choose the kth group as the reference group, $\eta_1 = 1$, $\eta_2 = 0$, $\cdots$, $\eta_{k-1} = 0$ for group 1, $\eta_1 = 0$, $\eta_2 = 1$, $\cdots$, $\eta_{k-1} = 0$ for group 2, so on and $\eta_1 = 0$, $\eta_2 = 0$, $\cdots$, $\eta_{k-1} = 0$ for group k. Thus each value of Z denotes a category which can be represented by (l, m) for $l = 0, 1$ and $m = 1, \cdots, k$. This correspondence is denoted by $Z \simeq (l, m)$ for convenience. For $l = 0, 1$ and $m = 1, \cdots, k-1$, category (l, m) is defined by Z with $Z_{(1)} = l$, $\eta_m = 1$ and $\eta_j = 0$ for $j \neq m$, $j = 1, \ldots, k$, and category (l, k) is defined by $Z_{(1)} = l$ and $\eta_j = 0$ for $j = 1, \ldots, k-1$. Under model (22), the expected number of events for a subject in category (l, m) with the time-exposure interval $[0, T]$ is $T\exp(b_{lm})$, for $l = 0, 1$ and

$m = 1, \cdots, k$, where $b_{1k} = b_0 + b_1$, $b_{0k} = b_0$, $b_{1m} = b_0 + b_1 + \theta_m$ and $b_{0m} = b_0 + \theta_m$ for $1 \le m \le k-1$. The parameter b_1 represents the log-relative rate of the exposed versus the unexposed adjusted for other factors.

The following notations are used to show that the estimators of β and their variance estimators can be calculated using the automated information at the categorical levels. Let $V(a,l,m)$ denote the set of subjects in V with $(A = a, Z \simeq (l,m))$, $V(l,m)$ for the set of subjects in V with $(Z \simeq (l,m))$, n_{alm} for the number of subjects with $(A = a, Z \simeq (l,m))$, n^v_{alm} for the number of subjects in $V(a,l,m)$, n^v_{lm} for the number of subjects in $V(l,m)$, $\lambda_{alm} = n_{alm}/n^v_{alm}$, y_{alm} for the number of events for subjects in $V(a,l,m)$, y_{lm} for the number of events for subjects in $V(l,m)$, t_{alm} for the total exposure time for subjects with $(A = a, Z \simeq (l,m))$, $t_{2,alm}$ for the total squared exposure time for subjects with $(A = a, Z \simeq (l,m))$, t_{lm} for the total exposure time for subjects with $Z \simeq (l,m)$, α_{lm} for the number of automated events for subjects with $Z \simeq (l,m)$.

Estimation with $\hat{\beta}_{I1}$ using the automated data. The validation probability π^z_i can be estimated by $1/\lambda_{alm}$ when $A_i = a$, $Z_i \simeq (l,m)$. It can be shown that the estimating equation for $\hat{\beta}_{I1}$ is equivalent to the following nonlinear equations for $\{b_{lm}$, for $l = 0,1$, $m = 1, \cdots, k\}$,

$$\sum_{m=1}^{k}\left(\sum_{a\in A} y_{alm}\lambda_{alm} - e^{b_{lm}}\sum_{a\in A} t_{alm}\lambda_{alm}\right) = 0,$$
$$\sum_{l=0,1}\left(\sum_{a\in A} y_{alm}\lambda_{alm} - e^{b_{lm}}\sum_{a\in A} t_{alm}\lambda_{alm}\right) = 0,$$

for $l = 0,1$ and $m = 1, \ldots, k-1$. When $k > 1$, the equations have no explicit solutions.

In the following, we show that the asymptotic variance of $\hat{\beta}_{I1}$ can be consistently estimated by only using the automated information at categorical levels. The information matrix is a $(k+1) \times (k+1)$ symmetric matrix given by

$$I(\beta) = E(Z_i Z_i' T_i \exp(\beta' Z_i))$$
$$= \begin{bmatrix} \sum_{l,m} q_{lm} & \sum_{m=1}^{k} q_{1m} & q_{11}+q_{01} & \cdots & q_{1r}+q_{0r} \\ \sum_{m=1}^{k} q_{1m} & \sum_{m=1}^{k} q_{1m} & q_{11} & \cdots & q_{1r} \\ q_{11}+q_{01} & q_{11} & q_{11}+q_{01} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{1r}+q_{0r} & q_{1r} & 0 & \cdots & q_{1r}+q_{0r} \end{bmatrix},$$

where $r = k-1$ and $q_{lm} = E(T_i e^{b_{lm}} I\{\text{individual } i \text{ in category } (l,m)\})$. The consistent estimator, $\hat{I}(\hat{\beta})$, of $I(\beta)$ is thus obtained by replacing q_{lm} with $\exp(\hat{b}_{lm}) t_{lm}/n$.

Under model (23), the expected number of false events for a subject in category (l,m) with the time-exposure interval $[0,T]$ is $T\exp(d_{lm})$, for $l = 0,1$ and $m = 1, \cdots, k$, where $d_{1k} = a_0 + a_1$, $d_{0k} = a_0$, $d_{1m} = a_0 + a_1 + \gamma_m$ and $d_{0m} = a_0 + \gamma_m$ for $1 \le m \le k-1$. The conditional distribution of Y given $A = a$, T, and $Z \simeq (l,m)$ is Binomial (a, p_{lm}), where $p_{lm} = \exp(b_{lm})/(\exp(b_{lm}) + \exp(d_{lm}))$ for $a \ge 1$. Then $\mathrm{Var}(Y|A = a, Z \simeq (l,m))$ can be estimated by $a\hat{p}_{lm}(1-\hat{p}_{lm})$, where $\hat{p}_{lm} = e^{\hat{b}_{lm}}/(e^{\hat{b}_{lm}} + e^{\hat{d}_{lm}})$, and $\mathrm{Var}(T|A = a, Z \simeq (l,m))$

can be estimated by $v_{a,l,m} = t_{2,a,l,m}/n_{alm} - (t_{alm}/n_{alm})^2$. By (24) and the discussion that follows, $\Sigma_{A1}(\beta)$ can be estimated by

$$\hat{\Sigma}_{A1}(\hat{\beta}) = \sum_{a,l,m} \hat{\rho}(a,l,m) \frac{1-\hat{\rho}^v(a,l,m)}{\hat{\rho}^v(a,l,m)} G_{lm}\{a\hat{p}_{lm}(1-\hat{p}_{lm}) + v_{a,l,m}\}, \tag{26}$$

where $\hat{\rho}(a,l,m) = n_{alm}/n$, $\hat{\rho}^v(a,l,m) = n^v_{alm}/n_{alm}$ and G_{lm} be the value of $G_i = z_i^{\otimes 2}$ when subject i belongs to the category (l,m). Hence the covariance matrix of $\hat{\beta}_{I1}$ can be estimated by $\hat{I}^{-1}(\hat{\beta}) + \hat{I}^{-1}(\hat{\beta})\hat{\Sigma}_{A1}(\hat{\beta})\hat{I}^{-1}(\hat{\beta})$ using the automated data.

Estimation with $\hat{\beta}_{A2}$ using the automated data. The estimating equations (25) are equivalent to the following nonlinear equations for $\{b_{lm}$, for $l = 0, 1$, $m = 1, \cdots, k\}$,

$$\sum_{m=1}^{k} \left\{ \sum_{a\in A} y_{alm}\lambda_{alm} - e^{b_{lm}} t_{lm} + \frac{e^{b_{lm}}}{e^{b_{lm}} + e^{\hat{d}_{1m}}} \left(\alpha_{lm} - \sum_{a\in A} a n_{alm}\lambda_{alm} \right) \right\} = 0,$$

$$\sum_{l=0,1} \left\{ \sum_{a\in A} y_{alm}\lambda_{alm} - e^{b_{lm}} t_{lm} + \frac{e^{b_{lm}}}{e^{b_{lm}} + e^{\hat{d}_{lm}}} \left(\alpha_{lm} - \sum_{a\in A} a n_{alm}\lambda_{alm} \right) \right\} = 0,$$

for $l = 0, 1$ and $m = 1, \ldots, k-1$.

Since $\mathrm{Var}\left\{S_\beta(Y|X)|Z \simeq (l,m), T, A = a\right\} = a p_{lm}(1-p_{lm})G_{lm}$, $\Sigma_{A2}(\beta)$ can be consistently estimated by

$$\hat{\Sigma}_{A2}(\hat{\beta}) = \sum_{a,l,m} \hat{\rho}(a,l,m) \frac{1-\hat{\rho}^v(a,l,m)}{\hat{\rho}^v(a,l,m)} a\hat{p}_{lm}(1-\hat{p}_{lm})G_{lm}.$$

Hence the covariance matrix of $\hat{\beta}_{A2}$ can be estimated by $\hat{I}^{-1}(\hat{\beta}) + \hat{I}^{-1}(\hat{\beta})\hat{\Sigma}_{A2}(\hat{\beta})\hat{I}^{-1}(\hat{\beta})$ using the automated data.

Remark In the special case where $\rho(\alpha,l,m) \approx 0$ for $\alpha \geq 2$, a much simpler formula for the variance estimator of the log relative risk can be derived. For example in the vaccine safety study, the adverse-event rate is very small. Let

$$w_m = \frac{\alpha_{0m}\alpha_{1m}y_{0m}y_{1m}}{\alpha_{0m}y_{0m}n^v_{1m} + \alpha_{1m}y_{1m}n^v_{0m}} \Bigg/ \sum_{m=1}^{k} \frac{\alpha_{0m}\alpha_{1m}y_{0m}n^v_{1m}}{\alpha_{0m}y_{0m}n^v_{1m} + \alpha_{1m}y_{1m}n^v_{0m}}.$$

Then an estimate of variance of $\hat{b}_1$ is given by

$$\widehat{\mathrm{Var}(\hat{b}_1)} = \sum_{m=1}^{k} w_m^2 \left(\frac{1}{y_{1m}} - \frac{1}{n^v_{1m}} + \frac{1}{\alpha_{1m}} + \frac{1}{y_{0m}} - \frac{1}{n^v_{0m}} + \frac{1}{\alpha_{0m}} \right), \tag{27}$$

which is the weighted sum of the estimated variances for the estimated log relative rate of the exposed versus the unexposed over k groups. The details of deviation are given in the Appendix B.

5 A simulation study

We conduct a simulation study to examine the finite sample performance of the IPW estimator $\hat{\beta}_{I1}$ and the AIPW estimator $\hat{\beta}_{A2}$. We consider the Poisson regression model (22). The covariates Z_1 and Z_2 are generated from the Bernoulli distributions with the probability of success equals to 0.4 and 0.5 respectively. The exposure time T is generated from a uniform distribution on $[0, 10]$. Given $Z = (Z_1, Z_2)$ and T, the outcome variable Y follows

a Poisson distribution with mean $T\exp(b_0 + b_1Z_1 + \theta Z_2)$ where $b_0 = -0.5$, $b_1 = -0.8$ and $\theta = -0.6$, and W follows a Poisson distribution with mean $T\exp(a_0 + a_1Z_1 + \gamma Z_2)$ where $a_0 = -1.3$, $a_1 = -1.1$, $\gamma = -1$. We set $A = Y + W$.

Four models for the validation sample are considered. Under Model 1, the validation sample is a simple random sample with probability $\pi_i = 0.4$. Model 2 considers $\pi_i = 0.6$. In Model 3, the validation probability only depends on A through the logistic regression model $\text{logit}\{\pi_i(X, A)\} = A - 0.5$ where $X = (Z, T)$. In Model 4, the validation probability depends on A and Z_1 through the logistic regression model $\text{logit}\{\pi_i(X, A)\} = A - Z_1 - 0.5$.

Tables 1 and 2 present the simulation results for $n = 50, 100, 300, 500$ and 800. Each entry of the tables is based on 1000 simulation runs. Tables 1 and 2 summarize the bias (Bias), the empirical standard error (SSE), the average of the estimated standard error (ESE), and the empirical coverage probability (CP) of 95% confidence intervals of $\hat{\beta}_{I1}$ and $\hat{\beta}_{A2}$ for $\beta = (b_0, b_1, \theta)$. We also compare the performance of the estimators $\hat{\beta}_{I1}$ and $\hat{\beta}_{A2}$ with the complete-case (CC) estimator $\hat{\beta}_C$ obtained by simply deleting subjects with missing values of Y_i. As a gold standard, we present the estimation results for the full data where all the values of Y_i are fully observed. Table 1 presents the results under Model 1 and 2, and Table 2 shows the results under Model 3 and 4.

Table 1 shows that under Model 1 and Model 2, the bias of all estimators is very small at a level comparable with that of the full data estimator. The bias decreases with increased sample size and the increased level of the validation probability. The empirical standard errors are in good agreement with the corresponding estimated standard errors, except for the IPW estimator when $n \leq 100$ and $\pi \leq 0.6$. Among them, AIPW has the smallest standard errors for all parameters and sample sizes concerned. The coverage probabilities of the confidence intervals for b_0, b_1 and θ are close to the nominal level 95%. When the sample size and the validation probability are both small, for example, $n = 50$ and $\pi = 0.4$, the IPW has large bias and is unstable but the AIPW still performs well.

Table 2 gives the results under Model 3 and Model 4. The bias remains small for $\hat{\beta}_{I1}$ and $\hat{\beta}_{A2}$. The empirical standard errors are also close to the corresponding estimated standard errors. The coverage probabilities remain close to the nominal level 95% for all IPW and AIPW estimators. However, the complete-case estimator yields larger bias and incorrect coverage probability because of the association between the validation probability and the auxiliary variable A and/or the covariate Z_1, in which case the missing is not missing completely at random. The AIPW performs better than IPW with smaller standard errors.

6 An Application

A community-based, nonrandomized, open-label influenza vaccine (CAIV-T) study was conducted in Temple-Belton, Texas during the 2000-2001 influenza season. The total 11,606 healthy children aged 18 months - 18 years were involved in this study and about 20% of them received a single dose of CAIV-T in 2000. The primary clinical outcome was based on an nonspecific case definition called medically attended acute respiratory infection (MAARI), which included all International Classification of Diseases, Ninth Revision, Clinical Modification diagnoses codes (ICD-9 codes 381-383, 460-487) for upper and lower respiratory tract infections, otitis media and sinusitis. MAARI outcomes and demographic data were extracted from the Scott & White Health Plan administrative database. For each visit, one or two International Classification of Diseases, Ninth Revision, Clinical Modification diagnosis codes were listed. Visits for which asthma diagnosis codes alone

Table 1 Simulation comparison of the IPW estimator $\hat{\beta}_{I1}$, the AIPW estimator $\hat{\beta}_{A2}$ and the complete-case (CC) estimator $\hat{\beta}_C$ under various sample sizes and selection probabilities

		b_0				b_1				θ			
n		**Bias**	**SSE**	**ESE**	**CP**	**Bias**	**SSE**	**ESE**	**CP**	**Bias**	**SSE**	**ESE**	**CP**
						Model 1: $\pi_i = .4$							
50	IPW	-.0415	.3561	.1839	.851	-.2175	1.6737	.3354	.864	-.1610	1.2201	.2962	.847
	AIPW	-.0110	.2213	.1664	.890	-.0062	.3099	.2873	.943	-.0186	.3076	.2551	.929
	CC	-.0246	.3398	.2738	.938	-.1515	1.6082	.4730	.968	-.1038	1.1709	.4187	.959
100	IPW	-.0650	.1815	.1404	.870	-.0548	.3120	.2458	.891	-.0249	.2653	.2161	.898
	AIPW	-.0094	.1376	.1187	.914	-.0024	.2284	.1988	.926	.0027	.1994	.1780	.925
	CC	-.0240	.1728	.1685	.948	-.0086	.3086	.2981	.960	.0031	.2556	.2581	.946
300	IPW	-.0368	.0936	.0874	.931	-.0209	.1535	.1460	.946	-.0022	.1419	.1286	.929
	AIPW	-.0027	.0732	.0712	.946	-.0028	.1233	.1165	.940	.0005	.1130	.1046	.938
	CC	-.0092	.0919	.0935	.960	-.0012	.1627	.1634	.952	.0040	.1438	.1432	.952
500	IPW	-.0183	.0698	.0671	.938	-.0172	.1159	.1128	.943	-.0083	.1069	.0993	.933
	AIPW	.0022	.0566	.0550	.936	-.0022	.0956	.0902	.943	-.0068	.0867	.0811	.930
	CC	.0006	.0704	.0716	.949	-.0059	.1268	.1255	.949	-.0046	.1135	.1103	.942
800	IPW	-.0126	.0538	.0527	.942	-.0134	.0862	.0889	.950	-.0029	.0759	.0779	.947
	AIPW	.0011	.0433	.0435	.952	-.0047	.0720	.0713	.956	-.0020	.0638	.0640	.951
	CC	.0002	.0562	.0565	.948	-.0051	.0974	.0990	.958	-.0013	.0844	.0869	.958
						Model 2: $\pi_i = .6$							
50	IPW	-.0316	.2079	.1714	.926	-.0934	.8426	.3112	.944	-.0563	.3320	.2690	.937
	AIPW	-.0072	.1723	.1591	.941	-.0105	.2893	.2772	.950	-.0172	.2653	.2440	.948
	CC	-.0126	.1973	.1949	.959	-.0594	.8369	.3512	.967	-.0278	.3213	.3044	.959
100	IPW	-.0366	.1399	.1259	.926	-.0420	.2363	.2192	.944	-.0100	.2103	.1911	.925
	AIPW	-.0121	.1206	.1133	.941	-.0107	.2069	.1921	.944	.0078	.1764	.1700	.940
	CC	-.0142	.1370	.1345	.947	-.0216	.2379	.2370	.961	.0030	.2103	.2072	.949
300	IPW	-.0138	.0742	.0728	.944	-.0194	.1267	.1250	.957	-.0049	.1064	.1096	.964
	AIPW	-.0030	.0650	.0651	.948	-.0044	.1136	.1093	.949	.0005	.0960	.0974	.956
	CC	-.0017	.0763	.0759	.946	-.0118	.1345	.1328	.951	-.0035	.1147	.1169	.957
500	IPW	-.0069	.0571	.0555	.945	-.0096	.0946	.0965	.947	-.0094	.0866	.0844	.953
	AIPW	.0029	.0495	.0496	.942	-.0032	.0856	.0841	.947	-.0076	.0757	.0749	.955
	CC	.0013	.0577	.0581	.947	-.0034	.1024	.1019	.949	-.0086	.0906	.0899	.954
800	IPW	-.0072	.0437	.0438	.954	-.0069	.0754	.0763	.956	-.0025	.0692	.0664	.947
	AIPW	-.0011	.0401	.0393	.951	-.0019	.0693	.0665	.943	-.0015	.0626	.0590	.931
	CC	-.0012	.0452	.0460	.958	-.0026	.0805	.0806	.952	-.0024	.0723	.0709	.952
						Full data: $\pi_i = 1$							
50		-.0079	.1510	.1466	.952	-.0182	.2691	.2618	.948	-.0104	.2264	.2263	.957
100		-.0079	.1068	.1024	.943	-.0075	.1841	.1798	.948	-.0039	.1560	.1577	.949
300		-.0019	.0596	.0583	.950	-.0081	.1032	.1023	.936	.0001	.0934	.0898	.932
500		.0006	.0452	.0450	.951	-.0041	.0783	.0788	.950	.0014	.0656	.0693	.960
800		-.0004	.0343	.0356	.951	.0025	.0612	.0622	.938	-.0006	.0532	.0547	.955

Table 2 Simulation comparison of the IPW estimator $\hat{\beta}_{I1}$, the AIPW estimator $\hat{\beta}_{A2}$ and the complete-case (CC) estimator $\hat{\beta}_C$ under various sample sizes and selection probabilities

		b_0				b_1				θ			
n		**Bias**	**SSE**	**ESE**	**CP**	**Bias**	**SSE**	**ESE**	**CP**	**Bias**	**SSE**	**ESE**	**CP**
						Model 3							
50	IPW	.0081	.1609	.1502	.949	-.0034	.5535	.2790	.954	-.0116	.2592	.2400	.963
	AIPW	-.0070	.1543	.1486	.950	-.0134	.2715	.2690	.958	-.0185	.2364	.2330	.955
	CC	.0230	.1529	.1504	.938	.0798	.5441	.2835	.940	.0648	.2367	.2414	.952
100	IPW	-.0052	.1145	.1077	.948	-.0001	.2073	.2014	.959	.0030	.1789	.1724	.948
	AIPW	-.0124	.1077	.1041	.947	-.0085	.1869	.1840	.957	.0050	.1636	.1606	.948
	CC	.0221	.1074	.1054	.939	.1023	.1830	.1937	.924	.0828	.1625	.1664	.915
300	IPW	-.0011	.0617	.0614	.951	-.0044	.1176	.1157	.952	.0019	.1009	.0993	.953
	AIPW	-.0023	.0582	.0588	.956	-.0051	.1056	.1036	.954	.0022	.0936	.0910	.944
	CC	.0295	.0577	.0596	.924	.1051	.1070	.1095	.824	.0823	.0925	.0946	.861
500	IPW	.0018	.0451	.0473	.958	-.0037	.0853	.0895	.958	-.0069	.0765	.0767	.945
	AIPW	.0009	.0430	.0452	.957	-.0032	.0793	.0798	.947	-.0066	.0689	.0702	.951
	CC	.0317	.0429	.0459	.903	.1077	.0788	.0844	.763	.0737	.0704	.0730	.839
800	IPW	-.0006	.0374	.0375	.951	-.0030	.0671	.0708	.962	.0004	.0617	.0605	.946
	AIPW	-.0003	.0362	.0358	.949	-.0031	.0623	.0631	.954	-.0012	.0577	.0554	.935
	CC	.0315	.0353	.0364	.863	.1065	.0630	.0667	.633	.0786	.0568	.0576	.721
						Model 4							
50	IPW	.0053	.1627	.1504	.948	.0825	.3531	.2832	.913	-.0057	.2736	.2405	.948
	AIPW	-.0085	.1549	.1489	.950	-.0122	.2746	.2752	.966	-.0138	.2395	.2340	.961
	CC	.2295	.2640	.0855	.531	.4513	.3805	.1760	.517	.2954	.3285	.1409	.536
100	IPW	-.0050	.1168	.1085	.939	.0481	.2350	.2130	.922	.0016	.1884	.1761	.940
	AIPW	-.0130	.1083	.1043	.943	-.0067	.1920	.1885	.950	.0066	.1648	.1613	.949
	CC	.0196	.1077	.1063	.943	.2010	.1946	.2087	.820	.0900	.1645	.1702	.910
300	IPW	-.0001	.0630	.0624	.945	-.0001	.1323	.1311	.955	-.0011	.1043	.1038	.946
	AIPW	-.0020	.0588	.0588	.951	-.0052	.1095	.1059	.952	.0012	.0950	.0913	.931
	CC	.0271	.0582	.0601	.930	.2020	.1060	.1173	.576	.0894	.0939	.0965	.863
500	IPW	.0012	.0457	.0480	.951	-.0007	.0966	.1010	.966	-.0054	.0801	.0799	.948
	AIPW	.0006	.0433	.0453	.956	-.0010	.0821	.0813	.941	-.0059	.0697	.0704	.950
	CC	.0291	.0434	.0463	.912	.2047	.0817	.0903	.364	.0815	.0711	.0745	.820
800	IPW	-.0006	.0381	.0380	.949	.0004	.0761	.0794	.967	.0000	.0636	.0630	.947
	AIPW	-.0002	.0362	.0359	.949	-.0016	.0640	.0641	.955	-.0014	.0574	.0555	.942
	CC	.0288	.0356	.0367	.885	.2039	.0644	.0714	.166	.0864	.0570	.0588	.673

were noted, without another MAARI code, were excluded. More details about this study can be found in Halloran et al. (2003).

Any children representing with history of fever and any respiratory illness were eligible to have a throat swab for influenza virus culture. The decision to obtain specimens was made irrespective of whether a patient had received CAIV-T. The specific case definition was culture-confirmed influenza. Table 3 taken from Halloran et al. (2003) contains information on the number of children in three age groups, the number of children who are vaccinated versus unvaccinated, the number of nonspecific MAARI cases, the number of cultures performed, and the number of cultures positive for each group.

With the method developed in Section 4 for Poisson regression, we compare the risk of developing MAARI for children who received CAIV-T to the risk for children who had never received CAIV-T using the automated information provided in Table 3. The number of nonspecific MAARI cases extracted using the ICD-9 codes is the auxiliary outcome A, whereas the actual number of influenza cases Y is the outcome of interest. Let Z_1 be the treatment indicator (1=vaccine and 0=placebo). Let $Z_2 = (\eta_1, \eta_2)$ be the dummy variables indicating three age groups, where $\eta_1 = 1$ if the age is in the range 1.5–4, $\eta_1 = 0$, otherwise, and $\eta_2 = 1$ if the age is in the range 5–9, $\eta_2 = 0$, otherwise. The reference group is the age 10–18. The exposure time for all children is taken as $T = 1$ year.

Consider a Poisson regression model with mean $T \exp(b_0 + b_1 Z_1 + \theta_1 \eta_1 + \theta_2 \eta_2)$. Using the IPW estimator $\hat{\beta}_{I1}$, the estimates (standard errors) are $\hat{b}_0 = -0.7659$ ($\hat{\sigma}_{b_0} = 0.1046$), $\hat{b}_1 = -1.5830$ ($\hat{\sigma}_{b_1} = 0.5017$), $\hat{\theta}_1 = -0.5572$ ($\hat{\sigma}_{\theta_1} = 0.2111$) and $\hat{\theta}_2 = -0.0199$ ($\hat{\sigma}_{\theta_2} = 0.1472$). The age-adjusted relative rate (RR) in the vaccinated group compared with the unvaccinated group equals $\exp(\hat{b}_1) = \exp(-1.5830) = 0.2054$, which means that the rate of developing MAARI for the vaccinated group is 20% of that for the unvaccinated group. In terms of the vaccine efficacy $\text{VE} = 1 - \text{RR} = 0.7946$, this represents about 80% reduction in the risk of developing MAARI for the vaccinated group compared to the unvaccinated group. The 95% confidence interval of RR obtained by using the delta method is $(0.0768, 0.5490)$, showing clear evidence that the vaccinated children have less risk of influenza than the unvaccinated children. The 95% confidence interval for VE is $(0.4510, 0.9232)$.

Table 3 Study data for influenza epidemic season 2000-01, by age and vaccine group (from Halloran et al. 2003)

Age group (years)	Vaccine	No. of children	No. of MAARI cases	No. of MAARI cases cultured	No. of positive cultures
1.5-4	CAIV-T	537	389	16	0
	None	1844	1665	86	24
5-9	CAIV-T	807	316	17	2
	None	2232	1156	118	53
10-18	CAIV-T	937	219	19	3
	None	5249	1421	123	56
Total	CAIV-T	2281	924	52	5
	None	9325	4242	327	133

Using the AIPW estimator $\hat{\beta}_{A2}$, the estimates (standard errors) are $\hat{b}_0 = -2.0703$ ($\hat{\sigma}_{b_0} = 0.0851$), $\hat{b}_1 = -1.8072$ ($\hat{\sigma}_{b_1} = 0.3786$), $\hat{\theta}_1 = 0.6452$ ($\hat{\sigma}_{\theta_1} = 0.1966$) and $\hat{\theta}_2 = 0.6235$ ($\hat{\sigma}_{\theta_2} = 0.1265$). The age-adjusted relative rate (RR) is $\exp(\hat{b}_1) = \exp(-1.8072) = 0.1641$. The estimated VE is 0.8359 and the 95% confidence interval is $(0.6553, 0.9219)$. The estimator $\hat{\beta}_{A2}$ yields smaller standard errors and confidence intervals with more precision than using $\hat{\beta}_{I1}$.

This data was analyzed by Halloran et al. (2003) and Chu and Halloran (2004). Assuming the binary probability model for $P_\beta(Y|X)$ where X includes the vaccination status and age group indicators, and using the mean score method, Halloran et al. (2003) found that the estimated VE based on the nonspecific MAARI cases alone was 0.18 with 95% confidence interval of $(0.11, 0.24)$. The estimated VE by incorporating the surveillance cultures was 0.79 with 95% confidence interval of $(0.51, 0.91)$. Halloran et al. also reported sample-size-weighted VE= 0.77 with 95% confidence interval of $(0.48, 0.90)$. Chu and Halloran (2004) have developed a Bayesian method to estimate vaccine efficacy. By Chu and Halloran (2004), the estimated VE was 0.74 with 95% confidence interval $(0.50, 0.88)$ and estimated VE by the multiple imputation method was 0.71 with 95% confidence interval $(0.42, 0.86)$.

Our estimates of the vaccine efficacy are in line with the existing methods. The estimator $\hat{\beta}_{A2}$ yields smaller standard errors and therefore confidence intervals are more precise than the existing methods of Halloran et al. (2003) and Chu and Halloran (2004). Compared to the binary regression, Poisson regression model allows multiple recurrent MAARI cases for each child. Although for this particular application the exposure time is fixed at one year time interval, the proposed method is applicable to the situation where the length of exposure time may be different for different children.

7 Conclusions

In this paper, we investigated the mean score method, the IPW method and the AIPW method for the parametric probability regression model $P_\beta(Y|X)$ when outcome of interest Y is subject to missingness. The asymptotic distributions are derived for the IPW estimator and the AIPW estimator. The selection probability often needs to be estimated for the IPW estimator, and both the selection probability and the conditional expectation of the score function needs to be estimated for the AIPW estimator. We investigated the properties of the IPW estimator and the AIPW estimator when the selection probability and the conditional expectation are implemented differently.

An AIPW estimator is said to be fully augmented if the selection probability and the conditional expectation are estimated using the full set of observed variables; it is partially augmented if the selection probability and the conditional expectation are estimated using a subset of observed variables. Corollary 1 shows that the fully augmented AIPW estimator is more efficient than the partially augmented AIPW estimator. Corollary 2 shows that the AIPW estimator is more efficient than the IPW estimator. However, when the selection probability depends only on a set of discrete random variables, the IPW estimator obtained by estimating the selection probability nonparametrically with the cell frequencies is asymptotically equivalent to the AIPW estimator augmented using the same set of discrete random variables. Proposition 1 shows that the IPW estimator, the AIPW estimator and the mean score estimator are equivalent if the selection probability and the conditional expectation are estimated using same set of discrete random variables.

Applying the developed missing data methods, we derived the estimation procedures for Poisson regression model with missing outcomes based on auxiliary outcomes and a validated sample for true outcomes. By assuming the selection probability depending only on the observed discrete exposure variables, not on the continuous exposure time, we show that the IPW estimator and the AIPW estimator can be formulated to analyze data when only aggregated/summarized information are available. The simulation study shows that for a moderate sample size and selection probability, the IPW estimator and AIPW estimator perform better than the complete-case estimator. The AIPW estimator is more efficient and more stable than the IPW estimator. The proposed methods are applied to analyze a data set from for an influenza vaccine study conducted in Temple-Belton, Texas during the 2000-2001 influenza season. The data set presented in Table 3 only contains summarized information at categorical levels defined by the three age groups and vaccination status. The actual number of influenza cases (the number of positive cultures) out of the number of MAARI cases cultured, along with the number of MAARI cases, are available for each category. Our analysis using the AIPW approach shows that the age-adjusted relative rate in the vaccinated group compared to the unvaccinated group equals 0.1641, which represents about 84% reduction in the risk of developing MAARI for the vaccinated group compared to the unvaccinated group.

Appendix A

Proof of Proposition 1.

Since

$$\sum_{i=1}^{n}(1-\xi_i)\hat{E}\left\{S_\beta(Y|X_i)|Z_i,A_i\right\} = \sum_{i\in\bar{V}}\sum_{j\in V(Z_i,A_i)} S_\beta(Y_j|X_j)/n^V(Z_i,A_i)$$
$$= \sum_{i\in V}\left\{n^{\bar{V}}(Z_i,A_i)/n^V(Z_i,A_i)\right\} S_\beta(Y_i|X_i),$$

we have

$$\sum_{i=1}^{n} W_i^{E1} = \sum_{i\in V}\left(1+\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)}\right) S_\beta(Y_i|X_i) = \sum_{i=1}^{n}\frac{\xi_i}{\hat{\pi}_i^z} S_\beta(Y_i|X_i) = \sum_{i=1}^{n} W_i^{I1}. \quad \text{(A.1)}$$

This shows that the mean score estimator $\hat{\beta}_{E1}$ is the same as the IPW estimator $\hat{\beta}_{I1}$. Further, since

$$\sum_{i=1}^{n}\left(1-\frac{\xi_i}{\hat{\pi}_i^z}\right)\hat{E}\{S_\beta(Y|X_i)|Z_i,A_i\}$$
$$= \sum_{i\in\bar{V}}\sum_{j\in V(Z_i,A_i)}\frac{S_\beta(Y_j|X_j)}{n^V(Z_i,A_i)} - \sum_{i\in V}\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)}\sum_{j\in V(Z_i,A_i)}\frac{S_\beta(Y_j|X_j)}{n^V(Z_i,A_i)}$$
$$= \sum_{i\in V}\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)} S_\beta(Y_i|X_i) - \sum_{i\in V}\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)} S_\beta(Y_i|X_i) = 0,$$

we have $\sum_{i=1}^{n} W_i^{A1} = \sum_{i=1}^{n} W_i^{I1}$. Thus the AIPW estimator $\hat{\beta}_{A1}$, the IPW estimator $\hat{\beta}_{I1}$ and the mean score estimator $\hat{\beta}_{E1}$ are equivalent to each other.

Note that

$$\begin{aligned}
&\sum_{i=1}^{n}\left(1-\frac{\xi_i}{\hat{\pi}_i^z}\right)\hat{E}\{S_\beta(Y|X_i)|X_i,A_i\}\\
&=\sum_{i\in\bar{V}}\sum_{j\in V(X_i,A_i)}\frac{S_\beta(Y_j|X_j)}{n^V(X_i,A_i)}-\sum_{i\in V}\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)}\sum_{j\in V(X_i,A_i)}\frac{S_\beta(Y_j|X_j)}{n^V(X_i,A_i)}\\
&=\sum_{i\in V}\frac{n^{\bar{V}}(X_i,A_i)}{n^V(X_i,A_i)}S_\beta(Y_i|X_i)-\sum_{i\in V}\frac{n^V(X_i,A_i)}{n^V(X_i,A_i)}\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)}S_\beta(Y_i|X_i)\\
&=\sum_{i\in V}\left\{\frac{n^{\bar{V}}(X_i,A_i)}{n^V(X_i,A_i)}-\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)}\right\}S_\beta(Y_i|X_i),
\end{aligned}\tag{A.2}$$

which is not zero unless Z_i^c is linearly related to Z_i and in this case β is not identifiable. Hence the AIPW estimator $\hat{\beta}_{A2}$ is different from the AIPW estimator $\hat{\beta}_{A1}$.

By (A.1) and (A.2), we have

$$\begin{aligned}
\sum_{i=1}^{n}W_i^{A2}&=\sum_{i\in V}\left\{\frac{n(Z_i,A_i)}{n^V(Z_i,A_i)}+\frac{n^{\bar{V}}(X_i,A_i)}{n^V(X_i,A_i)}-\frac{n^{\bar{V}}(Z_i,A_i)}{n^V(Z_i,A_i)}\right\}S_\beta(Y_i|X_i)\\
&=\sum_{i\in V}\left\{1+\frac{n^{\bar{V}}(X_i,A_i)}{n^V(X_i,A_i)}\right\}S_\beta(Y_i|X_i)=\sum_{i=1}^{n}W_i^{I2}.
\end{aligned}$$

Following the same arguments leading to (A.1), we also have $\sum_{i=1}^n W_i^{E2}=\sum_{i=1}^n W_i^{I2}$. Hence, the estimators $\hat{\beta}_{I2}$, $\hat{\beta}_{E2}$ and $\hat{\beta}_{A2}$ are equivalent. By following the steps in (A.2), we also have $\sum_{i=1}^n\left(1-\frac{\xi_i}{\hat{\pi}_i}\right)\hat{E}\left\{S_\beta(Y|X_i)|X_i,A_i\right\}=0$. Hence, $\hat{\beta}_{A3}$ is the same as $\hat{\beta}_{I2}$. Therefore, these are essentially two different estimators. □

Proof of Theorem 1.

Applying the first order Taylor expansion, $\tilde{\pi}_i-\pi_i=(\partial\pi(X_i,A_i,\psi_0)/\partial\psi)'\left(\hat{\psi}-\psi_0\right)+o_p\left(n^{-1/2}\right)$. From (13), we have

$$n^{-1/2}U_I=n^{-1/2}\sum_{i=1}^{n}\frac{\xi_i}{\pi_i}S_\beta(Y_i|X_i)-n^{-1/2}\sum_{i=1}^{n}\frac{\tilde{\pi}_i-\pi_i}{\tilde{\pi}_i\pi_i}\xi_iS_\beta(Y_i|X_i)\tag{A.3}$$

The second term of (A.3) is

$$\begin{aligned}
&n^{-1/2}\sum_{i=1}^{n}\frac{\tilde{\pi}_i-\pi_i}{\tilde{\pi}_i\pi_i}\xi_iS_\beta(Y_i|X_i)\\
&=n^{-1/2}\sum_{i=1}^{n}{\pi_i}^{-2}\xi_iS_\beta(Y_i|X_i)\left(\frac{\partial\pi(X_i,A_i,\psi_0)}{\partial\psi}\right)'\left(\hat{\psi}-\psi_0\right)\\
&=E\left\{{\pi_i}^{-2}\xi_iS_\beta(Y_i|X_i)\left(\frac{\partial\pi(X_i,A_i,\psi_0)}{\partial\psi}\right)'\right\}n^{1/2}\left(\hat{\psi}-\psi_0\right)+o_p(1)
\end{aligned}\tag{A.4}$$

By (11), (A.3) and (A.4), we have

$$n^{-1/2}U_I=n^{-1/2}\sum_{i=1}^{n}\left(\frac{\xi_i}{\pi_i}S_\beta(Y_i|X_i)-O_i\right)+o_p(1)=n^{-1/2}\sum_{i=1}^{n}Q_i^I+o_p(1).$$

Now consider the AIPW estimator $\hat{\beta}_A$ based on solving the estimating equation (14). For simplicity, we denote $E_a\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ by E_i and $\tilde{E}\left\{S_\beta(Y|X_i)|X_i,A_i\right\}$ by $\tilde{E}_i$. We note that

$$\begin{aligned} n^{-1/2}U_A &= n^{-1/2}\sum_{i=1}^{n}\left[\frac{\xi_i}{\pi_i}S_\beta(Y_i|X_i)+\left(1-\frac{\xi_i}{\pi_i}\right)E_i\right] \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\left[\left(\frac{\xi_i}{\tilde{\pi}_i}-\frac{\xi_i}{\pi_i}\right)\left\{S_\beta(Y_i|X_i)-E_i\right\}+\left(1-\frac{\xi_i}{\tilde{\pi}_i}\right)\left(\tilde{E}_i-E_i\right)\right]. \end{aligned}$$

Suppose that $\tilde{\pi}_i$ and $\tilde{E}_i$ are the estimates of π_i and E_i based on some parametric or nonparametric models. Then it can be shown using Taylor expansion and standard probability arguments that the second term is at the order of $o_p(1)$ under MAR I. Hence

$$n^{-1/2}U_A = n^{-1/2}\sum_{i=1}^{n}\left[\frac{\xi_i}{\pi_i}S_\beta(Y_i|X_i)+\left(1-\frac{\xi_i}{\pi_i}\right)E_i\right]+o_p(1).$$

It can be shown that under MAR I, $n^{-1}\partial U_I/\partial\beta \xrightarrow{P} I(\beta)$ and $n^{-1}\partial U_A/\partial\beta \xrightarrow{P} I(\beta)$. By routine derivations, we have

$$\begin{aligned} n^{1/2}\left(\hat{\beta}_I-\beta\right) &= I^{-1}(\beta)n^{-1/2}U_I+o_p(1) = I^{-1}(\beta)n^{-1/2}\sum_{i=1}^{n}Q_i^I+o_p(1), \\ n^{1/2}\left(\hat{\beta}_A-\beta\right) &= I^{-1}(\beta)n^{-1/2}U_A+o_p(1) = I^{-1}(\beta)n^{-1/2}\sum_{i=1}^{n}Q_i^A+o_p(1). \end{aligned}$$

By the central limit theorem, both $n^{1/2}\left(\hat{\beta}_I-\beta\right)$ and $n^{1/2}\left(\hat{\beta}_A-\beta\right)$ have asymptotically normal distributions with mean zero and covariances equal to $I^{-1}(\beta)\mathrm{Var}\left(Q_i^I\right)I^{-1}(\beta)$ and $I^{-1}(\beta)\mathrm{Var}\left(Q_i^A\right)I^{-1}(\beta)$, respectively.

Next, we examine the covariance matrices $\mathrm{Var}\left(Q_i^I\right)$ and $\mathrm{Var}\left(Q_i^A\right)$ to understand the efficiency gain of $\hat{\beta}_A$ over $\hat{\beta}_I$. Note that $Q_i^I=\xi_i/\pi_iS_\beta(Y_i|X_i)-O_i$ and $Q_i^A=\xi_i/\pi_iS_\beta(Y_i|X_i)+(1-\xi_i/\pi_i)E_i$. Denote $A_i=\xi_i/\pi_iS_\beta(Y_i|X_i)$ and $B_i=(1-\xi_i/\pi_i)E_i$. Then $Q_i^I=Q_i^A-B_i-O_i$. Under MAR I, $\mathrm{Cov}\left(Q_i^A,O_i\right)=\mathrm{E}\left(Q_i^AO_i\right)=E\left\{E(Q_i^A|\xi_i,X_i,A_i)O_i\right\}=E\{E_iO_i\}=E\{E_iE(O_i|X_i,A_i)\}=0$, and

$$\begin{aligned} \mathrm{Cov}(Q_i^A,B_i) &= E\left[\left\{\frac{\xi_i}{\pi_i}S_\beta(Y_i|X_i)+\left(1-\frac{\xi_i}{\pi_i}\right)E_i\right\}\left(1-\frac{\xi_i}{\pi_i}\right)E_i\right] \\ &= E\left[\left(1-\frac{\xi_i}{\pi_i}\right)^2E_i^2\right]-E\left[\left(1-\frac{\xi_i}{\pi_i}\right)^2S_\beta(Y_i|X_i)E_i\right] \\ &\quad +E\left[\left(1-\frac{\xi_i}{\pi_i}\right)S_\beta(Y_i|X_i)E_i\right]=0. \end{aligned}$$

Hence, $\mathrm{Cov}\left(Q_i^A,B_i+O_i\right)=0$. It follows that $\mathrm{Var}(Q_i^I)=\mathrm{Var}\left(Q_i^A\right)+\mathrm{Var}(B_i+O_i)$. Since $Q_i^A=S_\beta(Y_i|X_i)-\left(1-\frac{\xi_i}{\pi_i}\right)\left\{S_\beta(Y_i|X_i)-E_i\right\}$ and the two terms are uncorrelated under MAR I, we have $\mathrm{Var}\left(Q_i^A\right)=\mathrm{Var}\left(S_\beta(Y_i|X_i)\right)+\mathrm{Var}\left((1-\xi_i/\pi_i)\left\{S_\beta(Y_i|X_i)-E_i\right\}\right)$, where the first term equals $I(\beta)$. This completes the proof of Theorem 1. □

Proof of Corollary 1.

Let $Q_i^{A1} = \xi_i/\pi_i^z S_\beta(Y_i|X_i) + (1-\xi_i/\pi_i^z) E\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}$, and $Q_i^{A2} = \xi_i/\pi_i^z S_\beta(Y_i|X_i) + \left(1-\xi_i/\pi_i^z\right) E\left\{S_\beta(Y|X_i)|X_i, A_i\right\}$. By (16),

$$\begin{aligned}
\mathrm{Var}\left(Q_i^{A1}\right) &= \mathrm{Var}\left(S_\beta(Y_i|X_i)\right) + \mathrm{Var}\left(\left(1-\frac{\xi_i}{\pi_i^z}\right)\left[S_\beta(Y_i|X_i) - E\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}\right]\right), \\
\mathrm{Var}\left(Q_i^{A2}\right) &= \mathrm{Var}\left(S_\beta(Y_i|X_i)\right) + \mathrm{Var}\left(\left(1-\frac{\xi_i}{\pi_i^z}\right)\left[S_\beta(Y_i|X_i) - E\left\{S_\beta(Y|X_i)|X_i, A_i\right\}\right]\right).
\end{aligned}$$

The second term of $\mathrm{Var}\left(Q_i^{A1}\right)$ equals $\Sigma_{A1}(\beta)$ and the second term of $\mathrm{Var}\left(Q_i^{A2}\right)$ equals $\Sigma_{A2}(\beta)$. Then it follows from the main results in Theorem 1 that (17) and (18) hold.

Also by Theorem 1, the difference in the variances of Q_i^{A1} and Q_i^{A2} contributes to the difference in the asymptotic variances of $\hat{\beta}_{A1}$ and $\hat{\beta}_{A2}$. Since $E\left\{\left(1-\xi_i/\pi_i^z\right)^2|Y_i, X_i, A_i\right\} = E\left\{\left(1-\xi_i/\pi_i^z\right)^2|X_i, A_i\right\} = (1-\pi_i^z)/\pi_i^z$ under MAR I,

$$\begin{aligned}
\Sigma_{A2}(\beta) &= E\left(\frac{1-\pi_i^z}{\pi_i^z}\left\{E\left\{S_\beta^2(Y_i|X_i)|X_i, A_i\right\} - \left[E\left\{S_\beta(Y|X_i)|X_i, A_i\right\}\right]^2\right\}\right) \\
&= E\left(\frac{1-\pi_i^z}{\pi_i^z}\left\{E\left\{S_\beta^2(Y_i|X_i)|Z_i, A_i\right\} - \left[E\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}\right]^2\right\}\right) \\
&\quad -E\left(\frac{1-\pi_i^z}{\pi_i^z}\left\{\left[E\left\{S_\beta(Y|X_i)|X_i, A_i\right\}\right]^2 - \left[E\left\{S_\beta(Y|X_i)|Z_i, A_i\right\}\right]^2\right\}\right) \\
&= \Sigma_{A1}(\beta) - E\left(\frac{1-\pi_i^z}{\pi_i^z}\mathrm{Var}\left(\left[E\left\{S_\beta(Y|X_i)|X_i, A_i\right\}\right]|Z_i, A_i\right)\right),
\end{aligned}$$

which is less than $\Sigma_{A1}(\beta)$ if the covariates Z_i is a proper subset of X_i. □

Proof of Corollary 2.

Consider the definitions of B_i and O_i given following (16). We note that

$$\begin{aligned}
&E\left\{{\pi_i}^{-1} S_\beta(Y_i|X_i)\left(\frac{\partial\pi(Z_i, A_i, \psi_0)}{\partial\psi}\right)'\right\} \\
&= E\left[{\pi_i}^{-1} E\left\{S_\beta(Y_i|X_i)|Z_i, A_i\right\}\left(\frac{\partial\pi(Z_i, A_i, \psi_0)}{\partial\psi}\right)'\right] \\
&= \sum_{z,a}\rho(z,a)\pi^{-1}(z,a,\psi_0)E\left\{S_\beta(Y|X)|Z=z, A=a\right\}\left(\frac{\partial\pi(z,a,\psi_0)}{\partial\psi}\right)',
\end{aligned}$$

and

$$\begin{aligned}
n^{-1/2}\sum_{i=1}^n S_i^\psi &= n^{-1/2}\sum_{i=1}^n \frac{(\xi_i - \pi(Z_i, A_i, \psi_0))}{\pi(Z_i, A_i, \psi_0)(1-\pi(Z_i, A_i, \psi_0))}\frac{\partial\pi(Z_i, A_i, \psi_0)}{\partial\psi} \\
&= \sum_{z,a} n^{-1/2}\sum_{j=1}^n \frac{\left(\xi_j - \pi(z,a,\psi_0)\right) I\left(Z_j = z, A_j = a\right)}{\pi(z,a,\psi_0)(1-\pi(z,a,\psi_0))}\frac{\partial\pi(z,a,\psi_0)}{\partial\psi}.
\end{aligned}$$

From the discussions preceding Corollary 2, $\psi = \{\psi_{z,a}\}$ and $\pi(z,a,\psi) = \psi_{z,a}$, where $\psi_{z,a} = P(\xi_i = 1|Z_i = z, A_i = a)$ for all distinct pairs (z,a). Hence, $\frac{\partial\pi(z,a,\psi_0)}{\partial\psi}$ is a column vector with 1 in the position for $\psi_{z,a}$ and 0 elsewhere. And I^ψ is a diagonal matrix and its inverse matrix is also diagonal.

We have

$$n^{-1/2}\sum_{i=1}^{n} O_i = E\left\{\pi_i{}^{-1} S_\beta(Y_i|X_i)\left(\frac{\partial \pi(Z_i, A_i, \psi_0)}{\partial \psi}\right)'\right\}(I^\psi)^{-1} n^{-1/2}\sum_{i=1}^{n} S_i^\psi$$

$$= \sum_{z,a} E\{S_\beta(Y|X)|Z = z, A = a\} n^{-1/2}\sum_{j=1}^{n} \frac{\xi_j - \pi(z, a, \psi_0)}{\pi(z, a, \psi_0)} I\left(Z_j = z, A_j = a\right),$$

and

$$n^{-1/2}\sum_{i=1}^{n} B_i = -n^{-1/2}\sum_{i=1}^{n} \frac{(\xi_i - \pi(Z_i, A_i))}{\pi(Z_i, A_i)} E_i$$

$$= -n^{-1/2}\sum_{i=1}^{n} \frac{(\xi_i - \pi(Z_i, A_i))}{\pi(Z_i, A_i)} E\left\{S_\beta(Y_i|X_i)|X_i, A_i\right\}$$

$$= -\sum_{z,a} n^{-1/2}\sum_{j=1}^{n} \frac{\xi_j - \pi(z, a, \psi_0)}{\pi(z, a, \psi_0)} I\left(Z_j = z, A_j = a\right) E\left\{S_\beta(Y_j|X_j)|Z_j = z, Z_j^c, A_j = a\right\}.$$

Then (21) holds. It follows that $\hat{\beta}_A$ is more efficient than $\hat{\beta}_I$ unless $Var\{S_\beta(Y_j|X_j)|Z_j = z, A_j = a\} = 0$ for all (z, a) for which $P(Z_i = z, A_i = a) \neq 0$. □

Appendix B

Proof of the simplified variance formula (27)

The information matrix $I(\beta)$ is a $(k+1) \times (k+1)$ symmetric matrix given by

$$I(b_0, b_1, \theta) = \begin{bmatrix} \sum_{l,m} q_{lm} & \sum_{m=1}^{k} q_{1m} & q_{11} + q_{01} & \cdots & q_{1r} + q_{0r} \\ \sum_{m=1}^{k} q_{1m} & \sum_{m=1}^{k} q_{1m} & q_{11} & \cdots & q_{1r} \\ q_{11} + q_{01} & q_{11} & q_{11} + q_{01} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{1r} + q_{0r} & q_{1r} & 0 & \cdots & q_{1r} + q_{0r} \end{bmatrix},$$

where $r = k - 1$ and $q_{lm} = E(T_i e^{b_{lm}} I\{\text{individual } i \text{ in category } (l, m)\})$. For ease of presentation in the following, we drop the augments (b_0, b_1, θ) and use I for $I(b_0, b_1, \theta)$. Let I_{ij} be the cofactor of the (i, j)th element of I.

First we need to find the elements on the second row of the information matrix $I(b_0, b_1, \theta)$. Note that for a matrix A, $((a_{ij})_{n\times n})^{-1} = (A_{ji})_{n\times n}/|A|$, where A_{ij} is the cofactor of a_{ij} in the matrix A and $|A|$ is the determinant of A. Also note that for a block matrix

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

the determinant $|B| = |B_{22}||B_{11} - B_{12}B_{22}^{-1}B_{21}|$. We have

$$\begin{aligned}
|I| &= \prod_{m=1}^{k-1}(q_{1m}+q_{0m}) \\
&\quad \left| \begin{pmatrix} \sum_{lm} q_{lm} & \sum_{m=1}^{k} q_{1m} \\ \sum_{m=1}^{k} q_{1m} & \sum_{m=1}^{k} q_{1m} \end{pmatrix} - \begin{pmatrix} \sum_{l=0}^{1}\sum_{m=1}^{k-1} q_{lm} & \sum_{m=1}^{k-1} q_{1m} \\ \sum_{m=1}^{k-1} q_{1m} & \sum_{m=1}^{k-1} \frac{(q_{1m})^2}{q_{1m}+q_{0m}} \end{pmatrix} \right| \\
&= \prod_{m=1}^{k-1}(q_{1m}+q_{0m}) \begin{vmatrix} q_{1k}+q_{0k} & q_{1k} \\ q_{1k} & q_{1k} + \sum_{m=1}^{k-1} \frac{q_{0m}q_{1m}}{q_{1m}+q_{0m}} \end{vmatrix} \\
&= \prod_{m=1}^{k}(q_{1m}+q_{0m}) \sum_{m=1}^{k} \frac{q_{0m}q_{1m}}{q_{1m}+q_{0m}},
\end{aligned}$$

and

$$\begin{aligned}
I_{21} &= -\begin{vmatrix} \sum_{m=1}^{k} q_{1m} & q_{11}+q_{01} & \cdots & q_{1r}+q_{0r} \\ q_{11} & q_{11}+q_{01} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ q_{1r} & 0 & \cdots & q_{1r}+q_{0r} \end{vmatrix} \\
&= -\prod_{m=1}^{k-1}(q_{1m}+q_{0m}) \left[\sum_{m=1}^{k} q_{1m} - \sum_{m=1}^{k-1} q_{1m} \right] \\
&= -q_{1k} \prod_{m=1}^{k-1}(q_{1m}+q_{0m}).
\end{aligned}$$

Hence, the $(2,1)$th element of I^{-1} is

$$\left(I^{-1}\right)_{21} = \frac{I_{21}}{|I|} = -\frac{q_{1k}}{W(q_{1k}+q_{0k})},$$

where $W = \sum_{m=1}^{k} q_{0m}q_{1m}/(q_{1m}+q_{0m})$.

To calculate $(2,2)$th element of I^{-1}, we have

$$\begin{aligned}
I_{22} &= \begin{vmatrix} \sum_{l,m}^{k} q_{lm} & q_{11}+q_{01} & \cdots & q_{1r}+q_{0r} \\ q_{11}+q_{01} & q_{11}+q_{01} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ q_{1r}+q_{0r} & 0 & \cdots & q_{1r}+q_{0r} \end{vmatrix} \\
&= \prod_{m=1}^{k-1}(q_{1m}+q_{0m}) \left[\sum_{l,m} q_{lm} - \sum_{l=0}^{1}\sum_{m=1}^{k-1} q_{lm} \right] = \prod_{m=1}^{k}(q_{1m}+q_{0m}).
\end{aligned}$$

Hence, the $(2,2)$th element of I^{-1} is $\left(I^{-1}\right)_{22} = I_{22}/|I| = 1/W$.

To calculate $(2,3)$th element of I^{-1}, we have

$$\begin{aligned} I_{23} &= -\begin{vmatrix} \sum_{l,m} q_{lm} & \sum_{m=1}^{k} q_{1m} & q_{12}+q02 & \cdots & q_{1r}+q_{0r} \\ q_{11}+q_{01} & q_{11} & 0 & \cdots & 0 \\ q_{12}+q_{02} & q_{12} & q_{12}+q_{02} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{1r}+q_{0r} & q_{1r} & 0 & \cdots & q_{1r}+q_{0r} \end{vmatrix} \\ &= -\prod_{m=2}^{k-1}(q_{1m}+q_{0m}) \left| \begin{pmatrix} \sum_{lm} q_{lm} & \sum_{m=1}^{k} q_{1m} \\ q_{11}+q_{01} & q_{11} \end{pmatrix} \right. \\ &\quad \left. - \begin{pmatrix} \sum_{m=2}^{k-1}(q_{1m}+q_{0m}) & \sum_{m=2}^{k-1} q_{1m} \\ 0 & 0 \end{pmatrix} \right| \\ &= -\prod_{m=2}^{k-1}(q_{1m}+q_{0m})(q_{0k}q_{11}-q_{1k}q_{01}). \end{aligned}$$

Hence

$$(I^{-1})_{23} = \frac{I_{23}}{|I|} = -\frac{q_{11}}{W(q_{11}+q_{01})} + \frac{q_{1k}}{W(q_{1k}+q_{0k})}.$$

To calculate $(2,4)$th element of I^{-1}, we have

$$I_{24} = \begin{vmatrix} \sum_{l,m} q_{lm} & \sum_{m=1}^{k} q_{1m} & q_{11}+q_{01} & q_{13}+q_{03} & \cdots & q_{1r}+q_{0r} \\ q_{11}+q_{01} & q_{11} & q_{11}+q_{01} & 0 & \cdots & 0 \\ q_{12}+q_{02} & q_{12} & 0 & 0 & \cdots & 0 \\ q_{13}+q_{03} & q_{13} & 0 & q_{13}+q_{03} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{1r}+q_{0r} & q_{1r} & 0 & 0 & \cdots & q_{1r}+q_{0r} \end{vmatrix}.$$

By switching the 2nd and the 3rd row, we have

$$I_{24} = -n^{k} \begin{vmatrix} \sum_{l,m} q_{lm} & \sum_{m=1}^{k} q_{1m} & q_{11}+q_{01} & q_{13}+q_{03} & \cdots & q_{1r}+q_{0r} \\ q_{12}+q_{02} & q_{12} & 0 & 0 & \cdots & 0 \\ q_{11}+q_{01} & q_{11} & q_{11}+q_{01} & 0 & \cdots & 0 \\ q_{13}+q_{03} & q_{13} & 0 & q_{13}+q_{03} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{1r}+q_{0r} & q_{1r} & 0 & 0 & \cdots & q_{1r}+q_{0r} \end{vmatrix}.$$

Hence

$$\left(I^{-1}\right)_{24} = \frac{I_{24}}{|I|} = -\frac{q_{12}}{W(q_{12}+q_{02})} + \frac{q_{1k}}{W(q_{1k}+q_{0k})}.$$

In general, to calculate $I_{2(m+2)}$ for $m = 1, \cdots, k-1$, we can obtain a matrix with a $(k-2)\times(k-2)$ diagonal right lower block by switching rows of $I_{2(m+2)}$ even number of times when m is odd and by switching rows odd number of times when m is even. Then similar to calculating $\left(I^{-1}\right)_{24}$, we have

$$\left(I^{-1}\right)_{2(m+2)} = -\frac{q_{1m}}{W\left(q_{1m}+q_{0m}\right)} + \frac{q_{1k}}{W(q_{1k}+q_{0k})}.$$

For $l = 1$ and $1 \le m \le k$, the (i,j)th element of G_{lm} is $g_{ij} = 1$ for $(i,j) = (1,1),(1,2),(1,m+2),(2,1),(2,2),(2,m+2),(m+2,1),(m+2,2),(m+2,m+2)$, and $g_{ij} = 0$ elsewhere. We have the $(2,2)$th element of $I^{-1}G_{lm}I^{-1}$ as

$$\begin{aligned}&\left(I^{-1}\right)_{21}^2 + \left(I^{-1}\right)_{22}^2 + \left(I^{-1}\right)_{2(m+2)}^2 + 2\left(I^{-1}\right)_{21}\left(I^{-1}\right)_{2(m+2)}\\ &+2\left(I^{-1}\right)_{22}\left(I^{-1}\right)_{2(m+2)} + 2\left(I^{-1}\right)_{21}\left(I^{-1}\right)_{22}\\ &= \frac{1}{W^2}\left(\frac{q_{0m}}{q_{1m}+q_{0m}}\right)^2.\end{aligned}$$

Since for $l = 0$ and $1 \le m \le k$, $g_{ij} = 1$ for $(i,j) = (1,1),(1,m+2),(m+2,1),(m+2,m+2)$, and $g_{ij} = 0$ elsewhere, in this case the $(2,2)$th element of $I^{-1}G_{lm}I^{-1}$ is

$$\begin{aligned}&(I^{-1})_{21}^2 + (I^{-1})_{2(m+2)}^2 + 2(I^{-1})_{21}(I^{-1})_{2(m+2)}\\ &= ((I^{-1})_{21} + (I^{-1})_{2(m+2)})^2\\ &= \frac{1}{W^2}\left(\frac{q_{1m}}{q_{1m}+q_{0m}}\right)^2.\end{aligned}$$

Hence the $(2,2)$th element of the asymptotic covariance matrix of $\hat{\beta}_{A2}$ is given by $\sigma_{b_1}^2 = 1/W + U/W^2$, where

$$U = \sum_{\alpha,l,m} \rho(\alpha,l,m)\frac{1-\rho^V(\alpha,l,m)}{\rho^V(\alpha,l,m)}\alpha p_{lm}(1-p_{lm})\left(\frac{q_{(1-l)m}}{q_{0m}+q_{1m}}\right)^2.$$

Note that $P(Y_i = 1|A_i = 1, i \in \text{ category } (l,m))$ can be estimated by $y_{lm}/n_{lm}^v$and $\rho(l,m)$ by α_{lm}/n. Thus q_{lm} can be estimated by $(\alpha_{lm}/n)(y_{lm}/n_{lm}^v)$. Then we can estimate W by $\hat{W} = n^{-1}\sum_{m=1}^k \alpha_{0m}\alpha_{1m}y_{0m}y_{1m}/(\alpha_{0m}y_{0m}n_{1m}^v + \alpha_{1m}y_{1m}n_{0m}^v)$. By replacing $\rho_{l,m}^v$ with n_{lm}^v/α_{lm} and p_{lm} with y_{lm}/n_{lm}^v, we can estimate U by

$$\begin{aligned}\hat{U} &= n^{-1}\sum_{l,m}\alpha_{lm}\frac{\alpha_{lm}-n_{lm}^v}{n_{lm}^v}\frac{y_{lm}}{n_{lm}^v}\frac{n_{lm}^v-y_{lm}}{n_{lm}^v}\left(\alpha_{(1-l)m}\frac{y_{(1-l)m}}{n_{(1-l)m}^v}\right)\left(\alpha_{1m}\frac{y_{1m}}{n_{1m}^v}+\alpha_{0m}\frac{y_{0m}}{n_{0m}^v}\right)^{-2}\\ &= n^{-1}\sum_{m=1}^k\left[\frac{\alpha_{1m}-n_{1m}^v}{n_{1m}^v}\frac{n_{1m}^v-y_{1m}}{n_{1m}^v}\alpha_{0m}\frac{y_{0m}}{n_{0m}^v}+\alpha_{0m}-n_{0m}^v n_{0m}^v\frac{n_{0m}^v-y_{0m}}{n_{0m}^v}\alpha_{1m}\frac{y_{1m}}{n_{1m}^v}\right]\\ &\quad\times\left(\alpha_{1m}\frac{y_{1m}}{n_{1m}^v}\alpha_{0m}\frac{y_{0m}}{n_{0m}^v}\right)\left(\alpha_{1m}\frac{y_{1m}}{n_{1m}^v}+\alpha_{0m}\frac{y_{0m}}{n_{0m}^v}\right)^{-2}\\ &= n^{-1}\sum_{m=1}^k\left[\left(\frac{1}{y_{1m}}-\frac{1}{n_{1m}^v}+\frac{1}{\alpha_{1m}}-\frac{n_{1m}^v}{\alpha_{1m}y_{1m}}\right)+\left(\frac{1}{y_{0m}}-\frac{1}{n_{0m}^v}+\frac{1}{\alpha_{0m}}-\frac{n_{0m}^v}{\alpha_{0m}y_{0m}}\right)\right]\\ &\quad\times\left(\frac{\alpha_{1m}y_{1m}\alpha_{0m}y_{0m}}{\alpha_{1m}y_{1m}n_{0m}^v+\alpha_{0m}y_{0m}n_{1m}^v}\right)^2.\end{aligned}$$

From $\hat{W}$ and $\hat{U}$, we obtain an estimate of $\sigma_{b_1}^2$ as follows

$$\begin{aligned}&\hat{\sigma}_{b_1}^2\\ &= \frac{1}{\hat{W}^2}\sum_{m=1}^k\left[\left(\frac{1}{y_{1m}}-\frac{1}{n_{1m}^v}+\frac{1}{\alpha_{1m}}\right)+\left(\frac{1}{y_{0m}}-\frac{1}{n_{0m}^v}+\frac{1}{\alpha_{0m}}\right)\right]\left(\frac{\alpha_{1m}y_{1m}\alpha_{0m}y_{0m}}{\alpha_{1m}y_{1m}n_{0m}^v+\alpha_{0m}y_{0m}n_{1m}^v}\right)^2\\ &\quad+\frac{1}{\hat{W}}+\frac{1}{\hat{W}^2}\sum_{m=1}^k\left[-\frac{n_{1m}^v}{\alpha_{1m}y_{1m}}-\frac{n_{0m}^v}{\alpha_{0m}y_{0m}}\right]\left(\frac{\alpha_{1m}y_{1m}\alpha_{0m}y_{0m}}{\alpha_{1m}y_{1m}n_{0m}^v+\alpha_{0m}y_{0m}n_{1m}^v}\right)^2\\ &= \frac{1}{\hat{W}^2}\sum_{m=1}^k\left[\left(\frac{1}{y_{1m}}-\frac{1}{n_{1m}^v}+\frac{1}{\alpha_{1m}}+\frac{1}{y_{0m}}-\frac{1}{n_{0m}^v}+\frac{1}{\alpha_{0m}}\right)\right]\left(\frac{\alpha_{1m}y_{1m}\alpha_{0m}y_{0m}}{\alpha_{1m}y_{1m}n_{0m}^v+\alpha_{0m}y_{0m}n_{1m}^v}\right)^2.\end{aligned}$$

The variance of $\hat{b}_1$ is estimated by $\widehat{\mathrm{Var}}\left(\hat{b}_1\right) = n^{-1}\hat{\sigma}^2_{\hat{b}_1}$. □

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
YS proved Proposition 1, Theorem 1, Corollary 1 and 2, and wrote the manuscript. LQ developed the Poisson regression using the automated data with missing outcomes, performed the computations for the simulation study and for the application, and wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
The authors thank the two referees for their valuable comments. The authors also thank Yuriko Nagano for some help with the data analysis. This research was partially supported by the National Science Foundation grant DMS-1208978, the National Institute of Health NIAID grant R37 AI054165, and a fund provided by The University of North Carolina at Charlotte.

Received: 24 July 2014 Accepted: 25 November 2014
Published online: 16 December 2014

References

Carpenter, JR, Kenward, MG, Vansteelandt, S: A comparison of multiple imputation and doubly robust estimation for analyses with missing data. J. R. Stat. Soc. A. **169**, 571–584 (2006)
Clayton, D, Spiegelhalter, D, Dunn, G, Pickles, A: Analysis of longitudinal binary data from multiphase sampling (with discussion). J. R. Stat. Soc. B. **60**, 71–87 (1998)
Chu, H, Halloran, EM: Estimating vaccine efficacy using auxiliary outcome data and a small validation sample. Stat. Med. **23**, 2697–2711 (2004)
Dempster, AP, Laird, NM, Rubin, DB: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. **39**, 1–38 (1977)
Ellenberg, SS, Hamilton, JM: Surrogate endpoints in clinical trials: cancer. Stat. Med. **8**, 405–413 (1989)
Fleming, TR: Evaluating therapeutic interventions: some issues and experiences. Stat. Sci. **7**, 428–456 (1992)
Halloran, EM, Longini, IM, Gaglani, MJ, Piedra, PA, Chu, H, Herschler, GB, Glezed, WP: Estimating efficacy of trivalent, cold-adapted, influenza virus vaccine (CAIV-T) against influenza A (H1N1) and B Using Surveillance Cultures. Am. J. Epid. **158**, 305–311 (2003)
Horvitz, DG, Thompson, DJ: A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**, 663–685 (1952)
Little, RJA, Rubin, DB: Statistical analysis of missing data. Wiley, New York (2002)
Pepe, MS, Reilly, M, Fleming, TR: Auxiliary outcome data and the mean score method. J. Stat. Plan. Inference. **42**, 137–160 (1994)
Prentice, RL: Surrogate endpoints in clinical trials: definition and operational criteria. Stat. Med. **8**, 431–440 (1989)
Rubin, DB: Multiple imputation for nonresponse in surveys. Wiley, New York (1987)
Robins, JM, Rotnitzky, A, Zhao, LP: Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc. **89**, 846–866 (1994)
Scharfstein, DO, Rotnitzky, A, Robins, JM: Adjusting for nonignorable drop-out using semiparametric nonresponse models: rejoinder. J. Am. Stat. Assoc. **94**, 1135–1146 (1999)

doi:10.1186/s40488-014-0023-3
Cite this article as: Qi and Sun: **Missing data approaches for probability regression models with missing outcomes with applications.** *Journal of Statistical Distributions and Applications* 2014 **1**:23.