A general instrumental variable framework for regression analysis with outcome missing not at random

Eric J. Tchetgen Tchetgen[1,2] and Kathleen Wirth[2]

Departments of Biostatistics[1] and Epidemiology[2],

Harvard University

Correspondence: Eric J. Tchetgen Tchetgen, Departments of Biostatistics and Epidemiology, Harvard School of Public Health 677 Huntington Avenue, Boston, MA 02115.

Abstract

The instrumental variable (IV) design has a long-standing tradition as an approach for unbiased evaluation of the effect of an exposure in the presence of unobserved confounding. The IV approach is also well developed to account for covariate measurement error or misclassification in regression analysis. In this paper, the authors study the instrumental variable approach in the context of regression analysis with an outcome missing not at random, also known as nonignorable missing outcome. An IV for a missing outcome must satisfy the exclusion restriction that it is not independently related to the outcome in the population, and that the IV and the outcome are correlated in the observed sample only to the extent that both are associated with the missingness process. Therefore, a valid IV must predict a person's propensity to have an observed outcome, without directly influencing the outcome itself. Under an additional assumption that the magnitude of selection bias is independent of the IV, it is shown that the population regression in view is nonparametrically identified. For inference, we propose to fit in a complete-case analysis, the regression of interest, modified to include an additional covariate carefully constructed as a function of the IV to account for selection bias. The approach is developed for the identity, log and logit link functions, and a sensitivity analysis technique is also described which allows one to assess the extent to which a violation of the identifying assumption might affect inference. For illustration, the methods are used to account for selection bias induced by HIV testing refusal in the evaluation of HIV prevalence in the Zambian Demographic and Health Surveys.

The instrumental variable (IV) approach typically refers to a set of methods used to recover, under certain assumptions, an unbiased estimate of the causal effect of an exposure in the presence of unobserved confounding (Wright,1928, Golderberger, 1972, Robins, 1994, Angrist et al, 1996, Heckman, 1997). Instrumental variable methods are also available for regression analysis with mismeasured or misclassified covariates (Amemiya, 1985, Schennach, 2007, Carroll et al, 2006, Hu, 2008, Buonaccorsi, 2010). Another complication of regression analysis is that the outcome may be unobserved for a subset of the sample. In such settings, the missing data mechanism is said to be not at random, or nonignorable when it depends on the underlying value of the missing outcome upon adjusting for fully observed covariates (Little and Rubin, 2002).

To ground ideas, consider a study of sexual behavior in India using data from the MEASURE DHS (Demographic and Health Surveys) project which administered nationally-representative, household-based surveys on HIV knowledge, attitudes, and behavior. These surveys were conducted with face-to-face interviews (DHS, 2013). Although such interviews can be economically efficient and optimal in terms of validity in certain populations (eg, low literacy), they also have been shown to sometimes introduce bias to the measurement of behaviors perceived as socially undesirable (Turner et al, 1998, Rogers et al, 2005, Tideman et al, 2007). Suppose that one aims to characterize using such data, the association between a (male) participant's frequency of sexual encounter with a (female) sex worker, and various of the male's demographic and other behavioral outcomes. Due to the sensitive nature of such a query, it is not surprising that a number of participants had a missing value for frequency of sexual encounter with a sex worker. Bias due to item nonresponse in this setting may occur if the average response of males who completed the survey item differ systematically from the average response of those who did not complete the item, i.e. nonresponse is nonignorable. Therefore, a valid analysis of such data must account for the potential selection bias due to nonresponse.

Existing strategies which have previously been used to account for some degree of selection bias due to missing data, such as inverse probability weighting (Robins and Rotnitzky and Zhao,1995, van der Laan and Robins, 2003, Tsiatis, 2007), or outcome multiple imputation (Rubin, 1987, Little and Rubin, 2002), typically rely on an assumption that missingness can effectively be rendered independent of the outcome, upon conditioning on a sufficiently rich set of observed covariates. This is the assumption that is most often made in practice, which formally entails an assumption that the response is missing at random. This assumption is strictly untestable without imposing an additional assumption and may be questionable in applications such as the DHS example, primarily because systematic differences between respondents and nonrespondents to a query such as frequency of sexual contact with a sex worker will likely prevail despite any covariate adjustment. Therefore, the outcome is likely to be missing not at random.

Analytic strategies that have sometimes been used for outcome missing not at random, include methods that rely for identification on parametric assumptions (Diggle and Kenward, 2004, Wu and Carroll, 1988, Roy, 2003, Rotnitzky and Robins, 1997) and therefore may be sensitive to small deviations from the assumed model. Sensitivity analysis techniques have also been proposed (Robins et al, 1999), and in some simple cases, worst case scenarios of such analyses produce bounds for certain population parameters of interest.

In this paper, the authors follow an alternative strategy, and develop an IV approach for regression analysis when the outcome is missing not at random. A valid IV in this context must satisfy two conditions, which we formally define in the next section and summarize below,

(i) first, the IV must not be directly related to the outcome in the underlying population, conditional on covariates in the regression model,

(ii) second, the IV must be independently associated with the missingness mechanism conditional on the covariates in the regression model.

Therefore, a valid IV must predict a person's propensity to have an observed outcome, without directly influencing the outcome itself.

Similar to IV for causal effects, (i) and (ii) essentially amount to a form of exclusion restriction such that the IV and the outcome in view are correlated in the observed sample only to the extent that the missingness mechanism is potentially influenced by both. A valid IV for the missingness mechanism may not always be easy to find, however, as we show below, if a valid IV is successfully observed, i.e. a variable that satisfies (i) and (ii), such an IV may potentially be used to account for nonignorable missingness of the outcome in regression analysis.

Returning to the DHS study of sexual behavior in India, suppose that the interviewer's gender were recorded with each interview. Then, one might expect a female interviewer to experience a potentially different nonresponse rate than her male counterpart for queries related to male sexual behavior. If this were indeed the case, the interviewer's recorded gender would clearly satisfy condition (ii). Furthermore, as the interviewer's gender is unlikely to have directly influenced the participant's sexual behavior, condition (i) would also be satisfied. In this case, the interviewer's recorded gender constitutes a valid IV for nonresponse to queries about sexual behavior in the DHS India sample. Note that, other interviewer characteristics may likewise serve as a valid nonresponse IV, say for example interviewer's age, provided they satisfy conditions (i) and (ii).

The idea that data on auxiliary variables known to satisfy certain exclusion restrictions can potentially be used to adjust for nonrandom selection is not entirely new and is a familiar concept, particularly in the social sciences (Heckman, 1979, Dubin and Rivers, 1990, Winship and Mare, 1992). The notion that in a survey study, an interviewer's characteristics, or other operational features of the study could serve as an IV (as long as they satisfy conditions (i) and (ii)) for nonresponse to sensitive queries was recently used in a groundbreaking analysis by Bärnighausen et al (2011), to correct HIV prevalence estimates for survey nonparticipation. They demonstrate

quite convincingly that the interviewer's identity in survey studies, generally satisfies exclusion restrictions (i) and (ii) and therefore can be used when available in the observed sample, to account for selectivity in nonresponse. However, as in most IV settings, assumptions (i) and (ii) only, do not generally suffice for identification and an additional assumption is needed. The standard analytic framework in the social sciences was proposed by Heckman (1976, 1979), whereby identification is obtained under assumptions (i) and (ii), and additional parametric assumptions. Bärnighausen et al (2011) used a Heckman-type selection model that becomes identified under (i) and (ii), and parametric specifications that involve both linearity of the effects of covariates, and bivariate Gaussian latent error terms. However, it is well known that Heckman's selection model can be sensitive to these parametric assumptions (Arabmazar & Schmidt 1981, Winship and Mare, 1992, Puhani, 2000), although recent work has made significant strides towards relaxing (albeit partially) the parametric assumptions made by Heckman's original model (Manski, 1985, Stolzenberg and Relies, 1990, Powell, 1987, Newey et al, 1990, Cosslett,1991, Das et al 2003, Newey, 2009). Nonetheless, a general analytic framework for the IV model for missing data remains of keen interest in several disciplines, including economics, sociology and epidemiology.

In this paper, a straightforward identification strategy is proposed, which entails restricting the nature of selection bias due to nonresponse, but allows the observed data distribution to a priori remain unrestricted. To fix ideas, consider a regression with identity link function. Bias due to selective nonresponse can then be encoded as the difference in the average outcome comparing the subset of individuals with complete data to individuals with missing outcome as a function of covariates and the IV. Our identifying assumption for the additive scale states that,

(iii) for a fixed covariate value, the magnitude of selection bias does not vary on the additive scale, with changes in value of the IV.

Assumption (iii) states that selection bias on average, remains constant on the additive scale

across values of the IV. Thus, in the DHS example, assumption (iii) requires for a fixed covariate value, that differences in the average outcome for respondents versus nonrespondents, is unrelated to interviewer's gender, and therefore that the magnitude of selection bias due to nonresponse for fixed covariates is additively constant, for interviewers of different gender. Note that the assumption does not rule out differences in the average outcome for the subgroup of nonrespondents interviewed by a male versus those interviewed by a female, in fact the assumption is perfectly compatible with this average outcome for nonrespondents varying with covariates and the IV.

Under assumptions (i)-(iii), the authors establish that the regression function in view is nonparametrically identified. This means that the regression curve is identified under these assumptions regardless of its underlying functional form, whether parametric, semiparametric or nonparametric. We emphasize this fact, as an attractive feature of the proposed framework, because it essentially guarantees that one will in general be able to assess the goodness-of-fit of a model for the population regression curve, even if the outcome is missing not at random, provided assumptions (i)-(iii) hold. The identifying assumptions (i)-(iii) are formalized below.

Next, we give our main identification result for regression analysis with identity link. Focusing on parametric models, mainly to simplify the exposition, we then propose a strategy for estimation and inference based on a complete-case regression analysis, in which the regression model of interest is modified by introducing a special covariate, carefully constructed in terms of the IV to account for selection bias due to nonresponse. We compare the proposed approach to a nonparametric formulation of Heckman's selection model due to Das et al (2003), which allows us to key in on core differences in the underlying identifying assumptions made by each approach. Next, the proposed approach is shown to extend to regression analysis with log and logit link functions. For illustration, the methods are used to account for bias due to HIV testing refusal in the evaluation of HIV prevalence rates in the Zambian Demographic and Health Surveys. Finally, we present a

sensitivity analysis that may be used in practice, to assess the extent to which a violation of the key identifying assumption (iii) might impact inference.

1 Notation, Assumptions and Preliminary Result

Suppose we have observed n independent and identically distributed observations $(\mathbf{X}, RY, R)$ with $\mathbf{X}$ fully observed, R the indicator of whether the person's outcome Y is observed. Suppose that, one aims to estimate the population regression function $\mu(\mathbf{X}) = g\{E(Y|\mathbf{X} = x)\}$ encoding the relation between $\mathbf{X}$ and the corresponding mean of Y, with g the identity, log or logit link. Until otherwise stated, we will focus on the identity link typically used for a continuous outcome. Let $\widetilde{\pi}(\mathbf{X}, Y) = \Pr(R = 1|\mathbf{X}, Y)$ define the probability that Y is observed given $(\mathbf{X}, Y)$. Under missing at random, it is customary to assume that $\widetilde{\pi}(\mathbf{X}, Y)$ does not further depend on Y, so it can be dropped as an argument of $\widetilde{\pi}$, in which case, $\mu(\mathbf{X})$ is nonparametrically identified without an additional assumption. Here we do not make such an assumption, and we allow $\widetilde{\pi}(\mathbf{X}, Y)$ to depend on Y, such that the missingness process is nonignorable, and therefore, the regression function $\mu(\mathbf{X})$ is not identified from the observed data without an additional assumption.

The following result characterizes the bias due to nonignorable missingness, in terms of the following selection bias function $\widetilde{\delta}(\mathbf{X}) = E(Y|R = 1, \mathbf{X}) - E(Y|R = 0, \mathbf{X})$ which encodes on the mean difference scale, the extent to which the outcome mean differs in the subsample with observed outcome from that of the subsample with unobserved outcome. Thus, $\widetilde{\delta}(\mathbf{X}) = 0$ encodes the null hypothesis of no selection bias given $\mathbf{X}$. Then,

$$
\begin{aligned}
E(Y|\mathbf{X}) &= E(Y|\mathbf{X}, R=1)\Pr(R=1|\mathbf{X}) + E(Y|\mathbf{X}, R=0)\Pr(R=0|\mathbf{X}) \\
&= E(Y|\mathbf{X}, R=1) - E\left\{(Y|\mathbf{X}, R=1) - E(Y|\mathbf{X}, R=0)\right\}\Pr(R=0|\mathbf{X}) \\
&= E(Y|\mathbf{X}, R=1) - \widetilde{\delta}(\mathbf{X})\Pr(R=0|\mathbf{X})
\end{aligned}
$$

Thus, the bias between $E(Y|\mathbf{X})$ and the complete case regression $E(Y|\mathbf{X}, R=1)$ is

$$
E(Y|\mathbf{X}, R=1) - E(Y|\mathbf{X}) = \widetilde{\delta}(\mathbf{X})\Pr(R=0|\mathbf{X}) \tag{1}
$$

which vanishes if either $\widetilde{\delta}(\mathbf{X}) = 0$ or equivalently if $\widetilde{\pi}(\mathbf{X}, Y) = \widetilde{\pi}(\mathbf{X})$, i.e. if data is missing at random, or if $\Pr(R=1|\mathbf{X}) = 1$ and therefore there is no missing data.

In the presence of nonignorable nonresponse, neither of the above conditions will hold. Nonetheless, we can make progress, if in addition to $\mathbf{X}$, we also observe a valid instrumental variable Z known to satisfy assumptions (IV.1)-(IV.3) given below. Let $\pi(\mathbf{X}, Z) = \Pr(R=1|\mathbf{X}, Z)$ denote the propensity score for the missingness mechanism given $\mathbf{X}$ and Z. Our assumptions entail,

(IV.1) Exclusion restriction: $E(Y|\mathbf{X}, Z) = E(Y|\mathbf{X})$ almost surely,

(IV.2) Non-null relation between Z and R: $\pi(\mathbf{X}, z) - \pi(\mathbf{X}, z') \neq 0$, almost surely, for $z \neq z'$.

(IV.3) Homogeneous additive selection bias: $E(Y|R=1, \mathbf{X}, Z) - E(Y|R=0, \mathbf{X}, Z) = \delta(\mathbf{X})$ almost surely.

The exclusion restriction (IV.1) states that the IV and the outcome are conditionally independent on the mean scale, given $\mathbf{X}$ in the underlying population. This assumption is similar to the assumption of no direct effect of the IV on the outcome, typically made in the IV context of causal

effects. The second assumption (IV.2) requires that Z is independently associated with R. Note that in spite of (IV.2), assumption (IV.1) implies that Z cannot reduce the dependence between R and Y. Consequently, $\Pr(R = 1|Y, \mathbf{X}, Z)$ remains a function of y even after conditioning on Z and $\mathbf{X}$. The last assumption implies that the magnitude of selection bias measured on the additive scale does not depend on Z. Thus, for all practical purposes, it is as if the IV were randomized with respect to the degree of selection bias within levels of $\mathbf{X}$. To motivate assumption (IV.3), the following result describes a relatively large class of possible data generating mechanisms for which (IV.3) is shown to hold. To state the result, let $\epsilon = Y - E(Y|R = 0, \mathbf{X}, Z)$ with corresponding conditional moment generating function $t \longmapsto M(t; \mathbf{X}, Z) = E\left(e^{t\epsilon}|\mathbf{X}, Z, R = 0\right)$.

Result 1: Suppose that R follows the logistic regression model

$$\operatorname{logit} \Pr(R = 1|Y, \mathbf{X}, Z) = \alpha_Y(\mathbf{X})\, y + \alpha_{zx}(Z, \mathbf{X})$$

where α_y and α_{zx} are unrestricted, and therefore the model is solely restricted in that the association between R and Y on the log odds ratio scale is linear in Y and does not depend on Z, i.e. there is no interaction between Z and Y in the linear log odds ratio association of Y with R within levels of $\mathbf{X}$. Further assuming that

$$M(t; \mathbf{X}, Z) = M^*(t; \mathbf{X})$$

does not depend on Z, implies that assumption (IV.3) holds, with

$$\delta(\mathbf{X}) = E(Y|R = 1, \mathbf{X}, Z) - E(Y|R = 0, \mathbf{X}, Z) = \left.\frac{\partial \log M^*(t; \mathbf{X})}{\partial t}\right|_{t=\alpha_y(\mathbf{X})}$$

Proof: Following Tchetgen Tchetgen, Robins and Rotnitzky (2010) one can show that

$$
\begin{aligned}
E(Y|R=1,\mathbf{X},Z) &= \frac{E(Y\exp\left(\alpha_y\left(\mathbf{X}\right)Y\right)|R=0,\mathbf{X},Z)}{E(\exp\left(\alpha_y\left(\mathbf{X}\right)Y\right)\ |R=0,\mathbf{X},Z)} \\
&= \frac{\partial E(\exp\left(tY\right)|R=0,\mathbf{X},Z)/\partial t|_{t=\alpha_y(\mathbf{X})}}{E(\exp\left(\alpha_y\left(\mathbf{X}\right)Y\right)\ |R=0,\mathbf{X},Z)} \\
&= \partial\log E(\exp\left(tY\right)|R=0,\mathbf{X},Z)/\partial t|_{t=\alpha_y(\mathbf{X})} \\
&= \partial\log\left\{E\left(e^{t\epsilon}|\mathbf{X},Z,R=0\right)\exp\left(tE(Y|R=0,\mathbf{X},Z)\right)\right\}/\partial t|_{t=\alpha_y(\mathbf{X})} \\
&= E(Y|R=0,\mathbf{X},Z)+\left.\frac{\partial\log M^*(t;\mathbf{X})}{\partial t}\right|_{t=\alpha_y(\mathbf{X})}
\end{aligned}
$$

proving the result.□

For any $\alpha_y \neq 0$, the model described in Result 1 allows the outcome to be missing not at random, although the dependence on Y of the selection model for $\Pr\left(R=1|\mathbf{X},Z,Y\right)$ is assumed to be linear on the log odds ratio scale, and independent of Z within levels of $\mathbf{X}$. The missingness process is otherwise quite general, since α_y and α_{zx} are not restricted to follow a particular parametric functional form. Crucially, the above model cannot be refuted empirically, without a priori restricting α_{zx} (Robins et al, 1999). Note also that the distributional assumption for the outcome is quite weak, and essentially amounts to an assumption of homoscedastic error with respect to the IV in the subsample missing the outcome. This assumption is thus also not empirically refutable without additional assumptions. It is also worth noting that the class of models described in Result 1 is contained but does not span the model defined by assumptions (IV.1)-(IV.3), thus indicating that our assumptions may be satisfied for a broad range of settings in which the methods derived below would be useful. A simple and familiar choice for the density of ϵ that readily satisfies the conditions of Result 1 is $\epsilon|\mathbf{X},Z,R=0 \sim N(0,\sigma^2\left(\mathbf{X}\right))$, which gives $\delta\left(\mathbf{X}\right)=\sigma^2\left(\mathbf{X}\right)\alpha_y\left(\mathbf{X}\right)$. We should also note that the condition for the result can be relaxed somewhat, in that both α_y and

M^* can depend on Z, provided that $\partial \log M^*(t;\mathbf{X},Z)/\partial t|_{t=\alpha_y(\mathbf{X},Z)}$ does not, in which case, the result continues to hold, although such a data generating mechanism may not be easily construed.

2 Inference with identity link function

We are now ready to state our first identification result.

Result 2: Under assumptions (IV.1)-(IV.3), the regression function $\mu(\mathbf{X})$ is nonparametrically identified from the observed data $(\mathbf{X}, RY, R, Z)$, and the complete-case regression curve $m(\mathbf{X}, Z) = E(Y|Z, \mathbf{X}, R = 1)$ can be expressed explicitely as a function of $\mu(\mathbf{X})$, $\delta(\mathbf{X})$ and $\pi(\mathbf{X}, Z)$:

$$m(\mathbf{X}, Z) = \delta(\mathbf{X})\{1 - \pi(\mathbf{X}, Z)\} + \mu(\mathbf{X}) \tag{2}$$

Result 2 states that the regression curve $\mu(\mathbf{X})$ is identified in the presence of nonignorable nonresponse of the outcome, provided that Z satisfies conditions (IV.1)-(IV.3) of a valid IV. The identification result is nonparametric in the sense that assumptions (IV.1)-(IV.3) do not impose any restriction on the functional form of $\mu(\mathbf{X})$, $\delta(\mathbf{X})$ and $\pi(\mathbf{X}, Z)$. This in turn implies that no restriction is placed on $m(\mathbf{X}, Z)$, and thus that the model is just-identified without restricting the observed data likelihood.

Result 2 also gives an explicit parametrization of the complete-case regression function $m(\mathbf{X}, Z)$ in terms of the selection bias function, the missingness propensity score and the underlying regression curve of interest. It is natural to use this parametrization to make inferences about $\mu(\mathbf{X})$. To fix ideas, suppose that we aim to estimate the linear model, $\mu(\mathbf{X};\beta) = (1, \mathbf{X}')\beta$ and we likewise posit the following models for the selection bias function, $\delta(\mathbf{X};\eta) = (1, \mathbf{X})\eta$, and for the propensity score, logit $\pi(\mathbf{X}, Z;\alpha) = (1, \mathbf{X}', Z)\alpha$. Assuming that the residual $\varepsilon(\theta) = Y - m(\mathbf{X}, Z;\theta)$ is

normally distributed with variance σ^2, where $m(\mathbf{X}, Z; \theta) = \delta(\mathbf{X}; \eta)(1 - \pi(\mathbf{X}, Z; \alpha)) + \mu(\mathbf{X}; \beta)$ $\theta = (\beta, \alpha, \eta, \sigma^2)$. The maximum likelihood estimator $\widehat{\theta} = \left(\widehat{\beta}, \widehat{\alpha}, \widehat{\eta}, \widehat{\sigma}^2\right)$ solves

$$\arg\max_{\theta} \sum_i L(O_i; \theta) \tag{3}$$

$$\text{with } L(O_i; \theta) = R_i \log f_1\left(\varepsilon_i(\theta) | \mathbf{X}_i, Z_i; \sigma^2\right) + \log f_2\left(R_i | Z_i, \mathbf{X}_i; \alpha\right),$$

f_1 the normal density with mean zero and variance σ^2, and f_2 the Bernoulli density with mean $\pi(\mathbf{X}, Z; \alpha)$.

The variance-covariance matrix of $\widehat{\theta}$ is given by the inverse observed information matrix:

$$\left\{ -\sum_i \left. \frac{\partial^2 L(O_i; \theta)}{\partial \theta \partial \theta'} \right|_{\widehat{\theta}} \right\}^{-1}.$$

Furthermore inference based on the Wald, score or likelihood ratio statistics may be obtained under standard maximum likelihood theory.

It is straightforward to verify that the above approach is not sensitive to a violation of the normality assumption, and that the score equation under the normal model remains unbiased even if the assumption does not hold, provided the mean model, the selection bias model and the propensity score model are all correct. However, when normality does not hold, the variance-covariance matrix of $\widehat{\theta}$ can no longer be estimated using the expression in the previous display, but instead may be estimated using the standard sandwich formula:

$$\left\{ \sum_i \left. \frac{\partial^2 L(O_i; \theta)}{\partial \theta \partial \theta'} \right|_{\widehat{\theta}} \right\}^{-1} \left\{ \sum_i \left. \frac{\partial L(O_i; \theta)}{\partial \theta} \right|_{\widehat{\theta}}^{\otimes 2} \right\} \left\{ \sum_i \left. \frac{\partial^2 L(O_i; \theta)}{\partial \theta \partial \theta'} \right|_{\widehat{\theta}} \right\}^{-1},$$

where $A^{\otimes 2} = AA'$ for any matrix A.

An alternative, potentially less efficient estimation strategy follows a two stage approach, whereby in a first stage one computes $\widehat{\alpha}_2$ by maximizing the log partial likelihood function $\sum_i \log f_2 (R_i | Z_i, \mathbf{X}_i; \alpha)$, followed by a second stage, in which one uses $\pi (\mathbf{X}, Z; \widehat{\alpha})$ to estimate $m(\mathbf{X}, Z; \theta)$ via complete case ordinary least square regression of Y on $(1, \mathbf{X}', (1, \mathbf{X}) (1 - \pi (\mathbf{X}, Z; \widehat{\alpha})))$. For inference under the two stage approach, we recommend the nonparametric bootstrap.

A potential advantage of the two-stage approach is that it may more easily be performed using standard statistical software for regression analysis, provided that the corresponding software accommodates a user specified offset in the regression model.

3 Comparison to a Nonparametric Heckman Selection Model

Heckman's selection model is perhaps the most common strategy used in economics and other social sciences to address selection bias in regression analysis (Heckman, 1979). We adopt a nonparametric formulation of the model due to Das et al (2003) to ease a comparison to the proposed approach. This formulation assumes that the selection or missingness mechanism is generated under the latent variable threshold model:

(D.IV.1) $R = 1 \{U < \zeta (\mathbf{X}, Z)\}$ where U is a latent random variable.

(D.IV.2) The model further supposes that, $Y = \mu (\mathbf{X}) + \varepsilon$ where ε is a separable residual error with the joint density of (ε, U) assumed to be independent of $(\mathbf{X}, Z)$ but otherwise unrestricted.

Then, assuming that the CDF of U, $G_u(\cdot)$ is one-to-one, for $V = G_u(U)$, Das et al (2003) establish that

$$\begin{aligned} E(\varepsilon|\mathbf{X}, Z, R=1) &= E(\varepsilon|\mathbf{X}, Z, U < \zeta(\mathbf{X}, Z)) \\ &= E(\varepsilon|\mathbf{X}, Z, V < G_u(\zeta(\mathbf{X}, Z))) \\ &= \lambda(\pi(\mathbf{X}, Z)) \end{aligned}$$

where $\pi(\mathbf{X}, Z) = \Pr(R=1|\mathbf{X}, Z) = \Pr(U < \zeta(\mathbf{X}, Z)|\mathbf{X}, Z) = G_u(\zeta(\mathbf{X}, Z))$.

Assuming that (ε, U) are joint Gaussian with $Var(U) = 1$, gives $\lambda(\pi) = \sigma_{\varepsilon U}\ \phi(\Phi^{-1}(\pi))/\pi$ where $\sigma_{\varepsilon U} = Cov(\varepsilon, U)$, $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal CDF and $\phi(\cdot)$ is the standard normal density. Assumptions (D.IV.1), (D.IV.2), the Gaussian assumption together with a linear specification for $\zeta(\mathbf{X}, Z)$ and $\mu(\mathbf{X})$ yield Heckman's (1979) standard correction for selection, which is completely identified from the observed data. However, in the larger nonparametric model defined by assumptions (D.IV.1) and (D.IV.2), so that $\zeta(\mathbf{X}, Z)$ and $\mu(\mathbf{X})$ remain unrestricted, Das et al (2003) established that $\mu(\mathbf{X})$ becomes nonparametrically identified up to an additive constant, provided assumption (D.IV.3) below also holds.

(D.IV.3) $\mu(\mathbf{X})$, $\lambda(\pi)$, and $\pi(\mathbf{X}, Z)$ are continuously differentiable with continuous distribution functions almost everywhere and with probability one,

$$\partial(\pi(\mathbf{X}, Z))/\partial Z \neq 0.$$

Similar to (IV.2) assumption (D.IV.3) states that Z must be independently predictive of R, although the latter is restricted to a continuous IV. Note that for the complete-case sample, under

the nonparametric model given by assumptions (D.IV.1)-(D.IV.2), one can write

$$\begin{aligned} E\left(Y|R=1,\mathbf{X},Z\right) &= \mu\left(\mathbf{X}\right)+\lambda\left(\pi\left(\mathbf{X},Z\right)\right) \\ &= \mu\left(\mathbf{X}\right)+\delta^{*}\left(\mathbf{X},Z\right)\left\{1-\pi\left(\mathbf{X},Z\right)\right\} \end{aligned}$$

which using equation (2), implies that the model restricts selection bias to be of the following form:

$$\delta^{*}\left(\mathbf{X},Z\right)=\frac{\lambda\left(\pi\left(\mathbf{X},Z\right)\right)}{1-\pi\left(\mathbf{X},Z\right)}$$

Thus, we have learned that the nonparametric version of Heckman's selection model allows dependence of the selection bias function on both $\mathbf{X}$ and Z, but restricts such dependence to operate only through an unrestricted function of the propensity score. In contrast, in this paper, we have allowed under Assumption (IV.3), the selection bias function to be an unrestricted function of $\mathbf{X}$, however restricting it to not further depend on Z. Assumptions (IV.1)-(IV.3) give nonparametric identification of the function $\mu\left(\mathbf{X}\right)$, while assumptions (D.IV.1)-(D.IV.3) can only identify $\mu\left(\mathbf{X}\right)+C$ for an unknown constant C. This means that the intercept of the function $\mu\left(\mathbf{X}\right)$ is not identified under the latter conditions, while it is under the former. The intercept may itself be of interest, in settings such as in the previous DHS example where the outcome level for each value of $\mathbf{X}$ is of primary scientific interest. The intercept will also be key to recover a valid estimate of the average outcome $E(Y)=E\left[\mu\left(\mathbf{X}\right)\right]$. Interestingly, Newey (2009) also notes that, together with (D.IV.1)-(D.IV.3), further restricting $\mu\left(\mathbf{X}\right)$ to be a linear function of $\mathbf{X}$, and assuming that $\pi\left(\mathbf{X},Z\right)$ is a single index model still does not suffice to identify the intercept of $\mu\left(\mathbf{X}\right)$ and thus to identify $E(Y)$. This further clarifies that identification of the intercept in the original Heckman model is principally derived from the joint Gaussian assumption of (ε,U), a parametric assumption which together with

linearity assumptions, imposes strong restrictions on the observed data distribution, and thus, it should be of no surprise that, as reported in the literature, inferences about the intercept in this framework can be quite sensitive to the underlying identifying assumptions (Arabmazar & Schmidt 1981, Winship and Mare, 1992, Puhani, 2000).

4 Inference with the log link

In this section, we consider regression analysis with the log link function, and define the model of interest as $\mu(\mathbf{X}) = \log E(Y|\mathbf{X})$. We may proceed as with the identity link and first derive the multiplicative selection bias for the observed complete-case regression $E(Y|\mathbf{X}, R = 1)$,

$$\begin{aligned}\frac{E(Y|\mathbf{X}, R=1)}{E(Y|\mathbf{X})} &= \frac{E(Y|\mathbf{X}, R=1)}{E(Y|\mathbf{X}, R=0)} \Big/ \left\{ \sum_{r=1} \frac{E(Y|\mathbf{X}, R=r)}{E(Y|\mathbf{X}, R=0)} \Pr(R=r|\mathbf{X}) \right\}^{-1} \\ &= \widetilde{\nu}(\mathbf{X}) \left\{ \nu(\mathbf{X}) \Pr(R=1|\mathbf{X}) + \Pr(R=0|\mathbf{X}) \right\}^{-1}\end{aligned}$$

where $\widetilde{\nu}(\mathbf{X}) = E(Y|\mathbf{X}, R = 1)/E(Y|\mathbf{X}, R = 0)$ encodes the degree of association between Y and R given $\mathbf{X}$ on the mean ratio scale, and quantifies the amount of selection bias. Naturally, as before, $E(Y|\mathbf{X}, R = 1) = E(Y|\mathbf{X})$ if and only if $\widetilde{\nu}(\mathbf{X}) = 1$ or $\Pr(R = 1|\mathbf{X}) = 1$, that is if and only if there is no selection bias or no missing data. We say that Z is a valid IV for a log regression analysis with nonignorable missing outcome, if Z satisfies assumptions (IV.1) and (IV.2) and the following additional assumption,

(IV.3') Homogeneous multiplicative selection bias : $E(Y|R = 1, \mathbf{X}, Z)/E(Y|R = 0, \mathbf{X}, Z) = \nu(\mathbf{X})$ does not depend on Z.

Similar to assumption (IV.3), the new assumption (IV.3') states that the IV essentially behaves as if it were randomized relative to selection bias on the multiplicative scale conditional on $\mathbf{X}$. Note that both our current and previous definition of a valid IV are scale specific. Thus, a valid IV on the additive scale that satisfies assumption (IV.3) cannot in general simultaneously satisfy assumption (IV.3') and therefore cannot in general be a valid IV on the multiplicative scale, and vice-versa. Our identification result for the multiplicative scale is given next.

Result 3: Under assumptions (IV.1)-(IV.3'), the regression function $\mu(\mathbf{X})$ *is nonparametrically identified from the observed data* $(\mathbf{X}, RY, R, Z)$, *and the complete-case regression curve* $m(\mathbf{X}, Z) = E(Y|Z, \mathbf{X}, R = 1)$ *can be expressed as a function of* $\mu(\mathbf{X}), \nu(\mathbf{X})$ *and* $\pi(\mathbf{X}, Z)$ *as followed:*

$$\log m(\mathbf{X}, Z) = \log \nu(\mathbf{X}) - \overline{\nu}(\mathbf{X}, Z) + \mu(\mathbf{X}) \tag{4}$$

$$\text{where } \overline{\nu}(\mathbf{X}, Z) = \log\{\nu(\mathbf{X})\pi(\mathbf{X}, Z) + 1 - \pi(\mathbf{X}, Z)\} \tag{5}$$

Result 3 states that the regression curve $E(Y|\mathbf{X}) = \exp\{\mu(\mathbf{X})\}$ is identified from data $(RY, Z, \mathbf{X}, R)$ provided that Z is an IV satisfying assumptions (IV.1)-(IV.3'). Equation (4) gives an explicit representation of the complete-case regression $E(Y|Z, \mathbf{X}, R = 1)$ as a function of the regression of interest $\mu(\mathbf{X})$, the selection bias function $\nu(\mathbf{X})$ and the propensity score $\pi(\mathbf{X}, Z)$. Crucially, we note that $\overline{\nu}(\mathbf{X}, Z)$ in equation (4) is not a free parameter, but corresponds to a carefully crafted offset fully determined by the selection bias function and the missingness mechanism as displayed in equation (5).

Equation (4) suggests a simple strategy for estimating $\mu(\mathbf{X})$ in practice. To illustrate, suppose that Y is a count, and interest lies in the familiar log-linear model $\mu(\mathbf{X}, \psi) = (1, \mathbf{X}')\psi$. Further suppose that one specifies a similar log-linear model to encode selection bias $\log \nu(\mathbf{X}; \eta) = (1, \mathbf{X}')\eta$. Then, assuming that Y follows a Poisson distribution with mean computed using formula (4) under

the above model,

$$m(\mathbf{X}, Z; \eta, \alpha, \psi) = \exp((1, \mathbf{X}')\eta - \overline{\nu}\,(\mathbf{X}, Z; \eta, \alpha) + (1, \mathbf{X}')\psi)$$

$$\text{where } \overline{\nu}\,(\mathbf{X}, Z; \alpha, \eta) = \log\left\{\exp\left[(1, \mathbf{X}')\eta\right]\pi\,(\mathbf{X}, Z; \alpha) + 1 - \pi\,(\mathbf{X}, Z; \alpha)\right\}$$

The maximum likelihood estimator of $\phi = (\psi, \alpha, \eta)$ maximizes equation (3) upon replacing f_1 with the Poisson density with mean given in the previous display. Maximum likelihood inference then proceeds as previously described. A two-stage estimation strategy similar to the one proposed for the identity link can likewise be used for the log link and is easily inferred from the presentation.

5 Inference with the logit link

In this section, we consider regression analysis for a binary outcome using a logit link function, and we define the model of interest as followed,

$$\begin{aligned} \mu\,(\mathbf{X}) &= \text{logit}\,\Pr(Y = 1|\mathbf{X}) \\ &= \log ODDS\ (\mathbf{X}) = \log\frac{\Pr(Y = 1|\mathbf{X})}{\Pr(Y = 0|\mathbf{X})} \end{aligned} \tag{6}$$

Likewise, let

$$ODDS\ (\mathbf{X}, R = 1) = \frac{\Pr(Y = 1|\mathbf{X}, R = 1)}{\Pr(Y = 0|\mathbf{X}, R = 1)}.$$

We begin by deriving the odds ratio selection bias on the odds ratio scale, for the complete-case odds ODDS $(\mathbf{X}, R = 1)$, obtained from data $(RY, \mathbf{X}, R)$,

$$\frac{ODDS\ (\mathbf{X}, R=1)}{ODDS\ (\mathbf{X})} = \frac{ODDS\ (\mathbf{X}, R=1)}{ODDS\ (\mathbf{X}, R=0)} / \left\{ \sum_{r=1} \frac{ODDS\ (\mathbf{X}, R=r)}{ODDS\ (\mathbf{X}, R=0)} \Pr(R=r|\mathbf{X}, Y=0) \right\}$$

$$= \widetilde{\omega}\,(\mathbf{X}) \left\{ \widetilde{\omega}\,(\mathbf{X}) \Pr(R=1|\mathbf{X}, Y=0) + \Pr(R=0|\mathbf{X}, Y=0) \right\}^{-1}$$

where $\widetilde{\omega}\,(\mathbf{X}) = ODDS\ (\mathbf{X}, R=1)/ODDS\ (\mathbf{X}, R=0)$ and, where we have used the following key collapsibility property of the odds function (See Tchetgen Tchetgen, 2013),

$$ODDS\ (\mathbf{X}) = E\left\{ODDS\ (\mathbf{X}, R)|\mathbf{X}, Y=0\right\},$$

the function $\widetilde{\omega}\,(\mathbf{X})$ encodes the degree of association between Y and R given $\mathbf{X}$ on the odds ratio scale, and quantifies selection bias. Naturally, $\Pr(Y=1|\mathbf{X}, R=1) = \Pr(Y=1|\mathbf{X})$ if and only if $\widetilde{\omega}\,(\mathbf{X})$ or $\Pr(R=1|\mathbf{X}, Y=0) = 1$, that is if and only if there is no selection bias or no missing data. We say that Z is a valid IV for a logistic regression analysis with nonignorable missing outcome, if Z satisfies assumption (IV.1) and (IV.2) and the following additional assumption,

(IV.3†) Homogeneous odds ratio selection bias:

$$\log ODDS\ (\mathbf{X}, R=1, Z)/ODDS\ (\mathbf{X}, R=0, Z) = \omega\,(\mathbf{X})$$

does not depend on Z.

Similar to assumption (IV.3), the new assumption (IV.3†) states that the IV essentially behaves as if it were randomized relative to selection bias on the odds ratio scale conditional on $\mathbf{X}$. Our identification result for the odds ratio scale is given next.

Result 4: Under assumptions (IV.1)-(IV.3†), the regression function $\mu(\mathbf{X})$ is nonparametrically identified from the observed data $(\mathbf{X}, RY, R, Z)$, and the observed regression curve

$$\text{logit}n(\mathbf{X}, Z) = \text{logit}\Pr(Y = 1|\mathbf{X}, R = 1, Z)$$

can be expressed as an function of $\mu(\mathbf{X}), \nu(\mathbf{X})$ and $\pi(\mathbf{X}, Z)$ as followed:

$$\text{logit}n(\mathbf{X}, Z) = \text{logit}t(\mathbf{X}) + \omega(\mathbf{X}) - \overline{\omega}(\mathbf{X}, Z)$$

$$\text{where logit}t(\mathbf{X}) = \mu(\mathbf{X})$$

$$\overline{\omega}(\mathbf{X}, Z) = \log\{\exp(\omega(\mathbf{X}))\lambda(\mathbf{X}, Z) + 1 - \lambda(\mathbf{X}, Z)\}$$

and $\lambda(\mathbf{X}, Z) = \Pr(R = 1|\mathbf{X}, Z, Y = 0)$ satisfies

$$\pi(\mathbf{X}, Z) = \{1 - t(\mathbf{X})\}\lambda(\mathbf{X}, Z) + t(\mathbf{X})\left[1 + (1 - \lambda(\mathbf{X}, Z))\exp\{-\omega(\mathbf{X})\}/\lambda(\mathbf{X}, Z)\right]^{-1} \tag{7}$$

Result 4 states that the regression curve $t(\mathbf{X}) = \Pr(Y = 1|\mathbf{X}) = \text{expit}\{\mu(\mathbf{X})\}$ is identified from data $(RY, Z, \mathbf{X}, R)$ provided that Z is an IV satisfying assumptions (IV.1)-(IV.3†). The result gives an explicit representation on the logit scale, of the observed regression $\Pr(Y = 1|Z, \mathbf{X}, R = 1)$ as a function of the regression of interest $\mu(\mathbf{X})$, the selection bias function $\omega(\mathbf{X})$ and $\lambda(\mathbf{X}, Z)$. Note that although $\lambda(\mathbf{X}, Z) = \Pr(R = 1|\mathbf{X}, Z, Y = 0)$ is not directly observed, it is readily obtained under our identifying assumptions by the law of total probability (7).

For inference, one may use a maximum likelihood approach, which entails maximizing the log-likelihood

$$\sum_i R_i \left\{\log \Pr\left(Y_i = 1 | R_i = 1, \mathbf{X}_i, Z_i\right) + \log \pi(\mathbf{X}_i, Z_i)\right\} + (1 - R_i) \log(1 - \pi(\mathbf{X}_i, Z_i)) \tag{8}$$

using the parametrization of Result 4. For instance, suppose that one aims to estimate the logistic regression model

$$\text{logit} t(\mathbf{X}) = \mu\left(\mathbf{X}; \psi\right) = (1, \mathbf{X}')\psi \tag{9}$$

Further suppose that one specifies a similar linear log odds ratio model to encode selection bias

$$\omega\left(\mathbf{X}; \eta\right) = (1, \mathbf{X}')\eta, \tag{10}$$

and assuming that

$$\text{logit} \lambda\left(\mathbf{X}, Z; \alpha\right) = (1, \mathbf{X}')\alpha \tag{11}$$

produces the following complete-case model,

$$\begin{aligned}
\text{logit} \Pr\left(Y = 1 | R = 1, \mathbf{X}, Z; \psi, \alpha, \eta\right) &= (1, \mathbf{X}')\psi + (1, \mathbf{X}')\eta \\
&\quad - \log\left(\lambda\left(\mathbf{X}, Z; \alpha\right) \exp\left\{(1, \mathbf{X}')\eta\right\} + 1 - \lambda\left(\mathbf{X}, Z; \alpha\right)\right) \\
\pi(\mathbf{X}, Z; \psi, \alpha, \eta) &= \left\{1 - t(\mathbf{X}; \psi)\right\} \lambda\left(\mathbf{X}, Z; \alpha\right) \\
&\quad + t(\mathbf{X}; \psi)\left[1 + \left(1 - \lambda\left(\mathbf{X}, Z; \alpha\right)\right) \exp\left\{-(1, \mathbf{X}')\eta\right\} / \lambda\left(\mathbf{X}, Z; \alpha\right)\right]^{-1}.
\end{aligned}$$

The maximum likelihood estimator of (ψ, α, η) maximizes the loglikelihood (8) under the working model in the above display. Inference then proceeds using standard maximum likelihood theory.

6 Empirical Illustration

To illustrate the proposed instrumental variable methods, we obtained data from the 2007 Zambia Demographic and Health Survey to estimate HIV prevalence among adult men adjusting for non-ignorable, selective non-participation in the survey's HIV testing component. Further details regarding the sampling and data collection procedures of the Zambia DHS are available elsewhere (CSO, 2009). Briefly, this cross-sectional, population-based survey, carried out over a 6-month period from April to October 2007, employed a complex sampling scheme to assess the general health status and family welfare among households in Zambia. At the initial household visit, a representative from the household completed a short household interview which collected information on access to drinking water, toilet and cooking facilities, and household assets. The representative was also asked to list and provide basic demographic information on all usual household members and any visitors who stayed in the household the previous night. Of those listed, men aged 15-59 years and women aged 15-49 years were eligible for participation in an individual interview and HIV testing. In total, 7,146 eligible men were identified from 7,164 household interviews; 7,116 (>99%) men had complete information from the household interview. Of those with complete information, 5,145 (72%) provided a specimen for HIV testing. We note that the 1,971 (28%) eligible men without an HIV test result comprise both individuals who either could not be contacted or were contacted and refused to participate in all components of the survey including HIV testing (N=654) and those who were contacted and agreed to participate in the individual interview, but refused to be tested for HIV (N=1,317).

6.1 Instrumental variables

To select the candidate instruments, we adapted the approach used in the previously described analysis by Bärnighausen and colleagues (2011), who employed a Heckman-type selection model to correct HIV prevalence estimates for testing non-participation in the 2007 Zambia DHS survey. Specifically, we used household interviewer identity and an indicator variable for whether or not a household was visited on the first day of data collection within a cluster. As described earlier, interviewer characteristics such as gender, personality, and interpersonal skills may lead to different response rates. Similarly, the chances of encountering and enrolling eligible individuals are higher for those households reached early in data collection because there are more opportunities for repeat visits by data collectors. Given that both the specific interviewer deployed to a household and the timing of that visit are determined at random (or by a known algorithm), these factors are unlikely to directly influence an individual's HIV status. In the 2007 Zambia DHS survey, 54 distinct interviewers conducted 50 or more household interviews with men and 1,831 (36% of 5,130) households were reached on the first day of data collection within a cluster. Both of these factors were highly associated with HIV testing non-participation ($P<0.001$).

6.2 Propensity Score and Selection Bias Models

For estimation, we used the logistic regression (9) to model the population prevalence of HIV (Y), conditional on observed covariates $\mathbf{X}$ containing age, education, wealth quintile, and location type of household. Table 1 summarizes the model and indicates most factors are strongly predictive of HIV seropositivity in this population. We likewise used the logistic regression (11) to model the probability of participation in the survey HIV testing component (R) as a function of covariates $\mathbf{X}$ and the IVs Z consisting of household interviewer identity and visit on the first day of data

collection within a cluster. Table 2 provides a complete description of this model, and summarizes evidence of strong correlation between interviewer identity and participation rate, i.e. that assumption (I.V.2) holds in this sample. Finally, we modeled the selection bias function using equation (10) .Table 3 provides a complete description of this last model and suggests significant selection bias in the odds ratio association between education and household location type, and HIV prevalence, further justifying the need to adjust point estimates of HIV prevalence.

We computed the estimate $\widehat{p}$ of the marginal HIV prevalence $p = \Pr(Y = 1)$ for Zambia as a weighted average, of individual fitted values $\widehat{\Pr}(Y = 1|\mathbf{X}_i) = \widehat{t}(\mathbf{X}_i)$, with survey weights W_i, i.e. $\widehat{p} = \sum_i W_i \widehat{t}(\mathbf{X}) / \sum_i W_i$.

All statistical analyses were conducted using PROC NLMIXED and PROC IML within SAS software version 9.3 (SAS Institute, Cary, NC). We used standard Taylor-series expansion arguments to derive the following large-sample variance estimator for the resulting point estimate $\widehat{p}$ of HIV prevalence, which simultaneously acknowledges the uncertainty due to first stage estimation of $t(\mathbf{X})$, and the presence of sampling weights, $\widehat{\mathrm{Var}}(\widehat{p}) = \widehat{\Lambda} + \widehat{\Gamma}\widehat{\Omega}\widehat{\Gamma}'$, where

$$\widehat{\Lambda} = n^{-2} \sum_i W_i \left\{ \widehat{t}(\mathbf{X}_i) - \widehat{p} \right\}^2$$

$$\widehat{\Gamma} = n^{-1} \sum_j W_j \left(1, \mathbf{X}_j'\right) \widehat{t}(\mathbf{X}_j) \left(1 - \widehat{t}(\mathbf{X}_j)\right)$$

and $\widehat{\Omega} = \widehat{\mathrm{Var}}\left(\widehat{\psi}\right)$ was obtained from the inverse information matrix of the mle of (ψ, α, η) for the loglikelihood derived in the previous section. Note that the survey weights were only used in the second stage, because conditioning on the covariates $\mathbf{X}$, gave virtually the same results for the first stage whether the weights were included or not (see next paragraph). Finally, we used the above estimated standard errors to construct Wald-type 95% confidence intervals (CIs) for p.

6.3 Results

We observed an unadjusted (crude) estimate of HIV seropositive prevalence of 12.2% (95% CI: 11.2% to 13.1%), which was significantly lower than the IV-adjusted HIV prevalence estimate of 21.1% (95% CI: 16.2% to 25.9%) obtained using the proposed IV approach. As noted in the previous paragraph, applying the survey weights at both stages gave similar results with an estimated HIV prevalence of 19.5% (95%CI: 15.9% to 23.1%). It is also noteworthy that, the IV-adjusted point estimate obtained using our methods essentially agreed with the corrected-estimate of Bärnighausen et al (2011), obtained via a Heckman-type selection model for a binary outcome, and reported to be 21% (95% CI: 20% to 22%). This suggests that, at least in this specific empirical example, the IV results appear to be fairly robust to the assumptions underlying either adjustment strategy, and that the adjustment for selection bias with an IV appears to matter more than the specific IV analytic strategy used. However, one may note that the 95%CI of Bärnighausen et al (2011) is considerably narrower than the one obtained with our approach. The observed difference between these CIs may be primarily due to the fact that, while our 95%CI accurately reflects all sources of uncertainty including from the first stage estimation of $t(\mathbf{X})$, the 95%CI of Barnighausen et al (2011) apparently did not appropriately account for the uncertainty due to the analogous preliminary estimation of $\Pr(Y = 1|\mathbf{X})$ obtained with Heckman's model, and therefore the reported 95%CI is likely to have understated the actual uncertainty around Heckman's estimator.

7 Detecting the presence of selection bias

Interestingly, if Z is known to satisfy assumptions (IV.1) and (IV.2) but neither assumption (IV.3), (IV.3') nor (IV.3†), such a variable cannot generally be used to correct for selection bias in the presence of nonignorable nonresponse for the outcome on any of these three scales. However, as

we argue next, such a variable may still be useful as a tool for detecting the presence of selection bias. This is in fact the case since assumptions (IV.3) ((IV.3') and (IV.3†)) are trivially satisfied under the null hypothesis of no selection bias, i.e. if $H_0 : \delta(\mathbf{X}) = 0$ ($H_0^* = \log \nu(\mathbf{X}) = 0$ and $H_0^{\dagger} = \omega(\mathbf{X}) = 0$) for all $\mathbf{X}$ respectively. Therefore a test statistic of H_0 (H_0^* and $H_0^{\dagger}$) based on either the Wald, score or likelihood ratio tests using the likelihood framework previously described, constitutes under assumptions (IV.1) and (IV.2), a valid test statistic of the null hypothesis that selection bias is absent on a given scale. Furthermore, such a test will generally be consistent under the alternative hypothesis that selection bias due to missing data is present on a given scale, regardless of whether assumption (IV.3) ((IV.3') and (IV.3†) holds.

8 Sensitivity to heterogeneous selection bias

Assumption (IV.3) and likewise assumptions (IV.3') and (IV.3†), are not empirically testable and may only be approximately correct in a given application. For this reason, it is crucial in practice, to supplement the proposed IV approach with a sensitivity analysis to assess the degree to which a violation of the assumption might influence inference. Focusing on the identity link, the sensitivity analysis approach entails defining a new function $\delta(\mathbf{X}, Z)$ to replace $\delta(\mathbf{X})$, which allows the latter to depend explicitly on Z, thus effectively allowing for a violation of assumption (IV.3). Accordingly, let $\delta(\mathbf{X}, Z) = \delta^*(\mathbf{X}, Z) + \delta_0(\mathbf{X})$ so that

$$\delta^*(\mathbf{X}, Z) = E(Y|R=1, \mathbf{X}, Z) - E(Y|R=0, \mathbf{X}, Z) - E(Y|R=1, \mathbf{X}, Z=0) + E(Y|R=0, \mathbf{X}, Z=0)$$

encodes the degree to which selection bias varies with Z within levels of $\mathbf{X}$, and

$$\delta_0(\mathbf{X}) = E(Y|R=1, \mathbf{X}, Z=0) - E(Y|R=0, \mathbf{X}, Z=0)$$

encodes the magnitude of selection bias for a baseline value $Z = 0$ within levels of $\mathbf{X}$. The function $\delta^*(\mathbf{X}, Z)$ is clearly not identified without an additional assumption, therefore we propose to proceed by obtaining inferences for fixed $\delta^*(\mathbf{X}, Z)$ upon substituting $\delta(\mathbf{X}, Z)$ for $\delta(\mathbf{X})$ in the likelihood model, with $\delta_0(\mathbf{X})$ estimated from the data under a parametric model. A sensitivity analysis is then obtained by varying δ^* producing inferences under various forms of violation of assumption (IV.3).

A similar approach can be used to assess the degree of sensitivity of inference when using a log or logit link, to a potential violation of assumptions (IV.3') and (IV.3†) respectively, which is easily inferred from the exposition.

9 Final remarks

In this paper, we have considered the somewhat pernicious problem of selection bias in regression analysis, due to an outcome, missing not at random. We have shown that this seemingly intractable problem can be made more tractable with the aid of an instrumental variable for non-ignorable missing data. Simple, yet novel identification assumptions are obtained for this IV framework, which yield a simple strategy for estimation, appropriately accounting for the presence of selection bias. The approach was then illustrated in a data set from Zambia, to obtain an adjusted estimate of HIV national prevalence, accounting for selection bias due to testing refusal. A sensitivity analysis was also proposed to assess the extent to which a violation of a key identifying assumption

could bias the results, and the methods were developed for the identity, log and logit link.

Several interesting extensions could be explored in the future, including analogous methods for longitudinal data, as well as for dependent censoring of a survival outcome. It may also be of interest to extend the approach to a regression framework with covariate missing not at random.

APPENDIX

Proof of Result 2: The proof relies on the following decomposition:

$$
\begin{aligned}
& E\left(Y|Z,\mathbf{X},R\right) \\
& = E\left(Y|Z,\mathbf{X},R\right) - E\left(Y|Z,\mathbf{X},R=0\right) \\
& - \sum_{r=0}^{1}\left\{E\left(Y|Z,\mathbf{X},r\right) - E\left(Y|Z,\mathbf{X},R=0\right)\right\}\Pr\left(R=r|\mathbf{X},Z\right) \\
& + E\left(Y|Z,\mathbf{X}\right)
\end{aligned}
$$

Thus, under assumptions (IV.1)-(IV.3), we obtain for $R=1$

$$m(\mathbf{X},Z) = \delta\left(\mathbf{X}\right) - \delta\left(\mathbf{X}\right)\pi\left(\mathbf{X},Z\right) + \mu\left(\mathbf{X}\right)$$

Next, since $\pi\left(\mathbf{X},Z\right)$ *is identified from the partial likelihood of* R *given* $(\mathbf{X},Z)$*, we may take it as known. Then, we obtain the identification result by noting that for all* $\pi\left(\mathbf{X},Z\right)$ *that satisfy (IV.2),* $m^{*}(\mathbf{X},Z) = m(\mathbf{X},Z)$ if and only if $\delta^{*}\left(\mathbf{X}\right) = \delta\left(\mathbf{X}\right)$ and $\mu^{*}\left(\mathbf{X}\right) = \mu\left(\mathbf{X}\right)$, *where* $m^{*}(\mathbf{X},Z) = \delta^{*}\left(\mathbf{X}\right) - \delta^{*}\left(\mathbf{X}\right)\pi\left(\mathbf{X},Z\right) + \mu^{*}\left(\mathbf{X}\right)$. □

Proof of Result 3: Note that the regression function $E\left(Y|Z,\mathbf{X},R\right)$ *can be decomposed nonparametrically as followed:*

$$
\begin{aligned}
&E\left(Y|Z,\mathbf{X},R=1\right)\\
&=\frac{E\left(Y|Z,\mathbf{X},R=1\right)}{E\left(Y|Z,\mathbf{X},R=0\right)}\times\left\{\sum_{r}\frac{E\left(Y|Z,\mathbf{X},R=r\right)}{E\left(Y|Z,\mathbf{X},R=0\right)}\Pr(R=r|Z,\mathbf{X})\right\}^{-1}\\
&\times E\left(Y|Z,\mathbf{X}\right)
\end{aligned}
$$

Then under our assumptions, we have that

$$
\log E\left(Y|Z,\mathbf{X},R=1\right)=\log\nu\left(\mathbf{X}\right)-\overline{\nu}\left(\mathbf{X},Z\right)+\mu\left(\mathbf{X}\right)
$$

Finally, we obtain the identification result upon noting that $\Pr(R=1|Z,\mathbf{X})$ is nonparametrically identified from the partial likeihood for the missingness mechanism, and thus $\log E\left(Y|Z,\mathbf{X},R=1\right)=\log E^{*}\left(Y|Z,\mathbf{X},R=1\right)$ if and only if

$$
\begin{aligned}
\nu^{*}\left(\mathbf{X}\right)&=\nu\left(\mathbf{X}\right)\\
\mu\left(\mathbf{X}\right)&=\mu^{*}\left(\mathbf{X}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
\log E^{*}\left(Y|Z,\mathbf{X},R=1\right)&=\log\nu^{*}\left(\mathbf{X}\right)-\overline{\nu}^{*}\left(\mathbf{X},Z\right)+\mu^{*}\left(\mathbf{X}\right)\\
\overline{\nu}^{*}\left(\mathbf{X},Z\right)&=\log\left\{\left[\exp\left\{\nu^{*}\left(\mathbf{X}\right)\right\}\right]\pi\left(\mathbf{X},Z\right)+1-\pi\left(\mathbf{X},Z\right)\right\}
\end{aligned}
$$

Under assumptions (IV.1)-(IV.3'), the regression function $\mu\left(\mathbf{X}\right)$ is nonparametrically identified from the observed data $\left(\mathbf{X},RY,R,Z\right)$, and the observed regression curve $m(\mathbf{X},Z)=E\left(Y|Z,\mathbf{X},R=1\right)$

can be expressed as an function of $\mu(\mathbf{X}), \nu(\mathbf{X})$ *and* $\pi(\mathbf{X}, Z)$ *as followed:*

$$\log m(\mathbf{X}, Z) = \log \nu(\mathbf{X}) - \overline{\nu}(\mathbf{X}, Z) + \mu(\mathbf{X})$$

$$\text{where } \overline{\nu}(\mathbf{X}, Z) = \log\left\{\left[\exp\left\{\nu(\mathbf{X})\right\}\right]\pi(\mathbf{X}, Z) + 1 - \pi(\mathbf{X}, Z)\right\}$$

<u>*Proof of Result 4:*</u> *Note that the odds function*

$$ODDS\ (Z, \mathbf{X}, R = 1) = P(Y = 1|Z, \mathbf{X}, R = 1) / \Pr(Y = 0|Z, \mathbf{X}, R = 1)$$

can be decomposed nonparametrically as followed:

$$\begin{aligned}
&ODDS\ (Z, \mathbf{X}, R = 1) \\
&= \frac{ODDS\ (Z, \mathbf{X}, R = 1)}{ODDS\ (Z, \mathbf{X}, R = 0)} \times \left\{\sum_r \frac{ODDS\ (Z, \mathbf{X}, R = r)}{ODDS\ (Z, \mathbf{X}, R = 0)} \Pr(R = r|Z, \mathbf{X}, Y = 0)\right\}^{-1} \\
&\times ODDS\ (Z, \mathbf{X})
\end{aligned}$$

Then under our assumptions, we have that

$$\log ODDS\ (Z, \mathbf{X}, R = 1) = \omega(\mathbf{X}) - \overline{\omega}(\mathbf{X}, Z) + \mu(\mathbf{X})$$

Finally, we obtain the identification result upon noting that

$$
\begin{aligned}
&\Pr(R=1|Z,\mathbf{X}) \\
&\quad = \Pr(R=1|Z,\mathbf{X},Y=0)\Pr(Y=0|\mathbf{X},Z) + \Pr(R=1|Z,\mathbf{X},Y=1)\Pr(Y=1|\mathbf{X},Z) \\
&\quad = \Pr(R=1|Z,\mathbf{X},Y=0)\Pr(Y=0|\mathbf{X}) + \Pr(R=1|Z,\mathbf{X},Y=1)\Pr(Y=1|\mathbf{X}) \\
&\quad = \{1-t(\mathbf{X})\}\,\lambda(\mathbf{X},Z) + t(\mathbf{X})\lambda(\mathbf{X},Z)\exp\{\omega(\mathbf{X})\} \,/\, [\lambda(\mathbf{X},Z)\exp\{\omega(\mathbf{X})\} + \{1-\lambda(\mathbf{X},Z)\}]
\end{aligned}
$$

which implies that $\Pr(R=1|Z,\mathbf{X},Y=0)$ is identified from the observed data likelihood of R given $(\mathbf{X},Z)$.

References

[1] Amemiya Y., 1985. Instrumental variable estimator for the nonlinear errors-in-variables model, Journal of Econometrics, Elsevier, vol. 28(3), pages 273-289.

[2] Arabmazar A. and Schmidt P. (1981), Further evidence on the robustness of the Tobit estimator to heteroscedasticity. J. of Econometrics 17: 253-258

[3] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), Identification of Causal Effects Using Instrumental Variables, Journal of the American Statistical Association, 91, 444–472.

[4] Barnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Correcting HIV prevalence estimates for survey nonparticipation: using Heckman-type selection models. Epidemiology. 2011;22:27–35.

[5] Bollinger R., (1996). Bounding mean regressions when a binary regressor is mismeasured, Journal of Econometrics, Elsevier, vol. 73(2), pages 387-399.

[6] Buonaccorsi (2010). Measurement Error and Misclassification : Models, Methods and Applications. Chapman and Hall, C R C Press.

[7] Carroll R, Ruppert D, Stefanski L, Crainiceanu C (2006). Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition. Chapman and Hall, C R C Press.

[8] Cosslett, S. R. 1991. Semiparametric estimation of a regression model with sampling selectivity. In Nonparametric and Semiparametric Methods in Econometrics and Statistics, ed. W. A. Barnett, J. Powell, G. Tauchen, pp. 175-98. Cambridge: Cambridge Univ. Press.

[9] Central Statistical Office (CSO), Ministry of Health (MOH), Tropical Diseases Research Centre (TDRC), University of Zambia (UNZA), Macro International Inc. Zambia Demographic and Health Survey 2007. Calverton, MD: CSO, Macro International Inc.; 2009.

[10] Das M, Newey W K, Vella F (2003). Nonparametric estimation of sample selection models. Review of Economic Studies 70, 33-58.

[11] Diggle, P. D., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. Journal of the Royal Statistical Society: Series C. Applied Statistics, 43, 49–93

[12] Dubin JA, Rivers D. Selection bias in linear regression, logit and probit models. In: Fox J, Long JS, eds. Modern Methods of Data Analysis. Newbury Park, CA: Sage Publications; 1990:410–443.

[13] Newey W & James, L., 2003. Instrumental Variable Estimation of Nonparametric Models, Econometrica, Econometric Society, vol. 71(5), pages 1565-1578.

[14] van der Laan MJ, Robins JM. (2003). Unified Methods for Censored Longitudinal Data and Causality. Springer Verlag: New York.

[15] Measure DHS. Demographic and Health Surveys: HIV Corner. Available at: http://demo.measuredhs.com/measuredhs2/topics/hiv/start.cfm.Accessed. June 2013.

[16] HeckmanJ, . J. (1979). Samples election bias as a specification error. Econometrica 47:153-61.

[17] Heckman, J. J. (1997), Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations, Journal of Human Resources, 32, 441–462.

[18] Hu Y. (2008) Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. Journal of Econometrics. 144, 1, 27-61.

[19] Little, R. J. A., and Rubin, D. B. (2002), Statistical Analysis With Missing Data (2nd ed.), New York: Wiley.

[20] Newey, W. K., Powell, J. L., Walker, J. R. (1990). Semiparametric estimation of selection models: some empirical results. Am. Econ. Rev. 80:324-28.

[21] Manksi, C. F. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. J. Economet. 27:313-33

[22] Puhani PA.(2000) The Heckman correction for sample selection and its critique. J Econ Surv;14:53– 68.

[23] Powell, J. L. (1987). Semiparametric Estimation of Bivariate Latent Variable Models. Working Paper No. 8704. Madison, Wisc: Soc. Systems Res. Inst., Univ. Wisc.

[24] Rotnitzky A, Robins JM. (1997). Analysis of semiparametric regression models with non-ignorable non-response. Statistics in Medicine, 16:81-102.

[25] Robins, J (1994), Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models,Communications in Statistics, 23, 2379–2412.

[26] Robins JM, Rotnitzky A, Scharfstein D. (1999). Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In: Statistical Models in Epidemiology: The Environment and Clinical Trials. Halloran, M.E. and Berry, D., eds. IMA Volume 116, NY: Springer-Verlag, pp. 1-92.

[27] Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. Biometrics, 59, 829–836.

[28] Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons.

[29] Schennach S, (2007). Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models, Econometrica, Econometric Society, vol. 75(1), pages 201-239, 01.

[30] Stolzenberg, R. M., Relies, D. A. (1990). Theory testing in a world of constrained research design: the significance of Heckman's censored sampling bias correction for nonexperimental research. Sociol. Meth. Res. 18:395-415.

[31] Tchetgen Tchetgen E, Robins JM, Rotnitsky A. (2010). On doubly robust estimation in a semiparametric odds ratio model. Biometrika 97(1):171-180.

[32] Tchetgen Tchetgen, Eric J. (2013), A General Regression Framework for a Secondary Outcome in Case-control Studies. Biostatistics. In Press.

[33] Tideman RL, Chen MY, Pitts MK, Ginige S, Slaney M, Fairley CK (2007). A randomised controlled trial comparing computer-assisted with face-to-face sexual history taking in a clinical setting. Sex Transm Infect ;83:52–56.

[34] Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL.(1998) Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. Science;280:867– 873.

[35] Winship C. and Mare R. Models for sample selection bias (1992). Annu Rev Sociol;18:327–350.

[36] Wu, M. C., & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics, 44, 175–188.

Table 1. Log odds and 95% confidence intervals (CI) for HIV seropositivity among 7,116 adult men in Zambia, 2007.

	HIV seropositivity	
	Log odds	(95% CI)
Intercept	-3.369	(-4.824, -1.914)
Age (years)		
55 to 59	1.276	(0.249, 2.303)
50 to 54	1.988	(0.972, 3.004)
45 to 49	2.090	(1.079, 3.102)
40 to 44	1.954	(0.990, 2.918)
35 to 39	2.263	(1.297, 3.229)
30 to 34	2.023	(1.085, 2.960)
25 to 29	1.493	(0.610, 2.376)
20 to 24	0.941	(0.137, 1.744)
15 to 19	(ref)	--
Educational attainment (years)	0.042	(0.008, 0.077)
Wealth quintile		--
5th (wealthiest)	0.858	(0.191, 1.524)
4th	0.989	(0.363, 1.614)
3rd	0.494	(-0.107, 1.095)
2nd	0.394	(-0.214, 1.002)
1st (poorest)	(ref)	--
Location type of household		
Countryside	-0.801	(-1.218, -0.385)
Town	-0.305	(-0.638, 0.0288)
Small city	0.200	(-0.237, 0.637)
Capital, large city	(ref)	--

Table 2. Log odds and 95% confidence intervals (CI) for HIV testing participation among 7,116 adult men in Zambia, 2007.

	HIV testing participation	
	Log odds	(95% CI)
Intercept	-0.075	(-0.811, 0.660)
Age (years)		
55 to 59	0.485	(-0.089, 1.058)
50 to 54	1.015	(0.050, 1.980)
45 to 49	0.711	(0.015, 1.407)
40 to 44	0.345	(-0.116, 0.805)
35 to 39	0.748	(0.058, 1.439)
30 to 34	0.602	(0.069, 1.134)
25 to 29	0.201	(-0.171, 0.573)
20 to 24	0.178	(-0.164, 0.520)
15 to 19	(ref)	--
Educational attainment (years)	0.08845	(0.050, 0.127)
Wealth quintile		
5th (wealthiest)	0.037	(-0.503, 0.576)
4th	0.144	(-0.322, 0.610)
3rd	-0.241	(-0.601, 0.118)
2nd	-0.286	(-0.646, 0.075)
1st (poorest)	(ref)	--
Location type of household		
Countryside	0.554	(-0.082, 1.189)
Town	0.504	(-0.039, 1.046)
Small city	0.982	(0.056, 1.907)
Capital, large city	(ref)	--
Household visited on first day of data collection within a cluster (yes/no)	0.093	(-0.036, 0.222)
Interviewer identity		
106	-0.173	(-0.576, 0.229)
107	-0.327	(-0.678, 0.024)
108	0.069	(-0.318, 0.456)
215	-0.672	(-1.422, 0.079)
216	-0.169	(-0.630, 0.292)
217	-1.141	(-1.765, -0.517)
218	-1.170	(-1.697, -0.642)

223	-1.198	(-1.951, -0.444)
224	-0.332	(-0.934, 0.269)
225	-0.910	(-1.473, -0.348)
226	0.086	(-0.476, 0.648)
227	-0.290	(-0.833, 0.253)
228	-0.621	(-1.219, -0.022)
303	0.325	(-0.010, 0.750)
304	0.087	(-0.287, 0.460)
305	1.006	(0.512, 1.500)
306	0.480	(-0.032, 0.992)
307	0.097	(-0.424, 0.619)
403	0.442	(0.039, 0.845)
404	-0.084	(-0.442, 0.275)
405	-0.052	(-0.686, 0.582)
407	0.062	(-0.581, 0.706)
516	0.937	(0.372, 1.501)
517	-0.070	(-0.538, 0.399)
518	1.165	(0.623, 1.707)
523	0.580	(-0.113, 1.273)
524	1.354	(0.453, 2.255)
525	0.705	(0.027, 1.384)
526	0.426	(-0.053, 0.906)
527	0.256	(-0.269, 0.782)
528	-0.693	(-1.401, 0.016)
613	0.648	(-0.011, 1.306)
614	1.044	(0.348, 1.741)
616	0.661	(-0.081, 1.403)
617	0.809	(-0.030, 1.648)
623	-0.185	(-0.688, 0.317)
624	0.038	(-0.640, 0.715)
625	-0.561	(-1.095, -0.026)
626	-0.513	(-1.171, 0.144)
627	-0.993	(-1.562, -0.424)
628	-0.466	(-1.082, 0.150)
806	0.584	(0.196, 0.972)

807	-0.407	(-0.714, -0.099)
808	-0.041	(-0.394, 0.312)
903	0.788	(0.384, 1.193)
904	0.379	(-0.031, 0.789)
905	0.453	(0.002, 0.905)
1000	(ref)	--

Table 3. Linear log odds ratios (OR) and 95% confidence intervals (CI) for selection bias in HIV seropositivity due to nonignorable missingness among 7,116 adult men in Zambia, 2007.

	Selection bias	
	Log OR	(95% CI)
Intercept	-0.300	(-3.542, 2.942)
Age (years)		
55 to 59	0.056	(-1.946, 2.059)
50 to 54	-1.113	(-3.166, 0.941)
45 to 49	-0.219	(-2.198, 1.760)
40 to 44	0.837	(-0.933, 2.607)
35 to 39	-0.414	(-2.228, 1.400)
30 to 34	-0.342	(-2.050, 1.367)
25 to 29	-0.281	(-1.898, 1.338)
20 to 24	-0.850	(-2.392, 0.693)
15 to 19	(ref)	--
Educational attainment (years)	-0.102	(-0.178, -0.026)
Wealth quintile		--
5th (wealthiest)	-0.686	(-2.529, 1.156)
4th	-0.579	(-2.320, 1.162)
3rd	0.166	(-1.543, 1.875)
2nd	0.142	(-1.545, 1.830)
1st (poorest)	(ref)	--
Location type of household		
Countryside	0.025	(-1.021, -1.021)
Town	-0.326	(-1.148, 0.496)
Small city	-0.923	(-2.112, 0.267)
Capital, large city	(ref)	--