HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2018 December 01.

Published in final edited form as:

Biometrics. 2017 December ; 73(4): 1123–1131. doi:10.1111/biom.12670.

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

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

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health

[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health

Abstract

The instrumental variable (IV) design is a well-known approach for unbiased evaluation of causal effects in the presence of unobserved confounding. In this paper, we study the IV approach to account for selection bias in regression analysis with outcome missing not at random. In such a setting, a valid IV is a variable which (i) predicts the nonresponse process, and (ii) is independent of the outcome in the underlying population. We show that under the additional assumption (iii) that the IV is independent of the magnitude of selection bias due to nonresponse, the population regression in view is nonparametrically identified. For point estimation under (i)–(iii), we propose a simple complete-case analysis which modifies the regression of primary interest by carefully incorporating the IV to account for selection bias. The approach is developed for the identity, log and logit link functions. For inferences about the marginal mean of a binary outcome assuming (i) and (ii) only, we describe novel and approximately sharp bounds which unlike Robins-Manski bounds, are smooth in model parameters, therefore allowing for a straightforward approach to account for uncertainty due to sampling variability. These bounds provide a more honest account of uncertainty and allows one to assess the extent to which a violation of the key identifying condition (iii) might affect inferences. For illustration, the methods are used to account for selection bias induced by HIV testing non-participation in the evaluation of HIV prevalence in the Zambian Demographic and Health Surveys.

Keywords

Instrumental Variable; Nonignorable Missing Data; Selection Bias; Complete-case Analysis

1 Introduction

A common complication in regression analysis is that the outcome may not be observed for a subset of the sample. In such settings, if missingness remains associated with the outcome even after adjusting for fully observed covariates, the missing data mechanism is said to be not at random and the regression of interest may not be identified without an additional

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

Supplementary Materials

Web Appendices referenced in Sections 1–4 are available with this paper at the Biometrics website on Wiley Online Library.

assumption (Little and Rubin, 2002). Consider a study of risky sexual behavior among men in India using data collected from the National Family Health Survey (NFHS); a nationally-representative household-based survey of HIV-related knowledge, attitudes, and behavior. Suppose that one aimed to estimate the prevalence of high-risk sexual behavior such as contact with a commercial sex worker. Due to the sensitive nature of this type of query and the face-to-face nature of the interview, it may not be surprising that participants often fail to respond (DHS, 2013). Bias due to item non-response may be present if the average response of males who completed the survey's item differs systematically from the average outcome of non-responders. A valid analysis of such data must account for potential selection bias due to non-response.

Existing strategies to address outcome data missing not at random in regression analysis rely for identification purposes on rather strong parametric assumptions (Diggle and Kenward, 2004, Wu and Carroll, 1988, Roy, 2003, Rotnitzky and Robins, 1997) and may be sensitive to small deviations from these assumptions. Sensitivity analysis techniques have also been proposed (Robins et al, 1999), and in simple settings, worst case scenarios of such analyses produce bounds for certain population parameters.

In this paper, we follow an alternative analytic strategy, and develop an instrumental variable (IV) approach for regression analysis when the outcome is missing not at random. A valid IV in this context must satisfy two conditions summarized below and more formally defined in the next section: (i) first, the IV must be conditionally independent of the outcome in the underlying population, given fully observed covariates, (ii) second, the IV must be associated with the nonresponse mechanism conditional on observed covariates.

Therefore, a valid IV must predict a person's propensity to have complete outcome data, without directly influencing the outcome itself conditional on covariates. Similar to IV assumptions in causal inference, assumptions (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 they are both associated with non-response. A valid IV for non-response may not always be easy to find, however, as we show below, if such a variable can be found, it may potentially be used to correct for selection bias. The use of variables satisfying the exclusion restriction to adjust for non-random selection or non-response is a familiar concept, particularly in econometrics and other social sciences (Heckman, 1979, Dubin and Rivers, 1990, Winship and Mare, 1992, Manski, 2003, Nicoletti, 2010). The notion that interviewer characteristics or other operational features of a study (e.g. interview mode, length and design of the questionnaire) could serve as credible IVs for non-response has gained prominence (Lepkowski and Couper, 2002, Manski, 2003, Schrapler, 2004, Nicoletti and Peracchi, 2005, Bärnighausen et al, 2011). Interviewer and other operational study characteristics are invariably strong correlates of non-response (condition (ii)), will often also satisfy the exclusion restriction (condition (i)) and are relatively straightforward to collect. Therefore, suitable IV options will often be more readily available for missing data than for endogenous treatments. However, meeting assumptions (i) and (ii) alone, do not generally suffice for identification; an additional assumption is needed. The standard analytic framework in the social sciences was proposed by Heckman (1976, 1979) who achieves identification under assumptions (i) and (ii), and additional parametric

assumptions. Yet, 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).

In this paper, a straightforward identification strategy is proposed, which involves restricting the amount of heterogeneity in the magnitude of selection bias due to non-response, but allows the full data distribution to a priori remain unrestricted. For instance, suppose one is interested in a population regression model defined with the identity link function. Then, seletion bias due to non-response can be encoded as the difference in the average outcome comparing individuals with complete data to individuals with incomplete data conditional on covariates and the IV. Our identifying assumption on the additive scale is stated in terms of this selection bias: (iii) the magnitude of selection bias does not vary on the additive scale, with changes in the value of the IV conditional on covariates. Assumption (iii) states that selection bias on the additive scale is independent of the IV given covariates. Thus, in the NFHS example, assumption (iii) requires that conditional on covariates, the difference in the prevalence of risky sexual behavior (the outcome of interest) for respondents compared to non-respondents, is independent of interviewer's gender. Note that the assumption does not rule out differences in the prevalence of risky sexual behavior for non-respondents interviewed by a male vs a female. In the next sections, we show that under assumptions (i)–(iii), the population conditional mean of the outcome given covariates is non-parametrically identified, allowing for either identity, log (see Supplemental Materials) or logit link function. Furthermore, we propose a simple approach for estimation and inference based on a complete-case regression analysis, in which the population regression model of interest is modified by introducing a special covariate, carefully constructed using the IV to account for selection bias due to non-response. For illustration, the methods are used to account for bias due to HIV testing non-participation in the evaluation of HIV prevalence among men in Zambia. For inferences about the population mean of a binary outcome assuming (i) and (ii) only, we describe new approximately sharp bounds which are closely related to sharp bounds proposed by Robins (1989) and independently by Manski (1990). An important feature of our bounds is that unlike the Robins-Manski bounds, they are smooth in model parameters, thus permitting a straightforward account of uncertainty. The approach delivers asymptotically valid confidence intervals for the population outcome mean. Although, as shown below, the point estimate obtained under (i)–(iii) may be useful in finite samples, to anchor the bounds and to ensure that they remain coherent in the sense that the estimated lower bound does not exceed the estimated upper bound for the mean.

Although the paper focuses primarily on inference in the context of parametric models both for practical convenience and to more clearly communicate the main ideas, out results do not rely on parametric restrictions; the extension to semi-parametric and non-parametric inference can be entertained, curse of dimensionality permitting, without presenting new challenges for identification. Additional results are relegated to the Supplementary Materials where we compare the proposed approach for point identification to a non-parametric formulation of Heckman's selection model according to Das et al (2003). We argue that certain parametric distributional assumptions not needed in our approach are essential for identification in Heckman's selection model.

2 Notation, Assumptions and Preliminary Result

Suppose that we have observed n independent and identically distributed data ($\mathbf{X}$, RY, R) with $\mathbf{X}$ fully observed and R the indicator of whether the person's outcome Y is observed. Suppose that, one aims to estimate the population regression function $E(Y|\mathbf{X} = x) = g^{-1}\{\mu(\mathbf{X})\}$ encoding the relationship between $\mathbf{X}$ and the corresponding mean of Y, with g the identity, log or logit link function: Until otherwise stated, we will focus on the identity link typically used for a continuous outcome: Let $\tilde{\pi}(\mathbf{X}, Y)=Pr(R=1|\mathbf{X}, Y)$ define the probability that Y is observed given ($\mathbf{X}$, Y): Under missing at random, it is assumed that $\tilde{\pi}(\mathbf{X}, Y)$ does not further depend on Y, so it can be dropped as an argument of $\tilde{\pi}$, in which case, $\mu(\mathbf{X})$ is nonparametrically identified without an additional assumption. Here we do not make such an assumption, and we allow $\tilde{\pi}(\mathbf{X}, Y)$ 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 selection bias which we define as $\tilde{\delta}(\mathbf{X})=E(Y|R=1, \mathbf{X}) - E(Y|R=0, \mathbf{X})$. The function $\tilde{\delta}(\mathbf{X})$ encodes on the mean difference scale, the extent to which the outcome mean differs between complete and incomplete cases. Thus, $\tilde{\delta}(\mathbf{X})=0$ corresponds to no selection bias given $\mathbf{X}$. We note that

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

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

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

which vanishes if either $\tilde{\delta}(\mathbf{X})=0$ or equivalently if $\tilde{\pi}(\mathbf{X}, Y)=\tilde{\pi}(\mathbf{X})$, i.e. if data are missing at random, or if $\Pr(R = 1|\mathbf{X}) = 1$ and therefore there are 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, in which case the observed data is $\mathbf{O} = (\mathbf{X}, RY, R, Z)$. As an IV, Z must 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) IV relevance: $\pi(\mathbf{X}, z) - \pi(\mathbf{X}, z') \neq 0$, 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 mean independent in the underlying population given **X**. This assumption is similar to the assumption of no direct effect of the IV on the outcome, typically made in causal inference. The second assumption (IV.2) requires that Z is associated with R conditional on **X**. In spite of (IV.2), assumption (IV.1) implies that Z cannot completely eliminate the dependence between R and the mean of Y, since to do so would require that Z be an intermediate variable between R and Y, which contradicts assumption (IV.1). Consequently, $\Pr(R = 1 | Y, \mathbf{X}, Z)$ remains a function of Y even after conditioning on Z and **X**. The last assumption (IV. 3) implies that the magnitude of selection bias measured on the additive scale does not depend on Z. For all practical purposes, it is as if the IV were randomized with respect to the degree of selection bias within levels of **X**. To motivate assumption (IV.3), in the Supplementary Materials, we describe a fairly large semiparametric shared parameter model for which (IV.3) can be shown to hold.

3 Inference with identity link function

We are now ready to state our first identification result. All proofs are relegated to the Supplementary Materials.

Result 1: Under assumptions (IV.1)–(IV.3), the regression function $\mu(\mathbf{X})$ is nonparametrically identified from **O**. Furthermore, the complete-case regression curve $m(\mathbf{X}, Z) = E(Y|Z, \mathbf{X}, R = 1)$ can be expressed explicitly as the following function of $\mu(\mathbf{X})$, $\delta(\mathbf{X})$ and $\pi(\mathbf{X}, Z)$:

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

Result 1 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). 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 1 also gives an explicit parametrization of the complete-case regression $m(\mathbf{X}, Z)$ 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 ground ideas, suppose that one aims to estimate the linear model, $\mu(\mathbf{X}; \beta) = (1, \mathbf{X}')\beta$. Suppose also that one posits the following models for the selection bias function, $\delta(\mathbf{X}; \eta) = (1, \mathbf{X})\eta$, and for the propensity score, logit $\pi(\mathbf{X}, Z; \alpha) = (1, \mathbf{X}', Z)\alpha$. Other model specifications equally apply with no real additional difficulty. 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 corresponding maximum likelihood estimator $\hat{\theta} = \left(\hat{\beta}, \hat{\alpha}, \hat{\eta}, \hat{\sigma}^2\right)$ solves

$$arg\max_{\theta}\sum_{i} R_i log\, f_1\left(\varepsilon_i(\theta)\,|\mathbf{X}_i, Z_i;\sigma^2\right) + log\, f_2\left(R_i|Z_i, \mathbf{X}_i;\alpha\right), \quad (3)$$

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 $\hat{\theta}$ is given by the standard inverse of the observed information matrix. 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, selection bias and the propensity score models are all correct. However, when $\varepsilon_i(\theta)$ fails to be normal, the variance-covariance matrix of $\hat{\theta}$ should be estimated using either the standard sandwich formula or the bootstrap. An alternative, potentially less efficient two-stage approach may also be considered whereby an estimate of α is obtained by maximizing the partial log-likelihood Σ_i log f_2 ($R_i|Z_i, \mathbf{X}_i; \alpha$), and the remaining parameters are obtained by maximizing $\Sigma_i R_i$ log f_1 ($\varepsilon_i(\theta)|\mathbf{X}_i, Z_i; \sigma^2$) with the maximum partial likelihood estimator of α subtituted in. A potential advantage of the two-stage approach is that it may more easily be performed using standard statistical software for regression analysis. Because the linear regression is a complicated nonlinear function of the parameter indexing the propensity score; the log likelihood function (3) may be multimodal at small to moderate sample sizes. Although bound to be asymptotically less efficient than the maximum likelihood estimator, the two-stage estimator may be particularly valuabe as providing a good starting value for optimization of the log likelihood function.

In the Supplementary Materials, we extend the methodology developed in this section to accommodate a log link function and give a result analogous to Result 1 for the log case.

4 Inference with the logit link

In this section, we consider regression analysis for a binary outcome using a logit link function, with the population model of interest now defined as: interest as followed,

$$\mu(\mathbf{X}) = logit\, Pr(Y=1|\mathbf{X}) = log ODDS(\mathbf{X}) = log\frac{Pr(Y=1|\mathbf{X})}{Pr(Y=0|\mathbf{X})}. \quad (4)$$

Likewise, let $ODDS(\mathbf{X}, R = 1) = \Pr(Y = 1|\mathbf{X}, R = 1)/\Pr(Y = 0|\mathbf{X}, R = 1)$. We begin by defining the selection bias on the odds ratio scale. Note that

$$\frac{ODDS(\mathbf{X},R=1)}{ODDS(\mathbf{X})} = \frac{ODDS(\mathbf{X},R=1)}{ODDS(\mathbf{X},R=0)} \Big/ \left\{\sum_{r=0}^{1}\frac{ODDS(\mathbf{X},R=r)}{ODDS(\mathbf{X},R=0)} Pr(R=r|\mathbf{X},Y=0)\right\}$$
$$= \tilde{\omega}(\mathbf{X})\{\tilde{\omega}(\mathbf{X})\, Pr(R=1|\mathbf{X},Y=0) + Pr(R=0|\mathbf{X},Y=0)\}^{-1},$$

where $\tilde{\omega}(\mathbf{X}) = ODDS(\mathbf{X}, R=1)/ODDS(\mathbf{X}, R=0)$ and, where we have used the following key property of the odds function (See Tchetgen Tchetgen, 2013), $ODDS(\mathbf{X}) = E\{ODDS(\mathbf{X}, R)|\mathbf{X}, Y=0\}$, the function $\tilde{\omega}(\mathbf{X})$ 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 $\tilde{\omega}(\mathbf{X})=1$ or $\Pr(R=1|\mathbf{X}, Y=0) = 1$, that is if and only if there is no selection bias or no missing data among noncases. 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; where $ODDS(\mathbf{X}, R=1, Z) = \Pr(Y=1|\mathbf{X}, R=1, Z)/\Pr(Y=0|\mathbf{X}, R=1, Z)$.

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

Result 2: Under assumptions (IV.1)–(IV.3†), the regression function $\mu(\mathbf{X})$ is nonparametrically identified from the observed data $\mathbf{O}$*, 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 a function of* $\mu(\mathbf{X})$, $\nu(\mathbf{X})$ *and* $\pi(\mathbf{X}, Z)$ *as follows:* $logitn(\mathbf{X}, Z) = \mu(\mathbf{X}) + \omega(\mathbf{X}) - \overline{\omega}(\mathbf{X}, Z)$*, where* $\overline{\omega}(\mathbf{X}, Z) = log\{exp\{\omega(\mathbf{X})\}\lambda(\mathbf{X}, Z) + 1 - \lambda(\mathbf{X}, Z)\}$*, and* $\lambda(\mathbf{X}, Z) = \Pr(R=1|\mathbf{X}, Z, Y=0)$ *satisfies*

$$\pi(\mathbf{X}, Z) = \{1 - expit\{\mu(\mathbf{X})\}\}\lambda(\mathbf{X}, Z) + expit\{\mu(\mathbf{X})\}[1 + (1 - \lambda(\mathbf{X}, Z))exp\{-\omega(\mathbf{X})\}/\lambda(\mathbf{X}, Z)]^{-1}.$$

Result 2 states that the regression curve $\Pr(Y=1|\mathbf{X}) = \text{expit}\{\mu(\mathbf{X})\}$ is identified from data $\mathbf{O}$ provided that Z is an IV which satisfies (IV.1)–(IV.3†). The result gives an explicit representation for the complete-case logistic regression $\pi(\mathbf{X}, Z)$ as a function of the regression of interest $\mu(\mathbf{X})$, the selection bias function $\omega(\mathbf{X})$ and $(\mathbf{X}, Z)$.

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

$$\sum_i R_i\{log\ Pr(Y_i=1|R_i=1, \mathbf{X}_i, Z_i) + log\,\pi(\mathbf{X}_i, Z_i)\} + (1 - R_i)\,log\,(1 - \pi(\mathbf{X}_i, Z_i)), \tag{5}$$

using the parametrization of Result 2. For instance, suppose that one aims to estimate the logistic regression model $\mu(\mathbf{X}; \psi) = (1, \mathbf{X}')$. Further suppose that one specifies a similar linear model for the selection bias log odds ratio function $\omega(\mathbf{X}; \eta) = (1, \mathbf{X}')$, and assuming that $\text{logit}\lambda(\mathbf{X}, Z; \alpha) = (1, \mathbf{X}', Z')\alpha$, gives the following complete-case model,

$$logit\ Pr\left(Y{=}1|R{=}1,\mathbf{X},Z;\Psi,\alpha,\eta\right)=\left(1,\mathbf{X}^{'}\right)\Psi+\left(1,\mathbf{X}^{'}\right)\eta$$
$$-log\left(\begin{array}{c} expit\left\{\left(1,\mathbf{X}^{'},Z^{'}\right)\alpha\right\}exp\left\{\left(1,\mathbf{X}^{'}\right)\eta\right\} \\ +1-expit\left\{\left(1,\mathbf{X}^{'},Z^{'}\right)\alpha\right\}\end{array}\right)$$
$$\pi\left(\mathbf{X},Z;\Psi,\alpha,\eta\right)=\left\{1-expit\left\{\left(1,\mathbf{X}^{'}\right)\Psi\right\}\right\}expit\left\{\left(1,\mathbf{X}^{'},Z^{'}\right)\alpha\right\}$$
$$+expit\left\{\left(1,\mathbf{X}^{'}\right)\Psi\right\}\left[1{+}exp\left\{-\left(1,\mathbf{X}^{'},Z^{'}\right)\alpha\right\}exp\left\{-\left(1,\mathbf{X}^{'}\right)\eta\right\}\right]^{-1}.$$

The maximum likelihood estimator of (ψ, α, η) maximizes the loglikelihood (5) under Pr(Y = 1|R = 1, **X**, Z; ψ, α, η) and π(**X**, Z; ψ, α, η). Inference may then proceed using standard maximum likelihood theory.

5 Detecting the presence of selection bias

Interestingly, if Z is known to satisfy assumptions (IV.1) and (IV.2) but not assumption (IV. 3) (nor (IV.3†) and (IV.3′) of the Supplementary Materials), it can no longer be used to correct for selection bias. However, as we argue next, such a variable may still be used as a tool for detecting the presence of selection bias. This is because (IV.3) ((IV.3′) and (IV.3†)) are trivially satisfied under the null hypothesis of no selection bias, i.e. if $H_0{:}\delta(\mathbf{X}){=}0(H_0^*{=}log\ \nu(\mathbf{X}){=}0$ in the Supplementary materials, and $H_0^{\dagger}{=}\omega(\mathbf{X}){=}0)$ for all **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 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 for most alternatives be consistent under the hypothesis that selection bias is present on that scale, regardless of whether assumption (IV.3) ((IV.3′) or (IV.3†)) holds.

6 Partial identification via smooth bounds for binary outcome

Assumption (IV.3†) is not empirically testable and may at best only be approximately correct in a given application. For this reason, it is crucial in practice, to supplement the proposed IV approach with IV inferences based only on assumptions (IV.1) and (IV.2) for validity. It is straightforward to check that for binary Y, the population mean is contained in the interval defined by so-called assumption-free bounds that do not make use of Z,i.e. $p = E(Y) \in$[Pr { Y = 1, R = 1}, Pr { Y = 1, R = 1}+Pr { R = 0}]. However, the length of this interval is equal to the proportion of missing data and therefore most informative when the proportion of missing data is negligible in which case selection bias is less of a concern. Robins (1989) and Manski (1990) independenlly proposed IV bounds for the mean of Y which can be considerably tighter than the assumption-free bounds. Specifically, in the case of binary outcome, polytomous Z and assuming for simplicity that (IV.1) and (IV.2) hold without conditioning on **X**, the mean of Y is contained in the interval [LB, UB] defined by Robins-Manski IV bounds: LB = max{Pr { Y = 1, R = 1|z}: z} and UB = min {Pr { Y = 1, R = 1|z} + Pr {R = 0|z} − max {− Pr { Y = 1, R = 1|z} − Pr {R = 0|z}: z}. There are two potential difficulties with implementing the Robins-Manski bounds in practice. To discuss these difficulties, note that in practice the bounds must be estimated from the observed data **O**. Let θ denote a finite dimensional parameter indexing models for Pr { Y = 1|R = 1, z; θ and Pr {R

= 1|*z*, *θ*}. Let $\overline{\theta}$ denote the maximum likelihood estimator of *θ*, $a\left(z;\overline{\theta}\right)=Pr\left\{Y=1,R=1|z;\overline{\theta}\right\}$, $q\left(z;\overline{\theta}\right)=a\left(z;\overline{\theta}\right)+Pr\left\{R=0|z;\overline{\theta}\right\}$, and $LB(\overline{\theta})$ and $UB(\overline{\theta})$ denote the estimated lower and upper bounds. The first difficulty arises in accounting for the uncertainty due to estimation of *θ*, because *LB* (·) and *UB* (·) are non-smooth functionals of *θ* so that the delta method cannot be used to obtain the large sample distribution of $LB(\overline{\theta})$ and $UB(\overline{\theta})$. The second difficulty is that in finite sample $LB(\overline{\theta})$ may be greater than $UB(\overline{\theta})$ due to sampling variation, in which case the estimated bounds are not coherent. We propose to resolve both difficulties by first replacing the nonsmooth max function with a smooth approximation known as the softmax function defined for fixed scalar parameter *ν* as the mapping $(b_1,\ldots b_K)\rightarrow softmax_\nu(b_1,\ldots b_K)=\nu^{-1}log\left\{\sum_k exp(\nu b_k)\right\}$ for *K* real numbers (b_1,…b_K). For large values of *ν*, softmax$_\nu$ (b_1,…b_K) ≈ max (b_1,…b_K), and while the right-hand side is not differentiable in b_k the left-hand side is. To address the second difficulty we leverage the availability of a point estimate of *E*(*Y*|*X*) under assumptions (IV.1)–(IV.3) (respectively (IV.3′) or (IV.3†)) as proposed in previous Sections. Because this point estimate is obtained under a submodel of the IV model defined by (IV.1) and (IV.2) only, it must be contained within the IV bounds with probability tending to one as sample size goes to infinity. We may refine the smoooth bounds by ensuring that this property holds in finite sample. In this vein, let $\hat{p}$ denote the IV estimator of *p* obtained using results from Section 4, e.g. $\hat{p}=n^{-1}\sum_i expit\left\{\left(1,\mathbf{X}_i'\right)\hat{\Psi}\right\}$. A smooth estimator of the IV lower bound may be defined as $SLB_{v,p}\left(\hat{p},\overline{\theta}\right)=v^{-1}\log\left\{\sum_z\phi\left\{p\left(\hat{p}-a\left(z;\overline{\theta}\right)\right)\right\}\exp\left[va\left(z;\overline{\theta}\right)\right]\right.$ where $\phi\left\{\rho(\hat{p}-a(z;\overline{\theta}))\right\}=expit\left\{\rho(\hat{p}-a(z;\overline{\theta}))\right\}$ is a smooth approximation of the indicator function $1\left\{\hat{p}\geq a(z;\overline{\theta})\right\}$ for sufficiently large *ρ*, ensuring that the lower bound does not exceed the point estimator $\hat{p}$. Likewise, we define a smooth estimator of the IV upper bound as $ULB_{v,p}\left(\hat{p},\overline{\theta}\right)=-v^{-1}\log\left\{\sum_z\phi\left\{p\left(q\left(z;\overline{\theta}\right)-\hat{p}\right)\right\}\exp\left[-vq\left(z;\overline{\theta}\right)\right]\right.$. For sufficiently large values of *ν* and *ρ*, the interval $SB_{\nu,\rho}(\hat{p},\overline{\theta})=[SLB_{\nu,\rho}(\hat{p},\overline{\theta}),SUB_{\nu,\rho}(\hat{p},\overline{\theta})]$ is asymptotically approximately equal to Robins-Manski bounds and therefore guaranteed to include the true population mean *p* under assumptions (IV.1) and (IV.2). Furthermore, assuming that $(\hat{p},\overline{\theta})$ is asymptotically normal, a straightforward application of the delta method implies that $[SLB_{\nu,\rho}(\hat{p},\overline{\theta})-1.96\hat{\sigma}^L_{\nu,\rho},SUB_{\nu,\rho}(\hat{p},\overline{\theta})+1.96\hat{\sigma}^U_{\nu,\rho}]$ is a valid asymptotic 95% confidence interval for *p*, where $\hat{\sigma}^L_{\nu,\rho}$ and $\hat{\sigma}^U_{\nu,\rho}$ are consistent estimators of the standard error of $SLB_{\nu,\rho}(\hat{p},\overline{\theta})$ and $SUB_{\nu,\rho}(\hat{p},\overline{\theta})$, respectively, obtained by a straightforward application of the delta method, or the nonparametric bootstrap. In practice, one may choose *ρ* and *ν* such that larger values of either parameter do not substantially change results.

7 Empirical Illustration

To illustrate the proposed instrumental variable approach, we obtained data from the 2007 Zambia Demographic and Health Survey to estimate HIV prevalence among adult men adjusting for selective non-participation in the HIV testing component of the survey study. Further details regarding 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 was 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. 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 agreed to participate in the individual interview, but refused to be tested for HIV (N=1,317).

7.1 Instrumental variables

In order to select instrumental variables, we adapted the approach used by Bärnighausen and colleagues (2011). 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. Likewise, the chances of encountering and enrolling eligible individuals are higher for households reached early in data collection because there are more opportunities for repeat visits by data collectors. Both the specific interviewer deployed to a household and the timing of that visit were determined at random (or by a known algorithm) (CSO, 2009). As a result, these factors are unlikely to directly influence an individual's HIV status upon adjusting for key correlate of HIV prevalence such as household geographic location. In the 2007 Zambia DHS survey, 53 distinct interviewers conducted 50 or more household interviews with men, the remaining interviewers were pooled into a 54th category, 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 ($P<0.001$).

7.2 Propensity Score and Selection Bias Models

For estimation, we used standard logistic regression 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. Likewise, we modeled the probability of participation in the survey HIV testing component (R) using the model described in Section 4 with 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, indicating that assumption (IV.2) holds in this sample. Finally, we modeled the selection bias function as linear in $\mathbf{X}$ on the log odds ratio scale: Table 3 provides a complete description of this last model and gives evidence of statistically

significant selection bias in the odds ratio association between education and household location type, with HIV prevalence.

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

All statistical analyses were conducted using SAS software version 9.3 (SAS Institute, Cary, NC). We applied standard Taylor-series expansion arguments to derive the following large-sample variance estimator for the resulting point estimate $\tilde{p}$ of HIV prevalence, which simultaneously acknowledges the uncertainty due to estimation of $t(\mathbf{X})$, and the presence of sampling weights, $\widehat{Var}(\tilde{p})=\hat{\Lambda}+\hat{\Gamma}\hat{\Omega}\widehat{\Gamma'}$, where $\hat{\Lambda}=n^{-2}\sum_i W_i\left\{\hat{t}(\mathbf{X}_i)-\tilde{p}\right\}^2$, $\hat{\Gamma}=n^{-1}\sum_j W_j(1,\mathbf{X}_j')\hat{t}(\mathbf{X}_j)(1-\hat{t}(\mathbf{X}_j))$, and $\hat{\Omega}=\widehat{Var}(\hat{\Psi})$ was obtained from the inverse information matrix of the mle of (ψ, α, η) for the loglikelihood derived in the previous section. The survey weights were not used to obtain $\hat{t}(\mathbf{X}_i)$ because conditioning on the covariates $\mathbf{X}$ gave virtually the same results for the first stage whether the weights were included or not (data not shown), although the weights were used to obtain $\hat{p}$. We used the above estimated standard errors to construct Wald-type 95% confidence intervals (CIs) for p.

7.3 A random effect formulation

In the above formulation of the IV model, interviewer identity is entered in the nonresponse regression as a fixed effect, and thus in principle, may not be well estimated for an interviewer with few interviews. To address this potential concern, we also explored an alternative random effect formulation, whereby interviewer identity is specified as a mean zero normal random effect in the nonresponse regression, i.e. Letting Z_1 denote indicator of visit on first day, and Z_2 interviewer identity, $logit\lambda(\mathbf{X}, z;\alpha)=(1,\mathbf{X}', z_1')\alpha+b_{z_2}, b_{z_2}$ is mean zero normal with variance $\sigma^2_{z_2}$ which is estimated jointly with the fixed effects (associated with $\mathbf{X}$) in models for Y and R by maximum likelihood using PROC NLMIXED in SAS.

7.4 Results

We obtained a complete-case crude estimate of HIV prevalence of 12.2% (95% CI: 11.2% to 13.1%) compared to the IV-adjusted estimate of 21.1% (95% CI: 16.2% to 25.9%) obtained using the proposed approach (Table 4). Furthermore, we computed three-well known estimators of p that rely on the assumption that HIV status is missing at random conditional on $(\mathbf{X}, Z)$: inverse probability weighting, outcome regression and doubly robust estimation (Wirth, Tchetgen Tchetgen and Murray, 2010). All three estimators produced results nearly identical with the complete-case analysis clearly failing to account for selection bias. The IV-adjusted point estimate obtained using our methods essentially agree with the IV-corrected estimates obtained via a parametric Heckman-type selection model for a binary outcome used by Bärnighausen et al (2011), which we replicated to produce a prevalence estimate of 21.0% (95% CI: 19.8% to 22.2%). 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 analysis. The 95% CI obtained implementing the Heckman-type estimator as outlined in Bärnighausen et al (2011) is considerably narrower than the one obtained with our approach. The observed difference is primarily due to the fact that our CI accurately reflects all sources of uncertainty including from the first stage estimation of $t(\mathbf{X})$, whereas the CI of Bärnighausen et al (2011) does not appropriately account for the uncertainty due to the analogous preliminary estimation of $\Pr(R = 1|\mathbf{X})$ obtained with Heckman's model. Therefore their reported CI likely understates the actual uncertainty around Heckman's estimator. Finally, the random effect formulation of the model produced an estimate of HIV prevalence aligned with the fixed effect estimate (see Table 4).

7.5 Smooth Bounds for Zambia DHS data

In order to account for uncertainty about assumption (IV.3†), we computed the proposed smooth bounds. For simplicity, we ignored covariates in computing these bounds which may lead to some efficiency loss but is unlikely to invalidate IV assumptions (IV.1) and (IV.2). Thus, lower and upper bounds were evaluated using interviewer-specific empirical estimates of proportions $\Pr\{R = 1, Y = 1|z\}$ and $\Pr\{R = 0|z\}$ defining $\overline{\theta}$. We report estimated bounds for the choice of $\rho = \nu = 100$ as larger values of either parameter did not substantially change results. The lower and upper bounds with corresponding 95% confidence intervals were $SLB_{\nu,\rho}(\hat{\mu}, \overline{\theta})=0.17$ (95%CI: 0.14–0.21) and $SUB_{\nu,\rho}(\hat{\mu}, \overline{\theta})=0.22$ (95%CI: 0.17–0.27) therefore producing a 95%CI of HIV prevalence in Zambia equal to (14%–27%), only slightly wider than the 95%CI obtained under assumptions (IV.1)–(IV.3†).

8 Final remarks

This paper has considered the somewhat pernicious problem of selection bias due to a regression outcome missing not at random. We have shown that this seemingly intractable problem can be made more tractable with the aid of an IV. Straightforward, 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. Straightforward bounds were also obtained allowing for inferences with a valid IV appropriately accounting for uncertainty about assumption (IV.3) needed for identification. Although not pursued here, our smooth bounds can easily be generalized to any bounded outcome.

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. Finally, in future work, we plan to explore semiparametric doubly robust estimators that are guaranteed to remain consistent even under partial model mis-specification of the observed data likelihood.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Arabmazar A, Schmidt P. Further evidence on the robustness of the Tobit estimator to heteroscedasticity. J of Econometrics. 1981; 17:253–258.

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. [PubMed: 21150352]

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.

Das M, Newey WK, Vella F. Nonparametric estimation of sample selection models. Review of Economic Studies. 2003; 70:33–58.

Diggle PD, Kenward MG. Informative drop-out in longitudinal data analysis. Journal of the Royal Statistical Society: Series C. Applied Statistics. 1994; 43:49–93.

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

van der Laan, MJ., Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. Springer Verlag; New York: 2003.

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

Heckman JJ. Samples election bias as a specification error. Econometrica. 1979; 47:153–61.

Heckman JJ. Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. Journal of Human Resources. 1997; 32:441–462.

Lepkowski, JM., Couper, MP. Nonresponse in the second wave of longitudinal household surveys. In: Groves, RM.Dillman, DA.Eltinge, JL., Little, RJA., editors. Survey Nonresponse. Wiley; New York: 2002.

Little, RJA., Rubin, DB. Statistical Analysis With Missing Data. 2nd. New York: Wiley; 2002.

Manski CF. Nonparametric Bounds on Treatment Effects," American Economic Review. Papers and Proceedings. 1990; 80:319–323.

Manski, CF. Partial identification of probability distributions. Springer-Verlag; 2003.

Nicoletti C, Peracchi F. Survey response and survey characteristics: Microlevel evidence from the European Community Household Panel. J Roy Statist Soc A. 2005; 168:119.

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

Rotnitzky A, Robins JM. Analysis of semiparametric regression models with non-ignorable non-response. Statistics in Medicine. 1997; 16:81–102. [PubMed: 9004385]

Robins, J. The Analysis of Randomized and Non-randomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies". Sechrest, L., editor. H.Public Health Service; Washington, DC: 1989. p. 113-159.

Freeman, Mulley, A., editors. Health service Research Methodology: A Focus on AIDS. U.S.:

Robins J. Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models. Communications in Statistics. 1994; 23:2379–2412.

Robins, JM., Rotnitzky, A., Scharfstein, D. Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models. In: Halloran, ME., Berry, D., editors. Statistical Models in Epidemiology: The Environment and Clinical Trials. NY: Springer-Verlag; 1999. p. 1-92.IMA Volume 116

Roy J. Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. Biometrics. 2003; 59:829–836. [PubMed: 14969461]

Rubin, DB. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons; 1987.

Stolzenberg RM, Relies DA. Theory testing in a world of constrained research design: the significance of Heckman's censored sampling bias correction for nonexperimental research. Sociol Meth Res. 1990; 18:395–415.

Tchetgen Tchetgen, Eric J. General Regression A Framework for a Secondary Outcome in Case-control Studies. Biostatistics. 2013; 15(1):117–128. [PubMed: 24152770]

Winship C, Mare R. Models for sample selection bias. Annu Rev Sociol. 1992; 18:327–350.

Wirth K, Tchetgen Tchetgen E, Murray M. Adjustment for missing data in complex surveys using doubly robust estimation: Application to commercial sexual contact among Indian men. Epidemiology. 2010; 21(6):863–871. [PubMed: 20881599]

Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics. 1988; 44:175–188.

Table 1

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

	Log OR	HIV seropositivity (95% CI)	P^a
Intercept	−3.369	(−4.824, −1.914)	<0.0001
Age (years)			<0.0001
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)	0.02
Wealth quintile		–	0.02
5^{th} (wealthiest)	0.858	(0.191, 1.524)	
4^{th}	0.989	(0.363, 1.614)	
3^{rd}	0.494	(−0.107, 1.095)	
2^{nd}	0.394	(−0.214, 1.002)	
1^{st} (poorest)	(ref)	–	
Location type of household			0.0001
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)	–	

[a] *P*-value from Wald χ^2 test.

Table 2

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

	Log OR	HIV testing participation (95% CI)	P^a
Intercept	−0.075	(−0.811, 0.660)	0.84
Age (years)			0.30
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)	<0.0001
Wealth quintile			0.28
5^{th} (wealthiest)	0.037	(−0.503, 0.576)	
4^{th}	0.144	(−0.322, 0.610)	
3^{rd}	−0.241	(−0.601, 0.118)	
2^{nd}	−0.286	(−0.646, 0.075)	
1^{st} (poorest)	(ref)	–	
Location type of household			0.11
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)	0.16
Interviewer identity[b]	–	–	<0.0001

[a] P-value from Wald χ^2 test.

[b] Log ORs and 95% CIs comparing each of the 47 interviewers with ≥50 household interviews to a combined reference group of all other interviewers with <50 household interviews not shown for the purpose of simplicity; data are available upon request from the corresponding author.

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.

	Log OR	Selection bias (95% CI)	P^a
Intercept	−0.300	(−3.542, 2.942)	0.86
Age (years)			0.16
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)	0.01
Wealth quintile		–	0.52
5^{th} (wealthiest)	−0.686	(−2.529, 1.156)	
4^{th}	−0.579	(−2.320, 1.162)	
3^{rd}	0.166	(−1.543, 1.875)	
2^{nd}	0.142	(−1.545, 1.830)	
1^{st} (poorest)	(ref)	–	
Location type of household			0.44
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)	–	

[a] *P*-value from Wald χ^2 test.

Table 4

Prevalence of and 95% confidence intervals (CI) for HIV seropositivity among 7,116 adult men in Zambia according to method of adjustment for HIV testing non-participation.

	Prevalence	(95% CI)
Unadjusted		
Complete case analysis	12.2%	(11.2%, 13.1%)
Adjusted[a]		
Inverse probability weighting	12.4%	(11.5%, 13.3%)
Outcome regression	12.5%	(11.8%, 13.2%)
Doubly robust estimator	12.5%	(11.6%, 13.4%)
Heckman-type selection	21.0%	(19.8%, 22.2%)
Instrumental variable		
Fixed effects	21.1%	(16.2%, 25.9%)
Mixed effects	21.2%	(12.6%, 29.8%)

[a]All adjustment methods considered the following set of covariates: age (in 5-year categories), wealth (in quintiles), educational attainment (in years) and location type of household.