Online Technical Appendix. National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models

Contents

1. Heckman-type selection model equations

Dubin and Rivers described the model equations that extend Heckman's original method to the case of a dichotomous outcome, such as HIV status.[1, 2] The equation that predicts participation in HIV testing for individual *i* (s_i) is the following probit model [3]:

$$s_i^* = \beta_s x_i + \phi z_i + u_i$$
$$s_i = 1 \; if \; s_i^* > 0, s_i = 0 \; otherwise$$

where *x* are observed characteristics, *z* are selection variables subject to an exclusion restriction, and *u* is random error. HIV status h_i is observed if $s_i = 1$. The equation for the HIV status of individual *i* (h_i) is predicted with a second probit model:

$$h_i^* = \beta_h x_i + \varepsilon_i$$
$$h_i = 1 \; if \; h_i^* > 0, h_i = 0 \; otherwise$$

where *x* are observed characteristics and ε is random error. The error terms *u* and ε are assumed to be distributed bivariate normal, and the parameter $\rho = corr(u, \varepsilon)$ measures the magnitude and direction of the correlation between participation and HIV status on the probit scale after controlling for the variables in *x.* A negative value of ρ would indicate that individuals who are more likely to be HIV positive are less likely to participate in testing, conditional on observed variables. Note that the conventional imputation probit model is nested within the bivariate probit selection model, and it can be thought of as a selection model that assumes ρ=0 with certainty.

2. Regression variables

The DHS system uses standardized questionnaires, and country specific questions are recoded to allow for comparisons across countries and surveys.[4] We used the same set of variables in conventional probit and selection model-based imputation regression models across surveys whenever possible, following previous work.[3] For those who completed an individual questionnaire, these variables included age, educational attainment, household wealth quintile as constructed from an index of household assets, urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behavior in the past year, condom use at last sex, sexually transmitted disease in the past year, tobacco and alcohol use, knowing someone with AIDS, willingness to care for a family member with AIDS, and having had a previous HIV test.[3] In some cases, we used only one of two variables when they were highly collinear (e.g., when there was nearly complete overlap between ethnicity and language). In a small departure from the Zambian 2007 analysis, we defined the "married" variable with three categories (i.e., never married, currently married, and formerly married), as widowed individuals may be at high risk for HIV infection. For those individuals for whom information was only available from the household questionnaire, we controlled for sex, age, education, wealth quintile, urban setting, and region. In Senegal 2005, which had low prevalence among men, we used wider age categories to ensure that there were HIV positive individuals in each category. Rates of missing observations for covariates were low across surveys, typically within the range of 2-4% of individuals missing at least one covariate observation on the individual questionnaire. We formed a single HIV status variable for surveys that reported HIV-1 and HIV-2 status.

For the selection models, we operationalized interviewer identity by creating a dummy variable for each interviewer. Interviewers who conducted at least 50 interviews were assigned their own dummy variable and those who conducted fewer than 50 interviewers were combined in an 'other interviewer' dummy variable.[3] Estimating the effect that interviewers who conduct very few interviews have on participation in testing is difficult and can lead to lack of identification or to numerical problems in obtaining estimates. In Malawi 2004 we used 30 interviews as the minimum threshold when assigning interviewers unique dummy variables, as many interviewers in these surveys did not complete at least 50 interviews. We explored using a threshold of 30 interviews across surveys but encountered model convergence issues with this approach in some settings.

3. Accounting for survey design

We employed household sampling weights to calculate nationally representative estimates of HIV prevalence for all three modeling strategies. The use of household weights is more appropriate than individual weights, which are adjusted for non-participation, as we correct for non-participation in our analysis. We incorporated sampling weights after estimating regression models, as the variables used to construct the sampling weights were included as regression covariates. Thus, for both imputation-based modeling strategies, regressions were fit without sampling weights, HIV status was predicted for those without a valid HIV test, and then a sampling-weighted average was calculated for those predictions. We accounted for survey strata and household clustering when estimating the covariance matrix of regression parameters.

4. Parametric simulation of 95% confidence intervals

We employed a parametric simulation approach to generate uncertainty intervals around imputation-based HIV prevalence estimates, which incorporates uncertainty about imputed HIV status and sampling variation.[5, 6] We simulated the sampling distribution of predicted prevalence for the two groups of people who were missing a valid HIV test—those who could not be contacted and those who refused consent—using the same procedure for conventional imputation and selection-model-based imputation strategies. First, we fit the regression model and saved the maximum likelihood estimates of the coefficients and their covariance matrix, which was adjusted to account for the complex survey design. In the case of the selection model, these coefficients included those from the selection and outcome equations and the correlation parameter ρ. Next, 10,000 regression parameter sets were drawn from a multivariate normal distribution parameterized by the coefficients and covariance matrix obtained in the first step.[5] For each set of regression parameter draws, we predicted HIV status and calculated sampling-weighted mean prevalence for those missing a valid HIV test. Aggregating these prevalence estimates across simulation draws approximated the sampling distribution of imputed prevalence for those missing a valid HIV test.

Obtaining 95% confidence intervals for national estimates of HIV prevalence required combining the uncertainty around imputed prevalence estimates for nonparticipants as described above with the sampling uncertainty around the prevalence estimate for those with observed HIV status. To incorporate uncertainty for the latter, we first simulated 10,000 prevalence values from a binomial distribution, parameterized with a probability equal to the complete case estimate for prevalence and a population size appropriate for the complex survey design. To approximate the sampling distribution for national HIV prevalence, the simulated values for HIV prevalence among those with a valid HIV test cannot be combined at random with the simulated values for imputed prevalence for those missing a valid HIV test because of correlated sampling uncertainty around these estimates. To address this, we induced correlation between the sets of simulated prevalence values with an empirical distribution copula method.[7] This procedure involves rank-ordering two vectors and then re-ordering them so as to induce a pre-specified amount of correlation in their values. We first used the copula method to combine the two vectors of imputed prevalence values (i.e., estimates for those who could not be contacted and those who refused consent). Then, we combined this vector with the simulated values from the sampling distribution for the complete-case analysis.

For the copula method, we used the average of the correlation coefficients calculated from comparisons of bootstrapped draws around prevalence estimates from analyses of the Cote d'Ivoire, Zambia and Zimbabwe surveys (correlation coefficients were similar across surveys and between men and women; see section below for description of bootstrapping procedure). For the conventional imputation analyses that relied on a probit regression, the correlation between imputed prevalence for those who refused consent and those who could not to be contacted was 0.66, and the correlation between the combined imputed prevalence for those who did not have a valid HIV test and those with a valid HIV test was 0.67. For the selection model analyses, the correlation between imputed prevalence for those who refused consent and those who could not to be contacted was 0.46, and the correlation between the combined imputed prevalence for those who did not have a valid HIV test and those with a valid HIV test was 0.17.

5. Participation rates

The proportion of eligible individuals participating in HIV testing in the 12 DHS surveys included in the final analysis ranged from 63 to 96% in men and 70 to 97% in women, with higher participation rates among women (Table 1). Non-consent was the more common cause of non-participation in HIV testing for women, while men had similar rates of non-participation due to non-consent and non-contact. Considering men and women separately, the span in non-participation outcomes between the most and the least successful interviewers, in terms of either non-consent or non-contact, had a median value of 30 or more percentage points in all cases. All surveys had at least one interviewer with a non-participation rate below 9%, with the exceptions of Zambia 2007 (for which the lowest non-contact rate for men was 13%) and Zimbabwe 2005-6 (where the lowest non-contact rate for men was 12%).

6. Bootstrapped confidence intervals

The parametric simulation approach to generating 95% confidence intervals for imputation-based prevalence estimates makes strong distributional assumptions. The bootstrap is a more robust approach but was not feasible to implement for many surveys, for example due to collinearity between interviewer identities and the region variable. For comparison to the parametric simulation approach, we obtained bootstrapped confidence intervals for HIV prevalence imputed with the selection modeling approach in the Cote d'Ivoire 2005, Zambia 2007, and Zimbabwe 2005-6 surveys. To construct a bootstrap data set, we resampled clusters of households within each stratum. Across these three surveys, the bootstrapped 95% confidence intervals for HIV prevalence from the selection modeling approach for those refusing consent, for those who could not be contacted, and for the total national estimate were less conservative than those obtained from the parametric simulation approach, as shown below:

	Cote d'Ivoire 2005		Zambia 2007		Zimbabwe 2005-6	
	Simulation	Bootstrap	Simulation	Bootstrap	Simulation	Bootstrap
Men						
No consent	4.1, 40.3	8.6, 24.6	21.9, 82.5	34.6, 66.2	5.3, 42.0	10.7, 30.9
No contact	3.4, 68.5	12.3, 43.4	1.4, 69.7	8.2, 46.8	1.4, 55.8	6.2, 35.0
National	3.5, 18.4	5.5, 12.3	13.7, 29.7	17.0, 25.0	10.9, 25.3	12.9 20.3
Women						
No consent	1.9, 32.7	6.0, 20.7	7.5, 50.1	14.2, 35.7	6.2, 67.7	14.5, 49.7
No contact	1.8, 26.4	5.2, 18.0	1.3, 51.4	0.1, 28.9	0.0, 84.0	0.0, 41.6
National	5.0, 11.9	6.1, 9.7	13.7, 23.5	15.1, 20.2	17.0, 32.7	18.4, 26.6

7. Semi-nonparametric selection model

The parametric selection model used in the main analysis assumes that the error terms in the selection model are distributed bivariate normal. If this assumption was violated, it could impact the accuracy of the model's imputation results. There are limited choices among existing software packages for implementing models that relax the bivariate normality assumption. For our application, we used a semi-nonparametric selection model that approximates the unknown densities of the two error terms by Hermite polynomial expansions.[8] This is implemented in Stata in the –snp2s- command.[8] This approach is somewhat limited for our purposes as the

intercepts are not identified and therefore cannot be used for imputation. Thus, we only used it to estimate the selection model correlation parameter, ρ, for comparison to the estimate from the parametric model used in the main analysis. The semi-nonparametric model is computationally intensive to fit, so we only replicated the *consent* regressions for the sensitivity analysis. For each regression, we compared models fit under two possible specifications for the orders of the polynomial expansions: 3 for the selection model and 3 for the outcome model vs. 4 for the selection model and 4 for the outcome model. The preferred model was selected based on a likelihood ratio test,[8] except in a few cases where only one of the two expansion specifications converged, in which case the results from the converged model were used. Semi-nonparametric estimates of ρ were modestly correlated with those from the parametric model, with a correlation of 0.27, and tended to be closer to zero. All semi-nonparametric estimates of ρ were covered the 95% CI for ρ estimated with the bivariate probit selection model. The estimate for men in Zambia 2007 was similar but slightly lower, with ρ=-0.58 as compared to ρ=-0.75 from the parametric model. Given the limitations of this particular semi-nonparametric model, further development of semi- and nonparametric selection models is needed to establish strong tests of the bivariate normality assumption, which is a promising area for future research.

8. Simulation experiment of selection model sensitivity

If interviewers differ in their effect on participation in HIV testing, it is worth considering the sensitivity of the selection model to more complex interactions between interviewers and eligible individuals. For example, interviewer impact on participation could vary with the HIV status of respondents, which would violate the assumption of a constant value for ρ. Here we consider the case in which more successful interviewers obtain higher consent rates among those with HIV as compared to those without HIV. We used simulation to explore how this form of selection bias would affect estimates obtained from the selection model in comparison to complete case and conventional imputation analyses.

For the simulation, we used a simplified set of parameters informed from the analysis of men who refused consent in the Zambian 2007 DHS. We specified that $\rho = 0$ and generated HIV status for 5,000 individuals as:

$$h_i^* = -1.28 + 0.30 x_{1i} + \varepsilon_i$$

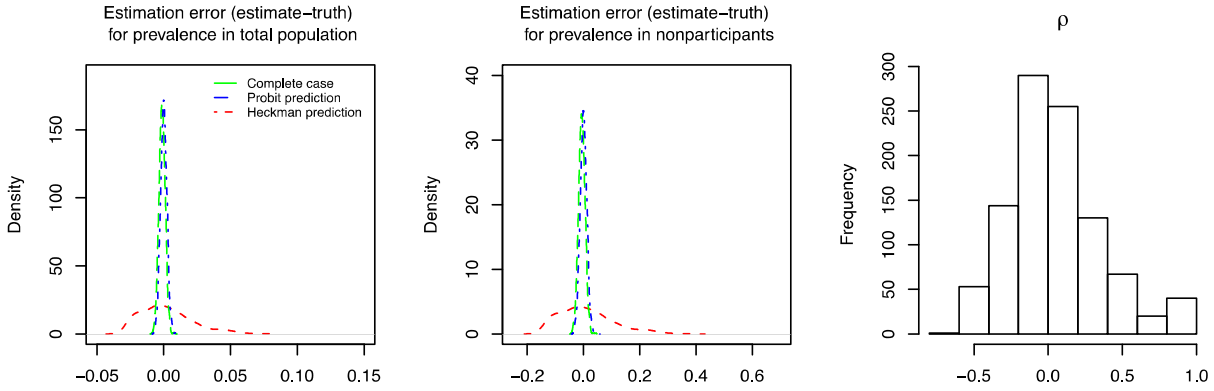
with ε~N(0,1) and $h_i = 1$ if $h_i^* > 0$. The variable x_1 denoted urban vs. rural regions, with 40% of the population located in an urban setting. For the base case, we generated participation status for the 5,000 individuals in the data set as follows, with each respondent assigned one of 34 interviewers who had unique effects on participation:

$$s_i^* = 0.7 + 0.24 x_{1i} + \boldsymbol{\phi}\mathbf{z}_i + u_i$$

where u ~ N(0,1), $\mathbf{z}_i$ is a vector that indicates which interviewer was assigned to respondent i, $\boldsymbol{\phi}$ is a vector with interviewer-specific participation effects, and $s_i = 1$ if $s_i^* > 0$. For half of the interviewers (group A, the successful interviewers), we assigned each interviewer j a unique participation effect ϕ_j ~ Uniform(0.28,0.68), and for the other half of the interviewers (Group

B), we drew ϕ_j ~ Uniform(-0.15,0.25). These specifications yielded an average participation rate of 86% in successful interviewer group A and 74% in interviewer group B, matching what was observed in the Zambian data set. They also yielded a distribution of participation rates across interviewers that was comparable to that observed in the Zambian data.

The data for this base case have no selection on unobserved factors and it is useful to compare the performance of the three modeling approaches explored in this paper in this context. As shown here in density plots of prevalence estimation error, comparing true sample means to those estimated with the three different modeling strategies across 1,000 simulated data sets, estimates from the Heckman-type selection model are unbiased but less precise than those obtained from either the complete case or standard probit imputation model:



The complete case analysis, which ignores the effects of x_1, leads to a slight underestimate of prevalence, as x_1 is associated with higher HIV prevalence and lower participation. For a small number of simulated data sets, the selection model estimated the correlation parameter ρ to be nearly equal to 1 (in many of these cases, the model failed to converge).

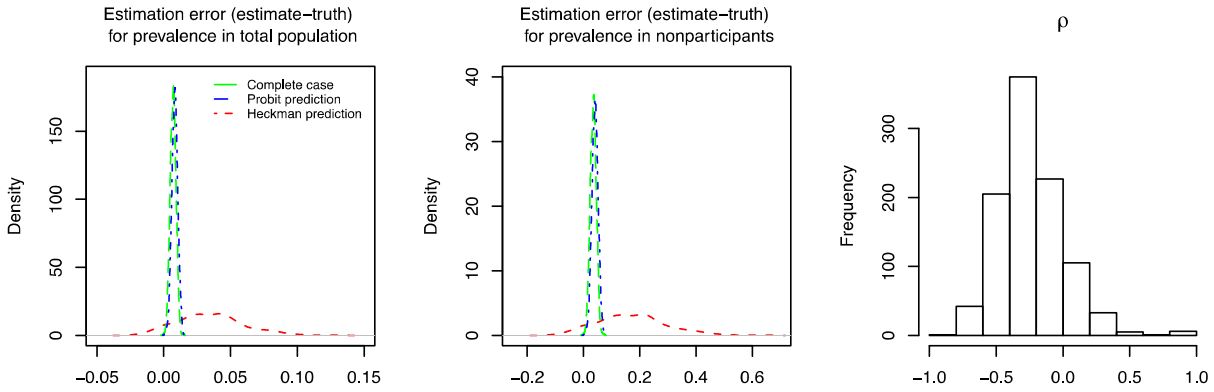
To explore the potential impact of differential interviewer effect by respondent HIV status, we regenerated participation status for respondents who had interviewers from the successful interviewers (group A) as follows:

$$s_i^* = 0.7 + 0.24\, x_{1i} + \lambda h_i \phi_{\mathbf{z}_i} + u_i$$

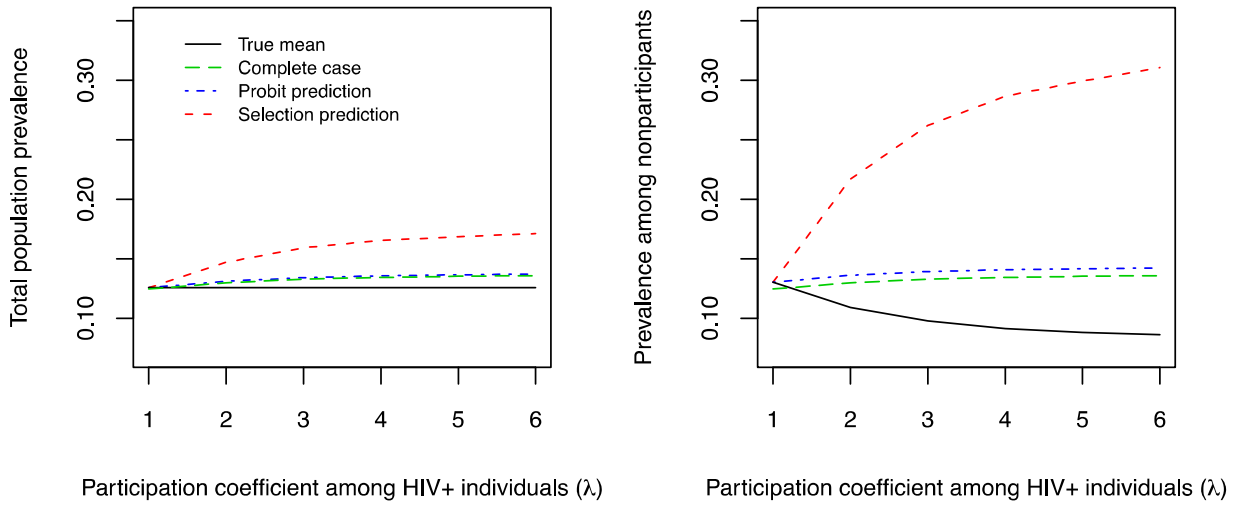
Larger positive values for λ yield higher participation rates for HIV positive individuals among successful interviewers. This mechanism generates selection bias in the data but the bias is of a different form than that which motivates the selection model. To maintain the same overall participation rates in interviewer group A across different values for λ, we reduced the absolute effect that each interviewer in group A had on participation by adjusting the uniform distribution for sampling values of ϕ_j. These distributions were parameterized as follows:

λ	Interviewer effect ϕ_j
1	ϕ_j ~ (0.28, 0.68)
2	ϕ_j ~ (0.23, 0.63)
3	ϕ_j ~ (0.21, 0.61)
4	ϕ_j ~ (0.19, 0.59)
5	ϕ_j ~ (0.18, 0.58)
6	ϕ_j ~ (0.17, 0.57)

These parameterizations result in a reduction in the proportion of HIV negative individuals in group A who participate as λ increases, maintaining an over all participation rate of 86% in group A. By way of example, if λ = 3, the data generated under these conditions leads to biased prevalence estimates in all three modeling strategies. The complete case and standard imputation analyses provide similar estimates, which are biased upwards. The selection model predictions are biased upwards to a greater extent than the complete case or conventional imputation model, as the model "corrects" in the wrong direction (i.e., ρ should be positive). The bias arises because there is relatively higher prevalence among consenters in the successful interviewer group, which leads to the model predicting higher prevalence among those who did not consent.



To systematically examine the relationship between λ and the amount of bias in predicted prevalence from different modeling strategies, we plotted mean estimates of prevalence across 1,000 simulations for the different values of λ. In most simulations, a value of λ=6 results in all HIV positive individuals participating within group A. The predicted prevalence estimates obtained from complete case, conventional probit, and selection model strategies are all biased for λ>1:



The magnitude of the difference between estimated and true prevalence increased nonlinearly with λ and suggests that systematic differences in interviewer consent rates by respondent HIV status do have the potential to lead to biased estimates of HIV prevalence with a selection model. However, the magnitude of the change in estimated prevalence in even the most extreme simulations was smaller than that estimated for adult men in the Zambia 2007 survey in the main analysis, suggesting that this violation of the model's assumptions, if it were to occur, would be unlikely to serve as an alternative explanation for our findings.

9. Software

Software commands implementing the bivariate probit model used in this study include: -heckprob- in Stata (StataCorp, College Station, TX), PROC QLIM in SAS (SAS Institute Inc., Cary, NC), and the sampleSelection (Henningsen and Toomet) and SemiParBIVProbit packages in R (Marra and Radice) in R (Foundation for Statistical Computing, Vienna, Austria).

References

1. Heckman J. Sample selection bias as a specification error. Econometrica 1979;**47**(1):153-62.
2. Dubin J, Rivers D. Selection bias in linear regression, logit and probit models. Sociological Methods and Research 1990;**18**(2 & 3):360-90.
3. Bärnighausen T, Bor J, Wandira-Kazibwe S, et al. Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. Epidemiology 2011;**22**(1):27-35.
4. Measure DHS. Demographic and Health Surveys (DHS) Final Reports. Secondary Demographic and Health Surveys (DHS) Final Reports. 2011. http://www.measuredhs.com.
5. King G, Tomz M, Wittenberg J. Making the most of statistical analyses: Improving interpretation and presentation. American Journal of Political Science 2000;**44**(2):341-55.
6. Timpone R. Estimating aggregate policy reform effects: New baselines for registration, participation, and representation. Political Analysis 2002;**10**(2):154-77.
7. Trivedi PK, Zimmer DM. Copula modeling: an introduction for practitioners. Foundations and Trends in Econometrics 2005;**1**(1):111.
8. Stover J, Brown T, Martson M. Updates to the Spectrum/EPP model to Estimate HIV Trends for Adults and Children [under review]. Sex Transm Infect 2012;**UNAIDS 2012 supplement**