Heckman selection model to correct HIV prevalence estimates for survey nonparticipation: Stata dofile - example DHS Zambia 2007

Version 22 December 2011

Introduction

We recently published revised estimates for national HIV prevalence among adult men and women from the Zambia 2007 Demographic and Health Surveys (DHS) using Heckman-type selection models to correct for potential selection bias due to non-participation in HIV testing during the DHS survey [1]. Our goal here is to briefly describe the rationale for the use of Heckman-type selection models in this context and provide annotated Stata code that can be used for conducting this type of analysis in the Zambia 2007 DHS and, after adaptation, in other datasets.

We currently have a paper under review for publication that applies this approach to all available DHS in sub-Saharan Africa and improves the calculation of confidence intervals for imputation-based HIV prevalence estimates [2]. Those interested in the development of these methods may wish to revisit this document in the future, as we may post additional information here.

Rationale

Low participation rates in HIV testing during national population-based surveys could lead to selection bias when estimating HIV prevalence if those infected with HIV are more or less likely to participate in HIV testing than those who are not infected with HIV. The conventional approach to correcting for potential selection bias is to impute missing HIV test results based on measured covariates, for example, using logistic or probit regression. For this approach to be valid, the expected HIV status of those missing an HIV test must be equal to those with an HIV test within the same strata of covariates. If this is assumption is violated, which could happen if HIV status itself influences an individual's propensity to test, then conventional imputation will not provide unbiased predictions of HIV status among non-participants.

Heckman-type selection models offer an alternative approach that has the potential to correct for selection on observed and unobserved covariates [1]. The bivariate probit regression used in our analyses can be thought of as a generalization of the probit regression model used for conventional imputation. Like a conventional probit model, it predicts HIV status conditional on observed covariates. However, it also contains an equation that predicts participation in HIV testing (termed the 'selection equation'), and allows for correlation between HIV status and HIV testing. The correlation parameter, which is commonly labeled ρ or rho, indicates the direction and magnitude of selection bias due to nonparticipation in HIV prevalence estimates, conditional on observed covariates.

To improve the identifiability of the Heckman-type selection model, one or more 'selection' variables that are subject to an exclusion restriction can be included in the selection equation. The exclusion restriction assumes that these variables determine selection into the sample (e.g.,

participation in HIV testing) but do not independently affect the outcome of interest (e.g., HIV status). The assumption that a variable predicts selection into the sample can be tested empirically, but it is not possible to test whether or not the variable affects the outcome other than through selection. In the Zambia 2007 DHS analysis, we identified two variables for which it is highly plausible that the two assumption are met – the identities of survey interviewers and an indicator variable for whether or not a household was visited on the first day that a survey team entered a cluster of households. Please see [1, 3] for a full discussion of the rationale for the use of variables and empirical tests of the assumption that the variables significantly determine HIV survey participation.

A second important assumption of the bivariate probit regression is that the error terms in the two regression equations are distributed bivariate normal. In future work, we plan to investigate non-parametric methods to determine the sensitivity of the model to this assumption and improve the robustness of model predictions.

Software

The bivariate probit regression employed for this analysis is available in common statistical software packages, including Stata (-heckprob-), SAS (PROC QLIM) and R (SemiParBIVProbit by Marra and Radice). As we used Stata for our analysis, we present the Zambia example that relies on the -heckprob- code available in Stata.

Data

The datasets for these analyses are available for free at
http://www.measuredhs.com/data/dataset/Zambia_Standard-DHS_2007.cfm
after application for permission to use the HIV datasets.

Stata code

The results that the code in the Stata dofile accompanying this description will produce are essentially the same as those shown in Bärnighausen et al. (2011) [1]. Minor differences are due to the following two changes: First, the DHS datasets have been updated since the work on the 2011 paper; second, we have used slightly different functional forms for a few of the control variables in this dofile, to make them consistent with the functional forms we used in the recent replication of this approach for all available DHS in sub-Saharan Africa [2].

Finally, please note that the dofile will produce regression results expressed as coefficients, while we show the marginal effects evaluated at the sample mean in the paper [1]. We have made this choice because in this analysis the Stata heckprob postestimation commands to produce marginal effects take a long time to execute on standard personal computers. Of course, the dofile user can easily adapt the Stata code to produce marginal effects.

References

1. Bärnighausen 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.
2. Hogan D, Salomon JA, Canning D, Hammitt JK, Zaslavksy A, Bärnighausen T. National HIV prevalence estimates for sub-Saharan Africa: revisiting the impact of survey non-participation. *(submitted)* 2011.
3. Bärnighausen T, Bor J, Wandira-Kazibwe S, Canning D. Interviewer identity as exclusion restriction in epidemiology. *Epidemiology* 2011,**22**:446.