Evaluation of Heckman Selection Model Method for Correcting Estimates of HIV Prevalence from Sample Surveys

via Realistic Simulation

Samuel J. Clark & Brian Houle

Working Paper no. 120
Center for Statistics and the Social Sciences
University of Washington

September 28, 2012

Contents

1 Correcting for Selection Bias in Estimates of HIV Prevalence from Sample Surveys

A common method of measuring HIV prevalence is to include HIV biomarker collection in household sample surveys primarily designed for a variety of other investigations. The usual procedure is:

1. Choose a sample.
2. Attempt to contact households and individuals in the sample. Some fraction of the sampled individuals are found and agree to be interviewed.
3. Interview the consenting members of the sample and in the course of the interview attempt to collect biomarkers. Some fraction of the interviewed individuals agree to provide biomarkers.
4. Conduct assays to provide HIV status for individuals who were i) found and agreed to interview and ii) agreed to provide biomarkers for testing.
5. Transform the HIV status of the individuals who were tested into *population-level* estimates of HIV prevalence.

There are two selection processes that separate the *population* from the individuals who agree to biomarker collection. At both stages, the selected sample can be (increasingly) systematically different from the population, resulting in bias in the population-level estimate of HIV prevalence. Under specific conditions the Heckman selection model is designed to identify the effect of selection in a single selection step. This ability can be used to identify a selection effect in the second stage of the process outlined above for a typical survey-based HIV prevalence study – the selection to test among those who are found and interviewed.

The correction is accomplished by:

1. Using the Heckman selection model to identify and quantify the selection bias in estimated HIV prevalence.
2. Using estimated results from the Heckman selection model to predict the HIV status of the individuals who did not agree to biomarker collection.
3. Combining the *real* data describing those who did agree to testing with the *predicted HIV status* of those who did not agree to testing to produce an estimate for the entire group of individuals who were found and agreed to be interviewed, *not* yet the whole population.

1.1 Aim: Does the Heckman Selection Model Correction Procedure Work?

The aim of this work is to test the Heckman selection model correction procedure. We want to characterize the effectiveness of the Heckman selection model procedure to correct estimates of HIV prevalence from sample survey data in a situation where we know the real prevalence.

1.2 Overview of the Heckman Selection Model

There are two processes at work. One mechanism leads to individuals being observed and another affects whether or not they are HIV+. These mechanisms work at the same time and may be related. The Heckman selection model assumes that we can model the selection process as a binary outcome with an individual either being observed $S = 1$ or not $S = 0$. Similarly, the process leading to being HIV+ can be modeled as a binary outcome with an individual either being HIV+ $H = 1$ or not $H = 0$. The catch is that the outcome H is only observed when the individual is selected. The Heckman probit technique models the binary outcomes as having normally distributed probabilities, hence both are probit models. After

accounting for the variation associated with known explanatory variables associated with both selection and outcome, the Heckman probit method assumes that each process is left with normally distributed errors. Possible correlation between selection and outcome is allowed by correlating these error terms. The Heckman probit estimation procedure simultaneously estimates the regression coefficients associated with the covariates explaining the selection and outcome processes *and* the value of correlation coefficient associated with the relationship between the error terms.

For the Heckman probit method to work well, there must be sets of explanatory variables associated with both selection and outcome, and optimally there should be one or more variables (the 'selection' variables) that only appear in the selection relationship. These selection variables should *not* affect the outcome. This yields two equations, one for selection that includes the selection variables, and one for the outcome: the 'selection' and 'outcome' equations.

There are standard procedures for estimating the biprobit Heckman selection model, and they strictly require:

1. the conditional joint error distribution be bivariate Normal, and
2. the 'instrumental' variables that appear in the selection equation, those that appear only in the selection equation, are not related to the outcome.

2 Method – Simulation Investigation

The biprobit Heckman selection model requires two strong assumptions, and in general it is not possible to test these assumptions using real data. Consequently we apply a simulation approach:

1. Define a number of simulation scenarios that fulfill the assumptions and break them in various ways.
2. Simulate datasets corresponding to each scenario.
3. Apply the Heckman selection model correction procedure to simulated datasets in each scenario.
4. For each scenario, compare the prevalence estimated using the Heckman selection model procedure with the real prevalence that we baked into the data.
5. Finally, compare how well the procedure does across the scenarios.

3 Simulations

3.1 Creating Realistic Virtual Populations Infected with HIV

We are interested in testing the Heckman selection model correction procedure using data with a structure similar to that of real populations along important dimensions over which we know HIV prevalence changes. To keep things tractable we include sex and age and generate datasets with sex-age-specific HIV prevalence rates that are similar to those found in high HIV prevalence populations living in Southern Africa. We generate simulated data using a model analogous to the data generating process assumed by the Heckman probit method. We create a virtual population with a sex-age composition similar to a real population living in Southern Africa, and we infect the virtual people with HIV so that the sex-age profile of HIV prevalence matches that same real population.

3.1.1 Data Generating Preliminaries, Age & Sex

We start by creating a variable number of virtual people and assign them ages (continuous variable) to match the age structure of the Southern African population shown in Table 1. In an independent process we then assign sexes with 52% female and 48% male.

Table 1. Population Age Structure

Age Group	Fraction of Population (%)
15-19	20.3
20-24	17.2
25-29	14.0
30-34	11.2
35-39	9.0
40-44	7.2
45-49	5.6
50-54	4.4
55-59	3.5
60-64	3.0
65+	4.0

3.1.2 Interviewer Variable

To model the interview and selection process we need a comparatively small number of interviewers to interview our virtual subjects. We want to test well-specified Heckman selection models and also break the assumptions underlying the model. Hence, variation in the 'effectiveness' of the interviewers can affect both the selection and outcome processes, only selection for simulated populations that fulfill the assumptions of the Heckman selection model, and both for populations that break those assumptions.

We randomly assign a variable number of interviewers to our virtual subjects. Each interviewer is assigned an ID I and given a random 'interview effectiveness' I_e drawn from a normal distribution with mean $\mu_{I_e} = 1$ and standard deviation $\sigma_{I_e} = 1$. Variation in this effectiveness is allowed to affect the selection and outcome variables in flexible ways that allow us to explore the ability of the Heckman selection models to accurately correct estimates of HIV prevalence.

3.1.3 Selection Variable

The selection variable is constructed by working a probit model backward, starting with the linear predictor and ending up with a binary variable that indicates selection.

For individual j the value of the linear predictor of selection $\kappa_j^\star$ is

$$\begin{aligned} \kappa_j^\star &= \mathbf{z}_j\boldsymbol{\gamma} + \mu_{sj} \\ \kappa_j^\star &= \omega + \theta_1(age_j - 40)^2 + \theta_2[sex_j = \text{male}] \\ &\quad + \theta_3[I_j = 2] + ... + \theta_{11}[I_j = 10] \\ &\quad + \mu_{sj} \end{aligned} \tag{1}$$

where indicator variables are noted using Iverson's bracket notation: $[\alpha] = \begin{cases} 1 & \text{if } \alpha \text{ is true} \\ 0 & \text{if } \alpha \text{ is false} \end{cases}$

The probability of being selected S_j for individual j is the Normal probability associated with the value of $\kappa_j^\star$

$$\Pr([S_j]|\mathbf{z}_j) = \Phi(\kappa_j^\star) \tag{2}$$

We observe the binary value (0,1) of the selection variable for individual j

$$S_j = \begin{cases} 1 & \text{if } \kappa_j^\star > 0 \\ 0 & \text{if } \kappa_j^\star \leq 0 \end{cases} \tag{3}$$

The values of ω, θ_{1-11}, and μ_s are adjusted to create different selection scenarios. Our model of the selection process includes the possibility of both a sex effect and a parabolic age effect centered at age 40. By adjusting ω and θ_{1-2} we can eliminate and/or adjust these effects.

The reference group for *sex* is 'female'. All of our simulations have ten interviewers, and the reference group for interviewer is $I = 1$.

3.1.4 Outcome Variable

The HIV status variable is constructed in a similar fashion by working another probit model backward. In this case the linear predictor is slightly more complicated because we want the HIV prevalence of the simulated population to reflect something realistic with respect to the sex-age distribution of the population.

We take measured sex-age-specific prevalence of HIV from the population living in Southern Africa and add a small amount of noise. We operationalize sex and age as indicator variables for sex and five-year age groups from 15 to 65+. Using the *real* data we estimate a regular probit model predicting HIV status to generate a set of coefficients for each sex-age category and their two-way interactions that we can use to simulate data that encode a sex-age pattern of HIV prevalence that matches the real population. These coefficients are inserted into Equation 4.

For individual j the value of the linear predictor κ_j of the outcome is

$$\begin{aligned} \kappa_j &= \mathbf{x}_j\boldsymbol{\beta} + \mu_{hj} \\ \kappa_j &= \lambda + \delta_1[sex_j = \text{male}] + \delta_2[age_j = 20-24] + ... + \delta_{11}[age_j = 65+] \\ &\quad + \delta_{12}[sex_j = \text{male} \wedge age_j = 20-24] + ... + \delta_{21}[sex_j = \text{male} \wedge age_j = 65+] \\ &\quad + \delta_{22}[I_j = 2] + ... + \delta_{30}[I_j = 10] \\ &\quad + \mu_{hj} \end{aligned} \tag{4}$$

The probability of being HIV+ H_j for individual j is the Normal probability associated with the value of κ_j

$$\Pr([H_j]|\mathbf{x}_j) = \Phi(\kappa_j) \tag{5}$$

Only when individual j is selected (i.e. when $S_j = 1$), we observe the binary value (0,1) of the outcome

$$H_j = \begin{cases} 1 & \text{if } \kappa_j > 0 \\ 0 & \text{if } \kappa_j \leq 0 \end{cases} \tag{6}$$

For the purpose of generating data, virtual individuals are given an HIV status regardless of their selection status, but when we estimate the Heckman selection models we only acknowledge the HIV status of individuals who *are* selected.

The values of λ and δ_{1-21} (the coefficients that generate the sex-age pattern of HIV prevalence) remain constant across all of our simulations. We adjust the values of δ_{22-30} and μ_h to create and vary an interviewer effect on HIV status.

The reference groups for *sex* and interviewer I are the same as for the selection generating process, and for *age* the reference group is '$15-19$'.

3.1.5 Correlated Errors

To create the type of relationship between selection and HIV status that the Heckman selection model is designed to identify and estimate, we correlate the error terms in the generating equations 1 and 4:

$$\mu_s \sim N(0,1) \tag{7}$$

$$\mu_h \sim N(0,1) \tag{8}$$

$$\text{corr}(\mu_s, \mu_h) = \rho \tag{9}$$

Consequently the probabilities of being selected and being HIV+ can be correlated with either sign and any magnitude – e.g. strong negative correlation implies that those who are selected are less likely to be HIV+ than those who are not, the worrying potential issue that motivates this work!

3.2 Simulation Dataset Specifications

The simulation scenarios are organized as follows:

1. Base: benchmark; all of the Heckman selection model assumptions are met.
2. Outcome scenarios: both the 'Outcome' scenarios break the assumption that the selection variable does not affect the outcome.
3. Non-normal scenarios: both the 'Non-normal' scenarios break the assumption that conditioned on the selection and outcome equations, the joint error distribution is bivariate Normal.
4. Sizes: this scenario varies the sample size of the virtual population.

For each of the first four simulation scenarios – Base, Outcome Strong, Outcome Weak, Non-normal Squared, Non-normal Cubed & Sizes – we generate 500 datasets with 5,000 observations each for each parameter set. For these scenarios we vary the Heckman ρ and the degree of selection through ω. The degree of selection results in various fractions of the virtual populations being selected for observation – i.e. results in situations with varying numbers of unobserved individuals for whom the Heckman selection method must produce a predicted probability of being HIV positive. For the remaining scenario – Sizes – we generate 500 datasets with varying numbers of observations for a fixed intermediate level of selection.

To keep things manageable we vary the Heckman ρ over nine values $-0.8(0.2)0.8$[1] and the level of selection over four values $0.5(1.0)3.5$. This yields 36 combinations of ρ and ω for each the first five simulation scenarios, which in turn requires 18,000 individual datasets for each simulation scenario; 90,000 datasets in total for the first five simulation scenarios. For the last scenario Sizes, there are 20 combinations of ρ and sample size with a fixed level of selection. This creates 10,000 more datasets to bring the grand total to 100,000 datasets for the complete simulation investigation.

3.2.1 Base

The Base simulation scenario corresponds to a high HIV prevalence population that fulfills the assumptions of the Heckman selection model. There is a good selection variable, Interviewer ID, and conditional on the selection and outcome equations, the joint error distribution is bivariate Normal. This is the 'base' scenario because it represents everything working well and is used as the benchmark to which to compare all of the variously pathologic scenarios that follow.

[1] At various points in the text and Table 2 sequences of values are specified using the notation $s(i)e$ where s is the start value, i is the increment value used to generate consecutive values in an additive fashion ($value_{(k+1)} = value_k + i$), and e is the end value – e.g. $4(2)8$ corresponds to the sequence $[4, 6, 8]$.

3.2.2 Outcome Strong

In the Outcome Strong scenario, the selection variable *Interviewer ID* has a strong effect on the outcome variable *HIV status*. This is accomplished by adding terms to Equation 4 with values for coefficients δ_{22-30} of $0.25 \times I_e$. This creates a strong relationship between *Interviewer ID* and the probability of being HIV positive.

3.2.3 Outcome Weak

In the Outcome Weak scenario, the selection variable *Interviewer ID* has a weak effect on the outcome variable *HIV status*. The δ coefficients in Equation 4 have values of $0.05 \times I_e$.

3.2.4 Non-normal Squared

In the Non-normal Squared scenario the conditional joint error distribution is distorted by squaring the error terms in the selection and outcome equations, Equations 7 and 8. This results in a joint distribution that is 'bunched-up' close to the origin for original error term values between -1 and 1 and greatly 'expanded' for original values with absolute value greater than 1. The whole distribution occupies the 1^{st} quadrant of the Cartesian plane and is concentrated near the origin with a long 'tail' toward increasing positive values.

3.2.5 Non-normal Cubed

In the Non-normal Cubed scenario the conditional joint error distribution is distorted by cubing the error terms in the selection and outcome equations, Equations 7 and 8. This results in a joint distribution that is greatly condensed and expanded depending on the exact values of the original error terms, and the result occupies potentially all four quadrants of the Cartesian plane.

3.2.6 Sizes

The Sizes scenario is the same as the Base scenario with varying 'sample sizes', i.e. numbers of virtual individuals. Five sample sizes are tested: 100(2,500)10,100.

3.3 Simulation Scenario Parameter Values

The specific parameter values used to generate datasets for each simulation scenario are listed in Tables 2 and 3.

Table 2. Simulation Selection, Correlation & Observations Specifications

			Selection Affects Outcome		Non-normal Error Distribution		
Coefficient	Variable	Base	Strong	Weak	Squared	Cubed	Sizes
			Selection				
ω		**0.5(1.0)3.5**	**0.5(1.0)3.5**	**0.5(1.0)3.5**	**0.5(1.0)3.5**	**0.5(1.0)3.5**	2.0
θ_1	age	0.008	0.008	0.008	0.008	0.008	0.008
-na-	*sex=female*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*
θ_2	sex=male	-0.300	-0.300	-0.300	-0.300	-0.300	-0.300
-na-	*I=1*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*
θ_3	I=2	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_4	I=3	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_5	I=4	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_6	I=5	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_7	I=6	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_8	I=7	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_9	I=8	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_{10}	I=9	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
θ_{11}	I=10	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$	$2.5 \times I_e$
μ_s'		-na-	-na-	-na-	μ_s^2	μ_s^3	-na-
			Heckman ρ and Number of Observations				
ρ		**-0.8(0.2)0.8**	**-0.8(0.2)0.8**	**-0.8(0.2)0.8**	**-0.8(0.2)0.8**	**-0.8(0.2)0.8**	**-0.8(0.2)0.8**
Observations		5,000	5,000	5,000	5,000	5,000	**100(2,500)10,100**

Table 3. Simulation Output Specifications

			Selection Affects Outcome		Non-normal Error Distribution		
Coefficient	Variable	Base	Strong	Weak	Squared	Cubed	Sizes
			HIV Status (Outcome)				
λ	constant	-1.522	-1.522	-1.522	-1.522	-1.522	-1.522
-na-	*sex=female*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*
δ_1	sex=male	-0.882	-0.882	-0.882	-0.882	-0.882	-0.882
-na-	*age=15-19*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*
δ_2	age=20-24	0.928	0.928	0.928	0.928	0.928	0.928
δ_3	age=25-29	1.220	1.220	1.220	1.220	1.220	1.220
δ_4	age=30-34	1.348	1.348	1.348	1.348	1.348	1.348
δ_5	age=35-39	1.383	1.383	1.383	1.383	1.383	1.383
δ_6	age=40-44	1.109	1.109	1.109	1.109	1.109	1.109
δ_7	age=45-49	1.135	1.135	1.135	1.135	1.135	1.135
δ_8	age=50-54	0.829	0.829	0.829	0.829	0.829	0.829
δ_9	age=55-59	0.848	0.848	0.848	0.848	0.848	0.848
δ_{10}	age=60-64	0.392	0.392	0.392	0.392	0.392	0.392
δ_{11}	age=65+	0.044	0.044	0.044	0.044	0.044	0.044
-na-	*sex=male, age=15-19*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*
δ_{12}	sex=male, age=20-24	-0.086	-0.086	-0.086	-0.086	-0.086	-0.086
δ_{13}	sex=male, age=25-29	0.407	0.407	0.407	0.407	0.407	0.407
δ_{14}	sex=male, age=30-34	0.877	0.877	0.877	0.877	0.877	0.877
δ_{15}	sex=male, age=35-39	0.878	0.878	0.878	0.878	0.878	0.878
δ_{16}	sex=male, age=40-44	1.042	1.042	1.042	1.042	1.042	1.042
δ_{17}	sex=male, age=45-49	0.700	0.700	0.700	0.700	0.700	0.700
δ_{18}	sex=male, age=50-54	1.076	1.076	1.076	1.076	1.076	1.076
δ_{19}	sex=male, age=55-59	1.126	1.126	1.126	1.126	1.126	1.126
δ_{20}	sex=male, age=60-64	1.110	1.110	1.110	1.110	1.110	1.110
δ_{21}	sex=male, age=65+	0.878	0.878	0.878	0.878	0.878	0.878
-na-	*I=1*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*	*-na-*
δ_{22}	I=2	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{23}	I=3	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{24}	I=4	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{25}	I=5	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{26}	I=6	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{27}	I=7	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{28}	I=8	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{29}	I=9	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
δ_{30}	I=10	-na-	$\mathbf{0}.25 \times I_e$	$\mathbf{0}.05 \times I_e$	-na-	-na-	-na-
μ'_h		-na-	-na-	-na-	μ_h^2	μ_h^3	-na-

3.4 Calculation and Comparison of HIV Prevalence Values

3.4.1 Real HIV Prevalence

Calculation of the real HIV prevalence for a virtual population is straightforward. The real HIV prevalence H_p is

$$H_p = \frac{1}{N}\sum_{i=1}^{N}[H_i = 1] \tag{10}$$

where N is the total number of virtual individuals and H_i is the HIV status of individual i.

3.4.2 Corrected HIV Prevalence using the Heckman Selection Model

Calculation of the corrected HIV prevalence for a virtual population consists of two parts, one for selected (observed) individuals and another for individuals who are not selected (observed). For selected individuals the procedure is as above for the calculation of real HIV prevalence. For those who are not selected, we use the Heckman selection model to estimate their probability of being HIV positive, and then summarize that and combine it with the real prevalence among the selected population to arrive at the corrected prevalence for the entire population.

For each of the simulated datasets we estimate the biprobit Heckman selection model specified in Equation 11 for selection and Equation 12 for HIV status using the `heckprob` procedure in the statistical package `Stata` 11.2 with the options `iterate(15)` and `difficult`.

$$\begin{aligned}
\Pr([S_i]|\mathbf{z}_i) &= \Phi(\eta_i^\star) \\
\eta_i^\star &= \Omega + \Theta_1[sex_i = \text{male}] \\
&\quad + \Theta_2[age_i = 20-24] + ... + \Theta_{11}[age_i = 65+] \\
&\quad + \Theta_{12}[I_i = 2] + ... + \Theta_{20}[I_i = 10] \\
&\quad + M_{si}
\end{aligned} \tag{11}$$

$$\begin{aligned}
\Pr([H_i]|\mathbf{x}_i) &= \Phi(\eta_i) \\
\eta_i &= \Lambda + \Delta_1[sex_i = \text{male}] \\
&\quad + \Delta_2[age_i = 20-24] + ... + \Delta_{11}[age_i = 65+] \\
&\quad + \Delta_{12}[sex_i = \text{male} \wedge age_i = 20-24] \\
&\quad + ... + \Delta_{21}[sex_i = \text{male} \wedge age_i = 65+] \\
&\quad + M_{hi}
\end{aligned} \tag{12}$$

The estimated 'Heckman ρ' R is

$$R = \text{corr}(M_s, M_h) \tag{13}$$

If an estimation attempt succeeds (converges on a high likelihood solution), we use `Stata`'s `predict` command with the options `pmarg, p10, p11 and psel` to generate and calculate the predicted values shown in Table 4.

Corrected HIV prevalence H_{pc} is

$$H_{pc} = \frac{1}{N}\left[\sum_{i=1}^{n}[H_i = 1] + \sum_{j=1}^{g} P(H = 1|S = 0)_j\right] \tag{14}$$

where N is the total number of virtual individuals, n is the number of selected (observed) virtual individuals, and g is the number of virtual individuals who are not selected (observed); $N = n + g$. H_i is the HIV status of individual i and $P(H = 1|S = 0)$ is calculated as in Table 4.

Table 4. Predicted Probabilities after `heckprob` Estimation

	Probability	`Stata` Predict Option or Formula
Marginal probability of being HIV+	$P(H=1)$	`pmarg`
Joint probability of being HIV+ and not selected	$P(H=1 \wedge S=0)$	`p10`
Joint probability of being HIV+ and selected	$P(H=1 \wedge S=1)$	`p11`
Probability of being selected	$P(S=1)$	`psel`
Probability of *not* being selected	$P(S=0)$	$1-P(S=1)$
Probability of *not* being selected given HIV+	$P(S=0\|H=1)$	$\frac{P(H=1,S=0)}{P(H=1,S=0)+P(H=1,S=1)}$
$\rightarrow$ *Probability of being HIV+ given* not *selected*	$P(H=1\|S=0)$	$\frac{P(H=1,S=0)}{1-P(S=1)}$

To actually calculate the corrected HIV prevalence we use the following procedure:

1. Create a new variable H_{pc} containing a value of 0 for each virtual individual.
2. For virtual individuals who were observed ($S=1$), replace H_{pc} with the value in their HIV status variable: 0 if they are negative and 1 if positive.
3. For the remaining unobserved virtual individuals, replace H_{pc} with the predicted probability of being HIV positive derived from the Heckman estimation: $P(H=1|S=0)$, calculated as in the bottom row of Table 4.
4. Calculate the corrected prevalence for the population of virtual individuals as the mean value of H_{pc}. The calculation of the mean is weighted correctly because the sum is over each individual in the virtual population.

3.5 Comparison of Real and Corrected HIV Prevalence values

We evaluate the success of the correction with a metric that compares the corrected and real HIV prevalences controlling for the overall level of HIV prevalence, or the 'size' of the epidemic, recognizing that an error that is small in an absolute sense is more problematic when HIV prevalence is low. Our metric is the proportional error E_{prop}

$$E_{prop} = \frac{H_{pc} - H_p}{H_p} \tag{15}$$

4 Results

Figures 1 and 2 display the mean proportional errors for each simulation scenario. The mean is taken over the 500 estimation attempts for each parameter set in each scenario. The plots display the proportional error by values of the Heckman ρ and the percent of individuals who were *not* selected. For some simulated datasets the estimation procedure for the biprobit Heckman selection model was not able to converge. The fractions of the 500 estimation attempts for each parameter set for each scenario that did not converge successfully are shown in Table 5.

4.1 Base

Panel 1 of Figure 1 displays results for the Base simulation scenario in which all of the Heckman selection model assumptions are valid. The magnitude of the proportional error never exceeds 0.4 percent, and there is little variation with the fraction of individuals who were not selected to test. There is some variation with the Heckman ρ such that there is a consistent bias toward slightly underestimating the true value of HIV prevalence when the absolute value of the Heckman ρ is small – roughly -0.5 to 0.5.

There is substantial variation in the fraction of estimation attempts that fail in the Base simulation scenario, see Table 5. The percent estimation failures increases dramatically as the fraction of individuals not selected increases and the Heckman ρ decreases toward -0.8. At the extreme values of these parameters, almost half of the estimation attempts fail – 47.4 percent.

4.2 Outcome Strong

Panel 2 of Figure 1 displays results for the Outcome Strong simulation scenario. This scenario models a strong effect of the selection variable on the outcome. The Heckman selection correction procedure performs *very badly* in this scenario. The magnitude of the proportional error is never less than 5 percent and reaches 20 percent for some parameter sets. As expected, the direction of the error is always positive because this effect is modeled with positive coefficients δ_{22-30} so that the selection variable has a positive effect on the outcome – i.e. increases the probability of being HIV positive. There is a clear deterioration in the correction procedure as the fraction of individuals not selected for testing increases, or as the percent missing goes up. Performance is slightly better for extreme values of the Heckman ρ in all cases.

Most of the estimation attempts succeed for this simulation scenario, see Table 5. The largest fraction of failures occur for highly negative values of the Heckman ρ and extreme values of the percent of individuals not selected for testing.

4.3 Outcome Weak

Panel 3 of Figure 1 displays results for the Outcome Weak simulation scenario. This scenario models a weak effect of the selection variable on the outcome. The results for this scenario are very similar to the Outcome Strong scenario, with magnitudes that are about $1/5^{\text{th}}$ of that in the Outcome Strong scenario. This suggests a linear relationship between the value of the parameters that model this effect and the outcome. Parameters δ_{22-30} in the Outcome Strong scenario are $5\times$ the corresponding values in the Outcome Weak scenario.

Most of the estimation attempts succeed for this simulation scenario, see Table 5. The largest fraction of failures occur for highly negative values of the Heckman ρ and extreme values of the percent of individuals not selected for testing.

4.4 Non-normal Squared

Panel 4 of Figure 1 displays results for the Non-normal Squared simulation scenario. This scenario models a violation of the requirement that the conditional error distribution be bivariate Normal. The correction procedure performs well in spite of the distorted error distribution; the magnitude of the proportional error is never more than 0.8 of a percent. However, in this scenario all of the errors are negative and *underestimate* prevalence. The errors become worse in a symmetric way as the magnitude of the Heckman ρ increases and as the fraction of individuals not selected increases.

Very few estimation attempts fail in this scenario, see Table 5. The largest fractions failing are found when few individuals are not selected for testing and this fraction never exceeds 5.5 percent.

4.5 Non-normal Cubed

Panel 5 of Figure 1 displays results for the Non-normal Cubed simulation scenario. This scenario models a different violation of the requirement that the conditional error distribution be bivariate Normal. The results are similar to the Non-normal Squared scenario but worse. Again the errors are consistently negative, but in this case the magnitude reaches 7 percent in the worst case. The symmetry seen for the Non-normal Squared scenario is not observed for this scenario. The errors are worse all around but much worse for negative values of the Heckman ρ. In contrast, there is less variation in the magnitude of the error as the fraction of individuals not selected for testing changes.

Almost a third of the estimation attempts fail for some parameter sets in this scenario, and there is a strong pattern in the failure profile with Heckman ρ and fraction of individuals not selected to test, see Table 5. The fraction of failures is low when both are moderate and increase rapidly as the Heckman ρ decreases away from 0 and the fraction individuals not selected increases *and* as the Heckman ρ increases from 0 and the fraction of individuals not selected decreases.

4.6 Sizes

Figure 2 displays results for the Sizes simulation scenario. This scenario investigates what happens at the 'sample size' changes. This is accomplished by varying the number of virtual individuals in the simulated datasets. The parameter values used in this scenario are the same as for the Base scenario except the fraction of individuals not selected for testing is fixed at an intermediate value. Not surprisingly, the results are very similar to the Base scenario for all sample sizes except the smallest, 100. When the sample was only 100 the method essentially failed and produced large errors that are strongly related to the value of the Heckman ρ.

The fraction of estimation attempts that fail in this scenario is also very similar to the Base scenario except when the sample size is 100, and in that case virtually all estimation attempts fail, see Table 5.

5 Discussion

This investigation used a simulation approach to test the Heckman selection model correction procedure designed to identify and correct for selection effects in sample surveys used to estimate HIV prevalence. The main results are very clear:

1. When the conditions required by the Heckman selection model are met, the method works very well.
2. When the two important assumptions of the Heckman selection model are violated, even mildly, the method does not work well.
3. A reasonable sample size of magnitude 1,000 is necessary for the procedure to function.

The results from the Base simulation scenario demonstrate how well the procedure functions when the conditional error structure is bivariate Normal and the selection variables do not affect the outcome at all. Results from the Outcome simulation scenarios demonstrate that the method is essentially useless if the selection variable affects the outcome even a little bit. Likewise, the Non-normal scenarios reveal that the shape of the conditional error distribution matters a lot.

Table 5. Estimation Failures, Percent of 500 Attempts

Heckman ρ	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8
Fraction Not Selected					Base				
Lots	47.4	28.6	12.8	3.2	7.0	1.4	0.4	2.8	5.4
Some	32.6	14.2	7.6	1.0	4.4	0.2	0.8	1.2	4.0
A few	13.0	7.8	4.4	0.8	1.4	0.6	1.0	2.0	3.6
Not many	9.4	4.8	2.4	2.0	2.0	1.0	1.0	3.2	8.6
Fraction Not Selected					Outcome Strong				
Lots	15.2	17.6	5.6	2.4	0.6	0.6	0.0	0.0	0.2
Some	6.4	5.6	4.8	2.2	0.4	0.4	0.2	0.0	0.4
A few	5.2	2.4	2.6	1.6	0.2	0.4	0.2	0.4	0.0
Not many	12.8	4.0	3.2	3.8	3.0	2.0	1.6	1.0	0.6
Fraction Not Selected					Outcome Weak				
Lots	40.2	35.0	14.2	2.0	3.4	2.2	0.6	1.8	3.4
Some	23.6	13.4	7.4	1.4	1.6	0.6	0.2	0.2	2.2
A few	11.8	5.2	3.4	0.8	0.6	0.4	0.0	0.4	2.0
Not many	7.4	5.2	3.2	1.8	1.2	0.6	1.0	2.0	5.2
Fraction Not Selected					Non-normal Squared				
Lots	0.8	0.2	0.6	0.2	1.0	0.4	0.2	0.8	0.6
Some	1.2	0.2	0.4	0.4	1.2	1.2	0.2	0.4	1.2
A few	2.4	0.6	0.6	1.0	1.4	0.4	0.2	0.0	2.0
Not many	5.2	1.8	1.2	2.6	2.8	1.8	1.6	0.4	4.2
Fraction Not Selected					Non-normal Cubed				
Lots	16.6	6.6	4.6	1.8	2.0	0.2	0.0	0.8	1.2
Some	6.8	2.2	1.4	0.4	0.8	0.2	0.2	1.2	6.0
A few	4.0	1.2	0.6	0.4	0.0	0.4	1.0	7.0	20.2
Not many	1.2	0.0	0.0	0.0	0.2	1.4	4.8	13.0	28.4
Sample Size					Sizes				
100	94.4	93.2	92.4	93.6	91.8	91.8	94.4	92.6	94.0
2,600	30.0	21.0	12.2	5.4	4.0	3.0	4.8	6.4	11.8
5,100	23.8	11.6	8.2	1.2	3.2	0.4	1.0	1.6	4.4
7,600	15.8	7.8	3.0	0.4	0.8	0.4	0.8	0.4	1.6
10,100	12.4	4.0	1.2	0.6	0.8	0.0	0.0	0.0	0.8

Figure 1. Mean Proportional Error between Corrected and True HIV Prevalence by Heckman ρ and Fraction of Individuals not Selected. Each point is the mean of values from 500 simulations, each with 5,000 virtual individuals. Simulation scenarios numbered 1-5 and named as in text.

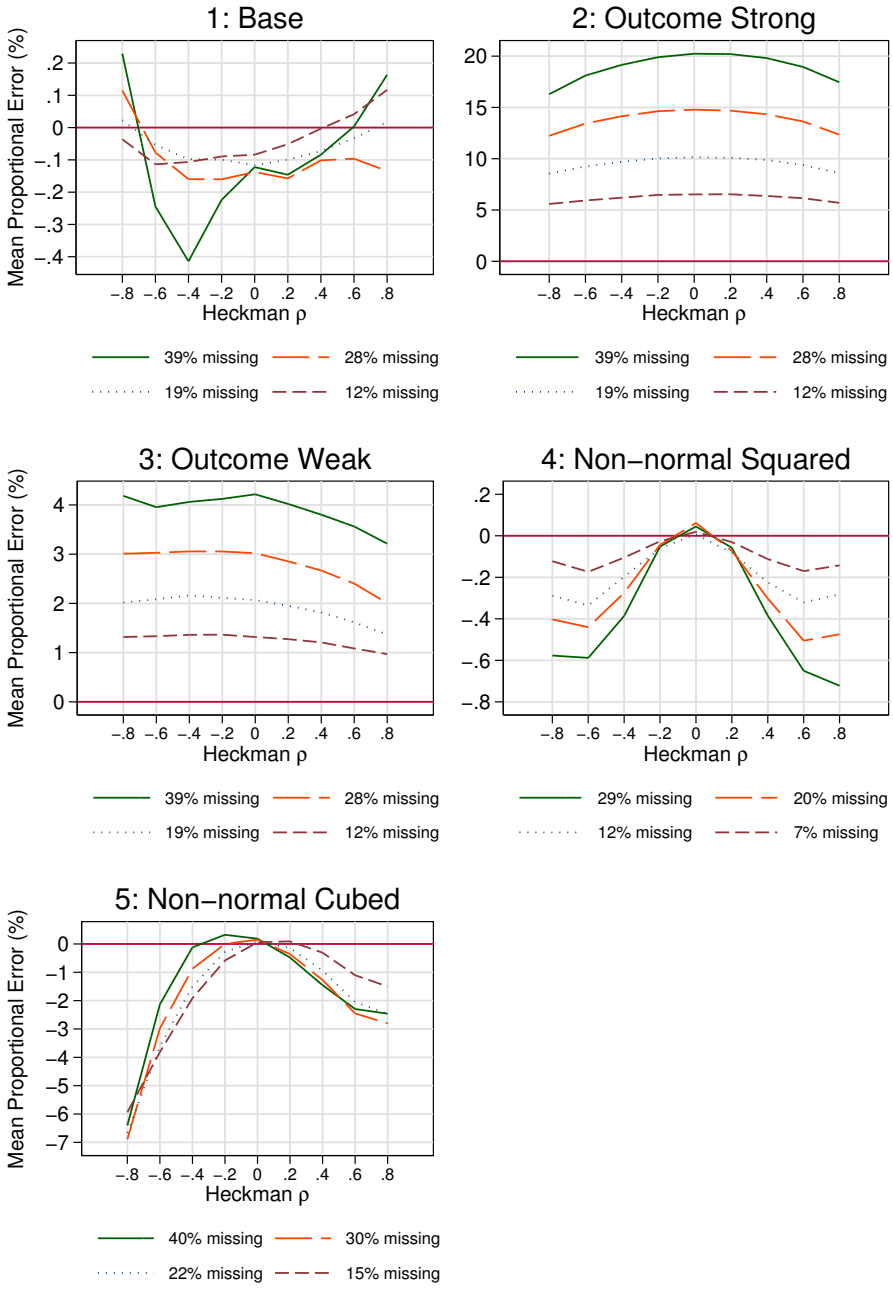
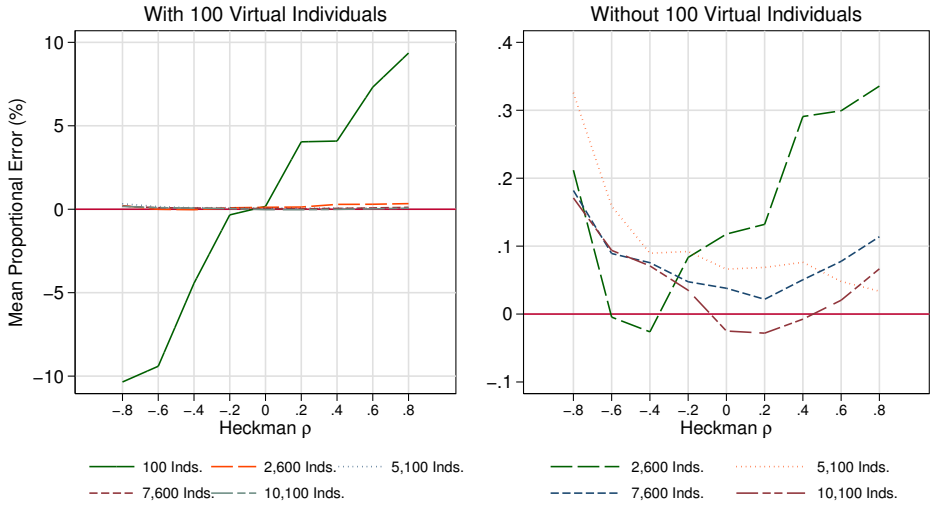


Figure 2. Mean Proportional Error between Corrected and True HIV Prevalence by Heckman ρ and Number of Virtual Individuals in the Sample. Each point is the mean of values from 500 simulations, each with varying number of virtual individuals. This is the SIZES simulation scenario. The left panel 'With 100 Virtual Individuals' displays all of the results; the right panel 'Without 100 Virtual Individuals' removes results from the simulation with 100 virtual individuals to expand the y-axis to reveal detail from the other levels of the sample size variable.



In general it is not possible to definitely test real data to determine if either of these assumptions is valid. The shape of the error distribution is only revealed through the estimation process, and is a sensitive function of the variables included in the selection and outcome equations. If some of the variables affecting selection and outcome are missing, then it will by definition be impossible to properly characterize the *conditional* error distribution, and since no survey will every have all the variables affecting either, it is purely a matter of story telling to ascertain how close to conditionally Normal the error distribution is.

Likewise, it is not possible in a general sense to be sure that selection variables are not related to the outcome, and as the OUTCOME WEAK scenario makes clear, even a weak association is enough to produce systematically incorrect results with large magnitude.

Given these results, our advice is:

1. Only use the Heckman selection model correction technique when 'good' selection variables are available and the sample size is relatively large, ∼1,000s.
2. Interpret the results in a limited sense:
   - If the Heckman selection model converges nicely and produces stable estimates *and* if the estimate of the Heckman ρ is statistically significant and of a reasonable magnitude, **take this as an indication that there may be a problem with selection in your data**.
   - ***Do not*** use the corrected HIV prevalence values as an indication of the true prevalence. Because the assumptions of the Heckman selection model are essentially untestable, there are effectively no circumstances in which you can be sure that they are met, and therefore given the *de facto* high likelihood of being wrong, the relatively large magnitude of the likely error, and the consequences

associated with being wrong, it is not advisable to 'correct' estimates of HIV prevalence with this procedure.