Original Research Article

Mark E. McGovern[1], Till Bärnighausen[2 3], Giampiero Marra[4], Rosalba Radice[5]

On the Assumption of Bivariate Normality in Selection Models A Copula Approach Applied to Estimating HIV Prevalence

Running Head: On the Assumption of Joint Normality in Selection Models

Word Count: 3,874

Conflicts of Interest: None Declared

The Program on the Global Demography of Aging receives funding from the National Institute on Aging, Grant No. 1 P30 AG024409-06.

Abstract

[1] Corresponding Author. Harvard Center for Population and Development Studies.
Email: mcgovern@hsph.harvard.edu. Tel: +1 857-600-8879.
Address: 9 Bow Street, Cambridge, MA 02138, USA.
[2] Department of Global Health, Harvard School of Public Health
[3] Africa Centre for Health and Population Studies, University of KwaZulu-Natal
[4] Department of Statistical Science, University College London
[5] Department of Economics, Mathematics and Statistics, Birkbeck

Background
Heckman-type selection models are potentially applicable in many contexts in epidemiology, particularly where the assumption of missing at random is not realistic. This approach has been applied to estimating HIV prevalence from nationally representative household surveys where rates of refusal to test are often high. A drawback of existing methods to control for selection on unobserved factors is that they typically rely on strong parametric assumptions.

Methods
We introduce a novel approach for relaxing joint normality in selection models. We apply this method to estimating HIV prevalence in the 2007 Zambian Demographic and Health Survey where 21% of men and 20% of women refuse to test, and using interviewer identity as the selection variable which predicts consent to test but not HIV status, we show how to allow for non-linear association between the participation and outcome equations using copula functions.

Results
HIV prevalence estimates are similar irrespective of the structure of the association between consenting to test and HIV status. For men, our estimation indicates a population HIV prevalence of 21%, compared to 12% among those who consent to test. For women, the corresponding figures are 20% and 16%.

Conclusions
Existing results indicating the presence of selection bias in the estimation of HIV prevalence for men and women in Zambia are robust to relaxing the assumption of joint normality. As misspecification results in inconsistent estimates, future research involving selection models to account for missing data should routinely conduct sensitivity analyses for alternative functional forms using this approach.

Abstract Word Count: 248

Introduction

Missing data is a common problem in epidemiological studies, and the mechanisms through which this missingness occurs can have an important impact on resulting estimates (Hernan et al., 2004). Therefore, in general the treatment of missing values requires careful consideration in order to minimise the potential for selection bias to affect results. One area which is particularly affected by this issue is the field of HIV research, due to the extent of attrition and missing information, and concerns surrounding the fact that individuals must actively choose whether to participate in HIV testing in order to be present in surveys. HIV status is more likely to be associated with social stigma and desire for confidentiality than other more routine parts of questionnaires, or even other biomarker data collection (e.g. Hosseinzadeh et al., 2012). Methods accounting for selection which are robust to the assumption of missing at random, such as Heckman-type estimators, are therefore highly suited to this context; however they typically require a strong set of assumptions. We introduce a novel methodology for improving the practical implementation of this approach, and demonstrate the methodology by estimating HIV prevalence whilst accounting for missing data.

This is an important policy-relevant and instructive application, as despite being the "gold standard" source of data for HIV prevalence estimation (Boerma et al., 2003), nationally representative household surveys commonly suffer from high rates of refusal to participate in HIV testing. If HIV prevalence among respondents who refuse to test differs from respondents who take the test, estimates solely based on the former will be biased. Recent research suggests that respondents may refuse to test if they have knowledge of their HIV status (Floyd et al., 2013; Bärnighausen et al., 2012; Reiners and Eaton, 2009). This has important implications for complete case analysis (i.e., only using information on individuals without missing data) and imputation models, which require that data are missing at random. Because HIV status is not observed among those who refuse to test, neither of these approaches is robust to systematic selection effects on unobserved factors (Donders et al., 2006). Rates of refusal to test for HIV can be substantial; e.g. up to 37% in the Demographic and Health Surveys (Hogan et al., 2012). Similar levels of refusal to participate in HIV testing can also occur in research. In a recent review of RCTs with an HIV outcome, Harel et al. (2012) found 26% missing HIV status data on average.

One potential solution to this problem is the adoption of Heckman-type selection models which can provide consistent estimates of the parameter of interest, even when missing data are systematically related to some unobserved characteristic of the individual (Heckman, 1979); Vella, 1998), such as HIV status itself. Due to their robustness to selection on unobserveables, these models have a potentially wide set of applications in epidemiology, especially where the untestable assumption of missing at random is unlikely to hold. However, their use in practice is affected by the fact that the implementation of this approach typically depends on two key assumptions. The first is the existence of an appropriate exclusion restriction or selection variable; a variable which predicts participation but not the outcome. Elements of survey design and implementation are often present in datasets in epidemiology, and are potential candidates if they are plausibly uncorrelated with the characteristics of the individual (Bärnighausen et al., 2011b). For example, in the case of HIV prevalence estimation, interviewer identity represents a plausible candidate for a variable which predicts consent to test but not HIV status. Previous research which has adopted this methodology has found evidence for selection bias in some contexts (Bärnighausen et al., 2011a; Hogan et al., 2012; McGovern et al., 2013; Clark and Houle, 2012a; Reniers et al., 2009), which is in contrast to results obtained from imputation, where the results are almost always very close to the complete case analysis (Hogan et al., 2012; Mishra et al., 2008). This is not a surprising finding if selection is

mainly taking place on unobserved characteristics. In addition, both the original formulation (Heckman, 1979), and previous literature in this context have relied on relatively strong parametric assumptions for identification. While the assumption of joint normality for characterising the relationship between consenting to test and HIV status is convenient and tractable, it is a serious limitation (Puhani, 2000). Arpino et al. (2013) note the importance of parametric assumptions in implementing the Heckman (1979) approach in the specific context of HIV prevalence estimation, and highlight this as an important drawback of this method. Results from selection models may not be robust to the particular choice of distribution, and therefore it is important to be able to evaluate the sensitivity of conclusions from this approach to alternative assumptions.

If both these conditions are met, the conventional bivariate probit estimated by maximum likelihood is consistent and asymptotically efficient. However, if the true distribution of the error terms does not meet the assumption of joint normality, results are likely to be inconsistent (De Luca, 2008). Simulation studies have indicated that HIV prevalence estimates from selection models may be sensitive to violations of this assumption (Clark and Houle, 2012b), however to date there is little evidence in practice, despite the growing literature on the use of Heckman selection models in epidemiological research. While Hogan et al. (2012) use a semi-nonparametric selection model based on Hermite polynomial expansions (De Luca, 2008; Gallant and Nychka 1987), the intercept is not identified in their model and so they do not estimate HIV prevalence *per se*.

As outlined in Geneletti et al. (2011), it is particularly important to evaluate the robustness of results obtained from surveys involving missing data due to the fact that we never observe the true HIV status of those who refuse consent. Therefore, the underlying assumptions in the analytic model are generally not possible to test, and the implementation of selection models can therefore be viewed as a sensitivity analysis to adjust for potential bias using alternative sets of assumptions about the underlying mechanisms causing data to be absent. If it can be demonstrated that the results from the particular method adopted are invariant to a variety of different assumptions, this lends credibility to the conclusions, and indicates that the extent of bias adjustment required is not just a function of the model imposed by the researcher. The lack of a flexible and practical method for evaluating the robustness of selection models to parametric assumptions is likely an important impediment to wider use of this approach.

This aim of this paper is to describe and illustrate a means of determining the sensitivity of results from selection models to alternative ways of characterising the functional form of the association between outcome and participation equations. We introduce and demonstrate a methodology for relaxing the assumption of joint normality in Heckman models that allows for non-linear association between participation (here, HIV testing) and the outcome of interest (here, HIV status). We show how copula functions can be used to define the dependence of the selection process by adapting the method of Marra and Radice (2013b) to a sample selection model context. We evaluate the robustness of estimates of HIV prevalence in Zambia which have indicated the presence of selection bias in previous research. Given the potential applicability of this approach to other contexts, we provide the computer code for this method in order to make this approach easily accessible to researchers working with surveys containing missing data. **Methods**

We begin by modelling consent for HIV testing in the context of a bivariate probit with two latent variables. Both consent to test and HIV status are considered simultaneously, an approach based on

the adaptation of the original Heckman estimator (Heckman, 1979) for binary outcomes by Dubin and Rivers (1989). For a survey of the literature, see Vella (1998).

Consent to test is given by:

$$Consent_i^* = X_i\beta + Z_i\alpha + u_i, i = 1, \dots n \quad (\mathbf{1})$$

$$Consent_i = 1 \; if \; Consent_i^* > 0, Consent_i = 0 \; otherwise \quad (\mathbf{2})$$

The observed consent for person i is the observed outcome arising from a latent variable $Consent_i^*$, measuring the respondent's propensity to test. $Consent_i$ is a dummy variable indicating acceptance to test, while X_i is a $p \times 1$ vector representing observed individual level characteristics with associated parameter vector β, Z_i is a $k \times 1$ vector of dummy variables representing the interviewer identity (the selection variable or exclusion restriction) with associated parameter vector α, and u_i is a random error term. Although in theory identification can be achieved through non-linearity, in practice the performance of selection models requires at least one selection variable to be present in the participation equation but not the outcome equation (Madden, 2008). In this case interviewer identity predicts consent to test but is assumed not to enter into the HIV equation directly.

The equation for the HIV status HIV_i of individual i is:

$$HIV_i^* = X_i\gamma + \varepsilon_i \quad (\mathbf{3})$$

$$HIV_i = 1 \; if \; HIV_i^* > 0, \;\; HIV_i = 0 \; otherwise \quad (\mathbf{4})$$

$$HIV_i \; observed \; only \; if \; Consent_i = 1, missing \; otherwise \quad (\mathbf{5})$$

where γ is a parameter vector and ε_i is a random error term. The structural assumption used in each of the studies which adopt selection models to estimate HIV prevalence (listed above) is that the error terms in both equations (u_i, ε_i) are normally distributed with means equal to zero, variances equal to one and correlation coefficient ρ; that is the joint distribution of (u_i, ε_i) is given by $F_2(u_i, \varepsilon_i) = \Phi_2(u_i, \varepsilon_i)$ where Φ_2 is the standardized bivariate normal cumulative distribution function (cdf). This model can be fitted using classic maximum likelihood.

In order to allow for non-linear associations between the consent and HIV status equations, we model the dependency of the error terms in the two equations using copulas. These are functions that connect multivariate distributions to their one dimensional margins, such that if F is a two-dimensional cdf with one-dimensional margins $(F_1(y_1), F_2(y_2))$, then there exists a two-dimensional copula C such that $F(y_1, y_2) = C(F_1(y_1), F_2(y_2); \theta)$, where y_1 and y_2 (in our case $Consent$ and HIV) are two random variables and θ is an association parameter measuring the dependence between the two marginals (e.g. Trivedi and Zimmer, 2007). A substantial advantage of the copula approach is that the marginal distributions may come from different families. This construction allows researchers to consider marginal distributions and the dependence between them as two separate but related issues. If the theoretical rationale for selection bias in this context is correct (namely that HIV positive individuals are refusing to test on the basis of knowledge of their HIV status), we would expect a value of ρ which is less than 0. In addition, results from the standard Heckman selection model also indicate the presence of negative correlation between testing and HIV status. Therefore, we consider the copulas which allow for at least some negative association.

These are the: (Gaussian (C_g), which is equivalent to the standard bivariate normal probit model; Frank (C_f); 90 and 270 degrees rotated Clayton ($C_{c_{90}}, C_{c_{270}}$); and Student-t (C_t)). Each of these copula functions are reported in Table 1**,** and also illustrated in Figure 1. While the Gaussian, Frank and Student-t copulas are symmetric, we also consider the use of the rotated Clayton copulas which allow for stronger negative dependence in the tails of the distribution. The 90 and 270 degrees rotated versions can be obtained using (e.g., Brechmann and Schepsmeier, 2013):

$$C_{90} = F_2(y_2) - C(1 - F_1(y_1), F_2(y_2); \theta)$$

$$C_{270} = F_1(y_1) - C(F_1(y_1), 1 - F_2(y_2); \theta)$$

These forms of dependence are particularly applicable in the context of HIV prevalence estimation as we might expect respondents with a strong negative score on the latent test variable to be of particularly high risk of being HIV positive. For example, this would be the case if respondents were refusing to test largely on the basis of their HIV status. Other copula functions which allow for asymmetric negative dependence in the tails of the distribution, such as 90 and 270 degrees rotated Gumbel and Joe copulas, could be employed. We did not employ these versions because they capture the tail dependence in a similar way to the Clayton, hence producing very similar estimates (e.g. Marra and Radice, 2013b).

[TABLE 1 HERE]

[FIGURE 1 HERE]

In the current sample selection context, the data identify the three possible events ($Consent_i = 1, HIV_i = 1$), ($Consent_i = 1, HIV_i = 0$) and ($Consent_i = 0$), with probabilities

$$P(Consent_i = 1, HIV_i = 1) = p_{11i} = C(\Phi(\beta + Z_i\alpha), \Phi(X_i\gamma); \theta)$$

$$P(Consent_i = 1, HIV_i = 0) = p_{01i} = \Phi(X_i\beta + Z_i\alpha) - p_{11i}$$

$$P(Consent_i = 0) = p_{0i} = 1 - \Phi(X_i\beta + Z_i\alpha)$$

where Φ is the cumulative distribution function of a standardized normal.

The log-likelihood function is therefore

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} Consent_i \times HIV_i \log(p_{11i}) + Consent_i \times (1 - HIV_i) \log(p_{01i}) + (1 - Consent_i) \log(p_{0i}) \quad (\mathbf{11})$$

where $\delta^{\mathrm{T}} = \left(\beta^{\mathrm{T}}, \alpha^{\mathrm{T}}, \gamma^{\mathrm{T}}, \theta\right)$.

Model (**1-5)** based on the joint normality assumption of the error terms is fitted by maximization of (11), employing a trust region algorithm which uses the analytical gradient and Hessian of the model (Marra and Radice, 2013a). The implementation used here proved to be more stable than the standard approaches (e.g., Newton-Raphson) adopted in the literature to estimate likelihood-based models.

We assess the degree of association between the consent and HIV status equations using a non-parametric measure of rank (Kendall's Tau, τ), which is more appropriate in the context of copulas than the correlation coefficient (ρ) as the dependence modelled by copulas is typically non-linear. τ can be interpreted in the same manner as ρ in the sense that it ranges between -1 and +1, therefore if individuals who refuse to test are more likely to be HIV positive, we would expect to see a value of τ which is less than 0. As the copula models are estimated in a maximum likelihood framework, we evaluate model fit using information criteria (specifically, the Bayesian Information Criteria, BIC).

We use data from the Zambian Demographic and Health Surveys from 2007 (which are publically accessible from http://www.measuredhs.com). We adopt the same explanatory variables and specification as Hogan et al (2012), the code for which is freely available online from http://hdl.handle.net/1902.1/17657. As outlined in model (**1**), interviewer identity enters into the consent equation as a series of dummy variables, one for each interviewer. As some interviewer fixed effects are collinear with other variables in the model (there is some matching of interviewers on gender, region and language in the Demographic and Health Surveys), interviewers with less than 50 interviewees or those with interviewer effects which are collinear are combined into a single category in order to achieve convergence. We focus on estimating selection models for individuals who refused to consent to test, as opposed to respondents who have missing HIV data due to non-contact, as there are relatively few of these individuals compared to those who refuse, and Bärnighausen et al. (2011a) find that their inclusion in the model has little impact on HIV prevalence estimates. Although the focus on respondents who refuse is sufficient for demonstrating the methodology we propose, it could also be applied to respondents who were not contacted. Table 2 illustrates the composition of the analysis sample for men and women separately; we stratify all analyses by sex. Excluding non-contacts, of the eligible 6,416 men, 1,318 (21%) declined to take a HIV test. Of the eligible 7,025 women in the survey, 1400 (20%) declined to take a HIV test. Table 2 also illustrates the HIV prevalence estimate based on the complete case analysis (i.e. only those respondents with a valid HIV test), which is estimated to be 13% for men and 17% for women.

All our estimates of HIV prevalence are weighted and take account of complex survey design. Statistical analyses were performed in the R environment version 3.01 (R Foundation for Statistical Computing, Vienna, Austria), using the package SemiParBIVProbit (Marra and Radice, 2013c) which implements the copula maximum likelihood approach to fit model (**1-5**).

[TABLE 2 HERE]

Results

Table 3 presents estimates for the rank association between consenting to test and HIV status (Kendall's Tau) for each of the four copula models employed, along with the corresponding 95% confidence intervals, which account for clustering at the level of the Demographic and Health Survey cluster. A measure of model fit is also presented in the final column of table 3 (the Bayesian Information Criteria). Although this measure is not adjusted for clustering, this is unlikely to affect the preferred ordering of the models (Dziak and Li, 2006). For men, there is support for the hypothesis of selection bias, with a negative association for each of the copula models, with the confidence interval for τ excluding zero in each case. The τ of -.53 for the normal model corresponds to a ρ (correlation coefficient) of -.73. On the basis of BIC, the model with the best fit is the C_{c270}.

For women, the measure of association between testing and HIV status is also negative, although the association is less strong than for men, with the 95% confidence intervals in most models including zero. The τ of -.19 in the normal model corresponds to a ρ of -.3. On the basis of BIC, the preferred copula specification for women is C_t.

Table 4 gives the corresponding HIV prevalence estimates. The point estimates for all copula models for men are similar, ranging from 20-22%, with the preferred model (C.270 copula) indicating a population HIV prevalence of 21% (with a corresponding confidence interval of 20%-22%). This is in contrast to an estimate of 13% based only on those with a valid HIV test (table 2).

As with men, the HIV prevalence estimates for women are not sensitive to the choice of functional form for describing the marginal distributions, with HIV prevalence estimates between 18% and 20%. The result for the preferred copula model (C_t) is 20% (with a confidence interval of 19%-21%). The population HIV prevalence estimated using women with a valid HIV test is 17% (table 2).

[TABLE 3 HERE]

[TABLE 4 HERE]

Discussion

Heckman-type selection models are potentially attractive for application in a wide variety of contexts in epidemiology, due to the fact that they allow for the recovery of consistent estimates even when data are not missing at random, such as when respondents systematically select out of HIV testing on the basis of knowledge of HIV status. However, their practical use has been limited by the strong assumptions required for their implementation (Puhani, 2000). This paper outlines a novel means of relaxing the commonly used parametric assumptions, and illustrates the approach using household surveys which incorporate HIV testing in their data collection.

Our method provides estimates of HIV prevalence which account for selection bias, but which do not rely on the assumption of joint normality for identification. In the specific context of the empirical application presented as an illustration of the methodology, this paper demonstrates the robustness of previous findings, and enhances the credibility of conclusions from selection models by demonstrating that identification does not rely on a specific functional form for HIV prevalence estimation in Zambia.

By demonstrating that existing results indicating the presence of selection bias in HIV prevalence estimation are robust to alternative assumptions regarding the association between testing and HIV status, this paper illustrates the value of Heckman-type selection models, particularly in relation to the potential alternative means of dealing with missing data. For example, imputation models have been found to produce results which are almost identical to the complete case analysis of respondents who have a valid HIV test (Hogan, 2012; Mishra et al., 2008; Zaidi et al., 2013) Therefore, this paper provides further evidence that the requirements for using imputation models (i.e. that the data are missing at random, Donders et al., 2006) may be unrealistic in the context of HIV prevalence estimation in household surveys where non-response is often substantial. We reiterate the finding that imputation models and complete case analysis cannot provide unbiased estimates when respondents select into testing on the basis of some unobserved characteristic (such as HIV status).

However, our main contribution is that we introduce a flexible and practical method for relaxing the structural assumption generally adopted in these models, which can be easily applied in a variety of contexts with missing data where selection models are potentially relevant, and will be particularly applicable if the missingness has a high probability of being non-ignorable. By weakening the parametric assumptions required to implement these models, we believe this makes the selection methodology even more viable as alternative to the assumption of missing at random, which is also strong and generally untestable. This method can be used to evaluate the sensitivity of results from selection models to alternative assumptions, which is important for allowing the researcher to draw conclusions about whether bias adjustment is required that are not dependent on a specific set of assumptions (Geneletti et al., 2011). Although we find that results are unaffected in our application, this is unlikely to be the case generally.

With this in mind, the methodology we outline is easily implemented in standard statistical software due to the SemiParBIVProbit package, which is publically available for the R environment (http://cran.r-project.org/web/packages/SemiParBIVProbit), and we provide the code for all the analysis discussed in this paper. Research based on selection models should routinely provide an investigation of the sensitivity of results to relaxation of the bivariate normality assumption due to the potential bias associated with incorrect specification of functional form (De Luca, 2008), not only in the specific context of HIV prevalence estimation, but also in other empirical applications which deal with the treatment of missing data. This is easily achieved with the approach we develop in this paper. An additional advantage of our approach is that it provides a means of identifying the most appropriate model in terms of information criteria.

There are a number of important avenues for future research. Further analysis should focus on establishing the validity of the other main assumption underlying the estimation of HIV prevalence in the presence of non-response, namely the exclusion restriction or selection variable, i.e. whether interviewers are related to the HIV status of respondents. While it is plausible that interviewer identity is a function of survey design, and not related to individual level characteristics, this is difficult to prove conclusively. As we never observe the HIV status of respondents who refuse to test, future research should aim to establish whether estimates based on selection models can be supported with objective external data, such as mortality records (Nyirenda et al., 2010), or RCTs where interviewers or incentives are allocated at random and can therefore be used as exclusion restrictions which are known not to affect HIV status.

There are also a number of other methodological issues to be addressed with selection models for estimating HIV prevalence. The use of interviewer fixed effects requires the pooling of interviewers who conduct few interviews, prevents the use of bootstrap standard errors and confidence intervals (Chiburis et al., 2012), and can result in convergence problems with certain models (Clarke and Houle, 2012b). On-going work aims to establish an alternative means of introducing interviewer identity into selection models (McGovern et al., 2013).

Finally, given the increasing focus on treatment-as-prevention in HIV research and policy, it is likely that the coverage and frequency of HIV testing will need to increase. Therefore, the issue of non-response bias will become increasingly important, especially if refusal to test is systematically related to prior testing or knowledge of HIV status. Moreover, knowledge of HIV status, and therefore the potential for selection bias in prevalence estimation, is likely to increase with the roll-

out of programmes focusing on treatment-as-prevention (Korenromp et al., 2013). Methods which allow for adjustment of results for selection bias are also likely to become increasingly important in this context. The development of appropriate methodologies to enable the researcher to make as few assumptions as possible when implementing the model of interest, and testing whether the conclusions are robust to alternatives is an important aim. The exposition of the use of copula functions in this context is one advance in that direction.

References

Arpino, B., Cao, E. D. and Peracchi, F. (2013). "Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random." Journal of the Royal Statistical Society: Series A (Statistics in Society), Forthcoming, doi: 10.1111/rssa.12027

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011a). Correcting HIV prevalence estimates for survey nonparticipation using Heckman-type selection models. *Epidemiology, 22*(1), 27-35.

Bärnighausen, T., Bor, J., Wandira-Kazibwe, S., & Canning, D. (2011b). Interviewer identity as exclusion restriction in epidemiology. *Epidemiology, 22*(3), 446.

Bärnighausen, T., Tanser, F., Malaza, A., Herbst, K., & Newell, M. L. (2012). HIV status and participation in HIV surveillance in the era of antiretroviral treatment: a study of linked population-based and clinical data in rural South Africa. *Tropical Medicine & International Health, 17*(8), e103-e110.

Boerma, J. T., Ghys, P. D., & Walker, N. (2003). Estimates of HIV-1 prevalence from national population-based surveys as a new gold standard. *Lancet, 362*(9399), 1929-1931.

Brechmann, E. C., & Schepsmeier, U. (2012). Modeling dependence with C-and D-

vine copulas: The R-package CDVine. *Journal of Statistical Software, 52*, 1-27.
Clark, S. J., & Houle, B. (2012a). An Application of the Biprobit Heckman Selection Model to Correct Estimates of HIV Prevalence from Sample Surveys at the Agincourt HDSS in South Africa. *Center for Statistics and the Social Sciences Working Paper No. 119, University of Washington*.
Clark, S. J., & Houle, B. (2012b). Evaluation of Heckman Selection Model Method for Correcting Estimates of HIV Prevalence from Sample Surveys via Realistic Simulation. *Center for Statistics and the Social Sciences Working Paper No. 120, University of Washington*.
De Luca, G. (2008). SNP and SML estimation of univariate and bivariate binary-choice models. *Stata Journal, 8*(2), 190.
Dizak, J., and Li, R. (2006). Variable Selection with Penalized Generalized Estimating Equations, The Methodology Center, Pennsylvania State University, Technical Report 06-78
Donders, A. R. T., van der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087-1091.
Dubin, J. A., & Rivers, D. (1989). Selection bias in linear regression, logit and probit models. *Sociological Methods & Research, 18*(2-3), 360-390.
Floyd, S., Molesworth, A., Dube, A., Crampin, A. C., Houben, R., Chihana, M., Price, A., Kayuni, N., Saul, J., French, N., Glynn, J. (2013). Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. *AIDS, 27*(2), 233-242.
Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the econometric society*, 363-390.
Geneletti, S., Mason, A., & Best, N. (2011). Adjusting for selection effects in epidemiologic studies: why sensitivity analysis is the only “solution”. *Epidemiology, 22*(1), 36-39.

Harel, O., Pellowski, J., & Kalichman, S. (2012). Are We Missing the Importance of Missing Values in HIV Prevention Randomized Clinical Trials? Review and Recommendations. *AIDS and Behavior, 16*(6), 1382-1393.
Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.
Hernán, M. A., S. Hernandez-Diaz, Robbins, J.A., (2004). "A structural approach to selection bias." Epidemiology **15**(5): 615-625.
Hogan, D. R., Salomon, J. A., Canning, D., Hammitt, J. K., Zaslavsky, A. M., & Bärnighausen, T. (2012). National HIV prevalence estimates for sub-Saharan Africa: controlling selection bias with Heckman-type selection models. *Sexually transmitted infections, 88*(Suppl 2), i17-i23.
Hosseinzadeh, H., Hossain, S. Z., & Bazargan-Hejazi, S (2012). "Perceived stigma and social risk of HIV testing and disclosure among Iranian-Australians living in the Sydney metropolitan area." Sexual Health **9**(2): 171-177.
Korenromp, E. L., Gouws, E., & Barrere, B. (2013). HIV prevalence measurement in household surveys: is awareness of HIV status complicating the gold standard? *AIDS, 27*(2), 285-287.
Madden, D. (2008). Sample selection versus two-part models revisited: the case of female smoking and drinking. *Journal of Health Economics, 27*(2), 300-307.

Marra, G., & Radice, R. (2013a). A penalized likelihood estimation approach to semiparametric sample selection binary response modeling. *Electronic Journal of Statistics, 7*, 1432-1455.

Marra, G., & Radice, R. (2013b). Copula Regression Spline Models for Binary Outcomes With Application in Health Care Utilization*. University College London Research Report, 321*.

Marra, G., & Radice, R. (2013c).. SemiParBIVProbit: Semiparametric Bivariate Probit Modelling. R package version 3.2-8

McGovern, M. E., Bärnighausen, T., Salomon, J. A., & Canning, D. (2013). Using Interviewer Random Effects to Calculate Unbiased HIV Prevalence Estimates in the Presence of Non-Response: a Bayesian Approach. *PGDA Working Paper 101*.

Mishra, V., Barrere, B., Hong, R., & Khan, S. (2008). Evaluation of bias in HIV seroprevalence estimates from national household surveys. *Sexually transmitted infections, 84*(Suppl 1), i63-i70.

Nyirenda, M., Zaba, B., Bärnighausen, T., Hosegood, V., & Newell, M.-L. (2010). Adjusting HIV prevalence for survey non-response using mortality rates: an application of the method using surveillance data from rural South Africa. *PloS one, 5*(8), e12370.

Puhani, P. (2000). "The Heckman Correction for Sample Selection and Its Critique." Journal of Economic Surveys **14**(1): 53-68.

Reniers, G., Araya, T., Berhane, Y., Davey, G., & Sanders, E. J. (2009). Implications of the HIV testing protocol for refusal bias in seroprevalence surveys. *BMC Public Health, 9*(1), 163.

Reniers, G., & Eaton, J. (2009). Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. *AIDS, 23*(5), 621.

Trivedi, P. K., & Zimmer, D. M. (2005). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics, 1*(1), 1-111.

Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources*, 127-169.

Winkelmann, R. (2012). Copula Bivariate Probit Models: With an Application to Medical Expenditures. *Health Economics, 21*(12), 1444-1455.

Zaidi, J., Grapsa, E., Tanser, F., Newell, M.-L., & Bärnighausen, T. (2013). Dramatic increases in HIV prevalence after scale-up of antiretroviral treatment: a longitudinal population-based HIV surveillance study in rural kwazulu-natal. *AIDS, forthcoming.*

Tables

Table 1 Definition of Copula Functions

Copula	$C(F_1(y_1), F_2(y_2); \theta)$
$Gaussian$: C_g	$\Phi_2(\Phi^{-1}(F_1), \Phi^{-1}(F_2); \theta)$
$Frank$: C_f	$-\theta^{-1}\ln(1 + \frac{(e^{-\theta F_1}-1)(e^{-\theta F_2}-1)}{(e^{-\theta}-1)})$
$Clayton$: C_c	$({F_1}^{-\theta} + {F_2}^{-\theta} - 1)^{-1/\theta}$
$Student$: C_t	$\mathrm{t}_{2v}(\mathrm{t}_v^{-1}(F_1), \mathrm{t}_v^{-1}(F_2); \theta)$

Note to table 1: $t_{2v}(.,.;\theta)$ denotes the cdf of a standard bivariate Student-t distribution with correlation coefficient θ and v degrees of freedom. ${t_v}^{-1}$ denotes the inverse univariate Student-t distribution function with v degrees of freedom.

Table 2 Summary Statistics for Zambia DHS 2007

	Men			
	%		N	%
HIV Prevalence	12	Consented to HIV Test	5,098	79
95% CI LL	11	Refused HIV Test	1,318	21
95% CI UL	13	Total	6,416	100

	Women			
	%		N	%
HIV Prevalence	16	Consented to HIV Test	5,625	80
95% CI LL	15	Refused HIV Test	1,400	20
95% CI UL	17	Total	7,025	100

Note to table 2: HIV prevalence estimates are based on analysis of respondents who have a valid HIV test and are adjusted for complex survey design. Non-contacts are excluded.

Table 3 Measures of Association between HIV Testing and HIV Status (Men and Women in Zambia 2007)

	Men				Women			
	Kendall's Tau	95%CI LL	95%CI UL	BIC	Kendall's Tau	95%CI LL	95%CI UL	BIC
Normal	-0.53	-0.77	-0.12	10,667.20	-0.19	-0.48	0.12	12,327.57
Frank	-0.58	-0.73	-0.22	10,662.66	-0.17	-0.43	0.17	12,327.83
Student-t	-0.53	-0.79	-0.07	10,669.87	-0.19	-0.51	0.18	12,328.66
Clayton 90	-0.31	-0.77	-0.04	10,671.94	-0.13	-0.58	-0.01	12,327.94
Clayton 270	-0.71	-0.83	-0.56	10,661.00	-0.27	-0.74	-0.04	12,327.57

Note to table 3: Estimates are presented for selection models based on the maximisation of model (**11**), and the copula functions defined in table 1. The exclusion restriction is a series of fixed effects for interviewer identity. Additional control variables include urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behaviour 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 as per Hogan et al. (2012). Non-contacts are excluded. Confidence intervals are adjusted for clustering at the level of the Demographic and Health Survey cluster.

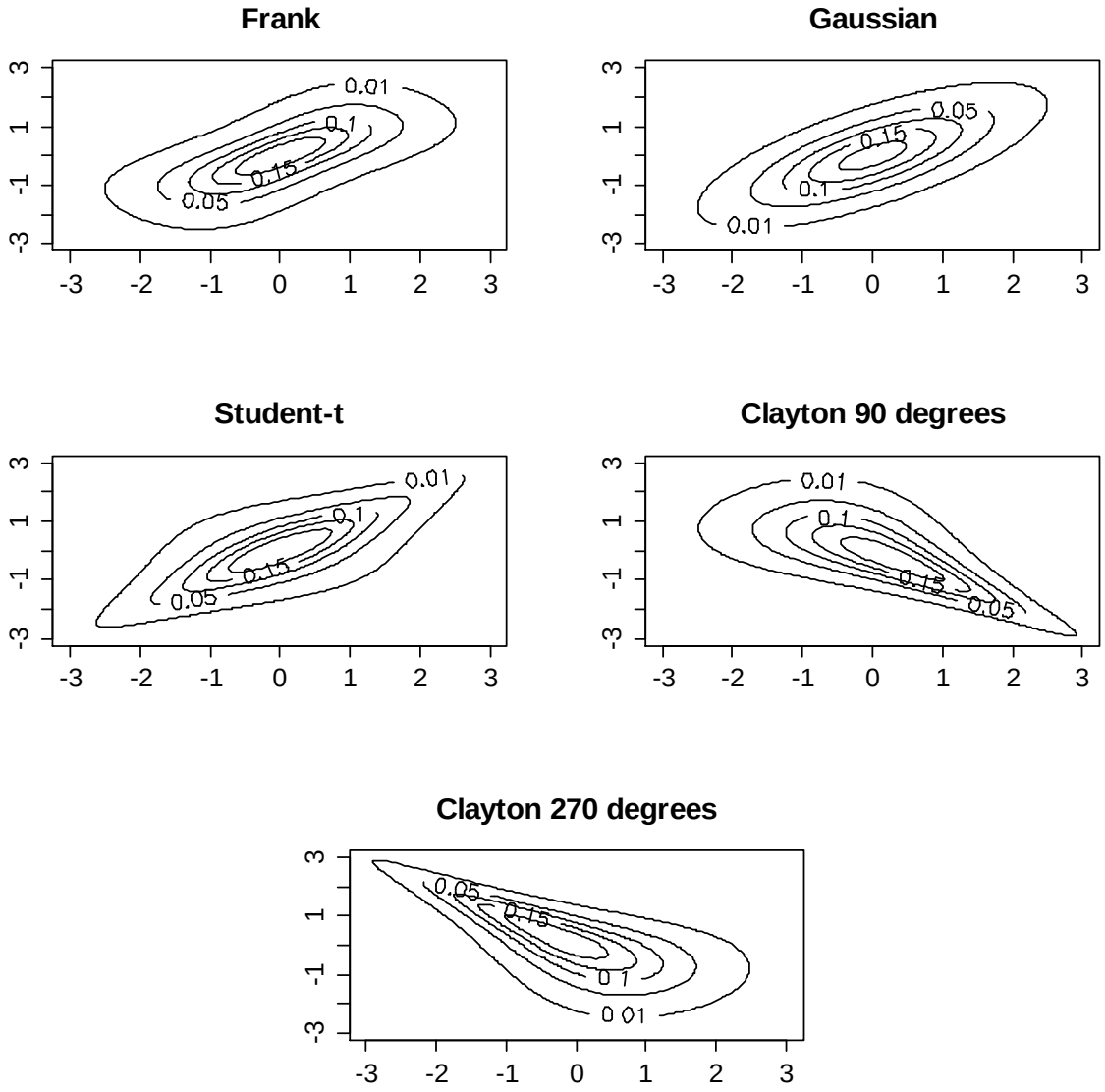
Table 4 HIV Prevalence Estimates (Men and Women in Zambia 2007)

	Men			Women		
	HIV Prevalence	95%CI LL	95%CI UL	HIV Prevalence	95%CI LL	95%CI UL
Normal	21	20	22	19	18	20
Frank	21	20	22	18	17	19
Student-t	22	20	23	20	19	21
Clayton 90	20	18	21	19	18	21
Clayton 270	21	20	22	18	17	19

Note to table 4: HIV prevalence is based on individuals who have a valid HIV test and predicted HIV status from selection models based on the maximisation of model (**11**), and the copula functions defined in table 1. Estimates are presented for selection models based on the maximisation of model (**11**), and the copula functions defined in table 1. Additional control variables include urban setting, region, interview language, ethnicity, religion, marital status, high-risk sexual behaviour 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 as per Hogan et al. (2012). Non-contacts are excluded. Estimates are adjusted for complex survey design.

Figure 1 Illustration of Modelling Dependence Using Copulas


Frank
Gaussian
Student-t
Clayton 90 degrees
Clayton 270 degrees
0.01
0.1
0.15
0.05
-3
-2
-1
0
1
2
3