

Accounting for Selection Bias: An Application to Marriage

David C. Ribar
The George Washington University

October 2003

Presentation to the GW Institute for Public Policy

Family Structure and Well-being

- Marriage associated with a host of positive outcomes (see, e.g., Waite and Gallagher 2000)
 - Better developmental outcomes for children
 - Better economic outcomes for adults and children
 - Better physical and mental health for adults
 - Results appear in numerous studies across time and across countries
- Do these outcomes represent an effect of marriage?
 - Does marriage make people better off?
 - Or are better off people more likely to marry? (Does the association reflect selection)
- Paper reviews theories, methods and evidence

Empirical model

- Consider the simplest possible case—the difference between people who are married and unmarried
- Goal research has been to estimate models of the form

$$Y_i = \alpha M_i + B'X_i + \varepsilon_i$$

- Estimate of α is biased if M_i is correlated with ε_i
- Correlation (and bias) could occur if
 - there are omitted variables that cause Y_i and M_i
 - Y_i causes M_i
 - M_i is mismeasured (won't focus on this case)

What to do?

- Very common problem—most researchers are aware of it
- Typical approach in the marriage literature (and others) has been to acknowledge the problem and caution readers to distinguish between correlation and causality
 - unfortunately, authors also often go on to interpret their results in causal terms
 - word “effect” is used repeatedly
- Recent research has used statistical approaches to address selection problem
 - different methods have different data requirements and rest on different assumptions
 - estimates sensitive to violations of assumptions

Adding variables

- Intuitive approach: can solve problems associated with omitted variables by including the relevant variables
- Advantages:
 - addresses omitted variables problem
 - no special statistical methodology required
- Disadvantages:
 - requires additional data
 - requires that the researcher know what all of the relevant omitted variables are
 - can lead to “kitchen sink” approach and multi-collinearity
 - does not address reverse causality

Circumstances without selectivity (natural experiments)

- Look for situations where M_i occurs independently of Y_i ; intuition is to form a set of natural comparison groups
- Advantages:
 - addresses general set of selection problems
 - does not require special statistical methods
- Disadvantages:
 - hard, if not impossible, to identify these types of situations (examples in marriage literature include disruptions associated with death or job/military separations)
 - may lead to other types of selectivity

Instrumental variables

- Similar intuition: look for variables that affect marriage but only affect well-being through marriage
- Advantages:
 - can be used to address omitted variables, reverse causality and measurement error
 - requires special estimation procedures, but these are available in many statistical packages
- Disadvantages:
 - data requirements: hard to find variables that satisfy necessary properties (divorce laws? marriage fees?)
 - variables need to be strong predictors (efficiency loss)
 - may not work if effects of marriage vary across people

Matching methods

- Intuition: construct comparison groups using married and unmarried people with identical observed characteristics
- Advantages:
 - methodology and assumptions are easy to convey to public and policymakers
 - does not require variable exclusions
- Disadvantages:
 - dimensionality problem—addressed by propensity score method
 - requires comparison observations
 - does not address selection based on unobservable characteristics (e.g., omitted variables)

Modeling the selection process

- Construct a joint statistical model of both the selection and outcome processes
 - examples are MLE selection, dummy endogenous variables and switching models and Heckman/Lee two-stage estimators for these models
- Advantages:
 - address selection based on unobservables
 - available in many statistical packages
- Disadvantages:
 - MLE models require strong distributional assumptions
 - less restrictive semi-parametric approaches available, but these require variable exclusions

Nonparametric bounds

- Manski's insight:

$$E(Y_{Mi}|X_i) - E(Y_{Ui}|X_i) =$$

$$E(Y_{Mi}|X_i, M_i=1) \Pr(M_i=1|X_i) + E(Y_{Mi}|X_i, M_i=0) \Pr(M_i=0|X_i) \\ - E(Y_{Ui}|X_i, M_i=1) \Pr(M_i=1|X_i) - E(Y_{Ui}|X_i, M_i=0) \Pr(M_i=0|X_i)$$

- several terms in this expression are unobserved
- can form bounds by making worst-case assumptions about the terms
- Does not require any assumptions (bounds can be narrowed with additional assumptions)
- Leads to wide bounds (useful only for specification tests)

Fixed effect models

- In data with longitudinal observations, use the information on a person at one point in time to control for unobserved characteristics at another
- Suppose

$$Y_i(t) = \alpha M_i(t) + \mathbf{B}'\mathbf{X}_i(t) + \mu_i + v_i(t)$$

and that μ_i is the source of bias

- Differencing dependent variable and independent variables over time eliminates μ_i
- Can be applied in other dimensions (e.g., within families)

Fixed effect models (cont.)

- Advantages:
 - researcher is not required to identify or measure μ_i
 - easy to apply and available in many packages
- Disadvantages:
 - require special data (longitudinal or family)
 - inference involving observed time-invariant characteristics (e.g., race) and predictions difficult
 - require a strong assumption about source of selection (can be partly relaxed)
 - do not address reverse causality
 - exacerbate measurement error bias
 - cannot generally be applied in non-linear models

Correlated random effect (latent factor) models

- Suppose

$$M_i(t)^* = \Gamma' \mathbf{Z}_i(t) + \theta_i + \zeta_i(t)$$

$$Y_i(t) = \alpha M_i(t) + \mathbf{B}' \mathbf{X}_i(t) + \mu_i + v_i(t)$$

where θ_i and μ_i are

- unobserved, time-invariant variables (random effects)
 - correlated with one another
 - uncorrelated with $\mathbf{Z}_i(t)$, $\mathbf{X}_i(t)$, $\zeta_i(t)$ and $v_i(t)$
- Restrictions
 - shares restrictions of the fixed effects approach
 - additional restrictions on correlations with observables

Correlated random effects models (cont.)

- Advantages:
 - can be applied to non-linear models
 - can examine time-invariant observed variables
 - can be used to predict
 - available in aML software
- Disadvantages:
 - impose many of the assumptions of the fixed effects models plus some additional restrictions
 - estimation methods are computationally intensive

Recommendations—no single, fool-proof method

- Use theory and available evidence to choose approach
 - for instance, if theories suggest that reverse causality is at work, several estimators would be ruled out
 - use evidence to check other assumptions
 - Freedman's (1991) "shoe leather" approach—combine theory with multiple methods
- Specification tests necessary
 - in some cases, models are nested; should test models to see which ones can be ruled out
 - test other assumptions (e.g., balancing in matching estimators, over-identification in IV estimators)
 - test for substantive differences in estimates