

Methods for Addressing Selection Bias in Observational Studies

Susan L. Ettner, Ph.D.

Professor

Division of General Internal
Medicine and Health Services
Research, UCLA

What is Selection Bias?

- In the regression context, selection bias occurs when one or more regressors is correlated with the residual term
- Recall that the residual captures the effects of all omitted and imperfectly measured variables
- Thus any regressors that are correlated with the unmeasured or mismeasured factors will end up proxying for them

The Problem

- In observational studies, especially those based on secondary data, it is common for important factors to be left out, i.e., subsumed into the residual
- If a regressor ends up proxying for those factors, we cannot interpret its estimated coefficient as the effect of that regressor per se, since it also captures part of the effect of the omitted or mismeasured variables

Selection Bias: Example #1

- Health outcomes among patients who see specialists vs. generalists
- If patients who see specialists are unobservably) more severely ill, then positive effects of specialty care on health outcomes may be understated.
- Severity of illness is captured in the error term of the outcome regression and is correlated with specialist

Selection Bias: Example #2

- Utilization among patients with and without health insurance
- If patients choose more generous insurance because they have a greater (unobservable) propensity to use care, then positive effects of insurance on health care use may be overstated.
- Propensity to use care is in the error term of the utilization regression and is correlated with insurance

Selection Bias: Example #3

- Earnings among persons with and without a college degree
- If persons with a college degree are more (unobservably) motivated or intelligent, then positive effects of college degree on earnings may be overstated.
- Motivation or IQ are in the error term of the earnings regression and are correlated with college degree

Selection Bias: Example #4

- Health care costs among HMO members vs. fee-for-service patients
- If patients enroll in HMOs because they have a lower (unobservable) propensity to use health care, then negative effects of HMO enrollment on costs may be overstated.
- Propensity to use care is in the error term of the cost regression and is correlated with HMO enrollment

Possible Solutions

- “Treatment effects” models
 - Requires strong distributional assumptions
- Instrumental variables methods
 - Two-stage least squares as a special case
 - Requires variable that is correlated with the “treatment” but uncorrelated with the outcome
- Propensity score methods
 - Fewer assumptions but does not necessarily help with bias due to unobservables
 - Used with dichotomous regressors

Treatment Effects Models

- The basic idea behind these models is to estimate two regressions simultaneously.
- The first is a probit regression predicting the probability of “treatment.”
 - Probit models are very similar to logit but assume normally distributed error terms.
- The second is either a linear or probit regression for the outcome of interest as a function of the “treatment” variable, controlling for observable confounders.

Treatment Effects Models (cont'd)

- To simultaneously estimate the two regressions, you assume that the error terms are jointly normally distributed and use maximum likelihood methods.
- Ideally you would have some explanatory variables in the treatment regression that do not belong in the outcome regression.
 - Strictly speaking, this is not necessary, but it helps to “identify” the effect of the treatment on the outcome and makes the estimates more robust (more on this later)

Treatment Effects Models (cont'd)

- Recall that selection bias arises because the “treatment” was correlated with the error term in the outcome equation.
- By estimating the two equations together, this method allows you to model that correlation directly, thereby eliminating the omitted-variable bias
- The tradeoff is that you have to assume a particular joint distribution for the error terms.

Selection Bias: Example #1 Revisited

- To get an unbiased estimate of the effect of provider specialty on health outcomes, simultaneously estimate the following:
 - probit regression of whether the patient saw a specialist vs. generalist
 - linear or probit regression of health outcome
- The simultaneous estimation is equivalent to controlling for the non-zero expectation of the error term in the outcome equation, eliminating the omitted-variable bias

Testing for Selection Bias

STATA will give you an estimate of rho (ρ , the correlation between the error terms of the two equations), sigma (σ , the standard error of the outcome regression if linear) and lambda ($\lambda = \rho * \sigma$).

If ρ is positive (negative), the estimated (positive) effect of treatment from single-equation estimation will generally be biased away from zero (towards zero).

Testing for Selection Bias (cont'd)

STATA will automatically test for you whether $\rho=0$ (or equivalently, whether $\lambda=0$, since $\sigma>0$). If $\rho=0$, there is no selection bias and you can present the single-equation estimates. If $\rho\neq 0$, there is bias and you should present the estimates from the treatment selection model instead. In other words, the diagnosis is also the cure. Of course, all of this is based on our model assumptions being correct.

STATA Notes

- If your outcome is linear, the STATA command is:
 - `treatreg depvar treatvar controlvars, treat`
(`treatvar = selectvars`)
- STATA will calculate several predicted values based on `treatreg`, so be sure to use the right option for what you want
 - The default `predict varname, xb` yields the expected value of the outcome for the entire sample, which is usually what you want

STATA Notes (cont'd)

- Instead of using maximum likelihood, you can estimate the model with a two-stage method by choosing the *twostep* option. This is useful if ML won't converge.
- If your outcome is dichotomous, the STATA command is:
 - biprobit (eq1:*depvar = treatreg indvars*)
(eq2: *treatreg = selectvars*)

Propensity Scores vs. Treatment Effects

- Propensity scores (Rosenbaum and Rubin) can be used in the same situations as treatment effects models
- The treatment effects models explicitly address bias caused by correlation of the regressor with omitted variables, by adding a term to the regression that represents the non-zero expectation of the error term.
- The PS approach is slightly different, as it does not explicitly address unobservables.

Propensity Scores vs. Treatment Effects (cont'd)

- Propensity scores “balance out” the groups being compared in terms of their covariates.
- Thus the propensity score approach seeks to do a better job of controlling for *observable* characteristics.
- Its main advantage over regular regression adjustment is that it avoids out-of-sample prediction due to linearity assumptions, e.g., if everybody in the first group is young and everybody in the second group is old.

Propensity Scores vs. Treatment Effects (cont'd)

- However, the ultimate goal is the same – to turn an observational comparison of the treatment and comparison groups into a quasi-randomized design.
- Also, if the observables are correlated with the unobservables, then you may be able to “balance out” the latter by doing a better job of balancing the former. R&R extend the model to allow assessment of sensitivity to an unobservable binary covariate.

Overview of Propensity Scores

- First run a logit (or probit) regression of the probability of treatment as a function of the observed covariates, X
- Use the resulting estimates to create a predicted probability of treatment for each patient; this is the propensity score
- There are several ways to use the propensity score to improve the estimated effect of the treatment on the outcome

Ways to Use the Propensity Score

- Matching
- Stratification
- Regression adjustment, with or without matching and/or stratification

Matching

- The propensity score can be used to divide patients into groups with a similar probability of receiving the treatment, regardless of whether they actually did
- As a result, the comparison of outcomes between treated and untreated patients is “quasi-randomized.”
- The matching method is used most commonly when the two groups (treatment and control) are of very different sizes.

Matching (cont'd)

- Randomly order the treated subjects
- Use the propensity score to match the first treatment patient to all control patients within a given caliper around the score
 - If multiple possible matches are made, pick the closest match
- Put the treatment and matched control observations into the new dataset and repeat the process in order for all other treated subjects.

Matching (cont'd)

- Although patients could be directly matched on the basis of observable covariates, this process is difficult with many covariates
- PS allow you to “balance” the covariates, so even if matched patients aren’t exactly alike in every way, they have a similar overall probability of being treated.
- Note that patients in either group who do not match to patients in the other group are eliminated from the sample altogether.

Stratification

- Stratify the sample by (say) quintile of the PS distribution and calculate separate treatment effects for each
- You can also calculate the weighted average of the five estimates to report an overall effect
- The variance of the overall estimate can be calculated using the formula $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$

Stratification

- The lowest or highest quintiles may drop out of the analyses altogether, if they contain only treated or only control patients.
- If there is too much overlap between the two groups in each quintile, then the model predicting treatment choice may not discriminate well between these groups.
- You should check how well the stratification “balances” treatment and control patients within each stratum.

Regression Adjustment: Options

- Control for continuous propensity score or for propensity score quintile, with or without controlling for a (subset of) the covariates used to calculate the propensity score
- Use the matching method to define the sample, then perform the above regression
- Use the stratification method to define subsamples based on the propensity score, then estimate separate regressions within each subsample

Instrumental Variables (IV)

- IV methods are used for the general case of endogeneity bias
- Exogenous variables are determined outside of the model, e.g., age, sex, race.
- Endogenous variables are determined within the model, i.e., as a simultaneous structural equations system.
- Bias arises when one endogenous variable is regressed on another.

Endogeneity Bias

- Although endogeneity bias is not identical to correlation of the regressor with the error term (Greene), it is similar and the easiest way to think about this.
- If random noise affecting the outcome in turn affects one of the regressors, then you have a bias that IV methods may help with.
- Thus IV can be applied to selection as well as reverse causality problems

More on Endogeneity Bias

- With single-equation estimation, the effect of an endogenous regressor on the outcome will be biased, as will the effects of other regressors that are correlated with it (and in nonlinear models, *all* regressor effects, whether correlated or not)
- The direction of the bias cannot be determined theoretically if there is more than one regressor, due to correlations among regressors

When Do You Have Endogeneity Bias?

Ask yourself the following:

- Is the dependent variable determined simultaneously with any covariates?
- Can the story be told in both directions?
- Are any of the regressors correlated with the error term?

If yes, need to test for endogeneity (using Hausman-Wu or augmented regression)

Endogeneity Bias in Action

Even after adjusting for observable baseline severity...

- Depression treatment was associated with worse outcomes in the PIC study
- Medicaid insurance was associated with higher mortality in the HCSUS study

Both of these results reversed themselves when IV methods were used.

A Simple Example of How IV Works

Q: Does employment reduce depression among women?

Stylized Fact #1: Rates of depression are lower among women who are employed.

But...does this mean that employment lowers depression, or that women who are depressed are less likely to seek or obtain a job?

A Simple Example (cont'd)

Stylized Fact #2:

Women whose mothers worked during their formative years are also more likely to work themselves.

Stylized Fact #3:

Women whose mothers worked during their formative years have lower rates of depression.

A Simple Example (cont'd)

If maternal employment during the daughter's formative years does not *directly* influence whether the daughter is depressed later in life, then the *only* way to explain the lower rates of depression among daughters whose mothers worked is through the daughters' own employment.

A Simple Example (cont'd)

We know:

Mother worked \Rightarrow daughter worked

Mother worked \Rightarrow daughter less depressed

before controlling for whether daughter worked

We assume:

Mother worked \nRightarrow daughter less depressed

after controlling for whether daughter worked

This means:

Daughter worked \Rightarrow daughter less depressed

A Simple Example (cont'd)

In other words, employment causally affects depression. The causal impact of employment on depression is “identified” through our assumptions that (1) the mother’s employment (the “instrument”) affects the daughter’s employment but (2) does not directly affect the daughter’s depression. Note that if either assumption fails, we cannot draw this causal inference.

Two-Stage Least Squares (2SLS)

- The instrumental variables method is illustrated using the special case of 2SLS, when both variables are continuous and regression analysis is used
- First, estimate a “reduced-form” regression of the endogenous regressor on all exogenous variables in the system, meaning all variables that explain either the endogenous regressor or the outcome

Two-Stage Least Squares (cont'd)

- Next, use the regression estimates to construct a predicted value for the endogenous regressor
- Substitute this predicted value for the actual value of the endogenous regressor in the outcome equation
- Then estimate as you normally would and adjust the SE's for the use of a predicted value (STATA does this for you)

Two-Stage Least Squares: Example

- Suppose you want to estimate the effect of informal caregiving hours on work hours.
- However, just as caregiving hours reduces work hours, work hours reduce caregiving hours, leading to potential endogeneity bias
- Estimate caregiving hours as a function of all variables that explain either caregiving hours or work hours
 - This is the “first-stage” or “reduced-form” regression

2SLS: Example (cont'd)

- Create predicted caregiving hours for each person
- Estimate work hours as a function of predicted caregiving hours instead of actual caregiving hours
 - This regression also controls for all other variables that affect work hours (but not variables affecting caregiving hours only)
 - This is the “second-stage” or “structural” regression

Intuition

- By using predicted rather than actual values of caregiving hours, you are “breaking” its correlation with the error term in the work hours regression, so you get an unbiased estimate of its effect on work hours.
- The predicted values are just linear combinations of the exogenous variables, so by construction are not correlated with the error term.

Instruments and the “Exclusion” or “Identifying” Restriction

- Instruments for informal caregiving are variables included in the first-stage but not the second-stage regression.
- To “identify” the effect of caregiving hours on work hours, we need at least one variable that affects caregiving hours, but does not directly influence work hours after controlling for caregiving hours and other covariates.

Intuition Behind “Identification”

- To determine the effect of caregiving hours on work hours, we want to see how work hours change when there is an exogenous shift in caregiving hours, *holding everything else determining work hours constant*.
- Thus, we need to find something that will shift around caregiving hours while the other regressors in the work hours equation remain unchanged.

Intuition Behind “Identification” (cont’d)

Suppose the instrument is parental health. If parental health does not actually affect caregiving hours, or if parental health has a direct influence on work hours (perhaps because the child needs to work more to pay for the parent’s health care), then we can’t separate out the effect of caregiving hours from the direct effects of the other regressors.

Randomized Controlled Trials: An IV Application

- Suppose many of the people randomized to the treatment group do not actually end up getting the treatment
- “Intent-to-treat” design may understate the true effect of the treatment (say, on HRQL)
- The random assignment can be used as the instrument for whether the person actually got the treatment in an “as-treated” analysis

RCTs: An IV Application (cont'd)

- Stage 1: Estimate whether the person got treatment as a function of treatment assignment and all other predictors of either treatment or HRQL
- Stage 2: Estimate HRQL as a function of the predicted probability of treatment, controlling for other determinants of HRQL
- Treatment assignment is assumed to influence HRQL only indirectly, through actual receipt of the treatment.

Interpretation

- For which patients does this procedure yield an unbiased estimate of the effect of treatment on HRQL?
- At a minimum, this estimate applies to the “marginal” people, i.e., those for whom a change in the instrument (treatment assignment) would change the endogenous regressor (treatment).
- If the treatment effect is homogeneous, the estimate generalizes to the entire population.

IV Assumptions

- The five IV assumptions are illustrated using the example of estimating the effect of having a usual source of care (USC) on the use of preventive care
 - USC is endogenous because people who (unobservably) care more about their health are both more likely to develop a relationship with a provider and to seek preventive care.
 - Length of residence in the geographic area is used as the instrument for having a USC

IV Assumptions: *Non-Zero Average Causal Effect*

Formally:

- The instrument must predict the endogenous regressor, controlling for the other covariates.

Example:

- Controlling for the other covariates, length of residence is a good predictor of having a USC.

IV Assumptions: *Exclusion Restriction*

Formally:

- The instrument has only a negligible direct influence on the outcome after controlling for the covariates, that is, the instrument is uncorrelated with the error term.

Example:

- Controlling for having a USC and the other covariates, length of residence does not predict use of preventive services.

IV Assumptions: *Monotonicity*

Formally:

- The instrument cannot increase the value of the endogenous regressor for some subjects, but decrease it for others.

Example:

- All respondents who would have had a USC if they lived in an area for a short time would also have one if they lived in the area for a long time.

IV Assumptions: *Random Assignment*

Formally:

- Subjects must be effectively randomized into the value for the IV within subgroups defined by the other covariates.

Example:

- Knowing a respondent's use of preventive services does not yield any information about that respondent's length of residence.

IV Assumptions:

Stable Unit Treatment Value Assumption

Formally:

- The outcome of one subject is not influenced by the value of the endogenous regressor for other subjects.

Example:

- The use of preventive services by one subject is not influenced by whether other subjects have a USC, and differences in effectiveness among types of USC are minor.

IV Assumptions: A Note

- Key question: Is the correlation of the IV with the endogenous regressor high *relative* to its correlation with the outcome?
- The greater the correlation of the IV with the outcome, the stronger its correlation with the endogenous regressor needs to be.
- Test the individual and joint significance of the instruments in the 1st-stage regression and if you have >1 possible instrument, test the overidentifying restrictions on the model.

Places to Look for Instruments (borrowed from Staiger)

- Geography (distance, rivers, small area variation)
- Legal/political institutions (laws, election dynamics)
- Administrative/program rules (wage/staffing rules, reimbursement rules, eligibility rules, mandates)
- “Natural” randomization (draft, birthdate, lottery, roommate assignment, weather)₆

Limitations of IV Analysis

- It's difficult to find a valid instrument
 - Sometimes instruments are so bad that IV is actually more biased than OLS
 - Can't test exclusion restriction unless you have multiple instruments, and even then, validity of test depends on having at least one good instrument
- You lose precision, especially if the instruments don't predict the endogenous regressor very well

Comparison: IV vs. Selection Models

- Selection models with valid exclusion restrictions (variables that predict treatment but not the outcome) and IV models with valid instruments (same thing) are closely related
- Selection models can be estimated without exclusion restrictions, but then you are relying heavily on (untestable) assumptions about the joint distribution of the error terms

Comparison (cont'd)

- Tradeoff between robustness and efficiency; if the joint distributional assumption holds, selection models are more efficient than IV, but they are more sensitive to the failure of the assumption
- IV methods are more complicated and can be problematic when either the endogenous regressor or dependent variable is dichotomous, so in that case, selection models are often preferable