

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

Bias from Using OLS to Estimate Treatment Effects

Regression equation:

$$Y = X\alpha + \theta T + \varepsilon$$

where T is a dummy for some “treatment” of interest and X represents observable confounders

$$E(Y \mid T=1) = Xa + \theta + E(\varepsilon \mid T=1)$$

$$E(Y \mid T=0) = Xa + E(\varepsilon \mid T=0)$$

Bias from Using OLS (cont'd)

The estimated coefficient on T captures

$$E(Y | T=1) - E(Y | T=0)$$

$$= \theta + E(\varepsilon | T=1) - E(\varepsilon | T=0)$$

= θ only if the OLS assumption that

$$E(\varepsilon | T) = 0 \text{ holds.}$$

This means that knowing whether patients got treated doesn't tell you anything about their unobservable characteristics. But this assumption often fails.

Bias from Using OLS: Example

Suppose you can't measure severity and more severely ill patients get treated, while healthier patients do not. If Y

represents a good health outcome, then

$E(\varepsilon | T=1) < E(\varepsilon | T=0)$, and

$$\theta + E(\varepsilon | T=1) - E(\varepsilon | T=0) < \theta$$

That is, the positive impact of treatment on health outcome will be understated.

More Examples of Selection Bias

- Health outcomes among patients who see specialists vs. generalists
- Utilization among patients with and without health insurance
- Earnings among persons with and without a college degree
- Health care costs among HMO members vs. fee-for-service patients

Direction of Bias: Examples

- 1) If patients who see specialists are (unobservably) more severely ill, then positive effects of specialty care on health outcomes may be understated.
- 2) 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.

Direction of Bias: Examples (cont'd)

- If persons with a college degree are more (unobservably) motivated or intelligent, then positive effects of college degree on earnings may be overstated
- 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.

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
 - Can only be used with dichotomous regressors

Treatment Effects Models

Probit model of selection into treatment:

$$T^* = W\beta + \eta \quad \eta \sim N(0,1)$$

T^* is a latent index, W are observable characteristics and η are unobservables

$$T = 1 \text{ if } T^* > 0, \quad T = 0 \text{ if } T^* \leq 0$$

$$\text{pr}(T=1) = \Phi(W\beta), \quad \text{pr}(T=0) = 1 - \Phi(W\beta)$$

Linear outcome regression: $Y = X\alpha + \theta T + \varepsilon$

Key assumption:

$$(\eta, \varepsilon) \sim \text{bivariate normal } [0,0,1, \sigma_\varepsilon, \rho]$$

Treatment Effects Models (cont'd)

We can use a formula for the joint density of bivariate normally distributed variables to substitute into our earlier formula:

$$\begin{aligned} E(Y \mid T=1) &= Xa + \theta + E(\varepsilon \mid T=1) \\ &= Xa + \theta + \rho \sigma_{\varepsilon} \frac{\phi(Wb)}{\Phi(Wb)} \end{aligned}$$

$$\begin{aligned} E(Y \mid T=0) &= Xa + E(\varepsilon \mid T=0) \\ &= Xa - \rho \sigma_{\varepsilon} \frac{\phi(Wb)}{1 - \Phi(Wb)} \end{aligned}$$

Treatment Effects Models (cont'd)

Thus, without controlling for selection, the coefficient on the treatment indicator will estimate $E(Y | T=1) - E(Y | T=0) =$

$$Xa + \theta + \rho \sigma_{\varepsilon} \frac{\phi(Wb)}{\Phi(Wb)} - Xa + \rho \sigma_{\varepsilon} \frac{\phi(Wb)}{1 - \Phi(Wb)} =$$

$$\theta + \rho \sigma_{\varepsilon} * \frac{\phi(Wb)}{\Phi(Wb) [1 - \Phi(Wb)]}$$

If ρ is positive (negative), the OLS coefficient on T will be biased upward (downward).

Treatment Effects Models (cont'd)

As with all omitted variable bias, the solution is to add the omitted variable to the regression. In this case, the omitted variable is the selection term (which will depend on the individual) and the estimated coefficient on this term will be $\lambda = \rho^* \sigma_\varepsilon$. Since σ_ε must be positive, the sign and significance of λ tell you whether there is selection bias. However, in this case, the diagnosis is also the cure.

Treatment Effects Models (cont'd)

STATA's *treatreg* command uses maximum likelihood methods to estimate the probit treatment equation and linear outcome equation simultaneously:

$$T^* = W\beta + \eta$$

$$Y = X\alpha + \theta T + \varepsilon$$

The estimate of θ should be consistent if the assumption that η and ε are jointly normally distributed holds.

STATA Notes

- Syntax for the treatment effects model 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 command `predict varname, xb` will give you the unconditional expected value of the outcome for the entire sample, which is usually what you want

STATA Notes (cont'd)

- Instead of MLE, you can estimate the model using a two-stage method if you give the *twostep* option in STATA. This is useful if ML won't converge.
- If your outcome measure is dichotomous, you can use the *biprobit* command to estimate a model equivalent to *treatreg*, again assuming a joint normal distribution
- Syntax 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

- $H = \alpha + \beta * X + \delta * T + \varepsilon$
- H is the health outcome of interest
- X is observable confounding factors
- T is a 0-1 indicator for whether the patient got a particular treatment
- ε represents unobservables that influence H

Overview of Propensity Scores (cont'd)

- The propensity score for patient i is the conditional probability of assignment to the treatment condition ($T=1$), given observed covariates, X .
- We run an auxiliary logit (or probit) regression to predict the probability of T :
$$\text{pr}(T=1) = \frac{e^{X\theta}}{1 + e^{X\theta}}$$
- The propensity score is the predicted probability based on the estimates of θ .

Ways to Use the Propensity Score

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

Matching

- The propensity score is used to divide patients into groups with a similar probability of receiving the treatment, regardless of whether they actually did, so 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 PS to match the first treatment patient to all control patients within a given caliper around the PS; if >1 possible match is 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, calculate separate treatment effects for each, and also average the five estimates into an overall effect:

$$\delta = \sum_{k=1}^5 (n_k/N) [\mu_{k, \text{treated}} - \mu_{k, \text{untreated}}]$$

where k indexes the stratum, N is total sample size, n_k is the sample size for stratum k , $\mu_{k, \text{treated}}$ ($\mu_{k, \text{untreated}}$) is the mean outcome for treated (untreated) patients in stratum k

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 only for continuous propensity score or only for propensity score quintile
- Control for the propensity score (either continuous or quintile) along with (a subset of) the covariates used to calculate the propensity score
- Define subsamples based on the propensity score, as with the stratification approach, and then estimate separate regressions within each subsample

Propensity Score (PS) vs. Ordinary Regression Adjustment

- PS adjustment can be useful to avoid overfitting the model with lots of regressors or quadratic and interaction terms. You can put all of these covariates into the 1st stage equation for treatment choice and leave them out of the 2nd-stage equation.
- There are technical problems with PS adjustment if the variance is larger in the untreated than treated group, so matching or stratification may be better.

Instrumental Variables (IV)

- IV methods can be 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 exogenous shocks affecting the outcome also affect any regressors, then $E(X'\varepsilon) \neq 0$ and you have a bias that IV methods might help with.
- Thus IV can be applied to selection as well as reverse causality problems

More on Endogeneity Bias

- Using single-equation estimation, the effect of an endogenous regressor on the outcome will be biased and inconsistent, 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 is theoretically indeterminate if there is >1 covariate, due to correlations among regressors

When Do You Have Endogeneity Bias?

Ask yourself the following about your regression:

- 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, use Hausman-Wu or augmented regression to test for endogeneity.

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 casual 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.

A Formal Example of IV Methods

- What is the effect of informal caregiving on work hours?
- Structural equations are:

$$W = \delta I + \tau X_w + \varepsilon_w \quad (1)$$

$$I = \lambda W + \gamma X_I + \varepsilon_I \quad (2)$$

We want a consistent estimate of δ .

- Treat both variables as continuous => special case of IV, known as two-stage least squares (2SLS).

Step 1: Run reduced-form regression and create predicted values for endogenous regressor

- Estimate the reduced-form regression of informal caregiving on all of the exogenous variables in the system:

$$I = \alpha X_W + \beta X_I + \eta$$

- A predicted value is then constructed using the regression estimates:

$$I^P = \alpha^{\text{hat}} X_W + \beta^{\text{hat}} X_I$$

Step 2: Estimate structural equation after replacing actual with predicted value for endogenous regressor

- Substitute the predicted value of informal caregiving from the 1st stage (I^P) for the actual value and then estimate:

$$W = \delta I^P + \tau X_w + \varepsilon_w$$

- If doing this by hand, standard errors must be adjusted for use of predictions.

Intuition

- By using predicted rather than actual values of the endogenous regressor, you are “breaking” its correlation with the error term, so you get a consistent estimate of its structural effect, δ .
- 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 X_I but not X_W .
- To “identify” the effect of I on W , we need at least one variable that affects I , but does not directly influence W after controlling for I and X_W . If X_I and X_W are identical, then I^P is collinear with X_W and its effect cannot be estimated.
- For example, parental health might be included in X_I but not X_W .

Identified Work Hours Equation

- For simplicity, suppose that the only variables in X_I are the instruments.

$$I = \alpha X_W + \beta X_I + \eta \quad \text{1st stage regression}$$

$$W = \delta I^p + \tau X_W + \varepsilon_W \quad \text{2nd stage regression}$$

$$= \delta(\alpha X_W + \beta X_I) + \tau X_W + \varepsilon_W$$

$$= (\delta\alpha + \tau)X_W + \delta\beta X_I + \varepsilon_W$$

- We know α and β from the 1st stage and we can estimate $(\delta\alpha + \tau)$ and $\delta\beta$ by regressing W on X_W and X_I . Thus δ and τ can be uniquely identified.

Identified Work Hours Equation: Intuition

- To determine the effect of I on W , we want to measure how W changes when there is an exogenous shift in I , *holding everything else determining W constant*.
- Thus, we need to find something that will shift around I while the other regressors in the equation for W remain unchanged. Otherwise, we can't separate out the effect of I from the direct effects of the other regressors.

Unidentified Work Hours Equation #1

Case 1. X_i does not influence informal caregiving (weak instruments)

$$I = \alpha X_w + \eta$$

$$\begin{aligned} W &= \delta I^p + \tau X_w + \varepsilon_w \\ &= \delta(\alpha X_w) + \tau X_w + \varepsilon_w \\ &= (\delta\alpha + \tau) X_w + \varepsilon_w \end{aligned}$$

We know α from the 1st stage and we can estimate $(\delta\alpha + \tau)$ from a regression of W on X_w , but that isn't enough to uniquely identify δ and τ .

Unidentified Work Hours Equation #2

Case 2. X_i has direct effects on both I and W (non-excludability)

$$I = \alpha X_w + \beta X_i + \eta$$

$$W = \delta I^p + \tau X_w + \theta X_i + \varepsilon_w$$

$$= \delta(\alpha X_w + \beta X_i) + \tau X_w + \theta X_i + \varepsilon_w$$

$$= (\delta\alpha + \tau)X_w + (\delta\beta + \theta)X_i + \varepsilon_w$$

We know α and β from the 1st stage and we can get $(\delta\alpha + \tau)$ and $(\delta\beta + \theta)$ by regressing W on X_w and X_i , but since we don't know τ or θ , we can't figure out δ .

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)

1st-stage regression:

$$\text{Treated} = \alpha \text{Treatment Assignment} + \delta X + \varepsilon$$

2nd-stage regression:

$$\text{HRQL} = \beta^* \text{Treated} + \lambda X + \eta$$

Treatment assignment is assumed to influence HRQL only indirectly, through actual receipt of the treatment.

Interpretation

- For which patients is β a consistent 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 IV (treatment assignment) would change the endogenous regressor (treatment).
- If the treatment effect can be assumed to be homogeneous, then the estimate generalizes to the entire population.

Main IV Assumptions

- *Non-zero average causal effect* The instrument must predict the endogenous regressor, controlling for the other covariates.
- *Exclusion Restriction* 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.

Main 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 IVs in the 1st-stage regression and if you have >1 possible instrument, test the overidentifying restrictions on the model.

Other IV Assumptions

- *Monotonicity* The instrument cannot increase the value of the endogenous regressor for some subjects, but decrease it for others.
- *Random assignment* Subjects must be effectively randomized into the value for the IV within subgroups defined by the other covariates.
- *Stable unit treatment value assumption* The outcome of one subject is not influenced by the value of the endogenous regressor for other subjects.

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)

- As always, tradeoff between robustness and efficiency; if the joint distributional assumptions hold, selection models are more efficient than IV, but the cost is less robustness if assumption fails
- 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