MLE: Categorical and Limited Dependent Variables
Unit 4-1: Endogenous Regressor and Sample-Selection Models

PS2730-2021

Weeks 12-13

Professor Steven Finkel

Endogenous Regressor Models

- In the censored outcome models we considered a few weeks ago, we saw that one way to estimate that model was to view it as potentially a form of *omitted variable bias*: factor(s) related to the likelihood of a unit having a censored observation (including X) were also related to the observed value of Y (given censoring).
- In the two-step tobit, we then estimated a variable corresponding to the Inverse Mills Ratio, or the likelihood of having a censored observation, and censoring, and included this into the analysis to obtain unbiased estimates of the (causal) effects of X on Y* (the true value of a dependent variable.
- Omitted variables are one source of the *endogeneity* in the relationship between independent and dependent variable that causes bias in the estimation of causal effects unless corrective steps are taken.
- In this unit, we'll extend our discussion to incorporate other causes of **endogeneity** and possible corrections in many different models with non-continuous variables

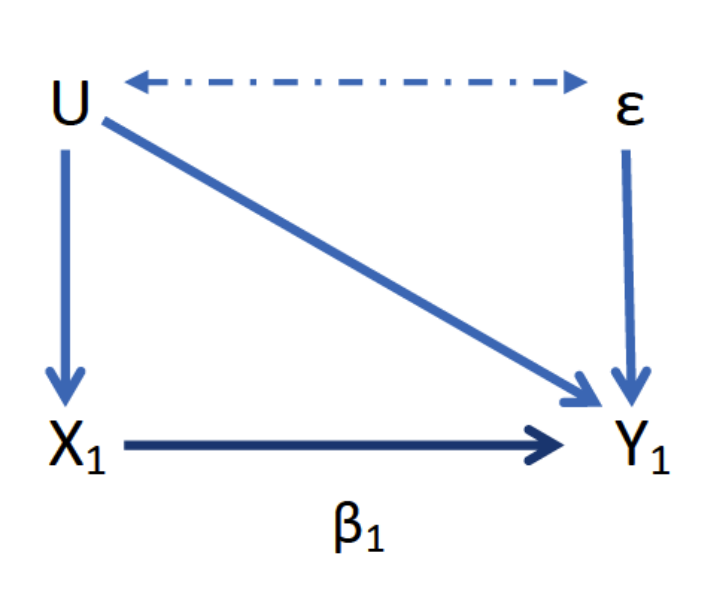
- One extension is in the estimation of *sample-selection* models, which involve a similar kind of omitted variable bias as in the censoring case. Unobserved factors that lead units to be in the sample in the first place are related to both X and Y. This is called ***"endogenous sample selection"*** – omitted variable bias related to inclusion in the sample.
- In other instances, we have an ***endogenous treatment***, such that we are trying to estimate the effect of an endogenous dichotomous "treatment" variable on a continuous outcome. For example, the "treatment" effect of UN peacekeeping on conflict deaths: whether the UN sends peacekeepers to a conflict setting likely depends on the severity of the conflict in the first place. This is a classic endogenous treatment!
- Or we may have endogenous ***continuous*** independent variables which impact ***non-continuous*** outcomes. Example: years of education (X) on employment status (Y). Unit-level unobservables – e.g., personality, grit or determination, family connections -- affecting the acquisition of education also affect the likelihood of employment.

- We can have endogenous treatments which are either dichotomous or some other form (ordered, multinomial, censored) affecting continuous or non-continuous outcomes.
- Let's discuss the main features of these models, the problems they create, and typical solutions
- We'll see that it is useful to consider most endogeneity processes in a multi-equation system, with one equation governing the "sample selection", the "treatment" or the "endogenous regressor" variable, and the other governing the outcome Y (or Y*).
- **The correlation between the disturbances of the two equations represents the extent of the endogeneity that needs to be taken into account**

- We will also see that it is advantageous to have additional variable(s) which affect the selection/ treatment/endogenous regressor variable but *not* the outcome variable directly
- These additional variables serve as ***instrumental variables***, similar to how instruments are used in the endogenous continuous regressor-continuous outcome models that you considered in PS2030
- Let's first discuss the general problem of endogeneity in more detail as applied to continuous outcomes, then move on to sample selection and other important endogeneity models in the non-continuous variables case.
- In the process, we'll (re-)introduce Stata's GSEM and the new ERM ("Extended Regression") modules for estimation purposes

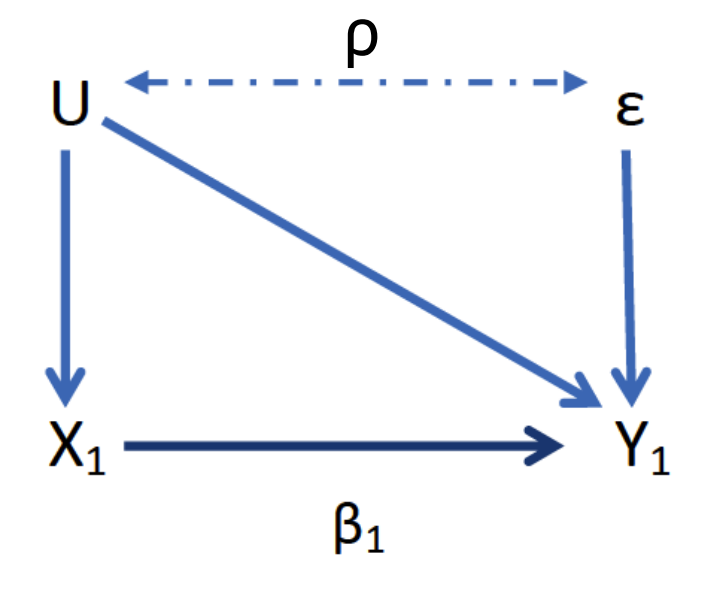
- Endogenous Regressors in Linear Models
  - The general problem of endogeneity in linear regression: whenever X related to the disturbance term in the outcome equation for Y, where X is an independent variable
  - Occurs when:
    1. Omitted variables are related to both X and Y
    2. There is measurement error in X
    3. There is reverse causality such that Y causes X as well as (or instead of) X causing Y
  - Any of these situations leads to a correlation between X and ε in the Y equation, which violates the *exogeneity* assumption in OLS that $E(X\,\varepsilon)=0$
  - This causes estimates of the β associated with X to be "biased", and the bias does not diminish as $N->\infty$ (so also "inconsistent")

Omitted variable bias in a bivariate causal system



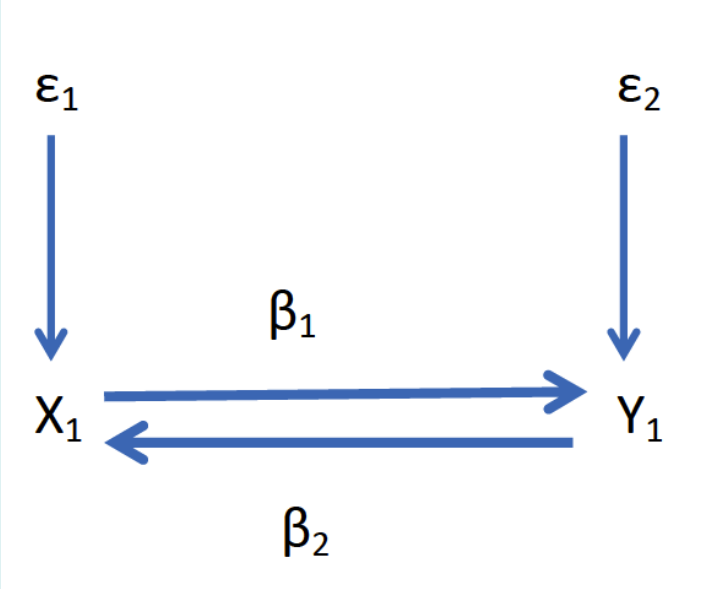
U causes both X_1 and Y_1
U is unobserved, so folded into ε
X is now related to ε and hence "endogenous"
OLS β_1 will give the true effect of X_1 *plus* some correlated effect from U

- Diagram shows the utility of looking at the problem through a "multi-equation" system
- U – what we call the "unobservables" affecting X and Y, is actually part (or all) of the error term in an equation predicting X, the independent variable.
- It then "causes" Y, but, since it is unobserved, it is folded into ε, and "causes" Y via its role as part of the error term for Y.
- So the endogeneity in the model is transmitted via the correlation between the disturbance terms for X and Y's respective equations
- NOTE: in the longitudinal case the stable unobservables (ζ_i) are also part of the composite error term. So there could be endogeneity in the (X ζ_i) correlation *and* other portions of the error term -- the (Xε_{it}) correlation.



- Back to the multi-equation perspective. Call the correlation ϱ.
- If ϱ=0, then there is no endogeneity **bias** – there may be omitted variables causing X but they are not also causing Y; or there may be omitted variables causing Y but they are not related to X
- But if ϱ≠0, then, if not corrected, β absorbs some of the correlated effects via ϱ and hence it is estimated with bias

Another Endogeneity-Inducing Problem: Reciprocal Causality

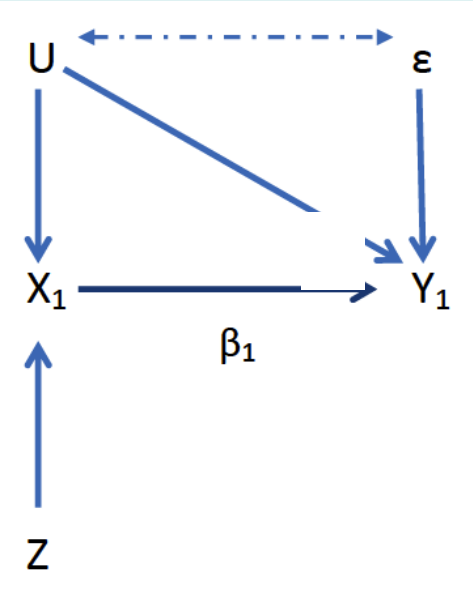


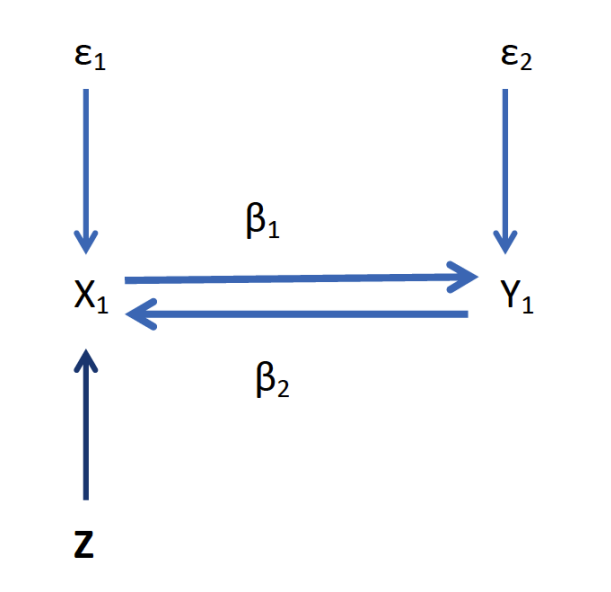
$$Y_1 = \beta_1 X_1 + \varepsilon_2$$

$$X_1 = \beta_2 Y_1 + \varepsilon_1$$

X_1 is a function of ε_2. But, in the equation predicting Y_1, OLS assumes that X_1 and ε_2 are unrelated. This is the $E(X\varepsilon)=0$ assumption for the OLS estimate of β_1 to be unbiased

Solution: Add information to the system in the form of additional "instrumental variables"





To estimate β, we need an "instrument" Z for X

Conditions that Z *MUST* Fulfill:

1) The "Exclusion Restriction": Z does not cause Y_1 except through X
2) The "Exogeneity Restriction": Z is unrelated to U and ε

Logic of IV Analysis

- The logic of IV analysis is as follows. Given an endogenous regressor **X** in an equation with some dependent variable **Y**, we find some *exogenous* variable **Z** that produces change in Y through one mechanism only -- *the mediating effect of* X.
- Because Z is:
  - (a) Exogenous; and
  - (b) Has no direct effect on Y, then:
- Any changes in Y that may result from changes in Z ***must be attributable to X***, and must also be **unrelated to the problematic endogenous part of the X-Y relationship**.
- So the instrumental variable gives us *exogenously-induced* changes in X, and we then test whether these changes produce subsequent changes in Y.

- Classic Example #1: The "return" on earnings from education. The kinds of people who seek education are likely to have unobservables that relate to earning power, over and above whatever education they get. So education and the error term in an earnings equation are likely to be correlated – and education is therefore "endogenous".
- Angrist and Krueger (1991). Uses "quarter of birth" as an "instrument" for education.
- Individuals born in the 4th quarter of a calendar year tend to stay in school longer than do individuals born in the 1st quarter of the year – they are either in earlier grades and need longer to complete the mandatory amount of school (in many states), and/or they turn 16 later and hence are legally compelled to stay in school longer than 1st quarter individuals
- So, if birth quarter can be assumed to be "randomly" determined (or "as if" randomly determined), it can be used as an instrument for educational attainment so long as birth quarter is not *directly* related to earnings (i.e., in ways unrelated to its effect on earnings through additional education)

- Classic Example II: Does military service affect earnings?
- Omitted variables/endogeneity problems: the kinds of individuals who select into military service may have different earning potentials than individuals who do not, and these differences may be unobserved
- Angrist (1990): uses Vietnam-era draft lottery number as an instrument for military service
- People with (randomly) low lottery numbers needed to serve, people with (randomly) high lottery numbers could avoid service; so this is an *exogenously-induced* change in military service. It is also unlikely that lottery number status had a *direct* impact on earnings – why should it?
- Finding: Military service is *negatively* correlated with earnings
- Applied in political science by Stoker and Erikson (2012 *APSR*); positive effects of having a low lottery number on antiwar political attitudes, Democratic party identification, etc. via "draft vulnerability".

- IV Estimation I: Two Stage Least Squares
  - use OLS estimation at two stages to generate estimates
  - First stage: regress X_1 against Z and all exogenous covariates and generate predicted X_1, which, by construction, is uncorrelated with U since all of the variables generating it are unrelated to U

$$X_1 = \pi_1 Z_1 + \beta_k X_k + \mathrm{U}_1$$

$$\hat{X}_1 = \pi_1 Z_1 + \beta_k X_k$$

  - Second stage: regress Y against predicted X and the exogenous covariates. All variables in this equation are unrelated to U and ε

$$Y = \hat{\beta}_1 \hat{X}_1 + \beta_k X_k + \varepsilon$$

- IV Estimation II: "Control Function" Regression, e.g. "Two Stage Residual Inclusion" (2SRI)
- First stage: regress X on Z and the exogenous covariates, and use those estimates to generate a predicted residual U* which by construction is the potentially *endogenous* portion of X (i.e. unrelated to all the things that are truly exogenous)

$$X_1 = \pi_1 Z_1 + \beta_k X_k + \mathrm{U}_1$$

$$\hat{X}_1 = \pi_1 Z_1 + \beta_k X_k$$

$$\hat{U}_1 = X_1 - \hat{X}_1$$

- Second stage: regress Y against predicted U, X and the exogenous covariates

$$Y = \beta_1 X_1 + \beta_2 \hat{U} + \beta_k X_k + \varepsilon$$

- Including predicted U as a "control" allows the β_1 effect for X to represent the "true" effect of the exogenous portion of X

Endogeneity in Models with Non-Continuous Variables

- So far, the exposition has been applied to continuous variables – both the endogeneous regressor and outcome
- How can these models and ideas apply to models with non-continuous variables?
- One application from our censoring discussion: we used a "control function" regression to model the effect of a variable on censored Y. We showed that X (education) was related to the probability of being censored on Y (tolerance), so *low educated* individuals had to have a large error term in the Y* equation in order not to be censored; therefore X was related to the disturbance term in censored Y.

- What did we do? **A variation of Two Stage Residual Inclusion (2SRI)**
  - Step 1: modeled the probability of being censored
  - Step 2: calculated a predicted Inverse Mills Ratio, or the instantaneous probability of being censored
  - Step 3: Included the predicted IMR as a variable in the outcome equation for continuous Y, where it controlled for the difference between E(Y*|X) – the uncensored true value of the DV, and E(Y|X), the censored DV, which allowed for unbiased estimation of the effect of X on Y*

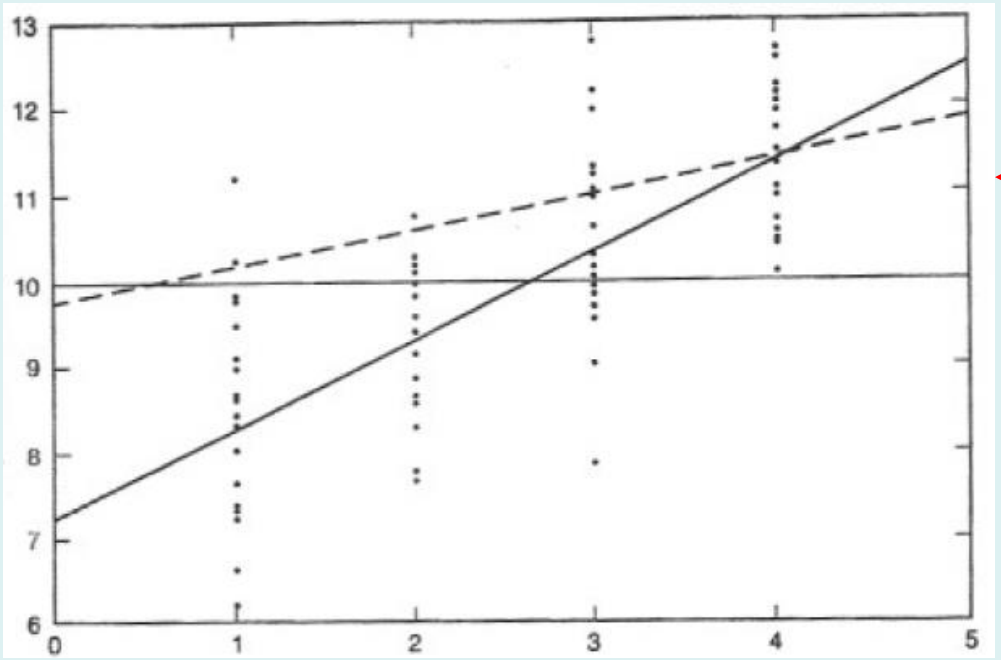
- Extensions: a) models with endogenous sample selection; b) endogenous treatment effects, c) dichotomous outcomes with endogenous regressors of all kinds

Endogenous Sample Selection Models

- An important problem related to the discussion of the censored regression model occurs when you observe a truncated sample not because of some intrinsic measurement process, but because a case is observed **only** when some other variable passes a given threshold – that is, the sample itself is non-random and the observation of Y depends on some outside factors.
- Example: Does the prestige of graduate school attended predict publication success after PhD? You gather a sample of assistant professors and measure graduate school prestige and publications.
- What is wrong with this? It ignores the **non-random** nature of the sample – that is, people who have assistant professor jobs may have come from better schools, and thus the sample as a whole is composed of disproportionately prestige-school type individuals who are also all likely to publish -- but within the sample itself the effect of prestige is not so strong
- Ignoring the **"endogenous sample selection"** process will produce erroneous results!

- This is a ubiquitous problem in social science --- whenever the sample is not random, and whenever the biases in the sample are related to the dependent variable, this is big trouble for estimating causal effects!!
- Example: Effect of education on women's wages – examine a sample of working women and see how variations in education lead to difference in wages. What is wrong? Ignores the facts that: working women are a non-random sample of all women; that women who choose to work may be the kinds of women who would make more money anyway; and that education makes it more likely that you would be a working woman
- Again, factor you are interested in (education) is related to the selection into the sample, and being in the sample makes you more likely to be high on the D.V. in the first place – so you will tend to underestimate the true effect of education on the process
- Examples: Is there a positive effect of democracy on winning wars? Selection problem: democracies may be less likely to engage in unwinnable wars

*Y Axis: Publications



Selected Sample: Assistant Profs

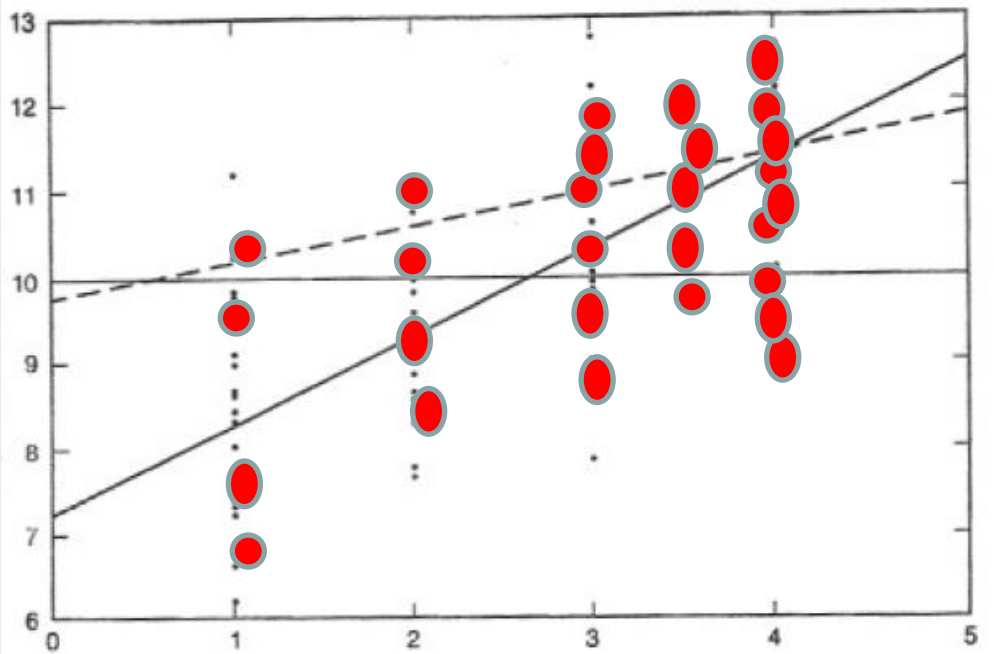
Not Selected Sample: Other PhDs

X Axis: Prestige of Graduate School

Can see the problem here

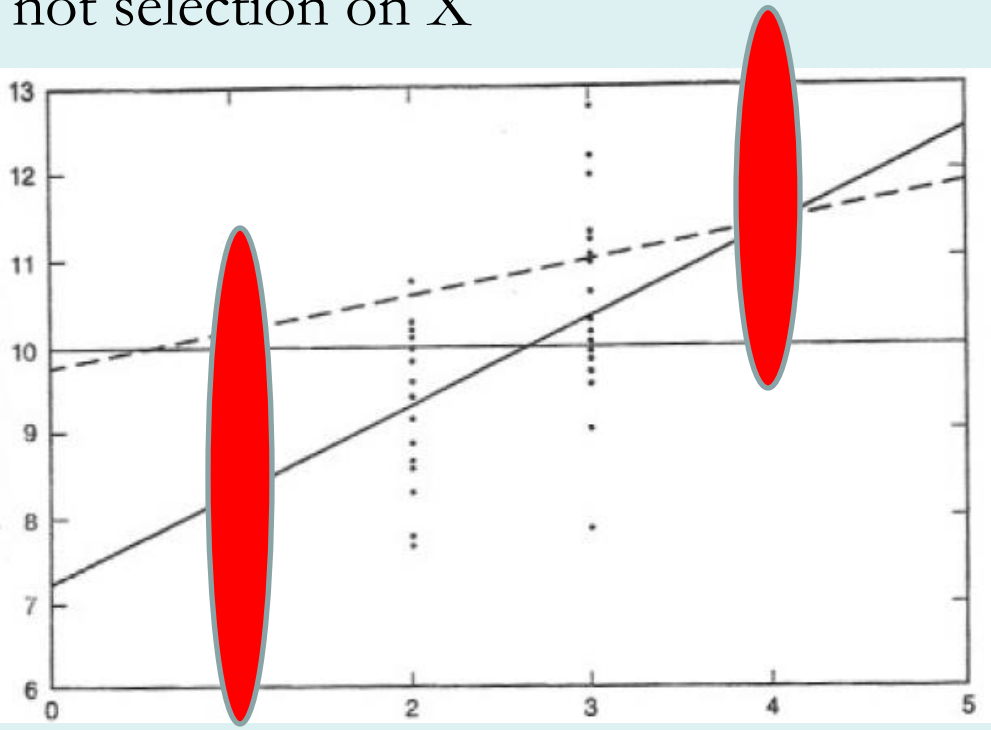
- Strong relationship between X and Y* among all PhDs
- Strong relationship between X and P(Observed Sample): more prestigious grad school PhD are more likely to be Assistant Profs
- Among Assistant Profs, weaker relationship between prestige of grad school and publications

- So problems whenever selection into the sample is related to Y* **and** X is related to Y* and to probability of selection
- If either condition doesn't hold, not a problem
- For example, prestige of grad school could be related to having an Assistant Professor job, but having an Asst Professor job is unrelated to publication output – in that case selecting only Asst Profs means we have a non-random sample but no bias on Y* and can still get accurate effects



If red dots are Assistants and you sample the red dots, no problem! You have a non-random sample but being in the sample is not related systematically to Y* . Still recover the true X->Y* effect

- Similarly, selection on X is not necessarily bad, insofar as you get a random sample of different values of X. Let's say we took a random sample of All PHD from low prestige schools and all PHDs from excellent schools, ignored the middle range – We select on X but this would have no effect on our results. This is why selection on Y* is the big issue, not selection on X

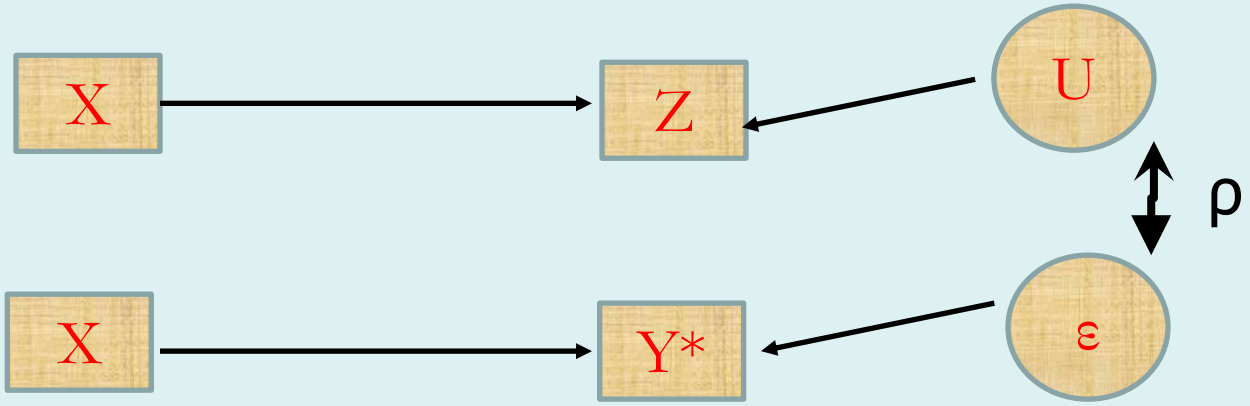


Red ovals: selected sample. Contains both low and high prestige grad schools, and you still recover the true X-Y* effect

- And of course, if X has no effect on the probability of being selected into the sample, then there is no problem either. In our case, if high prestige grad school PhDs are no more likely to be Assistant Professors than low grad school PhDs, no problem.
- And if X unrelated to Y*, obviously no problem either. If prestige of grad school unrelated to publications, then non-random subset of selected units unlikely to show it

- **WHEN THERE IS A NON-RANDOM SAMPLE THAT IS RELATED TO Y*, AND X IS RELATED TO SELECTION INTO THE SAMPLE AND TO Y* -- THEN YOU HAVE A BIG PROBLEM**

- Multi-equation way to look at the problem: one "selection" equation predicting whether a unit will be in the observed sample (Z); another "outcome" equation predicting the dependent variable Y*



- In the selected sample, X and U are related, because a unit is either in the sample due to X (which means low U), or due to an unusually high U at low levels of X. So X and U are *negatively* related in the selected sample. And if the selection is related to Y*, then U and ε are related, and X and ε are (negatively) related
- So we can't regress Y* on X in the selected sample without correcting for the U ε/X ε correlation. Classic endogeneity!!! X is related to the error term in its estimation equation

- As mentioned, similar to tobit analysis: when we model the conditional means of Y, given X, in the selected sample, we have *larger* Y than the true Y* at low levels of X because of selection (censoring in the tobit case). So OLS gives bias because the observed mean of Y at low levels of X will be higher than the true mean of Y*, just like in tobit.
- So we need to include a variable, as in two-step tobit, that stands for this bias in E(Y*). X will be negatively related to this bias term and so including it in the outcome equation will correct the coefficient on X to where is "should be" in terms of its effect on Y. We need to add a term that represents the difference in Y-mean from Y*-mean, based on the likelihood of being in the observation sample
- Moving now to multi-equation analysis, where we have a selection and outcome equation and possible correlation between error terms in the two equations to handle selection biases.
- This is the (Nobel prize winning economist James) Heckman's two-step sample-selection model

Sample Selection Model

Latent

$Y^* = XB + \varepsilon$

$E(\varepsilon \mid X) = 0, E(\varepsilon^2) = \sigma^2$

Selection

$z^*_i = W\gamma + u$

$z = 1$ if $z^* > 0$

$z = 0$ otherwise

Observed Outcome

$Y = Y^* = XB + \varepsilon$ if $z=1$

Y not observed if $z=0$

Problem: if selection and outcome are related, then corr(u, ε) not equal to 0. What need to do is somehow get a model that captures true β by taking into account that correlation, which we have referred to as "rho" or ρ .

Note: this is a more general and potentially more theoretically appealing model than tobit, since the variables in **X** and **W** can differ (or not).

In tobit the same Xs that determine censoring are also included (automatically) as predictors of the outcome.

- To estimate the sample selection model:
- Assume that errors for z and y are normally distributed with covariance:

$$Y^* = XB + \varepsilon$$

$$z^*_i = W\gamma + u$$

$$\begin{pmatrix} \varepsilon \\ u \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_\varepsilon & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & 1 \end{pmatrix} \right]$$

- So the covariance between errors is ϱ*standard deviation of the outcome error equation (while variance of the selection equation is set to 1)
- Steps: 1) Estimate the selection equation via probit
  2) Model the expected value of Y in the outcome equation, conditional on being in the sample (Z*>0)
  3) Put the model in a form to be estimated via OLS

- Step 1: Estimate the selection equation via probit

$$\mathrm{P}(z_i = 1) = \mathrm{P}(z_i^* > 0) = P(u > -W\gamma) = P(\varepsilon < W\gamma) = \Phi(W\gamma)$$

$$\mathrm{P}(z_i = 0) = 1 - \Phi(W\gamma)$$

- Step 2: Estimate the conditional expectation of Y from the X, given that the unit is in the observed sample (z=1)

$$\begin{aligned}\mathrm{E}(Y \mid z = 1, X) &= X\mathrm{B} + \mathrm{E}(\varepsilon \mid z = 1, X) \\ &= X\mathrm{B} + \mathrm{E}(\varepsilon \mid u > -w\gamma) \\ &= X\mathrm{B} + \lambda\rho\sigma_\varepsilon\end{aligned}$$

$$\text{where } \lambda = \frac{\phi(w\gamma)}{\Phi(w\gamma)} = \text{"Inverse Mills Ratio"}$$

- In this case the IMR is the instantaneous probability of *exclusion* from the sample; as this gets larger the difference in the conditional expectation of Y* from Y (observed) gets larger and larger, subject to the size of the correlation between the two equations' error terms (ϱ)

- So: Expectation of Y is XB, the true place it should be, plus a term which is based on a) the likelihood of exclusion from the sample; and b) the degree of correlation between the outcome and the selection equations
- IMR: height of the normal curve for case relative to exclusion point, divided by cumulative probability of inclusion
- In our example, at low values of X (high probability of exclusion), we will see high IMR, and if there is a high ϱ, will mean the Expected Y and Expected Y* differ by a lot, pulling the OLS line downwards. This new term corrects for it.
- Suggests a two-step estimation strategy, known as "Heckman Selection Model" or "Heckit"
- Step 3: **HECKIT** 1) Probit for selection, generate IMR for each case, and include it in a second step regression as control for "omitted variable", i.e., the probability of exclusion from the observed sample

- This is exactly like Tobit, but not necessarily based on X itself. Rather, a potentially different set of coefficients influencing Z (selection) and outcome (X), though the variables in each set can overlap

- It is desirable to have variables in the W set for the selection equation that are **not** in X; this identifies the model in the same way as instrumental variables do so in linear IV regression

- This is not **technically necessary** – but if W and X are the same then you are relying solely on the non-linearities of probit to identify the model. In a linear model with the same variables in X and W, X and the IMR would be perfectly collinear, so impossible to estimate. In the non-linear probit, this will not be the case, BUT without excess W variables, X and IMR could be very highly related in a given sample so that estimation will prove difficult due to the multicollinearity

- ML Estimation of the Selection Model
- As with Tobit model, we can derive the likelihood function for the outcome by considering the contribution of the sample-selected and the non-selected cases. In this case it is more complicated due to the fact that the two equations' errors may be correlated with value ϱ
- For the non-selected cases

$$P(z^* < 0) = P(\varepsilon < -w_i\gamma) = 1 - \Phi(w_i\gamma)$$

$$\prod_{\text{non-selected}} = \prod(1 - \Phi(w_i\gamma))$$

$$\ln L(\gamma) = \sum_{\text{non-selected}} \ln(1 - \Phi(w_i\gamma))$$

- For the selected cases, the likelihood contribution is their probability of being selected multiplied by the height of the normal curve at XB, subject to the correlation between the outcome and selection equations

$$\mathrm{LnL}_{i\text{selected}}(\beta,\sigma^2,\rho) = \sum_{\text{selected}} \log\Phi\left(\frac{w_i\gamma + \rho\left(\frac{y_i - XB}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)$$

$$+ \sum_{\text{selected}}\left(\left[-\frac{1}{2}\ln 2\pi\sigma^2\right] + \left(\frac{y_i - XB}{\sigma}\right)^2\right)$$

- Complicated!! (But Heckman won the Nobel prize for this!)
- First term is the adjusted probability of selection (Φ); second term the height of the normal curve at XB
- Put this together with the LnL component from the non-selected cases from the previous slide and you arrive at the full likelihood function

- Interesting result if $\rho=0$, that is, if there is no correlation between the errors in the selection and outcome equations

$$\mathrm{LnL}(\gamma,\beta,\sigma^2) = \sum_{\text{non-selected}} \ln(1-\Phi(w_i\gamma)) + \sum_{\text{selected}} \Phi(w_i\gamma)$$

$$+ \sum_{\text{selected}} \left(\left[-\frac{1}{2}\ln 2\pi\sigma^2 \right] + \left(\frac{y_i - XB}{\sigma} \right)^2 \right)$$

- Top line is the standard probit result – each case contributes the likelihood that they were selected or not selected based on the W
- Bottom line is the OLS result in ML format – the height of the normal curve given the deviation of y from predicted Y from the X
- This shows that sample selection is *ONLY* problematic if the factors (observed or unobserved) predicting selection are also related to the outcome. That's what we showed earlier (maybe more intuitively)!
- Next steps: Treatment effects models: continuous and non-continuous endogenous treatments with continuous and non-continuous outcomes