Dis 13: Exogeneity in Dynamic Causal Analysis; Second Half Recap

1 Exogeneity in Dynamic Causal Analysis

- Shifting gears: Now we care about causal interpretation again
- **Dynamic causal effect**: The relationship between X and Y is causal, and it's manifested through time, instead of through different groups (treated group vs. control group).

	Static Causal Effect (Experiments & Quasi-Experiments)	**Dynamic Causal Effect**
Type of data	Cross-sectional (for experiments) Panel (for quasi-experiments)	Time series
Groups	Treatment vs. Control	Before policy period vs. After
In each group	Different individuals in each group	Same individual / object in each group

- Dynamic causal effect is usually studied using a **distributed lag model**:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \ldots + \beta_{r+1} X_{t-r} + u_t$$

  - For the βs to have correct estimates under OLS, there cannot be omitted variable bias.
  - We often specify two types of exogeneity for such dynamic model:
    - **Exogeneity** (or **past and present exogeneity** / **contemporaneous exogeneity**):

$$E[u_t | X_t, X_{t-1}, \ldots, X_1] = 0$$

    - **Strict exogeneity** (or **past, present, and future exogeneity**):

$$E[u_t | \ldots, X_{t+1}, X_t, X_{t-1}, \ldots, X_1] = 0$$

  - Relationship between the two:

$$\text{Exogeneity} \rightleftharpoons \text{Strict Exogeneity}$$

  - Why do we need two types of exogeneity definitions?
    - Exogeneity gives us correct βs to interpret causal effect caused onto Y_t
    - Strict exogeneity gives us correct βs to interpret causal effect caused onto Y_t, along with future periods of Y

2 Second Half Recap

2.1 Topics since midterm

(Note that the following list is not exhaustive)

- Topics related to cross-sectional data
  - Binary response
    - Run a linear probability model (OLS regression + robust SE)
    - Perform logit regression
    - Perform probit regression
    - Interpret each probability model's result
  - Instrumental variable (IV)
    - How IV is chosen (relevance + exogeneity)
    - Calculate $\hat{\beta}_{IV}$
    - Perform 2SLS / TSLS
  - Big data
    - Adjust variables for a big data prediction model (demean Y; standardize all X)
    - How to find $\hat{\beta}_{Ridge}$ and $\hat{\beta}_{Lasso}$ in a small sample
    - Conceptually, how principal components are selected
    - Conceptually, how to perform stepwise selection (see PS 10 Q3)
- Topics related to panel data
  - Clustered standard error
  - Run a fixed effect model
  - Run a first difference model
- Topics related to time series data
  - Use time series data for prediction purpose
    - Definition of autocorrelation
    - Check whether time series data is stationary (plot time series; use ACF or PACF; use Dickey-Fuller test for stochastic trend; use Chow or QLR test for breaks)
    - What to do if data is not stationary? (often times, try taking first difference of Y_t: $Y_t - Y_{t-1}$)
    - What model to use if data is stationary? ($AR(p)$, $ADL(p,q)$)
    - Determine number of lags to include in your model? (AIC, BIC, use of ACF plot)
    - Estimate MSFE (mean squared forecast error)
  - Use time series data to interpret dynamic causal effect
    - Exogeneity vs. Strict exogeneity

2.2 Study resources

- Discussion handouts (the final is cumulative, so don't just focus on materials after the midterm)
- Problem sets (for Stata data exercise practice)
- Practice final exam on Canvas (scroll to the bottom of homepage)

3 Problems

1. Consider the following prediction model:

$$Y_i = \beta_1 X_i + u_i$$

We have the following observations of Y and Xs (let's not standardize variables for this question):

Y	X
22	4
-11	-2
52	10

(a) What is the OLS estimate $\hat{\beta}_{1,OLS}$?

(b) What are the Ridge estimates of $\hat{\beta}_{1,Ridge}$ when $\lambda_{Ridge} = 2$?

(c) Consider an alternative model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Suppose that X is endogenous in this model. A valid instrument Z has been proposed for X, and it has the following values corresponding to each row of data in the table: 7, -1, 12. What is $\hat{\beta}_{1,IV}$?

(d) Write down the first and second stage models you'd estimate to obtain $\hat{\beta}_{1,2SLS}$ under 2SLS procedure. Does the value of $\hat{\beta}_{1,2SLS}$ equal to $\hat{\beta}_{1,IV}$? Does the standard error? Explain.

2. In this exercise, we are going to combine two sets of data in order to study the determinant of college GPA.

   (a) The two sets of data record college and high school performance of 141 individuals who are uniquely identified by the `id` column. Merge the two datasets into one.

   (b) Someone proposes that a student's class standing (i.e. whether a student is a freshman / sophomore/ junior / senior) affects their college GPA level, and thus wants to include all four variables `fresh`, `sopho`, `junior`, and `senior` in their regression analysis. Is this feasible? Explain.

   (c) Say that you settled on the following regression model:

$$\begin{aligned}\texttt{colgpa}_i = \beta_0 + \beta_1 \texttt{hsgpa}_i + \beta_2 \texttt{act}_i + \beta_3 \texttt{soph}_i + \beta_4 \texttt{junior}_i + \beta_5 \texttt{senior}_i \\ + \beta_6 \texttt{campus}_i + \beta_7 \texttt{greek}_i + \beta_8 \texttt{alcohol}_i + u_i\end{aligned}$$

   where

   - `colgpa` records college GPA level
   - `hsgpa` records high school GPA level (all in the same 0-4 scale)
   - `act` records student ACT score
   - `soph` $= 1$ if sophomore
   - `junior` $= 1$ if junior
   - `senior` $= 1$ if senior
   - `campus` $= 1$ if live on campus
   - `greek` $= 1$ if a member of greek society
   - `alcohol` records average number of days per week drinking alcohol

   Plot residual-fitted value plot. Is there serious concern about heteroskedasticity in this model?

(d) After running the regression model, you're concerned that male and female students might see different level of impact from their high school GPA level to their college GPA. How would you address that in your regression model? Is there actually a difference between male and female students' effects from high school GPA onto college GPA at 5% significance level?

(e) Instead of use the level of college GPA as your dependent variable, create a binary variable, `highcolgpa`, which equals to 1 if college GPA is greater or equal to 3.5. Rerun the model proposed in (c) with this new binary dependent variable. What is this type of model called?

(f) Interpret your estimate of β_1 from (e).

(g) What's the drawback of the model used in (e)? Is there any alternative model that we can consider running?