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:

$$\text{Exogeneity} \underset{\leftarrow}{\not\rightarrow} \text{Strict Exogeneity}$$

  - 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 $\hat{\beta}_{1,OLS}$?

OLS estimator is obtained by minimizing sum of squared residuals. This means we want to solve

$$\min_{\beta_1} \quad \sum_{i=1}^{3} (Y_i - \beta_1 X_i)^2$$

Taking first order condition with respect to β_1:

$$\begin{aligned} 2\sum_{i=1}^{3}(Y_i - \beta_1 X_i)(-X_i) &= 0 \\ \sum_{i=1}^{3}(X_i Y_i - \beta_1 X_i^2) &= 0 \\ \beta_1 \sum_{i=1}^{3} X_i^2 &= \sum_{i=1}^{3} X_i Y_i \\ \beta_1 &= \frac{\sum_{i=1}^{3} X_i Y_i}{\sum_{i=1}^{3} X_i^2} \\ \beta_1 &= \frac{4 \times 22 + (-2) \times (-11) + 10 \times 52}{4^2 + (-2)^2 + 10^2} = 5.25 \end{aligned}$$

Thus, $\hat{\beta}_{1,OLS} = 5.25$

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

Ridge estimator is obtained by minimizing penalized sum of squared residuals:

$$\begin{aligned} &\min_{\beta_1} \quad \sum_{i=1}^{3}(Y_i - \beta_1 X_i)^2 + \lambda_{Ridge} \sum_k \beta_k^2 \\ \Leftrightarrow &\min_{\beta_1} \quad \sum_{i=1}^{3}(Y_i - \beta_1 X_i)^2 + \lambda_{Ridge} \beta_1^2 \end{aligned}$$

Taking first order condition with respect to β_1:

$$2\sum_{i=1}^{3}(Y_i - \beta_1 X_i)(-X_i) + 2\lambda_{Ridge}\beta_1 = 0$$

Since $\lambda_{Ridge} = 2$,

$$\begin{aligned}
2\sum_{i=1}^{3}(Y_i - \beta_1 X_i)(-X_i) + 2 \times 2\beta_1 = 0 \\
2\beta_1 &= \sum_{i=1}^{3}(X_i Y_i - \beta_1 X_i^2) \\
2\beta_1 &= \sum_{i=1}^{3} X_i Y_i - \beta_1 \sum_{i=1}^{3} X_i^2 \\
\left(2 + \sum_{i=1}^{3} X_i^2\right)\beta_1 &= \sum_{i=1}^{3} X_i Y_i \\
\beta_1 &= \frac{\sum_{i=1}^{3} X_i Y_i}{2 + \sum_{i=1}^{3} X_i^2} \\
\beta_1 &= \frac{4 \times 22 + (-2) \times (-11) + 10 \times 52}{2 + 4^2 + (-2)^2 + 10^2} \approx 5.16
\end{aligned}$$

Thus, when $\lambda_{Ridge} = 2$, $\hat{\beta}_{1,Ridge} \approx 5.16$. This also shows how Ridge is a shrinkage estimator, given that $\hat{\beta}_{1,Ridge} < \hat{\beta}_{1,OLS}$ whenever $\lambda_{Ridge} > 0$.

(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 $\hat{\beta}_{1,IV}$?

IV estimator for a simple linear model with intercept is

$$\hat{\beta}_{1,IV} = \frac{\widehat{Cov}(Z_i, Y_i)}{\widehat{Cov}(Z_i, X_i)} = \frac{\sum_{i=1}^{3}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{3}(Z_i - \overline{Z})(X_i - \overline{X})} \qquad (*)$$

We need to first estimate $\overline{Z}$, $\overline{Y}$, and $\overline{X}$:

$$\overline{Z} = \frac{1}{3}\sum_{i=1}^{3} Z_i = \frac{1}{3} \times (7 + (-1) + 12) = 6$$

$$\overline{Y} = \frac{1}{3}\sum_{i=1}^{3} Y_i = \frac{1}{3} \times (22 + (-11) + 52) = 21$$

$$\overline{X} = \frac{1}{3}\sum_{i=1}^{3} X_i = \frac{1}{3} \times (4 + (-2) + 10) = 4$$

Plug these back in equation (∗):

$$\hat{\beta}_{1,IV} = \frac{\sum_{i=1}^{3}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{3}(Z_i - \overline{Z})(X_i - \overline{X})}$$

$$= \frac{(7-6)\times(22-21)+(-1-6)\times(-11-21)+(12-6)\times(52-21)}{(7-6)\times(4-4)+(-1-6)\times(-2-4)+(12-6)\times(10-4)} \approx 5.27$$

Thus, $\hat{\beta}_{1,IV} \approx 5.27$

(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.

The 2SLS approach first projects X onto Z, then uses the projected $\hat{X}$ in the second stage for estimation. This means that in the first stage, we estimate

$$X_i = \delta_0 + \delta_1 Z_i + v_i$$

which gives us estimated (projected) X:

$$\hat{X}_i = \hat{\delta}_0 + \hat{\delta}_1 Z_i$$

In the second stage, our model is

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

When estimating the second stage, the $\hat{\beta}_{1,2SLS}$ we obtain is exactly the $\hat{\beta}_{1,IV}$ we calculate from the first stage. However, using this manual 2SLS approach, we cannot trust the standard error produced for $\hat{\beta}_{1,2SLS}$, since our separate second stage regression does not take into account that $\hat{X}_i$ is an estimated object from the first stage.

2. In this exercise, we are going to combine two sets of data in order to study the determinant of college GPA. Download this week's dataset, and don't forget to change your Stata working directory to where your data is located.

   (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.

   See Do-file solution. The function needed is `merge`, but a slight caveat for this question is that both datasets are stored as comma delimited values (.csv). In order to use the `merge` function, we need to save at least one of the two datasets as Stata data format (.dta), so that the `merge` function can correctly call the second file to merge onto.

   (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.

   No, it's not feasible. The reason is that including all four variables create perfect multicollinearity (also known as dummy variable trap in this case). This is because the four variables together form the following linear relationship:

$$\texttt{fresh} + \texttt{sopho} + \texttt{junior} + \texttt{senior} = 1$$

   Here, these four variables are perfectly collinear with the constant term (the intercept) in our regression. To address perfect multicollinearity, we can include three out of the four variables, or remove the intercept term from our regression model.

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

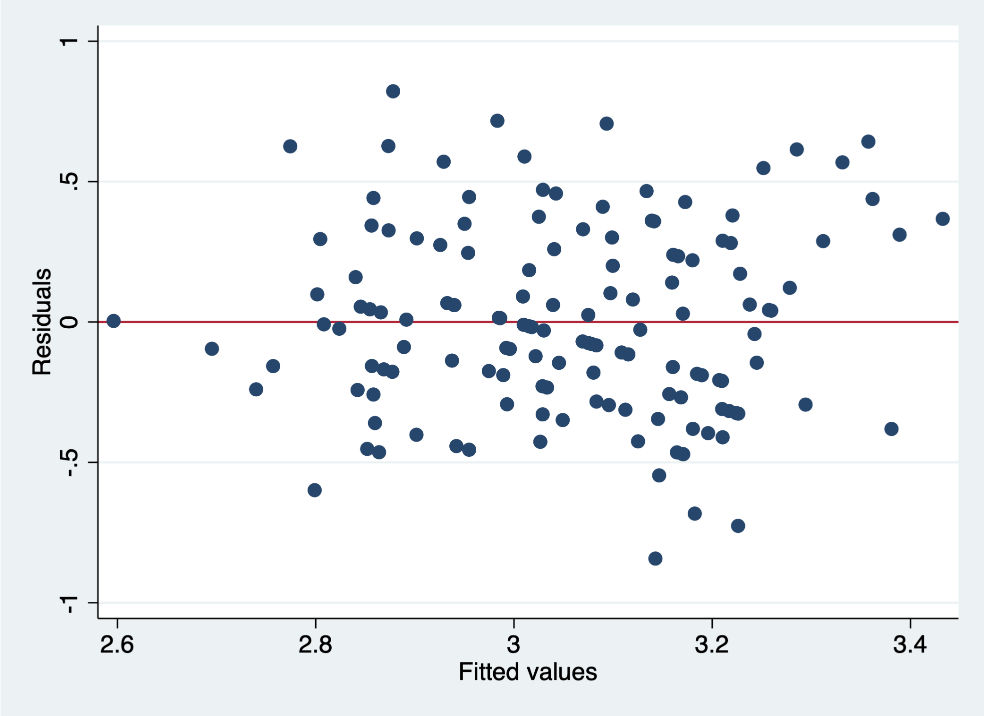
$$\begin{aligned}\texttt{colgpa}_i = \beta_0 + \beta_1\texttt{hsgpa}_i + \beta_2\texttt{act}_i + \beta_3\texttt{soph}_i + \beta_4\texttt{junior}_i + \beta_5\texttt{senior}_i \\ + \beta_6\texttt{campus}_i + \beta_7\texttt{greek}_i + \beta_8\texttt{alcohol}_i + u_i\end{aligned}$$

   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?

   See Do-file solution. The residual-fitted value plot looks like the following:



The residuals seem to be randomly distributed around mean 0, and variance of residuals is pretty much constant over all fitted values. This is consistent with homoskedasticity, so we do not have serious concern about heteroskedasticity in this case.

(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?

Interacting `hsgpa` with a binary variable `male` should address the concern. The model becomes

$$\begin{aligned}\texttt{colgpa}_i &= \beta_0 + \beta_1\texttt{hsgpa}_i + \beta_2\texttt{act}_i + \beta_3\texttt{soph}_i + \beta_4\texttt{junior}_i + \beta_5\texttt{senior}_i \\ &\quad + \beta_6\texttt{campus}_i + \beta_7\texttt{greek}_i + \beta_8\texttt{alcohol}_i + \beta_9\texttt{hsgpa}_i \times \texttt{male}_i + u_i\end{aligned}$$

Denote all terms between β_2 and β_8 as W. That is,

$$W = \beta_2\texttt{act}_i + \beta_3\texttt{soph}_i + \beta_4\texttt{junior}_i + \beta_5\texttt{senior}_i + \beta_6\texttt{campus}_i + \beta_7\texttt{greek}_i + \beta_8\texttt{alcohol}_i$$

Also denote all variables in the model in vector notation as X.

We can now write down the conditional mean of `colgpa` for male and female group:

$$\begin{aligned}E[\texttt{colgpa}_i|\texttt{male}_i = 0, X] &= \beta_0 + \beta_1\texttt{hsgpa}_i + W \\ E[\texttt{colgpa}_i|\texttt{male}_i = 1, X] &= \beta_0 + \beta_1\texttt{hsgpa}_i + W + \beta_9\texttt{hsgpa}_i \times 1\end{aligned}$$

This means that β_9 records the difference in effects from high school GPA onto college GPA between male and female students.

Estimating this model in Stata (see Do-file solution), the p-value of the interaction term is $0.868 > 0.05$. Thus, under 5% size, β_9 is not statistically significant, so there's likely not a difference in effect between male and female students from high school GPA onto college GPA.

(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?

See Do-file solution on how the binary variable is created. When we have a binary variable as the dependent variable, our model essentially becomes

$$\texttt{highcolgpa}_i = \beta_0 + \beta_1 \texttt{hsgpa}_i + W + u_i$$

where W is defined in the same way as the solution given for part (d). Consider taking conditional expectation on the new dependent variale,

$$E[\texttt{highcolgpa}_i|X] = \beta_0 + \beta_1 \texttt{hsgpa}_i + W$$

where X still denotes a vector of all independent variables in the model.

Since `highcolgpa` is binary,

$$\begin{aligned} E[\texttt{highcolgpa}_i|X] &= 0 \times Pr(\texttt{highcolgpa}_i = 0) + 1 \times Pr(\texttt{highcolgpa}_i = 1) \\ &= Pr(\texttt{highcolgpa}_i = 1) \end{aligned}$$

Thus, our model parameters are explaining the probability that `highcolgpa` equals to 1:

$$Pr(\texttt{highcolgpa}_i = 1) = \beta_0 + \beta_1 \texttt{hsgpa}_i + W$$

This type of model is called linear probability model. We discussed this style of model in Dis 9. Notice that when estimating a linear probability model, we always need to use the `robust` option in Stata, since the dependent variable being binary creates heteroskedastic error term (you can create residual-fitted value plot in here again to verify this claim).

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

From (e), we learned that model parameters are explaining the probability that `highcolgpa` equals to 1:

$$Pr(\texttt{highcolgpa}_i = 1) = \beta_0 + \beta_1 \texttt{hsgpa}_i + W$$

So β_1 is interpreted as the amount increased in the **probability** of having high college GPA when high school GPA increases by one unit.

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

Linear probability model's main drawback is that it doesn't constrain probability to be within 0 and 1. Alternative models include logit and probit, which map the fitted values from a linear probability model through a CDF (cumulative density function), and a CDF always has range between 0 and 1.

See Dis 9 handout for more information on how a logit and probit model is run.