Recitation 9

Seung-hun Lee

Columbia University

December 6th, 2021

Experiments in Economics

- You categorize some individuals under treatment group and controlled group and compare the differences across groups and times.
- You can do this in lab setting (randomized control trials) or find an exogenous event (natural experiments) which derives cross-sectional and time variation
- They provide a conceptual benchmark for assessing observational studies and can solve many validity threats in regular regressions.
- Do note that they have a validity threat of their own
- Further discussion of related topics - potential outcome frameworks, average treatment effect etc - are in the notes

Difference in Differences: Regression format

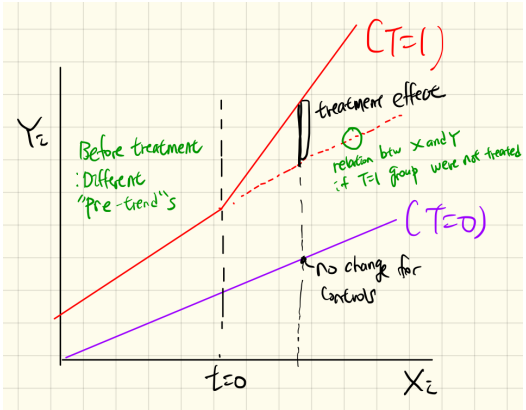
- Assume two time periods (before and after) and two groups (treated and controlled)
- You have the following regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 D_{it} + \beta_3 X_{it} D_{it} + u_{it}$$

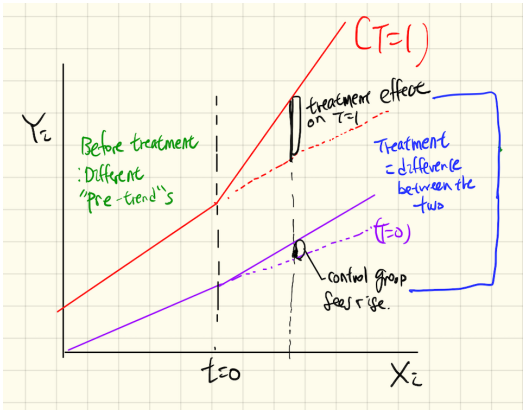
  where $X = 1$ if in treatment, $D = 1$ if after treatment
- You can get the following group averages
  - Treated, After: $E[Y_{it}|X_{it} = 1, D_{it} = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$
  - Treated, Before: $E[Y_{it}|X_{it} = 1, D_{it} = 0] = \beta_0 + \beta_1$
  - Controlled, After: $E[Y_{it}|X_{it} = 0, D_{it} = 1] = \beta_0 + \beta_2$
  - Controlled, Before: $E[Y_{it}|X_{it} = 0, D_{it} = 1] = \beta_0$
- Difference-in-differences $\hat{\beta}^{DD} = [\bar{Y}_{\text{treat,after}} - \bar{Y}_{\text{treat,before}}] - [\bar{Y}_{\text{control,after}} - \bar{Y}_{\text{control,before}}]$
  - In the above equation, we would be looking at $\hat{\beta}_3$
- Can be generalized to many time periods and many groups with TWFE, Event-studies etc (active area of research)

Difference in Differences: In pictures

Figure: DD estimator



(a) No change in control group



(b) Some change in control group

Big data

- It can simply mean data with large observations or large number of control variables.
- In general instances, atypical data such as text data, and satellite imagery are referred to as big data.
- In this class, we focus on the instance where there are many control variables, specifically on what to do if there are many control variables relative to the number of observations.
- Additionally, we focus on trying to get the best way to predict a result that is currently not in the dataset.

Characterizing MSPE

- Regression format is
$$Y_i^* = \beta_1 X_{1i}^* + ... + \beta_k X_{ki}^* + u_i^*$$
- X_{1i}^* is the "standardized" version of X_{1i}'s, Y_i^* the "demeaned" version.
- Note that we CANNOT have a constant term β_0 here because if we demean β_0, which is the same for all i's, they vanish.
- MSPE: the expected value of the squared error made by predicting Y for an observation not in the dataset.
$$MSPE = E[Y^{OS} - \hat{Y}^{OS}]^2$$
  - $\hat{Y}^{OS}$: obtained from the coefficients of β's made from the in-sample
  - Y^{OS} : realized value of Y outside of the sample

Characterizing MSPE

- With these notations, we can define the prediction error as

$$Y_i^{OS} - \hat{Y}_i^{OS} = (\beta_1 - \hat{\beta}_1)X_{1i}^{OS} + ... + (\beta_k - \hat{\beta}_k)X_{ki}^{OS} + u_i^{OS}$$

- Define $\sigma_u^2 = E[u_i^{OS}]^2$, then we can write MSPE as

$$MSPE = \sigma_u^2 + E[(\beta_1 - \hat{\beta}_1)X_{1i}^{OS} + ... + (\beta_k - \hat{\beta}_k)X_{ki}^{OS}]$$

- Oracle prediction: the smallest possible MSPE, σ_u^2
- However, we cannot predict β's perfectly.
- The more predictors we have, we generally end up having larger MSPE $\rightarrow$ need to reduce X's.
- Need Ridge, LASSO, and Principal Component method comes in

Methods

- Goal: Find other estimators that does not increase the MSPE compared to the rate in which the same rises in OLS estimators.
- Idea: Reduce the σ_u^2, the variance from the residual sums of squares, at the expense of introducing a small bit of bias.
- How: Providing a penalty for having a model with large number of regressors (what we formally call 'shrinkage').

Ridge

- Minimize a 'penalized' sum squared of residuals

$$\hat{\beta}_{Ridge} = \arg \min_{\beta_1,..,\beta_k} \left[\sum_{i=1}^{n} (Y_i - \beta_1 X_{1i} - .... - \beta_k X_{ki})^2 + \lambda_{Ridge} \sum_{j=1}^{k} \beta_j^2 \right]$$

- $\lambda_{Ridge} \sum_{j=1}^{k} \beta_j^2$: penalty for complexity.
- Introduce bias so that the variance term σ_u^2 will be reduced.
- Variance and the bias in MSPE moves in a trade-off relation
- Ridge estimator minimizes MSPE by reducing the variance term to the extent that the bias term does not rise too drastically.

LASSO

- Penalty term takes a different form:

$$\hat{\beta}_{LASSO} = \arg \min_{\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n} (Y_i - \beta_1 X_{1i} - .... - \beta_k X_{ki})^2 + \lambda_{LASSO} \sum_{j=1}^{k} |\beta_j| \right]$$

- The difference between the two lies in the degree of shrinkage.
  - When the OLS estimates are small, the LASSO shrinks those estimates all the way to 0.
  - Ridge also shrinks those coefficients close to 0, they do not exactly set them to 0.

Principal component

- You are using a linear combination of some subset of k variables so that you end up with $p < k$ number of regressors ('collapsing' the model)
- Solve the following problem to get the j'th principal component PC_j

$$\max var\left(\sum_{i=1}^{K} a_{ji}X_i\right) \text{ s.t. } \sum_{i=1}^{k} a_{ji}^2 = 1$$

  with another condition being that $corr(PC_j, PC_{j-1}) = 0$
- Solve the *maximization*: We want the X's to explain more of the variation
- $\sum_{i=1}^{k} a_{ji}^2 = 1$: Regularization method
- $corr(PC_j, PC_{j-1}) = 0$: We want to minimize the overlapping amount of information across different principal components.