

# Recitation 10: Big data and quasi-experiments in econometrics

Seung-hun Lee

Columbia University  
Undergraduate Introduction to Econometrics Recitation

December 1st, 2022

# Big data

# 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 (Mean Squared Prediction Error)

- Regression format is

$$Y_i^* = \beta_1 X_{1i}^* + \dots + \beta_k X_{ki}^* + u_i^*$$

- $X_{1i}^*$  is the "standardized" version of  $X_{1i}$ 's,  $Y_i^*$  the "demeaned" version.
- Note that we CANNOT have a constant term  $\beta_0$  here because if we demean  $\beta_0$ , 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  $\beta$ 's made from the in-sample
- $Y^{OS}$ : realized value of  $Y$  outside of the sample

## Characterizing MSPE

(brief)

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

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

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

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

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

elastic net

# Methods

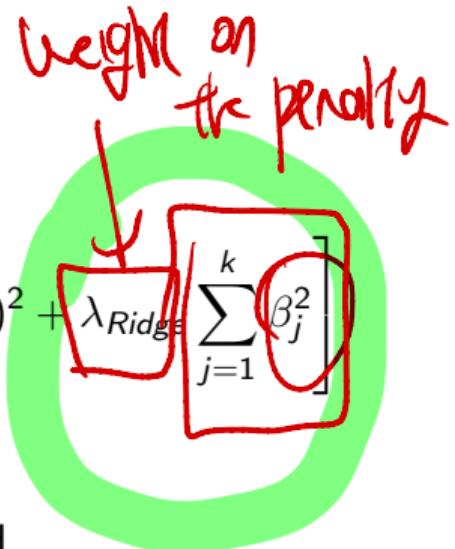
- Goal: Find other estimators that does not increase the MSPE compared to the rate in which the same rises in OLS estimators.
- Idea: Reduce the  $\sigma_u^2$ , the variance from the residual sums of squares, at the expense of introducing a small bit of bias.
- How: Providing a penalty for having a model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

# Ridge

- Minimize a 'penalized' sum squared of residuals

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

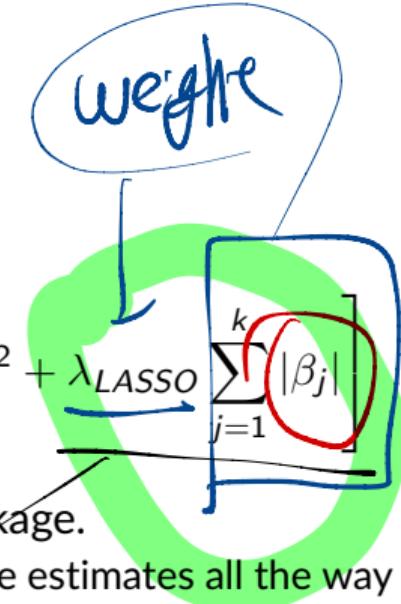
- $\lambda_{Ridge} \sum_{j=1}^k \beta_j^2$ : penalty for complexity.
- Introduce bias so that the variance term  $\sigma_u^2$  will be reduced.
- Variance and the bias in MSPE moves in a trade-off relation
- Ridge estimator minimizes MSPE by reducing the variance term to the extent that the bias term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



# LASSO

C Pulls more coeff. 0

- Penalty term takes a different form:

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


- 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.

Elastic Net  $\alpha \hat{\beta}_{LASSO} \sum |\beta_j| + (1-\alpha) \hat{\beta}_{Ridge} \sqrt{\sum \beta_j^2}$

# 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  $PC_j$



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

with another condition being that  $\operatorname{corr}(PC_j, PC_{j-1}) = 0$

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

# quasi-experiments in econometrics

# Experiments in Economics

## Motivation

- 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)
- 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

# Potential Outcome Framework

$$\underline{Y_{it}(1) - Y_{it}(0)}$$

## Setup

- Start by assuming that the treatment effect is identical for everyone (**constant treatment effect assumption**) ~~SUTVA~~
- Let  $Y_{i,t}$  be the **observed** outcome variable for individual  $i$  at time  $t$
- $X_{it}$  be the treatment variable: It is 1 if individual  $i$  is treated and 0 if otherwise.
- Let  $Y_{it}(0)$  denote a **potential outcome** if subject  $i$  is not treated at time  $t$  and  $Y_{it}(1)$  be the same if  $i$  is treated.
- Then  $Y_{it}$  can be split into

$$\begin{aligned} \rightarrow Y_{it} &= Y_{it}(1)X_{it} + Y_{it}(0)(1 - X_{it}) \\ &= Y_{it}(0) + (Y_{it}(1) - Y_{it}(0))X_{it} \end{aligned}$$

# Potential Outcome Framework

Two takeaways

↙ Cameron Trivedi says?

select few on  
unconfoundable

- **Fundamental problem of missing data:** We can only observe at most one of  $Y_{it}(1)$  and  $Y_{it}(0)$ , since individual  $i$  cannot be treated and untreated simultaneously at time  $t$ 
  - ▶ For us to make statements about the treatment effect, we need to be sure about what the missing outcome looks like.
  - ▶ Perfect randomization, randomization conditional on observables, instrumental variables on treatment variable  $X_{it}$ ...
- **Average treatment effect:** Average treatment effect is defined as

$$ATE = E[Y_{it}(1) - Y_{it}(0)] = \frac{1}{N} \sum_{i=1}^N (Y_{it}(1) - Y_{it}(0))$$

which is obtained from the coefficient of the  $X_{it}$  variable of the regression

↖ under random assignment  $Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$

# Average Treatment Effect

$$(Y_{it}(1), Y_{it}(0)) \perp X_i$$

Regressional form

- We will assume that the experiment is **perfectly randomized**
  - The treated individuals and controlled individuals are identical except for treatment status, or that  $E[u_{it}|X_{it}] = 0$  for all possible  $X_{it}$  values.
- Derive the expected value of  $Y_{it}$  separately for the treated and controlled individuals.  
For the controlled ( $X_{it} = 0$ , so that  $Y_{it} = Y_{it}(0)$ ):

$$E[Y_{it}|X_{it} = 0] = E[Y_{it}(0)] = \beta_0$$

and for the treated, ( $X_{it} = 1$ , so that  $Y_{it} = Y_{it}(1)$ )

*y<sub>it</sub> = β<sub>0</sub> + β<sub>1</sub>x<sub>it</sub> + u<sub>it</sub>*

$$E[Y_{it}|X_{it} = 1] = E[Y_{it}(1)] = \beta_0 + \beta_1$$

$$E(Y_{i|c} | X_i = 1) \quad \text{under RA.}$$

↓ P. 6

$$E(X_i Y_{i|c}) + (1 - X_i) E(Y_{i|c}(0) | X_i = 0)$$

↓  $X_i = 1$

RA

$E(Y_{i|c}(1) | X_i)$        $E(Y_{i|c}(0))$

Not observed for untreated

$$E(Y_{i|c} | X_i = 0)$$

↓

$E(Y_{i|c}(0) | X_i = 0)$        $E(Y_{i|c}(1))$

Not observed for treated

## Average Treatment Effect

$$E(Y_{it}) = E(Y_{it}|X_{it}=1)$$

Regressional form

- Then, the average treatment effect can be characterized as

$$ATE = E[Y_{it}(1) - Y_{it}(0)] = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

where subtraction among  $E[Y_{it}|X_{it} = 1]$  and  $E[Y_{it}|X_{it} = 0]$  is possible since we are assuming perfect randomization.

- Under perfect randomization, identifying the average treatment effect is equivalent to obtaining the  $\beta_1$  coefficient through an OLS process.
- However this is possible in rare circumstances...

# Validity Threats

## Validity threats

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 W_{it} + \epsilon_{it}$$

bad control variable

- We cannot automatically subtract  $E[Y_{it}|X_{it} = 1]$  and  $E[Y_{it}|X_{it} = 0]$  under **imperfect randomization**.
- If the **attrition rate is nonrandom** - if it differs by a treatment status - we end up with a biased estimate of the treatment effect. **(selection bias)**
- In terms of experimental procedure, **failure to comply to experiment protocol** and other **experimental effects** that rises through peculiar behaviors of the experimental and the experiment subject could also serve as a validity threat.
- Constant treatment effect assumption that we have imposed on ourselves may not be accurate.  
**Resafe Causality**
- There are **external validity** threats when the sample is not representative, is not replicable in other settings, and the results may have general equilibrium effects.

# Validity Threats: Unconfoundedness

- If the treated and the controlled differ in some observed characteristics, we can simply include control variables  $W_{it}$
- We assume that
  - ▶ Unconfoundedness:  $(Y_i(1), Y_i(0)) \perp X_i | W_i$
  - ▶ Overlap:  $0 < \Pr(X_i = 1 | W_i = w) = p(W_i = w) < 1$  (This is the propensity score)
- ATE can be identified using the fact that (separate page)  $\rightarrow$  *W/o assuming regression Model*

$$E\left[\frac{X_i Y_i}{p(W_i)}\right] = E[Y_i(1)], E\left[\frac{(1 - X_i) Y_i}{1 - p(W_i)}\right] = E[Y_i(0)] \implies E[Y_i(1) - Y_i(0)] = E\left[\frac{(X_i - p(W_i)) Y_i}{p(W_i)(1 - p(W_i))}\right]$$

and using the sample analogues (propensity score estimated using probit/logit)

- For ATT, we can use

$$E[Y_i(1) - Y_i(0) | X_i = 1] = E\left[\frac{(X_i - p(W_i)) Y_i}{\rho(1 - \rho)}\right]$$

where  $\rho$  is the overall probability of treatment (not controlling for observables)

Note ( $Y_i = X_i Y_i(1) + (1-X_i) Y_i(0)$ )

$$X_i Y_i = X_i - X_i Y_i(0) + X_i (1-X_i) Y_i(1)$$

0 anyway

$$= X_i Y_i(1) \quad X_i \neq 0$$

$$(1-X_i) Y_i = (1-X_i) X_i Y_i(0) + (1-X_i) (1-X_i) Y_i(1)$$

anyway

$$= (1-X_i) Y_i(0)$$

$$\begin{aligned} E\left[\frac{X_i Y_i}{P(W_i)}\right] &= E\left[E\left[\frac{X_i Y_i(1)}{P(W_i)}|W_i\right]\right] \\ &= E\left[\frac{E(X_i|W_i) E(Y_i(1)|W_i)}{P(W_i)}\right] \\ &= E(E(Y_i(1)|W_i)) \quad \text{V.I.E.} \\ &= E(Y_i(1)) \end{aligned}$$

$P(W_i) = \Pr(X_i=1|W_i)$   
 $= E(X_i|W_i)$

Similar for  $E\left(\frac{(1-X_i) Y_i}{P(W_i)}\right) = E(X_i(0))$

# IV in treatment effects (brief)

## Regression

- Let  $Z_{it}$  be an instrument for  $X_{it}$  (An exogenous variation dictating treatment assignment) and apply 2SLS

Puts weights  
on latent participation

$$X_{it} = \pi_0 + \pi_{1i}Z_{it} + e_{it} \text{ (First stage)}$$

$$Y_{it} = \beta_0 + \beta_{1i}X_{it} + u_{it} \text{ (Equation of interest)}$$

- Obtain  $\hat{X}_{it} = \hat{\pi}_0 + \hat{\pi}_{1i}X_{it}$  - the predicted probability of being treated.
- If  $\beta_{1i}, \pi_{1i}$  are independent of  $(u_{it}, v_{it}, Z_{it})$ ,  $E(\pi_{1i}) \neq 0$ , then

$$\hat{\beta}_1 \xrightarrow{p} \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})}$$

- Ultimately,  $\hat{\beta}_1$  can be written as

$$\hat{\beta}_1 \xrightarrow{p} E[\beta_{1i}] + \frac{cov(\beta_{1i}, \pi_{1i})}{E(\pi_{1i})} = \beta_1 + \frac{cov(\beta_{1i}, \pi_{1i})}{E(\pi_{1i})}$$

- Treatment effect is large for individuals for whom the effect of the instrument is large.

Local average treatment effect

## Difference in Differences: Regression format

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

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

where  $X = 1$  if in treatment,  $D = 1$  if after treatment *(Group Time)*

- 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:  $E[Y_{it}|X_{it} = 0, D_{it} = 0] = \beta_0$

- 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  $\hat{\beta}_3$

## Difference in Differences: Regression format

- Other way around is to difference the  $Y$  and get

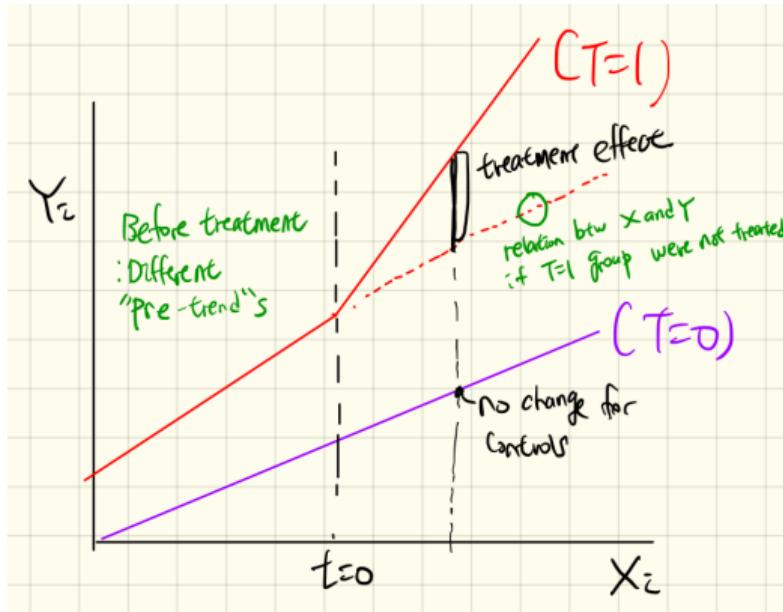
$$\Delta Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

where  $X = 1$  if in treatment

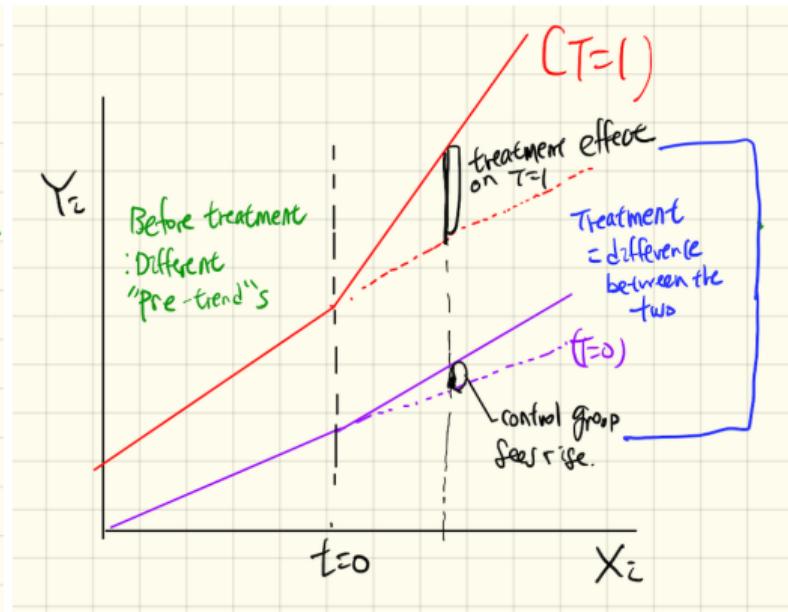
- You can get the following group averages
  - ▶ Outcome difference, treatment:  $E[\Delta Y_{it}|X_{it} = 1] = \beta_0 + \beta_1$
  - ▶ Outcome difference, control:  $E[\Delta Y_{it}|X_{it} = 0] = \beta_0$
  - ▶ In the above equation, we would be looking at  $\hat{\beta}_1$  to get the treatment difference
- 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

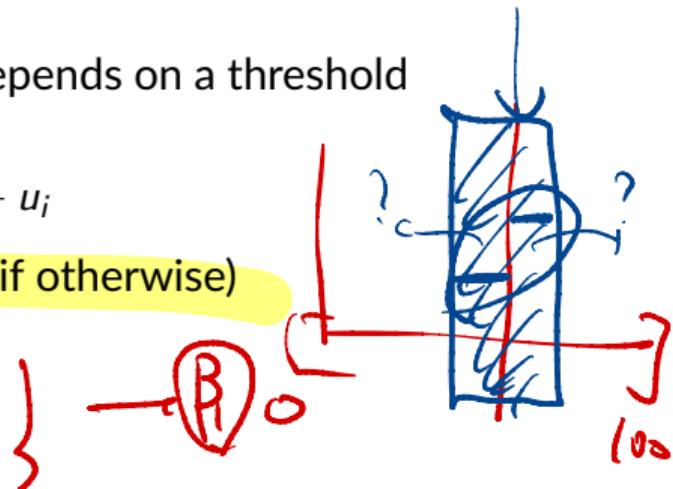
# Regression discontinuity

- This applies to case where the treatment status depends on a threshold
- Loosely speaking, we have

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

where  $X_i = 1$  if  $i$  passes threshold for eligibility (0 if otherwise)

- Back out the treatment effect using the difference
  - ▶ For those not in the program,  $E[Y_i | X_i = 0] = \beta_0$
  - ▶ For those in the program,  $E[Y_i | X_i = 1] = \beta_0 + \beta_1$
- Thus,  $\beta_1$  is the treatment effect
- Note: For RD to work, there should be no 'jump' in observables around the threshold except treatment status
- Loss of samples: RD is applied to a narrow bandwidth around the threshold
- There may be cases where threshold is not strict (fuzzy RD)



## Regression discontinuity: In pictures

