EC4305 Applied Econometrics

Lingjie, November 7, 2021

Objective: understanding causal relationships between variables.
Require: dataset with external shock to determine the causal impact
Example: Do hospitals make people healthier?

Acronyms

Y_i	: outcome for individual i
D_i	: treatment of individual i
Y_{1i}	: outcome for individual i when given treatment
Y_{0i}	: outcome for individual i when not given treatment
ATE	: $E[Y_{1i} - Y_{0i}]$
	: Average Treatment Effects
TOT	: $E[Y_{1i} - Y_{0i} \vert D_i = 1]$
	: Treatment effect Of the Treated
ITT	: $E(Y \vert$ compliers & non-compliers$) - E(Y \vert$ untreated$)$
	: Intent to Treat
counterfactual	: $E[Y_{0i} \vert D_i = 1]$
	: (unobserved) outcome for individual in an alternative universe without treatment
selection bias	: $E[Y_{0i} \vert D_i = 1] - E[Y_{0i} \vert D_i = 0]$
	: bias due to omitted variable
observed difference	: $E[Y_{1i} \vert D_i = 1] - E[Y_{0i} \vert D_i = 0]$
	: observed difference in average outcome

Causal Inference

Outcome:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i, \ D_i \in \{0, 1\}$$

Where $Y_{1i} - Y_{0i}$ is the causal effect for individual i

Regression results

RCT

$$Y_i = \alpha + \beta \cdot D_i + \epsilon_i$$

$\alpha = E[Y_i|D_i = 0]$
$\beta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$

```
reg Y X
```

FE

Note: $\lambda_t = \sum_m \rho_m \cdot I_m$

$$Y_{it} = \lambda_t$$

$\lambda_t = E[Y_{it}|t]$

$$Y_{it} = \alpha + \lambda_t$$

$\alpha = E[Y_{it}|t = \text{base}]$
$\lambda_t = E[Y_{it}|t] - E[Y_{it}|t = \text{base}]$

Dynamic Effect

Note: $\lambda_t \cdot D_i = \sum_m \rho_m \cdot D_i \cdot I_m$
(causal effect is same with or without constant)
$\lambda_t \cdot D_i = E[Y_{it}|t, D_i = 1] - E[Y_{it}|t, D_i = 0]$

$$Y_{it} = \lambda_t + \lambda_t \cdot D_i$$

$\lambda_t = E[Y_{it}|D_i = 0]$

$$Y_{it} = \alpha + \lambda_t + \lambda_t \cdot D_i$$

$\alpha = E[Y_{it}|t = \text{base}, D_i = 0]$
$\lambda_t = E[Y_{it}|D_i = 0] - E[Y_{it}|t = \text{base}, D_i = 0]$

Time Trend

$$Y_{it} = \alpha + \beta Post_t$$

$\alpha = E[Y_{it}|Post_t = 0]$
$\beta = E[Y_{it}|Post_t = 1] - E[Y_{it}|Post_t = 0]$

$$Y_{it} = \alpha + \beta D_i + \gamma Post_t$$

$\alpha = E[Y_{it}|D_i = 0, Post_t = 0]$
$\gamma = E[Y_{it}|D_i = 0, Post_t = 1] - E[Y_{it}|D_i = 0, Post_t = 0]$
$\beta = E[Y_{it} - \gamma \cdot Post_t|D_i = 1] - E[Y_{it} - \gamma \cdot Post_t|D_i = 0]$
Note: β relates to $-\gamma \cdot Post_t$ to control for time trend.
Control the fact that everyone is healthier/sicker in the post-period due to macroeconomics conditions.

DID

$$Y_{it} = \alpha + \beta D_i + \gamma Post_t + \delta D_i \cdot Post_t$$

$\alpha = E[Y_{i0}|D_i = 0]$
$\beta = E[Y_{i0}|D_i = 1] - E[Y_{i0}|D_i = 0]$
$\gamma = E[Y_{i1}|D_i = 0] - E[Y_{i0}|D_i = 0]$
$\delta = (E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1]) -$
$(E[Y_{i1}|D_i = 0] - E[Y_{i0}|D_i = 0])$ (DID estimator)
Note:

- always Post minus Pre (base), Treat minus Control (base)
- $E[Y_{it}|Post_t = 0] = E[Y_{i0}]$
- $E[Y_{it}|Post_t = 1] = E[Y_{i1}]$

DID with FE

$$Y_{it} = \alpha + \beta D_i + \lambda_t + \delta D_i \cdot Post_t$$

Note: $Post_t = \sum_m \lambda_t$ (post treatment time dummies)
Useful for granular control of time trend and smaller standard errors.

DID with two-way FE

$$Y_{it} = \alpha + \beta D_i + \lambda_t + \delta D_i \cdot Post_t$$

Note: $D_i = \sum_j \alpha_i$ (entity dummies)
Useful for granular control of time trend and smaller standard errors.

Selection bias

Observed difference = TOT + selection bias

$$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \\ &+ E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \end{aligned}$$

Selection bias = difference in outcome between treatment and control group, before any treatment
e.g. health status of older vs younger people before going to hospital

Key questions

$$E\left[Y_{0i}|D_i=1\right]-E\left[Y_{0i}|D_i=0\right]$$

- Direction
- Magnitude
- r/s between outcome and treatment selection

$$Cov(Y_{0i},D_i)$$

who is more likely to be selected for treatment

- Over-estimation of causal effect: $b>a$
- Under-estimation of causal effect: $b<a$

a := actual treatment effect
b := estimated treatment effect (might be biased)

ITT vs TOT

A : offered program, treated
B : offered program, not treated
C : not offered program

	ITT	TOT
name	Intent to Treat	Treatment on the Treated
measures	effect of making eligible for treatment	effect of taking treatment
causal effect	yes	no, unless 100% compliers
calculation	$E(Y\|A\ \&B)-E(Y\|C)$	$E(Y\|A)-E(Y\|C)$

$$y_{it}=\alpha+\beta_1(\text{PostOffer}_{it})+\beta_2(\text{PreOffer}_{it})+X_i\Gamma+\gamma_t+\epsilon_{it}$$

β_1 : ITT
β_2 : waiting list effect
Γ : control for other covariants
γ_t : time dummy

Dynamic effect

Effect of a treatment might change over time. We might be interested in the dynamics of the ATE.
Creating $N-1$ time dummies

$$I(Jul=1)+I(Aug=1)+I(Sep=1)=1$$
$$Y_i=\alpha+\beta_{Aug}I_{Aug}+\beta_{Sep}I_{Sep}+\epsilon_i$$

Coefficient of I_{Aug} measures the average outcome of Aug to the base group (Jul).

Dynamic Effect of a treatment

$$Y_i=\alpha+\sum_{m=0}^{3}\rho_m D_i I_m+\sum_{m=1}^{3}\beta_m I_m+\epsilon_i$$
$$\Leftrightarrow\alpha+\sum_{m=0}^{3}\rho_m D_i I_m+\gamma_t+\epsilon_i$$

ρ_m : dynamic effect, average treatment effect in month m
I_m : indicator function for month m
D_i : treatment of entity i
γ_t : monthly fixed effect
Note: we can include all interactions but not all month dummies.

Randomization

Randomization ensures no selection bias

$$\{Y_{0i},Y_{1i}\}\perp D_i$$

$$\Rightarrow E[Y_{0i}|D_i=1]=E[Y_{0i}|D_i=0]=E[Y_{0i}]$$
$$\Rightarrow E[Y_{1i}|D_i=1]=E[Y_{1i}|D_i=0]=E[Y_{1i}]$$

Therefore

$$\begin{aligned}&E\left[Y_{0i}|D_i=1\right]-E\left[Y_{0i}|D_i=0\right]=0\\&\Rightarrow E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=0]\\&=E[Y_{1i}-Y_{0i}|D_i=1]\\&=E[Y_{1i}-Y_{0i}]\end{aligned}$$

TOT = ATE under randomization

Limitation

1. Budget, ethical constraints
2. Impossible to conduct experiment
3. Analysing past data/ pilot program before experiment

Regression Analysis of Experiments

Regression model

$$Y_i=\beta_0+\beta_1 D_i+\epsilon_i$$

Interpretation

$$\begin{aligned}\beta_0&=E[Y_{0i}|D_i=0] && \text{(base effect)}\\\beta_1&=E[Y_i|D_i=1]-E[Y_i|D_i=0] && \text{(observed difference)}\\&=E[Y_{1i}-Y_{0i}|D_i=1] && \text{(TOT)}\\&=E[Y_{1i}-Y_{0i}] && \text{(ATE)}\end{aligned}$$

Balanced test

Objective : Check $E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]=0$
Limitation : $E[Y_{0i}|D_i=1]$ is unobserved
Solution : check pre-treated outcomes
Ideal : pre-treated outcomes are all balanced
Check

$$E[X_i|D_i=1]=E[X_i|D_i=0]$$

Note: This is sufficient but not necessary condition. i.e. No selection bias ⇒ balanced pre-treated outcomes

```
ttest X, by(treatment)
```

Conditional Independence Assumption

Random assignment conditional on a group
⇒ selection bias is zero within the group

$$\{Y_{0i},Y_{1i}\}\perp D_i|F_i$$

$$\Rightarrow E[Y_{0i}|D_i=1]\neq E[Y_{0i}|D_i=0]$$
$$\Rightarrow E[Y_{0i}|D_i=1,F_i=1]=E[Y_{0i}|D_i=0,F_i=1]$$

Regression analysis

Assume CIA
Short regression : $Y_i=\alpha_0+\alpha_1 D_i+u_i$
: $\alpha_1=\text{TOT}+\text{SB}$
Long regression : $Y_i=\beta_0+\beta_1 D_i+\beta_2 F_i+\epsilon_i$
: $\beta_1=\text{TOT}$

Adding irrelevant regressors

unconditional : adding any variables (including F) changes β little
CIA : β sensitive to inclusion of F (should include)
: β changes little when adding other variables (excluding F)

$$(1)\ Y_i=\rho_1 D_i+u_i$$
$$(2)\ Y_i=\rho_2 D_i+\delta F_i+u_i$$
$$(3)\ Y_i=\rho_3 D_i+\gamma C_i+u_i$$
$$(4)\ Y_i=\rho_4 D_i+\delta_2 F_i+\gamma_2 C_i+u_i$$

$\rho_1\approx\rho_3\neq\rho_2\approx\rho_4$

Test coefficient across regression

Note: do not read off regression table for this

```
reg Y x1
est store reg1
reg Y x1 x2
est store reg2
suest reg1 reg2
test [reg1_mean]D = [reg2_mean]D
```

Instrument Variable and 2SLS

IV enable causal effect estimates with omitted variable
Valid IV satisfies two conditions

- Relevance condition: $Cov(s_i, Z_i) \neq 0$
- Exclusion restriction: $Cov(\eta_i, Z_i) = 0$

Where s_i := treatment
Z_i := IV
η_i := error in short regression

IV and Causality

Assume $s_i \perp \nu_i | A_i$ (CIA), A_i not observed

$$\begin{aligned} \text{(short) } Y_i &= \alpha + \rho s_i + \eta_i \\ \text{(long) } Y_i &= \alpha + \rho s + \gamma A_i + \nu_i \\ \Rightarrow \eta_i &= \gamma A_i + \nu_i \text{ (OVB)} \end{aligned}$$

Omitted variable bias:
Comparing OLS and IV estimator gives sign of bias

$$\begin{aligned} \rho^{OLS} &= \frac{Cov(Y_i, s_i)}{Var(s_i)} \\ &= \frac{Cov(\alpha + \rho s + \gamma A_i + \nu_i, s_i)}{Var(s_i)} \\ &= \rho + \gamma \frac{Cov(A_i, s_i)}{Var(s_i)} \end{aligned}$$

Exclusion restriction

- Instrument is independent of potential outcomes (condition on covariates)
- Z has no effect on outcomes other than through S

Two-Stage Least Squares (2SLS)

2SLS allows adding covariates (controls) and combine multiple instruments
First stage

$$s_i = \pi_{11} Z_i + X_i' \pi_{12} + \xi_{1i}$$

Reduced Form

$$Y_i = \pi_{21} Z_i + X_i' \pi_{22} + \xi_{2i}$$

Second stage (modified short regression)

$$Y_i = \rho \hat{s}_i + X_i' \theta + (\eta_i + \rho \xi_{1i})$$

ρ is determined by

$$\rho = \frac{\pi_{21}}{\pi_{11}}$$

Estimate causal effect ρ

Causal effect is ratio of reduced form / first stage:

$$\begin{aligned} \rho &= \frac{Cov(Y_i, Z_i)}{Cov(s_i, Z_i)} \\ &= \frac{Cov(Y_i, Z_i)/V(Z_i)}{Cov(s_i, Z_i)/V(Z_i)} \end{aligned}$$

Arise from

$$\begin{aligned} \frac{Cov(Y_i, Z_i)}{Cov(s_i, Z_i)} &= \frac{Cov(\alpha + \rho s + \gamma A_i + \nu_i, Z_i)}{Cov(s_i, Z_i)} \\ &= \rho \frac{Cov(s_i, Z_i)}{Cov(s_i, Z_i)} = \rho \end{aligned}$$

Wald Estimator

Suppose IV Z_i is dummy variable

$$\rho = \frac{E(Y_i | Z_i = 1) - E(Y_i | Z_i = 0)}{E(S_i | Z_i = 1) - E(S_i | Z_i = 0)}$$

2SLS bias

Note:

- 2SLS estimator is consistent
  Large sample: $E(\hat{\rho}) \approx \rho$
- 2SLS estimator is biased
  Finite sample: $E(\hat{\rho}) \neq \rho$
- When first stage $F-$stat $= 0 \Rightarrow$ 2SLS = OLS
- Bias vanish when $F-$stat in first stage is large (> 10)
- Useless IV increase bias
  (var with no effect on first-stage R^2)

Identify IV

Good instruments come from institutional knowledge and idea about process determining the variable of interest

Test relevance condition $\Rightarrow Cov(s_i, Z_i) \neq 0$

F test > 10 for all IVs in the first stage regression

Test exclusion restriction $\Rightarrow Cov(\eta_i, Z_i) = 0$

Difficult as η_i not observed

1. Over identifying restriction test
   If $Q > K$ (num of IV > num of treatment var) and IV are chosen with the same logic

   H_0 : instruments are all valid
   H_1 : at least one of the IV are not valid

2. Qualitative argument
   Argue that Z, Y not related

3. Falsification test
   Find another non-treatment sample s.t.
   $Y_i = \beta Z_i + \epsilon_i \Rightarrow \beta = 0$

Note: $Q < K \Rightarrow$ need more IV. $Q = K \Rightarrow$ just enough IV

Panel Data and Fixed Effects

Fixed Effect model

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X_{it}' \delta + \epsilon_{it}$$

Estimated using: OLS + dummies, de-mean, first difference

Cross Section/Time Series/Panel Data

Cross section	: many subjects at the same point of time
Time series	: sequence of data points over a time interval
Panel data	: behavior of entities are observed across time (aka longitudinal data)
Balanced panel	: all data across the years are observed
Unbalanced panel	: missing data for some years

Solving OVB

To solve omitted variable bias in cross-sectional data

- Find a proxy
- Find a valid IV

Panel data further use

- Random effect (RE) models
  (Assumes no fixed effect)
- Fixed effect method
  (Eliminate time-invariant individual characteristics)
- IV, Differences in Differences

Fixed Effect Model: Causal Inference

treatment (D_{it}), treatment effect (ρ)
regression model (A_i not observed)

$$Y_{it} = \alpha + \lambda_t + \rho D_{it} + X'_{it}\delta + A'_i\gamma + \epsilon_{it}$$
$$u_{it} := A'_i\gamma + \epsilon_{it} \text{ (OVB)}$$

fixed effect model (solve OVB by absorb A_i)

$$Y_{it} = \alpha_i + \lambda_t + \rho D_{it} + X'_{it}\delta + \epsilon_{it}$$
$$\alpha_i = \alpha + A'_i\gamma \text{ (individual fixed effect)}$$

FE: Estimation

OLS regression with dummies

$$Y_{it} = \lambda_t + \rho D_{it} + X'_{it}\delta + \sum_{i=1}^{N-1} \alpha_i I_i + \epsilon_{it}$$

Within estimator (de-mean)

$$Y_{it} - \bar{Y}_i = (\lambda_t - \bar{\lambda}) + \rho(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)'\delta + (\epsilon_{it} - \bar{\epsilon}_i)$$
$$\bar{Y}_i = \alpha_i + \bar{\lambda} + \rho\bar{D}_i + \bar{X}'_i\delta + \bar{\epsilon}_i$$

First differencing

$$\Delta Y_{it} = \Delta\lambda_t + \rho\Delta D_{it} + \Delta X'_{it}\delta + \Delta\epsilon_{it}$$

Remarks

- With 2 periods, all 3 methods are algebraically the same
- First differencing introduces serial correlation of error terms (not recommended)
- Interpretation (all methods):
  for a given individual/firm/country, as X varies across time by one unit, Y increases or decrease by ρ units

FE vs OLS

- Compare OLS vs FE gives sign of selection bias
- pooled regression: OLS estimate without fixed effects
- we require variations in FE: cannot investigate time-invariant variable (e.g. union status unchanged across time)
- FE controls for all time-invariant differences between individuals

Differences-In-Differences

	Treatment	Control
Pre-Program	A $\bar{Y}^{Treatment}_{Pre}$	B $\bar{Y}^{Control}_{Pre}$
Post-Program	C $\bar{Y}^{Treatment}_{Post}$	D $\bar{Y}^{Control}_{Post}$

DID estimator

$$(\bar{Y}^{Treatment}_{Post} - \bar{Y}^{Treatment}_{Pre}) - (\bar{Y}^{Control}_{Post} - \bar{Y}^{Control}_{Pre})$$
$$(C - A) - (D - B)$$

Selection bias $:= A - B$
Time trend $:= D - B$
Treatment effect $:= (C - A) - (D - B)$
$:= (C - D) - (A - B)$

DID Regression

$$Y_{it} = \alpha + \gamma D_i + \eta Post_t + \beta D_i \cdot Post_t + \epsilon_{it}$$

D_i := indicator for observation is treatment group
$Post_t$:= indicator for time is after treatment
β := treatment effect (DID estimator)
α := pre-program mean in control group
γ := selection bias
η := time trend

DID Assumption

Parallel (common) Trends

Test $\beta_\tau = 0$ for $\tau < 0$

$$Y_{ist} = \alpha + \gamma D_i + \eta Post_t + \sum_{\tau=-m, \tau\neq -1}^{q} \beta_\tau W^\tau_t \cdot D_i + \epsilon_{ist}$$

W^τ_t := indicator of time is τ periods ago
:= Pre-period: $\tau < 0$, Post-Period: $\tau \geq 0$
$\tau \neq -1$:= base group, period right before treatment
β_τ := average $(Y - \gamma)$ difference between treatment and control group at $t = \tau$ (similar to dynamic effect model)

Common trends assumption:

- Selection bias relates to fixed characteristics of individuals

  [⇒] selection bias magnitude does not change over time
- Time trend is the same for treatment and control groups

Note:
Ideally test if untreated outcome of treatment and control are parallel in post-treatment. But counterfactual is not observed. Therefore, test for parallel trend before treatment.

No Omitted Variables that Correlate with Treatment Status

Estimation is biased if there are other factors affecting the difference in trends between treatment and control.
Difficult to test, recommend:

- Placebo (Falsification) test

  [1.1] Exploit a population not affected by policy

  [1.2] Use outcome variable not affected by policy, but affected by potential unobserved policy shocks
- IV strategy

  exogenous variation of treatment status

Endogenous Intervention

DID estimation is appropriate when interventions are as good as random, conditional on the controls

- treatment is randomised
- if possible endogeneity occur (treatment group is selected), DID is biased
- idea: the post treatment trend is expected to change even without policy (bias upwards)

Continuous treatment intensity

$$Y_{it} = \alpha + \delta S_i \cdot Post_t + \beta S_i + \eta Post_t + \epsilon_{it}$$

$S :=$ continuous variable, measuring treatment intensity.
$\delta :=$ measures treatment effect for the continuous treatment.

Adding fixed effects

$$Y_{it} = \alpha + \delta S_i \cdot Post_t + \theta_t + \delta_i + \epsilon_{it}$$

$S :=$ continuous variable, measuring treatment intensity.
$\delta :=$ measures treatment effect for the continuous treatment.
$\theta_t :=$ time fixed effects
$\delta_i :=$ individual fixed effects

DID with IVs

If S is endogenous (treatment group is associated with lower outcome), we can find an instrument for S
First-Stage DID:

$$S_i \cdot Post_t = \alpha + \delta Z_i \cdot Post_t + \theta_t + \delta_i + \epsilon_{it}$$

Reduced Form DID:

$$Y_{it} = \alpha + \delta^r Z_i \cdot Post_t + \theta_t + \delta_i + \epsilon_{it}$$

Second stage:

$$Y_{it} = \alpha + \gamma \hat{S}_i \cdot \hat{Post}_t + \theta_t + \delta_i + \epsilon_{it}$$

S can be continuous or discrete
$\gamma = \delta^r / \delta$ measures the causal effect

Clustered Standard Errors

- Panel data introduce serially correlated error problem within the cross section across years.
- Ignoring serial correlation underestimate the standard error. (exaggerate precision of regression estimates)

Bias in un-clustered standard errors

Bias: overly rejected $H_0 : \beta = 0$ (67.5% instead of 5%)

Solution1: Ignore time series data

Aggregating data into one pre and one post period

Solution2: Clustered standard error

Cluster standard error at state, year, state x year level

- Require clusters to be sampled randomly
- Require large clusters (> 10)

Regression Discontinuity

Identification assumption: all factors (other than assignment) are evolving "smoothly" with respect to X
$\Leftrightarrow E[Y_{0i}|X_i]$ and $E[Y_{1i}|X_i]$ are continuous in X_i at c.

Sharp RD

Treatment is deterministic function of assignment D

$$D = \begin{cases} 1, & X \geq c \\ 0, & X < c \end{cases}$$
$$Y = \alpha + \tau D + f(X) + \epsilon$$

$\tau :=$ treatment effect

RD Regression

$$E[Y_i|X_i] = E[Y_{0i}|X_i] + (E[Y_{1i}|X_i] - E[Y_{0i}|X_i])D_i$$
$$\tau := E[Y_{1i}|X_i] - E[Y_{0i}|X_i]$$
$$E[Y_{0i}|X_i] := \alpha + f(X_i)$$
$$\Rightarrow Y_i = \alpha + f(X_i) + \tau D_i + \epsilon_i$$

$D_i = I(X \geq c)$, $c =$ cutoff

Polynomial Method

Approximate $f(X) = \sum_{p=1}^{P} \beta_p X^p$, polynomial terms
$\tilde{X}_i = X_i - c$

$$Y_i = \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \cdots + \beta_{0p}\tilde{X}_i^p + \tau D_i + \beta_{11}^* D_i \tilde{X}_i + \beta_{12}^* D_i \tilde{X}_i^2 + \cdots + \beta_{1p}^* D_i \tilde{X}_i^p + \eta_i$$

$\beta_{0p} :=$ correlation between Y, X of the control group
$\beta_{1p}^* := \beta_{1p} - \beta_{0p}$
$:=$ incremental correlation between Y, X relative to control group
$\tau :=$ treatment effect

Local Linear Regression

Select only the data certain width h around cutoff c
$c - h \leq X \leq c + h$, $\tilde{X}_i = X_i - c$

$$Y_i = \alpha + \beta_0 \tilde{X}_i + \tau D_i + \beta_1^* D_i \tilde{X}_i + \eta_i$$

$\beta_0 :=$ linear correlation between Y, X of control group
$\beta_1^* := \beta_1 - \beta_0$, incremental linear correlation between Y, X of treatment relative to control

Note:

- Can be combined with polynomial method
- Larger bandwidth: more precise (lower variance) but higher bias

Model selection

AIC

Choosing order of polynomial p

$$AIC(p) = N ln(\hat{\sigma}^2(p)) + 2p$$
$$p_{AIC}^{opt} = \arg\min_p AIC(p)$$

$\hat{\sigma}^2(p) :=$ mean squared error of regression
$p :=$ num of regressors

Bin Dummies

Choosing order of polynomial p

Test $H_0 : \gamma_k = 0$ with $F-$test
increase polynomial order until $\gamma_k = 0$

$$Y_i = \alpha + \beta_0 \tilde{X}_i + \tau D_i + \beta_1 D_i \tilde{X}_i + \sum_{k=1}^{b} \gamma_k \cdot I_k + \eta_i$$

I_k	$:= I(X \in bin_k)$
bin_k	$:= [l, l+2h), [l+2h, l+4h), \cdots, [u-2h, u]$
l	$:=$ lowest X
u	$:=$ highest X
bin width	$:= 2h$
num of bin	$:= (u-l)/2h$

Cross-validation Procedure

Only method for both bandwidth choice h and polynomial p

$$CV(h) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{Y})^2$$
$$h_{CV}^{opt} = \arg\min_h CV(h)$$

Select only data within $c \pm h$ and conduct train/test split
Testing assumption horizontal line fit is suitable

Graphical Analysis

Divide assignment variable into a number of bins and plot the average outcome value against mid-points of the bins

Valid or Invalid RD: Sorting

Individuals influence the assignment variable. For example, individuals check their answers to avoid failing

	Marginal pass	Marginal fail
Case I	Type A and B	Only Type B
Case II	Type A and B	Type A and B

Case I: invalid RD
Case II: valid RD, individual has imprecise control

Testing Validity

- Test I: test whether the covariates W are balanced at the threshold.

  e.g. income, age, and observed characteristics not affected by treatment

  However, impossible to test unobserved characteristics

- Test II: test if density of X (assignment variable) is continuous

  jump in density indicate sorting

Test I: no Discontinuity in covariate

$x-$axis: assignment variable
$y-$axis: pre-determined variable

$$W = a + bD + f(X) + \epsilon$$

$b = 0 \Rightarrow$ pre-determined characteristics is balanced

Test II: no Discontinuity in assignment var density

$x-$axis: assignment variable
$y-$axis: density of assignment variable

$$Density(X) = a + bD + f(X) + \epsilon$$

$b = 0 \Rightarrow$ density of assignment variable is continuous

Fuzzy RD

Exploits Discontinuity in probability of treatment conditional on assignment D

$$P(D_i = 1|X_i) = \begin{cases} g_1(X_i), & X_i > c \\ g_0(X_i), & X_i \leq c \end{cases}$$

$$g_0(X_i) \neq g_1(X_i)$$

$$P(D_i = 1|X_i) = g_0(X_i) + [g_1(X_i) - g_0(X_i)]T_i$$

$$T_i = I(X_i \geq c)$$

The discontinuity T_i becomes an instrument variable for treatment status D

Imperfect Compliance: Fuzzy RD Design

First-stage

$$D = \gamma + \delta T + g(X) + \nu$$

Reduced Form

$$Y = \alpha_r + \tau_r T + f_r(X) + \epsilon_r$$

Second stage

$$Y = \alpha + \tau \hat{D} + f(X) + (\epsilon + \tau\nu)$$

$T := I(X_i \geq c)$ instrument for D
$\tau_r := \tau\delta$ Intent-to-treat (ITT) effect
$\tau := \frac{\tau_r}{\delta}$ treatment effect

Estimation

$\tilde{X}_i = X_i - c$
First-stage

$$D_i = \gamma_{00} + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p + \pi T_i + \gamma_{11}^* T_i\tilde{X}_i + \gamma_{12}^* T_i\tilde{X}_i^2 + \cdots + \gamma_{1p}^* T_i\tilde{X}_i{}^p + \epsilon_i$$

Second-stage

$$Y_i = \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i{}^2 + \cdots + \beta_{0p}\tilde{X}_i{}^p + \tau\hat{D}_i + \beta_{11}^*\hat{D}_i\tilde{X}_i + \beta_{12}^*\hat{D}_i\tilde{X}_i{}^2 + \cdots + \beta_{1p}^*\hat{D}_i\tilde{X}_i{}^p + \eta_i$$

$T_i, T_i\tilde{X}_i, \cdots, T_i\tilde{X}_i^p$ are instruments for $D_i, D_i\tilde{X}_i, \cdots, D_i\tilde{X}_i^p$

Assumption

Same assumption as standard IV framework (LATE)

- Monotonicity: assignment X must result in same direction on outcome
- Excludability: assignment X cannot impact outcome expect through receipt of treatment