Data Sources

Observational Studies:
- Panel data (multiple N, multiple t)
- Cross sectional data (large N, one t)

Quasi Experiments / Natural Experiment
- Diff-in-Diff
- Regression Discontinuity
- Instrumental Variables

Experiments
- Randomized Controlled Trials

Ordinary Least Squares

Measures of Fit
- R2: fraction of variance of Yi explained by Xi
- Standard Error of Regression (SER) : average distance btw values and reg line
- Root Mean Square Error (RMSE): average error

Homoskedasticity
- var(u | X) is constant
- Assumption: E[u | X] = 0

Heteroskedasticity
- var(u | X) not constant. u = u(X)
- Assumption: E[u | X] = 0
- Standard Error too small if not robust

Omitted Variable Bias
- Z is determinant of Y (Z part of u)
- Z correlated with regressor X

Assumptions
- unbiased estimator, E[u | X] = 0
- (Xi,Yi) are i.i.d
- X and Y have finite fourth moments

Multiple Linear Regression Model

Measures of fit
- R2, adjusted R2: penalizes R2 when too many X, show when data overfitted

Multicollinearity
- corr(X1, X2) = +-1 problem

Assumptions
- unbiased estimator, E[u | X1,…,Xn] = 0
- (X1i,…,Xni, Yi) are i.i.d
- Large outliers are rare
- No perfect multicollinearity

Non-Linear Regression Functions

Polynomials
- give table Delta Y for Delta X

Interpretation of Coefficients:
- **Lin-Log**: a 1% change in X associated with 0.01b1 change in Y
- **Log-Lin**: a unit change in X associated with 100b1% change in Y
- **Log-Log**: a 1% change in X associated with b1% change in Y

Interaction between Independent Variables
- bin-cont: create one regression line per group
- bin-bin: different slope for each dummy
- cont-cont: deltaY/deltaY = b1 + b3X2

Linear Probability Model
- Very simple to interpret

Disadvantage
- predicted probabilities >1 or <0
- assumption that b1 != b1(X)

Probit Regression

Advantage
- bounded probability and b1 = b1(X)

Interpretation of Coefficients
- b1 is the change in the z-value of unit change in X
- b0 + b1X = z-value
- To get probabilities evaluate z in cumulative standard normal distribution

Measures of Fit
- pseudo-R2: improvement in value of log likelihood relative to having no X

Logit Regression
- Same advantage as Probit
- Same interpretation of coefficient but evaluate z in logistic distribution
- Coefficients are odd rations

Validity

Internal Validity, E[u|X] != 0
- OVB
- Simultaneous Causality Bias
- Wrong functional form
- Errors in variable bias
- Sample selection bias

External Validity
- Generalization of data to other time
- to other country, urban area?

Panel Regression
- contains observation on multiple entities at two or more points in time
- balanced panel: have data for each entity for each time

Fixed Differences
- two time periods, unobserved variable Z can be controlled for

Fixed Effects
- Add constant shift alpha_i in intercept for each entity/time

Entity Fixed Effects
- Same slope for all entities, different intercepts
- Control for OV which varies across entities but not over time
- **Assumption**: covariance(X_it, alpha_i) != 0

Time Fixed Effects
- Control for OV which varies over time but not across entities

Assumptions
- E[u_it | X_i1,…,X_iT, alpha_i] = 0
- (X_i1,…, X_iT, u_i1,…, u_iT) are i.i.d
- (X_it, u_it) have finite fourth moments
- No perfect multicollinearity

Autocorrelation
- data is i.i.d across clusters but not within
- corr(Z_t, Z_(t+j)) != 0 for j != 0
- Use clustered standard errors (assume variables are not i.i.d within entities)

Limitations and Challenges
- unobserved variable determinant of Y but uncorrelated with X
- unobserved variable varies across entities and over time
- Data collection issues, non-response

Random Effect Regression
- if OV random and uncorrelated with regressors
- if OV time invariant and random
- **Assumption**: covariance(X_it, alpha_i) != 0
- Hausman Test to decide if random or fixed effects

Instrumental Variable Regression
- breaks X into two parts, one correlated with u, one not. Uncorrelated part is IV called Z_i.
- Endogeneity: variable correlated with u

- Exogeneity: variable uncorrelated with u

Condition for valid Instruments

- **Relevance**:
  - corr(Zi, Xi) != 0
  - at least one must be relevant
- **Exogeneity**: (Exclusion Restriction Principle)
  - corr(Zi, ui) = 0
  - all must be exogenous

Two Stage Least Squares

- First stage: regress X on the IV Z
- Second stage: regress Y on the estimated X
- Include control variables W in both steps
- Endogenous coefficient X is:
  - m IV, k endogenous variables
  - over/under/exactly-identified if m >/</= k

Checking Instrument Validity

- Relevance: at least one pi is nonzero
- Weak instruments:
  - all pi zero or close to zero
  - with weak instruments, 2SLS can be biased in direction of OLS estimator
  - check: compute F statistic (>10) drop weakest
- Exogeneity: only poss. if m>k, do J-test

Assumptions

- E[u | W1i,…, Wri] = 0 (exogenous regressors are exogenous)
- (Yi, X1i,…, Xki, W1i,…, Wri, Z1i,…, Zmi) are i.i.d
- (X,W,Z,Y) have finite fourth moments
- The instruments (Z1i,…, Zmi) are valid

Difference in Differences

Comparison Group

- Quality of comparison group determines quality of policy evolution
- Counterfactual: what would have happen to same people if policy not implemented

Diff-in-Diff Estimator

- difference between two before after differences
- Treatment effect isolable —> Causality

$$Y = \beta_0 + \beta_1 D_{post} + \beta_2 D_{treat} + \beta_3 (D_{treat} \times D_{post}) + \beta_4 DX + u$$

Weakness

- non random treatment
- biased estimation if other determinant of jump than policy
- Can never really know counterfactual

Assumption:

- Common trend (also parallel trends)
- Special cast of panel data, use clustered SE because of autocorrelation

Test Common Trend Assumption

- Placebo DD with fake treatment group (0 effect)
- Placebo DD with different outcome var (0 effect)
- Different comparison group (find same results)

Randomized Controlled Trial

Measurement error: precision

- Increase sample size to get rid of it

Systematic error: accuracy (bias)

- get better comparison grp (close to treatment grp)

Main Idea

- Treatment has causal effect on person
- Treatment X randomly assigned, so independent of u —> b1 is unbiased
- No OVB as X randomly assigned, indep of any W
- Having baseline (W) still increases precision

Mechanisms of Randomization

- **Pure**:(list of participants, computer)
- **Systematic**:(dice)
- **Oversubscription**:(take first who show/sign up)
- **Pipeline**:(all get treatment, randomize when)
- **Encouragement**:(Discount, when ethical hazards)
  - Run IV Reg. with getting encouragement as IV
- Think of which unit of randomization! -> cluster SE

Challenges with RCT

- Ethical concerns (vaccines)
- focus on programs easier to measure?

Remaining Threats Internal Validity

- Does the study provides unbiased estimate?
- Partial Compliance (fail to follow treatment protocol)
- Attrition (subject dropping out of study)
- Experimental effects (Experimenter bias)
- Spillover effects (Positive or Negative)
- Small Samples

Remaining Threats External Validity

- Can the study be generalized?
- Non representative sample (diff. btw. population)
- Non representative treatment (small-scale well monitored to large scale)
- General Equilibrium Effects (small experiment to large permanent changes economic environment)

Regression Discontinuity

- Impact evaluation method

Conditions/Assumptions

- Need continuous eligibility index W and clearly defined threshold w0.
- Eligibility index must be continuous
- Cutoff must be unique to the program
- Only driver of having the treatment is W score.

Main Idea

- Compare people just above and under threshold
- Treatment effect is difference around threshold
- Effect of treatment shown as jump in Y
- No need for control group
- W called running variable

Sharp RD Design

- Everyone above threshold gets treatment

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 (X_i * W_i) + u_i$$

- Interaction term allows for having two different curves left and right of threshold
- No OVB by definition, running variable determinant of getting treatment or not.

Fuzzy RD Design

- Crossing threshold changes probability to get treatment
- IV Regression with probability as IV

Challenges and Limitations

- Local average Treatment Effect
  - estimation around threshold point not always generalizable (not externally valid)
- Statistical Power
  - effect estimated at discontinuity, fewer observations than in experiment with same sample size
- Sensitivity to functional form
  - jump might be simply due to nonlinear functional form

Robustness Checks

- Functional Form (include polynomials)
- Statistical Power (change bandwidth)
- Placebo RD with other threshold (no jump)
- Placebo RD with other outcome var (no jump)
- Placebo RD with fake treatment group (no jump)
- Check for manipulation of data (plot)