Applied Statistics

Introduction

Data Sources and Types

- **Cross-Sectional Data**
  - Many observation (large N)
  - single point in time (one t)
  - For ex: General Social Surveys, Population surveys
- **Time Series Data**
  - Many point in time (large t)
  - Few units (N)
- **Panel Data**
  - Multiple entities (N)
  - Observed at two or more time periods (t)

Probability

- Random sampling: Y1…Yn selected at random are i.i.d: identically and independently distributed
  - identically means they belong to the same probability distribution
  - independently means Y1 has no information about Y2
- Correlation coefficient
  - measures linear association, useless when functional form not linear
- T-test
  - Purpose: Test the difference between two means
  - Dependent variable has to be measured on a continuous scale
  - Null Hypothesis H0 generally mean_1 = mean_2, reject if t large enough
- p-value
  - Given a significance level a, reject H0 if p-value<=a
  - the smaller the p-value, higher the evidence against H0
- Confidence Interval: Equivalent statements
  - 95% confidence interval for Delta doesn't include zero
  - Hypothesis that Delta = 0 rejected at the 5% level

Linear Regression

Ordinary Least Squares with a single regressor

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

- With observations (Xi, Yi), i=1,…,n
- u_i is the regression error
  - consists of omitted factors, generally factors that influence Yi other than variable xi.
  - also includes error in measurement of Y
- **OLS estimator:**
  - min_[b0,b1] u_i = min_[b0,b1] sum(Yi-(b0+b1Xi))^2
  - minimizes average squared difference between actual values of Yi and linear prediction
  - The result of the OLS is the estimated b_0, b_1, u_i and Y_i, written with hats
    - Yi = ^Yi + û_i
- **Measures of Fit**

    - Regression R2, measures fraction of variance of Yi explained by the regressors
    - SER, standard error of the regression, represents the average distance that the observed values fall from the regression line.
    - RMSE, Root Mean Squared error: on average we make a mistake of RMSE
- **Assumptions**
    - Conditional distribution of u given X has mean zero, E[u|X] = 0 —> b1 unbiased
        - corr(u,X) = 0 assumption. On average, the regression line is the mean of our data
    - (Xi,Yi), i=1,…n are i.i.d
        - true if (X,Y) collected by random sampling, from the same population, independently
        - Non i.i.d arises for Panel Data and Time Series Data
    - Large outliers in X and/or Y are rare
        - outliers can result in meaningless values of b1
- **Hypothesis Testing and confidence interval**
    - H_0 : b_1 = reference b_1 (generally 0) vs H1: b_1 != reference b_1
    - t = (^b_1 - ref. b_1)/SE(^b_1)
    - t-statistic for b_1 is N(0,1) for large samples

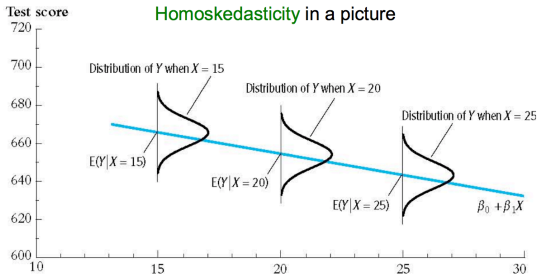
Ordinary Least Squares when X is binary

$$Y_i = \beta_0 + \beta_i D_i + \hat{u}_i$$

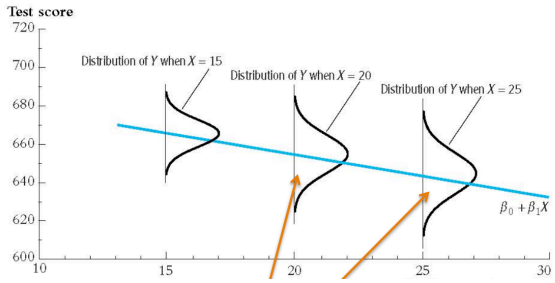
- Same mechanism as OLS but different interpretation
- Regression line makes no sense for dummy regressor
- b_1 not a slope, b_1 is a population difference in group means
    - b_1 = E[Yi|Di=1] - E[Yi|Di=0]

Heteroskedasticity and Homoskedasticity

- **Homoskedasticity**
    - var(u|X) is constant
    - E[u|X] = 0 (satisfies OLS Assumption 1)
    - If errors are homoskedastic, OLS estimators are BLUE, Best Linear Unbiased Estimators
    - Don’t need robust standard errors



- **Heteroskedasticity**
    - var(u|X) not constant
    - E[u|X] = 0 (satisfies OLS Assumption 1)
    - If errors heteroskedastic, OLS estimators are not BLUE
    - Need robust standard errors
    - SE are going to be too small if don’t include the robust standard errors



Omitted Variable Bias

- The error u here because factors that influence Y are not included in the regression
- Conditions for having a OVB Z
    - Z is a determinant of Y (Z is part of u)
    - Z is correlated with the regressor X
- Include the omitted variable in the regression

$$Yi = \beta 0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1,\ldots,n$$

Multiple Linear Regression Model

- **Interpretation of Coefficients**
  - b_1 = effect on Yi of a unit change in X1, holding X2 fixed (partial effect)
- **Assumptions**:
  - E[u|X1,…Xn] = 0, conditional distribution of u given Regressors has mean zero
  - (X1i,…,Xni, Yi) are i.i.d
  - Large outliers are unlikely
  - There is no perfect multicollinearity
- Multicollinearity
  - We cannot have more variable than data points
  - one regressor cannot be an exact linear regression of another (corr(X1,X2)=1 problem)
- **Measures of fit**
  - R2: fraction of variance of Y explained by the X
  - adjusted R2: R2 adjusted to degrees-of-freedom. Penalizes R2 as new variables are added. Adjusted R2 shows you when you overfit your data.
  - What R2 and adjusted R2 tell you:
    - Are you regressors a good explanation of your Y
  - What they don't tell you
    - Significance of the result
    - Causality between X and Y
    - If you are suffering from OVB
    - Wether you have the best set of regressors
- **Hypothesis Testing and Confidence Intervals**
  - For single coefficient
    - same recipe as for slope coefficient in a single-regressor model
    - Use t-statistic and confidence intervals (^b_1 +- 1.96* SE(^b_1))
  - Joint hypothesis testing (q regressors)
    - Need homoskedasticity
    - H0: b1 = 0 and b2 = 0 vs. H1: either b1 != 0 or b2 != 0 or both (for q=2)
    - F-test on b1 and b2 —> give F_2,infinity distribution, read in table
    - Interpretation: If value of F test above value corresponding to significance level given in the table of F_2,inf we can reject H0 that neither X1 nor X2 have an effect on Y (q=2 here)

Non-linear Regression Functions

- Case where Yi = f(X1i,… Xni) + u_i, i =1,…,n
- Polynomials and logarithmic transformation

Polynomials in X

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- Simple multiple regression model
- Regressors are powers of X
- **Interpreting the Coefficients:**
  - Don't say unit change in X leads to blablabla
  - Give table Delta Y for Delta X

Logarithmic functions of Y and/or X

- Transforms relations into percentage changes
- **Linear-log, (Y = b0 + b1ln(X) + u)**
  - A 1% increase in X (multiply by 1.01) is associated with a 0.01*b1 unit change in Y

- **Log-Linear, (ln(Y) = b0 + b1X + u)**
  - A unit increase in X is associated with a 100*b1 % change in Y
- **Log-Log, (ln(Y) = b0 + b1ln(X) + u)**
  - A 1% increase in X (multiply by 1.01) is associated with a b_1% change in Y
  - b1 is interpreted as an elasticity
  - b1 = %change in Y/%change in X

Interaction between Independent Variables

Interactions between two binary variables

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} * D_{2i}) + u_i$$

- Interaction term D1*D2 as regressor
- Allow effect of changing D1 to depend on D2. Different intercept and « slopes » for each case

Interaction between continuous and binary variable

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

- Creates two regression lines, one D=0 and one D=1
- Regression lines have different slopes and intercept
- Effectively creating one regression line per group (D)
- Two regression lines have same slope if b3=0, do hypothesis test (H0 b3 = 0)
- Two regression lines have same intercept if b1=0, do hypothesis test (H0, b1=0)
- Two regression lines are the same if b1 and b3 = 0, do joint hypothesis test (Ftest b1,b3 = 0)

Interaction between two continuous variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} * X_{2i}) + u_i$$

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

- Interaction X1*X2
- Same analysis as above

Regression with a binary dependent variable

Linear Probability Model (LPM)

$$\Pr(Y=1|X) = \beta_0 + \beta_1 X_i$$

- Dependent variable is binary
- Simple linear regression model is called linear probability model when dealing with binary dependent variables
- **Interpretation of coefficients:**
  - b1 is the change in probability that Y=1 for a unit change in X
- **Advantages**:
  - simple to estimate and interpret, same interpretation as for multiple regressors OLS
- **Disadvantages**:
  - predicted probabilities can be >1 or <0
  - Assumption that increase in probability linked with increase in X is the same for all X. (b1 != b1(X))
  - Use non-linear probability model

Probit Regression

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

- Is a non-linear probability model
- Advantage:

    - 0<=P(Y=1|X)<=1
    - P(Y=1|X) increases for X for b1>0
- Use cumulative standard normal distribution function evaluated at z=b0+b1X
- P(Y=1|X=x) = Phi(b0+b1*x), it is the area under the standard normal density to the left of z
- **Interpretation of coefficients:**
    - b1 is the change in the z-value for a unit change in X
    - b1 >0 means positive association between raising X and having a higher prob of Y=1
    - b0+b1X = z-value
    - evaluate z-value in the cumulative standard normal distribution function to have probability
- Evaluating the effect of a change in a regressor:
    - evaluate difference in Phi(z2) - Phi(z1) for zi = b0 + b1Xi
- Straightforward generalization to multiple regressors. z = b0 + b1X1 + b2X2, gives the dependent variable P(Y=1|X1=x1,X2=x2)
- **Measure of Fit:**
    - pseudo-R2: measures the improvement in the value of log likelihood relative to having no X

Logit (Logistic) Regression

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

- is a non-linear probability model
- evaluate the z-value under the cumulative logistic distribution function F
- **Interpretation of coefficients**:
    - mainly like for Probit.
    - The coefficients are odd ratios

Internal Validity

- OVB
- Wrong functional form?
- Errors in variable bias?
- Sample selection bias?

External Validity

- Data from the past apply to now?
- Data from other city apply to here?
- Data from urban area, comparable to rural areas? to other countries?

Panel Regression

A panel dataset contains observation on multiple entities. Each entity is observed at two or more points in time.

$$(X_{1it}, X_{2it}, \ldots, X_{kit}, Y_{it}),\ i = 1, \ldots, n,\ t = 1, \ldots, T$$

- **Advantages**:
    - we can control for factors that vary across entities but not over time
    - we can control for unobserved and unmeasured variables
    - if an omitted variable does not change over time, any change in Y over time cannot be caused by the omitted variable
    - More observations five you more information

- Balanced panel, have data for each entity for each year. Unbalanced, some data missing for some entities for some years.

Panel Data with Two time Periods, Fixed Differences

- Y_t2 = b0 + b1X_t2 + b2Z + u_t2
- Y_t1 = b0 + b1X_t1 + b2Z + u_t1
- (Y_t2-Y_t1) = b1 (X_t2 - X_t1) + (u_t2 - u_t1)
- The new error term is uncorrelated with X_t2 or X_t1
- This difference can be observed even though Z is not observed. The omitted variable Z doesn't change, so it cannot be a determinant of the change in Y

Fixed Effects Regression

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it},\ i = 1,\dots,n,\ T = 1,\dots,T$$

- t = time, i = entity (n of them)
- There are two ways to do this, n-1 binary regressors and fixed effects
- If you think that differences across time or entities has an effect on your regressors.

- **n-1 binary regressor**
  - integrate dummies D_i for each entity (i)
  - strictly equivalent to having a aplha_i. Changes the intercept of the regression line for each entity.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \dots + \gamma_n Dn_i + u_{it}$$

- **Fixed effects form**
  - integrate an entity effect as a constant for entity i
  - Note, b1 doesn't depend on the entity. We develop only one model which we want to apply to all entities. One model, one slope.
  - Assumption: cov(X_it, alpha_i) != 0

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

- **Time Fixed effects**
  - An omitted variable which varies over time but not across states (S_t) can be controlled for with time fixe effects.
  - Am omitted variable which varies across states but not across time (Z_i) can be controlled for with entity fixed effects

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

- **Both Time and Entity fixed effects**
  - Different intercept for each entity and years.
  - Check if the time effects are jointly statistically significant with an F test on the years dummies.
  - We are immune to OV which change over time but not over entities and which change over entities but not over time.
    - Still facing problem if variable varies over entities and over time.
    - Problem if OV is random.

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

- **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 no outlier
  - There is no perfect multicollinearity

Not same as OLS assumptions

Same as OLS assumptions

- **Autocorrelation**
  - Across clusters, the data is i.i.d but within clusters not.
  - A variable observed for the same entity across time is autocorrelated or serially correlated, if corr(Z_t, Z_t+j) != 0 for j != 0. Which is usually the case with panel data as you pick people randomly from regions but then follow them across time.
  - OLS errors assume that u_it is serially uncorrelated. For panel data, the u_it will be underestimated. We have to use **clustered standard errors**.
  - Example: gender is autocorrelated over time.
- **Clustered standard errors:**
  - estimate the variance when variable i.i.d across entities but possibly autocorrelated within.
  - we assume there is less variation in our data as in the real world, we assume the variables within entities are autocorrelated. Or conversely —> Assume variables are not i.i.d within entities.
- **Limitations and Challenges:**
  - Time lag effects can be important
  - need to use clustered standard errors (to insure against autocorrelation)
  - Data collection issues
  - Non-response in case of micro panels
  - We still have a problem if an unobserved variable is a determinant of Y but is not correlated with our regressors.

Random Effect Regression

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

- If the Omitted Variable is random and uncorrelated with the regressors.
- If the Omitted Variable is time invariant and random.
- **Assumption**:
  - covariance(X_it, alpha_i) = 0
- How to chose to use fixed effects or random effects? Hausman test

---

Instrumental Variable Regression

Threats to Internal Validity
- OVB from a variable correlated with X but unobserved
- Simultaneous causality bias (X causes Y, Y causes X)
- Errors-in-variables bias (X measured with an error)
- Result in E[u|X] != 0

What is IV
- break X into two parts, one possibly correlated with u, one not. First stage isolates part of the variation in X that is uncorrelated with u.
- Done by using an instrumental variable Z_i, which is correlated with X_i but not with u_i

An endogenous variable is correlated with u.
An exogenous variable is not correlated with u.

Usefulness of IV Regression
- When you have an endogenous variable, you look for an exogenous instrument.
- can eliminate bias when E[u|X] != 0

Conditions for valid Instruments

- **Instrument relevance**
  - corr(Zi,Xi) != 0.
  - At least one instrument must enter the population counterpart of the first stage regression
- **Instrument exogeneity**
  - corr(Zi,ui) = 0
  - All the instrument must be exogenous
  - (often called exclusion restriction)

Two Stage Least Squares

- **First Stage**: Isolate the part of X that is uncorrelated with u by regressing X on Z with OLS.
- **Second Stage**: Replace Xi by ^Xi and regress Y on ^Xi with OLS.
- ^Xi is uncorrelated with u_i

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

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

The general IV regression model

- Also happens in two phases. The W are the control variables. The X (of which there are k) are the (possibly) endogenous variables.
- There are m instrumental variables Z.
- You have to include the control variables in all the steps if the 2SLS.
- The endogenous coefficient is
  - overidentified if m>k
  - exactly identified if m=o
  - underifentified if m<k

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \ldots + \pi_{m+k} W_{ki} + u_i$$

$$Y_i = \beta_0 + \beta_1 \hat{x}_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

- **Assumptions:**
  - E[u|W1i,…Wri] = 0 —> The exogenous regressors are exogenous
  - (Yi, X1i,…, Xki, W1i,…, Wri, Z1i,…, Zmi) are i.i.d
  - There are no outliers for X, W Z and Y (nonzero, finite 4th moments)
  - The instruments (Z1i,…, Zmi) are valid —> Specific to IV

Checking Instrument Validity

- **Checking Relevance**:
  - Instruments are relevant if at least one of the pi are nonzero
- **Weak Instruments:**
  - Instruments are weak if all the pi are either zero or close to zero.
  - Weak instruments explain little of the variation in X beyond that explained by the Ws.
  - If weak instruments, 2SLS can be biased in the direction of OLS estimator
  - Checking by computing F statistic (>10). Drop the weakest instruments
- **Checking Exogeneity:**
  - Not possible if m=k
  - Possible if overidentified instrument, do a J-test of Overidentifying Restrictions.

Difference in Differences

Quasi Experiments

- Find a “natural experiment“ that allows to identify impact of a policy
- a quasi-experiment (or natural experiment) has source of randomization that is **as if** randomly assigned.
- The quality of the comparison group determines the quality of the policy evaluation
- **Comparison Group:**
  - Counterfactual: What would have happen to the same people if policy not implemented
  - a good comparison group constructs a good counterfactual that is as little biased as possible to be able to say something about causality

Difference in Differences

- **the DID estimator:**
  - the difference between two before-after differences, one for treatment group, one for control group
  - Unobserved factors that affect outcomes and changed with the treatment can be controlled for by double differencing. We can isolate the treatment effect.
- **Simple DD**
  - Two groups (one treatment, one control)
  - Two time Periods
  - DD = [(mean(Y_t1) | T=1) - (mean(Y_t0) | T = 1)] - [(mean(Y_t1) | T=0) - (mean(Y_t0) | T = 0)]
- Written as a regression:
  - Dpost = time dummy
  - b3 = DD estimate

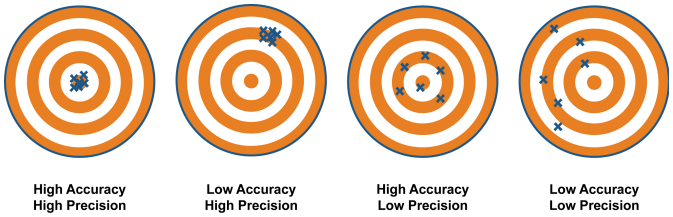
$$Y = \beta_0 + \beta_1 D_{post} + \beta_2 D_{treat} + \beta_3 (D_{treat} \times D_{post}) + \beta_4 DX + u$$

- **Requirements**
  - Need data on control and treatment group at two data points
- **Weakness**
  - Treatment is not random
  - you can’t control for all potential confounding factors
  - Any change you observe with you DD estimator will be attributed as the policy effect, if there are other factors driving this difference, the estimation will be biased
  - We can never really know the counterfactual.
- **Strength**
  - Control for effects that are common to control and treatment group
  - Effects that are common to all groups at one point in time (common trend)
- **Assumption:**
  - Common trend: The trend in control group approximates what would have happened int he treatment group in the absence of treatment
  - Also called parallel trends
- **Sensitivity Checks to test validity of common trends**
  - Use a Placebo DD with fake treatment group which you know was not affected
    - DD estimators should be zero
  - Use a Placebo DD with a different outcome variable, not affected by policy
    - DD estimator should be zero
  - Use different comparison group
    - DD estimator should reach similar results
- **Standard Errors**
  - DD is special case of estimation with panel data, need clustered SE because of autocorrelation

Randomized Controlled Trial

Precision and Accuracy:

- Measurement error: precision
  - Increase sample size to get rid of it
- Systematic error: accuracy (bias)
  - Get a better comparison group, should be as close as possible to treatment group



Establishing Causality:

- Problem of Counterfactual, we will never be able to have a perfect counterfactual as the same person cannot at the same time be inside and outside the treatment group.
- Counterfactual: What would have been the condition of the population at the time of the policy evaluation if the policy had not been implemented

Threats to Internal Validity in observational studies: —> E[u|X] != 0

- OVB from variable correlated with X but unobserved
- Sample selection bias (availability of the data related to Y)
- Simultaneous causality bias (X causes Y and Y causes X)
- Errors-in-variable bias (X measured with error)

Experiment:

- An experiment randomly assigns subjects to treatment and control groups

Randomized Controlled Trial

- A treatment has a causal effect for a given individual
- The average treatment effect is the population mean value of the individual treatment effects
- A RCT randomly assigns individuals to treatment and control groups
  - X randomly assigned, then X independent of u, so b1 is unbiased in an OLS regression
  - The causal effect is the value of b1 in an ideal RCT
- **Control variables**
  - As X is randomly assigned, we are not facing OVB because X is independent of any control variable W. (No systematic error, we are accurate)
  - However adding control variables reduces the error variance (could have less measurement error, be more precise)
- **Baseline:**
  - Do you need a baseline for RCTs? (set of variables measured before experiment)
  - By having a baseline you can improve the precision of your measurement.
- **Unit of Randomization:**
  - Pay attention to the unit of randomization you choose (for for SW book)
  - Do you randomize students within a class, different classes, different schools?
  - Careful with clustered SE if randomize in classes/schools
- **Checking for balance**
  - After randomization check it has worked
  - run t-test for all control variables W. Should have E[W|X=1] = W[W|X=0]
  - Run regressions of X on all W and conduct F-tests
- **Challenges with RCTs**

- Focus on programs easier to measure?
- Ethical concerns
- How many evaluations should be made across culture before having “common knowledge“

Mechanisms of Randomization

- **Pure randomization** —> Preferred solution if you have a list of participants
  - Problem: usually harder to get list for smaller entities.
- **Systematic Randomization** —> throw dice, lottery tickets
- **Oversubscription Randomization** —> Take the first who show up / sign up
- **Pipeline / Phase-In Randomization** —> Everybody gets the treatment but at different time, randomize when they get the treatment
- **Encouragement Randomization** —> Discount on something for certain persons
  - When you are facing ethically sensitive interventions, mechanisms which doesn’t exclude anyone from getting the treatment (for ex. Vaccines)
  - Need to run an IV regression with getting the incentive as IV for getting the treatment

Remaining Threats to Internal and External Validity

- **Internal Validity**
  - Wether the study provides unbiased and general estimate of what it claims to estimate
  - **Partial Compliance**
    - failure to follow treatment protocol, some controls get treatment and inversely
  - **Attrition**
    - Some subjects drop out of the study
  - **Experimental effects**
    - Experimenter bias
  - **Spillover effects**
    - Negative or Positive Spillover from treatment to the control group
  - **Small samples**
    - Not source of bias but of lower precision
- **External Validity**
  - Wether the results from the study can be generalized to other populations
  - **Non representative sample**
    - population studied and population of interest are different
  - **Non representative treatment**
    - small-scale and tightly monitored program could be different to large scale one
  - **General equilibrium effects**
    - turning small temporary experiment to large permanent could change economic environment.

Regression Discontinuity

Key Concept

- RD is an impact evaluation method
- Usable for programs that have continuous eligibility index (W) with clearly defined eligibility threshold (w0) (Like getting a scholarship based on test scores)

- **Conditions:**
  - Eligibility index (W) must rank people in a continuous way
  - Index must have a clearly defined cutoff (w0)
  - Cutoff must be unique to the program of interest and cannot be manipulated (not to any other program)
- **Main Idea:**
  - Compare population just above (treated) and just under (untreated) the threshold w0.
  - Treatment effect is the difference between individuals on both sides of the threshold.
  - Effect of treatment (w0) should show up as a jump in the outcome Y
  - Don't need control group —> Yaay.
- **Assumption**:
  - Observation on both sides of the threshold are very similar
  - The parameter value is the only driver of the assignment of a beneficiary to the treatment
  - The treatment is the only source of discontinuity in outcomes

Regression Discontinuity Design

- **Sharp RD Design**
  - Everyone above the threshold w0 gets the treatment
  - Treatment effect estimated by b1, X is treatment dummy
  - W is called the running variable
  - Allow for different slopes on left and right of the threshold. We have two curves, one with and one without Treatment.
  - **No OVB** per definition because only variable which affects X is W and we control for it.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

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

- **Fuzzy RD Design**
  - Crossing threshold only influences probability to get the treatment
  - Use IV regression.

Challenges and Limitations of RD

- **Local Average Treatment Effect (LATE)**
  - we estimate the effect around the cut-off point, not always generalizable to other population groups (Not Externally valid
- **Statistical Power**
  - Effect is estimated at the discontinuity, we have fewer observations than in a randomized experiment with same sample size (Low precision)
- **Estimated effect can be very sensitive to functional form**
  - Include nonlinear relationships
  - The effect might just be due to a nonlinear functional form

Robustness Checks

- **Functional Form**
  - Take into account nonlinearities, include polynomials
- **Statistical Power**
  - Move window around the threshold (narrower bandwidth)
  - Trade-off between bias and efficiency
- **Placebo RD, with other threshold** (w1). Should not see a jump.
- **Placebo RD, with other outcome variable**. Should not see a jump.
- **Placebo RD, with a fake treatment group**. Should not see a jump.
- **Check for manipulation**: plot.