EC4305

Causal Inference

- Observed Difference = $E[Y_1|D = 1] - E[Y_0|D = 0]$
- Average Treatment Effect on Treated (ATT) = $E[Y_1|D = 1] - E[Y_0|D = 1]$
- Selection Bias = $E[Y_0|D = 1] - E[Y_0|D = 0]$
- $E[Y_1|D = 1] - E[Y_0|D = 0] = E[Y_1|D = 1] - E[Y_0|D = 1] + E[Y_0|D = 1] - E[Y_0|D = 0]$
- i.e. $Observed\ Difference\ =\ ATT\ +\ Selection\ Bias$
- If the assignment is randomized, $E[Y_1|D = 1] = E[Y_1|D = 0], E[Y_0|D = 1] = E[Y_0|D = 0]$
- $\therefore\ Observed\ Difference\ =\ ATT\ =\ ATE$

OLS 1

- OLS Refresher
  - $\beta = \frac{Cov(y,x)}{Var(x)}$
  - $\beta =\ estimand, true\ population\ parameter$
  - $\hat{\beta} = estimator, random\ variable$
  - $1.55\ =\ estimates, a\ specific\ number$

- Standard Errors
  - measures how far the $\hat{\beta}$ is away from true β
  - $Var(\hat{\beta}) = \frac{1}{\sum_{i=1}^{n}(x-\bar{x})^2} \times \frac{1}{n-1}\sum_{i=1}^{n}\widehat{u^2}$
  - larger sample size, smaller std. err. $\because n - 1$
  - larger variance of x, smaller std. err. $\because (x - \bar{x})^2$
  - larger unexplained variation of y, larger std. err. $\because \widehat{u^2}$
  - p-value doesn't tell us anything about the effect and the unbiasedness of estimator
  - homoscedastic-only standard error and heteroscedastic-robust standard error

- Cluster Robust Standard Errors
  - ICC: Intra-Cluster Correlation
  - $ICC\ =\ 1\ \Rightarrow every\ observation\ in\ that\ cluster\ has\ the\ same\ value$
  - $ICC = 0\ \Rightarrow every\ observation\ in\ that\ cluster\ is\ completely\ different$
  - $ICC\ >\ 0\ \Rightarrow less\ variability\ within\ a\ cluster\ \Rightarrow smaller\ std. err$
  - Standard errors will be biased downward and show significant results too often
  - ESS: Effective sample size. The sample size that would give the same std. err as the clustered standard errors if there is no clustering
  - $ESS\ =\ \frac{M \times K}{1+ICC(M-1)}\ , where\ K\ =\ \#\ of\ clusters, M\ =\ \#\ of\ observations\ in\ a\ cluster$
  - $ICC\ =\ 0\ \Rightarrow ESS\ =\ K\ \times M\ =\ N$
  - $ICC\ =\ 1\ \Rightarrow ESS\ =\ K$

OLS 2

- Conditional Independence

- For a fixed characteristic, the causal effect is the outcome of getting treatment – not getting treatment. Here, we condition on the characteristic (i.e. holding some x constant)
- $y = \alpha + \beta D + \epsilon, \ \epsilon = \gamma x + \delta$
  - The error term ϵ is correlated with x, so $E[\epsilon|D = 1] \neq E[\epsilon|D = 0]$
  - But if we control for x, then $E[\delta|D = 1] = E[\delta|D = 0]$, because δ is uncorrelated with anything.
  - So, D is conditionally independent is that it is not correlated with the error term δ, or the only reason it is correlated with ϵ is because of γx

- Assumptions for OLS to be unbiased
  - A1: DGP has to be linear
  - A2: No perfect multicollinearity between independent variables
  - A3: The conditional mean of error term has to be zero ($E[\epsilon|x] = 0$)

- OLS Biases
  - Omitted Variable Bias
    - If there is OVB, x_2 has correlation β_2 with x_1 and γ with y
    - If there is OVB, the estimator we get will be $\beta_1 + \beta_2 \frac{Cov(x_1,x_2)}{Var(x_1)}$
    - Note: $\gamma = \frac{Cov(x_1,x_2)}{Var(x_2)}$, it is different from the estimator we get
    - If either $\beta_2 = 0 \ or \ \gamma = 0$, then there is no OVB
    - $\gamma \wedge \beta_2 \ have\ the\ same\ direction, bias\ is\ positive\ (\hat{\beta} > \beta)$
    - $\gamma \wedge \beta_2 have\ different\ direction, bias\ is\ negative\ (\hat{\beta} < \beta)$

  - Good and Bad Controls
    - Good controls: reduce bias
    - Bad controls: increase bias
    - Example of bad controls: x affects y through m, then m is a bad control because it cuts off the mechanism
    - Example of bad controls: colliders. x and y both affect z, then z is a bad control

  - Reverse Causality
    - y affects x but we run the regression as x affects y

  - Attenuation Bias
    - Measurement error
    - $\hat{\beta}\ will\ be\ biased\ towards\ 0$
    - But at least we know the direction of the bias
    - We can say the effect is at least $\hat{\beta}$ but potentially larger
    - Measurement error in y doesn't cause a problem in unbiasedness, it makes the standard errors larger
  - Non-linear Relationships
    - Polynomial regression allows x to have a different slope at different levels of x (i.e. the effect depends on x itself)
    - Binary Outcome Variables
      - we can use LPM, Logit, Probit model

            - LPM will go beyond 1 and below 0; Logit and Probit is much harder to interpret
        - Categorical Variables
            - It will be useful if we turn the “levels” into numerical values
            - But most of the time it is hard to compare the results between different scale systems
            - Instead, we standardize them using the formula: $\frac{x-\bar{x}}{sd(x)}$
            - The unit of the standardized value is standard deviation
        - Interaction between Binary Variables
            - β_3 estimates the difference in the effect of x_1 with x_2 and without x_2
        - Interaction between Continuous and Binary Variables
            - β_3 is the difference between binary variables in the association of the continuous variable

linear-log	$y = \beta_0 + \beta_1 ln(x) + u$	1% increase in x is associated with $0.01\times \beta_1$ change in y
log-linear	$ln(y) = \beta_0 + \beta_1 x + u$	A change in x by 1 unit is associated with $100 \times \beta_1\%$ change in y
log-log	$ln(y) = \beta_0 + \beta_1 ln(x) + u$	A change in x by 1% is associated with a $\beta_1\%$ *change in y*

Experiments

- Experiments Basics
    - In an experiment, β will be the same regardless of how many control variables are included because we randomly assign x. But it does improve the precision of β.
    - A treatment may affect x, but cannot be caused by x.
    - Be careful of what we control for, don’t control for any bad controls (colliders, mechanism)
    - Typically, we control for variables that is determined at the baseline, because they will not have correlation with the treatment (treatment hasn’t given at the baseline)
    - It is possible to measure the effect of one treatment on numerous outcome variables

- Heterogeneous Effects
    - Differences in the effect of the treatment on different groups
    - We can use interaction terms:
      $y = \alpha + \beta_1 \times Treatment + \beta_2 Female + \beta_3 \times Treatment \times Female$
    - This gives us the Std. Err of the difference (β_3)
    - However, only β_1 can be interpreted as causal effect because other control variables are not randomly assigned, might have OVB.

- Statistical Significance and Power

  - Power of Study: the probability of finding a significant result, given some parameter values of the experiment
  - What we need to find the power of study:
    - Effect Size
    - Standard deviation of outcome variable
    - Sample Size
    - Significance Level
  - Minimum Detectable Effect (MDE)
    - How large does the effect size have to be to get k% power?
    - We can set the sample size so that the MDE is as large as we think the effect is
    - MDE should be set at the level where the study will change the policy

- Internal Validity
  - Attrition Bias
    - If there is attrition bias, it violates the assumption of "the treatment is independent of potential outcomes"
    - If attrition is random, then it doesn't cause a problem
    - Attrition is a problem when it makes the endline groups uncomparable

  - Reactivity/Experimenter Demand Effects
    - Treatment group behave differently because they know they are in an experiment
    - Subjects respond to questions in the way they think the researcher wants them to
    - We can use placebo and double-blinding to address these issues

  - Spillovers
    - Who shouldn't be treated is indirectly treated
    - When spillovers go in the same direction as the treatment effect, there is a bias towards zero (absolute effect is smaller)
    - When spillovers go in different direction as the treatment effect, it creates a bias away from zero (absolute effect is larger)

  - General Equilibrium Effects
    - Experiment results might be valid only in partial equilibrium but not general equilibrium

- External Validity
  - Demographic groups
  - Over time
  - Geographically
  - By implementer

- ITT, ATE, ATT
  - ITT: Effect of being assigned to treatment group (difference of outcome between treatment group and control group)
  - $ITT = E(Y_i|T_i = 1) - E(Y_i|T_i = 0)$

  - ATE: Average of randomly selected patient taking the drug

- $ATE = E(Y_{1i} - Y_{0i})$

- Take-up: % of treatment group who took the treatment
- $Take - up \; < \; 100\% \; \Rightarrow ITT \neq ATE$
- $Take - up \; = \; 100\% \; \Rightarrow ITT \; = \; ATE$. (i.e. all subjects in treatment group take the treatment)
- $ATE \neq E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$, because D is not randomly assigned, only T is

- $ATT = E(Y_{1i} - Y_{0i}|D_i = 1)$
- If there is no spillovers, $ATT = \frac{ITT}{Take-up}$

$$\Leftrightarrow ITT \; = \; ATT \; \times Take - up$$

- Informal intuition: ITT is diluted because of those who didn't take up the drug. To get ATT, we need to inflate ITT by how much it was diluted.
- $ATT \; \neq ATE$. ATT is only for those who took the drug. They took it because it has an effect to them. ATE includes those who aren't affected by the drug regardless whether they take the drug. So ATE is smaller than ATT because ATE is diluted by these people.

- Publication Bias
  - Papers are only published when the result is significant

- Multiple-Hypothesis Testing
  - Do many hypotheses and then focus on those with statistically significant results
  - Bonferroni Correction: multiply the p-value by the number of tests

- P-hacking
  - Use different outcome variables until significant result is found
  - Pre-analysis plan makes p-hacking harder

Instrumental Variable

- Assumptions for IV
  - Relevance Condition: $Cov(x_i, z_i) \neq 0$
    - The instrument is affecting x directly or indirectly
  - Exclusion Restriction: $Cov(\epsilon, z) = 0$
    - Instrument is randomly assigned (uncorrelated with potential outcome)
    - Instrument can only affect y through x

- 2SLS
  - First stage: $x_i = \alpha_1 + \pi_1 z_i + \delta_i$
  - Second stage: $y_i = \alpha_2 + \beta\widehat{x_i} + \epsilon_i$
  - β^{2SLS} *is the consistent estimate of* β
  - We are only using the variation in x_i that is determined by z_i

- Reduced Form
  - $y_i = \alpha + \pi_2 z + \; \epsilon$
  - Effect of the instrument on outcome variable

- $\beta^{2SLS} = \frac{\pi_2}{\pi_1} = \frac{coeff\ of\ y\ on\ z}{coeff\ of\ x\ on\ z}$
- Another way to put it: $z \rightarrow x \rightarrow y$

  To get the effect of z on y, we multiply **the effect of z on x** and **the effect of x on y** sequentially. (i.e. $\pi_1 \times \beta^{2SLS} = \pi_2$). Rearranging terms, we obtain $\beta^{2SLS} = \frac{\pi_2}{\pi_1}$.

- More on Relevance Condition
  - Relevance condition says there has to be a first stage
  - First-stage F-stat needs to be greater than 10, then the bias will be small
  - Bias: $E[\beta^{2SLS} - \beta] = \frac{Cov(\epsilon,\delta)}{Var(\delta)} \times \frac{1}{1+F}$
  - The smaller the first-stage is, the more bias towards OLS Bias
  - Weak instruments also give imprecise estimation

- More on Exclusion Restriction
  - Independent of potential outcomes, conditional on covariates (z is determined outside of the system, and randomly assigned)
  - Instrument has no effect on outcomes other than through the first-stage channel

- LATE
  - Heterogeneous effect for different individuals: $Y_{1i} - Y_{0i} = \beta_i$
  - **A1 Independence**: $Y_i(d,z); \forall d, z, D_{1i}, D_{0i} \perp Z_i$
    - The instrument is as good as randomly assigned
    - $\Rightarrow E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = E[Y_i(D_{1i}) - Y_i(D_{0i}, 0)]$
    - We can measure the causal effect of the instrument on D_i because it is also randomly assigned
    - $\Rightarrow E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = E[D_{1i} - D_{0i}]$

  - **A2 Exclusion**: $Y_i(d,0) = Y_i(d,1)\ for\ d = 1\ and\ d = 0$
    - Conditional on d, Z has no additional effect on Y
    - Z affects Y only through the channel of D
    - Together A1 and A2, we have the exclusion restriction

  - **A3 Relevance**: $E[D_{1i} - D_{0i}] \neq 0$
    - The instrument must affect D_i in some way
    - D_{1i} got instrument, so it will be different from D_{0i}

  - **A4 Monotonicity**: $D_{1i} - D_{0i} \geq 0, \forall i$
    - The instrument may have no effect on some people, all those affected are affected in the same way
    - It assumes there is no defiers on individual level

  - If A1-A4 holds, then, **LATE theorem**:

$$\frac{E[Y_i|z_i = 1] - E[Y_i|z_i = 0]}{E[D_i|z_i = 1] - E[D_i|z_i = 0]} = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}] = LATE$$

$$\frac{Effect\ of\ z\ on\ y}{Effect\ of\ z\ on\ x} = \beta^{2SLS}$$

    - Detail Derivation of LATE:

$$E[Y(z_i = 1) - Y(z_i = 0)]$$

$$
\begin{aligned}
&= E[Y(z_i = 1) - Y(z_i = 0)|D_1 = 1, D_0 = 0]P[D_1 = 1, D_0 = 0] \\
&\quad + E[Y(z_i = 1) - Y(z_i = 0)|D_1 = 0, D_0 = 1]P[D_1 = 0, D_0 = 1] \\
&\quad + E[Y(z_i = 1) - Y(z_i = 0)|D_1 = 1, D_0 = 1]P[D_1 = 1, D_0 = 1] \\
&\quad + E[Y(z_i = 1) - Y(z_i = 0)|D_1 = 0, D_0 = 0]P[D_1 = 0, D_0 = 0] \\
&= E[Y(z_i = 1) - Y(z_i = 0)|D_1 = 1, D_0 = 0]P[D_1 = 1, D_0 = 0]
\end{aligned}
$$

$$E[Y(z_i = 1) - Y(z_i = 0)|D_1 = 1, D_0 = 0] = \frac{E[Y(z_i = 1) - Y(z_i = 0)]}{P[D_1 = 1, D_0 = 0]}$$

Note:

$$
\begin{aligned}
P[D_1 = 1, D_0 = 0] &= 1 - P[D_0|z_i = 1] - P[D_1|z_i = 0] \\
&= 1 - (1 - P[D_1|z_i = 1]) - P[D_1|z_i = 0] \\
&= P[D_1|z_i = 1] - P[D_1|z_i = 0] \\
&= E[D_i|z_i = 1] - E[D_i|z_i = 0]
\end{aligned}
$$

Plugging in this result into the denominator, we get:

$$E[Y(z_i = 1) - Y(z_i = 0)|D_1 = 1, D_0 = 0] = \frac{E[Y(z_i = 1) - Y(z_i = 0)]}{E[D_i|z_i = 1] - E[D_i|z_i = 0]}$$

$$E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}] = \frac{E[Y_i|z_i = 1] - E[Y_i|z_i = 0]}{E[D_i|z_i = 1] - E[D_i|z_i = 0]}$$

- LATE Theorem states that if A1-A4 are satisfied, the instrument can be used to estimate the average causal effect on compliers

- ATT is the weighted average of effects on always takers and compliers
  - $E[Y_{1i} - Y_{0i}|D_i = 1]$
    $= E[Y_{1i} - Y_{0i}|D_{0i} = 1]P[D_{0i} = 1|D_i = 1]$
    $+ E[Y_{1i} - Y_{0i}|D_{0i} > D_{0i}]P[D_{1i} > D_{0i}|D_i = 1]$
  - ATT = Effect of always takers × Share of always takers among D=1 + Effect of compliers × Share of compliers among D=1
- If there is no treatment among those z = 0, then ATT = LATE
  - z affects y only through D
  - If z absent but D is not, the effect of D on Y is biased
  - If z present but D is not, there is no way z can affect Y
  - In conclusion, we get the above statement

- When the whole population are compliers then ATE = LATE = ATT

- Notes:
  - When we use IV in an experiment/regression, the difference of the effect between those who was assigned IV and not assigned IV **automatically eliminates the effects of always-takers and always-non-takers**.
  - So we only left with compliers and defiers. We only need to **assume there is no defiers in the control group** because **defiers in the treatment group will not have effect on Y** since z can only affect Y through D, and they don't have D.
  - With that, now we are only left with compliers. Since all are **compliers**, then both are the **counterfactual of each other**. (control subjects are treated if not treated; treatment subjects are not treated if treated)

        - Therefore, the difference between them is **ATT**. Since we only left with compliers, we are also measuring the **LATE**.

    - Relationship between ATE, ATT, ITT, LATE
        - take-up = 100% ⇒ ITT = ATE
        - take-up = 100% ⇒ ATT = ITT
        - ∴ take-up = 100% ⇒ ITT = ATT = ATE
        - no spillover ⇒ We have only compliers (LATE)
        - Then LATE = ATT
        - ∴ take-up = 100% ⇒ ITT = ATT = ATE = LATE

- Weak Instruments
    - First-stage F ≅ 10 ⇒ weak instrument
    - A small first stage coefficient means a minor violation of the exclusion restriction can lead to a severe bias
    - Consider a Two-Stage:

$$x_i = \pi_0 + \pi_1 z_i + \delta_i$$
$$y_i = \alpha + \beta x_i + \gamma z_i + \epsilon_i$$

    - This is a violation of exclusion restriction because z can directly affect y without the channel of x.
    - When $\gamma \neq 0$, $\beta^{2SLS} = \beta + \frac{\gamma}{\pi_1}$. The bias is the relative size of the size of the violation of the exclusion restriction γ and the strength of the first stage π_1 .
    - If π_1 is small, a small γ can lead to huge bias. The bias doesn't disappear as the sample size increase because β^{2SLS} no longer consistent.

Fixed Effect

- Basic FE
    - Instead of controlling for variables as a continuous variable, we control for an indicator variable for each category.
    - can't control for FE at the level of your variable of interest
        - this results in perfect multicollinearity
        - e.g. variable of interest : school's teacher student ratio.
          we can't control for school because once we know the school, we know its school's teacher student teacher ratio
    - FE coefficients are interpreted as differences to the comparison group
    - Entity FE controls for variables that changes over time but remaining constant within each entity
    - When there are only 2 periods, first differencing and including entity FE yields the same results

$$\Delta GDP_i = \beta_i \Delta X_i + \Delta\epsilon_i$$

    - If there is something that changes and causes changes to both dependent variable and variable of interest, then there are OVB
    - Typically, we control for an indicator variable for each individual entity

$$Y_{it} = \alpha + \beta_1 X_{it} + \delta_1 FE_1 + \delta_2 FE_2 + \cdots + \epsilon_{it}$$
$$Y_{it} = \alpha + \beta_1 X_{it} + \delta_i + \epsilon_{it}$$

    - We don't care about δ_i as it indicates fixed effects at the i level.
- Two-Way Fixed Effects (TWFE)

  - If Y_{it} always grow, then we have OV that can be thought of as time or anything that changing over time, because we never control for time.

$$Y_{it} = \alpha + \beta_1 X_{it} + \delta_i + \tau_1 Time_1 + \tau_2 Time_2 + \cdots + \epsilon_{it}$$
$$Y_{it} = \alpha + \beta_1 X_{it} + \delta_i + \tau_t + \epsilon_{it}$$

  - Time FE are the same across all entities in that year, so it only has subscript t
  - To be an OV, it has to change over time and change differently for different countries.
  - TWFE cannot solve reverse causality since TWFE has no difference from adding control variables in usual regression.
  - TWFE cannot deal with attenuation bias. It makes it a lot worse.

- Standard Errors
  - Typically, there is a lot serial correlation in panel data.
  - Positive intra-cluster correlation in both outcome variable and the variable of interest.
  - Using usual robust standard error formula will give downward bias for the standard errors.
  - Thus, we should use cluster-robust standard errors.
  - When the number of clusters is small (<30), the cluster standard error is downward bias.

Difference-In-Differences

- Counterfactual
  - A pre-post comparison is useful when the time trend is very stable before the event that we are interested.
  - But when there is a trend over time, it will bias the pre-post comparison
  - If there is anything happened at the same time, this will also bias the comparison

$$Y_{it} = \alpha + \beta X_i \times Post_t + \delta X_i + \tau Post_t + \epsilon_{it}$$

  - β is the coefficient of interest. It measures the difference between before treatment and after treatment, compared to the one who didn't get the treatment.
  - We can include FE in DID to obtain a more precise estimate

$$Y_{it} = \alpha + \beta X_i \times Post_t + \delta_i + \tau_t + \epsilon_{it}$$

    This equation controlled for both $Post_t$ and X_i in the entity and time FE term. However, it is sufficient to control for X_i and $Post_t$ since these are the only thing that correlated to the interaction term at the entity level and time level respectively.

- Assumptions
  - Parallel trend assumption : In the absence of the treatment, the entities that got the treatment and the entities that didn't get the treatment would have changed in the same way.
    - treatment don't need to be randomly assigned
    - treatment and control group can be at different level
  - Nothing else happened at the time of the treatment differentially affecting treatment and control entities.
  - There are 2 methods to check these assumptions:
    - Method 1: Check the pre-trends are parallel

        - we can include “leads” to look for differential pre-trends

$$y_{it} = \alpha + \beta_1 Treat_i \times Post_t + \beta_2 Treat_i \times Lead1_t + \beta_3 Treat_i \times Lead2_t + \delta Treat_i + \tau_1 Lead1_t + \tau_2 Lead2_t + \tau_3 Post_t + \epsilon_{it}$$

        - If we find significant coefficients in the “lead” interaction term, then that is an evidence of differential pre-trend.

    - Method 2: Placebo test
        - This test checks that if there is something else happened at the time when treatment happened.
        - Take something that shouldn’t be affected by treatment as the outcome variable and run the regression.

$$Z_{it} = \alpha + \beta X_{it} \times Post_t + \delta_i + \tau_t + \epsilon_{it}$$

        - If $\beta \approx 0$, this means the study passed the placebo test. Otherwise, it means something happened with Z just around the time when treatment is given.

- Problems of Differential Trend
    - If there is a slight differential trend, it doesn’t bias our estimate much in the short-run.
    - However, the bias can build up over time.

- Addressing Problem of Differential Trend
    - If we assume the trends are linear and we have more than 2 time periods, then we can control for the linear trends

$$Y_{it} = \alpha + \beta X_i \times Post_t + time_t \times \delta_i + \delta_i + \tau_t + \epsilon_{it}$$

    - Adding the linear time trend term means that in the absence of the treatment, the treatment group and the control group would have changed away from their linear time trends in the same way.

- Addressing Problem of Something Else Happened in the Treatment Given Time
    - We can control for this variable.
    - But since there are TWFE in our regression model, what we can control for is those changes over time and over entities. Otherwise, there is collinearity.

- What Are We Estimating?
    - DID estimator estimates the ATT because we have the counterfactual for the treatment group.

- Dynamic Effects
    - When using the basic DID, we are measuring the average treatment effect over all the time periods in the post period.
    - But some effect depends on how long our post period is. Some may not show significant effect immediately but show in later period.
    - We can split the $Post_t$ variable into many variables

$$Y_{it} = \alpha + \beta_1 X_i \times Post1_t + \cdots + \beta_T X_i \times PostT_t + \delta_i + \tau_t + \epsilon_{it}$$

    - Now, each β is a DID estimate from the pre-period to a particular year in the post period.

    - If the β stabilize overtime, then the final β_T is our best guess for the long-run effect

Triple Difference-In-Differences

- For triple-difference to work, we need 2 characteristics (Treat1 and Treat2)
    - You are only treated if you have both characteristics and the time period is "post".

- DID

$$Y_{it} = \alpha + \beta_1 Treat_i \times Post_t + \delta Treat_i + \tau Post_t + \epsilon_{it}$$

- Triple DID

$$Y_{ijt} = \alpha + \beta_1 Treat1_i \times Treat2_j \times Post_t + \beta_2 Treat1_i \times Treat2_j + \beta_3 Treat1_i \times post + \beta_4 Treat2_j \times post + \delta_1 Treat1_j + \delta_2 Treat2_j + \tau Post_t + \epsilon_{ijt}$$

- If there is some omitted variable effecting the outcome and correlated with Treat1 that changes just around the start of the treatment, as long as it is not correlated with Treat2 it doesn't cause a bias.

- So to be an OV, it must be correlated with Treat1 and Treat2.

- Intuition of Triple-DID : Imagine 2 schools, one hire math TA and the other doesn't. We want to estimate the effect of math TA on math scores. However, other than math TA, there is also other things happening at the same time that affects math scores. How can we isolate the effect of math TA on math scores? We use English scores as another counterfactual. Since math TA doesn't bother with English scores, we just take the second difference of math scores subtract the second difference of English score, then we isolated the desired effect.

- First difference : pre-post
- Second difference : between entities
- Third difference : characteristics within entity

Event Studies

- Event studies can be used when we have an event happening at a specific time.

- We don't need to have a comparison group.

- We assume that our outcome variable wouldn't have changed in the absence of the event, and nothing would have happened to the outcome variable.

$$Y_{it} = \alpha + \beta_1 post_t + \delta_i + \epsilon_{it}$$

    - β_1 is the average pre-post difference.
    - This is basically a pre-post comparison.

- If we think that there are linear time trends, we can control for those

$$Y_{it} = \alpha + \beta_1 post_t + \beta_2 time_t + \delta_i + \epsilon_{it}$$

- Now we are estimating the effect from deviations from the time trend.

- We can allow for an effect on both the level and trend by including a $post_t \times time_t$ term
$$Y_{it} = \alpha + \beta_1 post_t + \beta_2 time_t + \beta_3 post_t \times time_t + \delta_i + \epsilon_{it}$$
  - β_1 is the immediate effect of treatment.
  - For each time period after the start of the treatment, the effect changed by β_3 from it's original level.
  - This specification can be used when we think that the treatment affects the trend and not just the level of the outcome.

- We can include Time FE for each time period. However, to avoid collinearity, we need to drop one time period. We would drop the time period just before the start of the treatment. Then τ_1 is the effect of the treatment one period after it started, and so on.

- We can even create 2 times, one relative to the event, one relative to the calendar time.
$$Y_{it} = \alpha + \cdots + \beta_{-2} Event - 2_{it} + \beta_0 Event0_{it} + \beta_1 Event1_{it} + \cdots + \tau_t + \delta_i + \epsilon_{it}$$
  - Now each period got each effect respectively.

Regression Discontinuity

- RD gets rid of OVB and reverse causality. You can think that it is similar to RCT when we are close to the cutoff.

- RD needs an assignment variable. If the assignment variable passes the cutoff, then it is assigned to treatment, otherwise, it is control group.
$$Y_i = \alpha + \tau D_i + \gamma Z_i + \epsilon_i$$

- Assumption
  - All factors other than the treatment are evolving "smoothly" with respect to assignment variable
  - Under this assumption, those that just below the cutoff can be a counterfactual for those who are just above the cutoff.
  - If this assumption holds, then the pre-determined characteristics of those observations in both groups should be similar.

- Bandwidth
  - the narrower the bandwidth, the more likely the assumption is to hold.
  - However, the narrower the region is, the less data we have.
  - Cross-validation can be applied to choose an appropriate bandwidth.

- Functional Form
  - there is no value of X at which you observe both treatment and control observations
  - Thus, we need to know the functional form of the regression model
  - It is reasonable to believe that $E[Y_{1i}|X_i]$ and $E[Y_{0i}|X_i]$ varies differently with X_i
  - To allow for this we can estimate the more flexible regression:
$$Y_i = \alpha + \beta D_i + \gamma_1 D_i \times X_i + \epsilon_i$$
The slope w.r.t. X depends on D, i.e. different slope for control and treatment group.

  - If the function takes in non-linear form, then we typically use $X_i - c$ instead of X_i

    $$Y_i = \alpha + \beta D_i + \gamma_1(X_i - c) + \delta_1 D_i \times (X_i - c) + \gamma_2(X_i - c)^2 + \delta_2 D_i \times (X_i - c)^2 + \epsilon_i$$

    This means we allow different slopes at different distance from the cutoff.
  - It is not recommended to use high-order polynomials.
  - Choosing a functional form means we let observations very far from the cutoff determine what we predict about observations very close to the cutoff.
  - But if we ignore those points far from the cutoff and the bandwidth is narrow enough, the treatment is completely random, we don't even need to control for running variable.

- Sharp RD
  - treatment is a deterministic function of an assignment variable X

    $$E[Y_{0i}|X_i] = \alpha + f(x_i)$$
    $$E[Y_{1i}|X_i] = E[Y_{0i}|X_i] + \tau$$

    where τ is the treatment effect.
  - RCT is basically an RD with flat trend
  - When zooming close enough to the narrow neighborhood of the cutoff, $E[Y_{1i}|X_i]$ and $E[Y_{0i}|X_i]$ are flat.

- Sorting
  - If subjects can choose to be just below or just above the threshold, then that will bias our outcome.
  - If subjects can choose to be below or above, but not just below or just above the threshold, then this is not a problem.

- How to test for sorting:
  - Test 1:
    - Test whether the observable covariates are balanced at the threshold
    - if it is balanced, then it is likely that unobservable characteristics are balanced.
  - Test 2:
    - Examination of the density of assignment variable itself
    - If there are a lot more observations in one side of the threshold, then it is possible that there is sorting problem.

- Who are we estimating the effect for?
  - We are only estimating the effect for individuals at the cutoff.
    - This is especially true when we allow different slopes at different running variable values.
  - In sharp RD, we are measuring ATE. This is just like an experiment, random assignment, 100% take-up.

- Biases that RD Eliminates
  - Since the treatment is as good as randomly assigned, there is no OVB.
  - There is no reverse causality as well since the treatment is effectively randomly allocated.
  - Attenuation bias could still exist.

  - Benefit of RD is that we can find evidence for a lot of treatments we cannot randomize.

- Fuzzy RD
  - exploits discontinuities in the probability of treatment conditional on an assignment variable
  - D is no longer deterministically related to crossing a threshold but there is a jump in the probability of treatment at c
  - i.e. if running variable passes the threshold, the probability that it gets the treatment increases.
  - It is similar to IV because it uses running variable to predict the probability of getting treatment.
    $$P[D_i = 1|Z_i] = g_0(Z_i) + [g_1(Z_i) - g_0(Z_i)] \times 1[Z_i > c]$$
    Now we can use $1[Z_i > c]$ as an IV
  - First stage:
    $$D_i = \alpha + \delta 1[Z_i > c] + g(Z_i - c) + u_i$$
    $g(Z_i - c)$ control for how far the score from cutoff
  - Second stage:
    $$Y_i = \alpha + \tau_{2SLS} D_i + f(Z_i - c) + \epsilon_i$$
  - Reduced from:
    $$Y_i = \mathrm{\alpha} + \mathrm{\tau}_{RF} 1[Z_i > c] + f_{RF}(Z_i - c) + \mathrm{\epsilon}_i$$
  - Since $1[Z_i > c]$ is the IV, it must satisfy all assumption of IV
  - Then we can interpret $\mathrm{\tau}_{2SLS}$ as LATE
  - $\mathrm{\tau}_{RF}$ can be thought as ITT