1 CEF

- Conditional expectation
  - $E[Y_i|X_i = X] = \int t f_Y(t|X_i = X)dt$ (continuous)
  - $E[Y_i|X_i = X] = \sum tP(Y_i = t|X_i = X)$ (discrete)
- LIE: $E[Y_i] = E_X[E[Y_i|X_i]]$ (i.e. weighted average of averages)
  Proof. Suppose Y_i and X_i are continuous and have joint density $f_{X,Y}(u,t)$, marginal densities $g_Y(t)$ and $g_X(u)$, and conditional distribution of Y_i, i.e. $f_Y(t|X_i = u)$.

$$\begin{aligned}
E_X[E[Y_i|X_i]] &= \int E[Y_i|X_i]g_X(u)du \\
&= \int \left[\int t f_Y(t|X_i = X)dt \right] g_X(u)\ du \\
&= \int \int t f_Y(t|X_i = X) g_X(u)\ du\ dt \\
&= \int t \left[\int f_Y(t|X_i = X) g_X(u)\ du \right] dt \\
&= \int t \left[\int f_{XY}(u,t)du \right] dt \\
&= \int t\ g_Y(t)dt \\
&= E[Y_i]. \text{ hom}
\end{aligned}$$

- CEF decomposition property
  - We can decompose a random variable Y_i into two parts
    $Y_i = E[Y_i|X_i] + \varepsilon_i$
  - Two properties
    (1) $E[\varepsilon_i|X_i] = 0$
    (2) $Corr(\varepsilon_i, h(X_i)) = 0$ for any function $h(\cdot)$
  - CEF is a good summary of the relationship between X_i and Y_i as it gives us the conditional mean.

- Why we would want to use linear regression to estimate CEF

1. Linear CEF theorem
   If the CEF is linear, then linear regression of Y_i on X_i estimates the CEF.
   Two common cases of linear CEF:
   (1) Joint normality of Y_i and X_i
   ⇒ This case has limited empirical relevance, since regressors and dependent variables are often discrete, while normal distributions are continuous.
   (2) Saturated regression models (i.e. having both main effects *and* the interaction terms)
   ⇒ By having a separate parameter for every possible combination of values that the set of regressors can take on. E.g., a regression model with only dummy variables.
2. Best Linear Predictor theorem
   The function $X_i'\beta$ is the best (i.e. min. MSE) linear predictor of Y_i given X_i.
   Proof. $\beta = E[X_iX_i']^{-1}E[X_iY_i]$ solves the population least squares problem, i.e. $\beta = \arg\min_b E[(Y_i - X_i'b)^2]$
   ⇒ Just as the CEF $E[Y_i|X_i]$ is the best (i.e. min. MSE) predictor of Y_i given X_i in the class of *all* functions of X_i, the population regression function is the best we can do in the class of *linear* functions.
3. Regression CEF theorem
   The function $X_i'\beta$ provides the MMSE linear approximation to $E[Y_i|X_i]$.

$$\beta = E[X_iX_i']^{-1}E[X_iY_i] = \arg\min_b E[(E[Y_i|X_i] - X_i'b)^2]$$

⇒ Even if the CEF is nonlinear, regression provides the best linear approximation to it. ⇒ A good way to motivate regression in line with the effort to describe the essential features of statistical relationships without necessarily trying to pin them down exactly.

2 Rubin's Causal Model

- Treatment effect:

$$\tau_i = Y_i(1) - Y_i(0),$$

⇒ The fundamental problem of causal inference is that we cannot observe both $Y_i(0)$ and $Y_i(1)$ for the one individual.

- Average Treatment Effect (ATE):

$$\overline{\tau}_i = E[\tau_i] = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

⇒ However, from actually observed we can get only $E[Y_i|D_i = 1]$ and $E[Y_i|D_i = 0]$. Therefore, we need RCT to calculate the ATE.

- Randomized Controlled Trial (RCT)
  Outcome can be expressed in potential outcomes

$$Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0).$$

Hence,

$$E[Y_i|D_i = 1] = E[D_i Y_i(1) + (1 - D_i)Y_i(0)|D_i = 1] = E[Y_i(1)|D_i = 1]$$
$$E[Y_i|D_i = 0] = E[D_i Y_i(1) + (1 - D_i)Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 0]$$

In a randomized controlled trial (RCT), the potential outcomes are independent of the treatment indicator. That is,

$$E[Y_i(1)|D_i = 1] = E[Y_i(1)] \quad \text{and} \quad E[Y_i(0)|D_i = 0] = E[Y_i(0)].$$

Now,

$$\begin{aligned}
\overline{\tau}_i &= E[Y_i(1)] - E[Y_i(0)] \\
&= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0] \quad \text{(by randomization)} \\
&= E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \quad \text{(by expressing as potential outcomes).}
\end{aligned}$$

- Selection bias

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{Observed difference}} = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

$$= \underbrace{E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]}_{\text{Average treatment effect on the treated}} + \underbrace{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]}_{\text{Selection bias}}$$

- It is important to look into the data generation process to discuss selection bias.

- Treatment on the treated (TOT):

$$\overline{\tau}_{TOT} = E[Y_i(1)|D_i = 1 - E[Y_i(0)|D_i = 1]]$$

⇒ Heterogeneity of treatment effects
- used when $Y_i(0)$ is independent of treatment but $Y_i(1)$ is not.
(i.e. $E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1] = E[Y_i(0)]$
but $E[Y_i(1)|D_i = 0] \neq E[Y_i(1)|D_i = 1]$)

- Stable Unit Value Treatment Assumption (SUVTA)
  Key assumption: no spillovers
  "If $D_i = D_i'$, then $Y_i(\mathbf{D}) = Y_i(\mathbf{D}')$,"
  where $\mathbf{D}$ is a N-length vector of treatment assignments.
  ⇒ In two different worlds, if the same individual has the same treatment assignment, the outcomes are the same, regardless of whether other individuals' treatments are different or the same.

3 Cautionary Notes

- Lalonde (1986)
  - Applied modern nonexperimental tools and compared them to benchmark (unbiased) estimates from RCT.
  - All of them failed.
  - Two fundamental problems
    1. Although assigned randomly, people selected into treatment. ⇒ people in two groups are likely to be systematically different).
    2. Even in the absence of treatment, those who had earnings below the cutoff can have mean reversion.
       (Note: in such case, controlling for pre-treatment earnings worsens the problem).
       (Note: a typical case is that several different treatments lead to the same outcomes)

- Freedman (1991)
  - Possibilities for Causal Inference

  (1) regression usually works but can go wrong

  (2) regression sometimes works but not for routine use

  (3) regression might work but hasn't yet

  (4) regreesion can't work

  - Freedman presents several problematic papers, but introduces one well-done piece:
  Snow got the correct "counterfactual" since people were not selecting into treatment but chose the water company before the quality improvement (treatment). Therefore, the treated and control groups look similar.

- Truncation Model

4 Selection on Observables

(aka. uconfoundedness assumption, conditional independence assumption)

- Selection on Observables

$$\big(Y_i(1), Y_i(0)\big) \perp D_i | X_i$$

  - In words: conditional on observed characteristics ($X_i's$), selection bias disappears.
  - This "weaker" assumption is necessary for justifying causal interpretation of regression estimates in observational studies, when random assignment ($(Y_i(1), Y_i(0)) \perp D_i$) clearly fails.
  - Combined with the overlap assumption ($0 < P(D_i = 1|X_i) < 1$: for any given X_i, there are both treated and untreated groups so that we can compare them), we refer to it as "strongly ignorable treatment assignment."
  - With SOO, we can have a causal interpretation of "conditional-on-X_i" comparisons:

$$E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0] = E[Y_i(1) - Y_i(0)|X_i]$$

- What X's to include
  - What covariates should be included in the SOO designs (e.g. OLS, p-score, matching, nonparametric regression)?
  - What covariates do not violate the SOO assumption?
  - Don't include anything being a channel from treatment to outcome (i.e. part of the causal chain).
  e.g. "alcohol tax → (drink alcohol) → fatalities" or "training program → (computer skills) → earnings"
  - Rule of thumb: anything determined (and measured) before treatment assignment will be okay.

- Caveat
  - SOO very often fails to get the right answer (e.g. Lalonde (1986, NSW), Kreuger (1993, computers), Arceneaux, Gerber and Green (2006,voters))
  - But sometimes SOO is the only option.
  - The key is to consider the sources of bias and acknowledge the limitations of my study.

5 Regression Adjustment

- To get consistent estimates of the treatment effect, we need two assumptions:

  (A1) **Selection on Observables**

  (A2) We know the functional form $h(\cdot)$ of CEF.

  $\Rightarrow$ (A2) can be solved with sufficient data. (A1) is more of a problem in empirical works.

- Multi-category D_i (i.e. not binary)
  Translating the potential outcome framework into classical linear regression model:

$$\begin{aligned} Y_i(0) &= \alpha + \varepsilon_i \\ Y_i(1) &= Y_i(0) + \beta + \beta_i \quad (\beta_i = 0 \text{ if constant treatment effect}) \end{aligned}$$

$$\begin{aligned} \Rightarrow Y_i &= D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0) \\ &= D_i \cdot (Y_i(0) + \beta) + (1 - D_i) \cdot Y_i(0) \\ &= \beta D_i + Y_i(0) \\ &= \alpha + \beta D_i + \varepsilon_i. \end{aligned}$$

  Here, we need $E[\varepsilon_i D_i] = 0$ to run OLS, which does not hold in observational studies.
  - So, we instead estimate this regression form

$$Y_i = \alpha + \beta D_i + \delta h(X_i) + \varepsilon_i \quad (\text{where} \quad h(X_i) = E[D_i | X_i]).$$

  By partialing out $h(X_i)$ from D_i, we have

$$Y_i = \alpha + \beta \eta_i + \varepsilon_i \quad (\text{where} \quad \eta_i = D_i - E[D_i | X_i]).$$

  To get consistent estimates of β, we need $E[\varepsilon_i \eta_i] = 0$, which can obtained by SOO (i.e. "$\varepsilon_i \perp D_i | X_i$").
  $\Rightarrow$ The key assumption here is the observable characteristics X_i are the only reason why ε_i and D_i are correlated.

- Regression adjustment in practice
  - Krueger (1993) conducts five robust analyses on treatment effect of computer skills on wage, but they turn out to be WRONG.
    The problem is not about functional forms, but not being able to adjust for omitted variables.
  - DiNardo and Pischke (1997) conduct a placebo test by measuring the treatment effect of pencil usage on wage.

- Heterogeneous treatment effects
  Now we have,

$$\begin{aligned} Y_i(0) &= \alpha + g_0(X_i) + \varepsilon_i \\ Y_i(1) &= Y_i(0) + \underbrace{\tau + g_1(X_i)}_{\text{treatment effect}} \end{aligned}$$

$$\begin{aligned} \Rightarrow Y_i &= D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0) \\ &= D_i \cdot \big(Y_i(0) + \tau + g_1(X_i)\big) + (1 - D_i) \cdot Y_i(0) \\ &= \alpha + \tau D_i + g_0(X_i) + g_1(X_i) D_i + \varepsilon_i \end{aligned}$$

  Then,

$$\begin{aligned} f_1(X_i) &= \alpha + \tau + g_0(X_i) + g_1(X_i) + \varepsilon_i \quad (\text{treatment group}) \\ f_0(X_i) &= \alpha + g_0(X_i) + \varepsilon_i \quad (\text{control group}) \end{aligned}$$

  Average treatment effect:

$$E[Y_i(1) - Y_i(0) | X_i] = f_1(X_i) - f_0(X_i)$$

  To estimate this,

$$\overline{\tau} = \frac{1}{N} \sum_{i=1}^{N} \big(D_i Y_i + (1 - D_i) \hat{f}_1(X_i)\big) - \big((1 - D_i) Y_i + D_i \hat{f}_0(X_i)\big)$$

  $\Rightarrow$ Taking the average of the vertical differences between an observation's outcome and its counterfactual (sometimes $\hat{f}_1(X_i)$, sometimes $\hat{f}_0(X_i)$)

6 Nonparametric Regression

Although (with SOO assumption) OLS gets us an approximation even when CEF is not linear, we use nonparametric regression to estimate: for example, $E[Y_i|D_i]$ when D_i is continuous or $E[D_i|X_i]$.
Three flexible ways to estimate a CEF:

1. Series regression (parametric approach)
   We want to estimate $E[Y_i|X_i]$.
   By Taylor approximation around 0 (i.e. Mclauren series), we get the regression equation
   $$Y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon_i.$$
   $\Rightarrow$ not parsimonious and weird behavior due to high order terms.
   Alternative: Spline

   $$Y_i = \beta_0+\beta_1 X_i+\beta_2 X_i^2+\beta_3 X_i^3+\beta_4 1(X_i > K_1)(X_i-K_1)^3+\beta_5 1(X_i > K_2)(X_i-K_2)^3$$

   - Cubic is the most commonly used.
   - Pick the "knots" (K_1 and K_2) where the third derivative can change.

$\Rightarrow$ parametric approaches cannot fit the data well when the underlying CEF is not continuous (e.g. $E[D_i|X_i]$) Alternative is fully nonparametric approach to estimating CEF.

2. Kernel regression
   We want to estimate the pdf $f(X)$.
   We can use histograms, but they are jumpy while the pdf is probably smooth.
   So, we try the kernel density estimator as a smooth approximation of $f(x)$.

   - Kernel density estimator:

   $$\hat{f}_h(X) = \frac{1}{N}\sum_{i=1}^{N} hK\left(\frac{X - X_i}{h}\right)$$

   where $K(\cdot)$ is the kernel.
   $\Rightarrow$ If we use a triangular kernel, we are basically picking a point (individual), drawing a triangle, and averaging the heights. (Basically doing the same things as a histogram.)

   - Two things to consider:

     1) Kernel choice (doesn't matter much)

     2) Bandwidth choice (more of an issue)
        $\begin{cases}\text{If too large, low variance but high bias (undersmooth)}\\ \text{If too small, low bias but high variance (oversmooth)}\end{cases}$

        How to pick h?

        (1) Visual inspection: plot $\hat{f}_h(X)$ with different bandwidths.

        (2) Minimize $ISE(X) = \int \left[\hat{f}_h(X) - f(X)\right]^2 dX$
        - The problem is that we usually don't know the true $f(x)$.
        - With a big assumption that $f(X)$ is Normal, we can get a Silverman plug-in. (This can provide a useful start off point for visual inspection).

        (3) Cross validation

        $$CV(h) = \frac{1}{N^2 h}\sum_i\sum_j \int K\left(\frac{X_i - X_j}{h} - t\right)K(t)dt - \frac{2}{N}\sum_i \hat{f}_{-i,h(X_i)}$$

        where $\hat{f}_{-i,h(X_i)}$ is $\hat{f}(\cdot)$ estimated with observation i dropped. - A data-driven procedure that is computationally intensive.

- Kernel regression

  - We want to estimate $f(Y|X)$ in the "$E[Y|X] = \int Y f(Y|X)dY$" without parametric assumptions.

  - The conditional density is

  $$\hat{f}_h(Y|X) = \frac{\hat{f}_h(X,Y)}{\hat{f}_h(X)}$$

  where

  $$\hat{f}_h(X,Y) = \frac{1}{Nh^2}\sum_{i=1}^{N} K\left(\frac{X - X_i}{h}, \frac{Y - Y_i}{h}\right)$$

  Note: We can put different bandwidths for X and Y if they have different variations.

- Our CEF is basically a weighted average of Y_i.

$$E[Y_i|X_i] = \int Y \hat{f}_h(Y|X)dY = \frac{\sum K\left(\frac{X-X_i}{h}\right) \cdot Y_i}{\sum K\left(\frac{Y-Y_i}{h}\right)} \quad \begin{matrix} \longrightarrow \text{weighted sum} \\ \longrightarrow \text{sum of weights} \end{matrix}$$

- Potential problems of kernel regression
  (1) problems with sparse areas (need h to be adaptive in distance)
  (2) sensitive to outliers (instead of average, need to use fitted values from weighted regression)
  (3) poor tail performance

  $\Rightarrow$ Lowess can solve these problems.

3. Lowess

1) For each data point X_i, run a weighted regression of Y_j on X_j and X_j^2 with weight $K(X_i, X_j) = 1\left(\left|\frac{X_i - X_j}{h_i}\right| < 1\right)\left(1 - \left(\frac{X_i - X_j}{h_i} < 1\right)^3\right)^3$ where h_i is the distance to the r-th nearest neighbor.
   $\Rightarrow$ Now, r is the tuning factor. (too large $r \rightarrow$ oversmooth; too small $r \rightarrow$ undersmooth.)
2) Let $\hat{\varepsilon}_i = Y_i - \widehat{E[Y_i|X_i]}$ (this lets us identify outliers).
   Define weight $\delta_j = 1\left(\left|\frac{\hat{\varepsilon}_i}{6s}\right| < 1\right)\left(1 - \left|\frac{\hat{\varepsilon}_i}{6s}\right|^2\right)^2$ where s is the median of $\hat{\varepsilon_j}$ (this downweights the outliers).
3) Generate estimates of $\widehat{E[Y_i|X_i]}$ for each observation by regressing Y_j on X_j and X_j^2 with weights $\delta_j K(X_i, X_j)$.
4) Loop over steps 2-3 t times (until $\delta_j K(X_i, X_j)$ stabilizes). On the last iteration, estimate $E[Y|X]$ at every level of X

$\Rightarrow$ When to use?

- Matching

- Assumption: Selection on observables

- For each treated unit, find a unit or units that are untreated with the same X.

$$\tau(X) = E[Y_i(1) - Y_i(0)|X_i = X]$$

- Matching constructs a valid counterfactual under the **SOO assumption**.

$$\hat{\tau}_m = \frac{1}{N_T} \sum_{i \in \{D_i = 1\}} [Y_i - \underbrace{\sum_{j \in \{D_i = 0\}} w_i(j) Y_j}_{\text{counterfactual}}]$$

where N_T is the number of treated units and $w_i(j)$ are weights.

- For 1-1 matching, pick $w_i(j) = 1$ for closet control unit.

- If X is higher dimension, use Euclidean space $(X_i - X_j)'(X_i - X_j)$

- We would want to use more than one unit to get counterfactual, since this reduces variance (since larger N) and reduces bias by interpolating.

$$w_i(j) = \frac{K(X_i - X_j)}{\sum_j K(X_i - X_j)}$$

7 Propensity Score

- Propensity score is the conditional probability of treatment.

$$P(X_i) = E[D_i|X_i]$$

- Key assumption:

(A1) $Y_i(0), Y_i(1) \perp D_i|X_i$ (Selection on observables)

(A2) $0 < P(D_i = 1|X_i) < 1$ (Overlap)

- Theorem:

If $Y_i(0), Y_i(1) \perp D_i|X_i$, then $Y_i(0), Y_i(1) \perp D_i|P(X_i)$

- Propensity Score Estimation
  - Estimate a flexible logit of D_i on X_i, X_i^2, and interactions.
  - Get predicted values $\Rightarrow \bar{p}(X_i)$
  - For this to work, we need:

  (1) Selections on observable (i.e. correct X's)

  (2) Logit needs to be right (i.e. correct functional form)

- How to use p-score

  - Blocking

  Split into K blocks and drop the blocks where there are either only treated or only controlled units (i.e. by imposing (A2) we can force common support). Get the weighted average of within-block average p-scores by the weight as the number of observations in that block.

$$\hat{\tau}_B = \sum_{k=1}^{K} \Big(\frac{N_{T,k} + N_{C,k}}{N} \Big) \hat{\tau}_k$$

  Note: $\hat{\tau}_k$ is equal to zero when there are only treated or only controlled units.

  - Nearest neighbor matching

  Pair each treated unit with nearest control unit.

$$\hat{\tau}_{NN} = \frac{1}{N_T} \sum_{i \in \{D_i = 1\}} (Y_i - Y_j)$$

  where Y_j is the nearest neighbor to Y_i.

  - Weighting with p-score

  Run WLS on $Y_i = \alpha + X_i\beta + \tau D_i + u_i$ with regression weights

$$w_i = \sqrt{\frac{D_i}{\hat{p}(X_i)} + \frac{1 - D_i}{1 - \hat{p}(X_i)}}$$

- Overlap assess with p-score

  - Histograms are useful when we have a single X. However, more than two covariates make the problem difficult. For example, if we use marginal distributions of $f(X_1, X_2)$, we might conclude there is shared support when there is none.

  - With p-score, we can reduce the dimension of the problem.

  - Fit the logit regression to estimate $p(X_i)$ very flexibly.

  - By assessing overlap, we can trim the data. (Large sample is not always good, especially when a large part of it is just producing bias).

8 Selection on Unobservables

- Comparison: SOO vs. SOU
  Talking in terms of correlation,
  - SOO removes all bad variation (i.e. the part of treatment affected by observable characteristics), leaving all good variation in the treatment. (recall $\eta_i = D_i - E[D_i|X_i]$)
  $\Rightarrow$ SOO designs: OLS, propensity score, matching, nonparametric regression

  - SOU admits we can't remove all bad variation in treatment.
  - Hence, SOU tend to be much less precise than SOO (esp. evident in IVs), because we are finding some subset of good variation.
  $\Rightarrow$ SOU designs: Instrument variable (IV), Diff-in-Diff (DD), Regression discontinuity (RD), synthetic control

9 Structural Estimation

- We should be very specific how the world works (e.g. functional form of our model, distribution assumptions, etc.)
- If the model is correct, it can do good estimation. However, if the belief goes wrong, it gets really messy.
- It is very sensitive to assumptions

10 Panel Models

- Panel (or longitudinal) data

  - N units and T time periods (usually, $N >> T$; when $N < T$, the problem turns from identification (i.e. getting correct β) to getting correct s.e.)

  - Inference issue: getting correct s.e. (consider "serial correlation")
  - Model:

$$Y_{it} = X_{it}'\beta + c_i + \varepsilon_{it}$$

  where X_{it} are (observed) time-variant covariates and c_i captures (unobserved) time-invariant individual effect.

- Key assumptions

  (A1) Strict exogeneity

$$E[\varepsilon_{it}|X_{i1}, X_{i2}, \ldots, X_{iT}, C_i] = 0$$

  (A2) Uncorrelated effect

$$E[c_i|X_{i1}, X_{i2}, \ldots, X_{iT}] = 0$$

  - In most empirical settings, (A2) is problematic if X_i (e.g. education attainment) and c_i (e.g. personal motivation) are correlated.
  - Not a problem if we have randomization.

  (A3)

$$\Omega = \sigma_\varepsilon^2 I + \sigma_c^2 i_T i_T'$$

  $\Rightarrow$ If (A1) and (A2) hold, use OLS, GLS, or FGLS, since they use all the variation in X. (A3) is necessary for GLS.

- OLS

  - Under (A1) and (A2), OLS estimator is consistent.

  - However the s.e. is wrong—typically too small, as OLS assumes independence (i.e. homoskedastic errors) within individual across time. (That is, by taking $N * T$ as the number of observations, OLS assumes there is randomization in every t so that there is no correlation across time.)

  - Therefore, we use **robust (cluster) variance**.

- GLS (Random effect) - [check 211A notes!!!]

  - Aka. uncorrelated effect in the sense that X_i and c_i are uncorrelated.

  - More precise than OLS.

  - Intuition: taking weighted sums of sample averages and putting more weight on those with more precision (e.g. inverse of variance as weights). In the nonpanel context, WLS is a version of GLS.

- Feasible GLS

  - Use FGLS when we don't know structure of var-cov matrix Ω.

  - Procedure:

  1) Run OLS.

  2) Set residuals $\hat{V}_{it} = Y_i - X_{it}'\hat{\beta}$.

  3) Set $\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N} \hat{V}_i \hat{V}_i'$.

$$\hat{\beta}_{FGLS} = (\sum_i X_i'\hat{\Omega}^{-1}X_i)^{-1}(\sum_i X_i'\hat{\Omega}^{-1}Y_i)$$

  - $\hat{\beta}_{FGLS}$ is still consistent but less precise than correctly specified $\hat{\beta}_{RE}$, since $\hat{\Omega}$ might be imprecise. (Note: $\hat{\beta}_{FGLS} = \hat{\beta}_{RE}$, if $T = 2$).

  - This is why we don't just jump into FGLS, but use RE when structure is known (since FGLS will be slower getting true).

11 Fixed Effect

- We use FE when "uncorrelated effect" assumption fails.

- Key assumptions for FE:

  (A1) $E[\varepsilon_{it}|X_{i1}, X_{i2}, \ldots, X_{iT}, C_i] = 0$ (strict exogeneity)

  (A2) $E[\varepsilon_{it}\varepsilon_{it}'|X_{i1}, X_{i2}, \ldots, X_{iT}, C_i] = \sigma^2 I_T$ (for the sake of mathematic simplicity)

  $\Rightarrow$ The idea behind FE is we either control for c_i or difference it out.

- FE with dummies

$$Y_{it} = X_{it}'\beta + R_i'C + \varepsilon_{it}$$

  where R_i is a vector of dummies for each individual.

  - Dummies absorb the fixed effect.

  - Estimated coefficients on the dummies are not consistent because they use only T obs.

  - This is computationally expensive

- Within estimator

$$\ddot{Y}_{it} = \ddot{X}_{it}'\beta + \ddot{\varepsilon}_{it}$$

  where $\ddot{Y}_{it} = Y_{it} - \overline{X}_i$.

  - Regressing deviations in Y (from unit-specific averages) on deviations in X.

  - Much faster than FE (with dummies)

- Difference estimator

$$\Delta Y_{it} = \Delta X_{it}'\beta + \Delta\varepsilon_{it}$$

  - When $T = 2$, this is identical to FE and within estimators.

  - Long differences: measurement error by time, can get less precise but more unbiased.

- FE vs. RE

  - FE is consistent under much weaker assumptions. That is, it doesn't need the (typically) implausible assumption of uncorrelated effect.

  - FE is much less precise as it uses less variation in X. As we difference much of it out, we use only time-variant variation in X.

- Between estimator

$$\overline{Y}_i = \overline{X}_i'\beta + \overline{\varepsilon}_i$$

- RE estimator

$$\hat{\beta}_{RE} = \hat{F}^W \cdot \hat{\beta}_{FE} + (1 - \hat{F}^W) \cdot \hat{\beta}_B$$

  where

$$\hat{F}^W = \left[S^W_{X'X} + \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \sigma^2_c} \cdot S^B_{X'X}\right]^{-1} S^W_{X'X}$$

$$S^W_{X'X} = \sum_i \sum_t (X_{it} - \overline{X}_i)(X_{it} - \overline{X}_i)'$$

$$S^B_{X'X} = \sum_i T(X_i - \overline{X})(X_i - \overline{X})'$$

  - As σ^2_c increases (i.e. more correlation in time-invariant c_i across periods), $\hat{F}^W \to 1$, or $\hat{\beta}_{RE} \to \hat{\beta}_{FE}$

  - It $\sigma^2_c = 0$ (i.e. no individual effect), $\hat{F}^W = \frac{S^W_{X'X}}{S^W_{X'X} + S^B_{X'X}}$ (the proportion of variation that is within)

- Measurement error

  - With measurement error, the OLS estimator underestimates.

$$plim(\hat{\beta}_{OLS}) = \beta \cdot \frac{\sigma^2_{X'X}}{\sigma^2_{X'X} + \sigma^2_\varepsilon}$$

  where σ^2_ε is the variation of the measurement error.

  - For FE estimator,

$$\begin{aligned} plim(\hat{\beta}_{FE}) = plim(\hat{\beta}_\Delta) &= \frac{Cov(\Delta X, \Delta Y)}{Var(\Delta X)} \\ &= \frac{Cov(\Delta X^* + \Delta\varepsilon, \Delta X^*\beta + \Delta u)}{Var(\Delta X^* + \Delta\varepsilon)} \\ &= \beta \cdot \frac{\sigma^2_X(1 - \rho_X)}{\sigma^2_X(1 - \rho_X) + \sigma^2_\varepsilon(1 - \rho_X)} \end{aligned}$$

  where $\rho_X = \frac{Cov(X^*_{it}, Y^*_{it})}{Var(X^*_{it})}$.
  $\Rightarrow$ What we want: $\sigma^2_X \uparrow$(more variation in X) and $\sigma^2_\varepsilon \downarrow$(less m.e.).

12 Difference in Differences

- When to use?

  - Two states: $s = 0$, $s = 1 \Rightarrow$ one of them is affected by a policy change

  - Two time periods: $t = 0$ (pre-policy change), $t = 1$ (post-policy change)

- Key assumption:

$$E[\varepsilon_{11} - \varepsilon_{10}] = E[\varepsilon_{01} - \varepsilon_{00}]$$

  where ε_{st} is state-time exogenous shock.
  $\Rightarrow$ in words: two states must be on similar trajectories

- DD estimator

$$\begin{aligned} \hat{\beta}_{DD} &= \text{treatment unit}\Delta - \text{control unit}\Delta \\ &= (\overline{Y}_{11} - \overline{Y}_{10}) - (\overline{Y}_{01} - \overline{Y}_{00}) \\ &= \tau + (\varepsilon_{11} - \varepsilon_{10}) - (\varepsilon_{01} - \varepsilon_{00}) \end{aligned}$$

  Implementation

$$Y_{ist} = \alpha + \tau D_{st} + \gamma_1 1(s = 1) + \delta_1 1(t = 1) + \varepsilon_{st} + u_{ist}$$

- Triple differences (difference in differences in differences)

$$\hat{\beta}_{DDD} = [(\overline{Y}_{111} - \overline{Y}_{110}) - (\overline{Y}_{101} - \overline{Y}_{100})] - [(\overline{Y}_{011} - \overline{Y}_{010}) - (\overline{Y}_{001} - \overline{Y}_{000})]$$

  Implementation

$$Y_{ista} = \alpha + \tau D_{sta} + \gamma_1 1(s = 1) + \gamma_2 1(t = 1) + \gamma_3 1(a = 1)$$

$$+\gamma_4 1(s = 1)1(t = 1) + \gamma_5 1(t = 1)1(a = 1) + \gamma_6 1(a = 1)1(s = 1) + \varepsilon_{sta} + u_{ista}$$

  where Y_{sta} is the outcome for state s, time t and crop a (e.g. 1 for corn; 0 for wheat).

13 Synthetic Controls

- When to use?

  - One treated unit vs. J potential (synthetic) control units

  - T time periods: $t = \underbrace{1, 2, \ldots, T_0}_{\text{pretreatment}}, \underbrace{T_{0+1}, \ldots, T}_{\text{posttreatment}}$

  (The treatment happens between T_0 and T_{0+1})

- Picking weights ($\mathbf{w}^*$)

  - Let $\mathbf{z_1}$ be a $(K+3) \times 1$ vector of covariates and outcomes for the treated unit in pretreatment period

$$\mathbf{z_1} = [\overline{X}_1, \underbrace{\overline{Y}_{1,1}, \overline{Y}_{1,T_0/2}, \overline{Y}_{1,T_0}}_{\text{three points of outcome}}]'$$

  where $\overline{X}_1$ is the average of covariates in pretreatment period.

  - We take only three points. If we took all the pretreatment outcomes, then it would be "overfitting."

  - $\mathbf{z_0}$ is a $(K+3) \times J$ matrix of possible control units

  - We want to find a $(J \times 1)$ vector $\mathbf{w}^*$ that minimizes the distance between $\mathbf{z_1}$ and $\mathbf{z_0}$. That is,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sqrt{(\mathbf{z_1} - \mathbf{z_0 w})' V (\mathbf{z_1} - \mathbf{z_0 w})}$$

- How to match (i.e. picking V)

  1) Normalized Euclidean distance

     Set V equal to diagonal with each element equal to variance of pre intervention/covariate.

  2) Set V equal to variance-covariance matrix of elements in $\mathbf{z}$.

  3) **Optimize our fit by examining out-of-sample prediction quality.**

     Choose V that minimizes

$$(Y_1^p - \hat{Y}_0^p)'(Y_1^p - \hat{Y}_0^p)$$

     where Y_1^p is a $T_0 \times 1$ vector of pretreatment outcomes and $\hat{Y}_0^p = Y_0^p W(V)$ is a $T_0 \times 1$ vector of pretreatment predicted outcomes.

     - Procedure: 1) Pick a V. 2) Solve for $\mathbf{w}$. 3) Compute the loss function. Repeat.

     $\Rightarrow$ We are basically checking how well our weight performs at least when we know the correct Y's.

     Plots

     - We can plot the $\mathbf{z_1}$ against $\mathbf{z_0 w}$ and visually check how the two fit in the pretreatment periods.

     - By comparing with the placebo test results, we can also make inference that the estimated gap for our treated unit is sufficiently large relative to the distribution of gaps for the states in the pool.

     $\Rightarrow$ As the hypothesis testing in frequentist inference, we're looking for sufficient distance/divergence to reject the null of no treatment effect. (an example of permutation test)

- Caveats

  - The more similarity in controlled units, the better. However, still the problem is that we have only a single treated unit. This causes a problem in terms of variance/generalization/precision.

  - Synthetic controls are similar to propensity scores/matching. Therefore, any reservations for them still apply here.

14 Instrument Variables

- Conditions for IV (+ Exclusion restriction)

  (A1) $Cov(D_i, z_i) \neq 0$

  (i.e. at least there should be some effect of z_i on D_i).

  (A2) $Cov(z_i, \varepsilon_i) = 0$

  (i.e. z_i is as good as randomly assigned).

  (A3) Exclusion restriction

  (i.e. z_i has no direct effect on Y_i;
  or z_i is uncorrelated with any other determinants of Y_i).

  - This can fail even with randomization.

  - The most attacked point of IV.

  - A remedy can be "double blinding". In some medical research, the treatment is "blinded" so that the treated individuals don't act differently.

- Reduced form ($z_i \to Y_i$)

  - z_i is assignment to treatment (in medical literature, it is called "intention to treat")

  - $\hat{\pi}_1$ is the difference in Y_i between groups with $z_i = 1$ and $z_i = 0$

$$\hat{\pi}_1 = \overline{Y}_1 - \overline{Y}_0$$

  $\Rightarrow$ This substantially <u>underestimate</u> the effect of z_i on Y_i.

  This is because not the whole but only a fraction in group with $z_i = 1$ got the treatment only because of the *treatment assignment*.

- First Stage ($z_i \to D_i$)

  - $\hat{\phi}_1$ is the fraction of population that got the treatment only because they were assigned to the treatment (i.e. "compliers").

$$\hat{\phi}_1 = \overline{D}_1 - \overline{D}_0$$

- IV estimator

$$z_i \xrightarrow[\text{First stage}]{\hat{\phi}_1} D_i \xrightarrow{\hat{\beta}_{IV}} Y_i$$

$$z_i \xrightarrow[\text{Reduced form}]{\hat{\pi}_1} Y_i$$

$$\hat{\phi}_1 \cdot \hat{\beta}_{IV} = \hat{\pi}_1 \Leftrightarrow \hat{\beta}_{IV} = \frac{\hat{\pi}_1}{\hat{\phi}_1}$$

  Recall $\hat{\beta}_{IV} = (Z'D)^{-1}(ZY)$

- 2SLS

  - Used when multiple instruments.

  <u>1st stage</u>

  Regress

$$D_i = \phi_0 + \phi_1 z_i + u_i$$

  and get fitted values

$$\hat{D}_i = \hat{\phi}_0 + \hat{\phi}_1 z_i$$

  <u>2nd stage</u>

  Regress

$$Y_i = \beta_0 + \beta_{2SLS} \hat{D}_i + \varepsilon_i$$

  - When single instrument, 2SLS is identical to IV

$$\hat{\beta}_{2SLS} = \frac{\hat{\pi}_1}{\hat{\phi}_1}$$

- Local average treatment effect (LATE)

  - Recall that IV was developed to deal with hetero. treatment effects. (Treatment effects are typically heterogeneous).

  - ATE: avg. treatment effect over entire population
  $\Rightarrow$ can be estimated unbiasedly by RCTs

  - TOT: avg. treatment effect for those treated
  $\Rightarrow$ We need counterfactual, and we can get this only when matching.

  - LATE: avg. treatment effect for compliers
  $\Rightarrow$ IV is the unbiased estimator

  Note: If homogeneous treatment effect, ATE=TOT=LATE.

- Population breakdown

  - Never takers (NT) $\Rightarrow \tau_{NT}$ small (or 0)

  - Always takers (AT) $\Rightarrow \tau_{AT}$ large

  - Compliers (C) $\Rightarrow \tau_C$ mid

  - Defiers (D) $\Rightarrow \tau_D$ negative

  Note: if we have randomization, the population proportions for the four groups are identical between control group and treatment group.

  - Compliers are the only people who change the reduced form (other than defiers; we assume defiers don't exist). IV gets the treatment effect for compliers.

$$\beta_{IV} = \frac{\text{Reduced form}}{\text{First stage}} = \frac{E[Y_i(1) - Y_i(0)|C] \cdot P_C}{P_C} = E[Y_i(1) - Y_i(0)|C]$$

  Note: if the P_C is very small, we should be concerned (weak instrument).

- Assumptions for IV

  1) (SUTVA) If $z_i = z_i^*$, then $D_i(\mathbf{Z}) = D_i(\mathbf{Z}^*)$.
     (i.e. no spillover effect in compliance with assignment)

  2) If $z_i = z_i^*$ and $D_i = D_i^*$, then $Y(\mathbf{Z}, \mathbf{D}) = Y(\mathbf{z}^*, \mathbf{D}^*)$.
     (i.e. no spillover in treatment effect)
     - e.g. easily violated when D_i is vaccination

  $\Rightarrow$ 1) and 2) are basically independence assumptions (i.e. no general equilibrium effects)

  3) z_i is randomly assigned.

  4) (Exculsion restriction) $Y(\mathbf{Z}, \mathbf{D}) = Y(\mathbf{Z}', \mathbf{D}) \quad \forall \mathbf{Z}, \mathbf{Z}', \mathbf{D}$
     (i.e. z_i has no direct effect on Y_i)
     - potentially the most concerned

  5) (covariance assumption) $E[D_i(1) - D_i(0)] \neq 0$
     (i.e. treatment assignment has some effect on treatment status)

  6) (monotonicity assumption) $D_i(1) \geq D_i(0) \quad \forall i = 1, \ldots, N$
     (i.e. no defiers)

  $\Rightarrow$ If there are defiers, the IV estimator becomes a disaster (e.g., $P_C \approx P_D$).

$$\beta_{IV} = \frac{\tau_C P_C - \tau_D P_D}{P_C - P_D}$$

- Multivalued treatments (e.g. still two groups, but ppl taking different amount of pills)

  - Instrument or treatment assignment: $z_i = 0$ or 1. ($\Rightarrow D_i(0)$ or $D_i(1)$)

  - Treatments: $D_i = 0, 1, \ldots, J$

  - Potential outcomes: $Y_i(0), Y_i(1), \ldots, Y_i(J)$

  - IV estimator:

$$\begin{aligned} \beta_{IV} &= \frac{E[Y_i|z_i = 1] - E[Y_i|z_i = 0]}{E[D_i|z_i = 1] - E[D_i|z_i = 0]} \\ &= \sum_{j=1}^{J} w_j E[Y_i(j) - Y_i(j-1)|D_i(1) \geq j \geq D_i(0)] \end{aligned}$$

  where $w_j \equiv \frac{P\left[D_i(1) \geq j \geq D_i(0)\right]}{\sum_{j=1}^{J} P\left[D_i(1) \geq j \geq D_i(0)\right]}$

  $\Rightarrow$ By counting each transitions (e.g. 0 to 1 pill, 1 to 2 pills, 2 to 3 pills, ...), we are taking a weighted sum of treatment effects.

- Multivalued instruments (e.g. assigned the number of pills)

  - $z_i = 0, 1, \ldots, K$

  - Weighted sum of treatment effects for each transitions

- Weak instruments

  (1) If z_i has small effect on D_i (i.e. weak first stage), this inflates the bias in the reduced form.

  Suppose we have bias in the reduced form (e.g. wealthy families may pull strings so that their children are more likely to get scholarships). Then,

$$plim(\hat{pi}_1) = \pi_1 + bias \Rightarrow plim(\hat{\beta}_{IV} = \frac{\pi_1 + bias}{\phi_1}) = \beta_{IV} + \frac{bias}{\phi_1}$$

  (2) If too many instrument variables, we have overfit in first stage.

Imagine a situation where we have 100 (slightly) different amounts of scholarship offers. In the 2SLS context, we would have overfitting in the first stage and the second stage would go to OLS.

$\Rightarrow$ this is not a problem in typical applications because we rarely have more than one instrument.

.

15 Regression Discontinuity

- Sharp RD design

  - D_i changes from 0 to 1 when X crosses the threshold c. ("experiment with full compliance")

  Assumptions

  (A1) $E[Y_i(0)|X_i = X]$ and $E[Y_i(1)|X_i = X]$ are continuous in X.

  - To estimate the effect of D_i on Y_i

$$\begin{aligned} \tau_{RD} &= \lim_{x \downarrow c} E[Y_i|X_i = X] - \lim_{x \uparrow c} E[Y_i|X_i = X] \\ &= E[Y_i(1) - Y_i(0)|X_i = c] \end{aligned}$$

- Fuzzy RD design

  - Probability of treatment changes at threshold

  Assumptions

  (A2) $0 < \lim_{x \downarrow c} P(D_i = 1|X_i = X) - \lim_{x \uparrow c} P(D_i = 1|X_i = X) < 1$.
  (This assumption ensures we do have a FS and rules out the Sharp RD case.)

  (A3) $D_i(X^*)$ is nonincreasing in X^* at $X^* = c$.
  (i.e. No defiers.) $\Rightarrow$ we are thinking in this way because it is often not reasonable to think that we can manipulate X for a certain individual

  - Estimate

$$\tau_{FRD} = E[Y_i(1) - Y_i(0) \mid \text{unit } i \text{ is a complier on } X_i = c]$$

  $\Rightarrow$ This gives results for a very specific subpopulation (e.g. kids with an IQ of 135 AND compliers) and thus have external validity issues. That is, it is difficult to generalize this result to the whole population.

- Before estimation
  scatter plots

- RD Estimation

  – local linear regression

  Choose bandwidth. Fit a linear regression on both side of the threshold.
  But this approach has problems related to s.e.

  – Thus in practice, we run the regression

$$Y_i = \alpha + \tau D_i + \beta_L(X_i - c) + \beta_R(X_i - c)D_i + u_i$$

  for $c - h < X_i < c + h$

  Note: we are recentering the data to get correct s.e. for $\hat{\tau}$. Use robust errors.

- Estimating fuzzy RD

$$FS: \quad D_i = \gamma_0 + \gamma_1 z_i + \gamma_2(X_i - c) + \gamma_3(X_i - c)z_i + u_i$$

$$RF: \quad Y_i = \pi_0 + \pi_1 z_i + \pi_2(X_i - c) + \pi_3(X_i - c)z_i + \varepsilon_i$$

$$\Rightarrow \hat{\tau}_{FRD} = \frac{\hat{\pi}_1}{\hat{\gamma}_1}$$

- Bandwidth choice:

  a) visual assess b) cross-validation c) plug in style
  make sure to show the robustness to h selection

- Specification testing
  Balance type table, McCrary density test

16 Inference

- Panel clustering (to deal with serial correlation)

  - Bertrand, Duflo, and Mullainathan (2004) note that many previous papers deal with panel data and assume independence within unit.

  Suppose we have the bivariate regression of a single unit observed in t time periods

  $$Y_t = \alpha + \beta X_t + \varepsilon_t$$

  Assume X_t and ε_t follow AR(1) processes with autocorrelation parameter $\lambda < 1$ and $\rho < 1$, respectively.

  As $T \to \infty$, the ratio of estimated variance (i.e. by default OLS in Stata) to true variance equals

  $$\frac{1-\rho\lambda}{1+\rho\lambda}$$

  If λ and ρ are nonzero (i.e. serial correlation), we will get a smaller s.e.

  This will lead to over rejection of the null hypothesis.'

  Bias also grows as we add more time periods. Therefore, reducing the time periods can lower the rejection rate.

- Solutions to serial correlation

  – TS route - model the AR processes

  Transformation of the data by purging the autocorrelation part.

  ⇒ this doesn't work because we are unlikely to get the structure rightly (i.e. dependence over time might be very hard to model.)

  – Collapse the data

  If the problem was too many time periods, we can collapse the data into pretreatment and postpreatment.

  a) Regress Y_{st} on state FE, year dummies and covariates and collect residuals $\tilde{Y}_{st}$

  b) Regress treatment indicator D_{st} on state FE, year dummies, and some covariates and collect residuals $\tilde{D}_{st}$.

  c) For the treated states, divide obs. into two groups: pre and post treatment. (So there are two observations per treated state)

  d) Regress $\overline{\tilde{Y}_{st}}$ on $\overline{\tilde{D}_{st}}$. Then we will get s.e. about the right size.

  – Clustered standard errors

  We now allow for an arbitrary variance-covariance matrix.

  To get reasonable estimates of this, we need G to be large. If G is too small, the estimate of matrix is poor.

  (rule of thumb: $G \geq 50$)

Rule of Thumb

- large G and small T: use clustered s.e.

- large G and large T: use clustered s.e.

- small G and large T: collapse the data

- small G and small T: collapse the data (a less challenging version of small G and large T)

- Randomization Inference (small sample):

  - Use when small sample (i.e. asymptotics not possible that we cannot use CLT or LLN)

  - Permutation test (Fisher exact test)

- Bootstrapping

  - We want to do inference on $\hat{\theta}$, our parameter of interest (e.g. sample mean)

  - What we *would* like to do is to take S estimates by randomly drawing from the population and compute

  $$\widehat{Var(\hat{\theta})} = \frac{1}{S-1}\sum_{s=1}^{S}(\hat{\theta}_s - \bar{\hat{\theta}})^2$$

  - However, we typically do not have access to that population.

  - Insight: a randomly drawn sample from the population is representative of the population.

  - Procedure: Randomly draw N obs. from the sample with replacement. (If replacement, we are taking some part of the sample distribution). Repeat this B times. Then, we have

  $$\widehat{Var(\hat{\theta})} = \frac{1}{1-B}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\hat{\theta}})^2$$

– General Procedure

Suppose we have a sample with observations $w_1, w_2, \ldots, w_N$ on which our statistic depends. So, $\hat{\theta}(w_1, w_2, \ldots, w_N)$.

1) Draw a sample from $w_1, w_2, \ldots, w_N$ with replacement.
   Call the sample $w_1^*, w_2^*, \ldots, w_N^*$.
2) Compute $\hat{\theta}$, the statistic of our interest, using the bootstrap sample.
3) Repeat 1) and 2) B times. Then compute $\widehat{Var(\hat{\theta})}$.
   We let $\hat{\theta}^* = \frac{\hat{\beta}^* - \hat{\beta}}{s^*_{\hat{\beta}}}$ define our rejection region. (i.e. we are bootstrapping the t statistic)

How large should B be?

- several thousands at least

- try several thousands twice and see if we get the same results.

– Paired bootstrap - draw pairs (Y_i, X_i)

- sample individuals keep their Y_i and X_i.

The basic idea here is that we assume independence across observations.

17 Multiple Inference problem

- What if we have multiple outcomes or treatments?

  - Suppose we have a series of M hypothesis tests. Assume all the nulls are true. Then, the probability of falsely rejecting (Type I error) at least one null hypothesis is

$$1 - (1 - 0.05)^M \approx 0.4 \quad (\text{if } M = 10)$$

  - We need a measure to generalize such Type I errors.

- False Discovery Rate (FDR)

$$FDR = E\left[\frac{\#\text{false rejections}}{\#\text{total rejections}}\right] = E\left[\frac{V}{V+U}\right]$$

  (Note: V is the number of false rejections and U is the number of correct rejections.)

- Family Wise Error Rate (FWER)
  : the probability that at least one of the true J hypotheses is rejected.

$$FWER = P(V > 0)$$

  (Note: We test M hypotheses and $J \leq M$ of them are true.

- FDR vs. FWER

  - If all the nulls are true (i.e. $U = 0$ and $J = M$), then $FDR = FWER$.

$$\begin{aligned} FDR &= E\left[\frac{V}{V+U}\right] \\ &= E\left[\frac{V}{V+U}|V=0\right] \cdot P(V=0) + E\left[\frac{V}{V+U}|V>0\right] \cdot P(V>0) \\ &= E\left[\frac{V}{V+U}|V>0\right] \cdot P(V>0) \\ &= P(V>0) = FWER \quad (\because E\left[\frac{V}{V+U}|V>0\right] = 1) \end{aligned}$$

  - When some of the nulls are false (i.e. $J < M$), then $FDR < FWER$

$$\begin{aligned} FDR &= E\left[\frac{V}{V+U}\right] \\ &= E\left[\frac{V}{V+U}|V=0\right] \cdot P(V=0) + E\left[\frac{V}{V+U}|V>0\right] \cdot P(V>0) \\ &= E\left[\frac{V}{V+U}|V>0\right] \cdot P(V>0) \\ &\leq P(V>0) = FWER \quad (\because E\left[\frac{V}{V+U}|V>0\right] \leq 1) \end{aligned}$$

  $\Rightarrow$ So if we are controlling FWER at the 0.05 level we are being more conservative than if we are controlling FDR at the 0.05 level. This increases the chances we will fail to reject null hypothesis that are not true.

- Solutions to Multiple inference problem

  1. Reduce the number of tests

     - If we have multiple testing on the RHS, we can do a single F-test. However, in this case, when we reject the joint null, we still don't know which treatment matters.

     - If we have multiple outcomes, collapse them into a single new variable. In this case, we need to know the clear directions of effect on each outcome. Again, when we reject the null that the treatment has no effect on the aggregate output, we still can't tell which of the outcomes are significant.

  2. Control FWER

     <u>Bonferroni correction</u>

     - Divide the significance level by the number of tests
     - Very conservative. Very high standard to reject the null.
     - We assume all nulls are true.
     - Doesn't allow for the possibility of dependence in outcome

     $\Rightarrow$ Limitation: very unlikely to correctly reject the null when it is false.

  3. Control FDR

     - Begin with the largest