Causal Inference

Introduction

- Comparative statics are theoretical descriptions of causal effects contained within the model. These kinds of comparative statics are always based on the idea of ceteris paribus – holding all else constant.
- Covariate balance: If we say that everything is the same except for the movement of one variable, then everything is the same on both sides of that variable's changing value.
- Summation property to remember: $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$
- Linear CEF Theorem: The conditional expectation function (CEF) is the mean of some outcome y with some covariate x held fixed, $E(y_i|x_i)$. Suppose that the CEF itself is linear. Then the population regression is equal to the CEF. This simply states that you should use the population regression to estimate the CEF when you know that the CEF is linear.
- Best Linear Predictor Theorem: Recall that the CEF is the minimum mean squared error predictor of y

given x in the class of all functions according to the CEF prediction property. Given this, the population regression function, $E\left(X'Y\right)E(X'X)^{-1}$, is the best that we can do in the class of all linear functions.

- Regression CEF Theorem: The function $X\beta$ provides the minimum mean squared error linear approximation to the CEF, ie. $\beta = \arg\min_b E\left\{\left[E\left(y_i|x_i\right) - x_i'b\right]^2\right\}$
- Regression Anatomy Theorem: Let our response be conditionally random on X and P other covariates. To estimate a causal effect, we need a sample with all of those P covariates and we need for X to be randomly assigned for a given set of the P covariates. Given regression model $i_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$, take an auxiliary regression in which variable x_{1i} is regressed on all the remaining independent variables and we let $\tilde{x}_{1i} = x_{1i} - \hat{x}_{1i}$. Then $\beta_1 = \frac{C(y_i, \tilde{x}_i)}{V(\tilde{x}_i)}$.

  Notice that again we see the coefficient estimate being a scaled covariance, only here the covariance is with respect to the outcome and residual from the auxiliary regression and the scale is the variance of that same residual. (This is our successive orthogonalization via Gram - Schmidt)
- RMSE = SE of the regression

DAGs

- To show reverse causality, one would need to create multiple nodes, most likely with two versions of the same node separated by a time index.
- DAGs may not be built to handle simultaneity
- Causal effects can be direct or mediated through a third variable.
- What makes the DAG distinctive is both the explicit commitment to a causal effect pathway, but also the complete commitment to the lack of a causal pathway represented by missing arrows.
- Why DAG? It is helpful for students to better understand research designs and estimators for the first time. A well-designed DAG can help you develop a credible research design for identifying the causal effects of some intervention.
- Backdoor path: An indirect path from X to Y when a direct path also exists. Similar to the notion of omitted variable bias in that it represents a determinant of some outcome that is itself correlated with a variable of interest. Leaving a backdoor open creates bias.
- For $X \rightarrow Y,\ X \rightarrow D,\ D \rightarrow Y$, X is a confounder since it jointly determines D and Y. X could also be unobserved (represented with dashed edges), and we say the backdoor path is open.
- Given a DAG, consider the possible edges that are absent - do these assumptions make sense? List all direct and indirect / backdoor paths from D, our covariate of interest, and Y.
- Colliders: Take DAG $D \rightarrow Y,\ D \rightarrow X \leftarrow Y$. We still count D as having two paths to Y, but here X is a collider instead of a confounder (D and Y's causal effect collide at X). X closes the backdoor path.
- Backdoor criterion: a set of variables X satisfies the backdoor criterion in a DAG if and only if X blocks every path between confounders that contain an arrow from D to Y.
  - Our goal is to close open backdoor paths, leaving a direct causal effect of D on Y. When all are closed, we say we have met the backdoor criterion. There are two ways to close: conditioning and colliders.
  - Condition on the confounder that has an open backdoor path. Conditioning requires holding the variable fixed using something like sub-classification, matching, regression, or some other method.

    - With colliders we do the opposite. Conditioning on a collider opens a backdoor path. Instead, we leave colliders alone and they already keep a backdoor path closed.
    - Satisfaction procedure:
        - Write down all paths between D and Y
        - Note open / closed for each backdoor path by checking for colliders and confounders
        - Check if backdoors can be closed through conditioning
- Notice all of this requires a theory of variable relationships and models. This is context specific knowledge.
- Collider bias
    - Controlling for tenure, job role, performance, there is no gender pay gap. But what if one of the ways in which gender discrimination creates gender disparities in earnings is through occupational sorting? Then naive regressions of wages onto a gender dummy controlling for occupation characteristics will be biased towards zero. If we condition on occupation, we close some backdoors but open others! No viable identification strategy
    - Sometimes the problem with conditioning on a collider, though, can be so severe that the correlation becomes statistically insignificant, or worse, even switches sign.

Potential Outcomes Causal Model

- Physical randomization: Causal inference, in this context, is a probabilistic idea wherein the observed phenomena is compared against permutation-based randomization called the null hypothesis, say using the Fisher Exact Test
- Potential outcomes: a causal effect is defined as a comparison between two states of the world. The difference between these two dimensions, if you would, at the same point in time represents the causal effect of the intervention itself.
- For simplicity, we will assume a dummy variable that takes on a value of one if a particular unit i receives the treatment and a zero if they do not.
- Each unit will have two potential outcomes, but only one observed outcome. Potential outcomes are defined as Y_i^1 if the unit received the treatment and Y_i^0 if the unit did not. We'll call the state of the world where no treatment occurred the control state.
- Observable outcomes, Y_i, are distinct from potential outcomes. Whereas potential outcomes are hypothetical random variables that differ across the population, observable outcomes are factual random variables. A unit's observable outcome is determined according to a switching equation: $Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$. $D_i = ifelse(treatment, 1, 0)$
- The causal effect is then defined as $\delta_i = Y_i^1 - Y_i^0$. Of course, we only observe one of these per individual
- Average treatment effect (ATE): $E[\delta_i] = E\left[Y_i^1 - Y_i^0\right] = E\left[Y_i^1\right] - E\left[Y_i^0\right]$ (still unknowable since indexed to individual i)
- ATE for treatment group (ATT): $E[\delta_i | D_i = 1] = E\left[Y_i^1 | D_i = 1\right] - E\left[Y_i^0 | D_i = 1\right]$. Still unknowable
- ATE for control (untreated) (ATU): $E[\delta_i | D_i = 0] = E\left[Y_i^1 | D_i = 0\right] - E\left[Y_i^0 | D_i = 0\right]$
- Given a dataset with (Y^1, Y^0) per individual, we can calculate δ_i and find the mean effect. Here we can fully know the ATE since we have individual data with both potential outcomes
- Now given a dataset of observed outcomes over all patients. We can find ATT and ATU by grouping and taking averages. Note $ATE = p \times ATT + (1 - p) \times ATU$

Simple Difference in Mean Outcomes (SDO)

- Simple difference in mean outcomes (SDO): take the observed values, calculate means to estimate parameter of interest, ATE. Also called naive average treatment effect (NATE). Given by $SDO = E\left[Y^1|D=1\right] - E\left[Y^0|D=0\right] \approx \frac{1}{N_T}\sum_{i=1}^{n}\left(y_i|d_i=1\right) - \frac{1}{N_C}\sum_{i=1}^{n}\left(y_i|d_i=0\right)$
    - We can have a postive ATE and a negative SDO. The SDO has a decomposition to understand this mismatch: $SDO = ATE + E\left[Y^0|D=1\right] - E\left[Y^0|D=0\right] + (1-\pi)(ATT - ATU)$
    - π is the share of individuals who received treatment.
    - The equation has three parts now: average treatment effect, selection bias, and the heterogeneous treatment effect bias.
    - Since we have ATE on the RHS, the other two terms must be the source of bias making SDO < ATE.
    - Selection Bias: the inherent differences between the two groups if they both were in control. Notice that the first term is a counterfactual, whereas the second is an observed outcome according to the switching equation.
    - Heterogeneous treatment effect bias: different returns to treatment for the two groups multiplied by the share of population in the control group.
- When assignment to treatment or control is independent of potential outcomes ($\left(Y^1, Y^0\right) \perp D$), SDO is a credible estimate of ATE. This kind of randomization of the treatment assignment would eliminate both the selection bias and the heterogeneous treatment effect bias.
- Stable unit treatment value assumption (SUTVA): the unit-level treatment effect ("treatment value") is fixed over the entire population, which means that the assignment of the treatment to one unit cannot affect the treatment effect or the potential outcomes of another unit. Implies homogenous dosage to all units, no externalities. This has significant problems for general equilibrium, since scaling an intervention to a population will surely have externalities
- Even with randomization, we may want to perform a multivariate regression. Randomization may be conditional on another variable (say which school you attend) and additional control variables, even if uncorrelated with the target effect, can increase precision of estimate of our effect since they may still highly correlate with Y and reduce the variance of our regression residuals.
- The conditional control in the regression is a fixed effect - the effect of being in a certain school is fixed across different students
- Attrition: what happends if people leave the experiment? If attrition is random, then it affects treatment and control alike and the SDO is unbiased. But attrition is often not random - good students more likely to leave if assigned to larger class size, exaggerating the effect of small class size on performance. We can imputation to fix this censored data problem, say using earlier test scores for students who drop out - if estimates are unchanged then we can infer little bias in our original data.
- Also issues when subjects switch treatment groups. Experimental design can help to make this costly. When this is infeasible, we can regress the changed groups in time 2 on the random assingment in time 1. As long as the original assignment is highly correlated with the switched assignments, we can treat the original assignment as an instrumental variable.

Matching and Subclassification

- Three different conditioning strategies (say for backdoor closures): subclassification, exact matching, approximate matching

Subclassification

- Weighting differences in means by strata-specific weights to adjust the differences in means to that their distribution by strata is the same as that of the counterfactual's strata.
- This method implicitly achieves distributional balance between the treatment and control in terms of that known, observable confounder.
- Conditional Independence Assumption (CIA): $\left(Y^1, Y^0\right) \perp D|X$ where X is the variable we are conditioning on. If CIA can be credibly assumed, then it necessarily means you have selected a conditioning strategy that satisfies the backdoor criterion. The expected values of Y^1 and Y^0 are equal for treatment and control group for each value of X.
- When treatment is conditional on observable variables, such that the CIA is satisfied, we say that the situation is one of **selection on observables**. If one does not directly address the problem of selection on observables, estimates of the treatment effect will be biased. But this is remedied if the observable variable is conditioned on.
- In smoking, had strong correlation with mortality but many worried about selection effects while randomization clearly impossible.
- How do we compare mortality among cigarette, pipe, and cigar smokers? We can use subclassification to ensure the means of the covariates are the same for each group, giving us **balanced covariates** and the groups are exchangeable wrt those covariates.
- Pipe smokers are older on average than cigarette smokers, leading to an imbalance in their death rates, since age here will be a confounder. We have covariate imbalance between groups, which we could correct in this case by conditioning on the imbalanced variable. Calculate the mortality rate by smoking type and age, then weight the rate for the treatment group by an age-specific weight that corresponds to the control group.
    - Specifically, we take the $\frac{N_t}{N}$, group totals over total people. If D_t = cig death rate for an age strata, then morality rate for smokers is $\sum_t D_t \frac{N_t}{N}$. Then to reweight to match the pipe smoker strata, if N_p are the age strata for pipe smokers, we take $\sum D_t \frac{N_p}{N}$ as the adjusted death rate for cig smokers.
- A covariate is usually a random variable assigned to the individual units prior to treatment. It is predetermined and therefore exogenous. It is not a collider, nor is it endogenous to the outcome itself (i.e., no conditioning on the dependent variable). A variable is exogenous with respect to D is the value of X does not depend on the value of D.
- Choosing the stratification variable - CIA tells us what to choose, as we want to condition on the covariate that allows all backdoors to be closed.
- If conditioning on a large number of categories (like specific age in years), you will suffer from curse of dimensionality - observations per strata will be low and there will not be common support between treatment and control. Problem compounds if we have to condition on multiple variables. Subclassification appears to be a less useful strategy here (note binning is not mentioned here but see coarsened exact matching)

Exact Matching

- What if we could impute the missing potential outcomes for each treatment unit using a control group

unit that was closest to the treatment group unit for a given confounder

- Simple matching estimator $\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - Y_{j(i)}\right)$ for $Y_{j(i)}$ the jth unit matched to the ith unit based on its closeness for X covariate. With many matches, we can average over those matches.
- This estimator is just for ATT, since we are only filling in treatment group missing values. We can estimate ATE by going in both directions - filling in treatment and control: $\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} (2D_i - 1) \left[Y_i - \left(\frac{1}{M} \sum_{m=1}^{M} Y_{jm(i)}\right)\right]$
- When two samples are exactly balanced on covariates, we say they are exchangeable

Approximate Matching

- With large number of covariates, then exact match across all of them is unlikely. Then can turn to distance based techniques like nearest neighbors. Often use normalized euclidean distance to account for different scales of predictors: $\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{V}^{-1} (X_i - X_j)}$ for $\hat{V}^{-1}$ the diagonal sample variance matrix.
- Mahalanobis distance also scale invariant: $\|X_i - X_j\| = \sqrt{(X_i - X_j)' \widehat{\Sigma}_X^{-1} (X_i - X_j)}$ replacing the diagonal sample variance matrix with the sample covariance-variance matrix.
- As the distance between a variable and its approximate match increases, we introduce bias into our estimation. Matching discrepancies tend to converge to zero as the sample size increases. The more covariates, the longer it takes for that convergence to zero to occur.
- The larger the dimension, the greater likelihood of matching discrepancies, the more data you need.
- These discrepancies are observed, so we have some idea how much bias is introduced. If we are just estimating ATT, we may be able to increase the size of control sample. Otherwise, we can apply some bias correction methods
- The bias correcting estimator given by $\hat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[\left(Y_i - Y_{j(i)}\right) - \left(\widehat{\mu}^0 (X_i) - \widehat{\mu}^0 \left(X_{j(i)}\right)\right)\right]$ where $\widehat{\mu}^0(\mathrm{X})$ estimates $E[Y|X = x, D = 0]$ via OLS.

Propensity Score Methods

- Propensity score matching is used when treatment is nonrandom but is believed to be based on a variety of observable covariates. Propensity score matching takes those covariates needed to satisfy CIA, estimates a maximum likelihood model of the conditional probability of treatment, and uses the predicted values from that estimation to collapse those covariates into a single scalar.
- This is only valuable if CIA is satisfied by conditioning on some X. If your data cannot satisfy the backdoor criterion, then you gain nothing with this technique. This of course is dependent on a complete and accurate DAG.
- The idea is to compare units who based on observable covariates had similar probabilities of being placed into the treatment group even though those units differed with regards to actual treatment assignment.
- If two units have the same propensity score, but one is the treatment group and the other is not, and the conditional independence assumption (CIA) credibly holds in the data, then differences between their observed outcomes are attributable to the treatment. The CIA says the assignment to treatment is as good as random in this case.
- Procedure
    - Estimate the propensity score

    - Select an algorithmic method to calculate ATE using propensity score. Often probit / logit models.
    - Calculate SEs
- We generate the estimated coefficients from the logit using $\Pr(D=1|X)=F\left(\beta_0+\gamma \text{ Treat }+aX\right)$. Best to use MLE to ensure values between 0 and 1. The propensity score used the fitted values from the maximum likelihood regression to calculate each unit's conditional probability of treatment regardless of their actual treatment status.
- Two indentifying assumptions: 1) CIA $\left(Y^0,Y^1\right)\perp D|X$ 2) Common support $0<\Pr(D=1|X)<1$, ie. that for each value of X, there is a positive probability of being both treated and untreated.
- $CIA \iff$ backdoor criterion met in data by conditioning on X $\iff$ conditional on X, the assignment of units to the treatment is as good as random $\iff \varepsilon_i \perp D_i|X_i$
- Note CIA is an assumption, not testable since it involved potential outcomes. However, common support is testable via histograms or data summaries.
- Propensity score theorem: if CIA ($\left(Y^0,Y^1\right)\perp D|X$) then $\left(Y^1,Y^0\right)\perp D|p(X)$ for $p(X)=\Pr(D=1|X)$, the propensity score. Conditioning on the propensity score is enough to have independence between the treatment and the potential outcomes.
- Given CIA, we can estimate average treatment effects by weighting appropriately the simple difference in means. Conditional on the propensity score, the probability that D = 1 does not depend on X any longer. That is, D and X are independent of one another conditional on the propensity score. Implies the balancing property $\Pr(X|D=1,p(X))=\Pr(X|D=0,p(X))$
- Again, all of this holds for observable covariates but does not imply we have balanced unobserved covariates, something especially pernicious in economics but maybe less so in medicine

Estimation via Propensity Score Matching

- Inverse probability weighting - weighting treatment and control units according to $\hat{p}(x)$. If propensity score is very small, we will get weights, so we trim the data of these enormous values.
- Nearest neighbor matching - you pair each treatment unit i with one or more comparable control group units j, where comparability is measured in terms of distance to the nearest propensity score. Then we take this control outcome as a matched sample and calculate $\widehat{ATT}=\frac{1}{N_T}\left(Y_i-Y_{i(j)}\right)$
- Coarsened exact matching - that sometimes it's possible to do exact matching if we coarsen the data enough. Create some cutpoints in a predictor's values and bin. Fewer strata result in more diverse observations within the same strata and thus higher covariate imbalance.
- Monotonic imbalance bounding - bound the maximum imbalance in some feature of the empirical distributions by an ex ante decision by the user. In CEM, this ex ante choice is the coarsening decision. Imbalance measured by L1(f,g) score, where f and g record the relative frequencies for the treatment and control group units. Perfect global balance is indicated by L1 = 0, while max imbalance L1 = 1.

Regression Discontinuity

- RDD is appropriate in any situation where a person's entry into the treatment group jumps in probability when some running variable, X, exceeds a particular threshold, c_0.
- We use our knowledge about selection into treatment in order to estimate average treatment effects. More specifically, since we know the probability of treatment assignment changes discontinuously at c_0, then we will compare people above and below c_0 to estimate a particular kind of average treatment

effect called the local average treatment effect, or LATE.

- The validity of an RDD doesn't require that the assignment rule be arbitrary. It only requires that it be known, precise and free of manipulation. We need a lot of data around the discontinuities which itself implies that the datasets useful for RDD are likely very large.
- Sharp - designs where the probability of treatment goes from 0 to 1 at the cutoff (Medicare enrollment). Fuzzy - designs where the probability of treatment discontinuously increases at the cutoff (class size function).
- People ordinarily think of RDD as a selection on observables observational study since a sharp RDD has treatment assignment as deterministic dependending only on $X > c_0$. Then the switching equation gives us $Y_i = Y_i^0 + \left(Y_i^1 - Y_i^0\right) D_i = \alpha + \beta X_i + \delta D_i + \varepsilon_i$
- The sharp RDD estimation is interpreted as an average causal effect of the treatment at the discontinuity, which is a kind of local average treatment effect (LATE). The key identifying assumption of RSS is the continuity assumption: $E\left[Y_i^0 | X = c_0\right], E\left[Y_i^1 | X = c_0\right]$ are continuous in X at c_0 meaning the population average potential outcomes Y^0, Y^1 are continuous. Absent the treatment, in other words, the expected potential outcomes wouldn't have jumped. The continuity assumption is not testable because it is based on counterfactuals.
- Often transform X to be shifted by c_0: $Y_i = a + \beta\left(X_i - c_0\right) + \delta D_i + \varepsilon_i$. Only changes the intercept interpretation since it absorbs βc_0
- If we have a nonlinear relationship with X, we can use a second degree polynomial in our local regression to decrease bias, ie $Y_i = f\left(X_i\right) + \delta D_i + \eta_i$. Alternatively can use kernel regression, but there can be bias at the boundary with this method
- Fuzzy RDD defined by $\lim_{X_i \to c_0} \Pr(D_i = 1 | X_i = X_0) \neq \lim_{c_0 \leftarrow X_i} \Pr(D_i = 1 | X_i = X_0)$. We use the same continuity assumptions, the conditional expectation of potential outcomes changes smoothly through c_0

Challenges to Identification

- **Continuity assumption** can be violated if it is known in advance, agents have motivation and time to adjust. The cutoff can be endogenous in that it is correlated with variables that affect the outcomes (often age cutoffs are like this)
- There are a number of tests to check the validity of RDD for your purpose
- McCrary Density Test: check for whether units are sorting on the running variable. Assuming a continuous distribution of units, manipulation would mean that more units are showing up just on the other side of the cut off. Formally, if we assume a desirable treatment D and an assignment rule $X \geq c_0$, then if individuals sort into D by choosing X to be above cutoff, then individuals are sorting on the running variable.
    - Under the null, the density should be continuous at the cutoff point. Under the alternative hypothesis, the density should increase at the kink.
    - Partition the assignment variable into bins and calculate frequencies. Treat the frequency counts as the dependent variable in a local linear regression. May be able to see an actual discontinuity in the density plot around the cutoff value
- Covariate Balance Test: For RDD to be valid in your study, there must not be an observable discontinuous change in the average values of the covariates around the cutoff. Unlike the outcome variable, the covariates should not show discontinuities. This test is basically what is sometimes called a placebo test. That is, you are looking for there to be no effects where there shouldn't be any.

- Arbitrary cutoffs - similarly, there shouldn't be effects on the outcome of interest at arbitrarily chosen cutoffs. Look at one side of the discontinuity, take the median value of the running variable in that section, and pretend it was a discontinuity, c_0'. Then test whether there is a discontinuity in the outcome at c_0'. You do not want to find anything.

Regression Kink Design

- Rather than the discontinuity creating a discontinuous jump in the treatment variable at the cutoff, it created a change in the first derivative.

Instrumental Variables

Natural Experiments

- While natural experiments are not technically instrumental variables estimator, they can be construed as such if we grant that they are the reduced form component of the IV strategy.
- An event that occurs naturally which causes exogenous variation in some treatment variable of interest
- The work may not come from the actual statistical analysis, but showing that other covariates are similar between the two groups of the natural experiments.

Intrumental Variable DAG

- Selection on unobservables - we have a backdoor path between D and Y that is caused by an unobserved variable, so we cannot use a conditioning strategy to close the backdoor.
- Now imagine there is a variable Z that acts upon D, ie $Z \rightarrow D \rightarrow Y$, and $U \rightarrow D,\ U \rightarrow Y$. Y varies when Z varies, but only because D also varies - D mediates the path from Z to Y. Z affects Y only through D.
- Imagine D consists of people making choices. Sometimes these choices affect Y, and sometimes these choices merely reflect changes in Y via changes in U. But along comes some shock, Z, which induces some but not all of the people in D to make different decisions. When those people's decisions change, Y will change too, because of the causal effect. But, notice, all of the correlation between D and Y in that situation will reflect the causal effect. The reason being, D is a collider along the backdoor path between Z and Y.
- Instrumental variables only identifies a causal effect for any group of units whose behaviors are changed as a result of the instrument. We call this the **causal effect of the complier** population. Secondly, instrumental variables are typically going to have larger standard errors, and as such, will fail to reject in many instances if for no other reason than because they are under-powered.
- **Exclusion restriction** - Z is independent of U. The IV estimator assumes that Z is independent of the variables that determine Y except for D.
- This relationship between Z and D is called the "first stage." Z and Y are correlated only because Z and D are correlated, as D is a collider along the path $Z \rightarrow D \leftarrow U \rightarrow Y$
- You can only identify a causal effect using IV if you can theoretically and logically defend the exclusion restriction, since the exclusion restriction is an untestable assumption technically - this requires a theory.
- Instruments are jarring precisely because of the exclusion restriction – these two things (gender composition and work) don't seem to go together. If they did go together, it would likely mean that the exclusion restriction was violated.

Homogeneous Treatment Effects and 2SLS

- Assumes all treatment effects are constant for all units. The parameter estimated through an IV methodology equals the ATE equals the ATT equals the ATU. The variance will still be larger, because IV still only uses part of the variation in D, but the compliers are identical to the non-compliers so the causal effect for the compliers is the same as the causal effect for all units.
- Instrumental variables methods are typically used to address omitted variable bias, measurement error, and simultaneity.
- With an omitted variable, our target D may be correlated with this missing covariate. Then in the reduced regression equation without this covariate, D is correlated with the error term, since the omitted effect is present in the error - becomes endogenous.
- Taking the effect of schooling on earnings with an omitted ability covariate, then $C(A, Z) = 0,\ C(\varepsilon, Z) = 0$. The estimated effect of schooling with IV $\hat{\delta} = \frac{C(Y,Z)}{C(S,Z)}$. If ability is independent of Z, then this first covariance is zero. And if Z is independent of the structural error term, ε, then it too is zero. This, you see, is what is meant by the "exclusion restriction": the instrument must be independent of both parts of the composite error term.
- We also need the instrument to be highly correlated with the endogenous variable, the higher the better. The numerator $\hat{\delta}$ is sometimes called the "reduced form", while the denominator is called the "first stage". Assuming our covariances above, $\hat{\delta}$ is a consistent estimator. But if that assumption fails or Z and S are not highly correlated, our estimate can be quite biased.

Two Stage Least Squares (2SLS)

- We set up a two stage regression model: $Y_i = \alpha + \delta S_i + \varepsilon_i,\ S_i = \gamma + \beta Z_i + \epsilon_i$
- See derivation of estimators on page 217.
- The 2SLS estimator used only the fitted values of the endogenous regressors for estimation. These fitted values were based on all variables used in the model, including the excludable instrument. And as all of these instruments are exogenous in the structural model, what this means is that the fitted values themselves have become exogenous too. Put differently, we are using only the variation in schooling that is exogenous.
- This exogenous variation in S driven by the instruments is only a subset of the total variation in the variable itself. Or put differently, IV reduces the variation in the data, so there is less information available for identification, and what little variation we have left comes from the complier population only. We are estimating the LATE, the causal effect for the complier population, not everyone affected by a change in D.
- Weak instruments - can add more instruments to the first stage regression to increase variation in the instrument and bring down the standard errors. The tradeoff is these instruments will have weaker correlation with the effect of interest, signal to noise drops. Adding more weak instruments causes the first stage F statistic to approach zero and increase the bias of 2SLS - it approaches the OLS bias in limit.

Heterogeneous Treatment Effects

- Explicitly based on the potential outcomes model, where each unit has a unique treatment effect. The causal effect itself is a random variable. Now we will allow for each unit to have a unique response to the treatment, or $Y_i^1 - Y_i^0 = \delta_i$
- Under homogenous treatment effects, there is no such tension between external and internal validity because everyone has the same treatment effect. But under heterogenous treatment effects, there is a

huge tension; the tension is so great, in fact, that it may even undermine an otherwise valid IV design.

- Y is a function of D and Z. Potential outcomes as we have been using the term refers to the Y variable, but now we have a new potential variable – potential treatment status. $D_i^1 = i's$ treatment status when $Z_i = 1$, etc. Treatment status switching equation: $D_i = D_i^0 + \left(D_i^1 - D_i^0\right) Z_i$

Identification Assumptions

- Five new assumptions needed for identification. View them through effect of military service on earnings using a draft lottery as the instrumental variable
    1. Stable unit value assumption (SUTVA): the potential outcomes for each person i are unrelated to the treatment status of other individuals. A violation of SUTVA would be if the status of a person at risk of being drafted was affected by the draft status of others at risk of being drafted.
    2. Independence assumption ("as good as random assignment"): the IV is independent of the potential outcomes and potential treatment assignments. Independence means that the first stage measures the causal effect of Z_i on D_i. For example, draft numbers assigned randomly through the population
    3. Exclusion restriction: that any effect of Z on Y must be via the effect of Z on D. An individual's earnings potential as a veteran or a non-veteran are assumed to be the same regardless of draft eligibility status - violated if low lottery numbers affected schooling by people avoiding the draft.
    4. First stage: Z is correlated with the endogenous variable such that $E\left[D_i^1 - E_i^0\right] \neq 0$ - Z has to have some statistically significant effect on the average probability of treatment. The first stage is testable as it is based solely on D and Z, both of which you have data on.
    5. Monotonicity: requires that the instrumental variable (weakly) operate in the same direction on all individual units. Draft eligibility may have no effect on the probability of military service for some people, like patriots, but when it does have an effect, it shifts them all into service, or out of service, but not both. Without monotonicity, IV estimators are not guaranteed to estimate a weighted average of the underlying causal effects of the affected group.
- What, then, is the IV strategy estimating under heterogenous treatment effects? Answer: the local average treatment effect (LATE) of D on Y. $\delta_{IV,LATE} = E\left[\left(Y_i^1 - Y_i^0\right) | D_i^1 - D_i^0 = 1\right]$ - the LATE parameters is the average causal effect of D on Y for those whose treatment status was changed by the instrument, Z. IV estimates the average effect of military service on earnings for the subpopulations who enrolled in military service because of the draft but who would not have served otherwise. It doesn't identify the causal effect on patriots who always serve or those who were exempted from service.

The LATE Framework

- The LATE framework partitions the population of units with an instrument into potentially four mutually exclusive groups. Seen through the eyes of attending private school

1. Compliers - those whose treatment status is affected by the instrument in the correct direction ($D_i^1 = 1$, $D_i^0 = 0$). Compliers go to the school if they win the lottery, and don't go to the school if they don't.
2. Defiers - those whose treatment status is affected by the instrument in the wrong direction ($D_i^1 = 0$, $D_i^0 = 1$). Defiers attend the school if they don't win, but don't attend the school if they do win.
3. Never takers - those that never take the treatment regardless of the value of the instrument ($D_i^1 = D_i^0 = 0$). They believe in public education, and so even if they win the lottery, they won't go to the private school.
4. Always takers - those that always take the treatment regardless of the value of the instrument (

$D_i^1 = D_i^0 = 1$). They always send their kids to private school, regardless of the number on their voucher lottery.

- With all five assumptions satisfied, IV estimates the average treatment effect for compliers. In homogeneous treatment effects, compliers have the same treatment effects as non-compliers. While we generally want to estimate the effects on a population, that is generally not possible with IV.
- You can immediately see why people find IV estimation less credible – not because it fails to identify a causal effect, but rather because it's harder and harder to imagine a pure instrument that satisfies all five conditions.

Panel Data

- Panel data estimators are designed explicitly for longitudinal data – the repeated observing of a unit over time. Repeatedly observing the same unit over time can overcome a particular kind of omitted variable bias, though not all kinds.
- Let's say that we have data on a column of outcomes, Y_i, which appear in three time periods (Y_{i1}, Y_{i2}, Y_{i3}). We have a matrix of covariates D_i that vary over time D_{i1}, D_{i2}, D_{i3}. There exists a single unit-specific unobserved variable, u_i, which varies across units, but which does not vary over time for that unit. i. Finally there exists some unit-specific observed time-invariant variable, X_i.
- D_{i1} causes its own outcome Y_{i1} is also correlated with the next period D_{i2}. Secondly, u_i is correlated with all the Y_{it} and D_{it} variables. There is no time-varying unobserved confounder correlated with D_{it} - the only confounder is u_i, which we call the **unobserved heterogeneity**. Past outcomes do not directly affect current outcomes (i.e., no direct edge between the Y_{it} variables). Past outcomes do not directly affect current treatments (i.e., no direct edge from $Y_{i,t-1}$ to D_{it}). Past treatments, $D_{i,t-1}$ do not directly affect current outcomes, Y_{it} (i.e., no direct edge from $D_{i,t-1}$ and Y_{it}).
- Often our outcome variable depends on several factors, some of which are observed and some of which are unobserved in our data, and insofar as the unobserved variables are correlated with the treatment variable, then the treatment variable is endogenous and correlations are not estimates of a causal effect. But if these omitted variables are constant over time, then even if they are heterogeneous across units, we can use panel data estimators to consistently estimate the effect of our treatment variable on outcomes.
- We are interested in the partial effects of variable D_j in the population regression function $E[Y|D_1, D_2, \ldots, D_k, u]$. We typically assume that the actual cross-sectional units (e.g., individuals in a panel) are identical and independent draws from the population. Our model for a randomly drawn cross sectional unit i is $Y_{it} = \delta D_{it} + u_i + \varepsilon_{it}, t = 1, 2, \ldots, T$

Pooled OLS

- Ignore the panel structure and regress $Y_{it} = \delta D_{it} + \eta_{it}; t = 1, 2, \ldots, T$. We need assumption $E[\eta_{it}|D_{i1}, D_{i2}, \ldots, D_{iT}] = E[\eta_{it}|D_{it}] = 0 \text{ for } t = 1, 2, \ldots, T$
- No correlation between Dit and hit necessarily means no correlation between the unobserved ui and Dit for all t and that is just probably not a credible assumption.
- An additional problem is that η_{it} is serially correlated for unit i since u_i is present in each t period. And thus pooled OLS standard errors are also invalid.

Fixed Effects

- If we have data on multiple time periods, we can think of ui as fixed effects to be estimated: $\left(\hat{\delta}, \widehat{u}_1, \ldots, \widehat{u}_N\right) = \underset{b, m_1, \ldots, m_N}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{1=1}^{T} \left(Y_{it} - D_{it} b - m_i\right)^2$ - this amounts to including N individual dummies
- Running a regression with the time-demeaned variables $\ddot{Y}_{it} \equiv Y_{it} - \bar{Y}_i$ and $\ddot{D}_{it} \equiv D_{it} - \bar{D}$ is numerically equivalent to a regression of Y_{it} on D_{it}. Time-demeaning eliminates the unobserved effects, so δ now consistent
- Implementation: can choose from
  - Demean and regress $\ddot{Y}_{it}$ on $\ddot{D}_{it}$
  - Regress Y_{it} on D_{it} and unit dummies
  - Regress Y_{it} on D_{it} using a fixed effect package
- Identifying assumptions
  - $E\left[\varepsilon_{it} | D_{i1}, D_{i2}, \ldots, D_{iT}, u_i\right] = 0; t = 1, 2, \ldots, T$ - regressors are strictly exogenous conditional on the unobserved effect.
  - $\operatorname{rank}\left(\sum_{t=1}^{T} E\left[\ddot{D}_{it}' \ddot{D}_{it}\right]\right) = K$ - regressors must vary over time for at least some i and not be collinear
- Fixed effects cannot address reverse causality or time-variant unobserved heterogeneity (demeaning will not change anything).
- It's thus the burden of the researcher to determine which type of unobserved heterogeneity problem they face.

Differences-in-Differences

- DD is basically a version of panel fixed effects, but can also be used with repeated cross-sections.
- The first difference, D1, does the simple before and after difference. This ultimately eliminates the unit specific fixed effects. Then, once those differences are made, we difference the differences (hence the name) to get the unbiased estimate of E.
- Relies on a parallel trends assumption - T, time effect, is the same for all units. Starting from initial positions, the slope for the two units change would be equal in time except for the treatment intervention that causes their paths to diverge.
- PA / NJ minimum wage change difference: Let Y_{ist}^1 be employment at restaurant i, in state s, at time t with a high minimum wage, and let Y_{ist}^0 be employment at restaurant i, state s, time t with a low minimum wage. Assume $E\left[Y_{\text{ist}}^0 | s, t\right] = \gamma_s + \tau_t$ - in the absence of a minimum wage change, employment in a state will be determined by the sum of a time-invariant state fixed effect, γ_s, that is idiosyncratic to the state, and a time effect τ_t that is common across all states.
  - ATE given by $E\left[Y_{ist}^1 - Y_{ist}^0 | s, t\right] = \delta$ and observed employment by $Y_{ist} = \gamma_s + \tau_t + \delta D_{st} + \varepsilon_{ist}$
  - To calculate the treatment effect, compute before and after differences for each state, and then difference those differences.
- If we want to control for other variables, can also do this through a regression framework, with something like $Y_{it} = \alpha + \beta_1 D_i + \beta_2 Post_t + \delta(D \times Post)_{it} + \tau_t + \sigma_s + \varepsilon_{st}$, D is a dummy whether the unit is in the treatment group or not, Post is a post-treatment dummy, and the interaction is the DD coefficient of interest.

- Parallel trends assumption tested by empiricists by looking at the trends before the treatment period - if parallel before, wouldn't they have continued on this path? Checking the parallelism of the pre-treatment trends is not equivalent to proving that the post-treatment trends would've evolved the same - it should just increase our confidence that our assumption might be valid. This is done by including lead terms in the regression.

Inference

- Often use data with longer time periods and the outcome variables are serially correlated. Conventional standard errors often severely understate the standard deviation of the estimators
- Potential solutions: Block bootstrapping standard errors, say sampling states with replacement for bootstrapping. Clustering standard errors at the group level. Aggregating the data into one pre and one post period - must have a single treatment date.

Threats to Validity

- Non-parallel trends
    - Often treatments are targeted towards the worst off - mean reversion may cause issues with parallel trends
    - Selection bias is a big problem with many types of treatments and policy interventions.
    - Correction - robustness checks are common, forms of placebo analysis. Looking at the leads, use a falsification test with an alternative control group, or falsification with outcomes that shouldn't be affected by the treatment.
    - Alternative control group - DDD differences in differences in differences. The logic of the DDD strategy is to use a within-city comparison group that experiences the same city-specific trends, as well as its own crime-specific trend, and use these within-city controls to net them out. This comes at a cost - more parallel trend assumptions and additivity assumptions
- Compositional differences - DD can be applied to repeated cross-sections, as well as panel data. But one of the risks of working with the repeated cross-section is that unlike panel data (e.g., individual level panel data), repeated crosssections run the risk of compositional changes.
- Long term effects vs reliability
- Functional form dependendence

Synthetic Control

- Synthetic controls models optimally choose a set of weights which when applied to a group of corresponding units produce an optimally estimated counterfactual to the unit that received the treatment.
- This counterfactual, called the "synthetic unit", serves to outline what would have happened to the aggregate treated unit had the treatment never occurred. Generalization of DD
- When the units of analysis are a few aggregate units, a combination of comparison units (the "synthetic control") often does a better job of reproducing characteristics of a treated unit than using a single comparison unit alone. The comparison unit, therefore, in this method is selected to be the weighted average of all comparison units that best resemble the characteristics of the treated unit(s) in the pre-treatment period.

- Advantages over regression based methods: control comparison within the same range as treatment data, precluding extrapolation beyond data support. Construction of the counterfactual does not require access to the post-treatment outcomes during the design phase of the study - data snooping not needed. The weights which are chosen make explicit what each unit is contributing the counterfactual, whereas regression weights are implicit. Bridges a gap between qualitative and quantitative types - choosing the counterfactuals gives power to those with detailed knowledge.
- Let Y_{jt} be the outcome of interest for unit j of J + 1 aggregate units at time t and treatment group j = 1. The synthetic control estimator models the effect of the intervention at time T_0 on the treatment group using a linear combination of optimally chosen units as a synthetic control. For the post-intervention period, the synthetic control estimator measures the causal effect as $Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$ for optimal weight vector w_j^*
- Take matching variables X_1, X_0 that are chosen as predictors of post-intervention outcomes and are unaffected by the intervention. Weights are minimizer of the norm $||X_1 - X_0 W||$ subject to $w_j \geq 0,\ j \in [2, J+1]$ and $\sum_j w_j = 1$
- Since $\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0) W}$ we are left to choose V, some symmetric, semi-definite matrix. Most people choose V that minimizes the mean squared prediction error: $\sum_{i=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^{I+1} w_j^*(V) Y_{jt} \right)^2$
- How do we determine whether the observed difference between the two series is a statistically significant difference?
    - Iteratively apply the synthetic control method to each country/state in the donor pool and obtain a distribution of placebo effects
    - Calculate the root mean squared prediction error RMSPE for each placebo for the pre-treatment period
    - Calculate the RMSPE for each placebo for the post-treatment period
    - Compute the ratio of the post-to-pre-treatment RMSPE. Sort ratio in descending order.
    - The treatment unit ratio in the distribution is p = rank / total
    - Create a histogram of the ratios, and more or less mark the treatment group in the distribution so that the reader can see the exact p-value associated with the model.
- Placebo robustness - rewind time from the treatment date and estimate the model on a placebo intervention date - should find no effect.