Causal Inference in Experiments and Observational Studies

Han Zhang

Mar 16, 2021

Outline

Logistics

Beyond Binary Treatment

Stratified Random Experiment

Heterogeneous treatment effect

Observational Studies

Ignorability and Matching

Logistics

- Exercise 2 due in one week (Mar 23); start early!
- Preliminary draft in one week
    - Double spaced, 4-6 pages. Describe the background, hypotheses, data, and methods you plan to use
    - I will give you feedback

Readings

Today's topics are drawn from:

- Joshua D. Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricists Companion* . Princeton University Press, 2009. (Chapters 2 - 3)
  - MHE later in this class
- Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics* . Cambridge University Press, 2019. (Chapters 6 - 7)
- Proofs:
  - Imbens, Guido W., and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015 (Chapter 6 - 7).

Categorical Treatment

- When there are categorical treatment D with more than 2 levels, simple
- Each group has its own treatment effect (defined with respect to the control group)
- You can still use Neyman estimator or linear regression estimator (treatments as dummies, and control group as the reference group)

Continuous Treatment

- When there are ordinal or continuous treatment D:
  - E.g., effect of years of schooling on future income
- One way to extend the Neyman-Rubin Causal Model

$$\begin{aligned} Y_i &= \begin{cases} Y_i^0 : D_i = 0 \\ Y_i^d : D_i = d \end{cases} \\ &= Y_i^0 + d(Y_i^d - Y_i^0) \end{aligned} \tag{1}$$

- This extension assumes that causal effects is linear in D at the unit level
- Assume we manipulate the treatment from d to d'.

$$ATE = E\left(\frac{Y_i^{d'} - Y_i^d}{d' - d}\right) \tag{2}$$

- If $d' - d' = 1$ (changes for one unit), this becomes to binary treatment

Continuous Treatment with Regression

- You can use regression to estimate ATE in continuous treatments
- Assume D is a continuous treatment, then
  - $Y_i = \alpha + \rho D_i + \epsilon_i$ (for random experiments)
  - $Y_{ij} = \alpha_j + \rho D_{ij} + \epsilon_{ij}$ (for cluster-randomized experiments)
- ρ will identify ATE, assuming causal effect is linear in treatment

Stratified Randomized Experiment

- Sometimes the treatment assignment are not completely random
- Whether small class size improves students' test scores?
- Tennessee Project STAR experiment: whether class sizes impact test scores
  - Within each school, we random select some classes to be in the treatment group (small class size), and other classes in regular class group (control group)
- This is known as cluster randomized, or stratified randomized experiments

Estimate ATE under Stratified Randomized Experiment

- Neyman estimator:
  - Estimate ATE within each group (unique combinations of X)
  - ATE is the weighted sum of group-specific ATE, with weights proportional to group size

$$A\hat{T}E_{neyman} = \sum_{j=1}^{J} \frac{\omega(j)}{\left(\sum_{j=1}^{J} \omega(j)\right)} \cdot A\hat{T}E_j \tag{3}$$

- Weights for each group is: $\omega(j) = \frac{N_j}{N}$
- S.E. of ATE is weighted sum of group-specific S.E., with weights proportional to $\omega(j)^2$
- This is equivalent to post-stratification estimator in missing data

Star Example: Neyman estimator

School	# Classes	Estimated Effect	$\widehat{\text{(s.e.)}}$
1	4	0.223	(0.230)
2	4	−0.295	(0.776)
3	5	0.417	(0.404)
4	4	0.748	(0.215)
5	4	−0.077	(0.206)
6	4	1.655	(0.405)
7	4	−0.254	(0.255)
8	6	0.429	(0.306)
9	4	−0.006	(0.311)
10	4	−0.014	(0.182)
11	4	−0.003	(0.605)
12	5	0.222	(0.309)
13	4	0.432	(0.179)
14	4	0.340	(0.336)
15	4	0.207	(0.396)
16	4	−0.306	(0.245)
overall		0.241	(0.092)

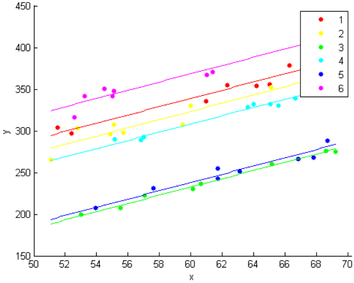
(4)

- Each school has its own ATE_j: average score difference between small vs large class within that school
- ATE is the weighted sum of school-specific ATE_j, with weights proportional to relative group size (N_j/N)
- S.E. of ATE is weighted sum of school-specific S.E., with weights proportional to $(N_j/N)^2$

Regression Estimator of ATE

- Regression estimator for overall ATE (still assume constant treatment effect)
    - Regression Y on D, with group fixed effects, that is, a separate intercept for each group j (more on fixed effects later)
    - $Y_{ij} = \alpha_j + \rho D_{ij} + \epsilon_{ij}$
- The regression coefficient for D is the estimate of ATE

Fixed effect



- Each group has its own intercepts
- But slope is the same
- More on this in next two weeks
- in R, use `plm` or `fixest` package

Use fixed effect regressions to estimate ATE

$$Y_{ij} = \alpha_j + \rho D_i + \epsilon_{ij} \tag{5}$$

- i is unit and j is group
- α_J are group-level fixed effects
- Still, constant treatment effect assumption; ρ is the overall ATE
- Estimated ATE is $\hat{\rho} = 0.238$, and its S.E. (0.103)
  - Compare this with ATE estimated using Neyman estimator
- In general, ATE estimates using Neyman mean estimator and regression estimator are different

Neyman vs Regression

- In fact, we can express ATE_{ols} as the following weighted mean:

$$\hat{ATE}_{OLS} = \hat{\rho} = \sum_{j=1}^{J} \frac{\omega(j)}{\left(\sum_{j=1}^{J} \omega(j)\right)} \cdot \hat{ATE}_j \tag{6}$$

- Weights for each group is proportional to:

$$\omega(j) = \frac{N_j}{N} \cdot P(D = 1|X = j) \cdot (1 - P(D = 1|X = j))$$

- N_j/N is relative size
- $P(D = 1|X = j)$ is the proportion of treated units in group j, or put it differently, group-specific treatment propensity

Neyman vs Regression

- $A\hat{T}E_{OLS}$ gives additional weights $P(D=1|X=j)\cdot(1-P(D=1|X=j))$
- Regression estimator puts most emphasis on group whose $P(D=1|X=j)=1/2$, that is, group with the same number of treated and control units (Read MHE 3.3 for details)
- In general: $A\hat{T}E_{OLS}$ is neither consistent nor unbiased
  - Note that we have not add covariates yet; adding covariates can make it worse
- $A\hat{T}E_{neyman}$ is consistent and unbiased
  - But harder to add covariates

	regression	Neym
unbiasedness and consistency	inferior	superi
standard error	smaller (with covariate balance)	larger
practice (with ctrl)	easier	harde

Heterogeneous treatment effect with subgroups

- Broadly speaking, heterogeneous treatment effect (HTE) just means that treatment effect varies
- For cluster-randomized experiment, Neyman estimator naturally gives heterogeneous treatment effect for each subgroup (e.g., each school)
- Regression estimator:
  - Fixed effect regression gives an overall ATE, but does not estimate heterogeneous treatment effect for each group
  - To estimate group-specific ATE, we fit linear regression with interactions between group dummy and treatment D (and no fixed effects)
  - (Interaction coefficients + the coefficient on treatment) captures the treatment effect for each subgroup
    - But this interaction model does not give us overall ATE

Heterogeneous treatment effect by covariates

- A more common case: treatment heterogeneity by covariates
- Formal notation: $\tau(x) = E(Y^1 - Y^0|X = x)$
- $\tau(x)$ is usually referred as conditional average treatment effect (CATE)
- In Project STAR example:
  - We think the important variation of treatment heterogeneity come from whether teacher is more experienced or not

Using interaction model to capture CATE

- To capture heterogeneity by covariate (CATE), the easiest approach is to add interaction between treatment and covariate:

$$Y_{ij} = \alpha_j + D_i\rho + \text{experienced}_i\beta + D_i\text{experienced}_i\gamma + \epsilon_{ij}$$

- γ and ρ together captures the treatment effect heterogeneity
  - ρ: *CATE* for classes with inexperienced teacher
  - $\gamma + \rho$: *CATE* for treated classes with experienced teacher

Using interaction model to capture CATE: shortcomings

- Using interaction model to capture CATE is the dominant approach in applied literature now, but it has many problems
- The interaction model assumes constant effect within covariate levels
  - It slightly relax the overall constant effect assumption, but still can be unrealistic
  - e.g., within class taught by experienced teacher, gender ratio still matters.
- If you did not add interaction for gender ratio, you are implicitly assuming that treatment is constant across classes with different gender ratio, which could be problematic
- [Question]: what regression model you should use if you believe that teacher experience and gender ratio all matters?
- [Question 2]: but we do not want to allow all possible interactions between covariates; why?

Using interaction model to capture CATE

- If the covariate has more levels, or is continuous variable, we are implicitly adding another assumption: linear interaction effect
- That is, CATE grows linearly in the covariate
  - (recall the Hainmuller et. al, article presented by Zhang Yi?)
- Is this realistic? certainly not.

Machine Learning for HTE

- Motivation
  - Add many interactions and then regularize, instead of assume only one or two covariates matter
  - Avoid strong linear interaction effect assumption
- Prediction vs causality:
  - prediction: use X to predict Y
  - causation: use X to predict Y^1 or Y^0 (depend on value of D)
  - One popular approach (for experiments)
    - Estimate $E(Y|D, X)$ for $D = 1$ and 0 (treatment and control) and X using some machine learning methods
    - then take the average of individual treatment effect for some subgroups

Example 1: using LASSO to select interactions

- Imai and Ratkovic. 2013. Ann. Appl. Stat.
- Add full interactions between treatment and covariates
- Combines SVM and LASSO
  - SVM: a commonly used machine learning model for making binary predictions
  - LASSO: regularize

Example 1: using LASSO to select interactions

- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others
  - Outcome: earning after 1/2 years
  - the (randomly selected) treatment and control groups consist of 297 and 425 such workers, respectively.
- Add interaction between treatment and
  - 7 pre-treatment variable, plus age and years of education (squared); total of 9 covariates
  - interactions between each of the above variable
- This lead to $9+ 9*8/2 = 45$ treatment-covariate interactions
- Then use LASSO to select treatment-covariate interactions that yield highest and lowest treatment effects

Example 1: using LASSO to select interactions

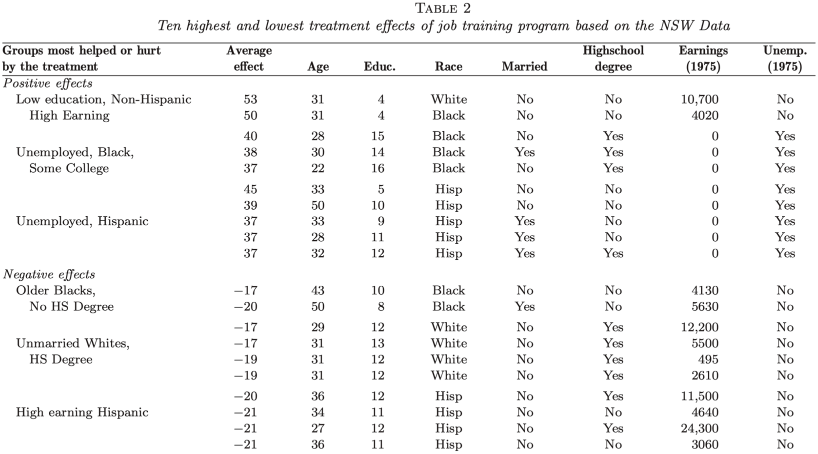


TABLE 2

Ten highest and lowest treatment effects of job training program based on the NSW Data

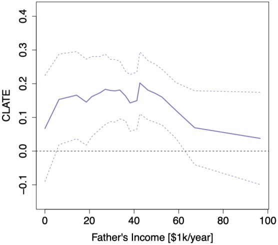
Groups most helped or hurt by the treatment	**Average effect**	**Age**	**Educ.**	**Race**	**Married**	**Highschool degree**	**Earnings (1975)**	**Unemp. (1975)**
Positive effects								
Low education, Non-Hispanic	53	31	4	White	No	No	10,700	No
High Earning	50	31	4	Black	No	No	4020	No
	40	28	15	Black	No	Yes	0	Yes
Unemployed, Black,	38	30	14	Black	Yes	Yes	0	Yes
Some College	37	22	16	Black	No	Yes	0	Yes
	45	33	5	Hisp	No	No	0	Yes
	39	50	10	Hisp	No	No	0	Yes
Unemployed, Hispanic	37	33	9	Hisp	Yes	No	0	Yes
	37	28	11	Hisp	Yes	No	0	Yes
	37	32	12	Hisp	Yes	Yes	0	Yes
Negative effects								
Older Blacks,	−17	43	10	Black	No	No	4130	No
No HS Degree	−20	50	8	Black	Yes	No	5630	No
	−17	29	12	White	No	Yes	12,200	No
Unmarried Whites,	−17	31	13	White	No	Yes	5500	No
HS Degree	−19	31	12	White	No	Yes	495	No
	−19	31	12	White	No	Yes	2610	No
	−20	36	12	Hisp	No	Yes	11,500	No
High earning Hispanic	−21	34	11	Hisp	No	No	4640	No
	−21	27	12	Hisp	No	Yes	24,300	No
	−21	36	11	Hisp	No	No	3060	No

Example 2: Causal trees (advanced topics)

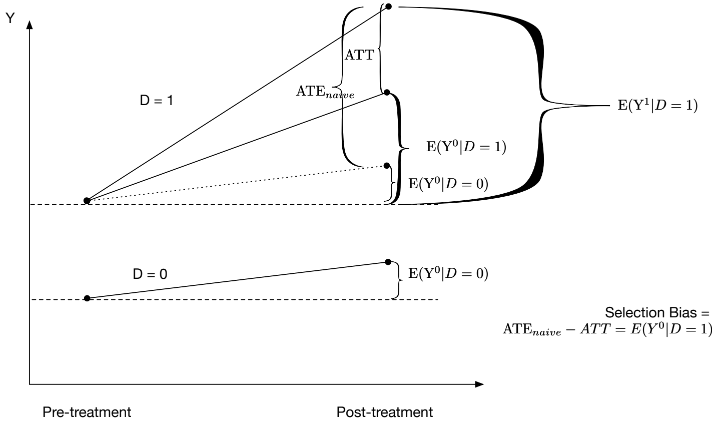
- Causal forests: a series of articles such as Athey and Imbens, 2016, PNAS; Wager and Athey, 2018, JASA; Athey, Tibshirani, and Wager, 2019, Ann. Stat.
- Basic idea: use random forests to predict counterfactual
  - tree models interactions even more flexibly
- How causal forests differ from random forests:
  1. each time, RF finds a split to best predict Y; CF finds a split to maximize HTE
     - Then the estimated causal effect for each leaf is the differences in means of treated and control units in that leaf
  2. honest split: use half data to grow a tree; and use the learned tree to calculate treatment effect
     - avoid over-fitting
- Benefits: have an estimation of individual level treatment effect $\tau_i = Y_i^1 - Y_i^0$
- Then you can easily calculate any HTE you want

Example 2: Causal trees (advanced topics)

- Athey, Tibshirani, and Wager, 2019, Ann. Stat.
- outcome is whether the mother did not work in the year preceding the census
- treatment is whether the mother had 3 or more children at census time
- Use causal forest to predict individual treatment effect τ_i; then plot τ_i against father's income
- Found an inverse-U shape HTE; cannot obtain this using interaction model



Observational Studies


Y
D = 1
ATT
ATE_{naive}
$\text{E}(\text{Y}^1|D=1)$
$\text{E}(\text{Y}^0|D=1)$
$\text{E}(\text{Y}^0|D=0)$
D = 0
$\text{E}(\text{Y}^0|D=0)$
Pre-treatment
Post-treatment

Selection Bias =

$$\text{ATE}_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

Observational Studies

- $ATE_{naive} = E(Y^1|D=1) - E(Y^0|D=0)$, which is the mean differences in “treatment” and “control” outcomes.
- In observational studies (without random assignment) ATE_{naive} neither estimates ATE nor ATT
- ATE_{naive} differs from ATT by selection bias

$$ATE_{naive} - ATT = E(Y^0|D=1) - E(Y^0|D=0)$$

- Selection bias becomes 0 and ATT is identified, if Y^0 is independent of D
  - in other words, if the treated units were not treated, their counterfactual outcome would be the same as that of the untreated users
  - e.g., college-educated would have the same earning as non college-educated, if college-educated did not go to college
- If we further assume Y^1 is independent of D
  - $ATT = ATE$

Design-based vs Model-based causal inference

- Design-based causal inference:
  - If you can, think and perform real randomized experiment
  - If you cannot, try to approximate an experiment by adding assumptions
    - A better study design uses assumption that makes your study more like an experiment
  - examples: natural experiment, matching, DID, modern IV, RD
  - Rule of thumb: the gold-standard is always randomized controlled experiment
- Model-based causal inference:
  - start from a regression and gradually add assumptions to the regression model (e.g., assumptions about endogenous or exogenous regressor)
  - example: traditional IV, fixed effects
- Historically people are more familiar with model-based causal inference
- The trend is leaning toward design-based causal inference

Natural Experiments: classical type of design-based causal inference

- Natural experiments seeks to find exogenous variations in the explanatory variable that is as if random
- Does political leaders (Presidents, Prime Ministers, etc) matter for regime type?
- The Neyman-Rubin model suggests that we need to manipulate the treatment variable, and compare counterfactual outcomes
- A better way to ask the question: do leadership changes lead to changes in regime type?
  - Problem: we do not know all factors that determine leadership changes
    - A non-exhaustive lists include economic growth itself, leadership personality, geospatial conditions, etc.

Natural Experiment Example

- Jones and Olken (2009): "Hit or Miss? The Effect of Assassinations on Institutions and War", AEJ.
- Assassination attempts provide exogenous variations in changing the leader:
    - failed assassinations are a control group for successful assassinations
    - e.g., compare the assassination of JFK to the assassination attempt on Ronald Reagan. Bullet killed JFK but missed Reagans heart by inches.
- Outcome is the change of regime type after the leadership transition
    - binary (democratic vs autocratic)
    - POLITY Score (democracy scores)

Estimating causal effects in natural experiment

- Because of the as-if-random assumption, observed and unobserved conditions, which may be related to treatment assignment, are randomized by treatment and control groups
- We can use Neyman estimator (compare changes in regime type in successful assassinations vs unsuccessful assassinations)
- We can also use regression estimator: $\hat{\rho}$ estimates ATE

$$Y = \alpha + \rho D + \epsilon$$

Always check pre-treatment balances

- The key difference between randomized experiments and natural experiments is that the former is controlled by researchers, while the latter is an assumption (that the exogenous shocks are really random)
- Because we assume the natural experiments are as-if-random, it is important to check pre-treatment balance to see whether treatment/control group are actually balanced
- If the assumptions is true, we would expect that other covariates are indeed balanced across treatment and control groups
  - That is, the other variables should be independent of D, if the assumption is correct
- But the reverse is not true: covariates are balanced $\not\Rightarrow$ exogenous shocks are truly random
- There is no substitute for a good research design (here, exogenous shocks)

Natural experiments with covariates

- Because natural experiments are not fully controlled by researchers, covariates can have additional help (other than checking pre-treatment balance)
- Say we think that within assassination attempts, those using guns are more likely to succeed than those using bombs
  - Assassinators also know this! So their decisions are conditioned on weapons
- We should add weapon types as additional controls

$$Y = \alpha + \rho D + \gamma \text{weapon} + \epsilon$$

- In analyzing natural experiments, it is recommended to use all other observed control variables you think are relevant to your outcome Y (no post-treatment controls, of course)
  - Another differences from the randomized experiments

Many different ways to define control groups

- Another thing worth noting of natural experiments is that there are often multiple ways to choose control groups; be careful about what you are actually comparing with
- Say, we compare successful assassinatons with failed assassinatons
- But failed assassinatons can be defined in multiple ways
    - Any assassinatons (including those still in secret planning?)
    - Any failed assassinatons planning that made to newspapers
    - Assassinations whose weapons were already discharged (serious attempts)
- Be clear what you are comparing with
- In randomized controlled experiments control groups were chosen with clear standard so there is no such problem

Selection on Observables/ Ignorability

- To identify causal effects without random assignment, we have to add strong assumptions (analogous to MCAR)

Definition (Selection on observable, or ignorability, or exogeneity)

- $Y_i^0, Y_i^1 \perp\!\!\!\perp D_i | X_i$ (Potential outcome is independent of treatment assignment, condition on observed X_i)
- $P(D = 1) > 0$ (non-zero treatment probability)

- Note that randomized experiments automatically satisfy this assumption

Regression estimator

- If ignorability assumption is true, and you assume the effect of treatment is constant on Y, we can use regression to estimate causal effects:

$$Y = \alpha + \rho D + \gamma X + \eta$$

- Estimate of ATE is the regression coefficient ρ
- Here we want as many X as possible
  - We are making assumptions that potential outcome is independent of D, conditional on X
  - More X increases the possibility that you do not missed anything important confounders

Regression as Imputation

- If ignorability is true, the regression estimator of ATE is *implicitly* making counterfactual imputation using linear regression.
- Hence the similar form of ignorability and MACR assumption

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

- Run a regression as $Y = \beta_0 + \beta_1 D_i + \beta_2 X_{[1]i} + \beta_3 X_{[2]i}$, and impute counterfactual outcome using the linear regression:

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 7$	2	1	1	7
2	5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 8 + \hat{\beta}_3 \cdot 2$	0	8	2
3	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 9 + \hat{\beta}_3 \cdot 3$	3	1	9	3
4	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 3 + \hat{\beta}_3 \cdot 1$	10	1	3	1
5	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 5 + \hat{\beta}_3 \cdot 2$	2	1	5	2
6	0	$\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 \cdot 7 + \hat{\beta}_3 \cdot 0$	0	7	0

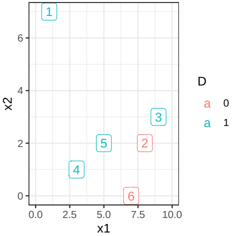
Matching estimator using original data

- With ignorability, we can also use matching estimator
  - it is very similar to the hot-deck imputation
- For unit i in the treatment ($D_i = 1$), we want to impute its Y_i^0
  - Find the j in the control group, whose X_j is the closest X_i
    - j is called the matched unit of i
  - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- ATT is estimated as the the difference between the mean of Y_i^1 and Y_i^0 for treated units
- We can do the similar things for control units:
  - For unit i in the control ($D_i = 1$), we want to impute its Y_i^1
    - Find the j in the control group, whose X_j is the closest X_i
    - Use the Y_j^0 associated with j as the imputed Y_i^0 value for i
- Then we can estimate ATE as the the difference between the mean of Y_i^1 and Y_i^0 for all units

Matching estimator using original data

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$
1	?	2	1	1	7
2	5	?	0	8	2
3	?	3	1	9	3
4	?	10	1	3	1
5	?	2	1	5	2
6	0	?	0	7	0

(8)



- Unit 3 is treated; it is closest to unit 2; unit 2 is the matched unit of unit 3
- $Y_3^0 \leftarrow Y_2^0 = 5$

Matching estimator using Propensity Score

- When we have multiple X, it will become hard to calculate distances between X.
  - The curse of dimensionality again
- Similar to missing data case, we have treatment propensity score (Rosenbaum and Rubin, 1983)

$$P(D = 1|X) \tag{9}$$

- Rosenbaum and Rubin proves (propensity score theorem)
  - If you have the correct propensity score
  - Then conditioning on X is equivalent to conditioning on $P(D = 1|X)$
- Treatment propensity score provides a single-number summary of treatment probability
- We should match treatment and control users with similar treatment propensity scores
  - This is usually called propensity score matching

Matching estimator using Propensity Score

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$	$p\left(D_i = 1 \vert X_i\right)$
1	?	2	1	1	7	0.33
2	5	?	0	8	2	0.14
3	?	3	1	10	3	0.73
4	?	10	1	3	1	0.35
5	?	2	1	5	2	0.78
6	0	?	0	7	0	0.70

(10)

Unit	Y_i^0	Y_i^1	D_i	$X_{[1]i}$	$X_{[2]i}$	$p\left(D_i = 1 \vert X_i\right)$
1	5	2	1	1	7	0.33
2	5	2	0	8	2	0.14
3	0	3	1	10	3	0.73
4	5	10	1	3	1	0.35
5	0	2	1	5	2	0.78
6	0	3	0	7	0	0.70

(11)

- Estimated *ATE* is $7/6$

Regression vs Matching

- Regression estimates of ATE in general will be different from matching estimates of ATE (MHE 3.3)

$$\hat{ATE}_{ols} = \hat{\rho} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x) \cdot P(D = 1|X = x) \cdot (1 - P(D = 1|X = x))$$

$$\hat{ATE}_{matching} = \sum_x \frac{\omega(x)}{(\sum_x \omega(j))} \cdot \hat{ATE}_x$$

$$\omega(x) = P(X = x)$$

- Regression estimator is generally inconsistent, while matching estimator approximates Neyman estimator and is consistent
- Regression give more weights to data whose propensity score is close to 0.5
  - These are observations whose treatment status cannot be predicted well by X, thus could have omitted variable bias

Regression vs Matching

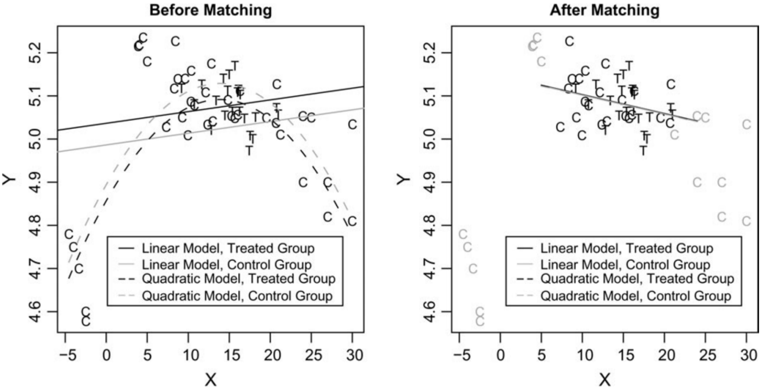
- Regression and matching estimates of causal effect share the same assumption: ignorability
- But they yield different estimates
- Regression
  - Pros: easier to work with (especially the standard error)
  - Cons: inconsistent; more weights to data whose treatment status we cannot predict
- Matching
  - Pros: consistent; explicitly approaches experimental ideal by predicting counterfactuals
  - Cons:
    - Do you have the correct propensity score? (if you use propensity score matching)
    - confidence intervals are hard to calculate analytically
    - The first theoretical work is by Abadie and Imbens, 2006, Econometrica.

Matching + Regression

- Matching and Regression are not mutually exclusive
- In practice, it is common to first perform matching, and then run regression on matched units
    - Basic idea: assume you have 100 treated units and 400 control units
    - Instead of running a regression with all of them
    - Find the 100 control units that are closet to 100 treated units
    - And then run regressions based on 200 units
    - This gives you the ATT estimates
- Ho, King, Imai and Stuart, "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference", *Political Analysis*, 2007
    - Matching reduces model dependency

Matching + Regression

- Ho, King, Imai and Stuart, “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference”, *Political Analysis*, 2007



Regression vs Matching

- In random experiments, covariates should be balanced, so Neyman and Regression in practice does not differ too much
- But in observational studies, matching and regression estimator can give very different estimates
- Dehejia and Wahba, 2002
- NSW: a random experiment (National Support Work) that randomly provide 9-12 months job training to some participants but not others
  - Outcome: earning after 1/2 years
  - the (randomly selected) treatment and control groups consist of 297 and 425 such workers, respectively.
- CPS: survey data that is much larger in size, but with similar variables
  - clearly no one in CPS received the work training so they served as a non-experimental control group
  - That is, construct a data as (NSW(treated), CPS); CPS replaced NSW control units

Regression vs Matching

- What the authors did:
  - Experiments: raw ATE vs. regression estimated ATE based on [NSW(treated), NSW(control)]
  - CPS: ATE_{naive} vs. regression adjusted ATE_{naive} based on [NSW(treated), CPS)]
  - Matching: match each unit in NSW treatment group with a control in CPS, then calculate ATE_{naive} and regress adjusted ATE_{naive} based on [NSW(treated), CPS(matched)]

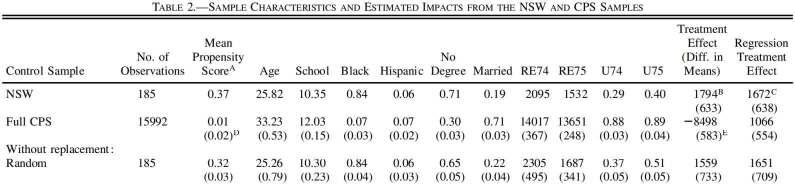


TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B] (633)	1672[C] (638)
Full CPS	15992	0.01 (0.02)[D]	33.23 (0.53)	12.03 (0.15)	0.07 (0.03)	0.07 (0.02)	0.30 (0.03)	0.71 (0.03)	14017 (367)	13651 (248)	0.88 (0.03)	0.89 (0.04)	−8498 (583)[E]	1066 (554)
Without replacement:														
Random	185	0.32 (0.03)	25.26 (0.79)	10.30 (0.23)	0.84 (0.04)	0.06 (0.03)	0.65 (0.05)	0.22 (0.04)	2305 (495)	1687 (341)	0.37 (0.05)	0.51 (0.05)	1559 (733)	1651 (709)

Do not condition on post-treatment variables

- We do not need to condition on all X we have in our data
- Some X can do more harm than good
- In particular, never condition on post-treatment X (Judea Pearl)
  - This is like cheating: if X is determined by your D, X would not be randomized even under random assignment of D, which recreates selection biases
- Example:
  - Effect of college education on future earning
  - And it will be dangerous to add *occupation* as a control variable, since occupation may be the result of treatment

Ignorability assumption in real world

- In practice, it is often unrealistic to make ignorability assumption
- Such that if you assume ignorability and claim you find causal effect, most people won't accept your argument
- But in rare cases, ignorability assumption can be convincing
- Hainmueller, Jens, and Dominik Hangartner. "Who gets a Swiss passport? A natural experiment in immigrant discrimination." APSR (2013): 159-187.
- Challenge in study determinants of discrimination:
    1. they won't tell you due to social pressure
    2. ignorability? in most survey data it cannot be true

Ignorability in real world

- Some Municipality in Switzerland ask every citizen to vote on whether giving immigrants citizenship or not
- Unless one knows the immigrant in person, he is most likely to make a decision based on “voting leaets summarizing the applicant characteristics were sent to all citizens usually about two to six weeks before”
- The voting is anonymous; no social pressure
- Therefore it is pretty convincing to assume ignorability in this article
- And a simple regression allows them to estimate causal effect

Selection on Unobservables

- Selection on observable/ignorability assumptions are often very strong
  - Basically, it says that you have to know all X that contribute to treatment assignment (D)
  - And, you have to have good measure for all of them
- There may always be unknown/unmeasured factors that contribute to treatment assignments
- That is, we are selecting on unobservables, or we have non-ignorability, or we have omitted variable bias
  - That is, we have selection bias due to unobservables
- General guideline when dealing with non-ignorability
  - Think about experiment ideal
  - Find a design that make your data your data like randomly assigned (as-if random)

Ignorability vs. Non-ignorability

- Randomized experiments:
  - Automatically satisfies ignorability by randomization
- Ignorability:
  - Very strong assumption; unrealistic in most settings
  - In this case, people just control a bunch of things but do not claim that they find any causal effect
  - Or avoiding using causal language; just say X predicts Y of X is associated of Y
- The third approach:
  - Do not assume ignorability (or assume non-ignorability)
  - Essentially admit that we cannot control everything; there are some unobserved variables we cannot control for

Econometric tools in working with non-ignorability

- Fixed effect and diff-in-diff
- IV
- RD