

ECMA 31100: Intro to Empirical Analysis II

Intro; Selection on Observables

Joe Hardwick

University of Chicago

Winter 2022

Causality

- So far we have considered counterfactuals using linear regression:
 - How much does earnings potential increase with another year of experience, holding all else fixed?
 - How much does an additional bedroom increase house price, holding floor area/lot size/air pollution (and all else) fixed?
- OLS estimates proportional to “partial correlations” between outcome and regressors.
 - ‘Holding fixed’ other observable factors in this way answers counterfactual question if $y = x'\beta + u$ and $E(xu) = 0$.
 - $E(xu) = 0 \implies \beta = E(xx')^{-1}E(xu)$: (Scaled) partial correlations are partial (causal) effects.
 - Partial correlations can be estimated with data.

$y = x'\beta + u$ – “All-causes” framework.

Rubin Causal Model

"Potential Outcomes Framework"

$y_i(1)$ - Wage with college degree^{of indiv.}

$y_i(0)$ - wage of indiv. i without college degree

- Can think about counterfactuals as 'potential outcomes' under different 'treatments':
 - Earnings with college degree, $y(1)$, and without $y(0)$.
 - Health outcome with drug A, $y(1)$, vs. drug B, $y(0)$.
 - Productivity with job training, $y(1)$ vs. without, $y(0)$.
- Define individual treatment effect for individual i as

$$y_i(1) - y_i(0).$$

$(y(1), y(0), D)$

have a
joint
distribution.

$$E(y(1) - y(0)) = \text{ATE}$$

$$E[y(1) - y(0)|D=1] = \text{ATT}$$

$$E[y(1) - y(0)|D=0] = \text{ATV}$$

"Average treatment effect on"

Store Upgrades

"Unbiased".

$$D_i = \begin{cases} 1 & \text{if } i \text{ upgraded} \\ 0 & \text{if } i \text{ not upgraded} \end{cases}$$

- $y_i(1)$ denotes sales store i *would* observe with an upgrade.
- $y_i(0)$ denotes sales store i *would* observe without an upgrade.
- Outcome observed

Observed outcome $\rightarrow Y_i = \begin{cases} y_i(1) & \text{if store } i \text{ is upgraded,} \\ y_i(0) & \text{if not.} \end{cases}$

- In other words:

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

Store Upgrades

Individual Causal effect → $y_{ic} - y_{oi}$

Store	D	y_{ii}	y_{oi}	$y_{ic} - y_{oi}$
1	✓		X	X
2	✓	✓	X	X
3	X	X	✓	X
4	X	X	✓	X
5	X	X	✓	X

Store Upgrades

- We never observe the individual causal effect $y_i(1) - y_i(0)$ for any store because only one of the two potential outcomes is ever observed.
- Can ~~try to~~ measure average effects across treated and untreated stores by comparing averages:
 - e.g. Average Treatment Effect in store population (ATE):
 $E(y(1) - y(0))$.
 - e.g. Average Treatment Effect on the treated (ATT):
 $E(y(1) - y(0) | D = 1)$.
 - e.g. Average Treatment Effect on untreated (ATU):
 $E(y(1) - y(0) | D = 0)$.

Which expectations can we estimate?

Can estimate $E(Y|D=1)$, $E(Y|D=0)$.

- First write out the conditional mean of $Y|D$:

$$\begin{aligned} E(Y|D=1) &= E(y(1)D + y(0)(1-D)|D=1) \\ &= E(y(1)|D=1); \end{aligned}$$

$$\begin{aligned} E(Y|D=0) &= E(y(1)D + y(0)(1-D)|D=0) \\ &= E(y(0)|D=0). \end{aligned}$$

- We can observe $E(y(1)|D=1)$ and $E(y(0)|D=0)$.
Suppose $E(y(1)|D=1) = 100$ and $E(y(0)|D=0) = 50$.

Averages for different groups $D=0$, $D=1$.

Average potential outcomes

$P(D=0)$ can be estimated.

	$D_i = 0$	$D_i = 1$	All
Avg. Sales if store is upgraded	$E[y(0) D=0]$ =?	$E[y(1) D=1]$ = 100	$E[y(1)] = ?$
Avg. Sales if not upgraded	$E[y(0) D=0]$ = 50	$E[y(0) D=1]$ =?	$E[y(0)] = ?$
Difference	$E[y(1)-y(0) D=0]$ =? $= ATT$	$E[y(1)-y(0) D=1]$ =? $= ATE$	$E[y(1)-y(0)]$ =?

Questions

- Should we upgrade stores that have not already been upgraded? -> ATU
- What was the effect of the upgrade on stores that have already been upgraded? -> ATT
- If upgraded all stores to begin with, what would be the average effect on sales be? -> ATE
- Data reveal an estimate of the 'naive comparison':

$$E(y(1)|D=1) - E(y(0)|D=0) = 50.$$

$$100 - 50 = 50.$$

Relationship between ATE,ATT,ATU

Law of total expectation.

- We can decompose the ATE as follows:

$$\begin{aligned} E(y(1) - y(0)) &= E(y(1) - y(0) | D = 1) \cdot P(D = 1) \\ &\quad + E(y(1) - y(0) | D = 0) \cdot P(D = 0) \\ &= ATT \cdot P(D = 1) + ATU \cdot P(D = 0) \end{aligned}$$

by the law of iterated expectation.

Regression

$$E(y(1)|D=1) - E(y(0)|D=0)$$

- Write out the conditional mean of $Y|D$:

$$E(Y|D=1) = E(y(1)|D=1);$$

$$E(Y|D=0) = E(y(0)|D=0).$$

It follows that

$$E(Y|D) = \beta_0 + \beta_1 D,$$

where

$$\beta_0 + \beta_1 = E(y(1)|D=1).$$

$$\beta_0 = E(y(0)|D=0)$$

$$\beta_1 = E(y(1)|D=1) - E(y(0)|D=0).$$

So β_1 equals the naive comparison, and we get an unbiased estimate by regressing Y on D .

Regression

- Recall: The slope coefficient in linear regression is proportional to correlation, and correlation may not reflect causality.
- What is β_1 ?

$$\begin{aligned}\beta_1 &= E(y(1)|D=1) - E(y(0)|D=0) \\ &= \underbrace{E(y(1)|D=1) - E(y(0)|D=1)}_{ATT} \\ &\quad + \underbrace{E(y(0)|D=1) - E(y(0)|D=0)}_{Selection\ Bias}\end{aligned}$$

- β_1 measures the average effect on stores that were upgraded (ATT) but is offset by “selection bias”, which measures differences in average sales in the absence of upgrades.

Treatment Assignment

- Treatment assignment tells us how units are assigned to treatment.
- Random assignment:

$$D_i \perp y_i(0), y_i(1).$$

- This says treatment choice/assignment is independent of outcomes under treatment and no treatment.
 - Often violated when treatment is chosen: Management will try to choose stores for which effect on sales $y_i(1) - y_i(0)$ is greatest.
 - Plausible in experimental setting: Randomized controlled trials.

Random Assignment

- Under random assignment:

$$\begin{aligned} E(y(0)|D) &= E(y(0)); \\ E(y(1)|D) &= E(y(1)); \end{aligned}$$

i.e. the treatment and control groups are representative of the entire population. Follows that

$$\begin{aligned} \beta_1 &= E(y(1)|D = 1) - E(y(0)|D = 0) \\ &= E(y(1)) - E(y(0)) \\ &= ATE. \end{aligned}$$

- Conclusion: Under randomized treatment assignment estimating a linear regression produces an unbiased estimate of the ATE.

$$Y = \beta_0 + \beta_1 D + \varepsilon.$$

Multiple possible treatments

- Suppose in a randomized clinical trial there is a control group and k possible treatments for a particular health condition.
- Suppose participants are randomly assigned to either control or one of the treatments. Let

$$x_{i1} = \begin{cases} 1 & \text{if } i \text{ receives treatment 1,} \\ 0 & \text{otherwise;} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i \text{ receives treatment 2,} \\ 0 & \text{otherwise;} \end{cases}$$

⋮

$$x_{ik} = \begin{cases} 1 & \text{if } i \text{ receives treatment } k, \\ 0 & \text{otherwise;} \end{cases}$$

Multiple possible treatments

- There are now $k + 1$ potential outcomes $y_i(0), y_i(1), \dots, y_i(k)$ and

$$Y_i = y_i(0) + x_{i1}(y_i(1) - y_i(0)) + \dots + x_{ik}(y_i(k) - y_i(0)).$$

See that:

$$E(Y_i | x_{i1} = 0, \dots, x_{ik} = 0) = E(y_i(0) | x_{i1} = 0, \dots, x_{ik} = 0)$$

(Random Assignment) = $E(y_i(0))$

⋮

$$E(Y_i | x_{i1} = 0, \dots, x_{ik} = 1) = E(y_i(k) | x_{i1} = 0, \dots, x_{ik} = 1)$$

(Random Assignment) = $E(y_i(k))$.

Multiple possible treatments

- Now let $x_i = (1, x_{i1}, \dots, x_{ik})'$. We have

$$E(Y_i|x_i) = \beta_0 + \beta_1 x_{i1} + \dots \beta_k x_{ik},$$

where

$$\beta_0 = E(y(0)), \quad \beta_j = E(y(j) - y(0))$$

⋮

$$\beta_k = E(y(k) - y(0)).$$

We can test whether these effects are statistically different from zero (or from each other) by forming appropriate hypotheses, e.g.

$$H_0 : \beta_k = 0; \quad \text{or} \quad H_0 : \beta_1 = \beta_2$$

Questions?

Job Training

- JTRAIN98 contains data on labor market outcomes for men, some of whom enrolled in a job training program in 1997.
- Data:
 - Whether or not individual took training in 1997: $train = 1$ if training taken Θ otherwise
 - Earnings in 1998 (in \$1000s): $earn98$
 - Earnings in 1996 (in \$1000s): $earn96$
 - Whether or not individual is married: $married = 1$ if married
 - Education in years of schooling: $educ$
 - Age in years: age

Job Training

$D = \begin{cases} 1 & \text{with training} \\ 0 & \text{without.} \end{cases}$

consistent est. of
naive comparison.

$$\hat{\beta}_1 = \bar{y}_{\text{train}} - \bar{y}_{\text{not train}} \quad \hat{\beta}_0 = \bar{y}_{\text{not train}}$$

- Regression output measuring differences in labor market outcomes for those with and without training:

```
. reg earn98 train
```

Source	SS	df	MS	Number of obs	=	1,130
Model	1054.41369	1	1054.41369	F(1, 1128)	=	17.91
Residual	66408.4778	1,128	58.872764	Prob > F	=	0.0000
Total	67462.8915	1,129	59.754554	R-squared	=	0.0156
				Adj R-squared	=	0.0148
				Root MSE	=	7.6729

earn98	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
train	-2.050053	.4844142	-4.23	0.000	-3.000507 -1.099599
_cons	10.6099	.279429	37.97	0.000	10.06164 11.15816

$\hat{\beta}_1$ is a comparison of means between individuals who did/did not take training, and indicates individuals who took training earned \$2050 less on average.

Covariate Balance Test

- One way random assignment is justified statistically by a “balance test”.
- Idea: See if individuals with particular covariates are more likely to select into treatment.
- Practically: Regress treatment dummy D_i on observable characteristics x_i :

$$D_i = \gamma_0 + x_i' \gamma_1 + \epsilon_i.$$

- Conduct F -test of $H_0 : \gamma_1 = 0$ vs. $H_1 : \gamma_1 \neq 0$.

Covariate Balance Test

- Test of overall significance rejects with p -value ≈ 0 :

```
. reg train age educ earn96 married
```

Source	SS	df	MS	Number of obs	=	1,130
Model	62.3769648	4	15.5942412	F(4, 1125)	=	93.06
Residual	188.511531	1,125	.167565805	Prob > F	=	0.0000
Total	250.888496	1,129	.222221874	R-squared	=	0.2486
				Adj R-squared	=	0.2460
				Root MSE	=	.40935

train	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age	-.0050343	.0012841	-3.92	0.000	-.0075539 -.0025147
educ	-.0123111	.0043717	-2.82	0.005	-.0208886 -.0037336
earn96	-.019028	.0011428	-16.65	0.000	-.0212703 -.0167858
married	-.1240124	.0289627	-4.28	0.000	-.1808395 -.0671854
_cons	.9464928	.0731975	12.93	0.000	.8028737 1.090112

Issues with Balance Tests

- If the significance level is 5%, we will reject H_0 5% of the time even when $\gamma_1 = 0$.
 - This means calling into question validity of 5% of experimental results.
- If the balance test fails to reject, it doesn't mean randomization holds. D_i may just be independent of observed covariates, leading to $\gamma_1 = 0$.
- What is really needed is

$$D_i \perp y_i(0), y_i(1).$$

Selection on Observables

- Since individuals choose whether to take training, it is unreasonable to assume they are randomly assigned.
 - Previous labor market outcomes, age and work experience all likely to impact selection into training.
 - In other words: $y_i(1), y_i(0)$ likely correlated with D_i .
- Treatment assignment is unconfounded if, conditional on observed covariates x_i , treatment randomly assigned:

$$D_i \perp y_i(0), y_i(1) | x_i.$$

$P(DEA, (y_0, y_1) \in B | X) = P(DEA | X) \cdot P(y_0, y_1 \in B | X)$

for any events A, B . Conditional independence does not imply independence. converse doesn't hold either.

Selection on Observables: Example

Suppose I want to estimate $E(w)$ w - wage. $D = \begin{cases} 1 & \text{if employed} \\ 0 & \text{otherwise} \end{cases}$
 Only observe wages/earnings potential for individuals that are employed.

$$E(w) = \underbrace{E(w|D=1)}_{\text{Known}} P(D=1) \xrightarrow{\text{Known}} + \underbrace{E(w|D=0)}_{\text{Unobserved}} P(D=0) \xleftarrow{\text{Known}}$$

$$E(w|D=0) \geq 0.$$

Assume $w \geq 0$: $E(w) \geq E(w|D=1) P(D=1)$

- Assume $F_w(\cdot | D=1) = F_w(\cdot | D=0)$

Dist. of wages for employed individuals same as for unemployed individuals.

$E(w | D=1) = E(w | D=0) \Rightarrow E(w)$ is now point identified.

Selection on Observables: Example

Suppose $D=1 \Leftrightarrow W \geq R$ where R is a reservation wage.

Does "missingness at random": $F_{W|D=1}(w) = F_{W|D=0}(w)$ imply missingness at random conditional on R also?

$$\begin{aligned} E(W | D=1, R=r) &= E(W | W \geq R, R=r) \\ &= E(W | W \geq r, R=r) \geq r. \end{aligned}$$

$$\begin{aligned} E(W | D=0, R=r) &= E(W | W \leq R, R=r) \\ &= E(W | W \leq r, R=r) \leq r. \end{aligned}$$

Unconfoundedness may fail to hold if we condition on too much.

Selection on Observables

- Idea: If we fix prior labor market experience, age, etc. participation is as good as random. \rightarrow Unconfoundedness.
- Implementation: Add in covariates to regression model to control for selection based on observable characteristics.
- Unconfounded treatment assignment implies:

$$E(y(1)|D, x) = E(y(1)|x);$$

$$E(y(0)|D, x) = E(y(0)|x).$$

- We can estimate

$$E(y(1)|D = 1, x); \quad E(y(0)|D = 0, x).$$

$$\begin{array}{ccc} \downarrow & & \downarrow \\ E(Y|D=1, x) & & E(Y|D=0, x). \end{array}$$

Selection on Observables

- Now we can estimate the ATE by aggregating over x :

$$\begin{aligned} \text{ATE} &= E(y(1) - y(0)) \stackrel{\text{ATE}}{=} E(E(y(1) - y(0)|x)) \\ &= E(E(y(1)|x) - E(y(0)|x)) \\ &= E(E(y(1)|D=1, x) - E(y(0)|D=0, x)) \\ &= E_x(E(Y|D=1, x) - E(Y|D=0, x)) \end{aligned}$$

Need to estimate $E(Y|D=1, x)$ and $E(Y|D=0, x)$.

Selection on Observables

$$E(Y|D=1, x) = \beta_0 + \beta_1 + x' \beta_2$$

$$E(Y|D=0, x) = \beta_0 + x' \beta_2$$

- We approximated conditional means with a linear model

$$E(Y_i|D_i, x_i) \approx \beta_0 + \beta_1 D_i + x'_i \beta_2.$$

This approximation gives $ATE = E(E(Y|D=1, x) - E(Y|D=0, x))$.

$$E(Y|D=1, x) - E(Y|D=0, x) \approx \beta_1.$$

This implies that $\beta_1 = E(y(1) - y(0))$ (which is the ATE).

- Summary: Controlling for x allowed us to identify the ATE because treatment assignment is random conditional on x .

Selection on observables + linear regression allows us to estimate ATE as the coefficient on D .

Selection on Observables

$$\begin{aligned} & E(y(1) - y(0)) \\ &= E(E(y(1) - y(0)|x)) \\ &= \alpha_1 - \alpha_0 \end{aligned}$$

- The approximation works if:

$$\begin{aligned} E(y(0)|x) &= \alpha_0 + x'\beta_2 \\ E(y(1)|x) &= \alpha_1 + x'\beta_2 \end{aligned}$$

so that by the LIE: Unconf: $E(y(0)|D, x) = \alpha_0 + x'\beta_2$.

$$ATE = E(y(1) - y(0)) = \alpha_1 - \alpha_0.$$

It follows by unconfoundedness that:

$$Y = y(1)D + y(0)(1-D)$$

$$E(Y|D, x) = E(y(0) + D(y(1) - y(0))|D, x)$$

$$\begin{aligned} E(D(y(1) - y(0))|D, x) &= \underbrace{\alpha_0}_{\beta_0} + x'\beta_2 + D\underbrace{(\alpha_1 - \alpha_0)}_{\beta_1} \\ &= \beta_0 + \beta_1 D + x'\beta_2. \end{aligned}$$

$$= D(\alpha_1 - \alpha_0).$$

Job Training After controlling for pre-program earnings + other covariates, a much different estimate of ATE is obtained. Conditional on X , program estimated to increase earnings on average by \$2,410.

- Regression controlling for observables greatly alters treatment effect estimate:

```
. reg earn98 train earn96 educ age married
```

Source	SS	df	MS	Number of obs	=	1,130
Model	27320.1797	5	5464.03593	F(5, 1124)	=	152.99
Residual	40142.7118	1,124	35.7141564	Prob > F	=	0.0000
Total	67462.8915	1,129	59.754554	R-squared	=	0.4050

Effect statistically significant and positive ($t=5.54$).

$$H_0: \beta_1 = 0$$

$$VSH_1: \beta_1 \neq 0.$$

Note that these standard errors are not robust to heteroskedasticity!

earn98	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
train	2.410547	.4352625	5.54	0.000	1.556528 3.264565
earn96	.3725384	.0186262	20.00	0.000	.3359923 .4090845
educ	.3628329	.064047	5.67	0.000	.2371678 .4884979
age	-.181046	.018875	-9.59	0.000	-.2180803 -.1440118
married	2.481719	.4262625	5.82	0.000	1.64536 3.318079
_cons	4.667042	1.145283	4.08	0.000	2.419908 6.914176

Interaction Terms

- Previously we assumed

$$\begin{aligned} E(y(0)|x) &= \alpha_0 + x'\beta_2 \\ E(y(1)|x) &= \alpha_1 + x'\beta_2 \end{aligned}$$

which implied, under unconfoundedness

$$\begin{aligned} E(Y|D, x) &= E(y(0) + D(y(1) - y(0))|D, x) \\ &= \alpha_0 + (\alpha_1 - \alpha_0)D + x'\beta_2 \end{aligned}$$

Interaction Terms

- May think that potential outcomes should have different β :

$$\begin{aligned} E(y(0)|x) &= \alpha_0 + x'\beta^0; \\ E(y(1)|x) &= \alpha_1 + x'\beta^1. \end{aligned}$$

- The ATE (τ) is therefore written as:

$$\begin{aligned} \text{ATE} \\ E(y(1) - y(0)) &= E(E(y(1)|x) - E(y(0)|x)) \\ &= E[(\alpha_1 - \alpha_0) + x'(\beta^1 - \beta^0)]. \end{aligned}$$



$$\begin{aligned} &= \hat{\alpha}_1 - \hat{\alpha}_0 + E(x') (\hat{\beta}^1 - \hat{\beta}^0) \\ &\approx \hat{\alpha}_1 - \hat{\alpha}_0 + \bar{x}' (\hat{\beta}_1 - \hat{\beta}_0). \end{aligned}$$

Interaction Terms

$$E(Y|D=1, x) \approx \hat{\alpha}_1 + x' \hat{\beta}^1$$
$$E(Y|D=0, x) \approx \hat{\alpha}_0 + x' \hat{\beta}^0.$$

- Under unconfoundedness:

$$E(y(0)|D=0, x) = \alpha_0 + x' \beta^0;$$
$$E(y(1)|D=1, x) = \alpha_1 + x' \beta^1.$$

- Then natural to estimate $E(y(1) - y(0))$ as

$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^n [\hat{y}_i(1) - \hat{y}_i(0)] = \underbrace{\hat{\alpha}_1 - \hat{\alpha}_0}_{\hat{\tau}} + \bar{x}'_n \left(\hat{\beta}^1 - \hat{\beta}^0 \right).$$

Interaction Terms

$$\begin{aligned} & E(D(y(1) - y(0)) | D, X) \\ &= D \cdot E(\alpha_1 - \alpha_0 + x'(\beta^1 - \beta^0) | D, X) \end{aligned}$$

- Having different slopes changes linear regression model:

$$\begin{aligned} E(Y|D, x) &= E(y(0) + D(y(1) - y(0)) | D, x) \\ &= \alpha_0 + \underbrace{(\alpha_1 - \alpha_0)}_{\tau} D + x' \beta^0 + D x' \underbrace{(\beta^1 - \beta^0)}_{\rho} \end{aligned}$$

- Estimating this model with the entire sample is equivalent to estimating the two regressions separately.

Regress Y on constant + X for $D=1$.
Regress Y on constant + X for $D=0$.

The Overlap Condition

$$0 < P(D=1|x=x') < 1$$

- Identified ATE as

$$ATE = E(y_1 - y_0) = E[E(Y|D=1, x) - E(Y|D=0, x)].$$

- Implicitly assumed overlap condition:
 $0 < P(D=1|x=x') < 1$ for all x' .
- If for some x' , $P(D=1|x') = 0$, then there are no observations of Y for which $D=1$ and $x=x'$, so $E(Y|D=1, x=x')$ is undefined.
- In practice, with several covariates, many pairs (d, x') will have very few observations (curse of dimensionality).

Non-parametric estimates of the ATE $E(Y|A) := \frac{E(Y1_A)}{P(A)}$

- If x is discrete, we can use non-parametric estimate which averages Y for each (D, x) pair:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

$$\begin{aligned} E(Y|D=1, x=x') &= \frac{E(Y\mathbf{1}(D=1, x=x'))}{P(D=1, x=x')}, \\ &\approx \frac{\frac{1}{n} \sum_{i=1}^n Y_i \mathbf{1}(D=1, x=x')}{\frac{1}{n} \sum_{i=1}^n \mathbf{1}(D=1, x=x')} := \hat{\mu}_1(x'). \end{aligned}$$

where

$$\mathbf{1}(A) = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

- Estimate $E(Y|D=0, x=x') \approx \hat{\mu}_0(x')$ analogously. Then

$$\begin{aligned} \widehat{ATE} &:= \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(x_i) - \hat{\mu}_0(x_i) \\ &\approx E[E(Y|D=1, x) - E(Y|D=0, x)]. \end{aligned}$$

Non-parametric estimates of the ATE

- Suffers from curse of dimensionality if x is high-dimensional.
- If x is continuous, can use a non-parametric kernel estimator (compute average by including observations in a range of values for x , weighting $x \approx x'$ higher):

$$\hat{\mu}_1(x') := \frac{\frac{1}{n} \sum_{i=1}^n Y_i \mathbf{1}(D = 1, x \in [x' - h, x' + h])}{\frac{1}{n} \sum_{i=1}^n \mathbf{1}(D = 1, x \in [x' - h, x' + h])}$$

Uniform Kernel

- h is called a “bandwidth” and $h \rightarrow 0$ as $n \rightarrow \infty$ to ensure consistency.

Matching

- Matching provides an alternate method of estimating ATE based on unconfoundedness.
- Idea: Match unit with $D = 0, x = x'$ to one with $D = 1, x = x'$.
“Mahalanobis Distance”
- For x continuous, can match units with $x_i \approx x_j$ and construct a proxy for $y_i(d)$ using, e.g. k nearest neighbors:

$$\hat{y}_i(d) = \begin{cases} y_i & \text{if } D = d, \\ \frac{1}{k} \sum_{j \in \mathcal{M}_i} y_j & \text{if } D \neq d, \end{cases}$$

where \mathcal{M}_i is the set of points “closest” to i .

Propensity Score Matching

- Asserting $E(y(d)|x) = \alpha_d + x'\beta_d$ alleviates curse of dimensionality (approximate conditional means with linear regression to avoid the “small-bin” problem).
- Another way to reduce the dimension of the set of conditioning variables is propensity score matching.
- Rosenbaum, Rubin (1983):

$$D_i \perp y_i(0), y_i(1) | x_i \implies D_i \perp y_i(0), y_i(1) | p(x_i),$$

where

$$p(x') = P(D = 1|x = x').$$

is the propensity score.

Propensity Score Matching

- Adding/removing covariates can violate unconfoundedness, but this works because:

$$\begin{aligned} P(D = 1 | y_0, y_1, p(x)) &\stackrel{\text{LIE}}{\equiv} E(P(D = 1 | y_0, y_1, p(x), x) | y_0, y_1, p(x)) \\ &= E(P(D = 1 | y_0, y_1, x) | y_0, y_1, p(x)) \\ &\stackrel{\text{Unc} \cancel{\text{ondid}} \text{ness}}{\equiv} E(P(D = 1 | x) | y_0, y_1, p(x)) \\ &= E(p(x) | y_0, y_1, p(x)) \\ &= p(x). \end{aligned}$$

- After conditioning on $p(x)$, D is independent of (y_0, y_1) .
- Identification of ATE via propensity score:

$$ATE = E[E(Y|D = 1, p(x)) - E(Y|D = 0, p(x))]$$

Propensity Score Matching

- Avoid curse of dimensionality because $p(x)$ is scalar even if x high dimensional.
- However, p still needs to be estimated, so curse of dimensionality now features in estimation of p .
- Exception: In a randomized experiment, $p(x)$ is known (by design).
- Suggests estimate:

$$E(Y|D=1, p(x) = p') \approx \frac{\frac{1}{n} \sum_{i=1}^n Y_i \mathbf{1}(D=1, \hat{p}(x) = p')}{\frac{1}{n} \sum_{i=1}^n \mathbf{1}(D=1, \hat{p}(x) = p')}.$$

Propensity Score Matching

- Hirano, Imbens and Ridder (2003) use representation of ATE as unconditional weighted average:

$$ATE = E \left(\frac{Y(D - p(x))}{p(x)(1 - p(x))} \right),$$

estimated by sample analog principle:

$$\widehat{ATE} = \frac{1}{n} \sum_{i=1}^n \frac{Y_i(D_i - \hat{p}(x_i))}{\hat{p}(x_i)(1 - \hat{p}(x_i))}.$$

- Show that (under regularity conditions) this estimate is asymptotically normal and efficient.

Propensity Score Matching

Propensity Score Matching

Questions?