STA305/1004 - Design of Scientific Studies

Tiffany Fitzpatrick

Department of Statistical Sciences, University of Toronto

August 2, 2018

Slides prepared by Lin Zhang

Thanks to Becky Wei Lin and Nathan Taback for the slides

Agenda Today

1/ SUTVA : a) 1 form trt / control b) No interfer- Among unit

2/ Fundamental problem - 1 PO observed for a unit.
↳ Even if SUTVA

3/ Assignment Mechanisms - how trt / control chosen
↳ Even if SUTVA
. ignorable = indep PO
. unconfounded: indep obs'd + unobs'd PO

- Introduction to Casual Inference
- Maimoides' Rule
- The propensity score - smoking cessation study
- The balancing property of propensity score
- Three different ways to calculate the effect of treatment using propensity score

Ex. perfect Dr - Almost perfect Dr
Non-ignorable + confounded

Ex. Lord's paradox - Dining Hall $P(D|x)=1$
unconfounder

Ex. Reading program
- teacher decision ≈ pretest

Introduction to Causal Inference

The design phase of an observational study

Good observational studies are designed.

An observational study should be conceptualized as a broken randomized experiment ... in an observational study we view the observed data as having arisen from a hypothetical complex randomized experiment with a lost rule for the [assignment mechanism], whose values we will try to reconstruct.

- Rubin (2007)

observe only
No Experimenters

The design phase of an observational study

> *Of critical importance, in randomized experiments the design phase takes place prior to seeing any outcome data. And this critical feature of randomized experiments can be duplicated in observational studies, for example, using propensity score methods, and we should objectively approximate, or attempt to replicate, a randomized experiment when designing an observational study.*
>
> - Rubin (2007)

prior to
should ignorable Mechanism

Looking for Causal Treatment Effects

- Compare survival in patients who are treated with a drug therapy versus surgery in a randomized control trial.
- If patients that undergo surgery live 8 years and patients that take drug therapy live 5 years then the treatment effect is 8-5=3. yrs

Observational versus Randomized Studies

- In observational studies the researcher does not randomly allocate the treatments.
- Randomization ensures subjects receiving different treatments are comparable.

Not true Always in observational data

Observational Studies

Table1

trt	drug	surg
	Y.	T.
Age		
sex		
#diseases		
⋮		

- Suppose that we have an outcome measured on two groups of subjects (treated and control).
- We want to make a fair comparison between the treated group and the control group in terms of the outcome.
- We can obtain covariates that describe the subjects before they received treatments, but we can't ensure that the groups will be comparable in terms of the covariates.

Importance of Randomization

- Randomization tends to produce relatively comparable or "balanced" treatment groups in large experiments.
- The covariates aren't used in assigning treatments in an experiment. Not Explicitly Done But Nice Benefit
- There's no deliberate balancing of the covariates – it's just a nice feature of randomization.
- We have some reason to hope and expect that other
  * (unmeasured) variables will be balanced, as well.

Smoking cessation and weight gain

trt / Exposure — outcome

What is the effect of smoking on weight gain?

- Data was used from The National Health and Nutrition Examination Survey Data I Epidemiological Follow-up Study (NHESFS) survey to assess this question.
- The NHESFS has information on sex, race, weight, height, education, alcohol use, and intensity of smoking at baseline (1971-75) and follow-up (1982) visits.
- The survey was designed to investigate the relationships between clinical, nutritional, and behavioural factors.
- A cohort of persons 25-74 who completed a medical exam in 1971-75 followed by a series of follow-up studies.

Ethical?

Would it be possible to conduct a randomized study to evaluate the treatment effect? Not feasible: current smk < Quit / Not

Smoking cessation and weight gain

Table 1	Quit	Didnt Quit
	Cessation (A=1)	No cessation (A=0)
age, years	46.2	42.8
men, %	54.6	46.6
white, %	91.1	85.4
university, %	15.4	9.9
weight, kg	72.4	70.3
Cigarettes/day	18.6	21.2
year smoking	26.0	24.1
little/no exercise, %	40.7	37.9
inactive daily life, %	11.2	8.9

Maimonides' Rule

- Maimonides' rule is named after the 12th-century rabbinic scholar Maimonides, who identified a correlation between class size and students' achievements. Smaller size = Better grades
- Today this rule is widely used in educational research to evaluate the effect of class size on students' test scores. Maimonides' rule states that a class size may rise to an upper limit of 40 students.
- Once this quota is reached the class is cut in half, so instead of one class with forty-one students there are now two classes: one with twenty students and one with twenty-one students.

source: `https://en.wikipedia.org/wiki/Maimonides%27_rule`

Maimonides' Rule

- Since 1969, Maimonides' Rule has been used to determine the division of enrollment cohorts into classes in Israeli public schools.
- Class size usually determined by affluence or poverty of a community, enthusiasm or skepticism about the value of education, special needs of students, etc.
- With class size of 40 and total students 40, 80 and 120 students in the grade cohort, the effects of an additional student enrolled in the cohort on class size are summarized below:

smallest classes ↓

Biggest ↓

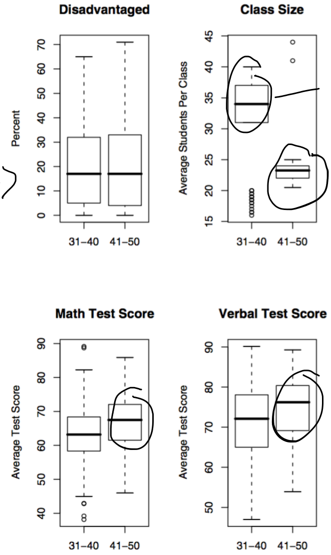
	Total	Class size	Total	Class size	Total	Class size
Original	40	40	80	40×2	120	40×3
+1 Student	41	21+20	81	27×3	121	$30 \times 3 + 31$

Maimonides' Rule

- Angrist and Lavy (1999) published a study of the effects of class size on academic achievement.
- Causal effects of class size on pupil achievement is difficult to measure.
- They looked at schools with fifth grade cohorts in 1991 with between 31 and 50 students, where the average class sizes might be cut in half by Maimonides' rule (an upper limit of 40 students per class).
- There were 211 such schools, with 86 of these schools having between 31 and 40 students in fifth grade cohort, and 125 schools having between 41 and 50 students in the fifth grade cohort.

original paper: *Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement*
`https://journals-scholarsportal-info.myaccess.library.utoronto.ca/details/00335533/v114i0002/533_umrtetocsosa.xml`

Preliminary results by Angrist and Lavy (1999)



Balance

~35

~21

Smaller classes = Better test score

Fig. 1.1 Eighty-six pairs of two Israeli schools, one with between 31 and 40 students in the fifth grade, the other with between 41 and 50 students in the fifth grade, matched for percentage of students in the school classified as disadvantaged. The figure shows that the percentage of disadvantaged students is balanced, that imperfect adherence to Maimonides' rule has yielded substantially different average class sizes, and test scores were higher in the group of schools with predominantly smaller class sizes.

Preliminary results by Angrist and Lavy (1999)

Among the 211 schools with between 31 and 50 students in fifth grade cohort, Angrist and Lavy (1999) investigated the correlation between class size, a factor with two levels (Level 1: class size 31-40, Level 2: class size 41-50).

- The correlation between class size and the percentage disadvantaged students is -0.10
- The correlation between class size and the performance on verbal test is -0.42
- The correlation between class size and the performance on math test is -0.55

For this reason, 86 matched pairs of two schools were formed, matching to minimize the total absolute difference in percentage of disadvantaged students. Balanced on disadvantage

Angrist and Lavy (1999)

trt = class size

86 pairs of Israeli schools that are matched as follows:

- Level 1 ($T = 0$): school with between 31 and 40 students in the fifth grade cohort, and
- Level 2 ($T = 1$): school with between 41 and 50 students in the fifth grade cohort.

The school pairs were matched for a covariate $x =$ percentage of disadvantaged students.

Impact of Maimonides' Rule on Class Size in Angrist and Lavy (1999)

The Israeli schools strictly adhere to to Maimonides' rule, then

- School with 31-40 students in the fifth grade cohort will be taught in one large class (class size between 31-40)
- School with 41-50 students in the fifth grade cohort will be taught in two small classes (class size between 20-25)

Adherence to Maimonides' rule was imperfect but strict enough to produce a wide separation in typical class size.

Angrist and Lavy (1999)

- $Y(T = 0)$ is the average test score in the fifth grade for 31-40 students in the cohort, that is, the schools with a smaller cohort and large class size ($T = 0$).
- $Y(T = 1)$ is the average test score in the fifth grade for 41-50 students in the cohort, that is, the schools with a larger cohort and a smaller class size ($T = 1$).

Question: What separates a school with 31–40 students in the fifth grade cohort ($T = 0$) and a school with 41–50 students in the fifth grade cohort ($T = 1$)?

$\approx$ 10 students

Answer: A handful of grade fifth students.

Angrist and Lavy (1999)

Ex1. distance to hospital → Heart attack (survival)

Ex2 proximity to university → PS → $

In the study by Angrist and Lavy,

- it seems plausible that whether or not a few more students enroll in the fifth grade is a haphazard event.
- An event that is not strongly related to the average test performance that the fifth grade students would exhibit with a larger or smaller cohort.
- It does seem reasonable to think that probability of a larger class are fairly close for the 86 paired Israeli schools.
- This might be implausible in some other context, say in a national survey in Canada where schools are funded by local governments, so that class size might be predicted by the wealth or poverty of the local community.

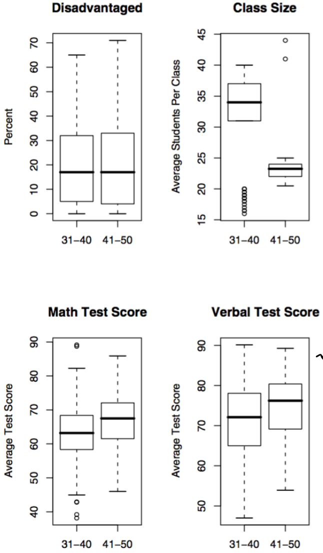
A 'natural experiment' is an attempt to find in the world some rare circumstance such that a consequential treatment was handed to some people and denied to others for no particularly good reason at all, that is, haphazardly.

INSTRUMENTAL VARIABLES

Ex. cig prices/tax → Smoke → lung cancer

Preliminary results by Angrist and Lavy (1999)

*IV imp when RCT isn't feasible



class size

The figure shows that treatment reflects the haphazard element, namely susceptibility to Maimonides' rule based on cohort size, rather than realized class size.

opposite "AS treated"

Defining treatment in this way is analogous to an 'intention-to-treat' analysis in a randomized experiment.

Assumption: Everybody assigned trt got it & no controls

The propensity score

- Covariates are pre-treatment variables and take the same value for each unit no matter which treatment is applied.
- For example, pre-treatment blood pressure or pre-test reading level are not influenced by a treatment that would alter blood pressure or reading level.

The propensity score is defined as

Prop to get trt

$$\underline{e(\mathbf{x})} = P\left(T = 1|\mathbf{x}\right),$$

#hrs read
Age
pretest

where $\mathbf{x}$ are observed covariates.

Propensity score is the probability that a unit receives treatment ($T = 1$), given all the covariates that are observed before the treatment.

The propensity score

Assignment Mechanism
Flip coin
prop (trt) = 0.5
prop (control) = 0.5

- In experiments, the propensity scores are known.
- In observational studies, they can be estimated using models such as logistic regression where the outcome is the treatment indicator and the predictors are all the confounding covariates.

Exercise: Why are the propensity scores known in experiments?

treated or not ~ all covariates that influence likelihood to get treated

The propensity score

Exercise 1: Consider a completely randomized design with $n = 2$ units and one unit is assigned treatment. The treatment assignment for the i^{th} subject is:

T_1	T_2	$P(T_1)$	$P(T_2)$
0	0	0	0
0	1	0	0.5
1	0	0.5	0
1	1	0	0

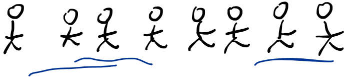
What is the propensity score of each unit?

$p = 0.5$
treated

The propensity score

$1/56$ $1/56$

$\binom{8}{3} = 56$



Exercise 2: Consider a completely randomized design with $n = 8$ units and 3 units are assigned treatment.

- What is the propensity score of each unit? $p = 3/8$
- What is the probability of an particular treatment assignment?

Exercise 3: Consider a completely randomized design with n units and m units are assigned treatment. $m \leq n$

- What is the probability that a unit receives the treatment? m/n
- What is the probability that a unit receives the control? $1 - m/n$
- What is the propensity score of each unit? $p = m/n$
- What is the probability of an particular treatment assignment?

$1/\binom{n}{m}$

Logistic Regression - Review

Consider a logistic regression model with one covariate X, and the response is T, whether a subject receives the treatment.

$$T_i | X_i = x_i \sim Bernoulli(p_i), \text{ where } p_i = \beta_0 + \beta_1 x_i$$

p_i is the probability that the i^{th} subject receives the treatment, that is,

$$p_i = P(T_i = 1 | X_i = x_i)$$

propensity score for ith person

Then, we can rewrite the above logistic regression model with,

log $\therefore p \approx (0,1)$

$$log\left(\frac{P(T_i = 1 | X_i = x_i)}{P(T_i = 0 | X_i = x_i)}\right) = \beta_0 + \beta_1 x_i$$

If we have k covariates $X_1, X_2, \ldots, X_k$, a general form of the logistic regression model is

$$log\left(\frac{P(T_i = 1 | X_i = x_i)}{P(T_i = 0 | X_i = x_i)}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_1 x_{ik}$$

Logistic Regression - Review

Consider the logistic regression model with one binary covariate X,

Eg. sex
$M = 0$
$F = 1$

$$log\left(\frac{P(T_i = 1|X_i = x_i)}{P(T_i = 0|X_i = x_i)}\right) = \beta_0 + \beta_1 x_i,$$

$M \rightarrow B_0$
$F \rightarrow B_0 + B_1$

it can be shown easily that

prop. to Be treated

$$\frac{(P(\boxed{T = 1}|x = 1)/P(T = 0|x = 1))}{(P(T = 1|x = 0)/P(T = 0|x = 0))} = exp(\beta_1)$$

where $exp(\beta_1)$ is the odds ratio of getting the treatment given $x = 1$ and with getting the treatment given $x = 0$.

$\frac{F}{M} = \log\left(\frac{B_0 + B_1}{B_0}\right)$
$= \log(B_1)$
$OR = e^{B_1}$

Logistic Regression - Review

Exercises: In a logistic model with one binary covariate, show that

1. $$\frac{(P(T=1|x=1)/P(T=0|x=1))}{(P(T=1|x=0)/P(T=0|x=0))} = exp(\beta_1)$$

2. $$\hat{p}_i = P(T_i = 1|X_i = x_i) = \frac{exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{1 + exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}$$

predicted ith propensity score

The propensity score - Doctor's Medical Records Example

- Consider a study that plans to use a doctor's medical records to compare two treatments ($T = 0$ and $T = 1$) given for a certain condition.
- Treatments were not assigned to patients randomly, but were based on various measured and unmeasured patient factors.
- The patient factors that were measured are age (x_1), sex (x_2), and health status before treatment (x_3).

other diseases

The propensity score can be estimated for each patient by fitting a logistic regression model with treatment as the dependent variable and x_1, x_2, x_3 as the predictor variables.

outcome = trt or not

$$log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3},$$

predictors x_1, x_2, x_3

where $p_i = P(T_i = 1 | x_{i1}, x_{i2}, x_{i3})$.

Exercise: Use the above equation to find $\hat{p}_i$.

How do we build a propensity score model?

- Usual tool is logistic regression model for the treatment allocation decision
- We therefore want to consider including any variables that have a relationship to the treatment decision (i.e. precede it in time, and are relevant)
- No information is included on the actual treatment received, or on the outcome(s).

Propensity model development

Q: Which diagnostics is more important for propensity score model development?

- Diagnostics for the successful prediction of probabilities and parameter estimates underlying those probabilities
- Diagnostics for the successful design of observational studies based on estimated propensity scores. Assessing 2 trt group Balanced

A: In propensity score model development the second point is important, but the first is not important .

When developing a propensity model, try to include

- all covariates that subject matter experts (and subjects) judge important when selecting treatments.
- confounder
  all covariates that relate to treatment and outcome, including any covariate that improves prediction (of exposure group).
- as much "signal" as possible.

The propensity score in Smoking Cessation Study

NHANES

cessation → wt gain

The probability of quiting smoking for each subject in the smoking and weight gain study can be estimated by fitting a logistic regression model. Quit

```
prop.model <- glm(qsmk ~ as.factor(sex) + as.factor(race)+
                  age+as.factor(education.code)+smokeintensity+
                  smokeyrs + as.factor(exercise)+as.factor(active) +
                  wt71, family = binomial(), data = nhefshwdat)
```

#cigs/day

The propensity score in Smoking Cessation Study

```
round(summary(prop.model)$coef, digits=3)
```

```
                             Beta
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -2.401      0.484  -4.961    0.000
as.factor(sex)1              -0.499      0.147  -3.406    0.001
as.factor(race)1             -0.778      0.207  -3.759    0.000
age                           0.046      0.010   4.672    0.000
as.factor(education.code)2   -0.066      0.196  -0.335    0.738
as.factor(education.code)3    0.053      0.176   0.300    0.764
as.factor(education.code)4    0.109      0.269   0.404    0.686
as.factor(education.code)5    0.466      0.224   2.080    0.038
smokeintensity               -0.027      0.006  -4.683    0.000
smokeyrs                     -0.028      0.010  -2.847    0.004
as.factor(exercise)1          0.360      0.179   2.013    0.044
as.factor(exercise)2          0.423      0.186   2.277    0.023
as.factor(active)1            0.045      0.132   0.342    0.733
as.factor(active)2            0.158      0.213   0.741    0.459
wt71                          0.006      0.004   1.396    0.163
```

The propensity score in Smoking Cessation Study

The propensity score for each subject is $\hat{p}_i$ is the predicted probability of quitting smoking from the logistic regression model.

```
##predicted propensity score based on logistic regression model
p.qsmk.obs <- predict(prop.model, type = "response")
qsmk_dat <- data.frame(subject=nhefshwdat$X, qsmk=nhefshwdat$qsmk,
                        propensity=p.qsmk.obs)
```

predicted

```
##first 6 subjects in the study. qsmk=1, if the person quit smoking.
head(qsmk_dat)
  subject qsmk propensity
1       1    0  0.1239035
2       2    0  0.1597305
3       3    0  0.1599358
4       4    0  0.3106921
5       5    0  0.3197595
6       6    0  0.1662245
```

~12% chance of quitting

Interpretation of the estimated propensity score:

- Subject 1's estimated probability of quitting smoking is 0.12
- Subject 1's estimated probability of not quitting smoking is 1-0.12=0.82

The propensity score in Smoking Cessation Study

```
predict.glm(prop.model)[1] #predicted value for the first subject
        1
-1.955973

exp(predict.glm(prop.model)[1])/(1+exp(predict.glm(prop.model)[1]))
        1
0.1239035

## use type="response" to get predicted probability
predict.glm(prop.model,type = "response")[1]
        1
0.1239035
```

print 1st obs'n

The balancing property of the propensity score

The balancing property of the propensity score says that treated ($T = 1$) and control ($T = 0$) subjects with the same propensity score $e(\mathbf{x})$ have the same distribution of the observed covariates, $\mathbf{x}$,

$$P(\mathbf{x} | T = 1, e(\mathbf{x})) = P(\mathbf{x} | T = 0, e(\mathbf{x}))$$

or

$$T \perp \mathbf{x} | e(\mathbf{x}).$$

x is set of covariates

This means that treatment is independent of the observed covariates conditional on the propensity score.

if Balanced

The balancing property of the propensity score

The balancing property says that if two units, i and j, are paired, one of whom is treated, $T_i + T_j = 1$, so that they have the same value of the propensity score $e(\mathbf{x}_i) = e(\mathbf{x}_j)$, then they may have different values of the observed covariate,

$$\mathbf{x}_i \neq \mathbf{x}_j,$$

but in this pair the specific value of the observed covariate will be unrelated to the treatment assignment since

$$P\left(\mathbf{x} \middle| T = 1, e(\mathbf{x})\right) = P\left(\mathbf{x} \middle| T = 0, e(\mathbf{x})\right)$$

The balancing property of the propensity score

The propensity scores for subject's 10 and 18 in the smoking cessation study are

```
qsmk_dat <- data.frame(subject=nhefshwdat$X, qsmk=nhefshwdat$qsmk,
                       propensity=p.qsmk.obs)
qsmk_dat[c(10, 18),]
   subject qsmk propensity
10      10    0  0.2941244
18      19    1  0.3197956
```

≈ 0.3

The difference between the two subject's propensity scores are 0.32-0.29=0.03. This could be set as a "caliper" or "tolerance" for what are considered equal propensity scores.

SD

Close enough.

The covariates for each subject are

```
x <- rbind(nhefshwdat[10,3:12],nhefshwdat[18,3:12])
colnames(x) <- c( "age","sex","race", "edu","smkint",
                 "smkyrs","exer","active","wt1971","qsmk")
x
   age sex race edu smkint smkyrs exer active wt1971 qsmk
10  43   0    0   2     20     25    2      1  62.26    0
18  48   1    0   3      2     30    1      1  62.03    1
```

The balancing property of the propensity score

T=1 T=0

Q: If the smoking cessation and smoking groups are balanced using the propensity score then both observed and unobserved covariates will have similar distributions in the two groups. Thus, this observational study has been turned into a randomized study by using propensity score methods.

Balanced only on obs'd
- Family support?
- # Attempts?

A. True

[✓ B.] False

Exercise: What is the difference in using propensity scores to form two groups vs. using randomization to form two groups? obs'd + unobs'd

A: In an observational study, using propensity score to form two similar groups all the observed covariates will be balanced, but the unobserved covariates may not be balanced.

Assessing balance

Sample size

- The difference in average covariate values by treatment status, scaled by their sample standard deviation. This provides a scale-free way to assess the differences.
- As a rule-of-thumb, when treatment groups have important covariates that are more than one-quarter or one-half of a standard deviation apart, simple regression methods are unreliable for removing biases associated with differences in covariates (Imbens and Rubin (2015)).

$\frac{1}{4} - \frac{1}{2}$ SD

close enough

Assessing balance

- $\bar{x}_t$: the sample mean of a covariate in the treated group
- s_t^2: the sample variance of a covariate in the treated group
- $\bar{x}_c$: the sample mean of a covariate in the control group
- s_c^2: the sample variance of a covariate in the control group

The pooled sample variance is

$$\sqrt{\frac{s_t^2 + s_c^2}{2}}. \qquad \approx \text{pooled SD}$$

The absolute pooled standardized difference (in percentage) is,

$$\frac{100 \times |\bar{x}_t - \bar{x}_c|}{\sqrt{\frac{s_t^2+s_c^2}{2}}}. \qquad \approx \text{2 sample t-test}$$

Exercise: Show that $\sqrt{(s_t^2 + s_c^2)/2}$ can be derived from s_p^2 in two-sample t-test assuming $n_t = n_c = n/2$.

Assessing balance - Smoking Cessation Study

- The absolute pooled standardized difference between the groups can be calculated for all the covariates using the function ‘MatchBalance‘ in the library ‘Matching‘.

```
mb <- MatchBalance(qsmk ~ as.factor(sex) + as.factor(race) +
                age + as.factor(education.code) +
                smokeintensity + smokeyrs + as.factor(exercise)+
                as.factor(active)+wt71,data=nhefshwdat,nboots=10)
```

- If the absolute value of the standardized mean difference is greater than 10% then this indicates a serious imbalance.
- Note that the absolute pooled standardized difference is expressed in percentage.

Assessing balance - Smoking Cessation Study

Output from ‘MatchBalance()’. (With some output omitted.)

```
***** (V3) age *****
before matching:
mean treatment........ 46.174
mean control.......... 42.788
std mean diff......... 27.714 >10%
```

- The absolute value of the standardized mean difference is greater than 10%, then this indicates a serious imbalance.
- Age has an absolute standardized mean difference of $|27.714|$, indicating serious imbalance between the groups in age.

Assessing balance - Smoking Cessation Study

```
***** (V2) as.factor(race)1 *****
before matching:
mean treatment........ 0.08933
mean control.......... 0.14617
std mean diff......... -19.905  > 10%
```

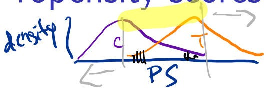
Balance in order to make causal statements

```
***** (V14) wt71 *****
before matching:
mean treatment........ 72.355
mean control.......... 70.303
std mean diff......... 13.13  > 10%
```

Exercises:

1. What are the absolute values of the standardized mean difference in *race* and *weight*?
2. Is there serious imbalance between the groups in *race* and *weight*?

Propensity scores and ignorable treatment assignment



Regression - Extrapolating

🟨 Area of common support

Assume that the treatment assignment T is strongly ignorable. This means that

if Balanced
$p(T=1)=p(T=0)$

$$P(T|Y(0), Y(1), \mathbf{x}) = P(T|\mathbf{x}),$$

or

$$\{T \perp Y(0), Y(1)\}| \, \boldsymbol{x}.$$

- It may be difficult to find a treated and control unit that are closely matched for every one of the many covariates in x,
- but it is easy to match on one variable, the propensity score, $e(\mathbf{x})$, and doing that will create treated and control groups that have similar distributions for all the covariates.

More feasible
if small N

Propensity scores and ignorable treatment assignment

Ignorable treatment assignment implies that

$$P(T|Y(0),Y(1),e(\mathbf{x})) = P(T|e(\mathbf{x})) \text{ or } \{T \perp Y(0),Y(1)\}|e(\mathbf{x})$$

- This means that the propensity score, $e(\boldsymbol{x})$ can be used in place of many covariates, $\boldsymbol{x}$. No convergence issues w/ small cells
- If treatment assignment is strongly ignorable then propensity score methods will produce unbiased results of the treatment effects.
- The treatment assignment mechanism has been reconstructed Rule using the propensity score.

Exercise: In the smoking cessation study, what does it mean for treatment assignment to be ignorable?

A: The potential outcomes for weight gain in the smoking cessation (treated) and smoking (control) groups are independent conditional on the propensity score.

Propensity scores and ignorable treatment assignment

- Suppose a critic came along and claimed that the study did not measure an important covariate (e.g., spouse is a smoker) so the study is in no position to claim that the smoking cessation group and the smoking groups are comparable.
- This criticism could be dismissed in a randomized experiment
  - randomization does tend to balance unobserved covariates
  - but the criticism cannot be dismissed in an observational study.
- This difference in the unobserved covariate, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in the unobserved covariate.
- The sensitivity of an observational study to bias from an unmeasured covariate is the magnitude of the departure from the model that would need to be present to materially alter the study's conclusions.
- There are statistical methods to measure how sensitive an observational study is to this type of bias.

Quantative Bias analysis

Propensity scores and ignorable treatment assignment

Q: If the smoking cessation study were a randomized study comparing weight gain in smokers versus quitters then a valid criticism is that the treatment effect could be due to an unobserved covariate?

A. True

[✓ B.] False. Balance wrt unobs'd AND obs'd covars
in RCT

* RCT GOLD STD

Using the propensity score to reduce bias

- The three most common techniques that use the propensity score are
  1. matching,
  2. stratification (also called sub-classification)
  3. regression adjustment.
- Each of these techniques is a way to make an adjustment for covariates prior to (matching and stratification) or while (stratification and regression adjustment) calculating the treatment effect.
- With all three techniques, the propensity score is calculated the same way, but once it is estimated it is applied differently.

Propensity score matching - Maimonides' Rule

1 trt : 1 control

1:2
1:3
1:4 } *

- In the Maimonides' rule study, assignment to a small/large was haphazard/random.
- If there is no opportunity to take advantage of this type of treatment assignment, then we can calculate the propensity score and use this to match.
- For each unit we have a propensity score. Student ~ small or large class

greedy matching

- Randomly select a treated subject, match to a control subject with closest propensity score (within some limit or "calipers"). optimal

  close PS ≈ 1/4 − 1/2 SD
- Eliminate both units from the pool of subjects until there is no acceptable match. Matching w/o replacement
- It's not always possible to match every unit treated to a unit that is not treated.

Propensity score matching - Smoking Cessation Study

```
prop.model <- glm(qsmk~as.factor(sex) + as.factor(race) + age +
              as.factor(education.code) + smokeintensity + smokeyrs
              + as.factor(exercise)+ as.factor(active) + wt71,
              data = nhefshwdat)
X <- prop.model$fitted;Y <- nhefshwdat$wt82_71; Tr <- nhefshwdat$qsmk
library(Matching)

rr <- Match(Y=Y,Tr=Tr,X=X,M=1); summary(rr)
```

caliper = 1/8 SD

```
Estimate...  3.1479
AI SE......  0.58127
T-stat.....  5.4155
p.val......  6.1104e-08

Original number of observations..............  1566
Original number of treated obs...............  403
Matched number of observations...............  403
Matched number of observations  (unweighted).  983
```

Not all people matched

* INTERNAL VALIDITY Most important

Propensity score matching - Smoking Cessation Study

- The treatment effect in the smoking cessation study is difference in weight gain between the group that stopped smoking and the group that did not stop smoking
- After matching on covariates the treatment effect is 2.93 with a p-value of 5.0087e-07.
- The 95% confidence interval of the treatment effect is

$$2.93 \pm t_{1009,0.05/2}(0.5838) = (1.84, 4.02)$$

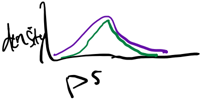
Note that $t_{1009,0.05/2} \approx z_{0.05/2}$

trt Effect
(cessatation)
signifcy ↑ wt

Propensity score matching - check covariate balance

trt + control groups comparable

```
MatchBalance(qsmk ~ as.factor(sex) + as.factor(race) +
                    age + as.factor(education.code) +
                    smokeintensity + smokeyrs  +
                    as.factor(exercise) +
                    as.factor(active) + wt71, data=nhefshwdat,
                    match.out=rr,nboots=10)
##some output
***** (V1) as.factor(sex)1 *****
                       Before Matching      After Matching
mean treatment........     0.45409           0.45409
mean control..........     0.53396           0.44901
std mean diff.........    -16.022            1.0204
```



>10%. <<10%. Balanced

```
***** (V3) age *****
                       Before Matching      After Matching
mean treatment........      46.174           46.174
mean control..........      42.788           45.439
std mean diff.........      27.714           6.0129
```

Balanced

Propensity score matching - check covariate balance

What if we do not adjust for imbalance?

```
t.test(nhefshwdat$wt82_71[which(nhefshwdat$qsmk==1)],
      nhefshwdat$wt82_71[which(nhefshwdat$qsmk==0)], var.equal=T)
  Two Sample t-test

data:  nhefshwdat$wt82_71[which(nhefshwdat$qsmk == 1)] and
nhefshwdat$wt82_71[which(nhefshwdat$qsmk == 0)]
t = 5.6322, df = 1564, p-value = 2.106e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.655796 3.425367
sample estimates:
mean of x mean of y
 4.525079  1.984498
```

little lower but similar estimate

BIAS

- The unadjusted treatment effect is 2.54 with a p-value of 0.
- Both analyses lead to the same conclusion that stopping to smoke leads to a significant weight gain.

  X
- The weight gain in the matched propensity score analysis is 0.39kg higher than unmatched analysis.

Propensity score sub-classification/stratification

|0.2|0.4|0.6|0.8 ⇒ 100-centiles 100 group

0.199 vs. 0.201   0-10 PS

0.01 and 0.19

- Propensity scores permit sub-classification on multiple covariates simultaneously. One advantage of this method is that the whole sample is used and not just matched sets.
- Cochran (1968) showed that creating five strata removes 90% of the bias due to the stratifying variable or covariate.
- Rosenbaum and Rubin holds for stratification based on the propensity score. Stratification on the propensity score balances all covariates that are used to estimate the propensity score, and often five strata based on the propensity score will remove over 90% of the bias in each of these covariates.

Stratification - Death Rates of Male Smokers

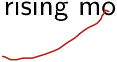
The following data were selected from data supplied to the U. S. Surgeon General's Committee from three of the studies in which comparisons of the death rates of men with different smoking habits were made (Cochran, 1968).
The table shows the unadjusted death rates per 1,000 person-years.

Smoking group	Canadian	British	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	20.5	14.1	13.5
Cigars, pipes	35.5	20.7	17.4

Conclusion: urge the cigar and pipe smokers to give up smoking and if they lack the strength of will to do so, they should switch to cigarettes.

Stratification - Death Rates of Male Smokers

- Are there other variables in which the three groups of smokers may differ, that are
  - related to the probability of dying; and
  - clearly not themselves affected by smoking habits?
- The regression of probability of dying on age for men over 40 is a concave upwards curve, the slope rising more and more steeply as age advances.



- The mean ages for each group in the previous table are as follows.

Smoking group	Canadian	British	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes only	50.5	49.8	53.2
Cigars, pipes	65.9	55.7	59.7

Stratification

- The table shows the adjusted death rates obtained when the age distributions were divided into 9 subclasses.
- The results are similar for different numbers of subclasses.

Adjusted for Age: Age-Adjusted Mortality

Smoking group	Canadian	British	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	29.5	14.8	21.2
Cigars, pipes	19.8	11.0	13.7

Comparing to the unadjusted death rates

Smoking group	Canadian	British	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes only	20.5	14.1	13.5
Cigars, pipes	35.5	20.7	17.4

Cochran (1968) showed that creating 5 or more strata removes 90% of the bias due to the stratifying variable.

Propensity score sub-classification/stratification - Smoking Cessation Study

```
prop.model <- glm(qsmk~as.factor(sex) + as.factor(race) + age +
              as.factor(education.code) + smokeintensity + smokeyrs
              + as.factor(exercise)+ as.factor(active) + wt71,
              data = nhefshwdat)

p.qsmk.obs <- predict(prop.model, type = "response")
```

SAME

```
##split the data into 5 strata
```

*

```
strat <- quantile(p.qsmk.obs,probs = c(.2,.4,.6,.8))
strat1 <- p.qsmk.obs<=strat[1]
```

Lowest PS to Quit

```
strat2 <- p.qsmk.obs > strat[1] & p.qsmk.obs <= strat[2]
strat3 <- p.qsmk.obs > strat[2] & p.qsmk.obs <= strat[3]
strat4 <- p.qsmk.obs > strat[3] & p.qsmk.obs <= strat[4]
```

Quintiles

```
strat5 <- p.qsmk.obs > strat[4]
```

Top PS to Quit

```
propmodel1 <- glm(wt82_71[strat1]~qsmk[strat1],data=nhefshwdat)
propmodel2 <- glm(wt82_71[strat2]~qsmk[strat2], data=nhefshwdat)
propmodel3 <- glm(wt82_71[strat3]~qsmk[strat3], data=nhefshwdat)
propmodel4 <- glm(wt82_71[strat4]~qsmk[strat4], data=nhefshwdat)
propmodel5 <- glm(wt82_71[strat5]~qsmk[strat5], data=nhefshwdat)
```

Propensity score sub-classification/stratification - Smoking Cessation Study

```
summary(propmodel1)$coef
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept)  3.586608  0.4441225 8.075718 1.476001e-14
qsmk[strat1] 1.568142  1.2143475 1.291345 1.975400e-01
summary(propmodel2)$coef
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept)  2.925217  0.4553337 6.424337 4.961353e-10
qsmk[strat2] 4.425085  1.0577621 4.183441 3.739459e-05
summary(propmodel3)$coef
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept)  1.832996  0.5421674 3.380868 0.0008147841
qsmk[strat3] 4.066539  1.0724093 3.791965 0.0001795346
summary(propmodel4)$coef
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept)  1.077701  0.5108151 2.109767 3.567709e-02
qsmk[strat4] 4.028424  0.9321204 4.321785 2.085639e-05
summary(propmodel5)$coef
               Estimate Std. Error    t value   Pr(>|t|)
(Intercept)  -0.4164234  0.5789714 -0.7192469 0.47252889
qsmk[strat5]  2.1904643  0.9018502  2.4288559 0.01571297
```

Propensity score sub-classification/stratification - Smoking Cessation Study

In summary the 5 quintiles produced treatment effects

Estimate	S.E.	P-value	PS Quintile
1.57	1.22	0.199	1
5.05	1.03	0.00	2
3.73	1.05	0.00	3
3.87	0.95	0.00	4
2.06	0.92	0.03	5

pool

- The overall treatment effect is 3.26, which can be obtained by averaging the estimates within each stratum.
- The overall treatment effect is a larger estimate compared to the treatment effect obtained by matching. using Everybody
- The treatment effect can also be estimated by fitting a linear regression model for the change in weight on the treatment variable and the quintiles of the estimated propensity score.

Propensity score sub-classification/stratification - Smoking Cessation Study

```
attach(nhefshwdat)
#create a variable to describe subclass to include in the model
stratvar <- numeric(length(qsmk))
for(i in 1:length(qsmk)){
  if(strat1[i]==T){stratvar[i] <- 1}
  if(strat2[i]==T){stratvar[i] <- 2}
  if(strat3[i]==T){stratvar[i] <- 3}
  if(strat4[i]==T){stratvar[i] <- 4}
  if(strat5[i]==T){stratvar[i] <- 5} }
stratmodel <- glm(wt82_71~qsmk+as.factor(stratvar),data=nhefshwdat)
```

Propensity score sub-classification/stratification - Smoking Cessation Study

```
summary(stratmodel)$coef
                       Estimate Std. Error    t value     Pr(>|t|)
(Intercept)           3.3551597  0.4370777  7.6763460 2.869213e-14
qsmk                  3.2984961  0.4544839  7.2576745 6.184322e-13
as.factor(stratvar)2 -0.2211815  0.6130499 -0.3607887 7.183063e-01
as.factor(stratvar)3 -1.3258585  0.6150994 -2.1555191 3.127378e-02
as.factor(stratvar)4 -2.0582474  0.6172615 -3.3344821 8.746234e-04
as.factor(stratvar)5 -4.2282480  0.6255308 -6.7594564 1.950785e-11
confint(stratmodel)
                         2.5 %     97.5 %
(Intercept)           2.498503  4.2118163
qsmk                  2.407724  4.1892681
as.factor(stratvar)2 -1.422737  0.9803742
as.factor(stratvar)3 -2.531431 -0.1202858
as.factor(stratvar)4 -3.268058 -0.8484371
as.factor(stratvar)5 -5.454266 -3.0022302
```

The linear regression yields the same treatment effect as averaging the estimates, but also provides an estimate of standard error, p-value, and confidence interval for the treatment effect.

Propensity score sub-classification/stratification - Smoking Cessation Study

✱ IMPORTANT ✱

We can investigate covariate balance within subclasses. In practice this should occur prior to looking at the outcome data. The number of subjects and average propensity score (shown in brackets) within each treatment group by subclass is shown in the table below.

Subclass	Smoking Cessation	No smoking cessation
1 lowest PS	42 (0.14)	272 (0.12)
2	59 (0.2)	254 (0.19)
3	82 (0.24)	231 (0.24)
4	92 (0.31)	221 (0.3)
5	128 (0.43)	185 (0.41)

For example, the percentage of males in each subclass are:

Subclass	Smoking Cessation	No Smoking Cessation
1	28.57%	22.79%
2	44.07%	43.31%
3	54.88%	46.32%
4	55.43%	59.73%
5	67.19%	70.81%

Abs diff

W/ matches?

Multivariate adjustment using the propensity score

- Another method for using the propensity score to adjust for bias is to use the propensity score itself as a predictor along with the treatment indicator. Continuous variable
- The treatment effect is adjusted by the propensity score.

```
prop.model.adj <- glm(wt82_71 ~ qsmk + p.qsmk.obs, data = nhefshwdat)
summary(prop.model.adj)$coef
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   5.575121  0.5157796 10.809115 2.586872e-26
qsmk          3.381171  0.4559755  7.415247 1.980993e-13
p.qsmk.obs  -14.793234  1.9128506 -7.733606 1.860811e-14
confint(prop.model.adj)
                 2.5 %     97.5 %
(Intercept)   4.564211   6.586030
qsmk          2.487476   4.274867
p.qsmk.obs  -18.542352 -11.044116
```

Similar conclusion as Before

The treatment effect is similar to the stratification method.

Comparing the three methods

The three propensity score methods yield similar results for the treatment effect.

highest internal validity
But not all Matched

Method	Average Treatment Effect	95% Confidence Interval
Matched	2.93	1.8-4.0 → widest ↓N
Stratified	3.26	1.7 - 3.4
Regression	3.40	2.5 - 4.3
Unadjusted	2.54 BIAS	1.7 - 3.4

same
concl
signif

The unadjusted analysis (two-sample t-test) underestimates the treatment effect by approximately 1kg.

Summary

Balancing score

causal inference

↓

- Use the propensity score in three different ways to calculate the effect of treatment in an observational study.
- Three methods: matching, stratification, direct regression adjustment.

  Most common ⇒ RCT
- Check that covariates are actually balanced using propensity methods. Not Balanced? Start over
- If covariates were not balanced using a method, then the treatment difference might be biased.

* PS DOES NOT Balance UNOBS(?) covars

Take-home Message

- Summary in previous page
- Run and understand the R code in the notes
- Work on the exercises and derivations in the lecture notes
- Work on the exercises in the online notes.
- Online notes: `http://utstat.toronto.edu/~nathan/teaching/STA305/classnotes/week5/sta305classnotes-week5.html`

See you next Tuesday!