STA305/1004 - Class 11

October 17, 2019

- Ignorable treatment assignment and the propensity score
- Three methods that use the propensity score to reduce bias: matching; stratification; and regression adjustment

- The propensity score is

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

where $\mathbf{x}$ are observed covariates.
- The propensity score is often estimated using logistic regression.
- The propensity score can be used to balance the distribution of covariates in two groups:

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

Propensity Scores and ignorable treatment assignment

1. If treatment assignment is ignorable then $T \perp (Y(0), Y(1)) | \mathbf{x}$.

The propensity score $e(\mathbf{x}) = P(T = 1|\mathbf{x})$ has the following property:

2. If the treatment assignment is ignorable then $T \perp (Y(0), Y(1)) | e(\mathbf{x})$.

If 1. and 2. are true then

$$P(T_i = 1 | Y(0), Y(1), \mathbf{x_i}) = P(T_i = 1 | \mathbf{x_i}) = P(T_i = 1 | e(\mathbf{x_i})).$$

Propensity Score and Ignorable Treatment Assignment

- In randomized studies the treatment assignment is ignorable (why?).
- In observational studies the treatment assignment is usually non-ignorable.

Propensity scores and ignorable treatment assignment

- The propensity score can be used in place of many covariates.
- If treatment assignment is ignorable then propensity score methods will produce unbiased results of the treatment effects.
- In the smoking cessation study what does it mean for treatment assignment to be ignorable?
- The potential outcomes for weight gain in the smoking cessation (treated) and smoking (control) groups are independent conditional on the propensity score.
- The treatment assignment mechanism has been reconstructed using 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.

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?

🖥 Respond at PollEv.com/nathantaback
📱 Text **NATHANTABACK** to 37607 once to join, then **A or B**

True	**A**
False	**B**

Total Results: 0

Using the propensity score to reduce bias

The three most common techniques that use the propensity score are

1. matching,
2. stratification (also called subclassification)
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.

- Educators are very interested in studying the effect of class size on learning.
- Does smaller class size cause students to achieve higher math and verbal scores?
- Angrist and Lavy (1999) published an unusual study of the effects of class size on academic achievement.
- Causal effects of class size on pupil achievement is difficult to measure. The twelfth century Rabbinic scholar Maimonides interpreted the the Talmud's discussion of class size as:
- "Twenty-five children may be put in charge of one teacher. If the number in the class exceeds twenty-five but is not more than forty, he should have an assistant to help with instruction. If there are more than forty, two teachers must be appointed."

Maimonides' Rule

- Since 1969 the rule has been used to determine class size in Israeli public schools.
- Class size is usually determined by other factors such as wealth of a community, special needs of students, etc.
- If adherence to Maimonides' rule were perfectly rigid, then what would separate a school with a single class of size 40 from the same school with two classes whose average size is 20.5 is the enrollment of a single student.

Number of children in grade 5	40	80	120
Class size with one extra student	20.5	27	30.25

Maimonides' Rule

- Angrist and Lavy matched schools where the number of grade 5 students are 31-40 to schools where the number of grade 5 students are 41-50.
- 86 matched pairs of two schools were formed, matching to minimize to total absolute difference in percentage disadvantaged.
- It's plausible that whether or not a few more students enrol in the fifth grade is a haphazard event.
- This is an example of natural experiment where students were haphazardly (randomly) assigned to small or large grade 5 classes.
- It was haphazard because it depended only on the number of grade 5 children at a school.

Maimonides' Rule
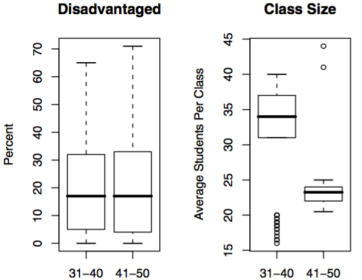### From Rosenbaum, 2010, pg.9

**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,
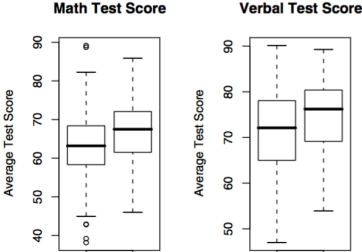
Propensity score matching

- 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.
- Randomly select a treated subject.
- Match to a control subject with closest propensity score (within some limit or "calipers").
- Eliminate both units from the pool of subjects until there is no acceptable match.
- It's not always possible to match every unit treated to a unit that is not treated.

Propensity score matching - the model

```
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)
```

Propensity score matching - the matching via `Match`

```
X <- prop.model$fitted
Y <- nhefshwdat$wt82_71
Tr <- nhefshwdat$qsmk
rr <- Matching::Match(Y = Y,Tr = Tr,X = X,M = 1)
summary(rr)
```

```
##
## Estimate... 2.9342
## AI SE...... 0.5838
## T-stat..... 5.026
## p.val...... 5.0087e-07
##
## Original number of observations.............. 1566
## Original number of treated obs............... 403
## Matched number of observations............... 403
## Matched number of observations  (unweighted). 1009
```

```
cat("Treatment Diff SE:", round(rr$se.standard,2))
```

```
## Treatment Diff SE: 0.56
```

Propensity score matching

After matching on covariates the treatment effect (difference in weight gain between the group that stopped smoking and the group that did not stop smoking) is 2.93 with a p-value of 0 (5.0087e-07) and 95% confidence interval (1.84, 4.02).

Propensity score matching -check covariate balance

Now, let's check covariate balance.

```
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)
```

Propensity score matching -check covariate balance

```
***** (V1) as.factor(sex)1 *****
                    Before Matching      After Matching
mean treatment........    0.45409           0.45409
mean control..........    0.53396           0.45331
std mean diff.........    -16.022           0.15703

NB: some output ommitted

***** (V3) age *****
                    Before Matching      After Matching
mean treatment........    46.174            46.174
mean control..........    42.788            46.595
std mean diff.........    27.714           -3.4504

NB: some output ommitted
```

Sex has an absolute standardized difference of 16 before matching and 0.16 after matching, and the absolute standardized difference of age has shifted from 27.71 to -3.45.

Propensity score matching -check covariate balance

How does this compare to not adjusting for imbalance?

```
#Unadjusted t-test
t.test(nhefshwdat$wt82_71[as.factor(nhefshwdat$qsmk)==1],
        nhefshwdat$wt82_71[as.factor(nhefshwdat$qsmk)==0],
        var.equal=T)
```

```
    Two Sample t-test

data:  nhefshwdat$wt82_71[as.factor(nhefshwdat$qsmk) == 1] and nhefshwd
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
```

The unadjusted treatment effect is 2.54 with a p-value of 0. So, both analyses
lead to the same conclusion that stopping to smoke leads to a significant weight
gain. Although the weight gain in the matched propensity score analysis is
0.39Kg higher.

Propensity score subclassification/stratification

Propensity scores permit subclassification 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 per cent 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 per cent of the bias in each of these covariates.

Stratification

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.

Are there other variables in which the three groups of smokers may differ, that (i) are related to the probability of dying; and (ii) are 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.

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

Compare 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 subclassification/stratification

```r
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)
p.qsmk.obs <- predict(prop.model, type = "response")
strat <- quantile(p.qsmk.obs,probs = c(.2,.4,.6,.8))
```

First Quintile

```r
strat1 <- p.qsmk.obs<=strat[1]
propmodel1 <- glm(wt82_71[strat1]~qsmk[strat1],data=nhefshwdat)
summary(propmodel1)
```

Second Quintile

```r
strat2 <- p.qsmk.obs > strat[1] & p.qsmk.obs <= strat[2]
propmodel2 <- glm(wt82_71[strat2]~qsmk[strat2], data=nhefshwdat)
summary(propmodel2)
```

Propensity score subclassification/stratification

Third Quintile

```
strat3 <- p.qsmk.obs > strat[2] & p.qsmk.obs <= strat[3]
propmodel3 <- glm(wt82_71[strat3]~qsmk[strat3], data=nhefshwdat)
summary(propmodel3)
```

Fourth Quintile

```
strat4 <- p.qsmk.obs > strat[3] & p.qsmk.obs <= strat[4]
propmodel4 <- glm(wt82_71[strat4]~qsmk[strat4], data=nhefshwdat)
summary(propmodel4)
```

Fifth Quintile

```
strat5 <- p.qsmk.obs > strat[4]
propmodel5 <- glm(wt82_71[strat5]~qsmk[strat5], data=nhefshwdat)
summary(propmodel5)
```

Propensity score subclassification/stratification

```
              Estimate Std. Error  t value      Pr(>|t|)
  (Intercept) 3.582883  0.4463651 8.026799 2.055049e-14
  qsmk[strat1] 1.571867  1.2204794 1.287909 1.987319e-01

              Estimate Std. Error  t value      Pr(>|t|)
  (Intercept) 2.700017  0.4466046 6.045654 4.258361e-09
  qsmk[strat2] 5.054241  1.0286540 4.913451 1.449627e-06

              Estimate Std. Error  t value      Pr(>|t|)
  (Intercept) 2.121351  0.5384292 3.939888 0.0001007005
  qsmk[strat3] 3.726930  1.0519470 3.542888 0.0004564504

               Estimate Std. Error  t value      Pr(>|t|)
  (Intercept) 0.9552357  0.5130865 1.861744 6.358234e-02
  qsmk[strat4] 3.8711676  0.9463872 4.090469 5.488916e-05

               Estimate Std. Error   t value    Pr(>|t|)
  (Intercept) -0.2892809  0.5878199 -0.4921251 0.62297810
  qsmk[strat5]  2.0550465  0.9192030  2.2356829 0.02608187
```

In summary the 5 quintiles produced treatment effects

Estimate (se)	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

The overall treatment effect is 3.26, which can be obtained by averaging the estimates within each stratum. This is a larger estimate compared to the treatment effect obtained by matching. The treatment effect and 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 subclassification/stratification

```r
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}
else
  if (strat2[i]==T) {stratvar[i] <- 2}
else
  if (strat3[i]==T) {stratvar[i] <- 3}
else
  if (strat4[i]==T) {stratvar[i] <- 4}
else stratvar[i] <- 5
}
stratmodel <- glm(wt82_71~qsmk+as.factor(stratvar),data=nhefshwdat)
summary(stratmodel)$coef
```

Propensity score subclassification/stratification

```
                     Estimate Std. Error   t value     Pr(>|t|)
(Intercept)         3.3564796  0.4373149 7.6752008 2.894022e-14
qsmk                3.2645028  0.4542610 7.1864036 1.027063e-12
as.factor(stratvar)2 -0.3191003 0.6134512 -0.5201723 6.030173e-01
as.factor(stratvar)3 -1.1139815 0.6157083 -1.8092683 7.060174e-02
as.factor(stratvar)4 -2.2229271 0.6172504 -3.6013378 3.265030e-04
as.factor(stratvar)5 -4.1403625 0.6255644 -6.6186033 4.971650e-11

                       2.5 %     97.5 %
(Intercept)         2.499358  4.21360105
qsmk                2.374168  4.15483802
as.factor(stratvar)2 -1.521443  0.88324196
as.factor(stratvar)3 -2.320748  0.09278456
as.factor(stratvar)4 -3.432716 -1.01313860
as.factor(stratvar)5 -5.366446 -2.91427883
```

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 subclassification/stratification

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	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%

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.

- 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)  5.560244  0.5090376 10.923052 8.104078e-27
qsmk         3.397278  0.4559641  7.450757 1.528381e-13
p.qsmk.obs -14.751531  1.8846521 -7.827190 9.128614e-15
```

```
confint(prop.model.adj)
```

```
                2.5 %     97.5 %
(Intercept)  4.562548   6.557939
qsmk         2.503604   4.290951
p.qsmk.obs -18.445381 -11.057680
```

The treatment effect is similar to the stratification method.

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

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

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