STA305/1004 Class Notes - Week 5

Nathan Taback

January 31, 2016

The design phase of an observational study

Good observational studies are designed. According to Rubin (2007)

> 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 propensity scores, whose values we will try to reconstruct.

Rubin (2007) also discusses the importance of a design phase of observational studies before seeing outcome data.

> 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. Propensity score methods are the observational study equivalent of complete (i.e., unrestricted) randomization in a randomized experiment. That is, these methods are intended to eliminate bias, but are not intended to increase precision. Of course, propensity score methods can only perfectly eliminate bias when the assignment mechanism is truly unconfounded, given the observed covariates, X, and when the propensity scores are effectively known, whereas randomization eliminates bias due to all covariates, both observed and unobserved. … no outcome data from the study are in sight when objectively designing either a randomized experiment or an observational study.

- The main part of the design stage is to assess the degree of balance in the covariate distributions between treated and control units, which involves comparing the distributions of covariates in the treated and control samples.
- The difference in average covariate values by treatment status, scaled by their sample standard deviation provides a scale-free way to assess the differences.
- 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)

Epidemiologic Follow-up Study

The NHEFS survey was designed to investigate the relationships between clinical, nutritional, and behavioural factors assessed in the first National Health and Nutrition Examination Survey NHANES I and subsequent morbidity, mortality, and hospital utilization, as well as changes in risk factors, functional limitation, and institutionalization. For more information see the survey website (http://www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm).

Individuals were classified as treated if they reported, being smokers at baseline in 1971-75, and having quit smoking in the 1982 survey. The latter implies that the individuals included in our study did not die and were not otherwise lost to follow-up between baseline and 1982 (otherwise they would not have been able to respond to the survey). That is, we selected individuals into our study conditional on an event (responding to the 1982 survey) that occurred after the start of smoking cessation. If smoking cessation affects the probability of selection into the study, we might have selection bias Hernan, Robins,2014 (https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2015/11/hernanrobins_v2.17.12.pdf).

The outcome in this study is weight change from 1981 to 1971 `wt82_71`. The covariates in this study are shown in the table below for each treatment group.

	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

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

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

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

The i^{th} propensity score is the probability that a unit receives treatment given all the information, recorded as covariates, that is observed before the treatment.

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 counfounding covariates.

Examples

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

Each unit’s propensity score is 0.5.

2. Consider a completely randomized design with $n = 8$ units and three units are assigned treatment. Each unit has a 3/8 chance of receiving treatment (and 5/8 of receiving control). Thus, each person’s propensity score is 3/8. The probability of an particular treatment assignment is $\frac{1}{\binom{8}{3}} = \frac{1}{56}$.

Remember that the overall treatment assignment is the collection of all 8 units’ treatment assignments. These can be generated using the R code below.

```
library(combinat)
i <- combn(1:8,3)
colnames(i) <- (nth <- paste0(1:56, c("st Trt Assig", "nd Trt Assig", "rd Trt Assig", rep("th Trt Assig", 53))))
i
```

```
     1st Trt Assig 2nd Trt Assig 3rd Trt Assig 4th Trt Assig 5th Trt Assig
[1,]             1             1             1             1             1
[2,]             2             2             2             2             2
[3,]             3             4             5             6             7
     6th Trt Assig 7th Trt Assig 8th Trt Assig 9th Trt Assig
[1,]             1             1             1             1
[2,]             2             3             3             3
[3,]             8             4             5             6
     10th Trt Assig 11th Trt Assig 12th Trt Assig 13th Trt Assig
[1,]              1              1              1              1
[2,]              3              3              4              4
[3,]              7              8              5              6
     14th Trt Assig 15th Trt Assig 16th Trt Assig 17th Trt Assig
[1,]              1              1              1              1
[2,]              4              4              5              5
[3,]              7              8              6              7
     18th Trt Assig 19th Trt Assig 20th Trt Assig 21th Trt Assig
[1,]              1              1              1              1
[2,]              5              6              6              7
[3,]              8              7              8              8
     22th Trt Assig 23th Trt Assig 24th Trt Assig 25th Trt Assig
[1,]              2              2              2              2
[2,]              3              3              3              3
[3,]              4              5              6              7
     26th Trt Assig 27th Trt Assig 28th Trt Assig 29th Trt Assig
[1,]              2              2              2              2
[2,]              3              4              4              4
[3,]              8              5              6              7
     30th Trt Assig 31th Trt Assig 32th Trt Assig 33th Trt Assig
[1,]              2              2              2              2
[2,]              4              5              5              5
[3,]              8              6              7              8
     34th Trt Assig 35th Trt Assig 36th Trt Assig 37th Trt Assig
[1,]              2              2              2              3
[2,]              6              6              7              4
[3,]              7              8              8              5
     38th Trt Assig 39th Trt Assig 40th Trt Assig 41th Trt Assig
[1,]              3              3              3              3
[2,]              4              4              4              5
[3,]              6              7              8              6
     42th Trt Assig 43th Trt Assig 44th Trt Assig 45th Trt Assig
[1,]              3              3              3              3
[2,]              5              5              6              6
[3,]              7              8              7              8
     46th Trt Assig 47th Trt Assig 48th Trt Assig 49th Trt Assig
[1,]              3              4              4              4
[2,]              7              5              5              5
[3,]              8              6              7              8
```

```
     50th Trt Assig 51th Trt Assig 52th Trt Assig 53th Trt Assig
[1,]              4              4              4              5
[2,]              6              6              7              6
[3,]              7              8              8              7
     54th Trt Assig 55th Trt Assig 56th Trt Assig
[1,]              5              5              6
[2,]              6              7              7
[3,]              8              8              8
```

Each column corresponds to the units that will be treated; so the units that will not be treated are not included in the column. For example in the first treatment assignment units 1, 2, 3 will be treated and units 4, 5, 6, 7, 8 will be given control. In the second treatment assignment units 1, 2, 4 will be treated and units 3, 5, 6, 7, 8 will be given control.

3. Consider a completely randomized design with n units and m units are assigned treatment. Each unit has probability $\frac{m}{n}$ of receiving treatment (and $1-\frac{m}{n}$ of receiving control). Thus, each person's propensity score is m/n. The probability of an particular treatment assignment is $\frac{1}{\binom{n}{m}}$.

4. 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). 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.

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

where $p_i=P(T_i=1)$. The predicted probabilities from the above equation are estimates of the propensity score for each patient.

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

5. The propensity score for each NHEFS subject can be estimated by fitting a logistic regression 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)

#Summary of propensity score model
summary(prop.model)
```

```

Call:
glm(formula = 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)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.4137  -0.8032  -0.6456   1.0843   2.2966  

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                -2.401228   0.484016  -4.961 7.01e-07 ***
as.factor(sex)1            -0.499080   0.146531  -3.406 0.000659 ***
as.factor(race)1           -0.778223   0.207032  -3.759 0.000171 ***
age                         0.046207   0.009889   4.672 2.98e-06 ***
as.factor(education.code)2 -0.065716   0.196123  -0.335 0.737567    
as.factor(education.code)3  0.052635   0.175523   0.300 0.764274    
as.factor(education.code)4  0.108653   0.269191   0.404 0.686486    
as.factor(education.code)5  0.466165   0.224106   2.080 0.037516 *  
smokeintensity             -0.026527   0.005664  -4.683 2.82e-06 ***
smokeyrs                   -0.028492   0.010009  -2.847 0.004417 ** 
as.factor(exercise)1        0.359557   0.178603   2.013 0.044098 *  
as.factor(exercise)2        0.422772   0.185657   2.277 0.022776 *  
as.factor(active)1          0.044928   0.131555   0.342 0.732717    
as.factor(active)2          0.158151   0.213435   0.741 0.458708    
wt71                        0.006099   0.004368   1.396 0.162630    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1786.1  on 1565  degrees of freedom
Residual deviance: 1694.5  on 1551  degrees of freedom
AIC: 1724.5

Number of Fisher Scoring iterations: 4
```

The propensity score for each subject is $\hat{p}_i$ for smokers and $1 - \hat{p}_i$, where $\hat{p}_i$ is the predicted probability in from the logistic regression. The predicted probabilities for the first 20 subjects are:

```
#Propensity scores for each subject
p.qsmk.obs <- ifelse(nhefshwdat$qsmk == 0,
                     1 - predict(prop.model, type = "response"),
                     predict(prop.model, type = "response"))
dat <- data.frame(1:20,nhefshwdat$qsmk[1:20], p.qsmk.obs[1:20])
colnames(dat) <- c("Subject","Quit Smoking", "Estimated Propensity Score")
knitr::kable(dat)
```

Subject	Quit Smoking	Estimated Propensity Score
1	0	0.8760965
2	0	0.8402695
3	0	0.8400642
4	0	0.6893079
5	0	0.6802405
6	0	0.8337755
7	0	0.7609109
8	0	0.7380972
9	0	0.7008263
10	0	0.7058756
11	1	0.2598566
12	0	0.8123821
13	0	0.6533319
14	0	0.8228333
15	1	0.2722928
16	0	0.7745581
17	0	0.6580986
18	1	0.3197956
19	0	0.7274018
20	0	0.7670320

Subject 1’s estimated probability of not quitting smoking (1-propensity score) is 0.88 (so the estimated probability of quitting smoking is 0.12) and subject 11’s estimated probability of quitting smoking (propensity score) is 0.26 (so the estimated probability of not quitting smoking is 0.74).

Imbalance versus lack of Complete

Overlap

In a study comparing two treatments (which we typically label "treatment" and "control"), causal inferences are cleanest if the units receiving the treatment are comparable to those receiving the control.

Suppose that treatment assignment is ignorable. There are two major ways in whcih the treatment and control groups may not be comparable, imbalance and lack of complete overlap.

Imbalance occurs if the distributions of relevant pre-treatment variables differ for the treatment and control groups.

Lack of complete overlap occurs if there are values of pre-treatment variables where there are treated units but no controls, or controls but no treated units. Lack of complete overlap creates problems because it means that there are treatment observations for which we have no counterfactuals (that is, control observations with the same covariate distribution) and vice versa. When treatment and control groups do not completely overlap, the data are inherently limited in what they can tell us about treatment effects in the regions of nonoverlap. No amount of adjustment can create direct treatment/control comparisons, and one must either restrict inferences to the region of overlap, or rely on a model to extrapolate outside this region. (Gelman and Hill, 2007)

Imbalance and lack of complete overlap are issues for causal inference largely because they force us to rely more heavily on model specification and less on direct support from the data.

When treatment and control groups are unbalanced, the simple comparison of group averages, $\bar{y}_1 - \bar{y}_0$, is usually not a good estimate of the average treatment effect. Although, it's possible to adjust for pre-treatment differences between the groups.

Lack of complete overlap is a more serious problem than imbalance. But similar statistical methods are used in both scenarios, so we discuss these problems together here.

The balancing property of the propensity score

The balancing property of the propsensity 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, $\mathbf{x}$,

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

or

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

This means that treatment is independent of the observed covariates conditional on 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 propesnity 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.
- If many pairs are formed way then the the distribution of the observed covariates will look about the same in the treated and control groups, even though individuals in matched pairs will typically have different values of x. Although it is difficult to match on 20 covariates at once, it is easy to match on one covariate, the propensity score $e(\mathbf{x})$, and matching on $e(\mathbf{x})$ will tend to balance all 20 covariates.

How can the degree of balance in the covariate distributions between treated and control units be assessed?

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)).

If $\bar{x}_t$, s_t^2 are the mean and variance of a covariate in the treated group and $\bar{x}_c$, s_c^2 are the mean and variance of a covariate in the control group then the pooled vaiance is

$$\sqrt{\frac{s_t^2 + s_c^2}{2}}.$$

The absolute pooled standardized difference is,

$$\frac{100 \times |\bar{x}_t - \bar{x}_c|}{\sqrt{\frac{s_t^2+s_c^2}{2}}}.$$

Example from the NHEFS

The following table shows the distribution of covariates (age, sex, race, etc.) in each treatment group.

	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

There are more men in the stop smoking group (A=1) compared to the smoking group (A=0) (55% vs. 47%). In addition, there are more white people, university graduates, and years of smoking in the group that stopped smoking.

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

```
library(Matching)
```

```
Loading required package: MASS
```

```
##
##  Matching (Version 4.9-2, Build Date: 2015-12-25)
##  See http://sekhon.berkeley.edu/matching for additional documentation.
##  Please cite software as:
##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
##   Software with Automated Balance Optimization: The Matching package for R.''
##   Journal of Statistical Software, 42(7): 1-52.
##
```

```
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)
```

```
***** (V1) as.factor(sex)1 *****
before matching:
mean treatment........ 0.45409
mean control.......... 0.53396
std mean diff......... -16.022

mean raw eQQ diff..... 0.079404
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.039935
med  eCDF diff........ 0.039935
max  eCDF diff........ 0.07987

var ratio (Tr/Co)..... 0.99779
T-test p-value........ 0.0057371


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

mean raw eQQ diff..... 0.057072
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.028422
med  eCDF diff........ 0.028422
max  eCDF diff........ 0.056844

var ratio (Tr/Co)..... 0.65287
T-test p-value........ 0.0012863


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

mean raw eQQ diff..... 3.3921
med  raw eQQ diff..... 4
max  raw eQQ diff..... 5

mean eCDF diff........ 0.068985
```

```
med  eCDF diff........ 0.074988
max  eCDF diff........ 0.12956

var ratio (Tr/Co)..... 1.0731
T-test p-value........ 1.6316e-06
KS Bootstrap p-value.. < 2.22e-16
KS Naive p-value...... 8.6584e-05
KS Statistic.......... 0.12956


***** (V4) as.factor(education.code)2 *****
before matching:
mean treatment........ 0.18362
mean control.......... 0.22872
std mean diff......... -11.633

mean raw eQQ diff..... 0.044665
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.022548
med  eCDF diff........ 0.022548
max  eCDF diff........ 0.045096

var ratio (Tr/Co)..... 0.85115
T-test p-value........ 0.049355


***** (V5) as.factor(education.code)3 *****
before matching:
mean treatment........ 0.38958
mean control.......... 0.41273
std mean diff......... -4.7408

mean raw eQQ diff..... 0.022333
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.011574
med  eCDF diff........ 0.011574
max  eCDF diff........ 0.023148

var ratio (Tr/Co)..... 0.98271
T-test p-value........ 0.41346


***** (V6) as.factor(education.code)4 *****
before matching:
mean treatment........ 0.07196
mean control.......... 0.079106
std mean diff......... -2.7616

mean raw eQQ diff..... 0.0074442
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.0035727
med  eCDF diff........ 0.0035727
max  eCDF diff........ 0.0071455

var ratio (Tr/Co)..... 0.91822
T-test p-value........ 0.6368


***** (V7) as.factor(education.code)5 *****
before matching:
mean treatment........ 0.15385
mean control.......... 0.098882
std mean diff......... 15.215

mean raw eQQ diff..... 0.054591
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.027482
med  eCDF diff........ 0.027482
max  eCDF diff........ 0.054964

var ratio (Tr/Co)..... 1.4633
T-test p-value........ 0.0062041


***** (V8) smokeintensity *****
before matching:
mean treatment........ 18.603
mean control.......... 21.192
std mean diff......... -20.874

mean raw eQQ diff..... 2.6849
med  raw eQQ diff..... 2
max  raw eQQ diff..... 20

mean eCDF diff........ 0.064175
med  eCDF diff........ 0.043336
max  eCDF diff........ 0.14366

var ratio (Tr/Co)..... 1.1679
T-test p-value........ 0.00025243
```

```
KS Bootstrap p-value.. < 2.22e-16
KS Naive p-value...... 8.6245e-06
KS Statistic.......... 0.14366


***** (V9) smokeyrs *****
before matching:
mean treatment........ 26.032
mean control.......... 24.088
std mean diff......... 15.26

mean raw eQQ diff..... 1.9752
med  raw eQQ diff..... 2
max  raw eQQ diff..... 5

mean eCDF diff........ 0.032783
med  eCDF diff........ 0.023244
max  eCDF diff........ 0.088511

var ratio (Tr/Co)..... 1.1846
T-test p-value........ 0.0072293
KS Bootstrap p-value.. < 2.22e-16
KS Naive p-value...... 0.018385
KS Statistic.......... 0.088511


***** (V10) as.factor(exercise)1 *****
before matching:
mean treatment........ 0.43672
mean control.......... 0.41702
std mean diff......... 3.9669

mean raw eQQ diff..... 0.019851
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.0098498
med  eCDF diff........ 0.0098498
max  eCDF diff........ 0.0197

var ratio (Tr/Co)..... 1.0135
T-test p-value........ 0.49202


***** (V11) as.factor(exercise)2 *****
before matching:
mean treatment........ 0.40695
mean control.......... 0.37919
std mean diff......... 5.6429
```

```
mean raw eQQ diff..... 0.027295
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.013878
med  eCDF diff........ 0.013878
max  eCDF diff........ 0.027756

var ratio (Tr/Co)..... 1.0269
T-test p-value........ 0.32766


***** (V12) as.factor(active)1 *****
before matching:
mean treatment........ 0.4665
mean control.......... 0.45314
std mean diff......... 2.6753

mean raw eQQ diff..... 0.014888
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.0066814
med  eCDF diff........ 0.0066814
max  eCDF diff........ 0.013363

var ratio (Tr/Co)..... 1.006
T-test p-value........ 0.64338


***** (V13) as.factor(active)2 *****
before matching:
mean treatment........ 0.11166
mean control.......... 0.089424
std mean diff......... 7.0522

mean raw eQQ diff..... 0.022333
med  raw eQQ diff..... 0
max  raw eQQ diff..... 1

mean eCDF diff........ 0.011119
med  eCDF diff........ 0.011119
max  eCDF diff........ 0.022239

var ratio (Tr/Co)..... 1.2202
T-test p-value........ 0.21198
```

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

mean raw eQQ diff..... 2.1872
med  raw eQQ diff..... 2.04
max  raw eQQ diff..... 14.75

mean eCDF diff........ 0.032352
med  eCDF diff........ 0.032386
max  eCDF diff........ 0.07

var ratio (Tr/Co)..... 1.0606
T-test p-value........ 0.022421
KS Bootstrap p-value.. 0.1
KS Naive p-value...... 0.10646
KS Statistic.......... 0.07


Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): age smokeintensity smokeyrs  Number(s): 3 8 9
```

If the absolute value of the standardized mean difference is greater than 10% then this indicates a serious imbalance. For example, sex has an absolute standardized mean difference of $|-16.022| = 16.022$ indicating serious imbalance between the groups in males and females.

Propensity scores and ignorable treatment assignment

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

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

or

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

- 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.

- Ignorable treatment assignment and the balancing property of the propensity score implies that (for a proof see Rosenbaum, 2010)

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

or

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

This means that the scaler propensity score $e(\mathbf{x})$ may be used in place of the many covariates in $\mathbf{x}$.

- The propsensity score can be used in place of many covariates.
- If treatment assignment is strongly 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 propesnity score.
- The treatment assignment mechanism has been reconstructed using the propensity score.
- 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 compareable.
- 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. (see Rosenbaum, 2010, pg. 76)

Using the propensity score to reduce bias

D'Agostino (1998) describes using the propensity score.

> In a randomized experiment, the randomization of units (that is, subjects) to different treat- ments guarantees that on average there should be no systematic differences in observed or unobserved covariates (that is, bias) between units assigned to the different treatments. However, in a non-randomized observational study, investigators have no control over the treatment assignment, and therefore direct comparisons of outcomes from the treatment groups may be misleading. This difficulty may be partially avoided if information on measured covariates is incorporated into the study design (for example, through matched sampling) or into estimation of the treatment effect (for example, through stratification or covariance adjustment) … Traditional methods of adjustment (matching, stratification and covariance adjustment) are often limited since they can only use a limited number of covariates for adjustment. However, propensity scores, which provide a scalar summary of the covariate information, do not have this limitation.

> Currently in observational studies, propensity scores are used primarily to reduce bias and increase precision. The three most common techniques that use the propensity score are matching, stratification (also called subclassification) and 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 scores are useful for these techniques because by definition the propensity score is the conditional probability of treatment given the observed covariates $e(\mathbf{x}) = P(T = 1|X)$, which implies that T and $\mathbf{x}$ are conditionally independent given $e(\mathbf{x})$. Thus, subjects in treatment and control groups with equal (or nearly equal) propensity scores will tend to have the same (or nearly the same) distributions on their background covariates. Exact adjustments made using the propensity score will, on average, remove all of the bias in the background covariates. Therefore bias-removing adjustments can be made using the propensity scores rather than all of the background covariates individually.

Matching - Maimonides' rule

- Educators are very intrested in studying the effect of class size on learning.
- Does smaller class size cause students to acheive 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."
- 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

- 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.

From Rosenbaum, 2010, pg.9

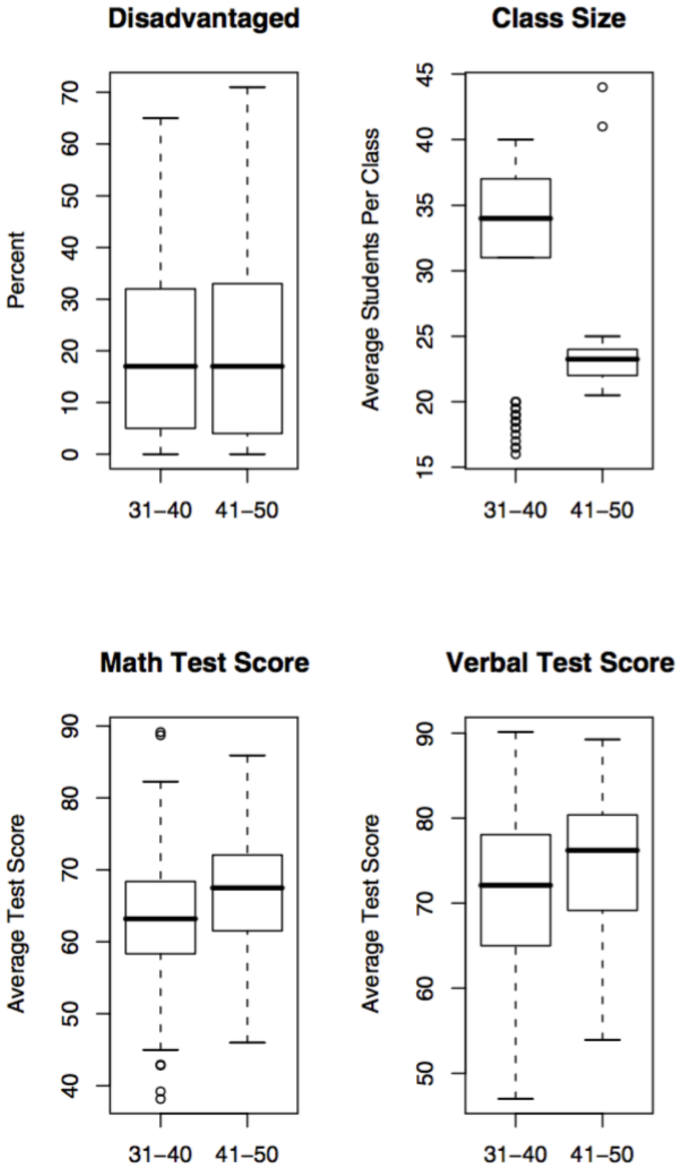

Fig. 1.1 Eighty-six pairs of two Israeli schools, one with between 31 and 40 students in the fifth grade, the other with between 41 and 50 students in the fifth grade, matched for percentage of students in the school classified as disadvantaged. The figure shows that the percentage of disadvantaged students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

Propensity score matching

- For each unit we have a propesnity score.
- Randomly select a treated subject.
- Match to a control subject with closest propesnity 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.

In R propensity score matching can be done using the `Match` function in the `Matching` library.

```
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)

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)
```

```

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
```

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).

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)
```

```
***** (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

mean raw eQQ diff.....   0.079404          0.0069376
med  raw eQQ diff.....          0                  0
max  raw eQQ diff.....          1                  1

mean eCDF diff........   0.039935          0.0034688
med  eCDF diff........   0.039935          0.0034688
max  eCDF diff........    0.07987          0.0069376

var ratio (Tr/Co).....    0.99779             1.0003
T-test p-value........  0.0057371            0.98136


***** (V2) as.factor(race)1 *****
                       Before Matching       After Matching
mean treatment........    0.08933            0.08933
mean control..........    0.14617           0.083561
std mean diff.........    -19.905             2.0202

mean raw eQQ diff.....   0.057072          0.0029732
med  raw eQQ diff.....          0                  0
max  raw eQQ diff.....          1                  1

mean eCDF diff........   0.028422          0.0014866
med  eCDF diff........   0.028422          0.0014866
max  eCDF diff........   0.056844          0.0029732

var ratio (Tr/Co).....    0.65287             1.0623
T-test p-value........  0.0012863            0.75212


***** (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

mean raw eQQ diff.....     3.3921            0.67294
med  raw eQQ diff.....          4                  1
max  raw eQQ diff.....          5                  2

mean eCDF diff........   0.068985           0.013693
```

```
med  eCDF diff........   0.074988          0.010902
max  eCDF diff........    0.12956          0.050545

var ratio (Tr/Co).....     1.0731           0.92406
T-test p-value........ 1.6316e-06           0.57566
KS Bootstrap p-value.. < 2.22e-16        < 2.22e-16
KS Naive p-value...... 8.6584e-05           0.15182
KS Statistic..........    0.12956          0.050545


***** (V4) as.factor(education.code)2 *****
                       Before Matching       After Matching
mean treatment........    0.18362           0.18362
mean control..........    0.22872           0.20084
std mean diff.........    -11.633           -4.4403

mean raw eQQ diff.....   0.044665          0.020813
med  raw eQQ diff.....          0                 0
max  raw eQQ diff.....          1                 1

mean eCDF diff........   0.022548          0.010406
med  eCDF diff........   0.022548          0.010406
max  eCDF diff........   0.045096          0.020813

var ratio (Tr/Co).....    0.85115           0.93399
T-test p-value........   0.049355           0.53095


***** (V5) as.factor(education.code)3 *****
                       Before Matching       After Matching
mean treatment........    0.38958           0.38958
mean control..........    0.41273           0.38093
std mean diff.........    -4.7408            1.7703

mean raw eQQ diff.....   0.022333          0.013875
med  raw eQQ diff.....          0                 0
max  raw eQQ diff.....          1                 1

mean eCDF diff........   0.011574         0.0069376
med  eCDF diff........   0.011574         0.0069376
max  eCDF diff........   0.023148          0.013875

var ratio (Tr/Co).....    0.98271            1.0084
T-test p-value........    0.41346           0.79553


***** (V6) as.factor(education.code)4 *****
                       Before Matching       After Matching
mean treatment........    0.07196           0.07196
mean control..........   0.079106          0.079115
std mean diff.........    -2.7616           -2.7652

mean raw eQQ diff.....  0.0074442         0.0059465
med  raw eQQ diff.....          0                 0
max  raw eQQ diff.....          1                 1

mean eCDF diff........  0.0035727         0.0029732
med  eCDF diff........  0.0035727         0.0029732
max  eCDF diff........  0.0071455         0.0059465

var ratio (Tr/Co).....    0.91822           0.91663
T-test p-value........     0.6368           0.68742


***** (V7) as.factor(education.code)5 *****
                       Before Matching       After Matching
mean treatment........    0.15385           0.15385
mean control..........   0.098882           0.15182
std mean diff.........     15.215           0.56014

mean raw eQQ diff.....   0.054591          0.019822
med  raw eQQ diff.....          0                 0
max  raw eQQ diff.....          1                 1

mean eCDF diff........   0.027482         0.0099108
med  eCDF diff........   0.027482         0.0099108
max  eCDF diff........   0.054964          0.019822

var ratio (Tr/Co).....     1.4633            1.0109
T-test p-value........  0.0062041           0.93208


***** (V8) smokeintensity *****
                       Before Matching       After Matching
mean treatment........     18.603            18.603
mean control..........     21.192             18.77
std mean diff.........    -20.874           -1.3479

mean raw eQQ diff.....     2.6849            1.4618
med  raw eQQ diff.....          2                 0
max  raw eQQ diff.....         20                20

mean eCDF diff........   0.064175          0.033126
med  eCDF diff........   0.043336          0.019822
max  eCDF diff........    0.14366          0.090188

var ratio (Tr/Co).....     1.1679            1.2535
T-test p-value........ 0.00025243           0.82363
```

```
KS Bootstrap p-value.. < 2.22e-16        < 2.22e-16 
KS Naive p-value...... 8.6245e-06         0.0005454 
KS Statistic..........    0.14366          0.090188 


***** (V9) smokeyrs *****
                       Before Matching       After Matching
mean treatment........     26.032            26.032 
mean control..........     24.088            26.437 
std mean diff.........      15.26            -3.176 

mean raw eQQ diff.....     1.9752            1.1655 
med  raw eQQ diff.....          2                 1 
max  raw eQQ diff.....          5                 6 

mean eCDF diff........   0.032783           0.01967 
med  eCDF diff........   0.023244          0.016848 
max  eCDF diff........   0.088511          0.050545 

var ratio (Tr/Co).....     1.1846            1.1132 
T-test p-value........  0.0072293           0.61403 
KS Bootstrap p-value..        0.1        < 2.22e-16 
KS Naive p-value......   0.018385           0.15182 
KS Statistic..........   0.088511          0.050545 


***** (V10) as.factor(exercise)1 *****
                       Before Matching       After Matching
mean treatment........    0.43672           0.43672 
mean control..........    0.41702           0.46081 
std mean diff.........     3.9669           -4.8493 

mean raw eQQ diff.....   0.019851          0.022795 
med  raw eQQ diff.....          0                 0 
max  raw eQQ diff.....          1                 1 

mean eCDF diff........  0.0098498          0.011397 
med  eCDF diff........  0.0098498          0.011397 
max  eCDF diff........     0.0197          0.022795 

var ratio (Tr/Co).....     1.0135           0.99007 
T-test p-value........    0.49202           0.49084 


***** (V11) as.factor(exercise)2 *****
                       Before Matching       After Matching
mean treatment........    0.40695           0.40695 
mean control..........    0.37919           0.36873 
std mean diff.........     5.6429            7.7689 
```

```
mean raw eQQ diff.....   0.027295          0.021804 
med  raw eQQ diff.....          0                 0 
max  raw eQQ diff.....          1                 1 

mean eCDF diff........   0.013878          0.010902 
med  eCDF diff........   0.013878          0.010902 
max  eCDF diff........   0.027756          0.021804 

var ratio (Tr/Co).....     1.0269            1.0368 
T-test p-value........    0.32766           0.25681 


***** (V12) as.factor(active)1 *****
                       Before Matching       After Matching
mean treatment........     0.4665            0.4665 
mean control..........    0.45314           0.46844 
std mean diff.........     2.6753          -0.38737 

mean raw eQQ diff.....   0.014888                 0 
med  raw eQQ diff.....          0                 0 
max  raw eQQ diff.....          1                 0 

mean eCDF diff........  0.0066814                 0 
med  eCDF diff........  0.0066814                 0 
max  eCDF diff........   0.013363                 0 

var ratio (Tr/Co).....      1.006           0.99949 
T-test p-value........    0.64338           0.95705 


***** (V13) as.factor(active)2 *****
                       Before Matching       After Matching
mean treatment........    0.11166           0.11166 
mean control..........   0.089424           0.09748 
std mean diff.........     7.0522            4.4974 

mean raw eQQ diff.....   0.022333          0.012884 
med  raw eQQ diff.....          0                 0 
max  raw eQQ diff.....          1                 1 

mean eCDF diff........   0.011119          0.006442 
med  eCDF diff........   0.011119          0.006442 
max  eCDF diff........   0.022239          0.012884 

var ratio (Tr/Co).....     1.2202            1.1275 
T-test p-value........    0.21198           0.51116 
```

```
***** (V14) wt71 *****
                       Before Matching       After Matching
mean treatment........     72.355            72.355
mean control..........     70.303            72.563
std mean diff.........      13.13           -1.3303

mean raw eQQ diff.....     2.1872            1.8433
med  raw eQQ diff.....       2.04              1.92
max  raw eQQ diff.....      14.75             14.75

mean eCDF diff........   0.032352          0.028802
med  eCDF diff........   0.032386          0.024777
max  eCDF diff........       0.07          0.078295

var ratio (Tr/Co).....     1.0606            1.0282
T-test p-value........   0.022421           0.84279
KS Bootstrap p-value.. < 2.22e-16        < 2.22e-16
KS Naive p-value......    0.10646         0.0041188
KS Statistic..........       0.07          0.078295


Before Matching Minimum p.value: < 2.22e-16
Variable Name(s): age smokeintensity wt71  Number(s): 3 8 14

After Matching Minimum p.value: < 2.22e-16
Variable Name(s): age smokeintensity smokeyrs wt71  Number(s): 3 8 9 14
```

The output shows the effectiveness of propesnity score matching in reducing imbalance. Sex has an absolute standardized difference of 16 before matching and 0.16 after matching, and the absolute standardized difference of race has shifted from 19.9 to 2.0.

How does this compare to not adjusting for imbalance?

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

```
##
##  Two Sample t-test
##
## data:  nhefshwdat$wt82_71[as.factor(nhefshwdat$qsmk) == 0] and nhefshwdat$wt8
2_71[as.factor(nhefshwdat$qsmk) == 1]
## 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:
##  -3.425367 -1.655796
## sample estimates:
## mean of x mean of y
##  1.984498  4.525079
```

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 propesnity score analysis is 0.39Kg higher.

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

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

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 (1984) show that Cochran’s result 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.

Propensity score subclassification/stratification

```
#nhefshwdat <- read.csv("~/Dropbox/Docs/sta305/2015/assignments/Assignment2/nhef
shw2dat.csv")
#Logistic regression of smoking cessation on covariates
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))

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

```
##
## Call:
## glm(formula = wt82_71[strat1] ~ qsmk[strat1], data = nhefshwdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -17.528   -3.882   -0.184    3.191   34.068
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.5829     0.4464   8.027 2.06e-14 ***
## qsmk[strat1]   1.5719     1.2205   1.288    0.199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 54.19378)
##
##     Null deviance: 16998  on 313  degrees of freedom
## Residual deviance: 16908  on 312  degrees of freedom
## AIC: 2148.8
##
## Number of Fisher Scoring iterations: 2
```

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

```
##
## Call:
## glm(formula = wt82_71[strat2] ~ qsmk[strat2], data = nhefshwdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -32.750   -3.946   -0.317    3.763   30.982
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.7000     0.4466   6.046 4.26e-09 ***
## qsmk[strat2]   5.0542     1.0287   4.913 1.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 50.66173)
##
##     Null deviance: 16979  on 312  degrees of freedom
## Residual deviance: 15756  on 311  degrees of freedom
## AIC: 2120.8
##
## Number of Fisher Scoring iterations: 2
```

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

```
##
## Call:
## glm(formula = wt82_71[strat3] ~ qsmk[strat3], data = nhefshwdat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -43.402   -3.707    0.263    4.807   41.663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.1214     0.5384   3.940 0.000101 ***
## qsmk[strat3]   3.7269     1.0519   3.543 0.000456 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 66.96828)
##
##     Null deviance: 21668  on 312  degrees of freedom
## Residual deviance: 20827  on 311  degrees of freedom
## AIC: 2208.2
##
## Number of Fisher Scoring iterations: 2
```

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

```
## 
## Call:
## glm(formula = wt82_71[strat4] ~ qsmk[strat4], data = nhefshwdat)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -25.578   -4.357    0.168    3.923   47.583  
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    0.9552     0.5131   1.862   0.0636 .  
## qsmk[strat4]   3.8712     0.9464   4.090 5.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 58.17997)
## 
##     Null deviance: 19067  on 312  degrees of freedom
## Residual deviance: 18094  on 311  degrees of freedom
## AIC: 2164.1
## 
## Number of Fisher Scoring iterations: 2
```

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

```
##
## Call:
## glm(formula = wt82_71[strat5] ~ qsmk[strat5], data = nhefshwdat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -28.7365   -3.9177    0.2878    4.9209   31.0088
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.2893     0.5878  -0.492   0.6230
## qsmk[strat5]   2.0550     0.9192   2.236   0.0261 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 63.92345)
##
##     Null deviance: 20200  on 312  degrees of freedom
## Residual deviance: 19880  on 311  degrees of freedom
## AIC: 2193.6
##
## Number of Fisher Scoring iterations: 2
```

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.

```
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)
```

```

Call:
glm(formula = wt82_71 ~ qsmk + as.factor(stratvar), data = nhefshwdat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-43.523   -3.971    0.019    4.212   47.405

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.3565     0.4373   7.675 2.89e-14 ***
qsmk                   3.2645     0.4543   7.186 1.03e-12 ***
as.factor(stratvar)2  -0.3191     0.6135  -0.520 0.603017
as.factor(stratvar)3  -1.1140     0.6157  -1.809 0.070602 .
as.factor(stratvar)4  -2.2229     0.6173  -3.601 0.000327 ***
as.factor(stratvar)5  -4.1404     0.6256  -6.619 4.97e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 58.89146)

    Null deviance: 97176  on 1565  degrees of freedom
Residual deviance: 91871  on 1560  degrees of freedom
AIC: 10835

Number of Fisher Scoring iterations: 2
```

```
# 95% confidence interval for treatment effect based on subclassification
confint(stratmodel)[2,]
```

```
   2.5 %   97.5 %
2.374168 4.154838
```

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.

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%

The other covariates were also investigated and subclassification balanced the 9 covariates within each subclass.

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.

```
prop.model.adj <- glm(wt82_71 ~ qsmk+ p.qsmk.obs, data = nhefshwdat)
summary(prop.model.adj)
```

```
## 
## Call:
## glm(formula = wt82_71 ~ qsmk + p.qsmk.obs, data = nhefshwdat)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -43.574   -3.977   -0.090    4.223   47.607  
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    5.560      0.509  10.923  < 2e-16 ***
## qsmk           3.397      0.456   7.451 1.53e-13 ***
## p.qsmk.obs   -14.752      1.885  -7.827 9.13e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 58.63809)
## 
##     Null deviance: 97176  on 1565  degrees of freedom
## Residual deviance: 91651  on 1563  degrees of freedom
## AIC: 10825
## 
## Number of Fisher Scoring iterations: 2
```

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

```
## Waiting for profiling to be done...
```

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

Comparing the three methods

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.