



Propensity Scores

Claire Miller
Tim Vigers
Elizabeth Wynn



Why use propensity score methods?

- Randomized controlled trials (RCTs) are the gold standard of study design
- Non-randomized observational studies are less expensive and easier to conduct
- Problems that arise in non-randomized studies:
 - Selection bias
 - Confounding
 - Difficult to establish causal relationship between exposure and outcome

Why use propensity score methods?

- Propensity score methods offer a solution to reduce effects of selection bias and confounding
- Seek to mimic RCTs in analysis and interpretation of results
- Allow for conclusions to be drawn about causal relationship between treatment/exposure and outcome





What are propensity scores?

- Propensity = “an inclination or natural tendency to behave in a certain way”
- Propensity score (PS) for each subject is the probability of treatment assignment conditional on observed baseline characteristics
- Usually estimated using logistic regression
- Incorporated into analysis through several different methods outlined in this presentation



Method and Process



Motivating Example using R:

- 400 male heart attack patients
- Some treated with new drug ($\text{trt}=1$), others received standard care ($\text{trt}=0$)
- Outcome variable is 30-day mortality ($\text{death}=1$)
- Covariates: age, admission severity, risk score



Selecting Baseline Covariates

- Only include variables that were measured at baseline
- Include variables associated with outcome or associated with both treatment and outcome
- Example: age, admission severity, and risk score are all associated with 30-day mortality
 - Note: Unknown yet whether these are associated with treatment



Estimating Propensity Scores

- The most common method for estimating PS is logistic regression of treatment group on the selected covariates.
- Estimated PS is the predicted probability of treatment from the regression model:

```
#Fit logistic regression model regressing on treatment/covariates  
pre.ps <- glm(trt ~ age + risk + severity, family="binomial",  
              data=data)  
#Solve for probabilities using predict function  
data$ps<-predict(pre.ps, type="response")
```



Alternative Estimation Methods

- Setoguchi et al. “simulated data for a hypothetical cohort study ($n = 2000$) with a binary exposure/outcome and 10 binary/continuous covariates with seven scenarios differing by non-linear and/or non-additive associations between exposure and covariates.”
- Compared logistic regression (LR) to recursive partitioning without pruning (RP1), RP with pruning (RP2), and neural networks (NN).
- NN and RP1 generally performed better based on C-index

Alternative Estimation Methods

	EPS MODELS			
	NN	RP1 (no pruning)	RP2 (pruning)	LR
(a) C-statistics				
Scenario A	0.77	0.79	0.67	0.76
Scenario B	0.78	0.79	0.66	0.75
Scenario C	0.78	0.80	0.66	0.74
Scenario D	0.80	0.81	0.69	0.78
Scenario E	0.80	0.81	0.68	0.78
Scenario F	0.80	0.81	0.70	0.77
Scenario G	0.80	0.81	0.70	0.75
(b) Standard error of the estimates				
Scenario A	0.17	0.18	0.16	0.17
Scenario B	0.17	0.18	0.16	0.17
Scenario C	0.18	0.19	0.16	0.17
Scenario D	0.18	0.19	0.16	0.17
Scenario E	0.17	0.19	0.16	0.17
Scenario F	0.17	0.19	0.16	0.17
Scenario G	0.17	0.19	0.16	0.17
(c) Bias of the estimates*				
Scenario A	0.011 (3%)	0.003 (1%)	0.107 (27%)	0.021 (5%)
Scenario B	0.011 (4%)	0.017 (4%)	0.060 (15%)	0.006 (2%)
Scenario C	0.011 (2%)	0.033 (8%)	0.006 (2%)	0.016 (4%)
Scenario D	0.011 (2%)	0.010 (3%)	0.081 (20%)	0.011 (3%)
Scenario E	0.011 (1%)	0.002 (<1%)	0.043 (11%)	0.025 (6%)
Scenario F	0.011 (4%)	0.005 (1%)	0.110 (28%)	0.011 (3%)
Scenario G	0.011 (3%)	0.017 (4%)	0.043 (11%)	0.044 (11%)



Alternative Estimation Methods

- Machine learning and data mining techniques do appear to slightly outperform logistic regression, but it's questionable how much better they really are.
- Is it worth setting up a neural network for a c-index of 0.77 instead of 0.76?
- There is also some evidence that while c-index is a good indicator of how well the PS model discriminates between treated and untreated, it may not necessarily tell you whether the PS model is correct.



Estimation Model Assumptions

All PS models have two big assumptions:

1. That all variables which could potentially affect treatment assignment have been measured.
 - a. This is also an assumption with regression approaches for estimating treatment effect, and not unique to PS
2. Every subject has a nonzero chance of receiving either treatment.



Propensity Score Methods

- There are a huge number of ways to match participants based on PS. (P.C. Austin's 2014 paper in *Statistics in Medicine* compares 12 different algorithms)
- The most common methods are:
 - 1 to 1 pair matching (most common)
 - Many to 1 matching
- Stratification
- Inverse probability of treatment weighting (IPTW)



One to One Matching

- With vs. without replacement
 - Without replacement ensures that each control is matched once at most
 - With replacement means that a control could be in multiple sets
 - Analysis must account for multiple matches

```
# Nearest neighbor method without replacement
```

```
match_dat<-match.data(matchit(trt ~ ps, data=data, ratio=1, replace=F,  
                             method = "nearest"))
```



One to One Matching

- Greedy vs. optimal matching
 - With greedy matching, a treated subject is picked at random, and matched with the closest control subject.
 - Does not matter if the control subject might be better matched to another treated subject.
 - Done until either the treated or control list is exhausted.
 - Optimal matching tries to minimize the total difference within pairs.



One to One Matching

- One study (Gu & Rosenbaum, 1993) found that optimal matching was actually no better than greedy matching.

```
# Optimal matching without replacement  
match_dat<-match.data(matchit(trt ~ ps, data=data, ratio=1, replace=F,  
                             method = "optimal"))
```



Nearest Neighbor Matching

- How do you define a “close” propensity score?
- Nearest neighbor matching selects the control subject with the closest PS, regardless of how close the scores actually are.
 - If there are multiple options with the same score, one is selected at random.
 - No restrictions on maximum difference



Caliper Matching

- Matching within caliper distance is similar to nearest neighbor matching, but the difference between PS must be within a predefined threshold (the caliper).
 - If there are no control subjects within the caliper, the unmatched treatment subject is excluded.



Caliper Matching

- What is a good maximum distance for caliper matching?
 - No generally agreed upon threshold
 - There are “theoretical arguments for matching on the logit of the propensity score, as this quantity is more likely to be normally distributed, and for using a caliper width that is a proportion of the standard deviation of the logit of the propensity score.” (Austin, 2011a)
 - A caliper that is 0.2 of the pooled standard deviation of the logit of the propensity score eliminates 99% of bias.



Caliper Matching

- This sounds like sort of a pain, but the MatchIt package does everything for you:

```
#Using nearest neighbor method (see documentation for details)  
#caliper indicates how close scores need to be to be matched  
match_dat<-match.data(matchit(trt ~ ps, data=data, caliper=.2,  
                             ratio=1, replace=F, method = "nearest"))
```




Stratification

- Group subjects into subsets by propensity score
 - Rank by score, then split into a predefined number of levels
 - Generally 5 groups is a good rule of thumb
 - Eliminates about 90% of bias
 - More groups only improves bias marginally

```
#Define groups based on propensity score quintiles  
data$ps_grp <- cut(data$ps, breaks=quantile(data$ps, prob=0:5*0.2),  
  labels=c("Q1", "Q2", "Q3", "Q4", "Q5"), include.lowest = TRUE)
```



Stratification

- Each strata can be treated like a separate RCT.
- Estimate treatment effect (differences in means or risk differences) in each group, and then average them.
 - Usually weighted by the number of participants in each group.
 - Same general approach for variance estimation, although this gets more complicated.
- Can use regression adjustment within each group if there are still some differences based on treatment assignment.



Inverse Probability of Treatment Weight (IPTW)

- Each subject is given a weight equal to the inverse probability of receiving the treatment that subject actually received.
- This allows calculation of the average treatment effect by comparing average weighted outcomes between treated and control groups.

Z_i = Indicator variable denoting treatment group

e_i = Propensity score

$$\text{Weight } w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}$$



IPTW

- Average treatment for the treatment and control groups can be calculated separately using different stabilized weights.
- Treatment group:

$$w_{i,ATT} = Z_i + \frac{(1 - Z_i)e_i}{1 - e_i}$$

- Control group:

$$w_{i,ATC} = \frac{Z_i(1 - e_i)}{e_i} + (1 - Z_i)$$



IPTW

- Weighting regression models using IPTW is part of a family of causal inference methods called the marginal structure model.
- Too complicated to go into detail, but there is a lot of research on these techniques.
- Lots of estimation methods for IPTW, which need to take the weighted sample into account.

#Create weights for each row of data

```
weight<-data$trt/data$ps+(1-data$trt)/(1-data$ps)
```



Covariate Adjustment

- Outcome is regressed on propensity score and treatment indicator variable using linear regression (continuous outcome) or logistic regression (binary outcome).
- Treatment effect is estimated from model coefficient of the treatment indicator variable
- Example code:

```
#Run logistic regression adjusting for propensity score  
mod1<-glm(death ~ trt+ps, family="binomial",  
           data=data)
```



Checking Balance Diagnostics

- After applying propensity score method check that there aren't systematic differences between groups for each covariate.
- For matching, stratification, and IPTW we check balance between treatment groups in:
 - Matching: matched sample
 - Stratification: individual strata
 - IPTW: weighted sample
- Many different methods of assessing goodness of fit for propensity score adjusted regression models (See Austin, 2008)



Balance Diagnostics: Standardized Difference

- Metric used to measure the difference in covariate between two groups:

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s_{treatment}^2 + s_{control}^2}{2}}}$$

- Good rule of thumb: $d < 0.1$

Standardized Difference Example

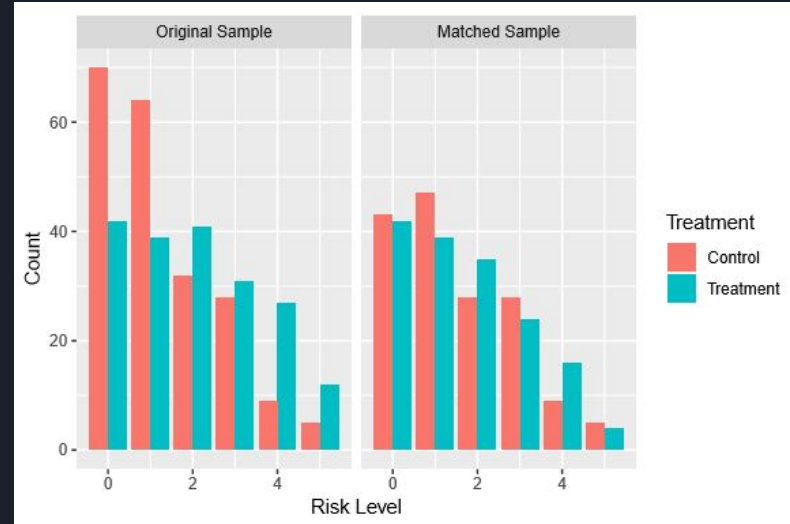
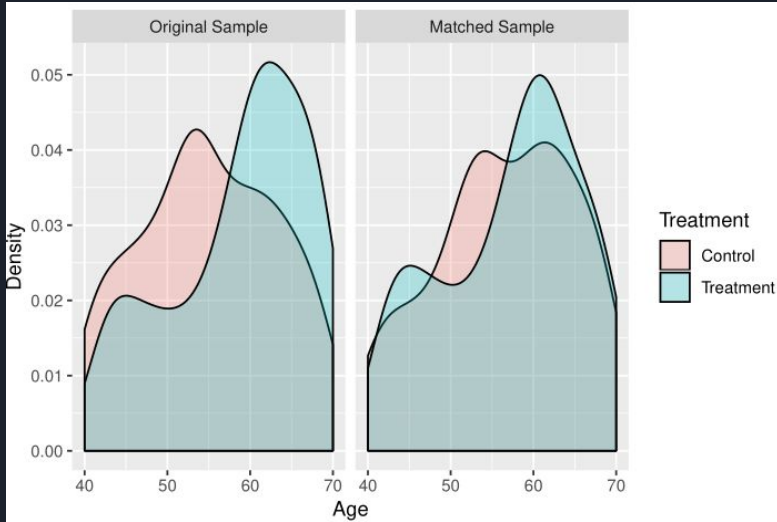
Using matched samples

	Stratified by trt				
	Control		Treatment		SMD
n	208		192		
age (mean (SD))	54.96	(8.27)	58.64	(8.26)	0.444
risk (mean (SD))	1.31	(1.31)	1.99	(1.54)	0.473
severity (mean (SD))	4.47	(2.03)	5.28	(2.16)	0.384

	Stratified by trt				
	Control		Treatment		SMD
n	160		160		
age (mean (SD))	56.42	(8.30)	57.23	(8.27)	0.097
risk (mean (SD))	1.52	(1.38)	1.66	(1.40)	0.099
severity (mean (SD))	4.83	(2.05)	4.94	(2.05)	0.058

Balance Diagnostics Plot Examples

Using matched sample





Estimate Treatment Effect

- Treatment effect estimation depends on propensity score method:
 - Matching: Compare outcome between groups in matched sample
 - Stratification: Estimate for each strata and then pool for overall treatment effect
 - IPTW: Compare average weighted outcomes between treated group and control group.
 - Covariate Adjustment: Estimate from model coefficient of the treatment indicator variable
- Additional significance testing and confidence intervals can be calculated using standard statistical methods.

Treatment Effect Analysis Example


- Using matched sample.
- Controversy over whether to treat groups as independent or not. We treat them as independent here.

```
#Difference in proportions between treatment groups, matched sample
treat_eff<-mean(match_dat$death[match_dat$trt=="Control"])-
  mean(match_dat$death[match_dat$trt=="Treatment"])

#Find number of people who died and totals for each treatment group
trt_died<-nrow(match_dat[match_dat$trt=="Treatment" & match_dat$death==1,])
trt_tot<-nrow(match_dat[match_dat$trt=="Treatment",])
contr_died<-nrow(match_dat[match_dat$trt=="Control" & match_dat$death==1,])
contr_tot<-nrow(match_dat[match_dat$trt=="Control",])

#2 Proportion z-test.
prop.test(c(contr_died, trt_died), c(contr_tot, trt_tot))
```

- Proportion who survived was 7% higher in the treated than the untreated group,
- Confidence interval is (-0.01, 0.15) so evidence that treatment works is inconclusive.



Alternative Method: Regression Covariate Adjustment

- Use regression and adjust for each baseline covariate separately in the model.
- Several limitations to this approach:
 - Harder to check regression goodness of fit diagnostics than propensity score balance diagnostics.
 - In regression the design of the study and analysis of treatment effect are integrated leading to possible bias in design by researchers.
 - Regression analysis is less effective when the outcome under study is rare or the treatment is common.



Limitations of Propensity Score Methods

- Do not always work well with small samples.
- If one covariate is missing, propensity score will also be missing.
- Only controls for measured covariates.



See Handout for References