

APPENDIX: PS MATCHING IN R

(with attached dataset and code)

Brian Lee (bklee@drexel.edu)

Assistant Professor

Dept. of Epidemiology and Biostatistics

Drexel University School of Public Health

SPER workshop June 17, 2013

Example 1

- 1987 National Medical Expenditures Survey
 - Persons 40+ with complete covariate data
 - Exposure: ever smoking
 - Control: never smoking
 - Outcome: lung cancer, laryngeal cancer, or COPD
 - N = 11,587
- *What is the effect of ever smoking on odds of lung cancer / laryngeal cancer / or COPD, as compared with never smoking?*

Variables

- **eversmk** (exposure)
- **lc5** (outcome)
- **LASTAGE**: age
- **MALE**: sex
- **RACE3**: race
- **beltuse**: seatbelt use
- **educate**:
- **marital**:
- **SREGION**: census region
- **POVSTALB**: poverty status
- 1/0 ever smoker / never smoker
- Lung / laryngeal CA / COPD
- In years
- 1/0 male / female
- Other / African American / Caucasian
- Rare / some / always
- college grad / some college/HS grad/other
- Married / widowed / divorced / separated / never married
- NE / MW / S / W
- Poor / near poor / low income/ middle income / high income

Selected sample characteristics

	ALL N = 11,587	Ever Smoke N = 6,564 (56.7%)	Never Smoke N = 5,023 (43.3%)
Cancer/COPD	1.9%	3.0%	0.5%
Age	60.2 years	59.1 years	61.7 years
Male	43.4%	55.4%	27.7%
Caucasian	78.1%	79.8%	75.8%
Rare use seatbelt	24.0%	26.0%	21.5%
College grad	14.2%	13.2%	15.4%
From South	36.8%	35.2%	38.9%
Poor	10.7%	10.2%	11.4%

1. Estimate the PS

- Goal: to achieve covariate balance on confounders so that they cannot bias results
- Take observed values of treatment (1/0), and build a model that estimates it using covariates k as predictors
 - Typically a parametric model used to estimate, e.g., logistic regression
 - `ps.model <- glm(treat ~ cov1 + cov2 + cov1*cov2 + cov3, data=dataset, family=binomial(link="logit"))`
 - `dataset$PS <- predict(ps.model, dataset, type="response")`
- Take as much care in building the PS model as you would an outcome model
 - Misspecification of the PS model can result in bias (although not as much as if you misspecify the outcome model)

1. Estimate the PS: Variable selection

- Begin with a DAG
- No post-treatment variables!
- Include variables that predict both treatment and outcome
- Little cost in including variables not related to the treatment but related to the outcome
- Exclude variables that are strong predictors of treatment with no obvious relation to the outcome
- Excluding a potentially important confounder can be costly in terms of bias
- PS analyses have included from a few to over 100 covariates
- General recommendation: theory-driven approach for variable selection

1. Estimate the PS: Variable parameterization

- Much like an outcome model, improper parameterization of a variable can result in residual confounding
- Use splines, polynomial terms, and interactions
- Still, misestimation of the propensity score is not a large problem as long as balance is obtained (e.g., exclusion of interactions or squares may be less severe for a PS model than for an outcome model) (Stuart 2010)

Machine learning

- Can use machine learning methods to estimate the PS (Westreich 2010)
 - E.g., neural nets, classification and regression trees (CART)
- Random forests and boosted CART work well for this (Lee et al. 2010)
- No need to specify non-linearities, interactions, etc. – these methods do automatically, just list the variables you want

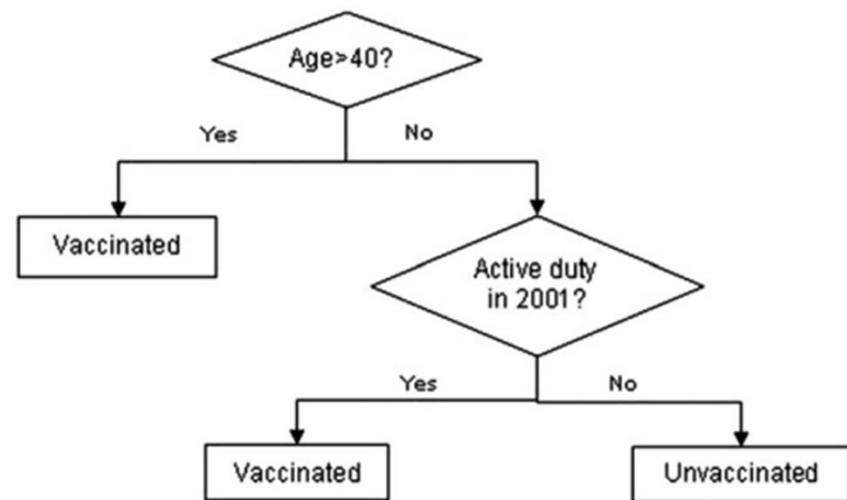


Fig. 3. A simple decision tree predicting smallpox vaccine status among military personnel.

Machine learning

- R: To implement boosted CART :
 - `library(twang)`
 - `ps.model <- ps(treatment ~ LASTAGE + MALE + educate + POVSTALB, data=data)`
 - `data$PS <- ps.model$ps[, 1]`
- R: To implement random forests:
 - `library(randomForest)`
 - `ps.model <- randomForest(treatment ~ LASTAGE + MALE + educate + POVSTALB, data=data)`
 - `data$PS <- ps.model$votes[, 2]`

1. Estimate the PS: missing data

- Observations with missing PS covariate data will be excluded from the PS model, and therefore also further analyses (i.e., the outcome model)
- Multiple imputation can be used to fill in missing data to estimate propensity scores but this has not been well-evaluated (Hayes, 2008)

Estimating the PS

Model 1

Logistic regression model

```
ps.model1 <- glm(eversmk ~  
LASTAGE +MALE +educate +beltuse  
+POVSTALB +marital +SREGION,  
data=a, family=  
binomial(link="logit"))
```

```
odds1 <- exp(predict(ps.model1,  
a))
```

```
a$PS <- odds1/(1+odds1)
```

Model 2

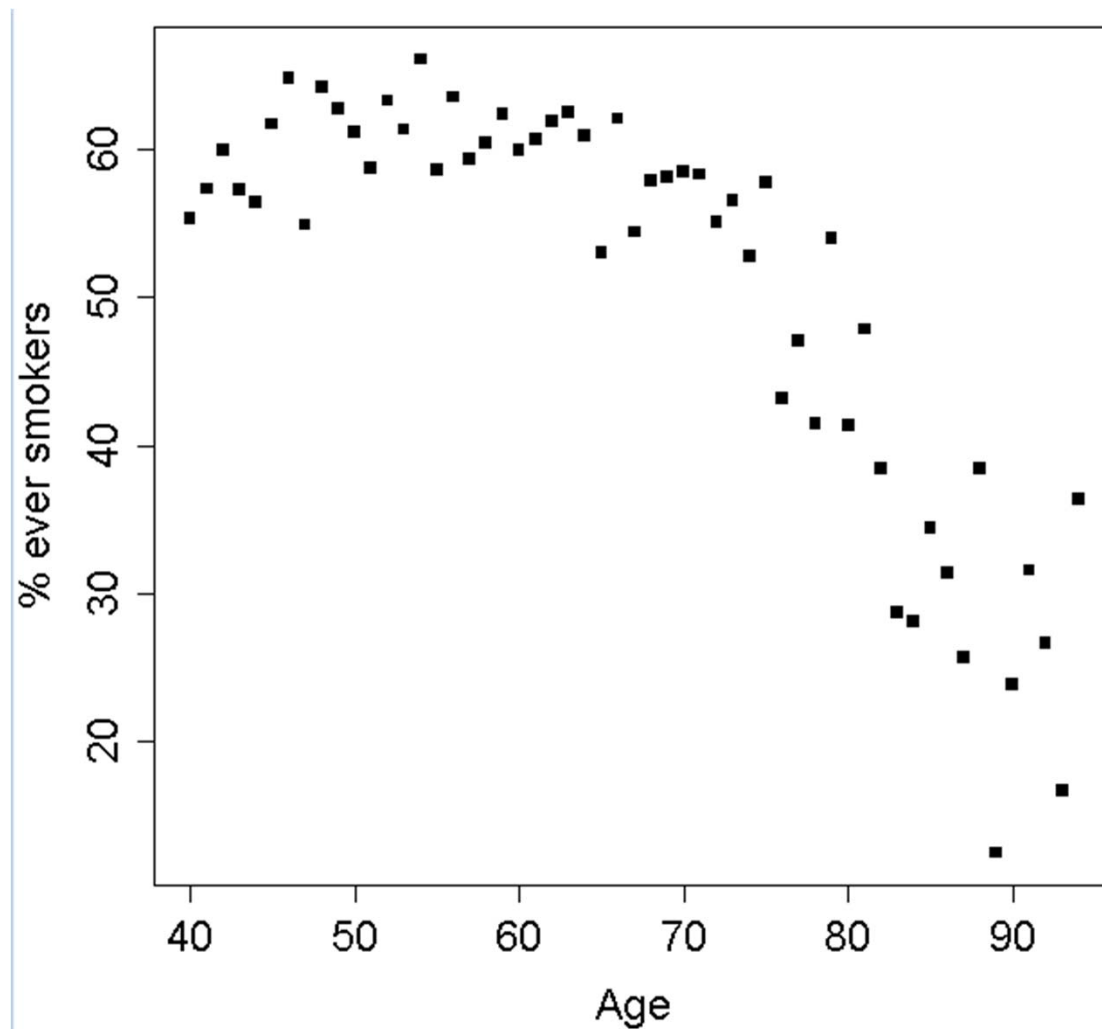
Boosted CART model

```
library(twang)
```

```
ps.model2 <- ps(eversmk ~  
LASTAGE + MALE + educate +  
beltuse + POVSTALB + marital +  
RACE3 + SREGION, data=a)
```

```
a$PS2 <- ps.model2$ps[, 1]
```

Is age linearly associated with ever smoking?



Matching attempt #1

- For logistic regression estimated PS (Model 1)...
- Let's try nearest neighbor 1:1 matching, greedy, no discarding of treated or controls

```
library(MatchIt)
nn1 <- matchit(eversmk ~ LASTAGE + MALE + educate + beltuse +
  POVSTALB + marital + RACE3 + SREGION, data=a, distance=a$PS,
  method="nearest")
nn1.data <- match.data(nn1)
summary(nn1, standardize=T)
```

Standardize in order to
show standardized balance

The *distance* option is not always necessary; if option is left out, Matchit can automatically calculate the PS based on the linear model as shown. For more complex PS models, e.g. with nonlinearities and such, estimate the PS beforehand and specify the resulting PS as the distance measure, like shown here.

Balance, matching attempt #1

- BEFORE matching: ASAM = 0.103

```
Summary of balance for all data:
      Means Treated Means Control SD Control Std. Mean Diff.
distance      0.6132      0.5055      0.1510      0.6945
LASTAGE      59.0891     61.6876     13.7372     -0.2134
MALE          0.5541      0.2773      0.4477      0.5567
educate1      0.1316      0.1545      0.3615     -0.0676
educate2      0.1507      0.1328      0.3394      0.0500
educate3      0.5178      0.4704      0.4992      0.0948
educate4      0.1999      0.2423      0.4285     -0.1060
beltuse2      0.2153      0.1915      0.3935      0.0578
beltuse3      0.5248      0.5939      0.4912     -0.1382
POVSTALB2     0.0457      0.0577      0.2333     -0.0576
POVSTALB3     0.1397      0.1457      0.3529     -0.0174
POVSTALB4     0.3167      0.3068      0.4612      0.0214
POVSTALB5     0.3961      0.3759      0.4844      0.0414
marital2      0.1371      0.2280      0.4196     -0.2641
marital3      0.1076      0.0685      0.2526      0.1261
marital4      0.0276      0.0219      0.1464      0.0347
marital5      0.0431      0.0581      0.2340     -0.0739
RACE32        0.1495      0.1666      0.3727     -0.0482
RACE33        0.7980      0.7581      0.4283      0.0993
SREGION2      0.2473      0.2491      0.4325     -0.0042
SREGION3      0.3524      0.3892      0.4876     -0.0771
SREGION4      0.1891      0.1857      0.3889      0.0085
```

Balance, matching attempt #1

- AFTER matching: ASAM = 0.163 (worse!)

Summary of balance for matched data:							
	Means Treated	Means Control	SD Control	Std. Mean Diff.	N's	C	T
distance	0.6797	0.5055	0.1510	1.1236			
LASTAGE	57.6311	61.6876	13.7372	-0.3331			
MALE	0.7237	0.2773	0.4477	0.8979			
educate1	0.1167	0.1545	0.3615	-0.1119	All	5023	6564
educate2	0.1553	0.1328	0.3394	0.0629	Matched	5023	5023
educate3	0.5415	0.4704	0.4992	0.1422			
educate4	0.1865	0.2423	0.4285	-0.1394			
beltuse2	0.2291	0.1915	0.3935	0.0915	Not matched	0	1541
beltuse3	0.4786	0.5939	0.4912	-0.2308			
POVSTALB2	0.0384	0.0577	0.2333	-0.0925			
POVSTALB3	0.1398	0.1457	0.3529	-0.0172	Discarded	0	0
POVSTALB4	0.3213	0.3068	0.4612	0.0312			
POVSTALB5	0.4027	0.3759	0.4844	0.0549			
marital2	0.0739	0.2280	0.4196	-0.4480			
marital3	0.1350	0.0685	0.2526	0.2146			
marital4	0.0342	0.0219	0.1464	0.0754			
marital5	0.0323	0.0581	0.2340	-0.1274			
RACE32	0.1412	0.1666	0.3727	-0.0715			
RACE33	0.8137	0.7581	0.4283	0.1383			
SREGION2	0.2437	0.2491	0.4325	-0.0125			
SREGION3	0.3317	0.3892	0.4876	-0.1204			
SREGION4	0.1931	0.1857	0.3889	0.0188			

So let's try something new

- Problems with balance in age so let's try the boosted CART model that should more accurately model the age – ever smoking relationship
- Because balance on gender was really bad, let's try exact matching on it
- And discard treated and controls beyond the range of overlap

Matching attempt #2

- For boosted CART estimated PS (Model 2)...
- Nearest neighbor 1:1 matching, greedy, discarding treated or controls beyond PS overlap, exact matching on sex
- `library(MatchIt)`
- `nn2 <- matchit(eversmk ~ LASTAGE + MALE + educate + beltuse + POVSTALB + marital + RACE3 + SREGION, data=a, distance=a$PS2, method="nearest", exact="MALE", discard="both")`
- `nn2.data <- match.data(nn2)`
- `summary(nn2, standardize=T)`

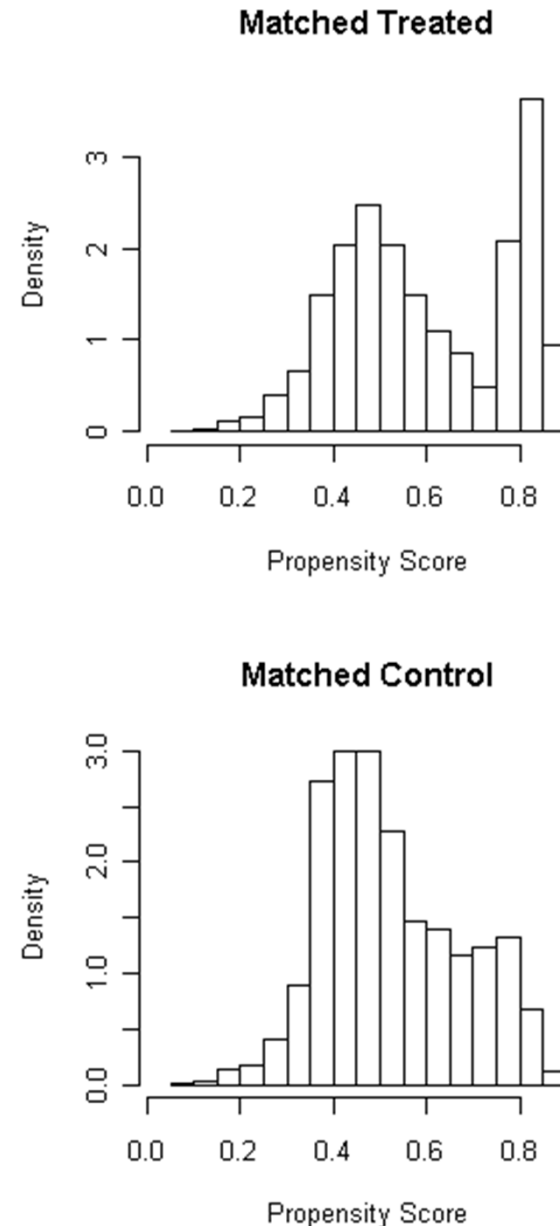
Balance, matching attempt #2

- AFTER matching: ASAM = 0.086..ok, but let's do better

```
Summary of balance for matched data:
      Means Treated Means Control SD Control Std. Mean Diff.
distance      0.5991      0.5226      0.1495      0.4705
LASTAGE      58.6778      59.5185     12.5502     -0.0690
MALE          0.3225      0.3225      0.4675      0.0000
educate1      0.0762      0.1685      0.3744     -0.2732
educate2      0.1463      0.1419      0.3490      0.0123
educate3      0.5903      0.4963      0.5000      0.1881
educate4      0.1873      0.1933      0.3949     -0.0150
beltuse2      0.2269      0.1988      0.3992      0.0681
beltuse3      0.4625      0.5956      0.4908     -0.2665
POVSTALB2     0.0532      0.0465      0.2106      0.0321
POVSTALB3     0.1574      0.1345      0.3412      0.0661
POVSTALB4     0.3287      0.3132      0.4638      0.0333
POVSTALB5     0.3368      0.4104      0.4920     -0.1505
marital2      0.1787      0.1782      0.3828      0.0013
marital3      0.1440      0.0769      0.2664      0.2167
marital4      0.0329      0.0229      0.1497      0.0608
marital5      0.0329      0.0528      0.2236     -0.0980
RACE32        0.1405      0.1544      0.3614     -0.0390
RACE33        0.8262      0.7829      0.4123      0.1078
SREGION2      0.2447      0.2488      0.4324     -0.0097
SREGION3      0.3366      0.3778      0.4849     -0.0862
SREGION4      0.1847      0.1866      0.3896     -0.0047
```

N's	C	T
All	5023	6564
Matched	4320	4320
Not matched	677	2213
Discarded	26	31

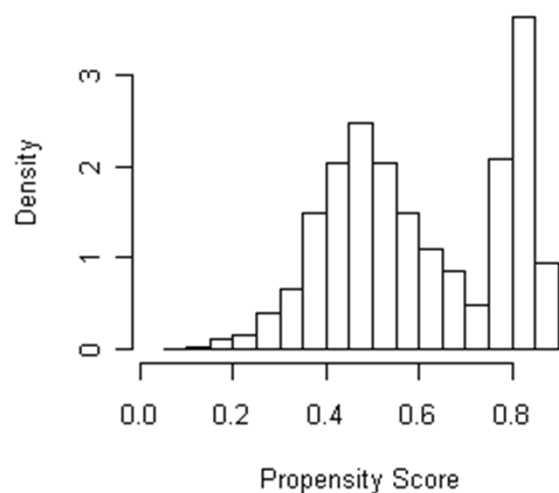
- There's a lot of treated units with high PS that don't seem to have good control matches that also have high PS
- Solution: try a caliper



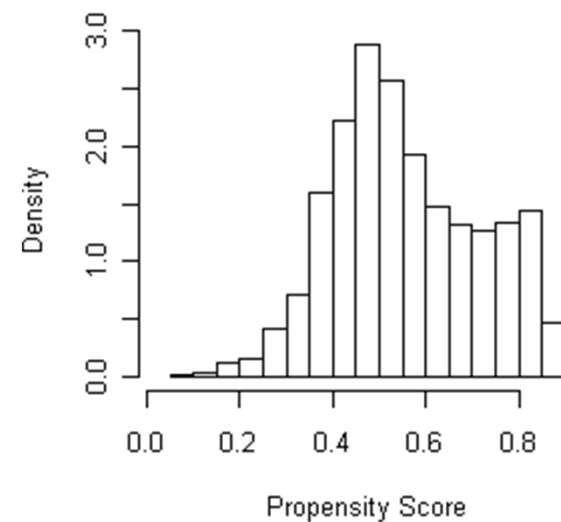
Matching attempt #3

- For boosted CART estimated PS (Model 2)...
- Nearest neighbor 1:1 matching, greedy, discarding treated or controls beyond PS overlap, exact matching on sex, caliper of 0.2 SD of the distance measure
- `library(MatchIt)`
- `nn3 <- matchit(eversmk ~ LASTAGE + MALE + educate + beltuse + POVSTALB + marital + RACE3 + SREGION, data=a, distance=a$PS2, method="nearest", exact="MALE", discard="both", caliper=0.2)`
- `nn3.data <- match.data(nn3)`
- `summary(nn3, standardize=T)`

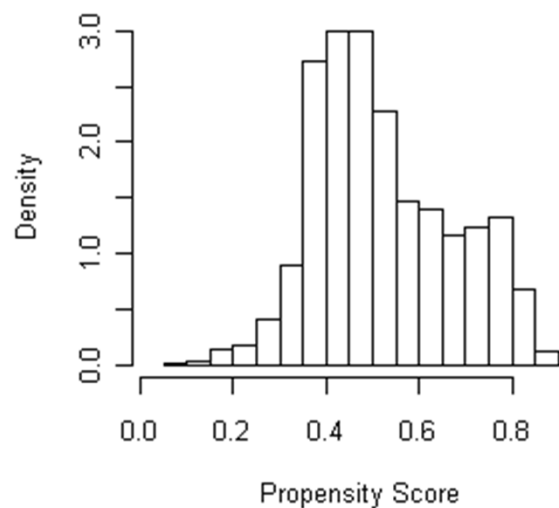
Match attempt #2
Matched Treated



Match attempt #3
Matched Treated

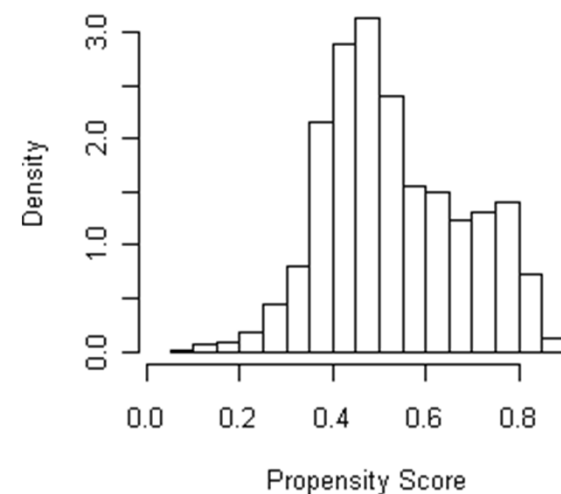


Matched Control



Going from #2 to #3: the caliper makes the controls look more similar to the treated according to the PS

Matched Control



Balance, matching attempt #3

- AFTER matching: ASAM = 0.027..excellent!

```
Summary of balance for matched data:
      Means Treated Means Control SD Control Std. Mean Diff.
distance      0.5579      0.5312      0.1494      0.1645
LASTAGE      58.9406     59.3036     12.5023     -0.0298
MALE          0.3345      0.3345      0.4719      0.0000
educate1      0.1487      0.1681      0.3740     -0.0573
educate2      0.1517      0.1467      0.3539      0.0137
educate3      0.5094      0.5016      0.5001      0.0157
educate4      0.1902      0.1836      0.3872      0.0166
beltuse2      0.2120      0.2032      0.4024      0.0215
beltuse3      0.5465      0.5887      0.4921     -0.0845
POVSTALB2     0.0501      0.0476      0.2130      0.0117
POVSTALB3     0.1433      0.1325      0.3391      0.0311
POVSTALB4     0.3092      0.3168      0.4653     -0.0164
POVSTALB5     0.3867      0.4069      0.4913     -0.0411
marital2      0.1828      0.1757      0.3806      0.0207
marital3      0.1038      0.0805      0.2721      0.0752
marital4      0.0267      0.0238      0.1525      0.0180
marital5      0.0486      0.0508      0.2196     -0.0109
RACE32        0.1593      0.1521      0.3592      0.0200
RACE33        0.7804      0.7907      0.4069     -0.0257
SREGION2      0.2380      0.2439      0.4295     -0.0137
SREGION3      0.3585      0.3740      0.4839     -0.0324
SREGION4      0.1953      0.1875      0.3903      0.0201
```

N's	C	T
All	5023	6564
Matched	4075	4075
Not matched	922	2458
Discarded	26	31

Estimate treatment effect

- In PS matched dataset (from match attempt #3), fit the outcome model
- `m1 <- glm(lc5 ~ everismk, data=nn3.data, family=binomial(link="logit"))`
- `summary(m1)`
- OR: 7.17 (95% CI: 4.27, 7.25)
- What does this estimate mean?
- To guard against residual confounding, may be a good idea to adjust for covariates in the outcome model

References

- Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in medicine*. 2006;25(12):2084-2106.
- Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006 Jun 15;163(12):1149-56.
- Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008 Sep 15;168(6):656-64.
- Hayes JR, Groner JI. Using multiple imputation and propensity scores to test the effect of car seats and seat belt usage on injury severity from trauma registry data. *J Pediatr Surg*. 2008;43(5):924-927.
- Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ, Jr. Splines for trend analysis and continuous confounder control. *Epidemiology*. 2011;22(6):874-875.
- Imai K, van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. *Journal of the American Statistical Association*. 2004;99(467):854-866.
- Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in medicine*. 2010;29(3):337-346.
- Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One*. 2011;6(3):e18174.
- Mansson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. *Am J Epidemiol*. 2007 Aug 1;166(3):332-9.
- Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci*. 2010 Feb 1;25(1):1-21.
- Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*. 2010 Aug;63(8):826-33.
- Westreich D, Cole SR, Funk MJ, Brookhart MA, Sturmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf*. 2010 Dec 9.