Propensity Scores

Tyler J. VanderWeele
Departments of Epidemiology and Biostatistics
Harvard T.H. Chan School of Public Health

Plan of Presentation

(1) Potential Outcomes and Confounding

(2) Propensity Score Properties

(3) Propensity Score Subclassification

(4) Example: Substance Disorders

(5) Propensity Score Matching

(6) Propensity Score Weighting

(7) Propensity Scores vs. Regression Methods

Potential Outcomes

Association is not causation; but we can formalize the relation

One common formalization used in this course will be "potential outcomes" (Rubin 1974 cf. Neyman 1923)

Consider a comparison of two treatments, for each individual we have two counterfactual outcomes (or "potential outcomes")

Y_1 = Would the individual have been cured if given treatment 1

Y_0 = Would the individual have been cured if given treatment 0

For each individual we only get to observe one of Y_1 and Y_0

We observe Y_1 if the individual received treatment 1

We observe Y_0 if the individual received treatment 0

We have no way to observe the other counterfactual outcome

Confounding

Even if the groups who received treatment 1 and those who received treatment 0 are not comparable…

It is possible that within strata of other variables (call them C) those who received treatment 1 and those who received treatment 0 are comparable

If so, then the proportions within strata of covariates C will reflect average counterfactual outcomes for the strata

We will use $X \perp\!\!\!\perp Y \mid Z$ to denote that X is independent of Y conditional on Z

Confounding: Formally, we say that the effect of treatment A on outcome Y is unconfounded given covariates C if for all values a:

$$Y_1 \perp\!\!\!\perp A \mid C \quad \text{and} \quad Y_0 \perp\!\!\!\perp A \mid C$$

Confounding

The “unconfoundedness” assumption is referred to in different ways:

- “No confounding” (Epidemiology and many disciplines)
- “Exchangeability” (Epidemiology)
- “Exogeneity” (Econometrics)
- “Selection on observables” (Econometrics and social sciences)
- “Ignorability” (Statistics)

Sometimes there are subtle differences:
Ignorability in statistics is $Y_a \perp\!\!\!\perp A \mid C$ and $0 < P(A=1|c) < 1$ for all c

In contrast “no confounding” or “exchangeability” is $Y_a \perp\!\!\!\perp A \mid C$
and… the second condition that $0 < P(A=1|c) < 1$ for all c
is taken as an additional assumption (sometimes called “positivity” or “experimental treatment assignment”)
And the condition that $Y_a=Y$ when $A=a$ is referred to as “consistency”

Causal Inference and Regression

If the effect is unconfounded conditional on C then within strata of the covariates C, the treatment groups are comparable (i.e. they have similar counterfactual outcomes) we can draw causal conclusions:

$$E[Y_a|C=c] = E[Y_a|A=a,C=c] = E[Y|A=a,C=c]$$

We can compute causal effects from the data;
If we have a regression model $E[Y|a,c] = \beta_0 + \beta_1 a + \beta_2 `c$

Then $E[Y_1|C=c] - E[Y_0|C=c] = E[Y|A=1,C=c] - E[Y|A=0,C=c]$

$$= \{\beta_0 + \beta_1 + \beta_2 `c\} - \{\beta_0 + \beta_2 `c\} = \beta_1$$

We can interpret our regression coefficient β_1 causally

An alternative (we will discuss advantages and disadvantages later) is propensity scores

Propensity Scores

If the effect of the treatment on the outcome is unconfounded, then we can estimate causal effects of treatment within strata of C

$E[Y_a|C=c] = E[Y|A=a,C=c]$

In principle we could just stratify by covariates

However, if C contains many variables, it may be difficult to stratify; there will be too many strata; and some of the strata will be sparse

Regression deals with this by extrapolation (same effect across C)

Propensity score methods *in some sense* allow one to reduce the problem of multivariate control to univariate control.

The initial work on propensity scores was carried out by Rosenbaum and Rubin (1983).

We will assume a binary treatment variable; the methods can however be extended (Imbens, 2000; Imai and van Dyk, 2004)

Propensity Scores

Propensity score: For treatment A and covariates C, the propensity score S(c) is defined by $S(c)=P(A=1|C=c)$.

In other words, the propensity score is the probability of a particular individual being assigned treatment conditional on some set of measured variables.

Example: Job training program is treatment; covariates are baseline income, age, gender, baseline health (ideally a sufficient set of confounders)

We will often use S for S(C)

The propensity score has certain theoretical properties which make it particularly useful in trying to control for various confounders.

Propensity Scores

The propensity score has certain balancing properties.

Result (Rosenbaum and Rubin, 1983):

$$\text{If } S = P(A = 1|C) \text{ then } P(C|A = 1, S) = P(C|A = 0, S) \text{ i.e. } C \coprod A|S$$

Within strata of S the treated subjects and the control subjects will have the same distribution of the measured covariates C.

In other words, if we stratify on the propensity score then within each stratum, all the measured covariates tend to be balanced between the treatment and control groups.

In effect, as far as concerns *the measured covariates*, it is as though we have conducted a randomized trial within each stratum.

Unlike a randomized trial, however, stratification by use of a propensity score balances only those covariates that have been measured.

It does not balance those covariates that have not been measured.

Propensity Scores

In fact, if treatment is unconfounded conditional on C then it is also unconfounded conditional on S

Theorem (Rosenbaum and Rubin, 1983): $\text{If } Y_a \coprod A|C \text{ then } Y_a \coprod A|S$

If it suffices to adjust for C to estimate the effect of treatment then it suffices also instead to adjust only for the propensity score S.

If the counterfactual outcomes are independent of treatment conditional on C then the counterfactual outcomes are also independent of treatment conditional on S.

The advantage of adjusting for the propensity score S rather than C itself is that S is a single variable and stratification on a single variable is relatively straightforward

C may contain dozens of variables, making stratification and other methods of adjustment difficult

Propensity Scores

We can use propensity scores to estimate treatment effects

Whenever $Y_a \coprod A|C$ for all a we also have $Y_a \coprod A|S$ for all a and thus

$$(1)\ \mathbb{E}[Y_a] = \sum_s \mathbb{E}[Y|A=a, S=s]P(S=s)$$

$$(2)\ \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \sum_s \{\mathbb{E}[Y|A=1, S=s] - \mathbb{E}[Y|A=0, S=s]\}P(S=s)$$

$$(3)\ \mathbb{E}[Y_1|S=s] - \mathbb{E}[Y_0|S=s] = \mathbb{E}[Y|A=1, S=s] - \mathbb{E}[Y|A=0, S=s]$$

$$(4)\ \mathbb{E}[Y_1|A=1] - \mathbb{E}[Y_0|A=1] = \sum \{\mathbb{E}[Y|A=1, S=s] - \mathbb{E}[Y|A=0, S=s]\}P(S=s|A=1)$$

Now we only have to adjust for one variable, the propensity score S

One could also estimate RR's or OR's within strata of the propensity score

Estimating Propensity Scores

In practice the propensity score is unknown and must be estimated.

This is most typically done by means of a logit model regressing the treatment/exposure A on the covariates C:

Propensity Score Model: $\text{logit}[P(A=1|c)] = \alpha + \beta' c$

Estimated Propensity Score: $e^{\alpha + \beta' c} / (1 + e^{\alpha + \beta' c})$

Surprisingly, it turns out that often the estimated propensity scores can perform just as well or even better than the true propensity scores.

An estimated propensity score cannot distinguish between systematic imbalance in the covariates and imbalance in the covariates that is due to chance and eliminates both

Adjustment for the true propensity score, even if it were known, would only remove systematic bias (Rubin and Thomas, 1996; Hirano, Imbens and Ridder, 2003)

Propensity Score Stratification

- Stratification on the propensity score allows for the control of all covariates simultaneously while stratifying on only one variable.
- When stratifying by the estimated propensity scores a decision needs to be made as to the number of strata to be used.
- Exact stratification cannot be used because typically each subject will have a different estimated propensity score.

- When stratifying by propensity scores it has become common practice to use five strata.
- Cochran (1968) proves various theoretical results showing that if there are reasonable numbers of both treatment and control subjects in each strata then five or six strata are usually sufficient to remove at least 90% of the bias of the crude estimates without any adjustment.
- Estimated propensity score quintiles, of both groups combined or in the treated or control group alone, are most often used in practice as boundaries for each strata.

Propensity Score Stratification

We define the block variable B by:
$B = 1$ if $S \in [0, q_1]$
$B = 2$ if $S \in (q_1, q_2]$
...
$B = n$ if $S \in (q_{n-1}, 1]$

Cochran's result then suggests:

$$\mathbb{E}[Y_1] - \mathbb{E}[Y_0] \approx \sum_c \{\mathbb{E}[Y|A=1, B=b] - \mathbb{E}[Y|A=0, B=b]\} P(B=b)$$

$$\mathbb{E}[Y_1|A=1] - \mathbb{E}[Y_0|A=1] \approx \sum_c \{\mathbb{E}[Y|A=1, B=b] - \mathbb{E}[Y|A=0, B=b]\} P(B=b|A=1)$$

$$\mathbb{E}[Y_1|A=0] - \mathbb{E}[Y_0|A=0] \approx \sum_c \{\mathbb{E}[Y|A=1, B=b] - \mathbb{E}[Y|A=0, B=b]\} P(B=b|A=0)$$

We can weight the stratum-specific effect estimates by the overall size of the strata, or by the size of these among the treated or untreated to get the average treatment effect, or the effect among the treated or untreated

We can also do this with other effect measures (e.g. risk ratio, odds ratio)

Propensity Score Stratification

- After estimating the propensity scores and fixing the strata boundaries, the distribution of propensity scores of the treatment and control groups can be compared to ensure that there is sufficient overlap

- Support: Subjects with propensity scores outside the range of the propensity score in the other group are often excluded from the analysis as there is no basis on which to make a comparison
- This changes the interpretation of the estimates: causal effects amongst subjects with common propensity score support

- It might seem best to use the distribution of estimated propensity scores of the group with the fewest number of subjects to set boundaries;
- A disadvantage of such an approach is that e.g. if the distribution of treatment group alone is used then the strata with the lowest propensity scores is likely to cover a wide range of propensity scores
- It also changes interpretation (e.g. effect on treated or untreated)
- In practice, the distribution of the propensity scores of both groups combined will often be used to determine boundaries for the strata.

Assessing Balance

If the propensity score model is correct, the covariates should be balanced

For each variable C_i and each block b we calculate $\mathbb{E}[C_i|A=1,B=b] - \mathbb{E}[C_i|A=0,B=b]$

Because $P(C|A=1,S=s) = P(C|A=0,S=s)$ these differences should be small

If these differences are large then either:

(1) The block sizes are too large

(2) The propensity score model is incorrect

- If there are large difference the propensity score model can be reformulated so as to include interaction or quadratic terms among the covariates.
- Balance can be checked again
- If it is not possible to obtain fairly good covariate balance within strata then it may be that covariate distributions do not overlap sufficiently

Example: Substance Disorders

Mojtabai R. and Zivin J.G. (2003). Effectiveness and cost-effectiveness of four treatment modalities for substance disorders: a propensity score analysis, Health Services Research 38:233-59.

Study Population: 3,047 patients in a random sample of 99 drug treatment facilities; four treatment facility types to treat substance disorders compared

We will focus on: Residential Facility vs. Outpatient Drug-Free Facility

Outcomes: Abstinence, Any Reduction in Use, Cost Effectiveness

Baseline Characteristics: Gender, Age, Education, Race, History of Homelessness, Type of Drug, Number of Drugs Abused, Heavy Use, Psychiatry Comorbidity, Involuntary Treatment

Major differences in all of these across treatment modalities

Example: Substance Disorders

Logistic Regression of treatment type on all pretreatment variables was used to estimate propensity scores

Quintiles of the estimated propensity score were used to define strata

Covariate balance was checked in each of the strata for each of the pair-wise treatment comparisons

For those covariates for which statistically significant differences remained, the interaction terms of that variables with all other variables were included in the logistic regression model, and the propensity score model was re-estimated

Assessment procedure was repeated until no statistically significant differences remained in any of the strata

Once strata with balanced covariates were obtained, odds ratios were used to compare outcomes (abstinence or reduction in use) for each strata

Example: Substance Disorders

Overall estimates slightly favor outpatient facilities but wide CI's

Propensity Score Strata	*N in Modalities*[a]		*Abstinence OR (95% CI)*	*Reduction in Use OR (95% CI)*
	Residential (RES) versus Outpatient Drug-free (OP)			
1 Propensity: ↓RES ↑OP	$N_{RES} = 34$	$N_{OP} = 138$	0.27 (0.08–0.97)*	0.50 (0.20–1.24)
2	70	103	0.52 (0.20–1.33)	0.91 (4.50–1.85)
3	85	85	1.57 (0.63–3.89)	0.66 (0.34–1.26)
4	119	53	1.77 (0.71–4.40)	0.95 (0.48–1.89)
5 Propensity: ↑RES ↓OP	136	35	1.52 (0.62–3.74)	1.27 (0.60–2.72)
Conditional logistic regression	444	414	*0.95 (0.66–1.36)*	*0.81 (0.61–1.09)*

Possibly some evidence for some advantage of outpatient facilities for those most likely to be in them

And conversely some for residential facilities for those most likely to be in them but CI's are all quite wide

Example: Substance Disorders

Costs for outpatient facilities were about half that of residential

Thus cost effectiveness ratios strongly favor outpatient facilities

Propensity Score Strata	*N in Modalities*[a]		*Abstinence Ratio*[b] *(95% CI)*	*Reduction in Use Ratio*[b] *(95% CI)*
	Residential (RES) versus Outpatient Drug-free (OP)			
1 Propensity: ↓RES↑OP	$N_{RES} = 34$	$N_{OP} = 138$	1.64 (0.80–4.10)	2.20 (1.21–4.19)**
2	70	103	2.92 (1.41–8.28)*	2.49 (1.49–4.29)***
3	85	85	1.88 (0.76–4.27)	3.41 (1.99–5.68)***
4	119	53	2.06 (0.72–4.40)	3.36 (2.28–5.06)***
5 Propensity: ↑RES↓OP	136	35	3.96 (1.64–7.83)*	3.74 (2.20–6.42)***
Total group	444	414	2.35 (1.71–3.09)***	2.67 (2.19–3.23)***

Authors conclude transitioning to this particular treatment modality could free resources to allow wider access to such care

Propensity Score Matching

Rosenbaum and Rubin (1985) "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score"

(1) Nearest Propensity Score Matching

- Randomly order treated subjects
- Select control subject with the nearest logit propensity score $\widehat{q}(c) = \log(\frac{\widehat{s}(c)}{1-\widehat{s}(c)})$ and remove the treated and control subjects from the pool

Note: logit propensity scores tend to be more normally distributed; normally distributed covariates have better properties in Mahalanobis matching

(2) Mahalanobis Matching with the Propensity Score

- For covariate vectors C the Mahalanobis distance is defined by between individuals i and j is defined by $d(i,j) = (c_i - c_j)'C^{-1}(c_i - c_j)$ where $C = Cov(C)$
- Include the logit propensity score $\widehat{q}(c) = \log(\frac{\widehat{s}(c)}{1-\widehat{s}(c)})$ in the covariate vector
- Randomly order treated subjects
- Select control with the smallest Mahalanobis distance and remove the treated and control subjects from the pool

Propensity Score Matching

Rosenbaum and Rubin (1985) "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score"

(3) Mahalanobis Matching within Propensity Score Calipers
- Randomly order treated subjects
- Select control with the smallest Mahalanobis distance within calipers of propensity score strata and remove the treated and control subjects from the pool
- Rosenbaum and Rubin (1985) suggest calipers 1/4 of the standard deviation of the estimated logit propensity score $\widehat{q}(c)$

Once a matched sample is constructed we can simply take sample averages of the outcomes as estimates of treatment effect

$\mathbb{E}[Y_1] - \mathbb{E}[Y_0] \approx \frac{1}{N}\sum_{i}^{N} Y_i^T - \frac{1}{N}\sum_{i}^{N} Y_i^C$ where Y_i^T denote the treated subjects, Y_i^C denote the control subjects and N denotes the total number of treated subjects

Rosenbaum and Rubin conclude approach 3 led to best covariate balance

Propensity Score Matching

- We can also easily assess and present results about covariate balance
- The matching methods described requires that there be many more control subjects than treated subjects (e.g. with equal sized groups matching will simply give the original sample)
- Analysis is simplified once matching is complete
- Matching can result in the discarding of subjects and a loss in efficiency
- If there are many more control subjects than treated subjects (or vice versa) one can do k:1 matching to improve efficiency
- Matching on one covariate results in $\sqrt{n}$ consistent estimator; matching on multiple covariates generally does not (Abadie and Imbens, 2006)
- Bootstrap generally fails with matching and alternative methods are needed for standard error (Abadie and Imbens, 2008, 2016)
- Propensity score matching has been criticized due to sensitivity to investigator choices (King and Nielson, 2016)
- Other matching methods are put forward e.g. Coarsened Exact Matching (e.g. Iacus et al., 2012) that are often more efficient but generally require fewer covariates

Propensity Score Weighting

A third alternative to use the propensity score to estimate causal effects is by weighting

Theorem: If $Y_a \coprod A|C$ then $\mathbb{E}[Y_1] = \mathbb{E}[\frac{YA}{S}]$ and $\mathbb{E}[Y_0] = \mathbb{E}[\frac{Y(1-A)}{1-S}]$.

If the covariates C suffice to control for confounding we can estimate

$$\mathbb{E}[Y_1] = \mathbb{E}[\frac{YA}{S}] \text{ by } \frac{1}{N}\sum_{i=1}^{N_T} \frac{Y_i^T}{s_i^T}$$

$$\mathbb{E}[Y_0] = \mathbb{E}[\frac{Y(1-A)}{1-S}] \text{ by } \frac{1}{N}\sum_{i=1}^{N_C} \frac{Y_i^C}{1-s_i^C}$$

where N_T, N_C and N are the number of treated subjects, the number of control subjects and the total number of subjects respectively, Y_i^T and Y_i^C denote outcomes for individual treated and control subjects and s_i^T and s_i^C denote the propensity scores for treated and control subjects

Propensity Score Weighting

Intuition: For one treated subject with P(A=1|C)=.2 we have P(A=0|C)=0.8; so there are 4 controls with S(C)=.2 and for each treated with S(C)=.2;

Thus there 5=(1/(.2)) subjects total with S(C)=.2 for each treated subject

We weight by 1/.2 = 5 to account for all 5 subjects

Properties of weighting

- There is no residual bias with propensity score weighing, unlike subclassification and matching
- If there are treated subjects with propensity scores close to 0 or control subjects with propensity scores close to 1 then the estimates of the average treatment effect can have very large standard errors and be very sensitive to the propensity score model specification;
- Often extreme weights are "trimmed" (eliminated) or "truncated" (set to values) e.g. $1^{st}/99^{th}$ or $5^{th}/95^{th}$ percentiles
- Propensity score weighting is closely related to the Horwitz-Thompson estimator (1952)
- The approach has been generalized to time-varying treatments by Robins (1998, 1999, 2000) using marginal structural models

Regression vs. Propensity Scores

(1) Model Specification:

- Propensity score methods do not impose any functional form on the relation between outcome and the covariates within each treatment group whereas linear regression models do.
- Linear regression forces the relation between outcome and the covariates to be linear; adjustment by stratification does not.
- Propensity Scores make assumptions about model specification between the exposure and the covariates
- But here we only need to get the propensity score itself (approximately) correct not the relations with any given covariate
- Propensity score methods may be especially helpful when it is difficult to model the outcome (e.g. cost data)
- So called “doubly robust” methods allow either to be correctly specified (e.g. Bang and Robins, 2005; van der Laan and Rose, 2011)

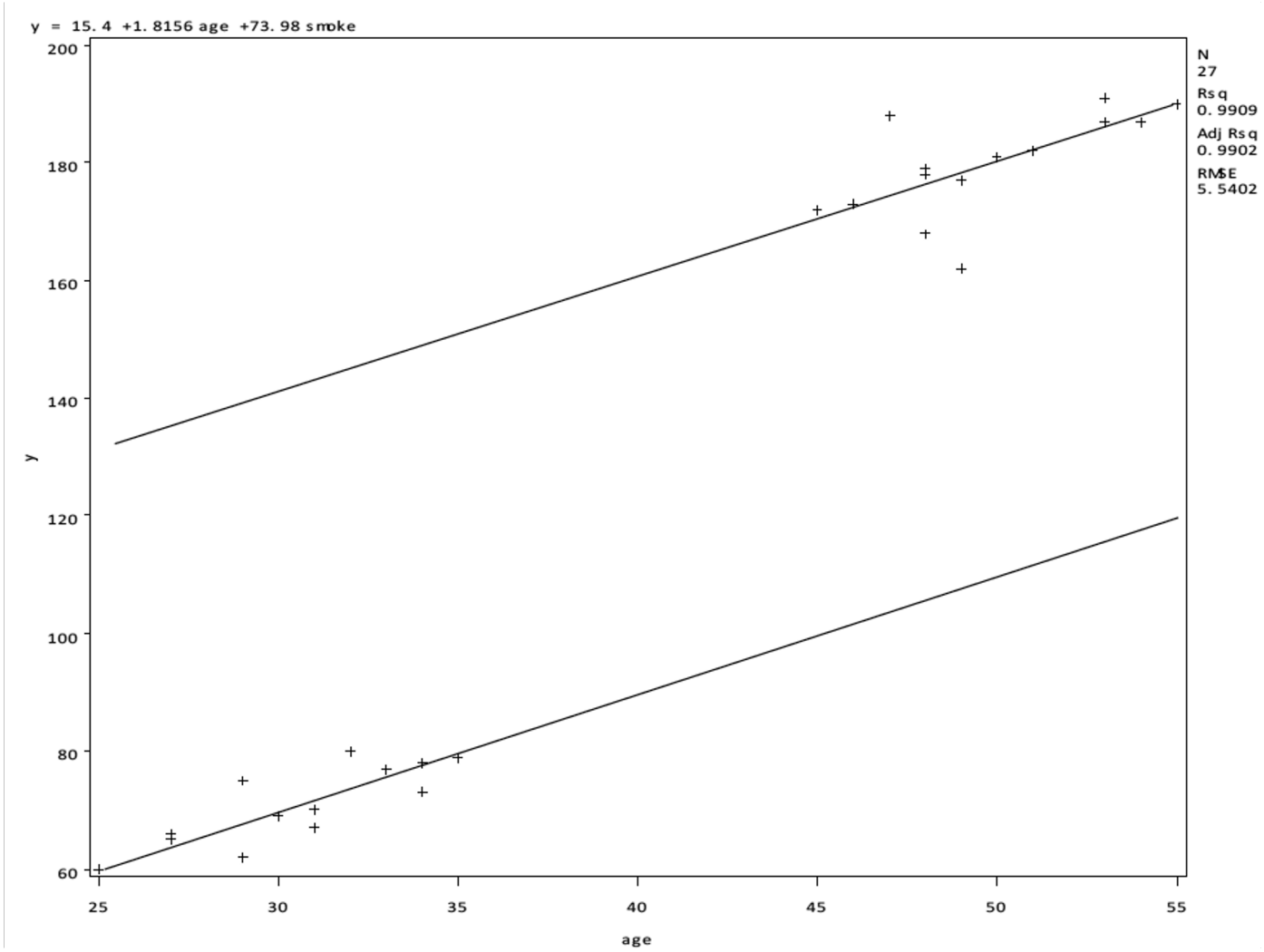
Regression vs. Propensity Scores

(2) Checking for Covariate Overlap:

- Propensity score methods encourage checking for overlap
- Standard methods of analysis such as linear regression and logistic regression can be deceptive because they provide no warning about when their use is or is not warranted
- Regression uses linearity assumptions and simply extrapolates
- It provides estimates even when there is little or no overlap in the covariate distribution between the treatment and control groups
- As an extreme example, Rubin (1997) notes that comparing survival rates among 70-year-old smokers and 40-year-old nonsmokers gives essentially no information about the effect of smoking or nonsmoking for either 70-year-olds or 40-year-olds.
- Nothing in the standard output of regression software will draw attention
- With propensity score stratification, insufficient overlap will be apparent; some matching techniques will also make this clear through inability to find matches; with weighting lack of overlap is evident in having very large standard errors

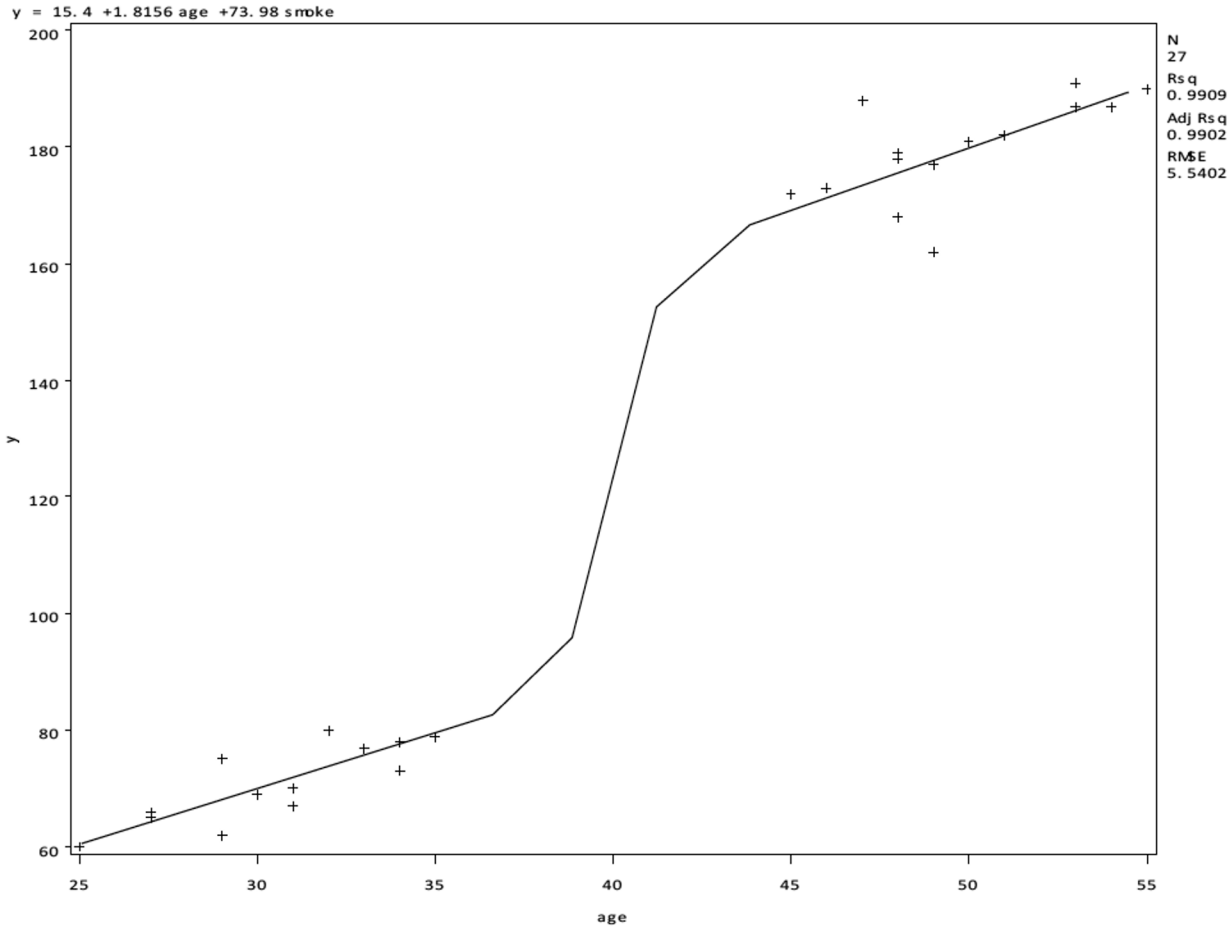
Regression vs. Propensity Scores

Suppose all the exposed are old and all the unexposed are young with data points given below. If we fit a linear regression we would get the following with a substantial effect:



Regression vs. Propensity Scores

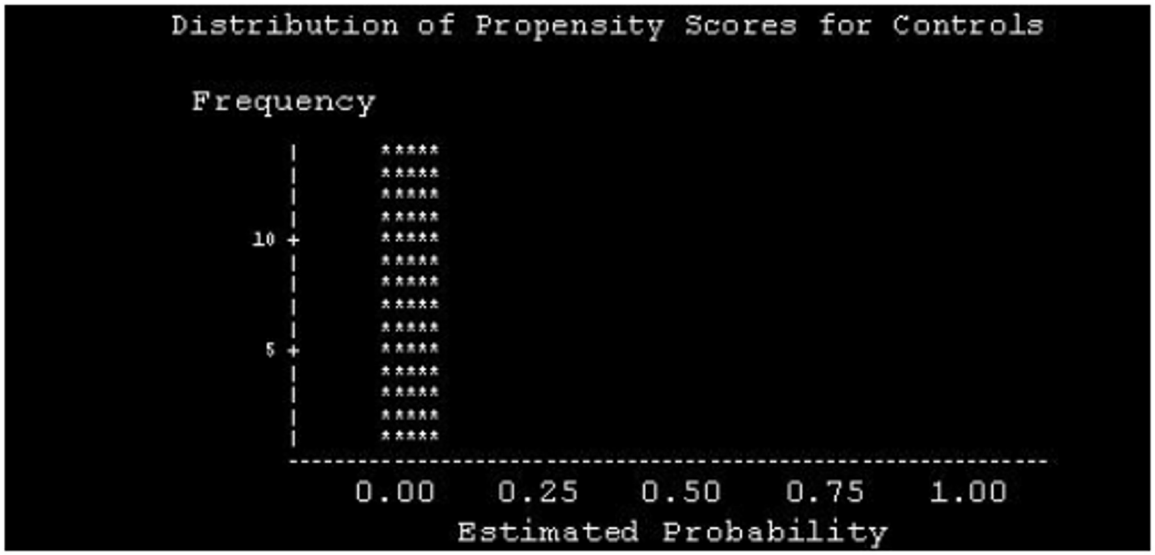
- The data are equally consistent with the following functional form where there is no effect

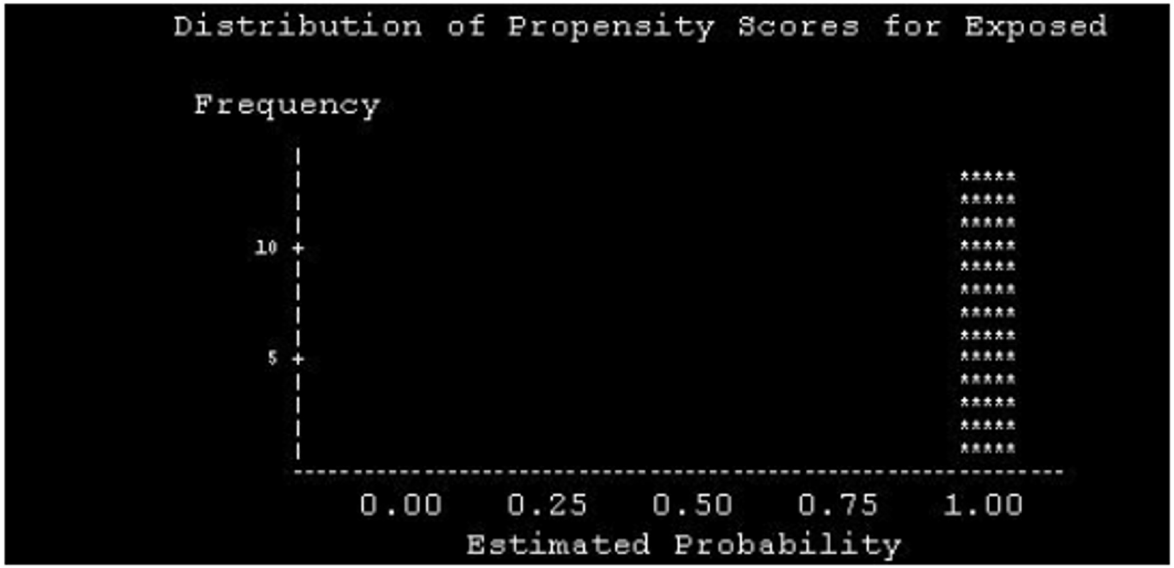


Functional form assumptions drive the result and the non-comparability of the exposed and unexposed led to extrapolation; Propensity scores will allow us to see this

Regression vs. Propensity Scores

We see lack of overlap immediately when using propensity score strata





Regression vs. Propensity Scores

(3) Frequency of Outcome vs. Exposure

- With rare binary outcomes but common treatment, the asymptotic properties of logistic regression will not apply; whereas propensity score methods can often still be used if treatment is common
- Simulations suggest that in the case of a binary outcome, when the outcome is comparatively rare (seven or less events per confounding variable), estimates from propensity score methods have better empirical properties than logistic regression estimates -- less biased, more robust, and more precise (Cepeda et al. 2003)
- Another way to think about these simulation results is that with rare outcome and common exposure propensity score methods will often allow for control for more confounding variables than logistic regression while still ensuring the sufficiently large sample size so that asymptotic properties hold.
- But with more common outcome and rare exposure, logistic regression will perform better

Regression vs. Propensity Scores

(4) Integrity of Results:

- With propensity scores modeling decisions can be made before examining the outcome data
- With propensity score methods there is perhaps less temptation to adjust the model and the covariates included until one attains the desired result

(5) Efficiency:

- Regression tends to generate smaller standard errors than propensity score matching and weighting, but at the price of assumptions of functional form *on the outcome*
- So-called "doubly robust" methods (model both outcome and exposure) capture some of the advantages of both approaches (Bang and Robins, 2005; van der Laan and Rose, 2011), but small sample performance is sometimes not good

Extensions

Non-Binary Treatment:

- Rosenbaum (2002) and Lu et. al. (2001) consider use of cumulative logistic proportional odds models for ordinal treatment
- With categorical treatments all two-way comparisons can be considered
- Imbens (2000) and Imai and van Dyke (2004) discuss further extensions to continuous treatment but the advantages over regression are somewhat fewer

Other Extensions:

- Large literature on matching
- Large literature on doubly-robust estimation
- Large literature on simulations concerning practices for estimating effects with propensity scores

Conclusions

(1) Propensity scores can be a helpful alternative approach to regression

(2) It more naturally forces checking of covariate overlap; there are fewer modeling assumptions with the outcome; it is easier to make modeling decisions without looking at outcome data

(3) Methods have been used widely; Rosenbaum and Rubin (1983) has been cited >34,000 times

(4) Matching methods are especially popular in the medical literature but have disadvantages of greater difficulty in estimating standard errors (often done incorrectly in the empirical literature) and making interpretation more difficult

(5) Propensity score methods can be extended to the time-varying treatment setting to deal with complex issues of time-dependent confounding i.e. marginal structural models (Robins et al., 2000)