An introduction to propensity scores

Federico Andreis
federico.andreis@gmail.com

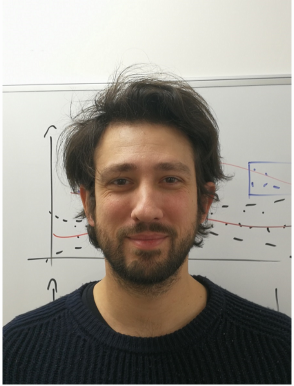
University of Milan-Bicocca

MEBRIC - *May 2021*

Something about me

Dr Federico Andreis,

- Previously Lecturer / Faculty Statistician @ Health Sciences and Sport, University of Stirling
- PhD in Statistics @ University of Milan-Bicocca / Karolinska Institutet, Stockholm
- sampling theory and applications, statistical modelling of electronic health records, non-standard Bootstrap, item response theory



Federico_Andreis · Chicco_Stat · Federico Andreis

Plan for the lecture

The course material for these lectures include:

- recording
- lecture slides
- references for further reading.

Learning outcomes

By the end of these lectures, you should be able to:

- describe the problem of drawing causal inference from observational studies
- define what propensity scores are and how to use them
- carry out basic propensity scores adjustments.

Exchangeability

In a **randomised trial**, we have that

$$\underbrace{P(Y_1 = 1|A = 1)}_{=P(Y=1|A=1)} = P(Y_1 = 1)$$

$$\underbrace{P(Y_0 = 1|A = 0)}_{=P(Y=1|A=0)} = P(Y_0 = 1).$$

Thanks to randomisation, Y_0 and Y_1 are independent of A:

$$(Y_0, Y_1) \perp\!\!\!\perp A.$$

We say that the exposed and unexposed are **exchangeable**. Under exchangeability, association = causation.

Exchangeability in observational studies

In an **observational study**, we have that

$$\underbrace{P(Y_1 = 1|A = 1)}_{=P(Y=1|A=1)} \neq P(Y_1 = 1)$$
$$\underbrace{P(Y_0 = 1|A = 0)}_{=P(Y=1|A=0)} \neq P(Y_0 = 1).$$

In other words, we **do not have exchangeability**, meaning Y_0 and Y_1 depend on A:

$$(Y_0, Y_1) \not\perp\!\!\!\perp A.$$

As a consequence, **association** $\neq$ **causation** and, for example, $RR \neq CRR$.

Three important questions

- What is the cause of non-exchangeability in observational studies?
- Can we identify non-exchangeability in a population/sample?
- How can we estimate causal effects in the presence of non-exchangeability?

What is the cause of non-exchangeability in observational studies?

Suppose that there is a covariate, L, which affects both A and Y. For example:

- $L =$ 'age'; older people have higher BMI (A) than young people, and are more likely to develop cancer (Y)

If so, we will find an association between A and Y, even if A has no causal effect on Y.

The association between A and Y suffers from **confounding** by L which, in turn, **causes non-exchangeability**.

Can we identify non-exchangeability in a population/sample?

We have non-exchangeability if (Y_0, Y_1) and A are not independent. That is, if

$$P(Y_1 = 1|A = 1) \neq P(Y_1 = 1)$$

or

$$P(Y_0 = 1|A = 0) \neq P(Y_0 = 1)$$

However, Y_1 is not observed for the unexposed $(A = 0)$, and Y_0 is not observed for the exposed $(A = 1)$. Thus, **the observed data can never tell us whether we have exchangeability or not**, or whether we have unmeasured confounding.

To judge whether exchangeability is plausible, we must rely on subject matter knowledge.

Conditional exchangeability

Adjusting for a potential confounder L produces a causal effect **if L is sufficient for confounding control**.

Technically, if we have conditional exchangeability, given L:

$$\underbrace{P(Y_1 = 1|A = 1, L)}_{=P(Y=1|A=1,L)} = P(Y_1 = 1|L)$$

$$\underbrace{P(Y_0 = 1|A = 0, L)}_{=P(Y=1|A=0,L)} = P(Y_0 = 1|L)$$

$$(Y_0, Y_1) \perp\!\!\!\perp A|L.$$

Conditional exchangeability cannot be tested, and must be judged by subject matter knowledge. Exchangeability can be achieved by adjustments, but can also be ‘destroyed’.

Average treatment effect

For dichotomous outcomes, the risk equals the average in the population, and we can rewrite the definition of association as

$$E[Y|A=1] \neq E[Y|A=0].$$

Under exchangeability, the **average treatment effect (ATE)**

$$E[Y_1] - E[Y_0]$$

can be estimated by its associational counterpart

$$E[Y|A=1] - E[Y|A=0].$$

Propensity scores

One of the most commonly employed parametric tools for causal inference from observational studies, introduced by Rosenbaum and Rubin in 1983.

Let $A \in \{0, 1\}$ be a dichotomous indicator of control/treatment. Now let

$$\pi(L) = P(A = 1|L)$$

denote the probability of being assigned to treatment given a set L of known covariates. $\pi(L)$ is referred to as a **propensity score (PS)**.

$\pi(L)$ measures the *propensity* to receive treatment, given the information available in L.

PS as balancing scores

PS is the simplest example of a **balancing score**. More generally, a balancing score $b(L)$ is any function of L such that

$$A \perp\!\!\!\perp L | b(L)$$

meaning that, for each value of $b(L)$, the distribution of the covariates L is the same in the treated and untreated.

Note: PS only balance wrt to measured L, not preventing residual confounding by unmeasured factors.

Positivity

The **positivity** assumption relates to the probability of being assigned to treatment/not treatment never being 0 or 1.

More formally, we say that we have **positivity** if, for all values l of covariates L in a population of interest

$$0 < P(A = 1 | L = l) < 1$$

holds true. Sometimes, positivity is referred to as **overlap** or **common support**.

Positivity is important, as it safeguards from conditioning on null-probability events.

Consistency

This assumption relates to the potential outcomes being well defined.

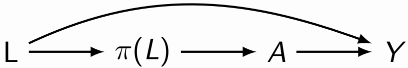
We say that we have **consistency** if assignment to treatment $A = a$ implies that the observed outcome Y equals the potential outcome Y_a

$$A = a \implies Y = Y_a.$$

A typical case where consistency can be violated, is when there are multiple versions of the same treatment.

Why are propensity scores useful?

In their seminal paper, Rosenbaum and Rubin proved that if it is sufficient to adjust for L, then it is sufficient to adjust for $\pi(L)$. Graphically, in DAG terms:



and adjusting for $\pi(L)$ blocks all paths from A to L, so that

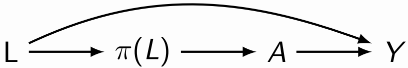
$$A \perp\!\!\!\perp L | \pi(L)$$

or, in other terms, treatment allocation is conditionally independent of pre-treatment covariates given $\pi(L)$.

Sounds familiar? This is similar to what happens in completely randomised designs!

Why are propensity scores useful?

In their seminal paper, Rosenbaum and Rubin proved that if it is sufficient to adjust for L, then it is sufficient to adjust for $\pi(L)$. Graphically, in DAG terms:



and adjusting for $\pi(L)$ blocks all paths from A to L, so that

$$A \perp\!\!\!\perp L | \pi(L)$$

or, in other terms, treatment allocation is conditionally independent of pre-treatment covariates given $\pi(L)$.

Sounds familiar? This is similar to what happens in completely randomised designs!

Propensity scores - trials and observational studies

In an ideal randomised trial in which half of the individuals are randomly assigned to treatment $A = 1$, $\pi(L) = 0.5$ for everyone, regardless L.

In observational studies, some might be more likely than others to receive treatment. Moreover, treatment status is *observed*, and thus beyond the control of the investigator.

For these reasons, the true PS are unknown, and need to be estimated from the data.

Propensity scores - trials and observational studies

In an ideal randomised trial in which half of the individuals are randomly assigned to treatment $A = 1$, $\pi(L) = 0.5$ for everyone, regardless L.

In observational studies, some might be more likely than others to receive treatment. Moreover, treatment status is *observed*, and thus beyond the control of the investigator.

For these reasons, the true PS are unknown, and need to be estimated from the data.

How to compute propensity scores

Being probabilities for a dichotomous outcome (treated/not treated) expressed as a function of a set of covariates, PS are typically estimated via logistic regression:

$$\text{logit}\widehat{\pi}(L) = \ln \frac{\widehat{\pi}(L)}{1-\widehat{\pi}(L)} = \widehat{\alpha} + \widehat{\beta}L$$

where $\widehat{\alpha}, \widehat{\beta}$ are model coefficients estimates.

Other approaches to estimating propensity scores can (and sometimes should) be used, examples include nonparametric methods, such as regression trees.

Quality assessment

Regardless of the method used to obtain them, the propensity scores should be inspected to assess whether they can be useful or not.

Specifically, we will be looking at the following two desirable aspects:

- the distribution of PS in the treated and not treated groups overlaps
- there is balance of covariates across blocks of propensity scores.

More on this later.

How to use

Under the assumption of exchangeability and positivity within levels of $\pi(L)$, PS can be used to estimate causal effects in a number of ways, including:

- stratification
- inverse probability weighting
- regression modelling
- matching.

Stratification

The conceptually simplest way to adjust for a potential confounder L is by **stratification**:

- the study population is partitioned into strata (groups), one for each level of L
- each stratum is analysed separately
- within strata, no variation in $L \implies$ no imbalance in L across exposure levels.

Stratification

For a dichotomous Y, the ATE among individuals with a particular value s of the propensity score $\pi(L)$

$$E[Y_1|\pi(L) = s] - E[Y_0|\pi(L) = s]$$

can be estimated, under the assumptions, via

$$E[Y|A = 1, \pi(L) = s] - E[Y|A = 0, \pi(L) = s].$$

Stratification - remarks

As $\pi(L)$ is generally a continuous variable in $(0, 1)$, it is unlikely that two individuals will share the exact same value for s of the true propensity score.

One possible approach is to create *strata* that contain individuals with similar, but not necessarily identical, values of $\pi(L)$.

A common choice is to use quantiles of the distribution of the estimated PS $\widehat{\pi}(L)$ to define the stratification, and then compute the causal effect within each of the strata.

Stratification - remarks

Stratification on quantiles or other functions of the PS raise a potential problem:

- in general, the distribution of $\pi(L)$ will differ, within some strata, between treated and untreated. This might lead to non-exchangeability in some of the strata.

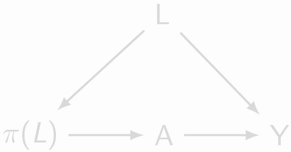
- The validity of our inference depends on the correct specification of the relationship between $\pi(L)$ and Y (which we are assuming to be linear in the simplest case).

However, as $\pi(L)$ is unidimensional, it is usually straightforward to safeguard against misspecifications by considering more flexible models (e.g. including spline terms).

Inverse probability weighting

The idea underlying **inverse probability weighting (IPW)** is to create a **pseudo-population** where the equivalence between association and causation holds.

We want to block the path from L to A by conditioning on the PS. IPW achieves this by weighting the observations with $\widehat{\pi}^{-1}(L)$.

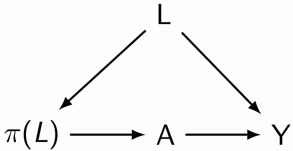


In practice, we under- and over-represent units based on their (estimated) probability of being treated, in the hope of obtaining a balanced version of the population.

Inverse probability weighting

The idea underlying **inverse probability weighting (IPW)** is to create a **pseudo-population** where the equivalence between association and causation holds.

We want to block the path from L to A by conditioning on the PS. IPW achieves this by weighting the observations with $\widehat{\pi}^{-1}(L)$.



In practice, we under- and over-represent units based on their (estimated) probability of being treated, in the hope of obtaining a balanced version of the population.

IPW - estimation

We can write the IPW estimator for $E(Y_a)$ as follows:

$$E[Y_a] = E\left[\frac{\mathbf{1}(A=a)Y}{P(A=a|L)}\right]$$

where $\mathbf{1}(A=a)$ is the indicator function that takes on value 1 if $A=a$, 0 otherwise.

We can then rewrite the expression for the ATE as follows:

$$E[Y_1] - E[Y_0] = E_{A=1}\left[\frac{Y}{\pi(L)}\right] - E_{A=0}\left[\frac{Y}{1-\pi(L)}\right]$$

where $E_{A=a}$ indicates that the expectation has been taken over those units with treatment level a.

Matching by propensity scores

Matching attempts to create a population where treated and untreated are exchangeable because they have the same distribution of $\pi(L)$.

Each treated individual is paired with one (or more) untreated individual with the same PS value. The subset comprised of the treated-untreated pairs is referred to as the **matched population**.

Under the assumptions, we recover the association $=$ causality equivalence and can estimate causal effects.

Matching - remarks

Once again, it is unlikely that two individuals will share the same value s of the propensity score, so the matching is usually approximate.

A general algorithm for PS matching would be:

1. estimate the $\pi(L)$ for all individuals in the study
2. for each treated, find one (or more) untreated with nearest $\widehat{\pi}(L)$ value
3. remove (prune) non-matched individuals.

How close units should be in terms of $\widehat{\pi}(L)$ to be matches is usually called the *caliper*.

Propensity scores matching - some criticism

Some authors maintain that propensity scores should not be used for matching, in that doing so would be sub-optimal with respect to the final inferential objective.

The two main points underlying the criticism are:

- PS matching aims to approximate a completely randomised trial, as opposed other methods that try to recreate a (more efficient) fully blocked design
- even in best-case scenario (all $\pi(L) = 0.5$) PS matching is sub-optimal, and increases model dependence and bias.

For a more complete discussion on the risks of matching by PS, refer to this excellent presentation by American political scientist Gary King and references therein.

Example - pneumonia and statin

To exemplify the propensity scores-based methods, we will use a dataset containing information on 7265 patients who had a diagnosis of incident pneumonia.

The treatment was statin $(A = 1)$ as opposed to no statin $(A = 0)$. The available covariates L include:

- some basic demographic information
- co-morbidities
- prior use of other medications.

The outcome Y is death within 6 months following diagnosis, and we are interested in the causal OR.

Variables in the dataset

variable	description
id	patient ID
statin	treatment: prescription of statin (1=yes, 0=no)
death	outcome variable: death within 6 months (1=yes, 0=no)
agecat	categorised age at pneumonia diagnosis
male	sex (1=male, 0=female)
smoke	smoking status (1=never, 2=ex, 3=current, 4=unknown)
alcohol	alcohol consumption (1=no, ..., 6=excessive, 7=unknown)
bmi	Body Mass Index (kg/m2; 1=<20, 2=20-25, 3=>25, 4=unknown)
diabetes	prior diagnosis of diabetes
cvd	prior diagnosis of cardiovascular disease
heartfail	prior diagnosis of heart failure
dementia	prior diagnosis of dementia
cancer	prior diagnosis of cancer
hyperlipid	prior diagnosis of hyperlipidaemia
aspirin	history of use of aspirin

Table 1: Variables included in the pneumonia dataset

Exploratory analysis

A preliminary look at the data reveals that

- approximately 14% of the sample has been prescribed statins ($A = 1$)
- as a crude measure of association, the OR of dying under statin is $OR \approx 0.64$
- patients with a diagnosis of diabetes or hyperlipidaemia are more likely to have a statin prescription
- age is fairly similar across treatment groups
- slightly fewer males receive statins.

Propensity scores estimation and assessment

We follow the procedure outlined earlier:

1. estimate PS by regressing treatment status on the available covariates
2. check the distribution of the estimated $\widehat{\pi}(L)$ by treatment group
3. check balance of covariates across blocks of PS
4. apply a PS-based adjustment methods.

Estimation

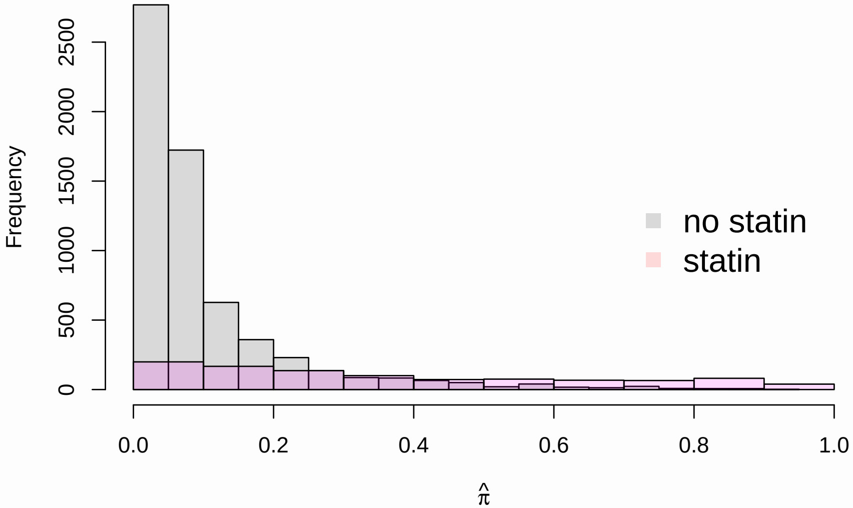
The functional form of the model we fit to obtain the estimates of the PS is

$$\begin{aligned}\text{logit}P(\texttt{statin}=1) &= \alpha_0+\alpha_1\texttt{agecat}+\alpha_2\texttt{male}+\alpha_3\texttt{smoke}+\alpha_4\texttt{bmi}+\alpha_5\texttt{alcohol}\\ &+\alpha_6\texttt{diabetes}+\alpha_7\texttt{cvd}+\alpha_8\texttt{heartfail}+\alpha_9\texttt{dementia}+\\ &+\alpha_{10}\texttt{cancer}+\alpha_{11}\texttt{hyperlipid}+\alpha_{12}\texttt{aspirin}\end{aligned}$$

A quick inspection of the estimated $\widehat{\pi}$ reveals (unsurprisingly, given the low number of treated), that the distribution is skewed towards smaller probabilities:

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0006338 0.0338194 0.0636313 0.1377839 0.1550677 0.9651635
```

Distribution of PS by treatment group



Balance of individual covariates by PS blocks

Checking for balance helps ensuring that the PS distribution is similar across groups. A few points:

- the initial distribution is likely to be imbalanced
- no agreed-upon best method of balancing the propensity score
- imbalance in the PS mean indicates the propensity scores need to be respecified
- even if balance in the mean, no indication about higher order moments.

Checking for balance is complex (and beyond the scope of this course). For those interested, start from Rosenbaum and Rubin (1985) and Austin (2009).

Stratification

Let $S_1, \ldots, S_5$ indicate the strata defined by the vigintiles of the PS distribution. The strata-specific causal OR can now be obtained via their associational counterparts:

Table 2: Strata-specific OR estimates with 95% CI

	estimate	95% CI
S_1	0.508	(0.147,1.352)
S_2	0.573	(0.218,1.247)
S_3	0.640	(0.295,1.226)
S_4	0.772	(0.480,1.195)
S_5	0.761	(0.567,1.017)

Stratification - remarks

- Table 2 has been obtained by running one logistic regression per stratum
- to obtain the ATE, we take a weighted average of the stratum-specific log-OR estimates and then exponentiate (for reference, this results in $OR \approx 0.642$)
- in order to account for the added uncertainty due to estimation of the PS, resampling methods can (and should) be used to construct confidence intervals (beyond the scope of this course).

Inverse probability weighting

Under IPW, the weights can be obtained by by taking the inverse of the estimated propensity scores

$$w = \frac{1}{\widehat{\pi}(L)}.$$

It is good practice to examine the distribution of IPW weights for extreme values:

Table 3: Quantiles of IPW weights by treatment group

	min	1%	5%	95%	99%	max
statin	1.036	1.064	1.119	25.868	48.464	104.841
no statin	1.001	1.004	1.010	1.581	2.828	13.302

IPW estimation

We have learned about the expression for the ATE under IPW in terms of risk difference:

$$E[Y_1] - E[Y_0] = E_{A=1}\left[\frac{Y}{\pi(L)}\right] - E_{A=0}\left[\frac{Y}{1-\pi(L)}\right]$$

To obtain an estimate of the odds ratio, we can resort to logistic regression. Specifically, we fit the following functional form:

$$\text{logit}\, P(\texttt{death} = 1|\texttt{statin}) = \alpha_0 + \alpha_1 \texttt{statin}$$

and specify $w = \widehat{\pi}^{-1}(L)$ to be used as weights (most statistical packages allow this).

The estimated odds ratio can then be obtained as $\widehat{OR} = e^{\widehat{\alpha}_1} \approx 0.633$ (see next slide).

IPW - model output

```
##
## Call:
## glm(formula = death ~ statin, family = "binomial", data = pneumonia,
##     weights = ps_weights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.2802  -0.6798  -0.6530  -0.6411  15.1454
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.49326    0.03032 -49.250   <2e-16 ***
## statinStatin -0.45717    0.04713  -9.701   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12313  on 7264  degrees of freedom
## Residual deviance: 12218  on 7263  degrees of freedom
## AIC: 11802
##
## Number of Fisher Scoring iterations: 5
```

Matching

Following the general algorithm outlined earlier for PS matching, we have to:

1. estimate the PS for all individuals in the study **[done]**
2. for each treated, find one (or more) untreated with nearest $\widehat{\pi}(L)$ value
3. remove (prune) non-matched individuals.

Finally, we can proceed to analyse the resulting matched dataset as we did previously.

Matching treated with untreated

This can be done using virtually any algorithm able to compute a distance and pair closest units with each other. Ad-hoc packages exist, such as `MatchIt` (Ho et al, 2011) in `R`, however they will tend to request the same key information:

- estimated PS and treatment indicator
- type of distance to be used (euclidean, mahalanobis, . . .)
- a *caliper*, or maximum distance to declare two units a match
- a *ratio* of how many untreated should be matched to each treated.

A new dataset can then be obtained, typically a subset of the original one, where treated individuals have one or more PS-matches in the untreated group.

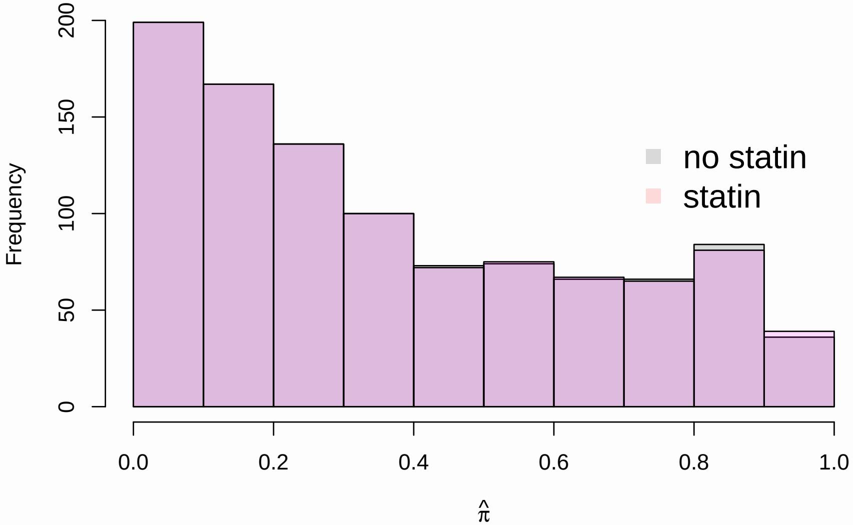
Exploring the matched dataset

As the PS approach strives to balance (observed) covariates between treatment groups, it is good practice to carry out interim checks that this is happening.

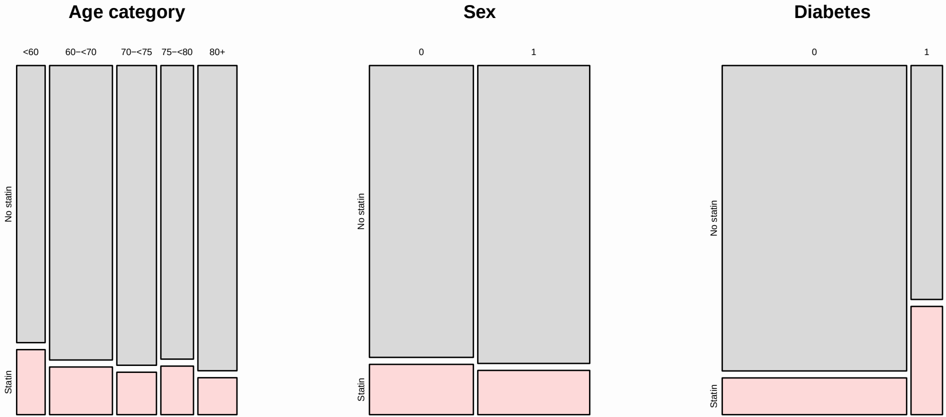
Let's take a look at the dataset resulting from using the a nearest-neighbour algorithm with a caliper of 0.2, and a treated/untreated ratio of 1.

The new dataset comprises 2002 individuals (half treated, half untreated).

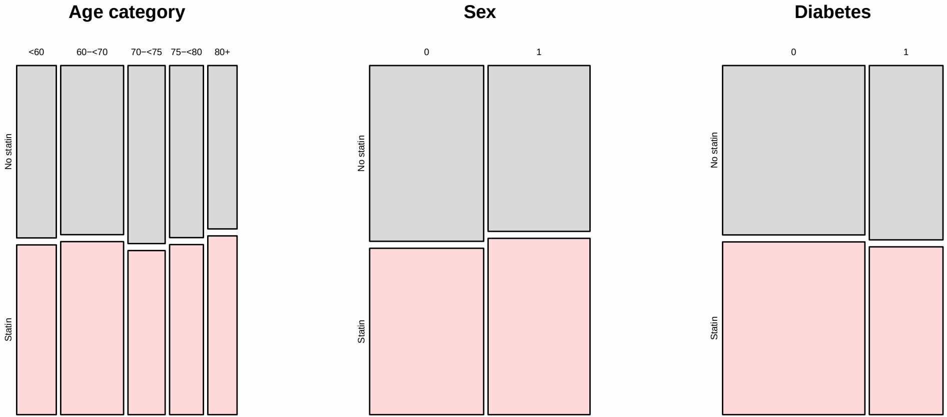
PS distribution by treatment status



Covariate balance by treatment status - original dataset



Covariate balance by treatment status - matched dataset



Age category

Sex

Diabetes

Estimation

```
## Call:
## glm(formula = death ~ statin, family = binomial, data = pneumonia_matched)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5615  -0.5615  -0.5253  -0.5253   2.0243
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.76749    0.08955 -19.738   <2e-16 ***
## statinStatin -0.14348    0.13007  -1.103     0.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1602.2  on 2001  degrees of freedom
## Residual deviance: 1601.0  on 2000  degrees of freedom
## AIC: 1605
```

The above results in a point estimate $\widehat{OR} \approx 0.866$. CI can and should be obtained via robust estimators or computationally intensive methods.

Final remarks

Propensity scores are a popular way to adjust for (observed) confounding and obtain causal estimates. A few points to wrap-up:

- aim is to approximate a completely randomised design
- many possible approaches, both parametric and nonparametric
- tricky to quantify uncertainty around estimates
- will **not** safeguard against unobserved confounding
- limited applicability to complex longitudinal data (time-varying treatments).

References

- Austin PC (2009). Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples. Statistics in Medicine, 28:3083—107.

- Austin PC (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res, 46(3):399–424. doi:10.1080/00273171.2011.568786.

- Dehejia RH and Wahba S (1999), Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94(448):1053–1062.

- Garrido MM, Kelley AS, Paris J, et al (2014). Methods for constructing and assessing propensity scores. Health Serv Res, 49(5):1701–1720. doi:10.1111/1475-6773.12182.

References

- Hernán MA and Robins J (2021). Causal Inference - What If. CRC Press url, book page.

- Ho DE, Imai K, King G and Stuart EA (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, 42(8), 1–28. https://www.jstatsoft.org/v42/i08/.

- Imai K and Ratkovic M (2014). Covariate balancing propensity scores. J R Stat Soc B, 76(1):243–263.

- Rosenbaum PR and Rubin DB (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70(1):41–55.

- Rosenbaum PR and Rubin DB (1985). Constructing a Control Group using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician 39 (1):33—8.