Statistical Modeling and Causal Inference with R

Week 6: Matching

Manuel Bosancianu

Max Schaub

October 12, 2020

Hertie School of Governance

Today's focus

- ✓ Subclassification: the mother of all matching methods
- ✓ Matching metrics and their application
- ✓ Common matching algorithms

Subclassification: the mother of all matching methods

Subclassification and the independence assumption

Data from study on internet use (0,1) on right-wing support (0,1), n=100

Raw numbers

Internet use	Right-wing support		
	yes	no	total
yes	40	33	73
no	11	16	27

Shares

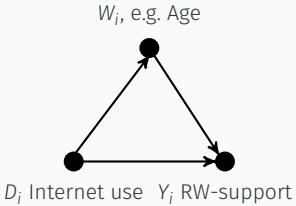
Internet use	Right-wing support
yes	0.55
no	0.41

Data on internet use and right-wing support (simulated)

$$NATE = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] = 0.55 - 0.41 = 0.14$$

Subclassification as an answer to confounding

But what about confounders, for example the age of the respondent?



Subclassification as an answer to confounding

Looking at the share of internet users by age group, we can observe stark differences.

Share internet users by age	
Aged 20-40	0.9
Aged 41-60	0.8
Aged 60+	0.5

Thus, the treatment is clearly not independent of age – the independence assumption $Y_{i0}, Y_{i1} \perp\!\!\!\perp D_i$ is violated.

Subclassification as a way to create conditional independence

From previous weeks we know that we can condition on a known confounder to close the back door path.

In other words, we can make use of the *conditional* independence assumption (CIA): $Y_{i0}, Y_{i1} \perp\!\!\!\perp D_i | W_i$.

One way of doing this is by using a regression model in the form $Y_i = \alpha + \beta D_i + \gamma W_i + u_i$.

We derived mathematically how regression can be used to control for confounders.

Here we will discuss other methods to achieve the same result.

Looking for twins

Rather than invoking regression mechanics to achieve *ceteris paribus*, we start more simply: by comparing like and like.

The aim is to find units that are similar in all respects but the treatment. We are looking for twins that are identical in all respects save for the fact that one received the treatment, but the other did not.

A first, most simple method is subclassification. The idea is to split our sample in strata of the potential confounder, and then to only compare treatment effects *within* these strata. Consider our example:

Raw numbers

Internet use	Right-wing support		
	yes	no	total
yes	40	33	73
no	11	16	27

Conditioning using subclassification

Subclassification gives us the following frequency table:

Numbers of right-wing supporters by internet use and age class						
Internet use	Aged 20–40 Right-wing support		Aged 41–60 Right-wing support		Aged 60+ Right-wing support	
	yes	total	yes	total	yes	total
yes	9	18	28	40	3	15
no	1	2	7	10	3	15
Shares						
yes	0.5		0.7		0.2	
no	0.5		0.7		0.2	

This allows us to compare rightwing support for individuals aged 20-40 that use and do not use the internet, for individuals aged 41-60 that use and do not use the internet, ...

We can see that when comparing right-wing support between ‘treated’ and ‘untreated’ units, the shares of right-wing supporters are actually identical within each of the three age-subgroups!

Conditioning using subclassification

In other words, $NATE_{20-40} = NATE_{41-60} = NATE_{60+} = 0$.

It follows (in this example) that the weighted overall, adjusted NATE is also zero:

$NATE_{overall} = NATE_{20-40} \times \frac{20}{100} + NATE_{41-60} \times \frac{50}{100} + NATE_{60+} \times \frac{30}{100} = 0$.*

* Note that the same result could be recovered with a regression model in the form RW-support $= \alpha + \beta$InternetUse $+ \gamma_1$Age$_{41-61} + \gamma_2$Age$_{61+} + u_i$.

Conditioning using subclassification

By subclassifying our data into strata and assuming that age is the only confounder we have satisfied the backdoor criterion, and $NATE_{overall}$ estimates the *ATE*. The estimated effect of internet on right-wing voting is zero.

But what about other confounders, e.g. gender, which is also plausibly related to internet usage and right-wing attitudes?

How can we use subclassification be used to condition on this second variable?

Subclassification with many dimensions

Introducing another dimension along which to classify increases the number of cells to consider twofold. There are now $2 \times 3 \times 2 = 12$ cells (of which 8 are shown in the table):

Numbers of right-wing supporters by age class and gender

Internet use	Aged 20–40				Aged 41–60				...
	Male RW-support		Female RW-support		Male RW-support		Female RW-support		...
	yes	total	yes	total	yes	total	yes	total	...
yes	2	5	2	4	15	23	10	17	...
no	–	–	1	1	2	5	10	10	...

There are several noteworthy observations in this table.

The curse of dimensionality

Already with this limited number of dimensions, we run into trouble! Note the circled cells:

There are no men aged 20–40 that do not use the internet! The respective cells are not populated. In other words, there are no observations that are equal with regard to gender and age that we can compare these men to. They lack a counterfactual.

This is the *curse of dimensionality* biting – the more variables we try to condition, the more cells we create, and the more likely it becomes that – with limited sample size – cells will not be populated.

If we wanted to control for additional variables like education, the problem would worsen. We would now need to create a table with $2 \times 3 \times 2 \times 3 = 36$ cells – of which many would surely be empty.

Common support (and lack thereof)

With such missing data, our subclassification scheme is no longer fully defined.

We say that the problematic cells lack common support.

Common support requires that for each strata, there exist observations in both the treatment and control group.

Without common support, we cannot calculate the group-specific treatment effects, and the ATE can no longer be estimated.

While subclassification therefore in principle allows for establishing conditional independence, its application is limited to simple cases.

One-sided lack of common support and the ATT

It matters which information is missing. Once any cell is not fully populated, we can no longer use subclassification to estimate the ATE.

However, as long as we have populated cells in the control condition, we can still estimate the average treatment effect for the treated, the $ATT = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$.

Because estimating the ATT requires less complete data, estimating the ATT is often more feasible, especially in cases where you have few treatment observations, but many control observations.

However, in our example, a control cell is empty. In this case, subclassification is 'broken'.

Matching metrics and their application

Matching philosophy

Matching techniques losen the requirement that each cell should have counterfactual in the respective control cell.

Instead, matching algorithms try to find the most plausible counterfactual within the data for a given observation in a treatment cell.

Where exact control cells are available, these exact matches will usually be chosen.

Where this is not the case, matching algorithms minimize some metric defining the distance between treatment and control cells, and choose control units that are closest to a given treatment unit lacking an exact match.

If there is more than one matched unit at the same distance, matching algorithms usually use the average of all matched units.

Choosing the best counterfactual

In our original example, we might for example match the young, internet-using men who lack a direct counterfactual (marked in orange) with a) women of the same age group, or b) men aged 41-60 (both marked in blue).

Numbers of right-wing supporters by age class and gender

Internet use	Aged 20–40				Aged 41–60				...
	Male RW-support		Female RW-support		Male RW-support		Female RW-support		...
	yes	total	yes	total	yes	total	yes	total	...
yes	2	5	2	4	15	23	10	17	...
no	–	–	1	1	2	5	10	10	...

Both are plausible counterfactuals that are ‘one step away’.

Matching algorithms will usually determine which of the two dimensions – gender or age – should be given more weight in the distance matrix, and will choose the match accordingly.

Or they will define some virtual counterfactual that mixes the outcomes for both plausible matching cells.

Basic matching estimators

Let's look again, in a more technical way, at the estimands that we are typically trying to recover from the data: the ATT, and the ATE, and how matching tries to approximate them.

As mentioned, the difference between the ATT and ATE from a matching perspective is that for the ATT, we only need to find counterfactuals for the treated units, while for the ATE, we also need counterfactuals for control units with missing treatment matches (Cunningham, 2021, 119).

Basic matching estimators

The matching estimator for the ATT is

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}) = \frac{1}{N_T} \sum_{D_i=1} (Y_i - [\frac{1}{M} \sum_{D_i=1} Y_{j_m(i)})$$

That is, our estimate for the ATT is the difference between the outcomes Y_i of all N_T treated units minus the outcome for the matched closest units $Y_{j(i)}$

If there is more than one match, the average oucome of the matched units s is used as a counterfactual. That's what the term $\frac{1}{M} \sum_{D_i=1} Y_{j_m(i)}$ means.

Basic matching estimators

The matching estimator for the ATE is

$$\hat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)(Y_i - Y_{j(i)}) = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)[Y_i - (\frac{1}{M}\sum_{D_i=1} Y_{j_m(i)})]$$

Different from the estimator for the ATT, this also finds matches for observations in the control condition, by having the term $2D_i - 1$ be 1 if $D_i = 1$ and -1 if $D_i = 0$, which reversing the order of the outcomes.

Common matching algorithms

Common matching algorithms

There are many different matching techniques out there, and even more ways of implementing them. We here discuss three of the most fundamental and most widely used:

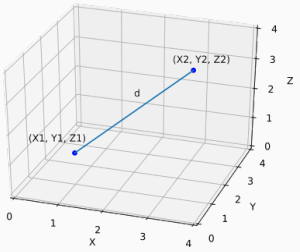
- ✓ Mahalanobis/nearest neighbor covariate matching
- ✓ Propensity score matching
- ✓ Coarsened exact matching

Another approach used in the social sciences is ‘genetic matching’ (Diamond & Sekhon, 2013), and there also exists a whole universe of matching algorithms used for automated text analysis (e.g. based on the Levenshtein distance).

Mahalanobis/nearest neighbor covariate matching

As stated, most matching techniques minimize a metric of the distance between treatment and control cells and choose control units that are closest to a given treatment unit lacking an exact match.

The most simple metric that can be minimized is the Euclidean distance $||W_i - W_j|| = \sqrt{\sum_{n=1}^{k} (W_{ni} - W_{nj})^2}$ – the shortest path between two points in n-dimensional space, or the closely related Mahalanobis distance.



Covariate matching in practice

In practice, when specifying a matching model, you will have to make have to make several decisions:

- ✓ The number of matches allowed
- ✓ The distance metric to use (e.g. Mahalanobis or Euclidean)
- ✓ Sampling with replacement or without, i.e. whether the same unit can serve as a match for only one or several units.
- ✓ Whether to enforce exact matching on some variables
- ✓ Whether to use a caliper – a maximum distance between treated and control units beyond which units are declared 'unmatchable', and are no further considered. The caliper defines which units are 'on common support' – here understood as reasonably close – or 'off common support' – no close enough. The process of discarding units off common support is also called 'trimming.'

How to make these decisions isn't fully defined, and a matter of trying out, reason, and experience. We will come back to practical aspects of matching towards the end.

Propensity score matching

Another very common metric used in matching (and other applications) is the propensity score.

The propensity score is defined as the probability of receiving the treatment given confounding variables:
Propensity score = $p(W) = Pr(D = 1|W)$.

Propensity score matching achieves a similar goal to nearest-neighbor matching on covariates, but in two steps steps. In a first step, all the potential confounder variables (that we usually refer to as W here) are used to estimate a single value: the probability or propensity to receive a treatment.

This is usually done with a maximum likelihood (logit or probit) model where we regress the treatment status D_i on the covariates. The predicted outcome from this regression – the predicted probability of receiving the treatment – is the propensity score.

Propensity score matching

A typical model used for estimating the propensity score would take the form

$Pr(D = 1|W) = \Phi(\beta \boldsymbol{W})$

which describes a standard probit model. $\boldsymbol{W}$ here is the matrix with all potential confounders, and Φ is the cumulative distribution function of the standard normal distribution.

The propensity score theorem holds that if the CIA $Y_0, Y_1 \perp\!\!\!\perp D|W$ applies, conditioning on the propensity score $p(W)$ will achieve conditional independence, i.e. $Y_0, Y_1 \perp\!\!\!\perp D|pr(W)$ (see: Cunningham (2021, 143)).

Propensity score matching

The propensity score thus sums up, in a single number, all the information contained in the control variables that are needed to make the potential outcomes independent from the treatment status!*

Instead of finding nearest neighbors with regard to all problematic covariates, we can find nearest neighbors in terms of the propensity score.

In so doing, all the practical issues mentioned above apply – especially regarding the use of the caliper/trimming – also apply here.

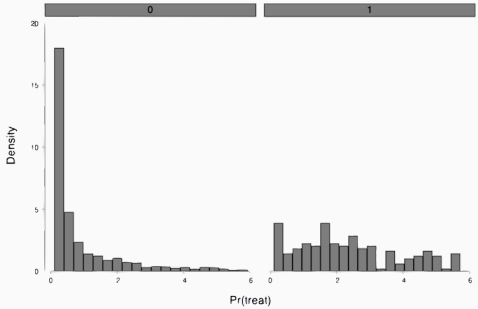
We can also use the propensity score for weighing our data – a method called inverse propensity score weighing (IPW).

* Note that this only holds, however, if your model for the estimating the propensity score is fully and correctly specified – and only your DAG can tell you which variables to include ...

Plotting propensity scores

The propensity score gives you a single number with a very intuitive interpretation: the probability of receiving treatment given the covariates.

It shows you how different treatment and control units in your data really are, and makes intuitively clear how much extrapolation your model will do.



A note on trimming

Trimming – discarding observations that are considered too different to have a plausible match in the data – is one of the most powerful aspects of the matching process.

It sets the matching logic apart from other conditioning methods such as regression, which usually uses all the data, and tends to 'create' counterfactuals even for observations that do not even have a 'remote cousin' in the data.*

It is by trimming that good balance on observable confounders can be achieved. A matched and trimmed dataset will often look like experimental data with regard to the observables (but of course not with regard to unobservables).

* But note that many matching methods will do the same by default.

A note on trimming

The estimand that can be estimated after trimming, appliying a caliper, or otherwise changing the sample is therefore no longer the ATT or ATE, however.

Because we have changed our sample, what we are estimating is the local ATT – the the treatment effect averaged over only the subset of treated units for which good matches exist among available controls.

Using the above matching techniques

The aim of a matching strategies discussed so far is to achieve balance on observable potential confounders. That is, after matching, the confounders should have the same average values in both the treatment and the control group.

In order to do so, in practice, several steps are taken (cp. Stuart (2010)):

1. Think carefully about the set of covariates to include in the matching procedure
2. Decide on the matching procedure (here assuming propensity score matching)
3. Specify the model used for matching, esp. the functional form of your covariariates (e.g. categorical, linear, quadratic, log), covariates to exactly match on, the maximum number of matches used to calculate the outcome etc.
4. Apply your matching algorithm
5. Inspect the propensity scores and consider common support: are the scores in treatment and control similar enough to justify using the whole sample/ estimate the ATT or ATE convincingly? If not, consider using a caliper.
6. Check your covariate balance in the matched sample. If balance is still poor, repeat steps 3 to 5.

Coarsened exact matching

A final matching method that has gained a lot in popularity in recent years is coarsened exact matching (CEM) (Iacus, King, & Porro, 2012).

Coarsened exact matching can be understood as a sophisticated form of subclassification. It proceeds in several steps:

1. In a first step covariates get split into temporary, coarse categories. For example, a continuous age variable might be split into three temporary age categories, while gender is split in two, and education in three categories.
2. Second, matches between treatment and control units are found within overlapping strata – just as in subclassification.
3. Third, data points that are alone in a class, i.e. that cannot be matched are 'pruned' – they are dropped from the data. Only treatment observations that have matches within their temporary strata are maintained.
4. Finally, the temporary categories are removed, i.e. the original data remains unchanged. Just the information which units do or do not have matches within their strata is maintained.

Coarsened exact matching

The pruned data set can then be analyzed with other statistical methods, like regression.

Coarsened exact matching gives up on trying to maintain a full sample – required to estimate ATT and ATE – but instead prioritizes establishing covariance balance.

CEM therefore paves the ground for estimating the *local* ATT, similar to nearest-neighbor or propensity score approaches with a caliper.

The aim of CEM is to achieve balance without making assumptions on the fuctional form as to how the covariates influence treatment assigment – a typical problem especially of propensity score matching.

Matching and regression

OLS regression can be seen as a matching technique that produces variance-weighted averages of treatment effects. See Angrist and Pischke (2009, 54ff.) for details.

It is therefore not too different from nearest-neighbor covariate or propensity score matching.

Arguably the main difference between the use of regression and that of matching is in in terms of research culture/approach.

While regression is often used mindlessly, when using matching, scholars much more explicitly discuss issues such as the how covariates should enter the model and to what extent common support holds.

If you practice the same care when specifying your regression model, the techniques will often give fairly similar results.

Regression is also commonly used after coarsened exact matching to further reduce the impact of potential confounders.

Matching, regression, and IV

While subclassification, matching, and regression share many similarities, they are quite different from both the experimental method and instrumental variables.

In experiments and IV strategies, we use random assignment to treatment to take care of imbalance – we are trying to create or isolate situations where the independence condition ($Y_{i0}, Y_{i1} \perp\!\!\!\perp D_i$) holds.

With subclassification, matching, and regression techniques we attempt to *explicity* balance the data. Rather than hoping for the independence assumption to hold thanks to (as-if) randomization, we are making our comparison work by invoking the *conditional* independence assumption $Y_{i0}, Y_{i1} \perp\!\!\!\perp D_i | W_i$.

Note that often these techniques are not mutually exclusive, but can (and often are) combined: in IV strategies, as-if random variation can be isolated *conditional on controls* implemented with regression or matching.

Thank you for watching, and see you next Monday!

References i

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Cunningham, S. (2021). *Causal Inference: The Mixtape*. New Haven, CT: Yale University Press.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932–945.

Iacus, S. M., King, G., & Porro, G. (2012, January). Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, *20*(1), 1–24.

Stuart, E. A. (2010, February). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, *25*(1), 1–21.