Matching

Christoph Hanck

Summer 2023

Matching

Definition: Matching

Matching is the process of closing back doors between a treatment and an outcome by constructing comparison groups that are similar according to a set of matching variables.

Matching is usually applied to binary treated/untreated treatment variables, where the control group is very similar to the treatment group.

Matching

Example: The effect of job-training on the chances of getting a good job

- The pool of unemployed people eligible for the job-training program was about 50% male and 50% female.
- People actually in the program were 80% male and 20% female due to heavily advertising to males.
- The back door: $Outcome \leftarrow Gender \rightarrow JobTrainingProgram$

Matching

Example: The effect of job-training on the chances of getting a good job

- For the matching approach, a **control group** is constructed such that it has 80% male and 20% female, to compare to the already 80-20 treated group.
- Now, the gender difference between the treated and untreated groups gets eleminated.
- This closes the back door.

Matching

Matching is a form of weighted averages

- Matching methods create a **set of weights** for each observation.
- The weights are designed to make the treatment and control groups comparable.
- To estimate the effect of treatment, a weighted mean of the outcomes for the treatment and control groups is calculated and compared.

Matching

Where do the weights come from?

- There are many different matching processes, each of which takes a different route to generating weights
- All approaches use a set of **matching variables** to construct a set of weights so as to close any back doors that those matching variables are on:

  The idea is to create a set of weights such that there is no longer any variation between the treated and control groups in the matching variables

Matching

Example: The effect of job-training on the chances of getting a good job

- Assume that 60 out of 80 men in the treated group end up with a job and 20 without. 12 out of 20 treated women end up with a job and 8 without.
- In the control group, out of 500 men, 350 end up with a job and 150 without. Out of 500 women, 275 end up with a job and 225 without.
- After raw comparison we get: $(60 + 12)/100 = 72\%$ of those with job training end up with jobs, while in the control group $(350 + 275)/1000 = 62.5\%$ end up with jobs.
- That is a treatment effect of 9.5 percentage points.

Matching

Example: The effect of job-training on the chances of getting a good job—ctd.

- There is a back door through $Gender$.
- The weights are constructed as:
  - Give a weight of 1 to *everyone* who is treated
  - Give a weight of $80/500 = 0.16$ to all untreated men
  - Give a weight of $20/500 = 0.04$ to all untreated women
- The treated group will still be 80% men and the untreated group becomes

$$\frac{(0.16 \cdot 500) + (0.04 \cdot 0)}{(0.16 \cdot 500) + (0.04 \cdot 500)} = 80\% \quad \Rightarrow \quad \text{backdoor closed!}$$

A single matching variable

Choices:

1. What will be the matching criteria?
2. Are matches being selected or a matched weighted sample constructed?
3. If matches were selected, then how many?
4. If a matched weighted sample is constructed, then how will the weights decay with distance?
5. What is the worst acceptable match?

Matching criteria

Distance Matching

Definition: Distance matching

Distance matching is matching based on the notion that observations are similar if they have similar values of the matching variables.

Distance matching

- Distance matching ensures that the treatment and control groups have very little variation in the matching variables between them. This **closes back doors**.
- The idea is to minimize the distance (in terms of of how far the covariates are from each other) **between the treatment observations and the control observations.**

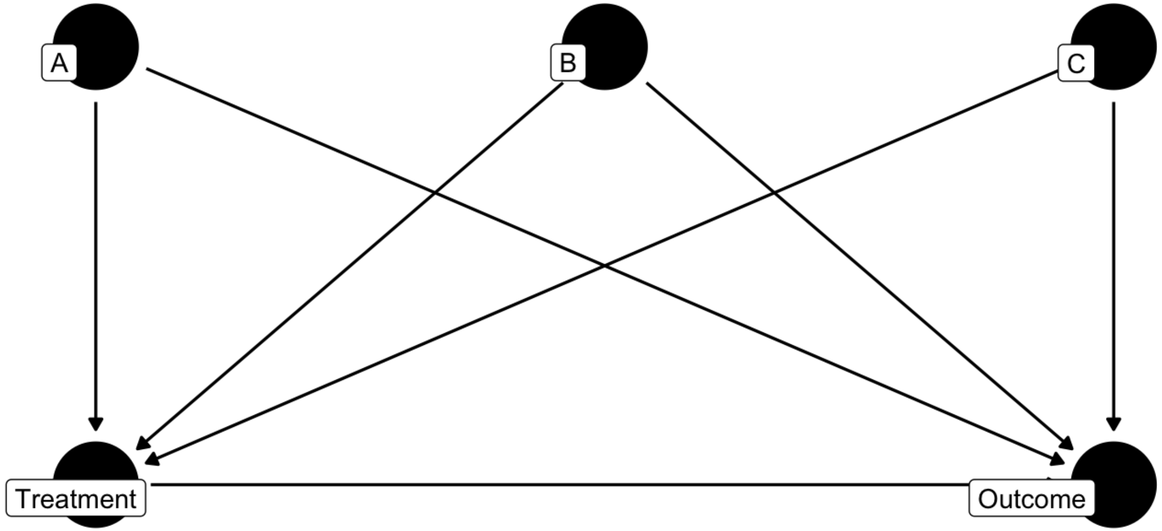

Figure 1: Causal Diagram

Distance matching

Example: Defaulting on credit card debt

- Data is on 30,000 credit card holders in Taiwan in 2005, their monthly bills, and their repayment status in April through September of that year
- We are interested in the effect of being late on the payment in April (treatment) on being late on the payment in September (outcome)
- An observation is chosen: a person having an April Bill of 91,630 New Taiwan dollars (NT$), their payment was late in April but their payment was not late in September.

Distance matching

Example: Defaulting on credit card debt—ctd.

- Here, the matching variable is Bill April.
- Untreated matching observations with Bill April values close to NT$91,630 should be looked for.
- A good choice of a single matching control observation for row number 10,305, would be row 27,281, which was not late in April (and so was untreated), and has a April bill of NT$ 91,609, (distance $|91639 - 91609| = 21$)
- This is the closest in the data to NT$ 91,630 among the untreated observations.

Matching criteria

Propensity score matching

Definition: Propensity score matching

Propensity score matching is a matching criterion based on the notion that *observations are similar if they were equally likely to be treated,* in other words have equal treatment *propensity*.

Propensity score matching

- Identifying the effect of the treatment on the outcome and the matching variables being on back doors are of interest.
- Propensity score matching takes this idea seriously and figures that if treatment propensity are being matched.

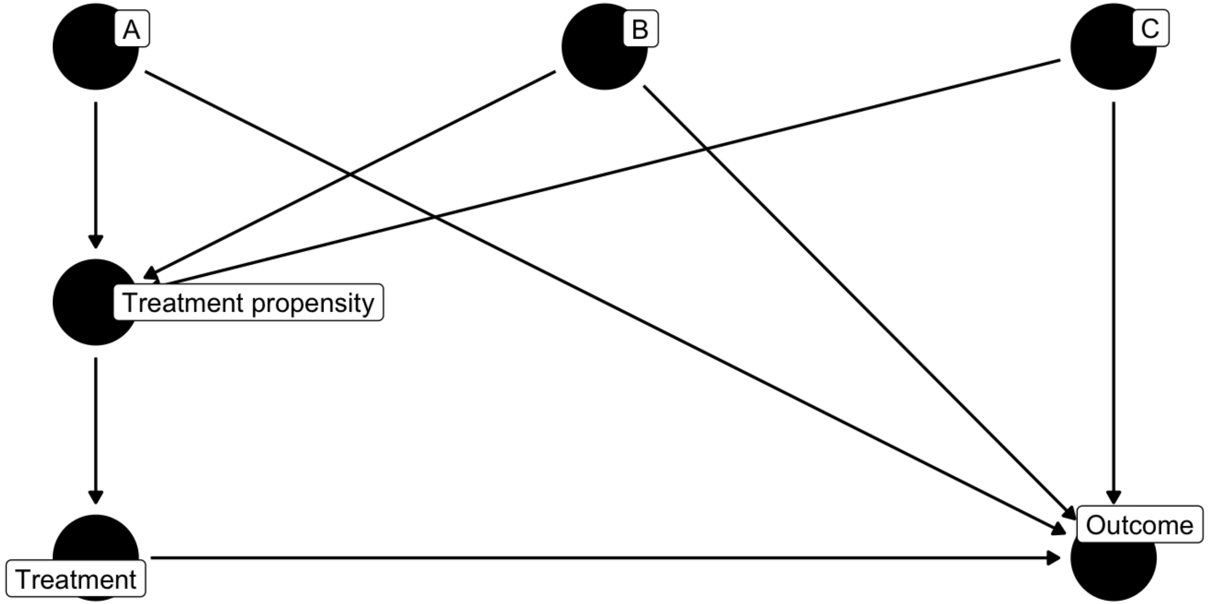

Figure 2: A DGP with treatment propensity

Propensity score matching

Example: Defaulting on credit card debt

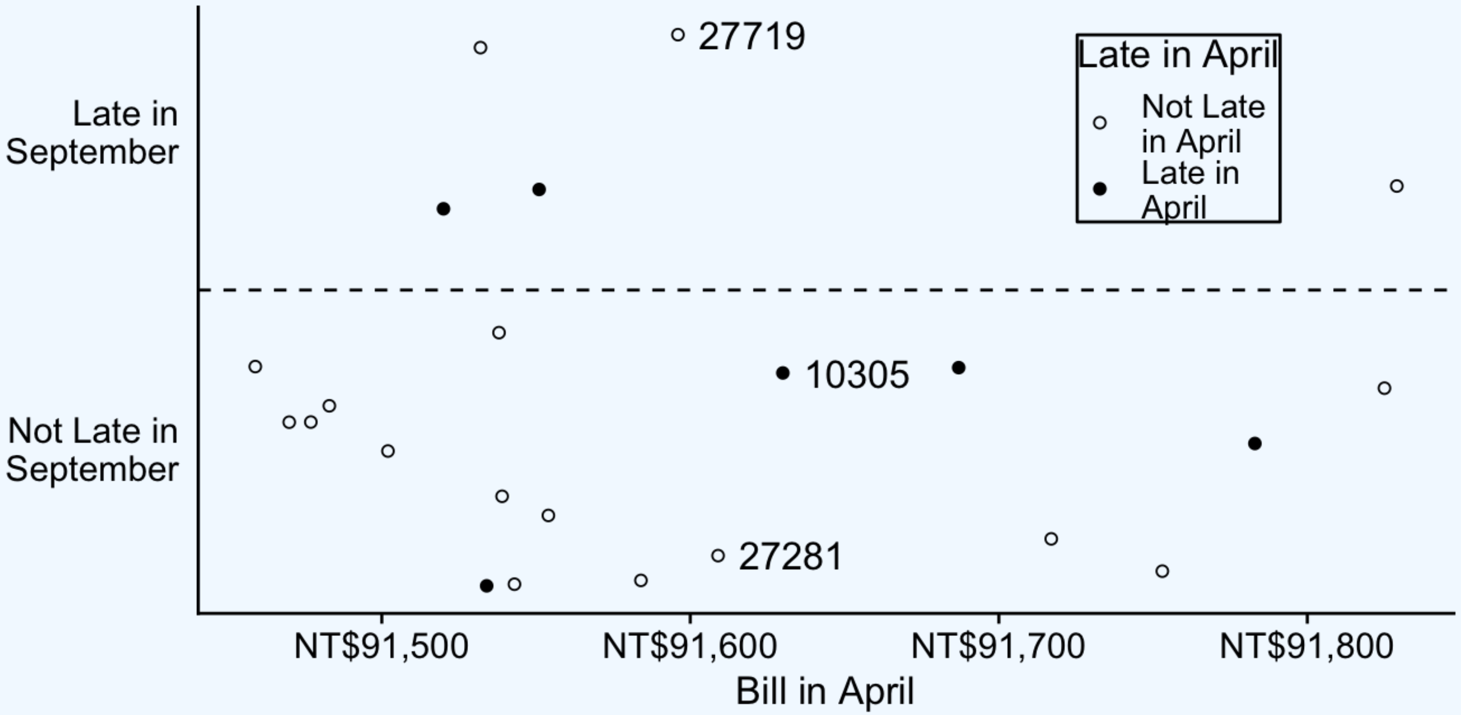

Figure 3: UCI Credit Card data

Propensity score matching

Example: Defaulting on credit card debt

- A logit regression of treatment on *Bill April* (thousands of NT$) yields an intercept of .091 and a coefficient on *Bill April* of .0003 can be obtained.
- If a match for the treated person in row 10,305 with a of NT$91,630 is searched for, then first, the probability of treatment for that person is predicted.
- Plugging in 91.630 for *Bill April* in thousands, a predicted probability of treatment of .116 can be obtained.

2. Matching or constructing a weighted sample?

It depends:

- The process of selecting matches means that control observations are picked to either be in or out of the matched control sample. If it is a good enough match, it will work, or else it will not.
- Everyone in the matched sample receives an equal weight
- Both approaches have their pros and cons and none of them dominates the other.

Selecting matches vs. weighted-group approach

Selecting-matches approach

- Easier to implement than methods with more fine-tuned weights. Also easier to interpret.
- Scenarios where one control observation is accidentally given an astronomical weight can be avoided

Weighted-group approach

- Less sensitive—“in” or “out” is a strong distinction, and so tiny changes in things like caliper or measurement error can change results significantly, making the selecting-matches approach a little noisier.
- Varying weights allow to account for the fact that different observations will simply have better or worse matches available in the data than others.
- Mainly used with propensity scores.

3. If matches were selected, then how many?

There are three main approaches to figure out how many control matches are to be picked for each treatment observation:

- To pick the one best match **(one-to-one matching)**

  The single best control match for each treated observation is picked.

- To pick the top k of best matches **(k-nearest-neighbor matching)**
- To pick all the acceptable matches **(radius matching)**

  (Acceptable matches are decided on. *All* the acceptable matches are matched.)

3. If matches were selected, then how many?

When to choose which approach?

It comes down to a tradeoff between bias and variance.

Example: One-to-one matching vs. 2-nearest-neighbors matching

The 2-nearest-neighbors match will include some observations in the control group that are not quite as closely comparable to the treated group.

- Both approaches are likely to be biased: we cannot claim confidently that all the back doors are closed.
- However, the more matches are there for each treated observation, the less noisy the estimate of the control group mean can be, and so the more precise the treatment effect estimate can be.

3. If matches were selected, then how many?

When to choose which approach?

Example: One-to-one matching vs. 2-nearest-neighbors matching—ctd.

- If there are 100 treated observations, the mean of 100 matched control observations will have a wider sampling distribution than the mean of 200 matched control observations.
- So, the choice of how many matches to do will be based on how important bias and precision are in the estimate, and how bad the matches will get if doing matches.

3. If matches were selected, then how many?

With replacement or without replacement?

- For matching without replacement, the control can only be a match for one of them, and the other treated observation needs to find someone different
- For matching with replacement, the same control can be used multiple times, giving it a weight equal to the number of times it has matched
- The choice between matching with replacement and without replacement hence also depends on bias variance tradeoff.

4. How will the weights decay with distance?

Kernel matching

Definition: Kernel Functions and kernel-based matching

- Kernel Functions are functions that a difference to is given, and it returns a weight. The highest value is at 0 (no difference), and then the value smoothly declines as the value move away from 0 in either direction
- Kernel-based matching estimators use a kernel function to produce weights.
- The weights given by the Kernel are used to calculate the weighted means.

4. How will the weights decay with distance?

Epanechnikov kernel

A frequently used kernel $K(x)$ is the *Epanechnikov* kernel: $K(x) = \frac{3}{4}(1 - x^2)$ on $[-1, 1]$, 0 else.

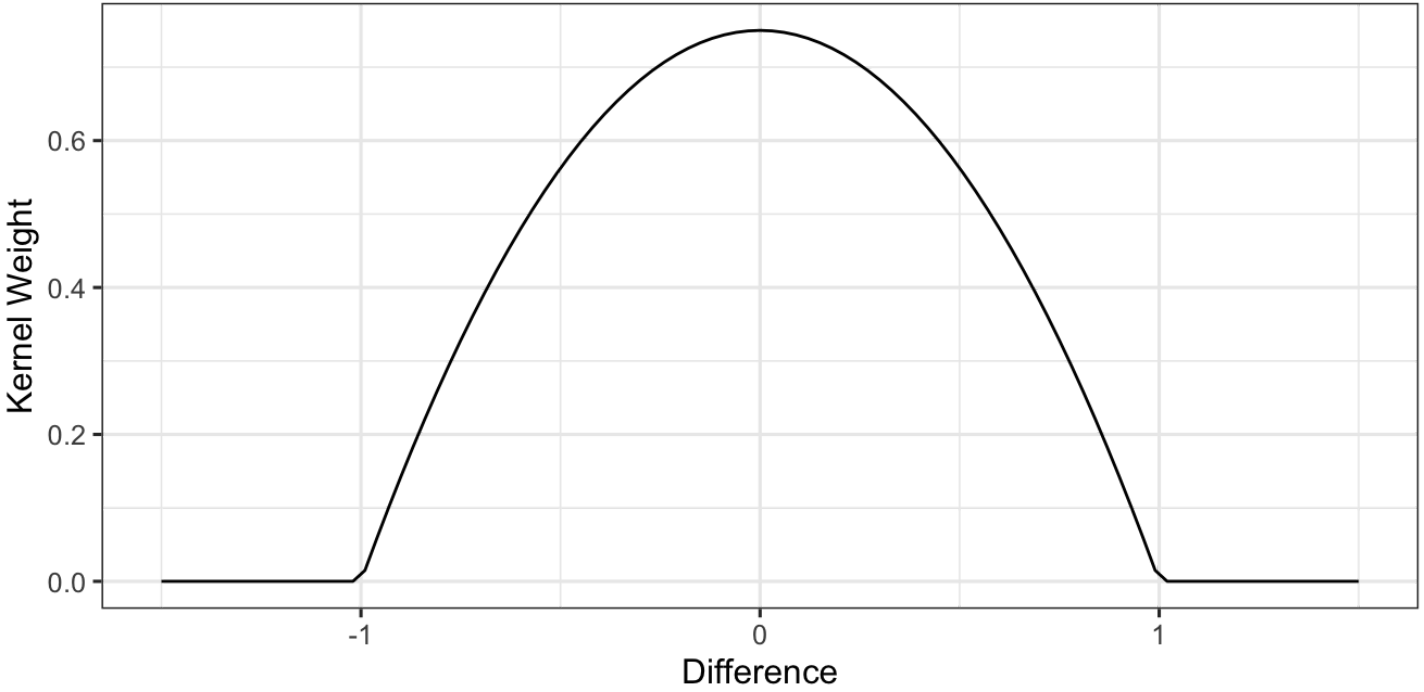

Figure 4: Epanechnikov kernel

4. How will the weights decay with distance?

Kernel matching

Example: Defaulting on credit card debt

Two differences are calculated:

- The difference between the $BillApril$ matching variable in rows 10,305 and 27,281 was $|91630 - 91609| = 21$.
- The difference between row 10,305 and row 27,719 was $|91630 - 91596| = 34$. These differences are standardized (why?) to $\frac{21}{59.994} = 0.35$ and $\frac{34}{59.994} = 0.56$.

4. How will the weights decay with distance?

Kernel matching

Example: Defaulting on credit card debt—ctd.

The Epanechnikov kernel gives weights for the standardized differences:

- $\frac{3}{4}(1 - 0.35^2) = 0.658$
- $\frac{3}{4}(1 - 0.56^2) = 0.515$

If the other controls are very close to each other, then the treated mean outcome $LateSept$ value of 0 (false) for row 10,305 is compared against the values of 0 for row 27,281 and 1 for 27,719, which are then averaged together with the above weights:

$$\frac{0.658 \cdot 0 + 0.515 \cdot 1}{0.658 + 0.515} = 0.561$$

The treatment effect ist $0 - 0.561 = -0.561$

4. How will the weights decay with distance?

Inverse probability weighting

- Inverse probability weighting is specially designed for use with a **propensity score**.
- It weights each observation by the inverse of the probability of the treatment value it had.
- The treatment group observations that get the **biggest weights are the ones that are most like the untreated group** and the ones with **small propensities are least likely to have gotten treated who got treated anyway.**
- Similarly, the control group observations with the biggest weights are the ones most like the treated group, who were most likely to have gotten treated but did not for some reason.

4. How will the weights decay with distance?

Inverse probability weighting

Example: Defaulting on credit card debt—ctd.

- The treated observation in row 10,305 had a .116 probability of treatment, as did the control observation in 27,281.
- Row 27,719, which has a propensity of .116 can be added as well.
- Row 10,305 is actually treated, so that observation is given a weight of 1 divided by the probability of treatment, or $\frac{1}{0.116} = 8.621$
- The control rows will both get weights of 1 divided by the probability of non-treatment, or $\frac{1}{1-0.116} = 1.131$.
- From here, weighted means on both the treated and control sides can be obtained.

5. What is the worst acceptable match?

Bandwidth selection

- Most approaches to matching use **caliper** or **bandwidth** to determine how far off a match can be before tossing it out.
- The main idea is:
    - First a number is picked which is caliper/bandwidth.
    - If the distance, or the difference in propensity scores, is bigger in absolute value than that number, then it is not counted as a match.
    - If a match is encountered, then that observation is lost.

5. What is the worst acceptable match?

Bandwidth selection

- Usually, the caliper/bandwidth is defined in terms of standard deviations of the value that is matching on rather than the value itself, to avoid scaling issues
- Some matching approaches end up using calipers/bandwidths naturally:

  Any kernel-based matching approach will place a weight of 0 on any match that is far away that the kernel function that is chosen sets it to 0.

  (For the Epanechnikov kernel that is a distance of 1 or greater)
- The wider the bandwidth, the more potential matches are obtained.
- However a wider bandwidth alows for more bad matches, which makes the match quality worse: the idea that the back doors are getting close less plausible!

5. What is the worst acceptable match?

Bandwidth selection

Summary: Bias variance tradeoff in matching

When making any sort of decision about matching, the choice is often between fewer, but better, matches that produce estimates with less bias but less precision, or more, but worse, matches that produce estimates with more bias but more precision.

Multiple matching variables

Key concept

- Take a lot of different variables and try to combine them to a single variable based on how *close* they are. For this we need a definition of *closeness*.
- Once a single variable obtained, matching can be done by previously discussed methods

Multiple matching variables

Example: Intrinsic movitations of American politicians

- **Research question:** Are black American politicians especially interested in supporting the black American community?
- **Experiment:** Broockman (2013) sent a lots of emails to (black and white) state legislators (politicians), simply asking for information on unemployment benefits.
  - Each email was sent by the fictional “Tyrone Washington,” which in the US is a name that strongly suggests it belongs to a Black man.
- **Expectation:** Black legislators responds more often to out-of-district emails from black people.

Multiple matching variables

Mahalanobis distance

The following steps guides to obtain Mahalonobis distance:

- Ensure that no variable ends up being weighted more heavily just because it's on a bigger scale: take each matching variable and divide its value by its standard deviation.
- Compute the distance: for a given treated observation A and a given control observation B, the Mahalanobis distance d is the sum of the squares of all the differences between A and B. Mahalanobis distance between two vectors x_1 and x_2 is

  $$d(x_1, x_2) = \sqrt{(x_1 - x_2)' S^{-1} (x_1 - x_2)}.$$

  S is the covariance matrix for all matching variables.

Multiple matching variables

Mahalanobis distance

Example: Defaulting on credit card debt—ctd.

- The example is extended to matching in two variables: $BillApril$ and Age.
- Comparing treated observation 10,305 and untreated observation 27,281, the distance between them in their $BillApril$ is 21
- The difference in ages is $|23 - 37| = 14$
- Ignoring the relationship between the matching variables, the standard deviation of $BillApril$ is NT$ 59,554.

Multiple matching variables

Mahalanobis distance

Example: Defaulting on credit card debt—ctd.

- Divide all the values of *Bill April* by the standard deviation, a new difference of $|1.5386 - 1.5382| = 0.0004$ sd is obtained.
- The standard deviation of *Age* is 9.22, giving a distance of $|2.49 - 4.91| = 1.52$ sd.
- The differences in standard-deviation terms are squared up and added and square rooted to get Mahalanobis distance of $d = \sqrt{0.0004^2 + 1.52^2} = 1.520$

Multiple matching variables

Mahalanobis distance

Example: Defaulting on credit card debt—ctd.

Now, the covariance matrix is taken into consideration:

- We have a difference of 21 in $BillApril$ and a difference of 14 in Age.
- Mahalanobis distance is

$$d(x_1, x_2) = \left(\begin{bmatrix} 21 \\ 14 \end{bmatrix}' \begin{bmatrix} 59544^2 & 26138 \\ 26138 & 85 \end{bmatrix}^{-1} \begin{bmatrix} 21 \\ 14 \end{bmatrix}^{-1/2} \right).$$

Multiple matching variables

Curse of dimensionality

Definition: Curse of dimensionality

The curse of dimensionality means that the more matching variables are added, the less likely we are to find an acceptable match for any given treated observation.

Multiple matching variables

Curse of dimensionality

- Mahalanobis distance suffers from curse of dimensionality.
- Ways to deal with it:
  - Limit the number of matching variables
  - To have a sample of observations
  - Compromise with the match quality, or with the value of the caliper/bandwidth, as the number of dimensions goes up

Coarsened exact matching

Definition: Coarsened exact matching

In coarsened exact matching we only match if observations exactly matches for each matching variable. The **coarsened** part comes in because continuous variables are **coarsened** first by putting them into **bins**, rather than matching on exact values.

Coarsened exact matching

Example: Defaulting on credit card debt

- The variable $BillApril$ can be cut into deciles
- For observation 10,305, instead of matching in that row's exact value of NT$ 91,630 (which no other observation shares), the value of NT$91,630 is between the 80th and 90th percentiles of $BillApril$
- So put it in the category of between NT$63,153 (80th percentile) and NT$112,110 (90th percentile)

Coarsened exact matching

Steps

1. Bin continuous variables.
2. Look for exact matches. Only keep *treated* observations with at least one exact match in the control group and drop the others. Only keep *control* observations with at least one exact match in the treatment group and drop the others.
3. Assign a weight corresponding to the number of treatment group observation matches it has, divided by the number of control observations matched to that treated observation and multiply by the total number of matched control observations divided by the total number of matched treatment observations:

$$\frac{TreatedMatches}{ControlMatches} \cdot \frac{TotalControlMatches}{TotalTreatedlMatches}$$

Coarsened exact matching

Curse of dimensionality

- Coarsened exact matching, if applied to moderately-sized samples (or any size sample with too many matching variables), can lead to lots of treated observations being dropped.
- This makes the treatment effect estimates much noisier, and can also lead the result to be a poor representation of the average treatment effect if certain kinds of treated observations are more likely to find matches than others.

Entropy balancing

Key Concept: Entropy balancing

Entropy balancing enforces restrictions in terms of moment conditions (means, variances, and so on) on the distance between treatment and control for matching variables: conditions demand no differences between treatment and control observation for these variables.

- Entropy balancing find sets of weights that satisfy those restrictions.
- Entropy balancing gives the assured-no-difference result that coarsened exact matching does, but without having to limit the number of matching variables or drop a bunch of treated observations.

Propensity score weighting with multiple matching variables

Common techniques

- Propensity scores based on **logit** or **probit regression** can aggregate multiple matching variables into a single value that can be matched on
- A good propensity score should serve the purpose of closing back doors

Other ways to estimate the propensity score

- Methods that can deal with high degrees of nonlinearity and high dimensions are particularly favored in the estimation.
- Two popular approaches to propensity score estimation using machine learning methods are:
    - **Regularized regression**
    - **Boosted regression**

Propensity score weighting with multiple matching variables

Boosted regression

Definiton: Boosted regression

Boosted regression is an iterative method that starts with a binary regression model (like a logit), checks which observations are particularly poorly predicted, and then runs itself again, weighting the poorly-estimated observations more heavily so the prediction model pays more attention to them, reducing their prediction error.

Propensity score weighting with multiple matching variables

Good propensity scores should close backdoors

Stratification test

1. Split propensity scores into bins.
2. Within each bin, check for each matching variable if it is related to treatment.

   If it is, then it is advisable to try adding some more polynomial or interaction terms for the offending matching variables.

Assumptions for matching

1. Conditional Independence

- The conditional independence assumption says that the set of matching variables that are chosen is enough to close all back doors.
- The entirety of the relationship between treatment and outcome is either one of the causal front-door paths that is wanted, or is due to a variable that is measured and included as a matching variable.

2. Presence of appropriate control observations to match with

- The assumption of **common support** says that there must be substantial overlap in the distributions of the matching variables comparing the treated and control observations.
- In the context of propensity scores, there must be substantial **overlap in the distribution of the propensity scores**.

Assumptions for matching

Two approaches to check for common support

1. The first approach to checking for common support is simply to look at the distribution of a variable for the treated group against the distribution of that same variable for the untreated group.
2. Another way to check for common support when matches are being selected is simply to see success rate at finding matches. If 1% of the treated observations fail to find an acceptable match, that is considered good. If 90% of the treated observations fail to find an acceptable match, then there is lack common support.

Assumptions for matching

Example: The effect of early-release programs on recidivism

- These are programs that allow people who have been sentenced to prison time to leave before their sentence is up, often under certain terms and conditions.
- There is a group of treated ex-prisoners who were released early, and a group of untreated ex-prisoners who served their full sentence.
- There is also information on whether they committed any additional crimes in the ten years after they got out.
- Matching variable: “behavior score” given by the prison review board gives. The score decides whether someone gets early release.

Assumptions for matching

Example: The effect of early-release programs on recidivism—ctd.

- When looking for matches for the treated observations, we end up trying to find ex-prisoners who got a score of 8-10 but did *not* get early release.
- Problem: the score (1 to 10), translates too directly to early release: everyone with an 8-10 gets early release, and everyone with a 1-7 does not.
- The analysis **lacks common support**: there simply are *no comparable control observations*!

Assumptions for matching

Dealing with support

- One way in dealing with support is to avoid trying to match where there is no support.
- Another way of dealing is treating the treated and control groups separately and looking for the range of values of the propensity score where the density distribution is zero. Then, dropping all observations in the other group that have propensity scores in the range that does not exist in the group.
- A common approach is to trim outliers—any observation with a propensity score in the top or bottom X% gets trimmed.

Assumptions for matching

2. Balance

- Balance is the assumption that the approach to selecting a matched group has closed back doors for the variables of interested
- In theory, the selected weights should lead to treated and control groups with the exact same values of the backdoor variables
- Worse matching of values leads to **bad balances** which in tern leads to **bias!**

Assumptions for matching

Balance table

A common way of checking for balance is a *balance table*.

Steps

- Take a bunch of variables that should have balance
- Then, show some of their summary statistics separately for the treated and control groups
- Finally, do a difference-of-means test for each of the variables to see if the treated and control groups differ

Assumptions for matching

Balance table

- Often, the balance table is created twice:
  - Once with the raw, un-matched data to show the extent of the balance problem that matching is to solve
  - Then again with the matched data to ideally show that the balance problem is gone.
- In a good balance, there are **no large diffferences in means.**

Assumptions for matching

Balance table: estimating mean differences

- Usually to estimate the effect of treatment, the weighted means of the outcome for the treated and control groups are often compared
- However, even if the treatment effect is that simple to estimate, the standard errors on that treatment effect will not be quite as simple
- The step of preprocessing the data to match observations and create weights introduces some uncertainty, and so increases the standard errors
- Incorporating this uncertainty into the standard errors is important

Estimation with matched data

Bootstrap standard errors

Bootstrap standard errors can be applied to matching estimates.

- Bootstrap standard errors can only be used with the **constructing a weighted matched sample** approach to matching.
- They do not work properly for the **selecting matches** approach, because the sharp in/out decisions that process makes do not translate into the bootstrap having the proper sampling distribution

Steps

- First randomly resample the data with replacement, then re-perform the matching from scratch, and finally estimate the treatment effect in the bootstrap sample.
- Repeat this process many, many times, and the standard deviation of the treatment effect across all the bootstrap samples is the standard error of the treatment effect.

Combining matching and regression

Regression adjustment

Our research design may depend on a specific functional form of the relationship between treatment and outcome that is difficult to capture with matching but easy with regression. But we still want to close back doors by matching.

Regression adjustment combines matching weights with a regression model. Outcomes of the matching procedure can be included in the regression model:

Propensity scores (or the inverse probability weight based on the propensity score) can be added to regression as a control: if controlled for, back doors from treatment to outcome should be blocked.

Combining matching and regression

Definition: Doubly robust estimation

Doubly robust estimation is a way of combining regression and matching that works even if there is something wrong with the matching or with the regression — as long as it is not both!

Double-robustness is a property that some estimators have rather than *being* a specific estimator.

Combining matching and regression

Doubly robust estimation

Steps

1. Estimate the propensity score p on each observation using the matching variables
2. Use the propensity score to produce the inverse probability weights: $1/p$ for treated observations and $1/(1-p)$ for untreated observations.
3. Using only treated observations, estimate the regression model, using the matching variables as predictors (and perhaps some other predictors too, depending on the purpose), and inverse probability weights

Combining matching and regression

Doubly robust estimation

Steps

4. Repeat step (3) but using only untreated observations
5. Use the models from steps (3) and (4) to produce “treated” and “untreated” predicted observations for the whole sample
6. Compare the “treated” means to the “untreated” means to get the estimate of the causal effect
7. To get standard errors, use bootstrap standard errors, or the heteroskedasticity-robust errors

Matching and treatment effects

Key facts

We can obtain the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the average treatment on the untreated (ATUT) using matching.

- For ATT: construct a control group that is as similar to the treated group as possible
- For ATUT: instead of matching *control observations to treated observations*, match *treated observations to control observations*
- The ATE can then be obtained as the average treatment affect for treated and untreated

  $$\text{ATE} = (p \times \text{ATT}) + (1 - p) \times \text{ATUT},$$

  where p is the proportion of treated individuals.

Matching and treatment effects

Example: Treatment effect on people

Name	Gender	Untreated outcome	Treated outcome	Treatment effect
Alfred	Male	2	4	2
Brianna	Female	1	5	4

- Sample: 500 untreated Alfreds and 500 untreated Briannas, as well as 1,000 treated Alfreds and 200 treated Briannas.

Matching and treatment effects

Example: Treatment effect on people—ctd.

- For the treated group the average outcome is

$$\frac{1000 \times 4 + 200 \times 5}{1200} \approx 4.17.$$

- If one-to-one matching with replacement is performed, for example, values just like in the treated group will be obtained, 1,000 Alfreds and 200 Briannas, but they are untreated this time.

Matching and treatment effects

Example: Treatment effect on people—ctd.

- So the average among the untreated group is

$$\frac{1000 \cdot 2 + 200 \cdot 1}{1200} \approx 1.83.$$

- This yields a treatment effect of $4.17 - 1.83 = 2.34$
- This is much closer to the male treatment effect of 2 than the female treatment effect of 4 because there are far more men in the treated sample than women—it is weighted by *how common* a given observation’s matching variable values are.

Matching and treatment effects

Inverse probability weighting

- Remember: inverse probability weighting does not match to anything: we weight using inverse propensity scores!
- The specific way which is chosen to turn propensity scores into inverse probability weights determines the kind of treatment effect average obtained.
- One approach to inverse probability weighting is to use a weight of $1/p$ for the treated group and $1/(1-p)$ for the untreated group, where p is the propensity score.

  This serves to both upweight the members of the untreated group that are most like the treated group and upweight the members of the treated group that are most like the untreated group.

  $\Rightarrow$ We obtain the ATE.