Matching

Introduction

- Matching is the final causal inference technique we will study.
- It can be thought of as a more sophisticated version of regression.
- Matching and regression both try to get at causal effects by controlling for other factors.
- We have already seen how this works in regression.
- In matching, the core idea is to match two people who have the same characteristics but one is 'treated' and the other is not, then compare the outcomes of each person.

What's Wrong With Regression?

- If the **true** model is really linear, and we have $E\left[U|X\right]=0$, then β is the causal effect.

- However, do we really need to assume the true model is linear?

- If it is not truly linear, a regression is the **best linear approximation** to the underlying model.

- So, in a sense, a regression with $E\left[U|X\right]=0$ is probably just an approximation of the causal effect.

Matching to the Rescue!

- The ability of matching to identify the causal effect does not depend on the linearity of the underlying model.
- While regression holds *the linear effect* of other factors constant, matching simply holds other factors constant.
- Interestingly, both strategies rely on the same assumption though: $E[U|X] = 0$.
- The key difference is in **how** they hold everything else constant.

Independence Assumption

- If you look back to the slides on treatment effects and potential outcomes, we showed that in our new setup, $E[U|X] = 0$ is the same as the selection bias being zero:

$$E\left[Y_{0i}\middle|D_i = 1\right] - E\left[Y_{0i}\middle|D_i = 0\right] = 0.$$

- When we have access to regressors, this becomes

$$E\left[Y_{0i}\middle|X_i, D_i = 1\right] - E\left[Y_{0i}\middle|X_i, D_i = 0\right] = 0.$$

- This is known as the **conditional independence assumption**, i.e. the outcome is independent of the treatment conditional on X.

- Because D_i is binary in this setup, it becomes an independence assumption rather than a mean independence assumption, but it is equivalent to $E[U|X] = 0$ in a standard regression context.

Matching - an Example

- It is easiest to think of binary treatments and discrete regressors (although everything works out with continuous regressors and treatment).
- Consider trying to determine the effect of entering the military on your future earnings (Angrist, 1998).
- Our parameter of interest may be the Average Treatment Effect on the Treated (ATET), given by $E[Y_{1i} - Y_{0i} | D_i = 1]$.
- This tells us the difference between what a veteran earns and what a veteran would earn if they hadn't served in the army (this is clearly a counterfactual).
- 'The glass cliff' gives another example...

The Process of Matching

- There are many different ways to determine a match. The most common method is **one-to-one nearest-neighbour matching with replacement**.

- For simplicity (we will make this more complicated later), suppose we are matching just on age.

- **One-to-one**: this means you only match with one other person. (k-to-one would match you with k other people and you would compare your outcome to their average outcome).

- **Nearest neighbour**: this means you match with someone who is closest in age to you (and has the opposite treatment assignment).

- **With replacement**: this means that when someone has been used as a match, they are 'put back' into the sample and can be used as a match again for someone else.

Matching - Obtaining a Causal Effect

- If we compare the earnings of veterans with those of non-veterans, we get a biased estimate of the ATET (selection bias):

$$\begin{aligned} E\left[Y_i \middle| D_i = 1\right] - E\left[Y_i \middle| D_i = 0\right] \quad &= \quad E\left[Y_{1i} - Y_{0i} \middle| D_i = 1\right] + \\ & \quad\; E\left[Y_{0i} \middle| D_i = 1\right] - E\left[Y_{0i} \middle| D_i = 0\right] \end{aligned}$$

- Using the mean independence assumption (see the lecture slides on treatment effects and potential outcomes):

$$\begin{aligned} \delta_{ATET} &= E\left[Y_{1i} - Y_{0i} | D_i = 1\right] \\ &= E\left\{ E\left[Y_{1i} | X_i, D_i = 1\right] - E\left[Y_{0i} | X_i, D_i = 1\right] \middle| D_i = 1 \right\} \\ &= E\left\{ E\left[Y_{1i} | X_i, D_i = 1\right] - E\left[Y_{0i} | X_i, D_i = 0\right] \middle| D_i = 1 \right\} \\ &= E\left\{ E\left[Y_i | X_i, D_i = 1\right] - E\left[Y_i | X_i, D_i = 0\right] \middle| D_i = 1 \right\} \\ &\equiv E\left\{ \delta_x \middle| D_i = 1 \right\} \end{aligned}$$

Matching Estimator

- Using this equation, we can write a matching estimator as

$$\hat{\delta}_{ATET} = \sum \delta_x P(X_i = x | D_i = 1)$$

- Given the discrete nature of the regressors, it is easy to calculate the probability term. Each combination of x is known as a **covariate cell**.

- δ_x is just the difference in mean outcome for treated and non-treated for different combinations of x.

- Of course, if we're interested in the Average Treatment Effect, we can use

$$\hat{\delta}_{ATE} = \sum \delta_x P(X_i = x)$$

- ATET is the causal effect for soldiers, ATE is the causal effect for the population from which our sample is from.

ATET Estimates (Angrist, 1998)

Race	Average earnings in 1988-1991	Differences in means by veteran status	Matching estimates	Regression estimates	Regression minus matching
	(1)	(2)	(3)	(4)	(5)
Whites	14537	1233.4	-197.2	-88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11664	2449.1	839.7	1074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

(Credit: Mostly Harmless Econometrics, Angrist and Pischke)

Matching Vs. Regression Estimates

- In Angrist and Pischke's book (pp. 55-56) they show that the OLS estimator and the matching estimator are constructed in a very similar way.
- In fact, they can be seen as the same estimator but
  - OLS puts most weight on covariate cells where the conditional (on X) variance of treatment status is highest. In general this is in cells where the number of treated and non-treated are equal.
  - Matching puts most weight on covariate cells which are most likely to be treated. (Angrist and Pischke, pp. 56)
- In the results on the previous slide, the matching estimates are lower. This is because those who are most likely to be treated (join the army) are those who benefit least from the treatment. (Army recruits, on average, are smarter and better educated.)

The Common Support Assumption

- If we think about the idea underlying the matching estimator, it is unsurprising that covariate cells which don't include both treated and untreated observations are not used (δ_x is undefined).

- This turns out to also be true for the OLS estimator as well.

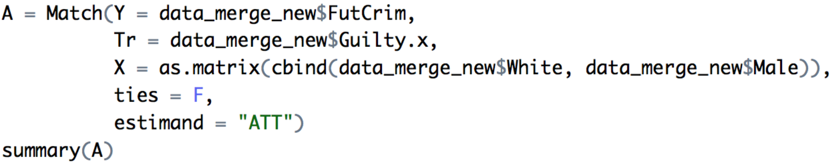
- In the derivation given in Angrist and Pischke, you can show that the OLS estimator can be written in the same form as $\hat{\delta}_{ATET}$ but with weights proportional to

$$P(D_i = 1|X_i = x)\left[1 - P(D_i = 1|X_i = x)\right].$$

- If this common support assumption is not met, the causal effect is only consistent for those covariate cells with common support.

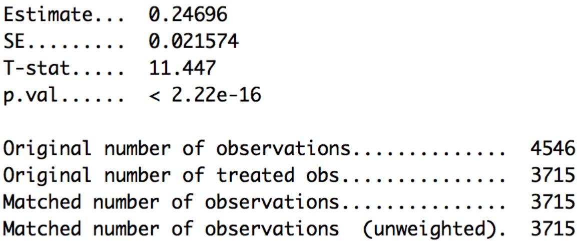
Example: Conviction and Future Crime

- We take a look at how to estimate the causal effect of being convicted on future criminal activity using matching estimators in R.



```r
A = Match(Y = data_merge_new$FutCrim,
          Tr = data_merge_new$Guilty.x,
          X = as.matrix(cbind(data_merge_new$White, data_merge_new$Male)),
          ties = F,
          estimand = "ATT")
summary(A)
```

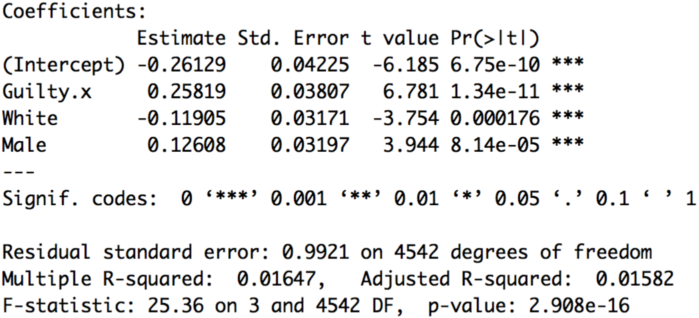
Example: Conviction and Future Crime



```
Estimate...  0.24696
SE.........  0.021574
T-stat.....  11.447
p.val......  < 2.22e-16

Original number of observations..............  4546
Original number of treated obs...............  3715
Matched number of observations...............  3715
Matched number of observations  (unweighted).  3715
```

Example: Conviction and Future Crime

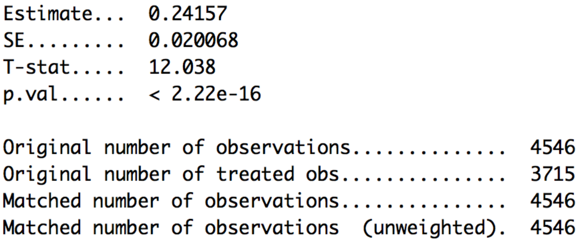
- In comparison to results from an OLS regression



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26129    0.04225  -6.185 6.75e-10 ***
Guilty.x     0.25819    0.03807   6.781 1.34e-11 ***
White       -0.11905    0.03171  -3.754 0.000176 ***
Male         0.12608    0.03197   3.944 8.14e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9921 on 4542 degrees of freedom
Multiple R-squared:  0.01647,    Adjusted R-squared:  0.01582
F-statistic: 25.36 on 3 and 4542 DF,  p-value: 2.908e-16
```

Example: Conviction and Future Crime

- Here we have results for the ATE, rather than ATET. Does it make sense that the ATE and ATET are very close in this example?



```
Estimate...  0.24157
SE.........  0.020068
T-stat.....  12.038
p.val......  < 2.22e-16

Original number of observations..............  4546
Original number of treated obs...............  3715
Matched number of observations...............  4546
Matched number of observations  (unweighted).  4546
```

- (Note, the results are subject to randomness, due to the stochastic nature of matches)

A Continuous Regressor

- In everything we have spoken about so far, we have restricted ourselves to discrete regressors.
- This makes the matching process quick and simple, it also means there are not too many covariate cells.
- What can we do when we have a continuous regressor though?
- With only one continuous regressor (and maybe some other discrete regressors), we find the 'nearest neighbour'.
- For example, if we are matching on age, I first try to find a match with someone of the same age, and if there are no matches, then I look for someone within one year, and so on...
- If there is little common support, i.e. not much overlap in the supports, the distance between two matches could be very large.

Multiple Continuous Regressors

- This is where things start getting tricky. Suppose you have age, education, and income as continuous regressors, as well as race and gender.
- A typical covariate cell will now be defined as all black women of 30-32 years of age with 14-15 years of education and income between \$100 000 - \$110 000.
- This is far too specific! We can't hope to have both a treated and an un-treated people in all of the covariate cells if they're like this!
- Is all lost for matching estimators?

The Propensity Score

- What if we could combine our regressors into a scalar, and then match based on this scalar instead of the (possibly) high-dimensional vector of regressors.
- This is a little bit like the idea of Principal Component Analysis (or other dimension reduction techniques), however, these techniques always result in a loss of information.
- Rosenbaum and Rubin (1983) provided a surprising and incredibly powerful solution:
    - *If potential outcomes are independent of treatment status conditional on a multivariate X, then potential outcomes are independent of treatment status conditional on the* **propensity score**, *defined as $p(X_i) \equiv E[D_i|X_i]$.*

Proof of the Propensity Score Theorem

- If we can show $P\left[D_i = 1 | Y_{ji}, p(X_i)\right]$ does not depend on Y_{ji}, then we have proved the result.

$$\begin{aligned}
P\left[D_i = 1 | Y_{ji}, p(X_i)\right] &= E\left[D_i | Y_{ji}, p(X_i)\right] \\
&= E\left\{ E\left[D_i | Y_{ji}, p(X_i), X_i\right] \Big| Y_{ji}, p(X_i) \right\} \\
&= E\left\{ E\left[D_i | Y_{ji}, X_i\right] \Big| Y_{ji}, p(X_i) \right\} \\
&= E\left\{ E\left[D_i | X_i\right] \Big| Y_{ji}, p(X_i) \right\} \\
&= E\left\{ p(X_i) \Big| Y_{ji}, p(X_i) \right\} \\
&= p(X_i)
\end{aligned}$$

- Where we have used the fact that treatment status is independent of potential outcomes conditional on X.

Benefits of The Propensity Score Theorem

- This theorem is incredibly powerful for us. First, it means that we don't have to match on some high-dimensional vector, we can match on a simple univariate object. **The only variable that matters is the probability of being treated**.
- Second, it says that we don't need to try and match on every single possible thing, we only need to control for things which affect the probability of treatment.
- This is similar to what we learned from the OVB formula. That equation showed us that we only need to control for things which affect the treatment.
- Note, as we did in regression, we can (and probably should) include polynomial terms in continuous regressors to improve the estimate of the propensity score.

Propensity Score Matching in Practice

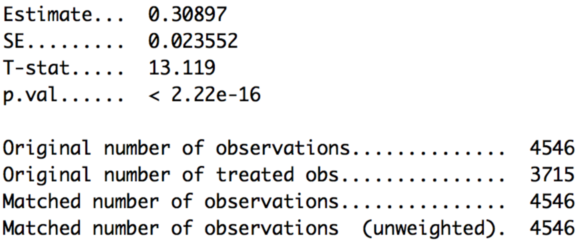
- Matching based on the propensity score proceeds in two steps.
    - First, we estimate the probability of treatment conditional on X_i. The first step is usually carried out using logit or probit.
    - Second, we construct the matching estimator based on this propensity score, i.e. people with similar propensity scores are matched together (rather than matching on age, gender, etc.).
- It's also possible to estimate the causal effect without going through the whole matching procedure. It can be shown that

$$E\left[Y_{1i}\right] = E\left[\frac{Y_i D_i}{p(X_i)}\right] \quad \text{and} \quad E\left[Y_{0i}\right] = E\left[\frac{Y_i\left(1 - D_i\right)}{\left(1 - p(X_i)\right)}\right]$$

and the causal effect is given by the difference of these two.

Justice Example (Again)

- We return to the estimation of the causal effect of conviction on future criminality.
- This time, we match based on the propensity score. In R, we have to do this manually (check the R Code on blackboard)...

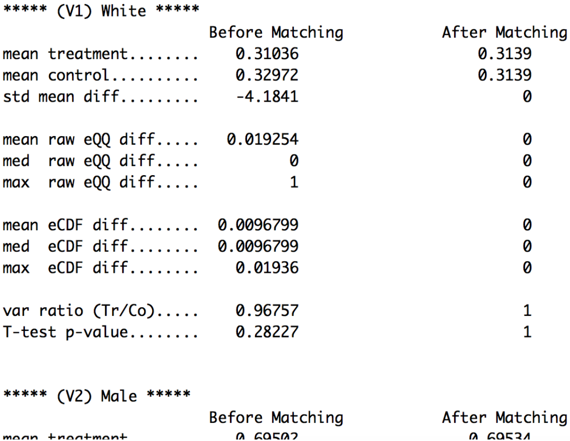


```
Estimate...  0.30897
SE.........  0.023552
T-stat.....  13.119
p.val......  < 2.22e-16

Original number of observations..............  4546
Original number of treated obs...............  3715
Matched number of observations...............  4546
Matched number of observations  (unweighted).  4546
```

Checking for Balance

- The whole point of everything we are doing here (and what we did in regression) is to compare like with like... apples with apples.
- One way we can check this is to **see if our covariates are balanced** in the treatment group and the control group.
- If the two groups are identical in all observable ways, we hope they may be the same in all unobservable ways too.
- Then the difference in mean outcomes between the two groups can be reasonably interpreted as a causal effect.
- R has some built in functions to do this...

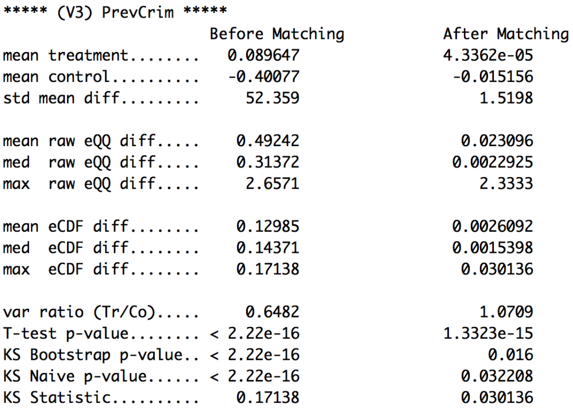
Checking for Balance in R



```
***** (V1) White *****
                       Before Matching          After Matching
mean treatment........    0.31036                  0.3139
mean control..........    0.32972                  0.3139
std mean diff.........    -4.1841                       0

mean raw eQQ diff.....   0.019254                       0
med  raw eQQ diff.....          0                       0
max  raw eQQ diff.....          1                       0

mean eCDF diff........  0.0096799                       0
med  eCDF diff........  0.0096799                       0
max  eCDF diff........    0.01936                       0

var ratio (Tr/Co).....    0.96757                       1
T-test p-value........    0.28227                       1


***** (V2) Male *****
                       Before Matching          After Matching
mean treatment            0.69502                 0.69534
```

- Matching on discrete regressors will result in perfect balance.

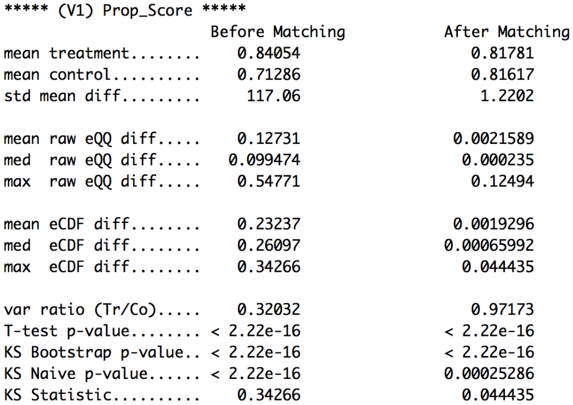
Checking for Balance in R



```
***** (V3) PrevCrim *****
                       Before Matching           After Matching
mean treatment........   0.089647                4.3362e-05
mean control..........   -0.40077                 -0.015156
std mean diff.........     52.359                    1.5198

mean raw eQQ diff.....    0.49242                  0.023096
med  raw eQQ diff.....    0.31372                 0.0022925
max  raw eQQ diff.....     2.6571                    2.3333

mean eCDF diff........    0.12985                 0.0026092
med  eCDF diff........    0.14371                 0.0015398
max  eCDF diff........    0.17138                  0.030136

var ratio (Tr/Co).....     0.6482                    1.0709
T-test p-value........ < 2.22e-16                1.3323e-15
KS Bootstrap p-value.. < 2.22e-16                     0.016
KS Naive p-value...... < 2.22e-16                  0.032208
KS Statistic..........    0.17138                  0.030136
```

- Now we don't get perfect balance, in fact, the p-value says we don't have balance.

Checking for Balance in R

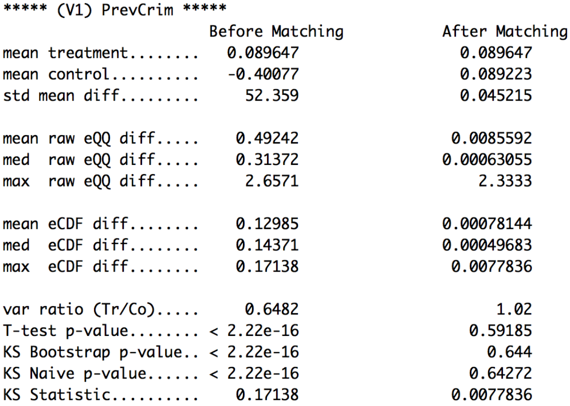


```
***** (V1) Prop_Score *****
                       Before Matching          After Matching
mean treatment........    0.84054                 0.81781
mean control..........    0.71286                 0.81617
std mean diff.........     117.06                  1.2202

mean raw eQQ diff.....    0.12731               0.0021589
med  raw eQQ diff.....   0.099474                0.000235
max  raw eQQ diff.....    0.54771                 0.12494

mean eCDF diff........    0.23237               0.0019296
med  eCDF diff........    0.26097              0.00065992
max  eCDF diff........    0.34266                0.044435

var ratio (Tr/Co).....    0.32032                 0.97173
T-test p-value........ < 2.22e-16              < 2.22e-16
KS Bootstrap p-value.. < 2.22e-16              < 2.22e-16
KS Naive p-value...... < 2.22e-16              0.00025286
KS Statistic..........    0.34266                0.044435
```

- Here, we have tried to achieve balance in the propensity score but failed :(

Genetic Matching

- There are even fancy genetic algorithms which find the matches which give the best balance possible. Below we have the results for the optimal matching on previous criminality



```
***** (V1) PrevCrim *****
                       Before Matching          After Matching
mean treatment........    0.089647                0.089647
mean control..........    -0.40077                0.089223
std mean diff.........      52.359                0.045215

mean raw eQQ diff.....     0.49242               0.0085592
med  raw eQQ diff.....     0.31372              0.00063055
max  raw eQQ diff.....      2.6571                  2.3333

mean eCDF diff........     0.12985              0.00078144
med  eCDF diff........     0.14371              0.00049683
max  eCDF diff........     0.17138               0.0077836

var ratio (Tr/Co).....      0.6482                    1.02
T-test p-value........ < 2.22e-16                  0.59185
KS Bootstrap p-value.. < 2.22e-16                    0.644
KS Naive p-value...... < 2.22e-16                  0.64272
KS Statistic..........     0.17138               0.0077836
```

- We have achieved balance now :)

Summary

- We have looked at an alternative to regression: matching.
- We saw how to estimate ATET and ATE using matching.
- We discussed the differences between matching and regression.
- We introduced the propensity score as a way to match on multiple continuous variables.
- Finally, we looked at how to check for balance in covariates to determine if the matching was successful.