LECTURE 5. MATCHING, PROPENSITY SCORES AND REGRESSION

Dr. Tom O'Grady

1. Selection on Observables and Average Treatment Effects

So far in this course we have studied the idealised world of experiments. When we can randomise a treatment and mitigate problems like attrition, experiments are often the best way to study causal questions in the social sciences. In many cases, though, randomisation is simply impossible. For instance, some characteristics like people's race or early childhood experiences simply can't be randomised, but both race and experiences in childhood are known to strongly influence people's economic, health and political behaviours. Likewise, some treatments can't be randomised for obvious ethical reasons but are extremely important for governments and policymakers to understand. In this lecture we'll look at two such examples: smoking and warfare.

One possibility in these situations is to look for natural experiments, where a treatment varies "in the wild" without the intervention of a researcher. These are rare, however. In this lecture we'll look at a very common situation, where the best that we can do is to conduct an observational study

> **Observational Research:** The opposite of experimental research: a study that involves observing the world and making inferences about it from our observations.

Uncovering the causal effect of a treatment in observational research requires that we account for all possible confounders. Recall that a confounder is a variable that affects both the outcome variable and selection into a treatment. We must be able to *observe* all confounders if we are to uncover a true causal effect. For that reason, in the field of causal inference the identification strategy behind observational research is often referred to as "selection on observables."

> **Selection on Observables:** A situation where assignment to treatment or control can be considered to be randomly assigned, conditional on observed covariates.

For concreteness, suppose we want to uncover the impact of smoking on a health outcome (like the incidence of heart disease, lung cancer etc.), and we have survey data on a large number of people that details their characteristics and their health. We have the following information:

- Y: our outcome of interest
- D: a treatment variable equalling 1 if the respondent smokes and 0 otherwise

- X: a collection of control variables like sex, income, and age that affect both the outcome and the decision to smoke or not

The selection on observables assumption means that X contains all possible confounders. Equivalently, it means that *amongst people with the same X, smoking should be as-if randomly assigned*. Smokers and non-smokers should on average have the same potential outcomes after controlling for X, i.e. they should be balanced conditional on X. Then, to find the effect of smoking on health, we need to compare the health outcomes of smokers and non-smokers with the same values of X.

Suppose for the sake of argument that the only confounders here are age and sex. For example, men and the elderly may be more likely both to smoke and to have worse health. If the selection on observables assumption is satisfied, then amongst people of the same sex and age, smokers and non-smokers would on average have the same potential outcomes.[1] In this case, causal inference would require us to compare the health outcomes of smokers and non-smokers who are balanced in terms of age and sex. Concretely, suppose that we only have 6 units in our sample, and they are divided into two strata, each of which contains people of the same age and sex (e.g. stratum 1 could be males aged 35 and stratum 2 could be females aged 70). In stratum 1 we have four units, two in the treatment condition ($D = 1$, smokers) and two in control ($D = 0$, non-smokers), while in stratum 2 we have two units, one in treatment and 1 in control. The outcome variable Y is an index of the person's health where higher values denote worse health.

The situation looks like this:

Stratum	Unit	D	Y
1	1	1	55
1	2	1	55
1	3	0	50
1	4	0	55
2	5	1	30
2	6	0	20

This is akin to a randomised experiment, because the treatment and control groups are balanced. Here, we are comparing "apples to apples" because both the treatment and control groups contain people from both strata, and therefore on average contain people with the same potential outcomes. Thus we can calculate the ATE of smoking on health as the difference in mean outcomes between the treatment and control groups. This gives an ATE of $(55 + 55 + 30)/3 - (50 + 55 + 20)/3 = 5$.

An exactly equivalent way to do it is to calculate a separate ATE within each stratum, and then take a weighted average across strata, weighting each stratum by the number of units it contains. With p strata, we have:

$$\hat{\tau} \;=\; \frac{1}{N}\sum_{k=1}^{p} ATE_k.n_k$$

where N is the total number of units in the study, n_k is the number of units in the stratum k and ATE_k is the stratum-specific ATE. These stratum-specific ATEs are as follows:

[1] Obviously this is an unrealistic example

Stratum	Unit	D	Y	ATE_k
1	1	1	55	
1	2	1	55	
1	3	0	50	
1	4	0	55	**2.5**
2	5	1	30	
2	6	0	20	**10**

where they are calculated in the same way as the overall ATE. The ATE using the weighted average is therefore $(2.5 * (4/6)) + (10 * (2/6)) = 5$.

The bottom line is that under selection-on-observables, the Average Treatment Effect can be thought of as an average of ATEs for each stratum (that is, for each group that has the same characteristics X that determine selection into treatment), weighted by the proportion of units in each one.

2. Multiple Regression: Model Dependence and Common Support

To estimate the ATE in an observational study with all confounders observed, we could also estimate a regression of the following form:

$$Y = \beta_0 + \tau D + \beta' X + u$$

where $\hat{\tau}$ will be our estimated ATE and X contains all of the confounders (e.g. sex and age). Is $\hat{\tau}$ the correct ATE in this case? The answer is a qualified yes: regression will usually come very close to the true ATE, when we control for all confounders. It can be shown to be an average of strata-specific ATEs that comes close to weighting each stratum in proportion to its size.[2] Thus in a situation where we observe all confounders, regression effectively ends up comparing the outcomes of treated and untreated units that have the same values of X. This is also only strictly true when, in the regression equation, X contains a dummy variable for every single stratum.

2.1. Model Dependence

Where this is impossible (e.g. there are too many strata because some X variables are continuous), an appropriate functional form must be chosen, and it must be the correct functional form. With many variables in X, this becomes more difficult to achieve. How can we possibly know not only which variables to include, but how they should be included too: as logs, with squared terms, or interactions? With even a reasonably small set of measured confounders, there are dozens of models that could be estimated from the same data, each of which can be supplied with a reasonable ex post justification. The problem is that small changes in modeling – which variables are included in a regression, and/or how they are included – can often lead to big changes in the estimated quantity of interest. Thus one problem with multiple regression is that results tend to be very *model dependent*.

[2]Technically, it assigns a slightly higher weight to strata with more variance in the treatment indicator. See Angrist and Pischke, *Mostly Harmless Econometrics* for details.

Model Dependence: A situation where the results of a statistical analysis are highly dependent on the choice of model. Often applied in particular to situations where one or models are equally plausible and defensible, but give very different results.

This is a major problem for causal inference. Remember that in this module, we are aiming to make causal inferences with as few assumptions as possible, ideally designing research up front so that our study yields a causal effect without strong assumptions. The problem is that for any given regression result, we must assume that the model is the correct one out of all possible models that could have been chosen – a very strong claim. Regression opens the door not only for post-hoc rationalisations (a researcher gets the result she wants with a particular model, and convinces herself in light of this that her model must have been correct), but also outright fraud or 'p-hacking'[3], where researchers deliberately make modeling choices that give them the best possible results, which might be necessary to get published in a good journal. Overall, "how do readers know that publications are not merely demonstrations that it is possible to find a specification that fits the author's favorite hypothesis?"[4]

2.2. Lack of Common Support

A second (and very related) major problem with regression is that recovering the correct ATE from a regression is only possible in a situation where there is substantial overlap in X between the treatment and control units. This is known as *common support* and corresponds to a situation like the one in Figure 1 below, where we have a single confounder X.

Common Support: A situation where there is (near) complete overlap in covariates between treated and control units

In Figure 1, we are comparing similar units to similar units - we are approximating a randomised experiment, given that X is the only confounder. Because D is binary, regression lines can be plotted separately for the two groups of units, and the ATE is simply the vertical distance between the two lines. Clearly in this case the treatment has a positive effect, on average, across all values of X.

A common situation, however, is lack of common support. That corresponds to a situation like Figure 2, where some of the control group bear no similarity to the treatment group in terms of X (in this case, at lower values of X). For those values of X, we are no longer comparing like with like. The regression function will nonetheless estimate a line for the treated units across the whole range of X, even where no data exists. In such a situation, we rely on *extrapolation* and *modeling assumptions* to estimate the ATE in this region without data on the treated units. Remember that it is perfectly possible to estimate non-linear regression functions with quadratic terms, interactions, and so on. Here we must ask: how do we know that the regression function for the treated units is linear in the region where we have no data? We are making a modelling assumption, relying on extrapolating the linear relationship that exists at the observed values

[3] "p-hacking" generally refers to trying to choose a specification that yields the lowest possible p-value for a quantity of interest, since statistically significant results are often needed for publication in journals

[4] Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis 15:, p.199

Figure 1: **Regression-Based Inference with Common Support at all values of X**

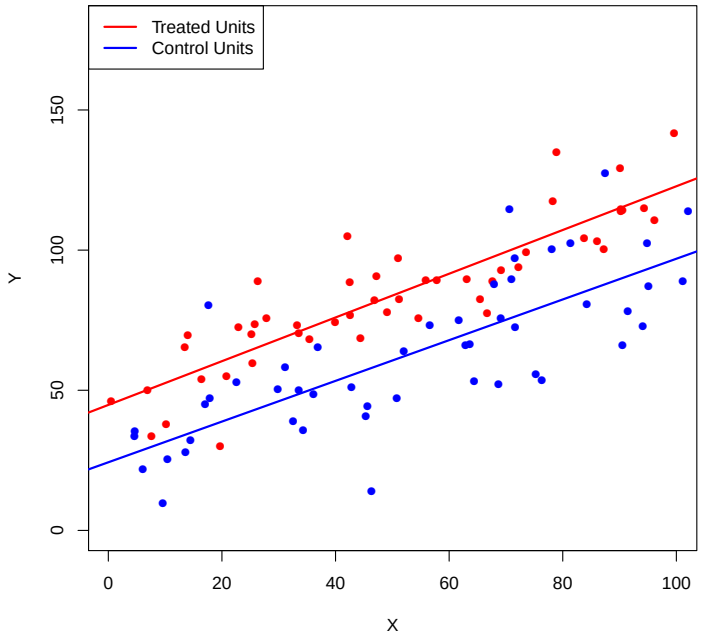
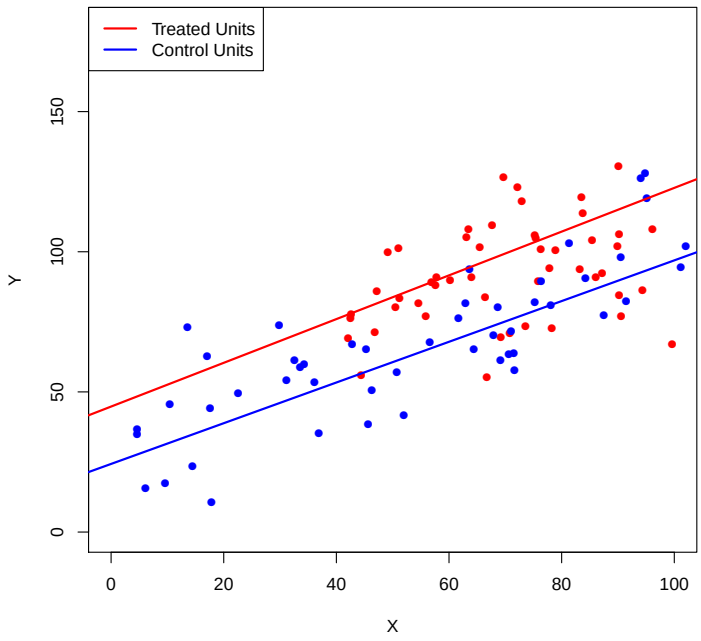


Figure 2: **Regression-Based Inference without Common Support at all values of X**



of X to the unobserved values. We are back again in the situation we are trying to avoid in this course: relying on modeling assumptions (one might even say guesswork) to make causal inferences. This is what Rubin is getting at in this week's reading on smoking, when he says (on pages 30-31):

> "A difference of means of more than a standard deviation (i.e., B =1.09) is simply too large to rely on modelling adjustments unless we are certain of the form of the model (e.g., we are sure the outcomes of interest are linearly related to covariates X), because of the extrapolation involved when fitting straight lines to such data... the groups are far apart, too far to trust adjustment based on linear models."

3. Matching

These problems of lack of common support and a heavy reliance on modeling are what motivate the use of matching estimators in place of regressions when carrying out observational studies with a binary treatment variable. Matching is an old idea that has become vastly more popular in the social sciences in recent years. The basic idea is to try to make sure that we are comparing "apples with apples" by cutting down our sample to only those units with similar observed values of baseline covariates. Specifically, in matching we take each of our treated units and match them to one or more control units with similar values for the X variables. In our smoking example we would take each treated person (i.e. smokers) and match them to a control person (non-smoker) who is of the same sex and close to the same age. We then create a dataset consisting only of these matched units. Having done that, we can then estimate an average treatment effect for these matched units only. The subset of matched untreated units provide a better counterfactual for the treated units than simply using all untreated units regardless of their similarity to the treated units. Rubin puts this nicely on page 30:

> "because we want to compare the current smokers to 'like' never smokers, if the never-smoker group includes individuals who look nothing like any current smokers with respect to the propensity score, or any other covariate, they should simply be discarded because they carry essentially no information about what a current smoker's health outcomes might be like if he instead were a never-smoker."

Thus matching can be thought of as a better approximation to the ideal world of randomised experiments than regression. But is matching *always* 'better'? This largely depends on what you are aiming to achieve. Matching is often considered superior to regression for causal inference, because it requires no modeling assumptions and it solves the problem of lack of common support. As Rubin also discusses, it is more transparent. It is more difficult to repeatedly try out different models and then report only the one that gives a low p-value. On the other hand, regression might be preferred for more descriptive or predictive purposes, where we want to carefully find the form of the relationship between dependent and independent variables, or predict the outcome for a unit based on its covariates (e.g. in economic or election forecasting).

For causal inference, matching and regression both rely on the same underlying selection-on-unobservables assumption. The observed covariates must account for all possible confounders. Often, this will not be the case because there are some confounders that are unobserved. Clearly, matching usually achieves better balance on observed covariates than regression. We might also

hope that balancing our treatment and control groups in terms of observed covariates will also achieve superior balance on unobserved covariates, but this is very far from certain.

An additional complication of matching is that it often means discarding some of the control units that lack similar X values to the treated units. In such a case, technically speaking we end up estimating a quantity known as the Average Treatment Effect for the Treated (ATT), since we are only focusing on the treated units rather than the whole population.

Carrying out matching in practice involves several steps. We'll go through each one in turn.

3.1. Choosing Variables to Match on

Often, in an observational dataset we will have many possible baseline variables to choose from. Which ones should we use to match treatment and control units? The answer is, in general, all variables that you think are likely to be confounders. That is, all variables that determine both treatment uptake and the outcome. It is usually safest to err on the side of including more rather than fewer variables, unless we have a very strong idea of what drives selection into treatment. Also, matching should clearly not be carried out using the *outcome variable*.

More subtly, it should also never be carried out using covariates that are themselves affected by the treatment. Matching should be done on *pre-treatment* covariates only. Otherwise, so-called "post-treatment bias" can occur.

> **Post-Treatment Bias:** Bias that results from controlling for a variable that is affected by the treatment

This is intuitive, when you think about it. Suppose your outcome of interest in the smoking study is the difference in incidence of heart disease between non-smokers and smokers, but you match treatment and control units based on their aerobic fitness. This is likely to under-estimate the causal effect of smoking on heart disease, because low aerobic fitness is itself a consequence of smoking. Suppose aerobic fitness is a binary variable. By comparing smokers with high aerobic fitness only to non-smokers with high aerobic fitness, you would be "controlling away" some of the treatment effect of smoking. You are now comparing two groups that are *not* similar to each other in terms of potential outcomes: those whose fitness is high in spite of smoking, and those whose fitness is high without smoking. The same is true when comparing smokers with low aerobic fitness only to non-smokers with low aerobic fitness. In both cases, controlling for fitness will underestimate the effect of smoking.

As an even more extreme example, suppose your outcome variable is whether or not someone dies prematurely, the treatment is smoking, and you control for lung cancer. Lung cancer is a consequence of smoking; it is one of the main ways that smoking *leads to* premature death, hence it is clearly post-treatment. If we only compare smokers with lung cancer to non-smokers with lung cancer (and smokers without lung cancer to non-smokers without lung cancer), we will vastly under-estimate the treatment effect.

3.2. Measuring 'Distance'

Having chosen the set of variables, next we must define what it means for a control unit to be a 'match' for a treated unit. In rare cases it may be possible to carry out exact matching.

Exact Matching: A matching algorithm where each treated unit is matched to a control unit with the exact same values for the covariates.

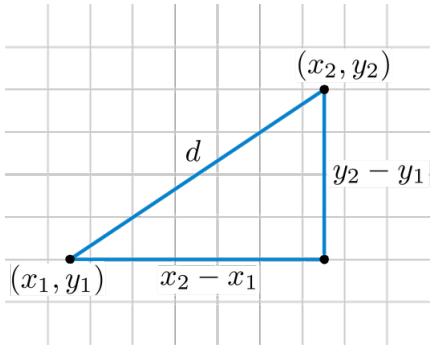
This would probably be feasible in our smoking example, where we simply have to match treated men and women with control men and women of the same age. But these sorts of situations are very rare. As soon as we add more variables, particularly continuous variables like income, it becomes impossible to find an exact match. Thus we usually need to choose a *distance metric* to define what it means for a treated unit to be similar to a control unit.

Distance Metric: A function that measures the distance between two sets of numbers. Loosely, a function that tells you how similar two sets of numbers are to each other

Distance metrics come from the mathematical field of geometry. Suppose we have only two covariates for matching. Then, we could plot any two treatment and control units against each other on two-dimensional axes, where each axis is one of the covariates. That allows us to calculate the *Euclidean Distance* between a treatment unit defined by (x_1, y_1) and a control unit defined by (x_2, y_2), where 1 and 2 represent the two units and x and y the two covariates.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Figure 3: **Euclidean Distance in Two Dimensions**



Geometrically, all we are doing is finding the hypotenuse of a right-angled triangle. In fact, in two dimensions the Euclidean Distance is just a re-statement of Pythagoras' Theorem, as shown in Figure 3. The formula generalises to more than two covariates (dimensions), comparing the distance between treatment and control units in n-dimensional space where the vectors TR and C represent the covariates of treatment and control units:

$$d = \sqrt{(TR - C)^\top (TR - C)}$$

If you are unfamiliar with this matrix notation, do not worry. The intuition of a distance metric is the important thing to understand. A commonly-used (and often superior) distance metric is given by the *Mahalanobis Distance*, which is similar to the Euclidean Distance except that it standardises distances into units of standard deviations. There are many different distance metrics out there, many of which are available in standard R packages.

3.3. With or Without Replacement?

Before moving on to estimation of the treatment effect, the analyst needs to choose whether to match with or without replacement. Matching with replacement means that control units can be used as a match for more than one treated unit. Say we went through our treated units one at a time, matching each one in turn to a control unit. Matching with replacement means that each control unit is "placed back" into the controls after being used once, so that it can potentially be used again. This has two key advantages. First, it often means that the treatment and control units after matching will be better balanced, because sometimes one control unit is naturally similar to more than one treated unit. Second, the matching algorithm is reduced in complexity, because the order in which we match the units does not matter. Without replacement, we could get different results depending on how we order the units. The only downside is that sometimes we may end up with lots of treatment units and very few control units. Nonetheless, it is most common to match with replacement, and that is what R does by default. There is nothing wrong in principle with having treatment and control groups of different sizes.

3.4. Breaking Ties

One also needs to decide what to do with "ties." It is sometimes the case that more than one control unit performs equally well as a match, particularly if only a small number of variables are used for matching. For example, two control units might have identical values for some set of control variables. In that case, by default R's matching function will include all of the 'tied' observations and weight them in the estimation. For example, if two control units are equally good matches for a particular treated unit, they will each be weighted one half in the ATT and standard error calculations.

Note that when using the option `ties==FALSE` in R's `Match` package, "ties" are broken at random. This means that if more than one control unit provides a potential match for a treated, then only one of the control units will be chosen at random to serve as the match. This means that different results will be obtained each time, depending on how R randomly breaks the ties, so that there is no unique answer. It is more principled to include all of the tied observations, since we do not arbitrarily 'throw away' any of the data, and the results do not depend on the particular randomisation that R carries out.

3.5. One-to-One vs. One-to-Many

As an alterative to worrying about ties, one can also use one-to-many matching instead of one-to-one. In one-to-one matching, each treated unit can be matched to only one control. In one-to-many matching, each one can be matched to more than one control. This can be useful in large samples where there are more control units than treated units, because the inclusion of more units will increase the precision of our estimates. However, often the second, third and fourth matches may be poorer than the first match, meaning that we may end up including control units that are not very similar to the treatment

3.6. Assessing Balance

Just as in a randomised experiment, after carrying out matching we should first carry out balance tests to compare the treatment and control units. If matching was successful, then by definition they should be very similar to each other in terms of their covariates. Balance tests are carried out in exactly the same way as before, using t tests or comparisons of densities. Balance tests are particularly useful in matching because they might be able to help us choose between different distance metrics or matching with vs. without replacement. A principled way to make those choices is to try multiple variations of matching, and select the procedure that gives the best balance after matching.

3.7. Estimating the Treatment Effect

Finally, having chosen a matching procedure that offers good balance, the only remaining step is to estimate the Average Treatment Effect (or more precisely, the Average Treatment Effect for the Treated) in the matched dataset. This can be done with a difference in means, as before.

The code for matching in R is relatively simple. First, to carry out matching using some set of covariates named "matching-var1", "matching-var2" etc., an outcome variable y and treatment variable *treat*, run:

```
library(Matching)
matching.vars <- cbind(matching_var1,matching_var2,...)
m <- Match(Y=y, Tr=treat, X=matching.vars, options...)
summary.Match(m)
```

And to compare balance before and after matching, run:

```
MatchBalance(treat ~ matching_var1 + matching_var2 + ..., match.out=m)
```

Note that just like with experiments, it is also possible to estimate the ATT using a regression. In that case, you first create a dataset of only the treated units and their matched counterparts amongst the controls, and then regress the outcome on the treatment for only this subset of the data. One advantage is that it allows you to control for variables where any imbalances remain - although this may lead to issues of a lack of common support.

4. Propensity Scores

One problem with matching is that often we have a very large number of potential covariates that we can use for matching. With more and more variables, it becomes increasingly impossible to find matches for the treated units that are in any way close to the control units across all of the covariates at once. This is sometimes known as the "Curse of Dimensionality" in the matching literature. One common way to solve this problem is to use propensity scores as the basis for matching. Propensity scores reduce this multidimensionality to a single dimension.

The propensity score for a unit i is simply the probability that i is assigned to treatment, given their covariates X:

$$\pi(x)_i = Pr[D_i = 1|X_i]$$

This can be estimated using a probit or logit regression of treatment status on the baseline covariates. For instance, we can simply carry out a logistic regression of D on X to estimate:

$$\pi_i = logit^{-1}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$$

This estimates the predicted probability of being in treatment for every unit. After choosing variables to include in the propensity score model, matching can then proceed as before using steps 2-7 above. That is, each treated unit is matched to a control unit with the most similar propensity score. The advantage of propensity score matching is that the curse of dimensionality goes away. All of the baseline covariates are summarized into one single-dimensional measure.

Conceptually, propensity scores make sense for causal inference because they again capture the logic of trying to re-create a randomised experiment conditional on observed confounders. They are just another way of balancing our treatment and control groups: we compare treated units to control units whose profile of covariates makes them equally likely to have selected into treatment. For example, if the outcome is smoking and the covariates with the biggest coefficients in the logistic regression are age and sex, that means that age and sex are highly predictive of smoking. Matching on these propensity scores should ensure that we end up comparing smokers and non-smokers of similar ages and sexes.

Propensity scores remain widely used, especially in the field of public health. However, recent work has cast strong doubt on whether they should be preferred as a method for matching.[5] I have included them here because they are so common in the type of literature that you are likely to read, but you should know that they are now considered a controversial technique.

5. Empirical Example: Ethnicity and Conflict in Chechnya

In a famous recent article from the political science field of conflict studies, Jason Lyall examines the conflict between Russia and Chechnya that occured in the second Chechen civil war from 1999 onward.[6] He seeks to understand a particularly pernicious tactic used by Russian forces to suppress resistance from native Chechens: enrolling Chechens themselves into the Russian forces. He argues that these Chechen forces fighting for the Russians may have been better able to quash Chechen rebellions because they enjoyed greater legitimacy amongst their own people. This has clear implications for understanding the tactics used by invading forces in many conflicts around the world, where the co-option of locals into invading forces, sometimes as mercenaries, is a common tactic.

Obviously this is a situation where it is impossible for an analyst to use any kind of experimental technique! Instead, to test the idea of a "co-ethnic" advantage in quashing rebellion in civil wars, he collected data on so-called "sweep" operations carried out in Chechnya against native Chechens by both Russian and pro-Russian Chechen units. These operations aimed to terrorize locals and encourage them to give up their struggle against Russia. He also collected data on the number of anti-Russian insurgent attacks carried out after these sweep operations, in

[5]The reasons are quite techical. See Gary King and Richard Nielsen (2019), "Why Propensity Scores Should Not Be Used for Matching." *Political Analysis* 27 (4), 435-454

[6]Jason Lyall (2010). "Are Coethnics More Effective Counterinsurgents? Evidence from the Second Chechen War" *American Political Science Review* 104 91)

order to gauge their effectiveness. Therefore his outcome variable is the number of "post-sweep" attacks by Chechen insurgents against Russian forces. His main treatment group consists of villages where sweeps were carried out by pro-Russian Chechen forces, and his control group consists of villages where sweeps were carried out by Russian forces.

This is difficult to study observationally because attacks by Russian vs. pro-Russian Chechen units were non-random. They targeted different types of areas, whose potential outcomes may have been very different. For example, pro-Russian Chechen forces may have been sent deliberately into "less friendly" areas who would have been more likely to launch subsequent insurgent attacks anyway. To get around this he uses matching, creating a matched dataset of only sweeps by Chechen and Russian forces on similar types of villages. He then uses regression on the matched dataset to find his treatment effect.

He finds that the Russian use of Chechen mercenaries was probably very effective: raids by the pro-Russian Chechen forces were marked by a decrease in subsequent Chechen insurgency, whereas no such decrease is seen after raids by the Russian army. These findings paint a disturbing picture of the deliberate recruitment of Chechens to fight against their own countrymen.[7]

[7]Nonetheless, Lyall also appears to make a significant error in his matching that may introduce post-treatment bias. See if you can spot it.