政治科學論叢／第 77 期／民國 107 年 9 月／頁 1-32

Gary King on Simplifying Matching Methods for Causal Inference*

Gary King
(Transcribed and Edited by Stephen B. Reynolds)

On May 30, 2018, Gary King, the Albert J. Weatherhead III University Professor at Harvard University, gave a speech at the International Conference about Innovations in Political Methodology and China Study, which was held at National Taiwan University. His second keynote speech focused on the benefits of using alternative matching methods for statistical analysis, and the dangers of using the most commonly used matching method, Propensity Score Matching (PSM).

King showed how to use matching in causal inference to reduce model dependence and bias. He introduced two matching methods: Mahalanobis Distance Matching (MDM) and Coarsened Exact Matching (CEM), and explained how these methods are simpler, more powerful, and easier to understand than existing approaches. The discussion that followed addressed some of the concerns with using matching methods, and why King believes matching to be the superior way to preprocess data before running regressions.

This speech was transcribed and edited by Stephen B. Reynolds, a Ph.D. Candidate at the Department of Political Science at National Taiwan University. For more information on *Simplifying Matching Methods for Causal Inference*, please refer to: GaryKing.org

(The original graphs found in this paper are in color: black, red and blue. To see the original color versions, please download the paper at: http://politics.ntu.edu.tw/psr/?hl=zh_tw)

---

* DOI:10.6166/TJPS.201809_(77).0001

Gary King on Simplifying Matching Methods for Causal Inference:

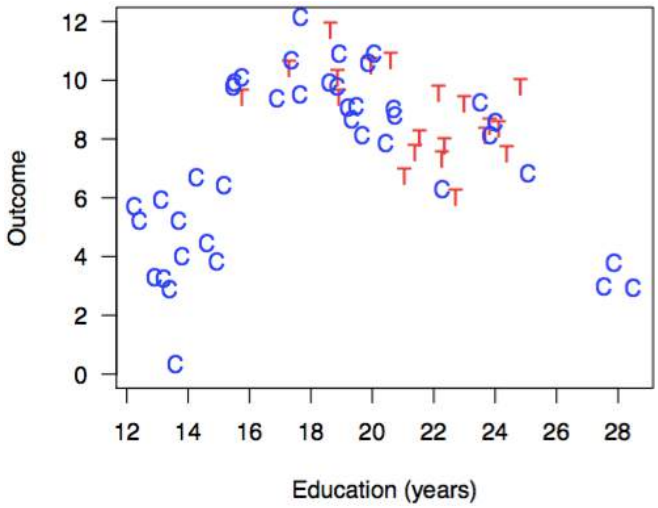
This speech is about causal inference and matching. First of all, the most popular method of matching is called Propensity Score Matching (PSM), and it sounds almost magical. It is very difficult to find two people who match in all respects, one who receives treatment and another who doesn't receive treatment. The more covariates you have, the more informative your analysis is, but the harder it is to find a match. The idea of PSM is that there is a way of taking the multitude of variables that describe each person, because it is very difficult to find another person who matches on all of the variables, and through the magic of propensity score, squash them down into one variable so that it is easier to find a match. It sounds magical, doesn't it? It always bothered me when I learned it, and I always thought that there was something wrong with it. It took us a long time, but we wrote an article: "Why Propensity Score Should Not Be Used for Matching." Propensity scores are great, and matching is great, but the two together are a disaster. I'll explain why later.

Do powerful methods have to be complicated? We came up with a method of matching that's very clear, very simple, and turns out to be statistically very powerful. It's called Coarsened Exact Matching (CEM). It's so simple that if I were going to teach an intro course again to new students, and I wanted to teach causal inference, I would teach CEM before I taught regression. Usually students learn causal inference, then regression, then the problems in regression, then try to correct for the problems in regression, and then they might do matching. However, I think it's actually much simpler to start with matching because it conveys very clearly what the control and treated groups are, and that's the essence of causal inference. When teaching freshmen how to do this, and they realize themselves that they are not going to be able to find an exact match, they will invent the idea of taking a nearby point. How do you take a nearby point? You find someone who is close. How do you measure close? You have to extrapolate. How do you extrapolate? Why not draw a line? I think this way is much more motivating, and this method is very simple and powerful.

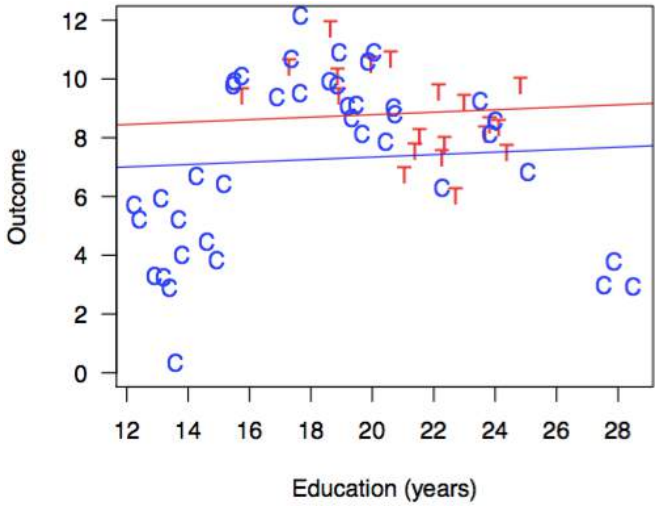
Matching methods tend to optimize balance between the treated group and the control group. The point of matching methods is to get a treated group and a control group that are the same prior to treatment. If you only give the healthy people the medicine, and the unhealthy people the control, then it's going to look like the treatment had an effect even if it didn't have an effect. You

want balance, but at the same time, the way you get balance is to prune observations. It's counterintuitive, but you throw away certain observations and then you can get the treated and control groups to be more alike. However, you don't want to throw away too many observations. The point of observational data analysis is to collect information, so most methods either optimize imbalance, or they try to get the largest number of possible observations, and you have to see whether you achieved any balance at the end of the procedure. You really should do both, and so we have a method that does both (The Matching Frontier).

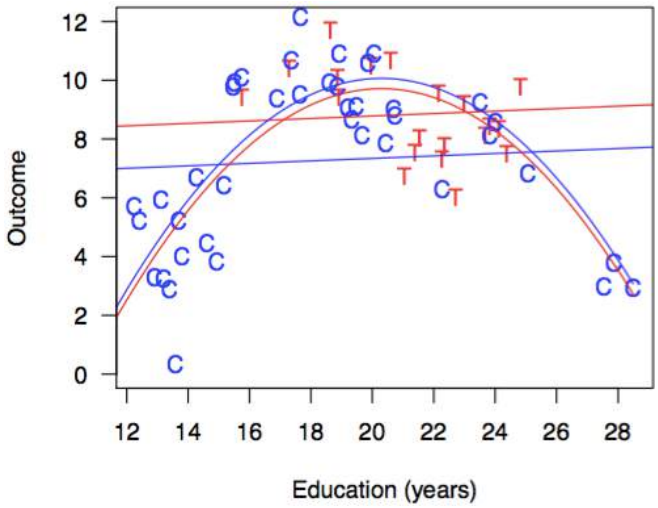
Let me start with matching to reduce model dependence. This is a figure from a paper that we wrote in *Political Analysis* in 2007:



Education is on the horizontal axis, and since this is a hypothetical dataset, I didn't even go so far as to make up a name for the outcome variable. There are some treated units, where the treatment variable is a "1", and there are some control units, where the treatment variable is a "0". The goal here is to figure out what the effect of T vs C (Treatment vs Control) is on the outcome, holding education constant. The traditional way, the way we would have learned in graduate school, is to run a regression. The key causal variable is T vs C, so it's a dichotomous variable, and the control variable is education. So after running that regression, this is what it looks like:



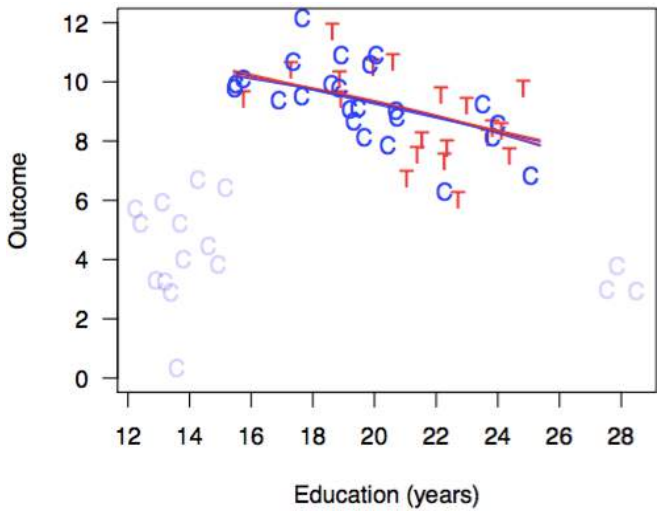
The Cs fit to the bottom line (blue) and the Ts fit to the top line (red). This means the causal effect is bottom (blue) to top (red). If you apply the treatment, the outcome goes up. Yet, you may think to yourself: "Wait a minute, I was going to write up an article where the effect was going to be negative, and this is positive," so what's a researcher to do? You may look at it and think that it doesn't really fit that well, so maybe it would be plausible to have a quadratic equation instead of a linear equation. The word "plausible" has gotten us into more trouble than any other word. So after running a quadratic equation, which looks like it fits a little better, the causal effect is now negative and runs from top (red) to bottom (blue):



This is convenient for the analyst, because we can pick whether we want to write up results with a positive effect or a negative effect. This is the definition of model dependence. It's the problem that matching really helps resolve. Let's forget these equations and think about how we can analyze the same data with matching.

Matching prunes observations. I don't think it should have been called "matching", rather it should have been called "pruning", because that's basically what it does. You prune away certain observations under very specific conditions so that you don't create selection bias, this results in a dataset with less model dependence, which means less changes to the results due to small decisions. After all, do you know whether it should be linear or quadratic in most applications? You don't really know, and there is no real theory that says it has to be quadratic. It's hard to really justify what it is, and it would be much better for us if the results did not depend upon deciding whether to put a squared term into the regression, or deciding whether to log the dependent variable, or whether to put a prior on the analysis, or whether to drop the outliers, or deciding what the sample period should be, and so on. We make these little decisions, and they have big effects, which end up giving us a lot of heartbreak.

The matching method pruned away the observations that have been grayed out, then fit the linear regression to the remaining data. The way we deleted the observations does not create selection bias because it only turns out to be a function of the explanatory variable and not the dependent variable:



For the linear regression, the causal effect goes from top (blue) to bottom (red), so there is no change. The causal effect is about zero now, meaning that T vs C controlling for education doesn't have an effect on the outcome. Our goal is to reduce model dependence, so we also ran a quadratic equation. It is already on the graph, so if you look really closely you can see it. The really interesting thing is that the linear model and the quadratic model are now giving basically the same answer. There is no more model dependence. We've gotten rid of it by doing matching, which is the point. It's not some complicated new procedure, it's relatively simple. Think of it as preprocessing. You have some data analysis procedure that you are going to do, and this is just preprocessing the data for it. Matching prunes the data in a very specific way, and then you do whatever it is that you were going to do originally with the dataset. It's very simple and convenient, and you don't have to change your method of calculating standard errors or anything like that. You get to say, when you are writing up your results, that you've reduced model dependence so those reviewers out there can't get you. Also, when you are a reviewer, you should make the author do this because then you can have more confidence in the results.

This is the general setup. Let's go a little deeper and explain it in more detail. Without matching, imbalance between the treated and control groups leads to model dependence. If you have exact matching between the treated and control groups, it doesn't matter what model you run, you are pretty much going to get the same answer. Imbalance leads to model dependence, and model dependence leads to researcher discretion. If there is model dependence, then we get to decide which model to write up: the linear one or the quadratic one, along with all of the millions of other decisions we have to make. Don't think of yourself, rather think of other researchers. That other researchers have too much discretion, because when they have discretion, it leads to bias. That's not a cynical remark, that's actually based on scientific psychological research. If somebody who is doing data analysis has discretion, they *will* bias the results in their favor. In the event that a qualitative choice is made among a whole bunch of estimators, or among a variety of unbiased models, you will get a biased estimator. Let me give you an example. Suppose you ran 50 randomized experiments, every one of which was done in exactly the right way. You are going to get different answers for the 50 experiments simply due to random variability. Now suppose you get to pick one to write up on the basis of the results. You could try to not be biased, but what's actually going to happen is that you will be slightly in favor of your a priori hypothesis.

So what happens if you try really hard to avoid biases while doing analyses? Typically, we don't have a lot of unbiased analyses to choose from, some are biased and some are not, and we don't know which are which, but we have to make a choice. We get to peek at the outcome and the result that we could write up. That peeking is going to cause bias. What happens if we understand this and try really hard to avoid biases? Psychologists have studied this and found that trying hard to avoid human biases does not work at all. This is because you don't have access to the mental processes that are causing you to make biased decisions. We all have the flaw of being human. What if we took experts and we had them choose among a whole bunch of analyses on the basis of the results? Well, it turns out that experts overestimate their ability to control their own personal biases even more than non-experts, and the most prominent experts are the most overconfident. So, if you are an expert that's a bigger problem. Suppose we take experts and we teach them about these psychological results, and we explain to them that we tend to bias things imperceptibly in our favor. As Nobel Prize winning psychologist Daniel Kahneman explained: "Teaching psychology is mostly a waste of time." Teaching experts about these things basically doesn't work.

So what happens with matching? With matching you get rid of imbalance, you get rid of model dependence, you get rid of researcher discretion, and you get rid of bias. Once you have balance, you're OK. There aren't a hundred decisions to make, and when there are, you can get the same answer regardless of what your decision is. Model dependence is an incredibly serious statistical problem. In fact, if you think of the goal of statistics, the purpose of quite a lot of it is automating away human discretion. That's the reason why we quantify things at the end of the day. If humans were good at looking at a bunch of data by themselves and picking out the right answers, we'd be fine, but that's not how it is.

Question:

It seems that one of the problems is that researchers get to pick the results before they choose what to write down. So if this is the case, the implication is that the entire journal review process is flawed, because the editors and reviewers also get to see the results before they decide if they should publish an article or not.

Gary King:

That's a good point. So there is a way around this which sometimes works and sometimes doesn't. It's called preregistration. The idea is that before

looking at the results and collecting the data, you decide what you are going to do, what data you are going to collect, and which analyses you are going to do, and that's it. It used to be that you had to do this in educational psychology. If you were writing a dissertation, you would write a decision tree and decide that if you ended up with a certain result, then you would conclude you were right, and if you ended up with another result, then you would conclude you were wrong, and that would be it. If you were wrong, you would still get your Ph.D. and then maybe write a concluding chapter talking about future research. I find that there are times when preregistration is very helpful, but I also tend to learn a lot during the process of doing data analysis. So, for example, for the talk I gave on "Reverse Engineering Chinese Government Information Controls", if I had preregistered that study, I would have just given you a talk on automated text analysis, and not even necessarily about China. So, preregistration is sometimes the wrong thing to do, but other times it can help protect you. For example, I did a big study in Mexico where we randomly assigned hospitals to different communities. I flipped coins in my office and with heads a community got hospitals and doctors, medicine and money to pay for it all, and with tails the community would just continue to not get healthcare for at least three or four more years. It was the evaluation of a program. I certainly wanted that study preregistered because I wanted to conclude that it was either working or not working depending upon the data, and not depending upon the throngs of local officials who were going to be fired if it didn't work. So, I brought them all into a room, and I drew a line and told them that if the results were below that line it was going to be: "It didn't work", and if the results were above that line: "It's going to work". So, as much as possible, I tried to tie my own hands, so that afterwards they wouldn't accuse me and tie me up.

Question:

By doing that, do you also create an incentive for the local officials to put in every effort to make it work? Doesn't that then create a problem of scaling up?

Gary King:

That's an excellent point. We definitely did give them an incentive to make it work. The officials in that program already had incentives because their jobs depended on it, and the healthcare of Mexicans depended on it. We also gave them an incentive to comply with the experimental treatment and to stop non-compliance. We would randomly assign a community to receive healthcare. It was easy to verify whether or not the hospital had been built, but people in the community also had

to go affiliate themselves and their families to the program, while in the control groups they had to not be affiliated, and so the officials had to make sure that these things happened. We tried to set up incentives and use them to our advantage.

Question:

If you have a good theory that can guide you in testing whether you are supposed to use a linear or non-linear relationship, could that reduce the risk of model selection bias?

Gary King:

Yes, absolutely. It's not only about having a theory, but that theory has to be right. If we have a theory that we really believe to be true, ideally on the basis of prior empirical evidence, then there wouldn't be a problem. The problem is that in the social sciences we're doing so many different things at the same time. I remember when during the AIDS crisis in the 1980s, I was on a plane sitting next to somebody studying HIV, and he was going to a conference where there were 16,000 people studying one virus. Yet I was going to the American Political Science Association conference with 6,000 people studying every conceivable thing having to do with politics. That means that basically every person attending the conference was studying something different. If you had something similar to the HIV conference where everyone was focused on one thing, then you might be able to develop theories that are really thoroughly tested, in which case they would be fine.

So what is matching? Well, we have a dependent variable which is Y, and we have a treatment variable which is the treatment or the control, which we can generalize. We could have more categories, and there are other things that we could do, but I'm going to stick to the simplest case for the purposes of this talk. There are also pretreatment confounders which are things that are causally prior to the assignment of treatment vs control, and they account for all the important differences that treatment and control have on the outcome. It's easy for me to say: "We have all of the confounders," but that would actually be a huge decision. We have to make sure that we really know what the differences are between them. If we don't know them, then it would be nice to be able to run a randomized experiment, where T is then random. If it's random, then it's unrelated to any X that you thought of, and also unrelated to all of the Xs that you didn't think of. We can't always do this, however, so we are stuck with choosing our own Xs. This is our setup:

$$\begin{aligned} \mathrm{TE}_i &= Y_i(1) - Y_i(0) \\ &= \text{observed} - \textit{unobserved} \end{aligned}$$

This is the treatment effect (TE) for the treated observation, which is *i*. The TE for one observation is the difference in the two potential outcomes. One is observed, and one is not observed. For a treated observation, *i*, "1" means it's treated, and "0" means it's not treated. They are not two different observations, rather they are the same observation. In fact, we can just get rid of the "1" because we can see that it is the value of the dependent variable, and this is the value that it would've been if that observation had not become inept.

One nice way to think about matching is: "We don't know *that*; we need to estimate *that*; where are we going to get *that* from?" We look two observations which are the same based on the pretreatment confounders. One got treated and one didn't. If it's possible, that's what we would like to find, but we could estimate it if we have a big enough control group. We don't estimate for every single observation, because there is a lot of randomness, so we'll look at the mean of the treatment effects over all of the treated units, which is the Sample Average Treatment effect on the Treated (SATT), or you could look at it over all of the units. There is one other quantity of interest to focus on, which other than the SATT, is the Feasible SATT, or FSATT.

The FSATT is basically the same as the SATT except that because some of the treated units don't have a control unit nearby, you prune those treated units too. However, you have to be careful, because you are changing the quantity of interest. It's a little weird, but it's part of the statistical procedure and it's perfectly fine. Do you know the story of the guy looking for his keys under the lamppost? He is a drunk looking for his keys, and he's looking under the lamppost and they ask: "Why are you looking under the lamppost?" And he answers: "Because there is light here." He's actually doing the right thing. He has no chance of finding his keys anywhere else, so he looks in the place where he could conceivably find them. That's what we do as social scientists. We can then follow this preprocessing step with whatever statistical procedure we would have otherwise used without matching, and all of the influential statistics then follow. We prune observations like a game of musical chairs, that is, we prune the ones that don't get matched. This makes the control variables, or confounders, matter less. Therefore, the modeling that is necessary to control for co-variables doesn't really matter, and we reduce

imbalance, model dependence, researcher discretion and bias. That was the slightly more technical explanation of matching.

Let me give you another perspective on what matching is. You have a big, messy observational dataset, and you would like to randomly assign treatment and control, but instead you do matching. So, one way of thinking about matching is that we look inside this big, messy dataset for a subset of the data which is pretty close to what it would have been if we had randomized. Matching is looking for the pristine, hidden, randomized experiment inside the observational dataset. There are many types of experiments, but there are two that are particularly important in this case. The first is called "complete randomization". It's the standard, classic approach. For example, we could take one person and flip a coin: heads the person gets the medicine, and tails the person gets the control, and we flip one coin for each person. That's how complete randomization works.

The other experiment is the "fully blocked" experiment, a special case of which is the matched-pair experiment. This is where we take one person and then look for another person who is exactly the same in all respects. We then flip a coin for the two of them: heads one person gets the treatment and the other person gets the control, and for tails the reverse. For example, let's say two people match just on gender: let's say two males. We flip a coin, and then we have one male in the treated group and one male in the control group, so we match on that exactly. If there are ten variables that they match on before flipping a coin, then those ten variables would match exactly. If we instead did complete randomization, all of the men could just by chance end up on one side and all of the women on the other side; or all of the healthy people could end up on one side and all of the unhealthy people on the other side. If n is larger, this is not going to happen, but we only have so much time and money, we'd like to finish one project and get on to the next one, and we can't collect data forever.

The goal of these two experiments is to balance covariates. There are two kinds of covariates: the ones you know about and measured – the observed ones – and the ones you either don't know about or didn't measure – the unobserved ones. Under complete randomization, on average the observed data is balanced if you get enough observations, and thanks to the miracle of randomization, you also balance the unobserved data, which are the things you didn't see or didn't think of. However, a matched-pair experiment is even

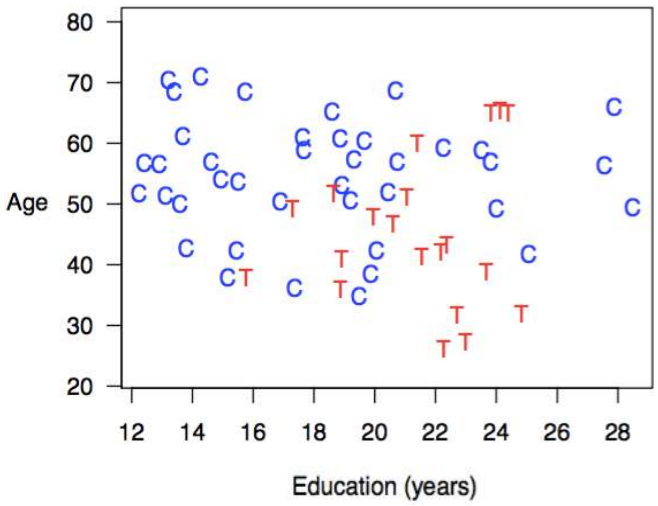
better. This is because there is exact matching on the observed variables, and on average still balance between the unobserved variables.

Fully blocked experiments dominate complete randomization because they have less imbalance, less model dependence, more power, more efficiency, less bias, more robustness, and fewer research costs. If you are running an experiment in graduate school, you will get out of graduate school faster if you use a fully blocked experiment. In the paper where we described the experiment that we ran in Mexico (Imai, King, Nall 2009), we were able to estimate how big the standard errors would be under each of these two designs, and we found that the standard errors were as much as 600% smaller with a matched-pair experiment than with a completely randomized experiment.

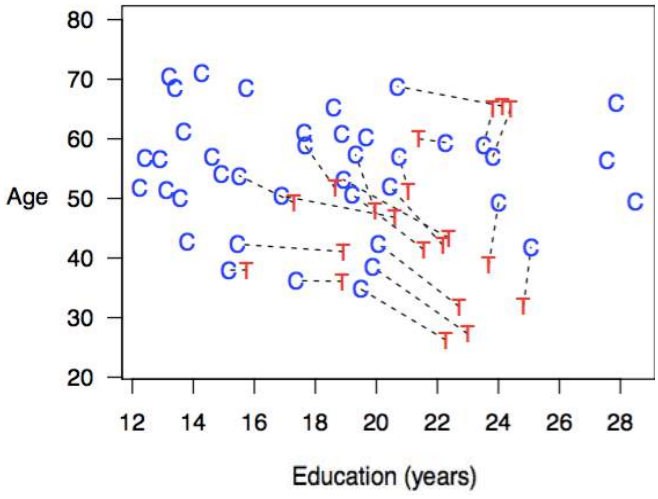
If Propensity Score Matching (PSM) works really well, its goal is complete randomization. For every other matching method, the goal is to achieve a fully blocked experiment. So, PSM has lower standards and doesn't try to achieve as much. Of course, if you were able to get to complete randomization, that would be pretty good for an observational study, but if you could get to a fully blocked experiment, that would be even better. Other methods dominate PSM – they are just uniformly better. However, it gets even worse for PSM, in a surprising way that took us years to understand.

I will now describe three methods of matching. The first is Mahalanobis Distance Matching (MDM). It approximates a fully blocked experiment, not merely complete randomization. The general idea is to first preprocess the data with matching, and then run whatever statistical procedure you are going to run afterwards. The way it works is that you take each person, or unit, that received treatment and you figure out the distance to another unit that didn't receive treatment. You measure distance with Mahalanobis distance, which is standardized distance. In general, you shouldn't standardize your variables, because standardization basically throws away the substance, and you don't really want to do that, so it's usually better to use Euclidean distance instead. But in any event, this is just a mathematical way of taking each observation, which has a certain amount of variables, or pretreatment confounders, and figuring out the balance between their variables. You then match each treated unit to the nearest control unit. Control units are not used more than once and are pruned if they aren't used, and once all the matches are set up, we then put a caliper in place, which is the largest distance we are willing to tolerate. Beyond that caliper, we also delete the treated units. That's the basic idea.

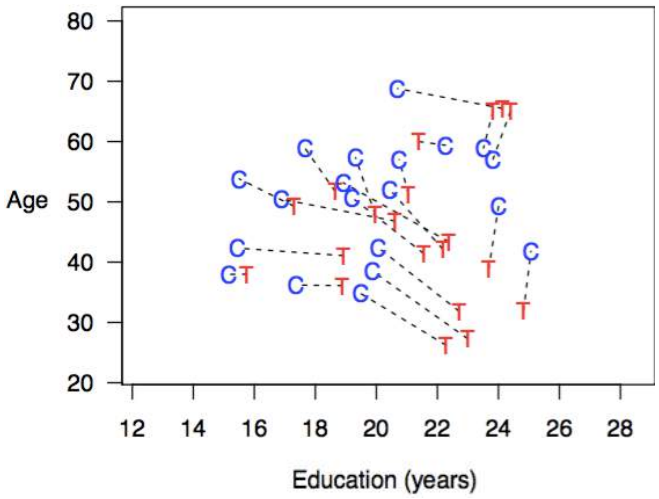
Let me show you a visual version of this. I have two explanatory variables, or two confounders: age and education. Here are some treated units and some control units:



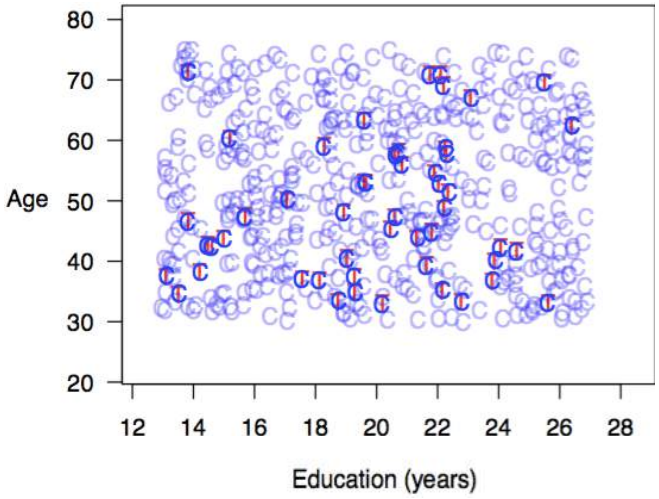
In this graph there is no dependent variable; these are two explanatory variables. We have treated units and control units. We take each treated unit and find the nearest control unit within the pretreatment confounders, which in this case is simply age and education – measured in Mahalanobis distance. This is the answer:



The ones that don't get matched lose the game of "musical chairs", so they go away. This is now our pruned dataset, and you can now do whatever you want with it:

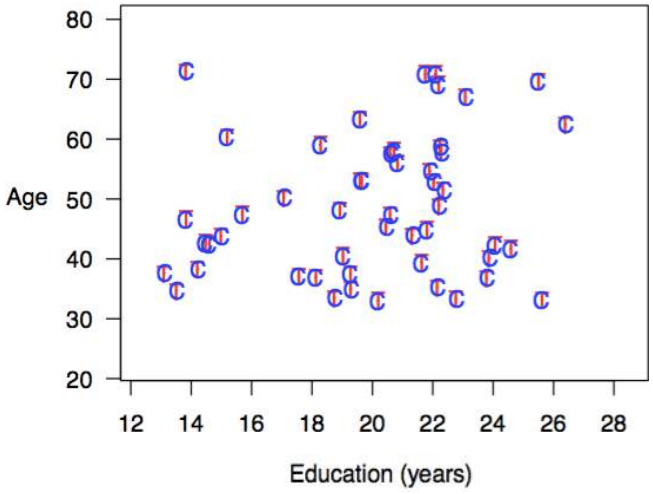


Let me give you one more example of MDM, to give you a feel for the best case scenario:



Again, within age and education, for every treated unit there is a close control unit. They sit right on top of them. There is also a bunch of other

control units which we really don't need, so we'll just prune them and make them go away, and this is our dataset:

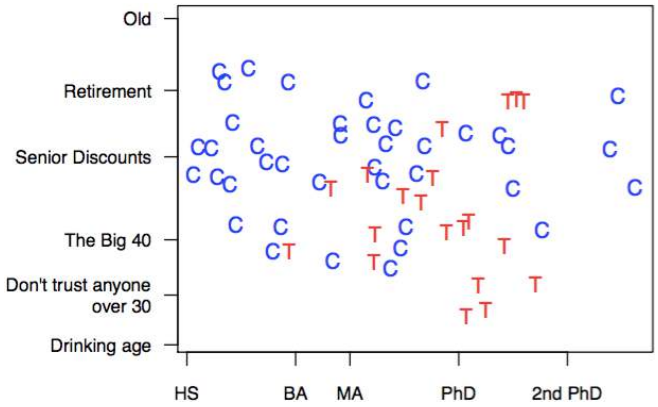


This is a wonderful and beautiful dataset, because no matter what model you run, you are not going to be able to predict which unit is treated or controlled. Education and age have no effect on how likely the unit is going to be T or C. That's the advantage of this method and this is how it reduces model dependence. This is the best case for MDM and everything works just as you'd expect it to. It's not going to be this way for PSM, however.

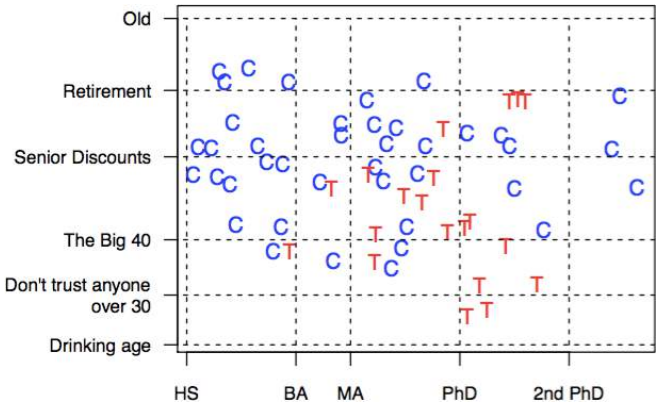
The second method is Coarsened Exact Matching (CEM). It's an easy method, and I think it's the most powerful and easiest to use approach. It approximates a fully blocked experiment, not merely a completely randomized one. It works by preprocessing, matching, and then estimating. Here is how matching works for CEM. You temporarily coarsen X as much as you are willing to. For example, if "years of education" is your variable, we in data analysis would sometimes coarsen this to "grade school", "high school", "college", "graduate school", etc. It's a serious data analytic choice as to whether you are willing to coarsen it or not, but I think we all understand what it is. Why coarsen? Well, because you can't find an exact match within the original variables. Since it's very difficult, we make it easier in this very specific way. Sometimes doing it this way is better because you wouldn't want to match a college dropout with a first year graduate student, but you might be willing to match a graduate student with somebody who has almost gotten through college. So, the points where you decide to cut can be really important. You then do exact matching within the coarsened X,

meaning you take the coarsened variables and find two people that are both in graduate school, for example, and you match them. You do this not only with that variable, but with all the other variables as well. You then sort the observations into strata, each with unique values for all the control variables on the coarsened scale. You prune any stratum that has either zero treated units or zero control units. This is how CEM works. There is one slight difference when estimating which is that you may have to use some weights.

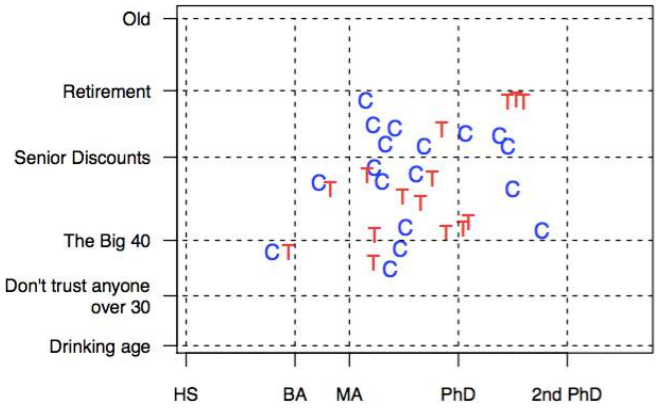
In this example, with age and education, instead of looking at each treated unit and finding the nearest control unit, we break up age and education into these coarsened limits. These categories are intended to be funny:



You have to set meaningful categories, and once you have them, you throw away the original scale temporarily. This works not only with two variables, but also with any number of variables.



If within a bin there is at least both one treated unit and one control unit, we keep all of the units in the bin. So we are going to keep these:

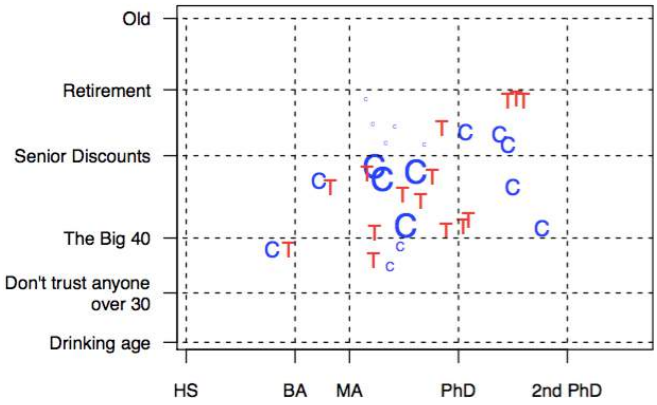


If there is only one unit in a bin, then there is nothing to compare it to, and this also applies to the unmatched Ts, since all control units are infinitely far away from them, we throw them away too.

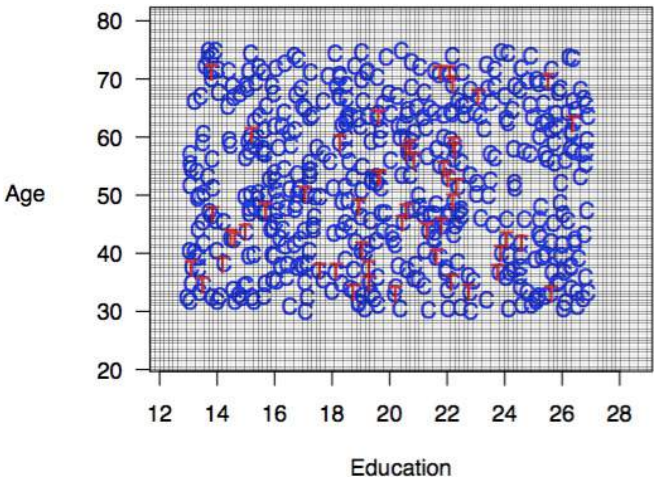
Question: So far you have always had more units in the control group and fewer units in the treated group. Is there any reason for this?

Gary King: There are a few reasons for this. Of course, you could just swap C and T, so it doesn't really matter, but it's easier for me to say that I'm going to keep all of my Ts and only prune the Cs, which keeps the quantity of interest the same. There is nothing saying that the treated units are the people receiving the medicine, and that the people receiving the placebo are the control units, so you could simply switch them. However, the reason I'm doing it this way is that I'm making my quantity of interest a causal effect for the treated units while pruning away some controls, and this way I don't change the definition of the quantity of interest. Yet, I'm also telling you that it's OK to change the quantity of interest. It's often the case that we are in a situation similar to this, but not always.

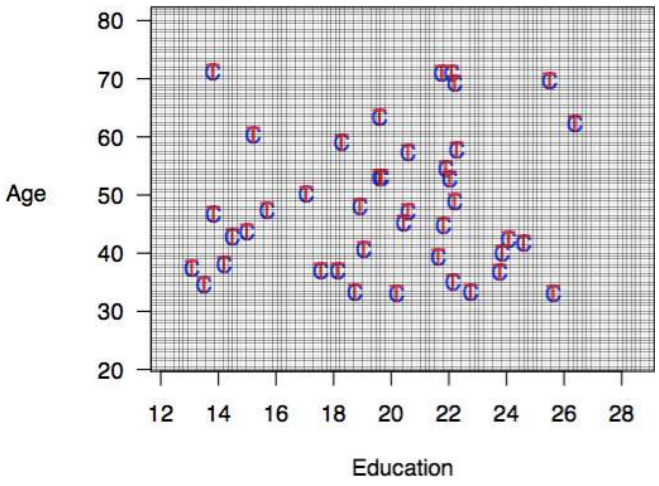
What's the causal effect within a bin? For example, in the bin with one T and five Cs, you could take the average value of the Cs and subtract it from the value of just that one T, and that would be the causal effect:



You could either average them within each bin, or you could tag one whole dataset with the weights. So that's CEM. Let me show you the best case scenario for this method:



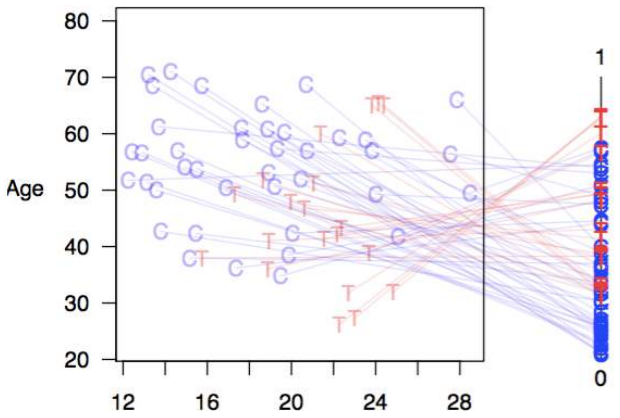
Every treated unit is essentially on top of a control unit. We have good matches and throw away anything that doesn't match. This is a beautiful dataset where the confounders have no effect on predicting what's treated and what's control. This is what it ends up as:



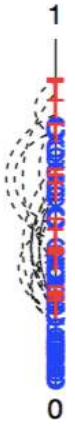
The last method is Propensity Score Matching (PSM). This is the most commonly used method, and it has been used in roughly 100,000 scholarly articles. It approximates a completely randomized experiment, not a fully blocked experiment, so it has lower standards. We use it to preprocess, and then do estimation by any method we want. How does it work? First we reduce all of our covariates into one variable. How do we do that? We run a logit of the treatment variable. The treatment variable is temporarily the dependent variable, and the explanatory variables in this logit are all of the confounders. The predicted value of treatment is called the propensity score, which means it's the propensity to receive treatment. This is the equation for the logit model:

$$\pi_i \equiv \Pr(T_i = 1 \mid X) = \frac{1}{1 + e^{-X_i \beta}}$$

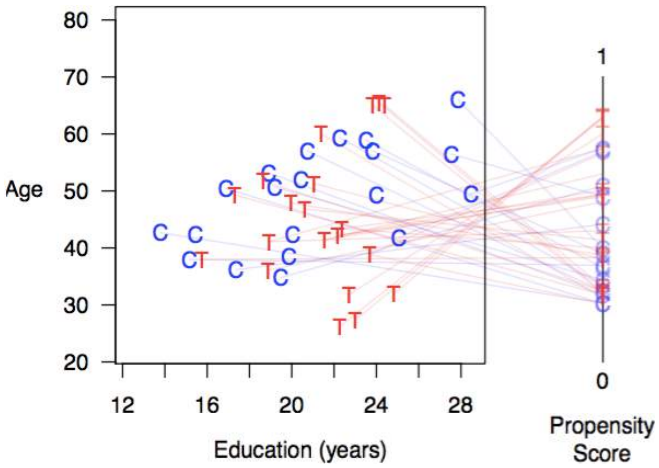
You then calculate the distance, which is the absolute distance between a treated unit and a control unit on the propensity score scale, and you then match each treated unit to the nearest control unit on that scale. Control units are not used more than once, and we prune matches if the distance is greater than some caliper, which is how large of a distance you are willing to tolerate, and then there are many other adjustments to be made. Let me show you this visually. This is the estimated probability of receiving treatment. It goes from low to high. We run the logit, and it predicts T vs C as a function of age and education. We project the units over to the scale:



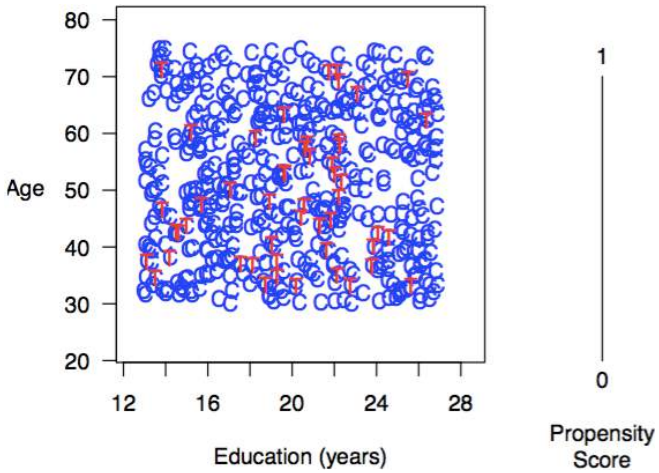
Every unit is now on the scale. We then temporarily throw away all of the data and only pay attention to this one dimension. So, as you can see, something that was two-dimensional has been reduced to one dimension. We might have something like 50 variables, and they would all be reduced to one dimension, so we are throwing stuff away, right? Once we have them all projected onto the scale, we then take each treated unit and match them to the nearest control unit:



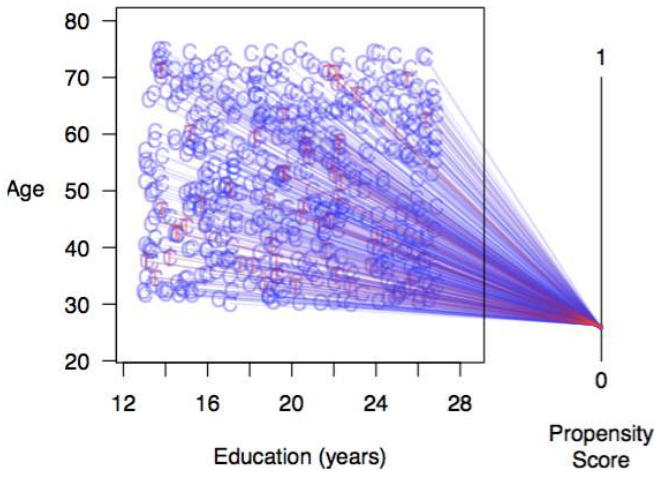
We then throw away the units that don't get matched, and this is now our dataset:



We project it back and now there are fewer observations. So, we ignore the original multi-dimensional scale for doing the matching, we match on the propensity score scale, and that's basically how it works. Let's look at the best case scenario for PSM. This is a beautiful dataset where each treated unit is right on top of a control unit:



We then project all of them over to the propensity score scale:



The best case, in theory, is that they all have the same propensity score. If you are running a randomized experiment where you are flipping coins, the propensity score is .5, meaning every unit has a .5 probability of receiving treatment. Now, I'm going to imagine in this particular example that everyone has a propensity score of .2, meaning that one out of five units receive treatment, and the rest receive the control. We are going to match them, but what does matching look like if all the propensity scores match exactly? How would you know which units to keep and which units to drop? You'd have to delete observations at random, but that doesn't seem like a good idea, does it? That's the problem. If propensity scores match exactly, which is supposed to be a good thing, then you'd be deleting observations at random. What dataset do you have where you would want to delete observations at random before running an experiment? There isn't any way that could be helpful, right? So, we started out with a beautiful dataset, there was an exact match for every observation, then we did matching by propensity score, and the dataset that we got as a result has Ts that aren't exactly on top of Cs. We used propensity score matching exactly the way it was designed to be used, and the result is that we didn't get exact matches. If you have a big complicated model, you could predict which of these units are in which place on the basis of covariates, to some degree, but this leaves some extra model dependence for no reason. Even though PSM is achieving its objectives, it's leaving model dependence on the table, and that's what we are trying and avoid. It's suboptimal.

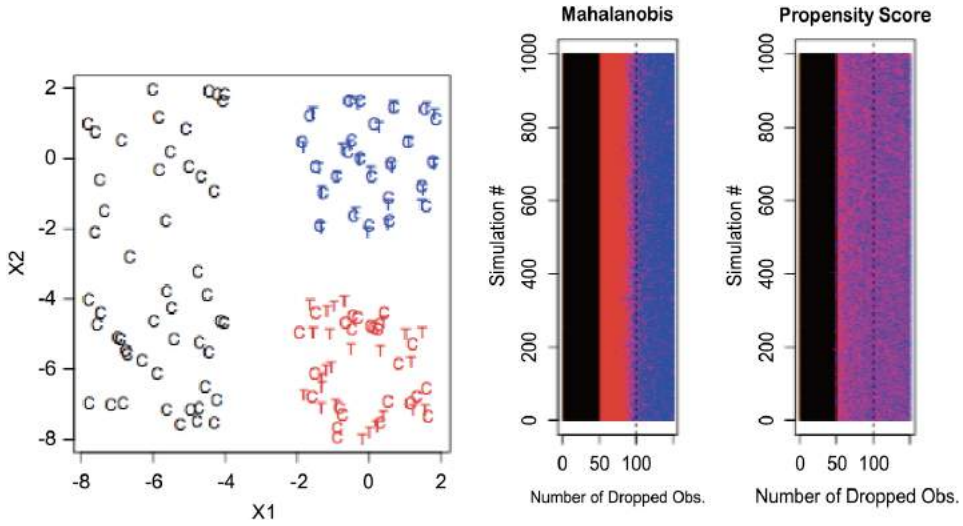
I will summarize. PSM's statistical properties have low standards. It sometimes helps, but it never optimizes. It is efficient relative to complete randomization, but it is inefficient relative to the more powerful fully blocked experiment. With PSM, if you have exact matching between the control units and the treated units, that implies that you have exact matching on the propensity score. It does not work the other way around. In the literature, everyone interprets this as working the other way around, but it does not work that way. It is not true that if you have exact matching on the propensity scores that you get exact matching on the covariates.

Second, and this took us a lot of time to figure out, is what we call the "propensity score paradox": when you do better you do worse. To begin with, random matching increases imbalance. Remember when I said PSM randomly deletes observations? Randomly deleting observations doesn't keep the treated and control groups about the same, rather it spreads them further apart, which is a surprise. Think of how far you are to the nearest person in the room. Now, imagine half of you, or every other person, randomly left the room. Now, think of how close you would be to the next person. You'd be further away, so the matches wouldn't be as good. If you get to all the propensity scores being the same, then you are literally matching at random. This means that you are pruning at random, which is producing imbalance, which leads to inefficiency, which leads to model dependence, which leads to bias. I'm making very strong claims here as to what around 100,000 scholarly articles have done, but let me show you an example.

If the data didn't have any good matches then the paradox wouldn't be a problem, but then you would be in trouble anyway because there wouldn't be any good matches in the data. People ask me at this point, doesn't PSM solve the curse of dimensionality problem? No, the curse of dimensionality problem is not something that can be solved, it's just a fact of the universe. In fact, the more covariates you have, the worse the paradox gets.

PSM is blind where other methods can see. Here is a simulated example where there is covariate one and covariate two. In the bottom right-hand corner, I created one set of randomly generated treated units and one set of randomly generated control units. This is a completely randomized experiment: random with respect to the first covariate and random with respect to the second covariate. They are not exactly matched, but it's still pretty good. Above that (the top right corner) is a matched-pair experiment where each

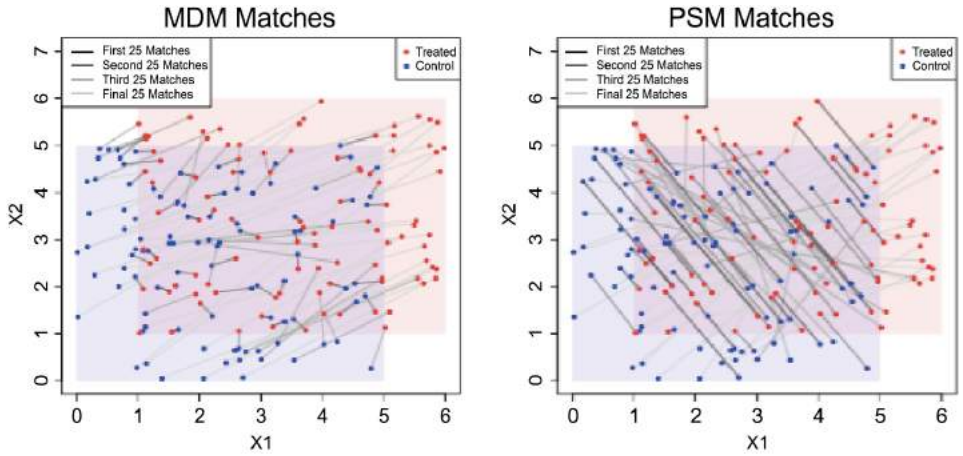
treated unit is matched almost exactly to each control unit (with a little bit of random variability). I also added a whole bunch of messy control units that are far from the treated units (on the left). I put these all together and imagined that we are going to analyze them all together. Let me explain what these two figures are on the right-hand side:



The first row of pixels corresponds to this particular dataset, and the other rows correspond to the other thousand datasets that I created. I only need to describe the first row, because the others all came out the same. Mahalanobis distance does the right thing: it first prunes the Cs on the left (black), just as you would want, because these Cs are very far away from the Ts. Then it prunes the ones in the lower right (red), because these Ts are somewhat far from the Cs. It makes sense: first you prune the ones on the left, then you prune the ones on the lower right, and then what remains are the ones on the upper right (blue). That's Mahalanobis distance from the left to the right.

What happens with propensity score? It starts with the left, and it does the right thing at first by pruning these units on the left. As for the two areas that remain, PSM can only see a completely randomized experiment and it can't see that a matched-pair experiment is better, so it doesn't know the difference between these two experiments. If you ask it which of these two to prune, it won't be able to tell the difference. As you can see, the ones in the lower right and upper right are all sort of random, and so they look the same to PSM. It's

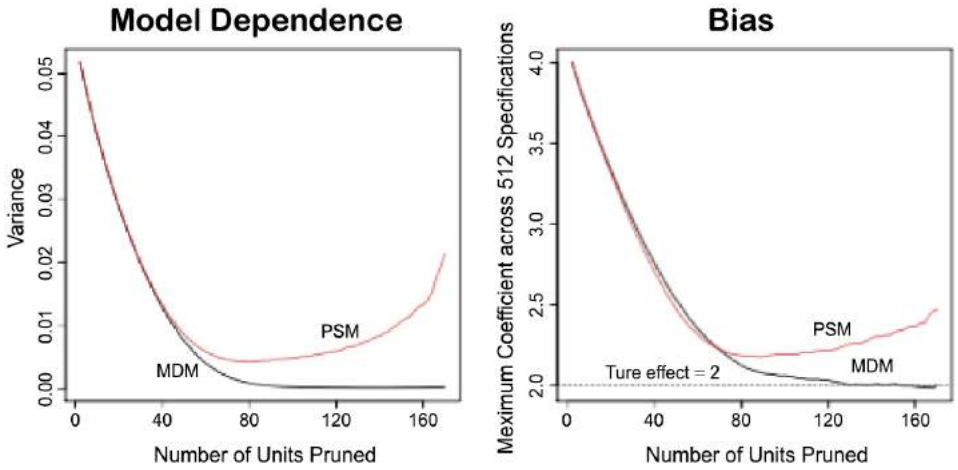
calculating distances based on the propensity score, not distances based on the data in the spaces. Let me give you another example:



Controls: $X_1, X_2 \sim$ Uniform(0,5)
Treateds: $X_1, X_2 \sim$ Uniform(1,6)

I created this lower left square (blue), and I randomly put control dots in that square (blue). The control dots (blue) are randomly spaced out within the square. In the upper right square (pink), I slightly moved it to the top right, and I put treated dots in it (red). The treated dots (red) are randomly scattered in this square. In the overlapping area, I have both control and treated dots (blue and red). We want the matching method to find the overlapping area, so this seems like a good case for PSM because there are no experiments that are exactly matched here. It is a completely randomized experiment, just as PSM intends to find. As for MDM, it matches each treated dot (red) to the nearest control dot (blue). It's a little hard to see, but the darkness of the line is the order in which they are pruned. It does exactly what you'd expect: when they are close, it's much less likely to prune them, while the ones it prunes have the longest distances, which are the treated dots (red) and the control dots (blue) which have to go far to find a match. This is what we would expect.

PSM is blind. It can only see in one dimension. Dots are matched even though there are nearby dots that are closer. It doesn't make sense, but it makes sense to PSM because it can only see in one dimension. Now let me go one step further, taking the same data and making a dependent variable with this model:



$$Y_i = 2T_i + X_{1i} + X_{2i} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, 1)$$

I am trying to measure model dependence, so how am I going to do that? I simulated what we might do in the privacy of our own office or dorm room. We would go spend two years or so collecting data, then we would set up the analysis, run our regression once and run it again and again with some adjustments, we would do all kinds of things that could generate model dependence, and they are for good reasons, but we know that discretion can lead to bias. How do I simulate this? I came up with a whole bunch of regressions that are simulated. Here's one regression, which is the top line (red). Then I ran a regression with X1 squared, then with X2 squared, then with X1 times X2, then with X1 times X2 and X squared. I ran a total of 534 regressions. I don't know what the truth is, but in each one of these 534 regressions, the point is to estimate the number "2".

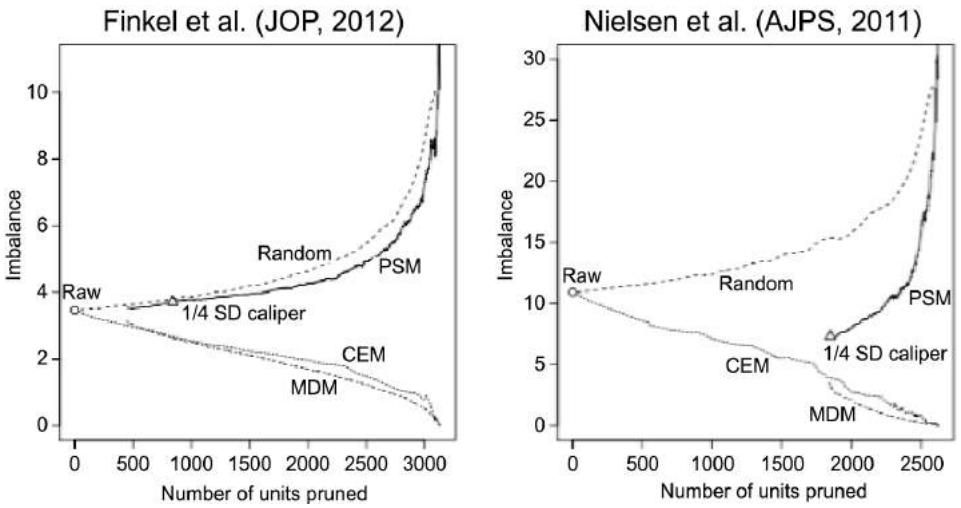
So what's model dependence? Model dependence is the variation across the 534 regressions. I ran the 534 regressions, then I calculated the variance across all of them. According to MDM, the worst match was found, I pruned it, then I reran all 534 regressions, and I calculated the variance across the estimates of the causal effect, and it dropped a little. I then rounded the worst match that was left and I deleted it. I ran the 534 regressions again, and I calculated model dependence again. As we deleted observations that were badly matched, we reduced the variance across the 534 regressions, and model

dependence was reduced. Eventually we got rid of all model dependence. That's what MDM does.

At the beginning PSM did the right thing, it went all the way down to the point where it reaches approximate complete randomization. However, at that point it's blind. What happened? It kept matching on the basis of random deletions and so it started to increase imbalance. By randomly deleting observations, it makes things worse. Rather, you should only be throwing away observations in order to make things better. It's like walking into a shoe store, giving them some money, and the proprietor saying: "May I please have your shoes now?" You expect a pair of shoes from him, but instead you give him money and you also give him your shoes. That doesn't seem like a very good procedure, but that's what PSM does.

The vertical axis is the variance across the estimates of "2", and it is reduced as you prune. Let's look at bias, which means estimating for the number "2". If we run 534 regressions, which one will you pick? I'm going to say that you might be slightly biased in favor of your a priori hypothesis. How do I model this in the simulation? I imagine that you just ran 534 regressions and picked the maximum one, which is the worst bias that you could have with all the extra discretion. With no matching, the estimate is about "4", even though it should be "2". With MDM, as you prune the worst possible observations or the worst possible matches, you reduce the bias essentially all the way down to "2". For PSM, it's definitely doing good things at first, and then it starts to take your shoes away, meaning it starts to make things worse and diverge away from "2". Now you could ask to me: "OK you had a lot of fun making up data, but what about real data?"

Here is some real data:



This is an actual article in the *Journal of Politics* where we started off on the horizontal axis with the number of units pruned, and on the vertical axis is imbalance. With MDM, we prune the worst observations, we calculate imbalance, and it goes down and continues to go down, which is exactly what you would expect. CEM is pretty much the same. Now, for comparison, I tried just randomly deleting observations, and of course imbalance got worse. Now, imagine if we had a sample survey and we were estimating the mean. What would happen as we randomly added observations to our dataset? The standard error is sigma over the square root of M, so as we get more observations, the standard error drops. It's the same thing in reverse. As we delete observations randomly, the standard error, variance, model dependence, and imbalance all increase. So, that's what randomness does. Now, what about with PSM? It also basically increases imbalance as we delete observations. This is an article from 2012, and this is an article from 2011 by my coauthor Rich Nielson. With PSM, as you continue to prune, it just makes things worse very fast.

When we first saw this graph, we really didn't understand what was going on. This is a method that has been used in around 100,000 articles, after all. So we advertised on the internet, asking: "Is anybody doing matching? Is anybody estimating causal effects? Would you like some help? Send us your datasets, and we promise not to publish them or tell anyone you sent them to us. We will do an analysis and send it back to you." We did this just so we could get more experience

with a diverse array of datasets. We got more than 20 datasets from people that wanted us to do their data analysis for them, and they all looked similar to this. The line always went up at the end. In real data this is actually very common. We then knew there was a systematic pattern, and then we figured out what it was. In observational data analysis, we push ourselves as hard as we can, but using PSM just makes things worse for us. So I'll stop here and take questions.

Question:

I lost faith in matching due to all of the problems that you described with PSM, but basically the method is trying to find the best counter factor. What are we losing by pruning the unused data? If you think of the question of continuity, you get a local average treatment effect by looking at different factors. What are we losing? Are we getting something close to a local average treatment effect with this method? What about counter factors?

Gary King:

The local average treatment effect we can almost think of as matching, except for where there is discontinuity. By local average treatment effect, what we mean is that the quantity of interest has been potentially changed. The effect after the discontinuity might be different, which we can't really test. Since it was done in an area where a reliable answer could be gotten, that's a useful contribution to knowledge. The things we are deleting are probably very similar to the observations that are further away.

The two methods can go together. At the discontinuity, there are typically layers of all of the variables. The key variable you would be focusing on is the discontinuity variable. The matches are quite good, but they are not perfect because some are above the line and some are below the line. It turns out that typically you may have twenty other variable measurements, and if it is the case that the ones before the discontinuity are the healthy people, and the ones afterwards are the unhealthy people, then that's a bad match. I would do both methods. If you have a discontinuity variable, that's a great thing, but I would also check with matching.

Question:

You describe the best case scenarios, but I keep thinking about the worst case scenarios. Your methods seem to be very data intensive, meaning that you

need to have a large *n*. You only use two dimensions here, but with other cases where there are more than two dimensions, like ten variables, then it's even harder to find a good match, so you need to have a lot of data to use this method, right? Even though you have a semi-large *n*, you talk about this method getting rid of human discretion, but sometimes I find that people choose the number of variables to match. So, if I cannot get a good match, then I can choose the number of covariates. So, there is still some human discretion when people choose the number of covariates. What would be your suggestion for this?

Gary King:

You need good matches. A lot of data may help you, but not necessarily. In some ways there is no way around any of this. If you have a good set of matches, then everything is good, and you don't have to worry about models. Suppose you don't have a good set of matches, perhaps because you didn't start with a large enough number of observations, perhaps because it just wasn't a good dataset. What do you do at that point? You would have to make a compromise with how you match. You could coarsen more, you could deal with Mahalanobis distances that are further, or you could drop some of the variables that you've decided are not as important. However, any of these moves would leave model dependence on the table. What do you do at that point? At that point you have to model. You've reduced some model dependence, but you are left having to justify the model. Theory is going to be the only answer. It's the only way that you are going to make any progress at that point. If you don't do matching, then you have more model dependence, period. We are in this situation together. It's not like you get out of this. If we don't have enough observations, that's just too bad. We would need more data.

Question:

Obviously in this case we are talking about a set of observed covariates. Suppose you have a situation where the observational dataset is purely observational, so there is no argument that it's an actual experiment. You match in an appropriate manner, you run your model, and you find out that the unobserved confounder would have to have a really strong impact in order to overcome the results. What do we do with observational data with no plausible experimental variation?

Gary King:

Just to put your great question in context, like I said at the beginning, everything here is conditional upon the chosen set of covariates. If those are wrong, then we are spending our time controlling for things that don't really matter, while the things that do matter we haven't controlled for. If this is the case, then we have a real problem. The best thing to do is to go get better data. If there are better covariates, or better confounders, then we should go measure those. If we know that there is an important confounder, but we can't measure it, then we should do a sensitivity analysis to see what the effect of that might have been. I think this is a great procedure. We should first focus on the data that we have. If we have data, and we have measured confounders that we know are confounders, ones that are related to the treatment and affect the outcome, and if we have not matched them to deal with the problem that we have actually observed, then we have no business trying to deal with the problem that we have imagined. Unless we can do a randomized experiment, we can't really be sure we have all of the confounders.

I notice that in most fields there are three or four variables that everybody agrees are big and important confounders, and then there are like 50 others that we have no idea about. In political science, party ID is an important one, right? It has a coefficient of .8, and it doesn't matter what the dependent variable is. In medical fields, prior health status, age, education, how healthy you are: excellent, good, fair, poor – those are the usual measures. Those things have a coefficient of .8. Then they've also got hair color and all kinds of dummy variables. In most fields it's actually sort of like that. Of course, that's based upon the knowledge of prior experiments, and all of the prior experiments may have ignored a big covariate, such as some genetic structure that has never been measured or something like that.

Question:

In the field of machine learning, there is a method of unsupervised learning which could help us to constrain or measure the distance between different observations in our data. I would like to know what you think of this kind of method and if you think it could help us do matching or not. Do you think this method could be a disaster like PSM?

Gary King:

I think unsupervised methods are really useful. I actually wrote an article on these with Justin Grimmer, a former student of mine, called "Computer-assisted Clustering and Conceptualization." That's really what it is, right? They help us come up with ideas, which would sound shocking because we generally think that's what humans do and that computer do the other stuff. What the clusters are is actually fundamentally important. In the talk I gave on "Reverse Engineering Chinese Government Information Controls", the clusters were the categories "criticism" and "collective action", which we just didn't think to separate. We came up with a theory eventually, that is we made it up, but we got it from the data. We thought about it and went back to prior evidence, and then we came up with the idea. So, interacting with the data is the best way to come up with these ideas. If it helps you figure out the metric of what observation is near a non-observation, it could be really useful.

Question:

Could it be possible that if you are throwing away all of the unnecessary data, you could end up not having a sufficient number of observations for your analysis? Your advice is to go out and generate more data, but sometimes it is very difficult to do so, especially when dealing with Chinese studies.

Gary King:

Well the ultimate answer is to always get more data, but actually the part of the talk that I didn't cover is about dealing with imbalance and the number of observations at the same time. So if you look up "The Matching Frontier", this is when we figured out how to optimize both. Of course, it still may be the case that you don't have enough observations. "The Matching Frontier" method will optimize it as best as possible, given the observations that we have, but it would still be best if you just got more data.

For more information, articles, & software, visit: GaryKing.org