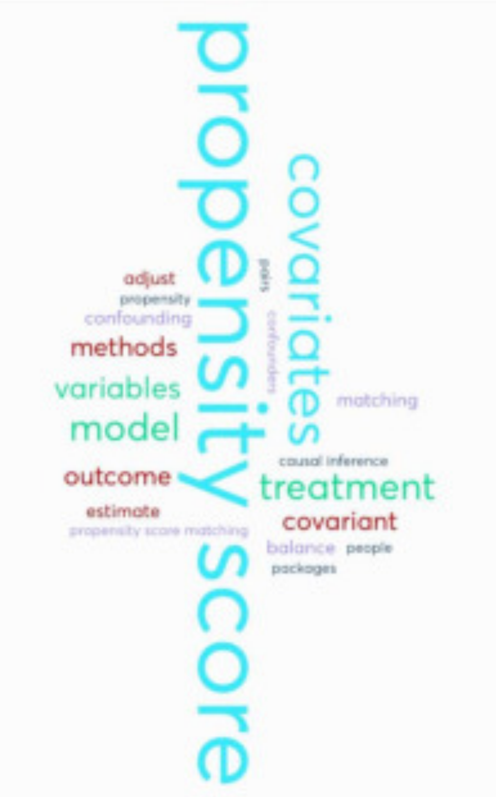
The Podcast *Quantitude*

with Greg Hancock & Patrick Curran

Season 3, Episode 27:

Propensity Scores -- I Meant to do That!

Published Tuesday April 19, 2022 • 57:15



SUMMARY KEYWORDS
propensity score, covariates, treatment, model, variables, methods, outcome, covariant, adjust, estimate, confounding, balance, matching, propensity score matching, confounders, , causal inference, pairs, propensity, packages

Patrick 00:00
Hi, my name is Patrick Curran and along with my matched covariant balanced friend Greg Hancock, we make up quantity food. We're a podcast dedicated to all things quantitative ranging from the irrelevant to the completely irrelevant. In this week's episode, Greg and I get a Hangout with Noah griefer from the Institute for quantitative Social Sciences at Harvard University to delve into the fascinating world of propensity scores, what they are, how we obtain them, and how these can be thoughtfully used to strengthen our causal inferences. Along the way. We also discuss Popular Woodworking. I meant to do that fixing things on the back end, hiding your own Easter eggs. At now we're warnings, easy undergrad majors, Myers Briggs career predictions, picking an ideal advisor, SAS holes, D myelination, hearing asterisks, being in COVID Missed the 11th commandment, what keeps you up at night? Answering the question you want to and it's all a bunch of BS. We hope you enjoyed this week's episode.

Greg 01:05
I had something to thank you for really?

Patrick 01:07
Did Goldie tell you to do this, because usually you get your sticky note in the morning of emotions to express and gratitude to convey why did Goldie tell you to thank me?

Greg 01:21
God, you just kept talking talking? I don't remember what it was now. That is so Okay, no, no, no, I remember I remember I was teaching about second order growth models where you're looking at growth in latent variables rather than growth and measured outcomes. But where you have different measures at different time points, right, where they're sort of shifting, so that people age out or develop

out of certain measures over time, but develop into or age into other measures, and was talking about these types of designs and how you sort of are clamping together the measures at each time point through constraints so you can get this larger trajectory. And 13 years ago, you sent me a picture that is completely on point to this. Do you know what picture I'm thinking of?

Patrick 02:05
Can you narrow it down? Was it appropriate or inappropriate?

Greg 02:09
If you cropped it, it was appropriate.

Patrick 02:13
The usual I do not I do not know what you're talking

Greg 02:17
about has to do with some clamps that you were using for woodworking.

Patrick 02:21
Oh, one of my first publications. It really was. I think that that was my second peer reviewed publication. It was in Popular Woodworking magazine. I was doing a project in the garage and I was working on a screen door and I needed to glue the base of it. The door was 36 inches wide, but I only had a 24 inch clamp but I had a pair of 24 inch clamps. I rigged them together so that it covered the full 36 inches. I was sitting there looking at it and I thought I kind of like that. So I took a picture of it. And they have a thing in woodworking magazine, which I subscribed to it was Tips of the Trade I printed off a picture at Walgreens and sent it in and they published it and it was titled to shorts make a long Wow. Now the funny thing about it is a naive reader might imply that I had made that door, especially since it was in popular woodworker. I had actually bought it at Home Depot. I hung it but made an error in hanging it and I didn't want to fix the error. So I tried to shave off the bottom of the door so that I didn't have to rehang and I split the wood while shaving the bottom, no reader looked at it and said, Well, what a beautiful work of art. Oh, there's nothing I can do about right?

Greg 03:49
Well, I admired your back end Jerry rigging of something. And it really was the perfect image for the type of design where I was talking about because in those models, you don't have all the same variables at different time points, but you often have the same measures that adjacent time points. And that's like you are clamping together time one and time two through a constraint. And then using a different variable for time two and time three. And that picture is exactly what I needed. So I wanted to thank you for that. I haven't thought about that in a while. One of the things that I like most about how it was framed is that it's as if you meant to do this all along. It's not absolutely true to the rest of the world. It's not covering your ass right. It's what clever for thought he exhibited in this. This is a

Patrick 04:32
general rule more broadly, if you're going to go back and fix something up to try to make things how they should have been in the first place. And you bolt together things on the back end. You insist from the outset that that's exactly how you would plan it.

Greg 04:49
Oh absolutely. A bunch of years ago I don't know might have been 1718 years ago. I was writing a chapter for some sage handbook. I don't remember it was on Layton means models. We actually had an episode on Layton means models earlier this season we did

Patrick 05:01
indeed, I'm gonna pretend like I remember that because I know I still keep texting your ideas for episodes and all you text back is s to e 18. And I'm like, Alright, forgot about that. Yeah, go ahead,

Greg 05:15
you're on Easter eggs. So I had written up this example of Layton means models. But it was for a design that was non experimental that I had these two groups that I was just comparing on some latent dimension. And the person who was editing the chapter, who was also the editor of the book had written back and said, You know what you should do because this design is not experimental, you should find a way to work propensity scores into it to really make the inference richer. And of course, my response is, what a great idea. Yeah. But it was not unlike you having a conversation with Niels Waller about fungibility, where you're going to Yeah.

Patrick 05:53
propensity scores, of course, no duck,

Greg 05:57
I was thinking the same thing. The more I got into learning about propensity scores, which is not to say that I understand propensity scores, but the more I got in, the more it looked like this thing that you were Jerry Regan on the back end to try to make it look like you had this beautiful experimental design, which of course you didn't. And then in the end, it almost feels like I'm meant to do that, right. And I tried to make the paper look as though somehow that was intentional from the start.

Patrick 06:24
So that is not unlike where we find ourselves today. In principle, we should have selected this topic on propensity scores and months ago, arranged for a content expert to come and join us. I forgot and did not months ago, contact the content expert to come and join us two simple words in the English language.

Insert 06:47
I forgot.

Patrick 06:49
I might have emailed him yesterday. In that spirit, I'm going to say I meant that the whole time. That's exactly how I had it planned. Because our guest today I have found over the years performs much better when we throw things at him unexpected and unprepared. And so from the outset, I decided to pull them in at the last minute so that we could get real time reactions from him. So we entirely meant to give him an 18 hour window of warning. I am so pleased to introduce Noah griefer, who is going to join us and talk to us about propensity scores, no and give us a little bit of your origin stories. Where did you come from? And how did you end up where you are today?

Noah 07:34
Sure. I'm from Los Angeles, California, my humble beginnings and in small town, rural California. And I went to UC San Diego for undergrad I started off as a math major because I was like a mathy kind of guy. But then I just kind of wanted to have a chill time in college. And so I switched to a psychology degree, which is a bit easier, and studied philosophy as well and had a nice time in college and was ready to go to grad school for cognitive neuroscience. I just felt like I was destined to be a professor. I took one of those like career should you do type of tests and it was like, be a professor. Okay,

Patrick 08:09
I will Oh, wow, I was supposed to be a park ranger.

Greg 08:12
After all the sciences and math and stuff I taken in high school, I was told I should be a photographer.

Patrick 08:19
Okay, so Myers Briggs said you were gonna be a professor. And then

Noah 08:23
my senior year, I just kind of had a total reversal and said, I actually don't care about cognitive neuroscience anymore. And I was just into statistics and had taken statistics classes in undergrad and even ta them in psychology. And I was like, I want to study statistics. You can't really go into a statistics PhD program with a psych degree. So I talked to people in my department, and one of those people was the late great Mark Applebaum. And he recommended that I apply to quantitative psychology programs, and that in particular that I applied to UNC so I applied to a bunch of grad schools and UNC was the only one I got into and I was on the waitlist and Patrick took me off the waitlist, which is very nice of him.

Greg 09:06
I just found your rejection letter for Maryland.

Noah 09:12
I applied to work with someone who wasn't taking students. And that was actually a theme in some of my grad school applications.

Greg 09:21
At some point, it's you know, I think this is an important realization.

Noah 09:26
Miraculously, Patrick took me off the waitlist and I decided to go to UNC I didn't know what I wanted to study. And I remember in my application, I was extremely vague and mentioned something silly, like small sample sizes, not that I knew anything about. I eventually fell into causal inference and propensity scores and decided that this was an area of interest for me. So I took a bunch of classes in that and wrote my master's on using propensity scores with a moderated nonlinear factor analysis eventually, did my dissertation And I applied to a postdoc at Johns Hopkins with Liz Stewart after a year in my postdoc, I was ready to go and be a professor. But I found that I was like really struggling in this postdoc to do the things that a professor had to do, like manage my own time and write papers. And I just wanted to basically do my passion, which was writing our packages, which I know is like a funny thing to call your passion. But it really wasn't mine. Like that's what I was doing. In my free time. I was recruited by Harvard to be a statistician for them. And now I work at Harvard as a statistical consultant and programmer.

Greg 10:31
So I'm confused on the professors are supposed to manage their time and write papers is that that's kind of news

Noah 10:38
to me too. And that's my I guess, this aptitude tests that I should be that despite like having this problem, entire life of procrastinating

Patrick 10:48
one thing we all love about you is you're very modest, as well, as you had some major successes in grad school as you were an exemplary teacher. You didn't interns at SAS and Rand? No, it was just a joy to have. You know, one of the reasons why I pulled you off of that waitlist is one is I was very interested in your application to begin with. But the second is, I was very close with Mark Applebaum. He was a senior mentor to me for 25 years. And he sent me a one line email that said, admit no. And I was like, Okay. Mark had his fingerprints on a lot of things. Thank you so much for joining us today, especially given how much effort we put into the scheduling and organizing of this that dates back. I don't know what Greg about 18 and a half hours ago. But we want to pick your brain about propensity scores, and I want to hear about the packages as well as it is a passion. And when your first day at Carolina, I realized that was a passion, if for nothing else than your incredulity that I didn't use our you were questioning your decision. At the very outset your passion was propensity scores. And our and I don't study propensity scores or use are. But other than that I was an ideal advisor for you. He's a

Greg 12:13
diehard Sasol? Well, let's set the stage for this. Now, what can you help situate us with regard to the benefits of good design and when you can't necessarily have good design, and then the role that propensity scores might play in helping us out of that hole?

Noah 12:29
When you're interested in estimating the causal effect of something a treatment or policy and exposure, um, just means that we're treatment on an outcome. And this is actually not the only way that psychologists think about causality psychologists are often interested in figuring out what are the causes of variability in some outcome or some behavior. And that is not the context of propensity scores are used, they're used in the context of estimating the causal effect of a known treatment or exposure on an outcome. So it's not about discovery, it's about estimation of an effect, we usually think about a binary treatment, that's the easiest case to think about, where we have two groups a treatment and control, they don't have to be treatment in control. They don't have to be two groups. But that's just the simplest context. And so we'll talk about propensity scores with respect to that context. Ideally, if you're interested in the effect of a binary treatment or an outcome, you would randomly assign your participants into the treatment and control group and then compare the differences in their observed outcomes. Hopefully, there's no drop out. And hopefully everything is measured well with no error. Although we know that's never the case. The benefit of randomization is that on average, the two groups are the same on all of the features, the characteristics, that would cause variability in the outcome. If you randomly assign it's not the case that you would have one group that's a lot older than the other, or one group has a lot more men than the other randomization in large samples guarantees that the groups resemble each other. And when they resemble each other, not just on the variables that you've collected, but on all possible variables on all features of the individual randomization balances those characteristics.

Greg 14:10
The way I think about that is in terms of path models, and Patrick and I both have pictures on the brain all the time. The way I think about that is that there might be a lot of other variables that influence the outcome. But by random assignment to two groups, what you're doing essentially is disconnecting that treatment variable from all the other variables, meaning that there's no other avenue for the treatment variable to have a connection with your outcome. Is that a reasonable way to think about that?

Noah 14:38
Yes, that's exactly right. There are graph theoretical approaches to causal inference and that criterion that there is no connection between the treatment and the outcome except through the causal effect of the treatment is exactly the criterion that is used to establish causality. Often, of course, we can't randomly assign because it's impossible to or unethical to think In this example, if you're interested in the causal effect of smoking, which we know is like a potentially harmful stimulus, you can't randomly assign people to smoke or not smoke. Now there are randomized trials you could do with smoking, like randomly assigning people to quit smoking, but it may just be impossible to randomly assign, what can you do. The problem with not randomly assigning is that the two groups are going to look different from each other on these important qualities that are going to affect the outcome. For example, it might be that those who take the treatment are older or are more susceptible to disease, or are more desperate for a treatment, if you simply compare the outcomes of the treatment and control group, when you don't randomly assign any difference that you observe might not just be due to the treatment, but might be due to these other factors as well, which means that simple comparison is not a valid estimate of the causal effect. It's biased,

Patrick 15:51

going back to Greg's comment about the path model, because one feature of the podcast is we try to use as much visualization as possible in an audio based format. Back in the day, when I learned this, if for whatever reason, you couldn't achieve randomization, as long as you have those covariates in the model, those would and I'm going to use air quotes, adjust the outcomes for these differences. And then when you looked at the unique effect of that binary grouping variable that was above and beyond the covariate effects, and there was almost a magical quality to it, don't worry your pretty little head about randomization, we can fix that on the backend. What are the problems with that so many

Noah 16:37

problems. The biggest problem and this is people's biggest critique with propensity scores and causal inference methods that rely on covariate adjustment in general, is that you really have to collect every single variable that causes both the treatment and the outcome in order to fully adjust for confounding. confounding is just what we call this situation where there are variables that cause the treatment and the outcome. And the whole point of randomization is that because there's no variables that cause the treatment, there's no confounding. If you can collect all these variables and adjust for them in a specific way, then you can claim that your effect estimate is causal, you can adjust for these variables arrive at an estimate that corresponds to the unique effect of that variable. Of course, the biggest critique is that you can never measure all of these variables, there are just so many causes of things in a complicated complex world that you simply cannot measure all of the causes of treatment in the outcome. Does that mean we're completely stuck? And the causal inference is over if you can't randomly assign? Some people say yes, but the whole field of causal inference says no, there are things you can do. You can get things approximately right, you can rely on substantive knowledge to collect and adjust for the best and most important confounders you can assess sensitivity. For example, how strong would a confounder that you didn't measure have to be before it would change your inference? It is true that this is a really strong criterion. But we can still do the best we can we can still learn from our data. And there are ways to assess how sensitive our inferences are to a violation of the assumption that we have collected all confounders

Patrick 18:10

so you said something really interesting there of covariates, not only associated with the outcome, but with the treatment condition.

Noah 18:18

It's not just about being associated with it's about causing that is confounding is a causal criterion, not a statistical one, the covariates have to cause the outcome in the treatment. And the reason that's important is because there are variables that are associated with the treatment and the outcome that are not confounders, and in fact, adjusting for them would be the wrong thing to do. The idea that a variable might cause selection into treatment. The way that manifests is that the distribution of that variable differs between the treatment groups, the way we describe that, as we see that there's imbalance in the distribution of that covariance between the treatment groups. And that imbalance is the problem. And that's the whole point of randomization is it achieves balance on all covariates. What we would love to do is find a way to achieve balance on all those covariates. And propensity scores are one way of doing that. I want to stress that propensity scores are one implementation of a class of methods. And that class of methods is methods that rely on covariant adjustment, but there are other statistical methods that you can use to adjust for confounding. And we're not going to talk about those today. But just an example of some of those are instrumental variables and difference in differences, which are two statistical methods that rely on different assumptions, not the assumption that you've measured all confounders even the methods that do use a propensity scores, which we'll talk about later matching and weighting. You can do them without propensity scores. So propensity scores truly are a narrowing down of one implementation of one type of method of one class of methods that relies on one assumption. I think a lot of people think of propensity scores as the iconic method to adjust for confounding but it's just one of many

Greg 19:58
when you say that you have a whole bunch of confounding variables or control variables. Of course, what comes to my aging is a demyelinating is that was the word

Patrick 20:06
Demyelination would best describe you.

Insert 20:09
But the best thing about getting old is you're not responsible for remembering things anymore. That's a lot of fun. You look around the dining room table, and you say, Who are you people, and where's my horse?

Greg 20:21
My demyelinating mind is an Cova. The idea of using the general linear model to try to incorporate these variables, and I'm going to assume that we've identified them reasonably on theoretical bases and all of that. But then when you start seeing all the variables that starts really nervous, that there's so much information I don't even know how to balance when I've got an entire boatload of these kinds of variables. My thinking is that that's probably where propensity scores come in. With all of these control variables are confounders I like

Noah 20:52
to think of propensity scores actually, as really an alternative to N Cova. That attempts to address the problems with an Cova. The problems with an Cova Are you are a assuming a linear relationship between the covariates and the outcome given the treatment, and you are assuming that the relationship between the covariates and the outcome does not depend on treatment, which is to say that there's no interactions between the treatment and the covariates. Now, you might say, Okay, well, I'll just put an interaction between the covariates and the treatment. And I'll just use a maybe a flexible generalized additive model, or splines, and that'll take care of all of it, it might, but you have no way of knowing whether it did or not, you can examine the residuals. So by that time, you've already estimated the treatment effect. And the residuals are a broad diagnostic that doesn't exactly tell you whether there are components of the variables that are still relating to the outcome given the treatment and the covariates, you included in your model regression may adjust for confounding, but you don't know if it did, how we would know if it did is if we had our two groups, and we looked at the distribution of the covariates within those groups, and the distributions of those covariates were identical. under that

circumstance, we know that we adjusted for all of the variables correctly. propensity scores and methods related to propensity scores are really aimed at this problem of the inability not only to correctly specify the outcome model, but to justify the complete adjustment of covariates by whatever method you use, whether that be the regression model or some other method. If you still haven't defined propensity scores, I'm intentionally resisting that because I really do want to paint this picture that propensity scores are one way to do what I'm about to describe, rather than adjusting for covariates. By modeling the outcome, you adjust for covariance, by manipulating the sample of the treatment and covariates. Without reference to the outcome. In the causal inference world, we call that a design based procedure as opposed to an analysis based procedure analysis is just the word for a model based on the outcome. And design is this kind of analogy to experimental design, where you are adjusting the sample or the selection features to eliminate confounding. An example of that would be designing and implementing a randomized trial. What propensity score methods do is they attempt to mimic a randomized trial in your observational study, they do that by attempting to balance the covariance between the treatment and control group. And what's cool about that is that doesn't involve the outcome at all. Which means that once you've done this process of adjustment, and you've manipulated your sample, you can just do a T test, for example, on your outcomes, there's no modeling assumption about the relationship between the outcome and the covariates that is otherwise an integral part of an Cova and other regression adjustment based approaches. Okay, now, let's talk about the propensity score, I think because I've like avoided talking about it,

Insert 23:42
get on with it, yeah, get a

Noah 23:48
different score is a variable that summarizes the covariates to be adjusted for the ideal way to balance the covariates would be to exactly match on them, which is to say you find treatment and control units that have identical covariant values, and you form pairs of those, basically, you implement a twin study, you throw out everyone else, because everyone else is irrelevant to the causal contrast. And what you're left with is two identical groups. And in those two identical groups, it's almost as if you have randomization. The problem when you have many covariates is that it's impossible to find perfect twins in your two groups, especially as the number of covariates increases, and it often does in order to satisfy this assumption that you've measured all your confounders you may end up having like a ton of variables like 50, and there's no way you can find perfect twins on 50 variables. This incredible discovery by Rosenbaum and Rubin in 1983, is that if it's sufficient to adjust for a specific set of variables to remove confounding, it's also sufficient to adjust for a propensity score. What is the propensity score the propensity score is a function In have all of those covariant values, it's kind of a one dimensional summary of all of the covariates. The way it is defined is it's the conditional probability of receiving treatment given the covariates. An example of how we might compute a propensity score would be to run a logistic regression of the treatment on the covariates. And then use the predicted probabilities from that model, the idea is that each individual receives either treatment or control. But if they hadn't, what was their probability of receiving treatment, so there might be people who did receive the treatment, but people like them typically don't receive the treatment, in which case of their predicted probability, their propensity score would be low, whereas there may be people who did receive the treatment, and they're very much like others who received the treatment and their propensity scores

would be high. The idea is that if you can adjust for the propensity score, you are adjusting for confounding. And this is a huge discovery because it means we don't need to exactly match on these 50 confounders, we can just match on the propensity score. And that's sufficient for achieving covariant balance between our treatment control group and therefore adjusting for confounding,

Patrick 26:09
this is a fascinating change your perspective, because going back to the N Cova, you're talking about taking an optimal linear combination of your set of covariates, adjusting for those in the dependent variable, and then looking at what is leftover in comparing the treatment and the control. And if I understand correctly, you're shifting that back, where you're taking an optimal linear combination or nonlinear as the case may be in predicting the treatment group membership. So you're building a model for selection, and then using that in a subsequent analysis in examining the group differences on the outcome of interest? Is that a right way of thinking about that? Yes, that's exactly right.

Greg 26:55
A lot of the things that you're saying I either hear asterisks in your voice, we know are footnotes that that we should drill down into, or things that make me go

Insert 27:09
things that make you go.

Greg 27:12
And one of the things that has made me go so far was the idea of the model that actually gets you to the propensity score, I could imagine that there are a ton of different possible models. You mentioned logistic regression. But there could be other things too, right?

Noah 27:27
Oh, yeah. Logistic Regression is definitely I think what people think of is like a canonical example of a model that's used to predict the probability of a binary outcome, in this case, the treatment, you can use anything, you can use a machine learning method, you can use other generalized linear models. And of course, if you'd have more than two treatments, you're using a multinomial model. One of the cool things about propensity scores is that they are amenable to machine learning, which means you can use a complicated model to predict the treatment selection and generate the predictive values from that. What's cool about the propensity score model is there's no value in interpreting the propensity score model, its sole purpose is to create these propensity scores, which means you can have any model it can be way overfit, it can be over parameterized, it can be completely uninterpretable. But as long as you're adjusting for the right variables, and as long as that propensity score yields covariant balance, after adjusting for it, you're golden,

Greg 28:21
technically doesn't have to be the right model for all of this to work out.

Noah 28:25
Yeah, technically, you do need the right propensity score model. But that's what's so cool about the ability to use machine learning, there's a much better chance that you're going to approach the right propensity score model, if you can use whatever model you want to use that can flexibly adapt to the data and possibly incorporate all interactions and nonlinear terms. And there's basically just like, no hope of you doing that with a linear outcome model. And even there's essentially no hope of you doing that with a logistic regression treatment model. And so that's like, why machine learning is so important for propensity scores?

Greg 28:56
Is there a danger here of overfitting? If you put all 130 potential confounders into your propensity sausage maker? Can that cause some problems?

Noah 29:07
That's a great question. The answer is kind of No, because the way the propensity score model is evaluated is on how well the resulting propensity score achieves balance when adjusting for it. It's not about parsimony at all it may be that a parsimonious model is the one that achieves the best balance. But it's often the case that it's not. And really, you need a rich model that adjusts for the covariates in a complete way. And that's the type of propensity score that yields covariant balance after you adjust for it. That is how propensity scores are evaluated. That's their sole criterion. And that's very much unlike any other model that we use in statistics really, which is maybe based on its predictive ability or its interpretability or the distribution of residuals or anything like that. And that's just not the case with propensity scores. They're evaluated solely on their ability to achieve balance.

Patrick 29:54
It makes me think back to Harold Hotelling in the early 1930s. With principal components analysis, which is he had a larger amount of information, he wanted to distill it down to a smaller amount of information, he did not care about rescaling Eigen vectors or rotating or interpreting, it was just a bit of a Thor's hammer to take a large dimensional space and reduce it to a small dimensional space. That seems like what one of the goals is here

Noah 30:26
is definitely related. The thing about propensity scores is it really is about this theorem, which demonstrates that adjusting for the propensity score in particular is sufficient to remove confounding. It's not just adjusting for any summary of the covariates. So it's not just about summarizing the covariates into a single dimension, it's specifically about the propensity score, the predicted probability of treatment that has this property that when you adjust for it, you remove confounding.

Patrick 30:51
So let's go there, we have a set of covariates, we have a binary outcome for simplicity, we'll have two groups in the treatment for simplicity, we'll do logistic, you have 500 people and every person gets a propensity score representing some probability of membership in the treatment group, what do you do, then

Noah 31:10
there are a few ways to adjust for the propensity score, there's like four canonical ways. One way, which I don't recommend is you actually treat the propensity score as a confounder. And you adjust for it using an Cova. So that's one reason that I don't like that is because that's just going back to the old using an Cova, which we're not about, we're going to leave that behind.

Greg 31:29
Are you an ANCOVist? You sound like an ANCOVist.

Noah 31:33
I will unabashedly say that I am not into ANCOVA.

Patrick 31:36
Even if estimated in SAS, I mean a SAS ANCOVA really is state of the art

Greg 31:41
done and done.

Noah 31:44
Let's talk about the other three methods of adjusting for the propensity score, the canonical one is matching. And so this method of propensity score matching is by far the most popular application of propensity scores. The idea is that each treated unit as a propensity score, and so does each control unit for each treated unit, you find a control unit, with a closest propensity score, you create pairs of units that have near identical propensity scores. And any unit that's not paired is just discarded from your sample. A lot of people have problems with that, by the way, because you're literally throwing away data. And the response is yes, but that data is not relevant for the causal comparison. Once you've created these pairs, these units may be identical on the propensity score are very close on their propensity score. But that doesn't actually mean they're close on specific covariate values, which is to say that you and I could have the same propensity score, and be completely different. Otherwise,

Greg 32:38
they're just different cocktails of covariates, that lead us to have a certain propensity

Noah 32:42
Exactly. Despite that, if you've created these pairs based on the propensity score, the distribution of covariates in the sample overall will be similar. And that's because of this propensity score theorem. propensity score matching allows you to achieve covariant balance in this way, people don't like this because you're throwing away data there. Of course, more sophisticated methods of matching that don't involve throwing away as much data like you can match two controls for each treatment, or three or a variable number.

Greg 33:07
The propensity scores aren't truth, I would imagine we could think about them as a measured variable and that there is even some underlying latent propensity, how much Fuzz is there in propensity scores? And how does that play into what you're describing?

Noah 33:20
The theorems are propensity scores all assume that you have the true propensity score for each individual, which of course we don't have in almost all situations. And so they need to be estimated. And so the theorems basically only apply approximately with respect to estimated propensity scores. If you can estimate them, well, you can achieve balance. Now this leads to an interesting idea in propensity scores, which is called the propensity score tautology, which is that a good propensity score achieves balance. But the way to assess whether a propensity score is good is whether it has achieved balance. This goes back to this idea that the propensity score is only assessed based on its ability to achieve balance. We don't need to rely on the theoretical properties of the propensity score to assess whether it's working right or to implement it, we just need to assess whether it achieve

Greg 34:06
balance. So then if there is this fuzz, though, you're doing some kind of matching and that fuzz has to play into matching, how do you decide how close is close? And to what extent if any, do we factor in this fuzz,

Noah 34:21
what most matching algorithms do is they don't consider how close is close, they just find the closest if I have a propensity score of point three, a matching algorithm will find the control unit with a propensity score closest to point three and assign that to me. And that's one way of doing propensity score matching. There are of course, more sophisticated algorithms that take into account the entire propensity score distribution and try to minimize some overall difference between propensity scores within matched pairs. But the basic kinds of nearest neighbor matching strategies to just find the nearest now that can actually lead to pairs that are very far on the propensity score. So if I have an extreme propensity score, and there's no By controlling it with a closed propensity score, I will be paired with someone who is very far from me. That is that funds that you're talking about where we're not really twins even on the propensity score, let alone on the covariates. And is that a problem? And the answer is yes, that is going to be a problem, it's possible to implement a restriction on the matching such that only pairs that are within some distance of each other are allowed to be matched. And anyone who does not have a potential twin that's close to them is simply not matched. And again, that's throwing away data. But who are you throwing away? Well, it's people who are irrelevant to the causal comparison, because they're unlike anyone else in the other group. And that is called using a caliper, which is just a bound on the allowable difference in propensity scores between units and a pair. And it turns out that using a caliper tends to dramatically improve the performance of propensity score matching, because you're left with just these closed pairs, and you don't have this problem with matching someone with someone else who is very unlikely among propensity score.

Greg 35:59
And so the caliper Gods determine how wide the caliper is,

Noah 36:03
you can guess how we choose the caliber, it's whichever caliber yields the best balance. So there are no caliber gods, it's you.

Insert 36:11
I'm a God, Your God, I'm a guy, not the guy.

Noah 36:15
I don't think you can try many different matching algorithms and many different calibers before we decide on the one that you want to proceed with the fact estimation, there are some guidelines of the caliber to use. But it's often said that 1/5 of the standard deviation of the logit of the propensity score is the ideal caliber with which of course is so specific, and it's going to vary depending on the specific context.

Greg 36:37
I think Moses had that on a tablet. Exactly. It's the 11th commandment,

Insert 36:42
the Lord Jehovah has given unto you these 1510 10 commandments, thou shalt use a calendar.

Patrick 36:53
You build a selection model, you get propensity scores, you set a caliper, you compare algorithms, you get pairs, you discard those who are not paired than what you

Noah 37:04
need to assess balance. This is the critical step, which is how well did you imagine do at balancing the covariates. And the classic way to do this is to compare the covariant distributions for each variable, you can use a summary statistic that assesses distributional balance beyond the means, like the Kolmogorov Smirnov statistic, which is a way of assessing whether two distributions are the same or not. And there are other multivariate methods. If you have achieved covariant balance to a sufficient degree, the next step is to estimate the treatment effect. The way we can do that is simply by comparing the outcomes in the two groups after matching, which could be a t test, we've already done the work of adjusting for the covariates in this design phase, which is the propensity score matching. And our outcome model can be very simple. And that's really the strength of this because you are not potentially engaging in this specification search to find the right outcome model that correctly adjusted for covariates. You don't need a model at all. You can also include covariates, in your outcome model, after matching that doesn't do the same thing that an Cova does of adjusting for the covariance, because they've already been adjusted for it just increases the precision of the resulting effect estimate because you're removing variance due to those covariates. And so all you're left with is a variance to the treatment. And that is just a way to increase the precision of your effect estimated but it's not necessary.

Patrick 38:22
What I see in my mind's eye is a set of covariates predicting a binary outcome in step one, and then you do all the work that you do. And in step two, we have the binary predictor of your dependent variable of key interest. Is treatment condition, a mediator?

Noah 38:40
No, that's absolutely right. The treatment is a mediator between the confounders and the outcome, it's just not really that valuable to think of it that way. Because you're not doing mediation analysis, you're interested in a single coefficient in this path model, which is the relationship between the treatment and the outcome. And the confounders are just these kind of nuisances that you need to adjust for. So it is true that the causal structure of the system is that of a mediation scenario. But treating this as mediation is not useful, because usually in mediation, we have a focal predictor, and then the mediator is also have substantive interest. In this case, only the treatment is of substantive interest. And the confounders we're just adjusting for they're not substantively important focal predictors in a mediation. I'm a congenital

Patrick 39:22
counter. We're on number two, but you said there were four, I'm starting to feel really uncomfortable.

Noah 39:31
We'll talk about the other two is after having talked about matching the other two, I think are much simpler. So that's why I wanted to go into a little bit of detail matching because the procedure is the same. The other two are propensity score weighting and propensity score subclassification. propensity score weighting is where you use the propensity score to compute sampling weights, and then you perform an analysis in the propensity score weighted sample and these sampling weights serve the same purpose as propensity score matching, after applying the weights the covariant distributions ideally are the same. The weighted difference in means is unbiased for the treatment effect, I love propensity score weighting because there are a lot of cool ways to estimate the weights. The formula that you use to estimate the weights from the propensity scores changes the interpretation of the treatment effect in cool ways by changing the reference or target population, the treatment effect estimate generalizes to in addition to using propensity scores to compute the weights, there are a ton of other really cool exciting methods that you can use to compute the weights, some of which bypass a propensity score and estimate the weights directly. And so those methods are increasing in popularity these days. The final method is propensity score subclassification. And this really is an older and easy to use method, but it tends to perform a little worse. This is where you create subclasses or strata based on the propensity score. So for example, everybody with a propensity score between 0.1 is placed into a stratum and everybody with a previous score between point one and point three is placed into a stratum, etc. And we have a bunch of strata, and each stratum contains both treated and control units with similar values of the propensity score. The idea is that within strata, there is covariant balance. So you can estimate the treatment effect within each stratum and then average across strata to arrive at a single treatment effect estimate.

Greg 41:19
That's like a randomized block design, exactly blocking on segments of your propensity score distribution.

Noah 41:25
That's exactly right. So those are the four methods natural and waiting are definitely like where the juice is like where the recent developments are, these are the most popular methods. propensity score matching is widely used in medicine and education and political science. Whereas propensity score weighting tends to be used a lot more in epidemiology and public health as a cultural phenomenon. Why it has gotten that way. There aren't specific reasons why one method should be preferred over the other.

Greg 41:50
So coming back to the matching condition, you said that the analysis on the back end could be as simple as a t test, you actually mean an independent samples t test, even though there has been matching that has gone on to ensure balance, which all of a sudden makes it feel more like it's paired or dependent is that

Noah 42:06
right? There has been debate about whether it should be an independent samples or a paired samples t test. The early research and propensity scores basically said you can just do an independent samples t tests because although you're creating these pairs, the pair's may not be close to each other on any variable, they're just close to each other on their on t score. And the point of creating these pairs is not really to create pairs is to create balanced samples. And so the goal is simulating a randomized experiment with no pairing in which case an independent samples t test should do the job. Subsequent research has indicated that accounting for the pairing is important. And not only is it important, it is necessary for valid inference, I am of the opinion that you should always account for pairing when you can, using the paired t test is one way to do that. You could use a multilevel model where you have a random effect for pair membership. You can use a fixed effects for a pair membership, which is going to be the same thing. You can use a cluster robust standard air, you can use a cluster bootstrap a bunch of ways to do it. But as long as you're adjusting for pair membership, you're golden, basically,

Greg 43:09
are there then other analytical options besides this most basic one?

Noah 43:14
Yes, definitely. I mentioned that you can do a regression of the outcome on the treatment and the covariates after doing the matching like an Cova or a more sophisticated model. But that leads me into this other kind of class of estimators which are called doubly robust estimators.

Greg 43:31
I don't care what it is, I want it. Yeah, exactly.

Insert 43:35
Yeah. Delta is already on probation. Oh, then as of this moment there on double secret probation.

Noah 43:41
A doubly robust estimator is an estimator that incorporates both a model for the treatment, which is to say a propensity score model, and a model for the outcome. The reason that these methods are doubly robust is that if either the treatment model or the outcome model are correctly specified, your effect estimate is consistent, which means in large samples, it's unbiased. And so the idea is you get two chances to get it right. Doubly robust estimators are most commonly used in the context of propensity score weighting, where you estimate your propensity score, you compute the weights, and you perform an analysis that incorporates the weights and an outcome model. And an example of such a model would be a weighted least squares regression model of the outcome on the treatment with the covariates in that model, but it's weighted by the propensity score weights. And those weights also incorporate the covariates you're doubly adjusting your accounting for the covariates in two ways. And in that sense, the method is doubly robust. And so that would be considered the way to go. Some people think that the people who developed these methods will definitely say there's no reason why you shouldn't use them. Exactly. One reason not to favor doubly robust methods is that you basically are not guaranteed to get either model correct. And so what is doubly robustness really buying you it's true that you get to choose answers, but there's a one infinity chance of getting either of them correct. And having two chances doesn't really make it that much better.

Patrick 45:08
But it's twice as much. Exactly. It's to infinity.

Greg 45:14
Geez, you still have the problem of multiple covariates. Right? So one of the initial things that we talked about is when you've got 100, covariates. And you're like, I'm pretty sure that this covers the spectrum. When you move to a model that's doubly robust. It seems like you still need some sort of principles, dare I say, to decide what covariates ought to be in there in this model, right?

Noah 45:32
Yeah, definitely. The principle of which covariates need to be adjusted for is a purely causal criterion, like more statistical question of like what variables need to go into this model to adjust for confounding without oversaturating. The model essentially, is kind of a different one. What's cool about these doubly robust estimators that really sets them apart is that you can use machine learning for both models, and yet use asymptotic normality in your inference, that allows you to use machine learning methods that perform regularization and feature selection as part of them to simplify the model. Again, you're not really interpreting these models. So it doesn't matter how you arrive at these estimates. But if you have 100 covariates, sometimes these machine learning models will just select out the variables that are less important to the outcome. You don't need to perform this active variable selection yourself or rely on dimension reduction methods. Because the machine learning method does that automatically. As part of it. The state of the art is a method called highly adaptive lasso, which is a form of lasso regression that operates on categorized versions of the covariance, it has this really interesting asymptotic

guarantee of convergence at a specific rate, it's possible to account for having a huge number of covariates in your outcome model by using machine learning

Patrick 46:47
as an expert in the field. What aspect of propensity scores keep you up at night?

Noah 46:53
I guess I'll maybe answer a different question before I get to that.

Patrick 46:58
Okay, everyone, this is a microcosm of five years of graduate study that I had with Noah. So please, answer a different question,

Greg 47:08
sir, what you want to answer?

Noah 47:10
I'll answer the question that I thought you were going to ask before you said the last word of your question I got really excited to answer. Okay, go for it. Which is what do I like about propensity score methods, these to keep me up at night, but for a different reason. You know, they keep me up at night, because I'm thinking, Oh, this is just so wonderful. What I like about them is that it's fun to use them. And it's because this process of finding the specification of the propensity score model, and the matching algorithm that gives you the best balance is fun, it's fun to think about, it's fun to implement, you get to try many things, you get to do a kind of exploratory data analysis in the context of a confirmatory analysis, like you're answering a confirmatory question, but you can use exploratory data analysis techniques by trying many different matching algorithms. And these algorithms are cool, it's cool to see how they work and how they all function differently. It's cool that you can not only use regression based techniques, but machine learning, which opens like a whole door to like education on machine learning. And so the only reason I know anything about machine learning is because of propensity score methods. Without it, it's just wouldn't have been part of my education. And if I was doing purely parametric type SEM statistics just might not have ever picked up machine learning, I just feel like that's been such a valuable skill. So it just opens the doors to all these cool methods into this singular goal of achieving covariant balance. That's why I love and enjoy propensity scores. And why I enjoy like being a scholar of this area is because it allows me to learn so much about other cool branches of statistics. So that's what keeps you up at night for

Greg 48:41
good reason,

Noah 48:42
for good reasons. Let's talk about keeps you up at night for bad reasons. One of the hardest things about propensity score matching is the inference is a little unclear. And by inference, what I mean is computing uncertainty around your effect estimate. I've mentioned that you can just use a cluster of a senator. And that's valid. But it's not exactly clear to me, if that is really accounting for the process of

finding the right match sample, there aren't asymptotic guarantees, it's relying a lot on intuitive arguments, there are proofs about why it's valid, but I still sometimes feel like it's a little fake, or that it would be really hard to justify rigorously, I get why econometricians don't always like some of these methods, because there's not a clear proof of why a specific standard error is valid. There's so many choices that you can make that are unverifiable, like which cluster robust standard error to use, or which outcome model to use after you've done the matching or if you have several matching specifications that all yield the same balance, but you'll different effect estimates. What do you do? Another thing is that people implement these methods really poorly, like people are not good at doing propensity score analysis. And I think that like every statistician feels that way about every single method that they studied. I'm sure you guys feel that nobody has ever run a structural equation model correctly or model correctly. I think that nobody's ever done it for funsies. One of my most popular tweets is that I really think that like nobody should be allowed to, except like the 12 experts in propensity score analysis, those are the only people who should be allowed to run the apostles. Right, exactly. And even then, if I'm one of those, there's still so many things, I don't know how to do correctly, that I kind of sometimes feel like, it's all both. There is truth. And there are good properties, the more I think about it, sometimes I do feel like there's too many unknowns. And we are definitely straying from the truly confirmatory asymptotic inference world.

Patrick 50:39
That's a really good pivot point into what do you see as areas of future work,

Noah 50:45
this is really for a statistician to figure out how you can incorporate the specification search and the matching algorithm and propensity score model in the uncertainty estimate of the treatment effect. If we were able to do that it would make justifying that process a lot easier and more valid for skeptics of this method, who often say that one of the problems with propensity score analysis is that there's so many ways to do it, there's so many ways to estimate a propensity score, there's so many ways to perform the matching. And those different choices can have big impacts on the resulting effect estimate. And so how do you have an honest uncertainty estimate. And usually, the way we do it now is we just pretend we didn't do all that stuff. And there's some reasons to think that that can be okay. And there's a whole philosophy, which is called matching is nonparametric pre processing, which basically assumes that the matching is not part of the inference, because the inference is only on the match sample as if that was the sample you started with. But of course, we know that's not true matching really is part of the analysis. I feel like if a statistician could really account for that, I think that would be really great. I think we really need better methods of assessing balance, because this whole thing relies on the propensity score tautology, which is like once you achieve good balance, your propensity score is good, and you can move forward. And the current methods for assessing balance are really poor. There are some cool new methods, but they just haven't really been explored very much. A method that I'm really excited about is called energy balancing, which has like a really cool sciency sounding name, which relies on this nonparametric measure of the association between the covariates and the treatment, and it's a multivariate measure. So it's a single number that describes the covariant balance of the entire sample after matching, I really think it's the statistical theory that would convince skeptics is really where the research needs to be done. One thing

Patrick 52:35
I've always admired about you is trying to take these very high level theories and procedures and get them out to people who can use them, you write packages, two big ones that I'm aware of our cobalt 10 match it. Tell us a little bit about those than anything else that you're working on.

Noah 52:55
I have written many our packages, and almost all of them are in service of propensity score analysis. In grad school, I started writing this package called cobalt, which was a way of assessing covariant balance both numerically and graphically. The reason I wrote it is because every our package to perform propensity score matching or weighting, and there are many had their own Balanced Assessment features that were not compatible with each other. And I was like, Wouldn't it be great if there was just one that at all cobalt became popular, which felt very cool to me as like a young grad student. And when I left grad school, I had no publications, but I have three very popular art packages under my belt. So I just took a different route, what it meant to be productive in grad school. So the packages that I wrote in grad school, the big ones were COBALT is a cool name sensor covariant balance tables. And I wrote this one called weighted which is a pun on an existing our package called match it which is like the our package for propensity score matching, which I didn't write after I finished grad school and started my postdoc with Liz Stewart, I asked if I could update match it, which had not been updated in a long time, I basically almost rebuild it from scratch. And now I definitely consider match it one of my packages, I am the maintainer of it, I definitely am not one of the original authors. And so I don't want to take credit for its inception. I have another package that implements an optimization based method of waiting, which is called off to wait at Harvard. My job is to write our packages for other people, which is just like the perfect job for me. It feels like this job basically. And some of those are packages are unrelated to things I know about like I've been asked to write our packages for phylogenetic analysis, which I don't understand at all and I'm just typing

Greg 54:33
doesn't stop us in academia.

Patrick 54:38
No, I don't know when the actual final episode minutes but we are at 96 minutes in recording time. There's so many more questions that I want to ask but we just don't have the time to explore it and we will have to give you another 18 hours notice and some other time, but I think it indicates how excited In this area of research really is yeah, I find it fascinating and so many different ways. We so appreciate you sharing your time very much importantly, we so appreciate you getting up several hours earlier than you usually do know is quite the night owl. So when I asked about what keeps you up at night, nothing keeps him up at night. He's already up at night. So thank you on all counts. Yeah,

Greg 55:27
we really appreciate it. No, thank you.

Noah 55:29
I'm so honored to be a part of this podcast, despite you only giving me 18 hours. That was enough for me to say yes, I think I said yes, in like 10 minutes.

Patrick 55:36
We've given you twice that amount of time.

Noah 55:40
I might have rethought my decision.

Patrick 55:43
That's just that's exactly right. Take care of everybody.

Greg 55:48
All right. Thanks. No, bye. Thank you. Bye. Thanks so much for joining us. Don't forget to tell your friends to subscribe to us on Apple podcasts, Spotify, or wherever they go for audio content within 1/5 of a standard deviation of their logit score. You can also follow us on Twitter where we are at quantity food pod and visit our website quantity food pod.org where you can leave us a message find organized playlists and show notes. Listen to past episodes and other fun stuff. And finally, you can get really cool quantity of merch like shirts, mugs, stickers and notepads from redbubble.com. We're all proceeds go to donorschoose.org to help support low income schools. You've been listening to quantitative, the podcast that has shown absolutely no propensity whatsoever to help anyone achieve balance. Today's episode has been sponsored by Badgett the dating app for Bachelors bachelorettes. And single people everywhere, simply enter your covariates and within a matter of seconds, you find your true propensity pow, Nero caliber matches available for Platinum members only. And by supermodel based inference. When you look at your model and just say to yourself, oh, and GE, that model like literally looks amazing. And finally, by survey statisticians, some of the only people who are actually happy when their kid comes home and says that they're going to earn a little extra money working the polls. This is most definitely not NPR