[SQUEAKING]

[RUSTLING]

[CLICKING]

ESTHER DUFLO: All right. So today is going to be a bit of a hodgepodge lecture. There are three things that i want to do-- one, to talk a bit about uses of randomization in real life. One is to show you-- spotlight another research project that could give you some ideas of use and then start talking about nonparametric comparison/nonparametric regression, which is also going to serve as an introduction to linear regression, which we are going to do next week. So that's sort of the game plan for today.

In terms of logistics, the problem set will be posted sometime later today. It's the last one. It's a little long, but it's a little long because it adds a question that's really about your project. So it has-- it would be of normal size. Plus, it had a question continuing one step further in thinking about your project. So-- and then you can say bye-bye to problem sets.

And I guess you don't even need to do it if you have done enough. But you still should do this part and give it to us because you probably want the feedback before we go too far-- you go too far down a path which might turn out to be impractical or something.

So I guess if we had taught this class-- last week's classes a few years ago, in particular in economics class-- we might have been different in a biostats class. But in an economics class, we would have said, well, this is how one would analyze randomized experiments. But really, they are not that frequent. And you have to think of them as some kind of standard to aspire to.

And now we're going to move to econometrics, which is going to give us a bunch of other tools to try to remove the selection bias. And the way, for example, I was trained as an applied econometrician is to think of the goals-- of the randomization as sort of the benchmark and think to myself, if this was an ideal experiment, this is how it would be constructed, this is how I am close to that or not close to that in a particular setting that I'm looking at.

But of course, things have changed a lot in the last 20 years-- since have changed progressively over the last several decades, but more rapidly in the last 20 years or so, such that now you can think of-- you can still think of randomization in this way, as a sort of analytical tool to help you think about situation where things are not randomized. But you can also just think about randomization as a tool that you can directly use.

And I want to spend just a few minutes talking about various domains where people are using randomized control trials. So the first one, of course, is clinical trials, where they have been around forever testing new drugs. They really are the standard rather than the gold standard, even for many years. The FDA and many other agencies require randomized trials before approving drugs-- several stages of randomized trials, including a last one with a population of live humans.

So you think, that's nice. That makes sense. But I guess the issue that the pharmaceutical or the FDA is facing-- or maybe the entire pharmaceutical industry is facing is that there are very, very strong incentives that are financial incentives. There is a lot of money on the line in developing new drugs and in testing new drugs. And whenever there is a lot of money on the line, they are also very strong incentives on the line.

And a very interesting topic from the point of view of an economist, which Sara is very familiar with, is thinking about the incentive of the pharmaceutical industry in various domains you could think about them. But one of the ways in which these incentives play out is interesting from a statistician's point of view, from an applied statistician's point of view-- is to think about how the incentive that the pharmaceutical world is facing affects the way that we interpret the experimental results.

So let's assume away outright cheating and implementation of the trials themselves and assume that this is sorted-- might be it's an assumption, but let's assume this away. We're still dealing with-- the pharmaceutical world still deal basically with two issues-- one is selective reporting of results from experiments. This is when people conduct a trial. They do it perfectly well, testing the new drug against a placebo. And they only report some of these trials. The results of some of these trials are published, and some are not published.

So what problems does it pose? And how can you think about this problem in the context of everything we've just been discussing about testing, hypothesis testing, the power on tests, and the like? Yeah?

AUDIENCE: It means that their confidence intervals would then be inaccurate because there's some probability that the results aren't due to chance. But that's being ignored because they only show the results that look significant.

ESTHER DUFLO: Right, exactly. So if you think about hypothesis testing, we construct a 95% confidence interval. And they only make sense to the extent that we sort of commit to report the results of all the experiments we are running. Because if we are only reporting the results where things happen to be significant-- and if you run enough experiments-- if you run 100 experiments, 5 of them will be-- and the effect is really 0.

Five of them will be significant at the 95-- will show you a significant result at the 95% interval. Now, 2.5% will be positive, and 2.5% will be negative. And you could report those two. So if you only report the two or three trials that give you significant results, then even a treatment that has no effect whatsoever is going to be reported as having an effect. Yes?

AUDIENCE: It's similar to survivorship bias in finance, where the funds that survive only post results versus the funds that went bankrupt do not. So you think that, [INAUDIBLE] hypothesis testing, they made good returns.

ESTHER DUFLO: So I think it is slightly different. It might be the same problem. I don't know this problem. I haven't thought about this problem in so much detail. But I would think this creates also a bias in the sense that only the good friends report results. So you are sampling from a population that is actually different when you sample from the one who have survived.

Whereas here, there is a distribution of possible-- suppose that the true effect is 0. Think about the setting where we are doing Fisher exact test. The true effect is 0. If we had done the experiment 100 times, even within the same setting, we would have a distribution of the results centered around 0 if the true effect is 0.

But if we only report the ones that actually do work out-- by definition, some of them will work out in the large sample. So there is no-- there might not be anything that these guys are-- the ones that are reported are selected from a different population, et cetera. But if only the positive ones are reported, we don't know how to interpret statistics anymore. So that is a pretty significant problem.

So what is a solution to this problem? or what's the-- in fact, it's not a vague solution. What is the solution that's been implemented for this problem? Yeah?

AUDIENCE: Some policy change when the FDA requires-- or some auditing to make sure that the companies are complying and publishing their full results.

ESTHER DUFLO: So--

AUDIENCE: So--

ESTHER DUFLO: But remember, I said it's publishing the results. Yes?

AUDIENCE: Isn't the clinicaltrials.gov--

ESTHER DUFLO: Exactly. So the solution is clinicaltrials.gov. It exists. You don't need a change. The solution is clinicaltrials.gov. Basically, the FDA refuses to-- will not approve a drug on the basis of a trial that has not been pre-registered. And since 2005, the leading medical journals will actually not publish papers based on trials that have not been registered.

So in principle, we have the universe of studies. That doesn't mean that the journals will publish everything that has been done, like-- and I actually personally think they shouldn't. Because it's not very interesting to publish things that don't work. But it means that the FDA or anybody who is interested in getting the full distribution of results for a particular drug can in principle get it by going to clinicaltrials.gov.

And the project that I want to highlight in a few minutes is exactly about what do you do when you have this bunch of results-- what do you make of it? So that's the solution, is clinicaltrials.gov. There are also other registries that are approved registries that-- basically, before starting a trial, you need to say, I'm going to start a trial. This is my drug. This is my sample. And this is my outcome of interest, is survival after six months-- or whatever is the main outcome you're interested in looking at. Yeah, yep.

So that's what you report in the-- now, it's not a perfect solution. So people are studying the solution. The problem is that the registries-- so the registry only works-- so people are very good at registering. Because if you're not registered, you're screwed. You can't publish. And the FDA will-- but people are very bad at reporting results on the registry.

So in practice, if you're looking for the results for a particular drug, you will have 100 studies on a particular molecule. You might have 100 studies that are registered on the registry, but 95 of them will have no results. They won't have a follow-up. So basically, the trial presumably has been done. Or maybe it's been abandoned. You don't know. There is just no information. So that's a fixable problem-- it's resources. But it remains a problem.

So that's the first problem you have, where it needs some extra work to force people to go back. But of course, at this point, you don't have any-- you can give people incentive to register. But it's much harder to give them incentive to report results of something. Because if something doesn't work, they're not interested in publishing it anyway. So what's the-- there is no bite to anything you could do.

So it's a game of cat-and-mouse anyway because incentives are strong enough that there will always be-- there is kind of disgrace. But I think this is where we are in the cat-and-mouse problem-- that the registry is there. It's followed, but people are not reporting their results.

Incidentally, there is now a registry for social science experiments as well that the American Economic Association started a couple of years ago. There's about 400 studies on it or something like that. There is much less of-- this problem is not absent in social science, but it's much less present, partly because the incentives are not there. Like, there is not so much money on the line, fortunately or unfortunately.

Another issue that clinical trials face, which is, again, a statistical issue-- and it's related-- is selective reporting of results by subgroups. So instead of not reporting-- so you're running your trials. And you're finding no effect on your population. But then you're finding that actually there is an effect for women but not for men. And you will maybe report the overall no-effect but also women and men.

And then you could say, well, the drug actually works on women. But it doesn't work on men. And you can expect to even come up with some theory why it would work for women and not for men. So what is the issue with selective reporting by subgroups? Lisa?

AUDIENCE: Would you be able to find the subgroup that has this, where--

ESTHER DUFLO: Yeah, it's exactly the same, except it's even easier because you don't have to run 100 trials. You have run one trial. If you have enough covariates, you can search enough. Your sample needs to be large enough to start with. But searching enough, you might end up finding a subgroup. And so if you find that the drug actually works for people who have green hair, then you can say, well, this works for the green hair people.

So the solution to that is something that in social sciences we call pre-analysis plan. Or I don't-- it's not a term they use in clinical trials. But it's basically the idea of, in the registry, when you register, you also need to say specifically which subgroup you're going to look at in advance. So you need to say that you're going to look at the effect separately for women and for men, or by age group, or stuff like that.

Even then, if you're planning, let's say, 10 comparisons, you will need to adjust your critical values for the fact that you're doing 10 comparisons at the same time. And hence, one of them by chance is more likely to be significant than if you did just one. But we do have statistical methods to fix the critical value-- none of them is perfectly fantastic, but there are some ways to deal with that.

The problem they have here-- and both of these are actually studies that have been published in *The New England Journal of Medicine* or-- and the problem they have here is that-- so there is a pre-analysis plan, a protocol that is registered. And then the study is sent to a journal. The problem is that nobody follows the protocol of what they were actually going to report.

They report many, many things that were not in the protocol in the first place without saying that they were not in the protocol in the first place. That means that referees and editor don't really go back to the protocol. So it's again like-- but that seems to be easier problem to fix because you can say that if it's not in the-- you could say that it's the job of even a secretary to check that whatever is reported was in fact what was in the protocol. But that's the situation now.

There's also a little bit of a movement in social science to get pre-analysis plans done. The issue is that, again, typically, the question is not that there are millions and billions of dollars on the line to show that a particular treatment works in a subgroup.

And there are many things that-- in social science, we often do experiments in order to learn about some behavior or-- and we may be interested in subgroups we have no idea in advance. So you can go a little too far. And if it's not-- you don't want to show effect versus no-effect where you're trying to learn something about the world. You might go a little too far in specifying in advance what you'd want to do. Yeah?

AUDIENCE: Are we going to talk about adaptive clinical trials in this class? And is that used in social sciences as well?

ESTHER DUFLO: No, we are not going to talk about adaptive medical trials. And it's not used in social science in any systematic form-- in a sense, that people do do stuff that look like that. But it's not theorized very well, et cetera.

But it's a very good question. Because in a sense, this trade-off between-- so in my view, there is absolutely no trade-off with registry. It makes a lot of sense to register that you are doing a trial on this particular stuff, the stuff being your drug or the stuff being an intervention, simply because someone who comes back down the line and wants to study the effect of this stuff-- they need to know the census of what has been done, ideally. So I think there is absolutely no trade-off in doing registries.

But this idea of-- but there is a trade-off between the danger that there is of-- so changing protocols in between creates statistical issues. But on the other hand, it might enrich the results or what we can do with the result from a point of view of understanding the world, be it in medicine or be it in reality.

Changing-- and specifying a lot of your analysis can serve some purpose. In particular, it will protect you against your partner if the partner really wants to see an effect somewhere and will push you to look at subgroups. So when you work with an NGO, they might be very excited. They might really want to see that it works for someone.

On the other hand, it might stifle the creativity and the ability to find something. Though, there are some trade-offs there. Some of the very exciting work on the theoretical level about thinking about experiments is happening exactly there, which is, how can you have rules that on the one hand give you kind of a path to go but at the same time leave you some explicit space for speculation?

That's sort of a whole set of questions around both the design of experiments, the combination of experiments and machine learning tools to think about-- for example, you could specify not the subgroups you're going to look at but how you're going to discover the subgroup with the machine learning rule, in which case you specify something in advance. But you specify a protocol. So you tie your hand in the sense of, this is the rule I'm going to follow, and therefore I'm not letting the data-- I'm not letting the convenience guide me towards the one thing that tests significant at the classical level without taking into account the searching process.

So all sorts of exciting stuff are happening exactly there in terms of the thinking about-- how do we think about this trade-off given the objective of a trial in the first place-- whether it's to give policy advice versus learn something about the world? How do we separate neatly what has been pre-specified and what has not? How do we have pre-specification rules that are flexible, in a way, without being too flexible? So a lot of interesting stuff is happening there, but it's a little bit beyond the scope of this class in the sense that a lot of it is not even done yet. But that's the type of questions people have in mind.

So that's for critical trials. Social experiments-- so these are sort of-- by that, I refer to experiments that are designed to test social policies. So as it have a long history in social science in testing social policies, in particular in the US.

Sort of the most well-known older ones from the '60s and '70s are the Rand Health insurance experiment, which is the first health insurance experiments that was done to test issues of adverse selection and moral hazard in having health insurance. There was a variation in how much people had to pay to be in the program and how much the copays was, et cetera.

So that's the well-known experiments people have looked at. The other old one is the negative income tax experiment, where there was-- people were given kind of a guaranteed income of varying level and different slopes of their wages-- so EITC-like. And there was variation in these two things.

So these were kind of the old one people have studied a lot. They have their issues. They are-- and we've learned a lot about how to do these things since then. But these issues were a boon, in a sense, because they helped us think about a lot of the statistics of these things.

Then MDRC, which is Manpower Development Research Corporation, was started in the '70s. Bob Solow was the chairman of the board. You might think of Bob Solow rightly as the father of growth theory, but turns out he was a very well-read person in general. And he was the chairman of this creation had a lot to do with it.

One of the first leader of it was a woman called Judy Gueron, who is a very interesting person to talk to. And they started to do welfare experiments with state welfare agencies-- so a lot of different experiments about how to help people who were on welfare get out of welfare, basically. And through kind of trials, and errors and political fights, and things like that, they sort of managed to establish the possibility of randomized controlled trials of social experiments on a pretty big scale over the years. And they continue to do it. And they have done a lot to increase the acceptability of the method as a way of testing experiment.

Judy Gueron has a book. And she also has a chapter in a forthcoming handbook on field experiment where she describes this in a very vivid and fun way. If people are interested, that's sort of a very nice reading.

So since then, of course, they have continued all this while and in the last 15 years or so have greatly expanded in scope. In geographical scope, it moved to developing countries, in particular. And in terms of the kind of teams that they talk about-- the size that they have-- well, some trials now involve millions of people. And the ambition in the sense of-- some are about government reforms, corruption, et cetera. I gave you an example with the regulation experiment, for example.

So J-PAL-- of course, the organization that I helped create and still co-direct and is here at MIT-- is only a subset of those experiments. But just J-PAL has today 729 evaluations in 67 countries in teams ranging from agriculture to crime, education, environment, finance/microfinance, health, labor market, and political economy.

You can see big blobs of point in India because that's kind of where it all started. But you can also see a bunch in North America. And this is-- J-PAL North America got started, like, literally five minutes ago. So they had a pretty large expansion when that started-- and sort of all over the world with a big-- I guess not too much going on in Australia. But the Australian people are found a lot in Indonesia. So it's not that they are absent from this trend.

So this has become sort of a much more at least discussed ways of doing things, with people against and for, et cetera. But all that to say, it's not just a theoretical construct. It is there-- and I would hope there to stay.

So that's for policy. And what we are going to do when we talk about experimental design-- we're going to talk a lot about how you design these type of policy experiments to learn what you're interested in-- to learn something about the questions you're interested in. Yep?

AUDIENCE: What are the factors that help you choose your J-PAL sites? Like, do you have relationships, it looks like, with a lot of universities in the area? But what about, like, administrative finance, for example, if you want to do a finance--

ESTHER DUFLO: So to run an RCT, you need a partner-- so someone who is willing to run the project. You need a researcher-- so someone who is willing to take the lead. It doesn't have to be a university researcher, but someone who is willing to take the intellectual lead. And you need money-- someone who is willing to pay for it.

And basically, each of these project needs to-- happens because there is these three things happening together. And the lead can come from either of the three. So sometimes it would be a partner-- for example, a government-- that is interested in doing something. Sometimes it's the researcher that convinces someone to do something. Occasionally it's the money that comes first. That's not the best situation usually, but that happens sometimes. Yeah?

AUDIENCE: Are there usually [INAUDIBLE] for RCTs?

ESTHER DUFLO: A combination of scientific foundations-- NIH/NSF in the US, for example; the equivalent foundations in Europe or in other countries; private foundations-- Gates Foundation, Hewlett Foundation, Sloan Foundation, Arnold Foundation; the partners themselves. So for example, we have a lot of work in Tamilnad, which is in the south of India. The government is paying for them. And so those are the three big sources of money for this type of work.

Of course, there are also randomized evaluations run. So randomization is also something that has relatively recently, in the last 5-10 years or so, made its appearance in a big way in firms as a way to do business. It's not that it was absent before, but it was not that frequent.

And one thing that has really changed that is the internet. Because it becomes so easy to do experiments, obviously. So this is-- so the internet experiments are A/B testing. So A/B testing, strictly speaking, is the comparison of two versions of a web page, version A and version B, where-- so users are-- so you design a new web page.

Maybe there is only one thing that changes-- it's a red button versus a green button. Users, when they try to go on the website, are randomly directed to either of the versions. And then you compare an outcome, which could be, for example, the number of clicks or the number of purchases, whatever the outcome is you wanted. So there are any number of forms and websites on the internet which will help you to do A/B testing.

And from what I-- I did some thoughts in preparation for this class. Basically, as of today, you know enough to start your own company or to do A/B testing. Basically, they will help you design-- think about an hypothesis. But typically, that's going to come from the client.

Then what is your proper sample size? You know that from power calculation. And do the statistics with the right level of-- with the proper tests. So you are all set to do that. There is another name called multivariate testing if you do something much more involved, which involves combining-- or comparing, say, four versions of a web page. So you can also do that.

So for natural reason, A/B testing has become standard practice in a lot of web-based businesses. It's almost free. The metric is natural-- click and stuff like that. It's very quantifiable. And there is a lot of stuff you don't know because we don't really have a theory that people prefer a red button to a green button, or to be shamed into taking an action, or to be encouraged into taking an action, et cetera. So there is many, many, many things that people don't know. And they can test.

One issue that is discussed that is pretty interesting-- because in a way, it's the same issue we are facing in policy-- is, should you try a very different web page-- two very different web pages? Or should you change just one thing at a time? The trade-off, of course, with two different ones is you're testing a package against another. You don't know what make the difference. And maybe you could do even better with another combination. So when you test one thing at a time, you understand exactly that you have tested that thing.

The problem is, without a theory to guide you on what are the type of things that are likely to make a difference, there is-- in even a very simple web page, there is in principle thousands, and thousands, and thousands of possibilities. So you need to do a little bit of thinking to decide what are the relevant variations, unless you have tons of traffic and tons of time and you can keep experimenting until you have something that you prefer.

So my understanding is that people use A/B testing to mean just about anything today as long as it involves the firm and it sounds cooler than RCT. So "A/B testing" is also used to refer to stuff that's not A/B testing in the sense of comparing pages, but that's business models more generally. So that's fine with me. I have no issue with that.

Any question on A/B testing? So-- yeah, go ahead.

AUDIENCE: So you mentioned that you have a lot of features in sort of one template [INAUDIBLE] time. Is automated A/B testing where you say, say, this button is of interest, and you can have some kind of software just make a whole bunch of templates of different sizes and colors, and you can do so for any kind of interface almost?

ESTHER DUFLO: So one could do that. And I think it would be pretty easy to design and to just have a lot of possible pages where people land to. The limitation is always going to be your sample size. So what you want to think about is your power plus the importance of the experiment. So always go back to sort of the basic principle of power calculation that we discussed at the end of the thing, which is, what is the effect size that would be large enough for me to prompt a different action?

So it's not clear to me that you want to necessarily use a lot of-- and then your sample size is your traffic. So if you're Google, you have lots, and lots, and lots, and lots of people. So you can test the Publish button against the purple button against the pinkish button because-- what the hell? But if you're not Google and you have only so many people, then you have to think about what are the hypotheses that-- to use some thinking-- could be some theory, implicit or explicit theory-- about what matters.

The theory could come from psychology, or it could come from economics, for deciding what you need to automatize. Because otherwise, you might also run into exactly the same problem we had with the testing a lot of versions of the thing. One of them will end up working even though-- just by pure chance. [INAUDIBLE] will end up if you don't adjust your-- if you do many, many tests and you don't adjust the standard errors, you're going to get false-- the critical value are false. And if you want to adjust the standard error for doing many, many tests at the same time, then your critical value are going to be so large that you're going to need infinite sample size.

So I'm not sure the automatizing of-- the issue that you have to face in how you set up your A/B testing is really not so much about the logistics of doing it but about what makes sense to test. And you of course have more luxury to test more things when you have larger samples or when you have larger traffic. On the other hand, that's presumably at the beginning, when you don't have so much traffic that it's more relevant to test things.

So I think the advice that they give-- like, I did some searching of how-- companies that will do it for you. The advice that they give is sort of a layman version of what I just said, which is, try to focus on the hypothesis that-- on the one thing that you think is really going to make a difference.

The bottom line of that is that randomization is not ever a substitute for thinking. Some people seem to believe that sometimes. And some people also seem to believe that randomization people think that. But neither of these is actually true.

Randomization is not a substitute for thinking. Once you have a good hypothesis, it's a very nice way to test it. But once-- but it's not going to give you your hypothesis per se.

So moving sort of one step beyond A/B testing is sort of-- other tests on business model. And of course, marketing is one where you kind of always had a certain-- you always had a little bit of randomization in firms. So Capital One was always well known for experimenting a lot with-- since they are mainly a business that have as clients who are sending letters, they were always changing the layout of their letters, et cetera. So they've always been known as being a sort of very randomization-friendly firm.

But beyond Capital One, the use of randomized control trial was surprisingly limited in marketing until fairly recently, when A/B testing has implanted the idea and also has made it easier. But what is interesting in the last few years-- so Duncan Simester, who is a professor here at Sloan, has a very nice paper where he review 61 experiments in marketing since the late-- since 1995. And of the 61 that he reviews that are published in the top marketing journals, 37 got published since 2010.

And what is interesting when you look at the growth is that, of course, a fraction of that growth is due to the fact that some experiments are now done with the internet. But there is also growth of the brick-and-mortar experiments as a maybe spillover of starting to work with the internet. And there are two big questions that people ask themselves in marketing. One is-- which might not be surprising. One is on pricing. And one set of questions is about pricing. And one set of questions is about advertising.

The set of questions about pricing is, how much discount-- what's the price elasticity in general? What's the price elasticity of some specific groups? So it's a lot about the targeting of discounts and stuff like that. And in advertising, it's how-- boosts sort of the shape of how an advertising company should look like, what are people drawn by, et cetera, et cetera.

The difference between what sort of the academic research on marketing and the business world in essence is [INAUDIBLE] try things out and then stick on what works, what appears to work. And in marketing research, they are always-- when they find the results, they always try and uncover what could be the mechanism, which usually involves going back and doing a lot of stuff that were not in their pre-analysis plan. So it goes back to the questions about the tension between pre-analysis plan that gives you a lot of statistical validity without any thinking but might not leave you with much leeway to think about your experiment.

Finally-- and that's kind of what I do for a living-- is use randomization to set up experiments to answer questions that, from a fundamental research point of view, you're interested in. I want to give you one example of that, which is the race question that we started the lectures on [INAUDIBLE] with a couple of lectures ago. And if you remember, the question-- the causal statement was, "she cannot get a job interview because she is Black." And I pointed out that, well, it's not very clear, a super well-defined question, unless we try and think about the counterfactual.

One possible way to set up the counterfactual is to say, conditional on everything that happened to her until the moment that went into building the person who she is how-- her college experience, her education, the fact that she went to ballet classes, or whatever. There is something in just the fact that she's Black which means that she doesn't get an interview. So this is about discrimination by the people who do the hiring.

So the research question would then be testing exactly the hypothesis-- that, in fact, there is no discrimination at the hiring stage-- against the alternative-- that there is discrimination at the hiring stage. That is, two identical people, in term of their background and everything, have less chances to be hired if they are African-American.

So one thing that people used to do and still do to some extent is called audit studies, which is to send trained actors to present themselves as the same people-- but some, say, female versus male trying to buy cars or African-American versus non-African-American trying to get a job-- and see how they were treated differently. The one issue with that is that, once you send a person, it's a little bit hard to keep everything constant except the color of their skin or their gender.

So the next step in that is a paper that was tremendously influential in the sense that they are-- we recently reviewed the literature that followed that paper. And there are literally hundreds and hundreds of studies that have followed this same protocol, is to build resumes. So basically, they made a databank of resumes, using templates from resumes that they could find on monster.com or something like that, and then changed just the first name of the person who sends the resume-- so "Lakisha" for a female African-American name and "Jamal" for a male, a white-sounding, like "Emily" and "Greg."

The way they figured out what sound African-American or white is by first starting with plausible names and then going to people waiting at the station and asking them, like, what do you think? And then they send those resumes in response to a real job application and set up-- and then the callback was going to an answering machine. And they're just counting the callbacks.

So there are pros and cons of that method. A clear pro is that it's pretty clear-- although not 100% clear, but it's reasonably clear that the only thing that differs between the resume is-- it is a fact that the only thing that differ between the resume is the name.

And it is pretty clear that the only thing that differ in the name is a signal of the race. The reason why I'm saying "pretty" is that you could also think that, as parents, deciding to go for a very white-sounding name also goes with sending them to prep school and sending them to ballet. Also, the prep school will be in the resume, but maybe not their ballet.

And there is a lot of discussion these days about cultural fit in organizations, which is often sort of a code word for maybe some form of discrimination-- but maybe not. Maybe it's people just like to be with people that they have always been around. So maybe the very white-sounding name or very Black-sounding name also go with a bunch of culture that actually have some real intrinsic differences. But take that aside, at least we just have the names.

One big drawback of that method, of course, is that you only have the callback. So the outcome is the callbacks. And that's what there is. Potentially, once people do get called back, things could be different at the interview stage. They could be more likely to be hired if they are African-American conditioned on getting an interview or less likely to be hired. We don't know. But--

PROFESSOR: Esther.

ESTHER DUFLO: Yeah?

PROFESSOR: Could you just say something-- just a word or two about the human subjects approval for doing a study like that?

ESTHER DUFLO: So actually-- [LAUGHS]-- this one is pretty interesting. Because you don't need any human subject approval because firms are not people. So for this one, you've not touched anybody. So for this particular one, you've not-- you have-- you sent your interviews to firms.

Of course, legally, you don't need any human subject approval. And in practice, I don't think they got any human subject approval. You might argue that there is a recruiter somewhere who is looking at these things. But of course, you never know who that person is. They'll never-- this information has no chance to be leaked in the first place because you just send it to the big anonymous number. You wouldn't even know who to contact.

So for this particular study, the human subject works or doesn't work-- which is, there are no human subjects involved. It's fake people applying to firms. And firms are not people-- contrary to what the Supreme Court thinks sometimes.

However, many versions of this you would need-- you would bump into human subjects because there is someone who is making the decisions. If you could identify that person, then that would matter. For example, I was involved-- I'm on the human subject committee at the NBER. And there was a study where they proposed to do something like that for financial advice.

And in that case, they are sending it to real people who are financial advisors. And there are two issues-- that, first of all, you're taking their time with people who are not real. And second of all, if they find out that they are likely to be discriminating, that kind of might distress them-- this type of stuff. So as long as the people can be identified who are taking the decisions, those people are humans. And you need to think about how to protect them.

The ones that I find deeply, actually, so potentially offensive-- people have done that on dating sites or sending fake project on dating site. And then of course you do have humans at the other end of it. So a team at Yale who did that without human subject approval-- that people were sent to a journal, actually. And the referees were like, we're not even touching that study because, like, how can you do that? Like, they are people who think they find this very attractive "Lakisha" person, and then they don't exist. So-- which I would agree with, personally.

All right. So what do they find in this case? They look at the callback rates. And they find that-- so on average, when you send 100 applications, you get 9.6 callbacks if you're a white-sounding name and 6.45 if you are a African-American-sounding name. So you basically need to send 50% more resumes to get one more callback. So that's a sign of discrimination.

And then you can look at it by educated, by type of jobs, et cetera, which helped them say a little bit more about what might be going on. In particular, one thing that is quite striking in that data is that the gap actually increases with education. One might think that education would protect you, but in fact-- these are all for administrative jobs.

So I don't know that it would be too-- for trying to become partner in a law firm. But in these positions, actually, being educated and African-American makes it even more difficult compared to-- your callback rate is higher, but the difference between the Black and the white one is even higher. So that can help us think about whether it's-- the theory of discrimination that might underlie this.

If you're interested in that type of things, I'm advertising this same handbook chapter. Marianne Bertrand, who was the co-author on the study, and myself wrote a whole big review on field experiment on discrimination in the-- for a handbook where we talk a lot about these studies, and their limitations/their advantages, and where the literature has gone, et cetera-- and what remains to be done, which is a lot.

The last point I'll make to that is that all of these things are not actually neatly separated in buckets. There really is a lot of interplay between these worlds. There is a blurring of boundaries between research, A/B testing, and policy. And largely, this is very good.

The one thing that it does raise-- the one issue it does raise is human subjects. Because if you remember when we talked about human subjects-- is that firms in principle do not human subjects to do A/B testing. If you want to launch a website tomorrow and you want to do A/B testing, all the power to you.

But if you want to publish the results, you will have to do a human subject-- you have to get human subject approval. Because then you're calling-- and that has not always been happening. And so one has to be a little bit careful in threading this.

But I just want to give one example of this sort of blurring and the fact that not only firms but also governments start to be a bit imaginative about thinking about firms. The White House has launched a Social and Behavioral Science Team. Sometimes they are called the Nudge Unit. I just described it from a White House announcement. Yeah?

AUDIENCE: Which is led by Cass Sunstein.

ESTHER DUFLO: So this was-- Cass Sunstein was at IRA. So Cass Sunstein was the boss of that, if you want. So he was certainly someone who was involved in getting it off the ground. Although, by the time it was actually off the ground, he might have been-- he might have left already. But it's very much the idea.

From the White House announcement-- "a group of experts in applied behavioral science that translates findings and methods from social and behavioral sciences into improvement in federal policies and programs for the benefit of the American people." So what I really like about that-- it's true that there is-- under the influence of Sunstein, there is kind of emphasis on behavioral things-- psychology and stuff. But there is also influence on translating research into action. That is broader than just behavioral traits in principle. And they put a lot of emphasis on trying things out.

So they have a series of descriptions. In their first year of activity, they launched a report. They have a series of description. And here is one example of something they did. They sent a reminder email to 100,000 borrowers and compared them to 100 thousands of non-borrowers who had missed their first payments. The reminder led to a 29% increase in the fraction of borrowers making a payment, from 2.7% making a payment after it was first missed to 3.5%. You might think it's a small difference. But again, sending a letter is extremely cheap.

So this is-- you can see that they worked on a large sample because they are interested in detecting a small difference. Because anybody who repays here instead of default is a huge benefit for the government and for this person-- that they don't have this unpaid debt to deal with. So this is just one example of the type of stuff that keeps happening.

Before moving to non-parametric, I want to give you one example of using-- what do you do when you have, say, a registry? So you have lots of results on one thing. And how do you combine them? How do we-- what do we make of several results on one thing?

I won't do that in detail, but this is also my advertisement section for student's work. So this is Rachel. And she took econometrics in the fall. She's great. She's interested in econometrics, and statistics, and development economics. And in particular, she's very interested in this question of, when you have a lot of RCTs on one subject, what do we make of seven different results? Like, how do we know that they are the same or they are different?

The example I want to give you today is a very topical, very important clinical trial example, which is on vitamin A. So in 1986, Sommer, who is a medical doctor who spends a lot of time in the field in Indonesia, made this startling discovery. It has always been known that vitamin A was good for vision.

And in particular, vitamin A deficiency led to night blindness. So they were giving vitamin A for night blindness. And at some point, he realized that the people to whom he had given the vitamin A-- kids were also more likely to survive.

So he sets up a first RCT in 1986. And it showed a very big, a very large effect on survival rates of infant of systemic vitamin A supplementation in developing countries where vitamin is-- where there is vitamin A deficiency. And then since, then there was a lot of replication, which I'll show you in a minute. But all of them show some effect but with some heterogeneity-- some more, some less, et cetera.

And then in 2013, the result of a large trial, the largest to date, was published. It's called the DEVTA trial. And they found a very precise zero effect. They found no effect. This was in Uttar Pradesh. They did de-worming and vitamin A. And they found no effect on deworming and no effect on vitamin A. It was a bit of a depressing trial, I guess.

And having found this, they decided that they were going to do a meta analysis. What's a meta analysis? It's a way of combining these results from randomized control trials. And now they found that it had a very little-- that vitamin A had very little effect. It got published in *The Lancet*-- huge, blah, blah, blah. And then people are starting to complain that maybe they didn't do their meta analysis right.

So it's pretty important. We're talking about a pretty big, big impact of-- and people don't seem to agree on how to do their meta analysis. So what I want to talk to you about is what the debate is about and maybe who is right in this particular case.

So this is the vitamin A literature. And you can see that the-- so the way doctors usually present results is not in difference but in the ratio, in ratio. So this is a survival ratio for someone-- so the risk ratio. It's the probability of death in the treatment group versus the control group. So it's the ratio in your probability of dying some months after the treatment-- I think it's a year after the treatment-- versus the control group. And so 1 means there is no effect. Anything below 1 means there is an effect.

You can see the first Sommer of 1986 is actually not the largest effect. But it found a pretty large and somewhat noisy effect. And then since then, you have studies that are almost all to the left of 1 but with the exception of the DEVTA. So there are three data that don't seem to be finding an effect-- one exactly at 0, the DEVTA slightly to the left, and that other one there.

So the question is, how do we aggregate these results? So first thing you could do is to say, well, let's just take the average of everything, treat every observation/every study as a data point, and then take the average. Would that be likely to be an efficient way to compare the data? No? Why wouldn't it?

AUDIENCE: Well, there was-- I guess just thinking through it, like, every experiment [INAUDIBLE] seven different methodologies and through sample sizes [INAUDIBLE] countries [INAUDIBLE] there's just way too many variations to just simply aggregate the data.

ESTHER DUFLO: Right. So if you unpack-- so in particular with the sample sizes-- so if we're thinking of the sample sizes, what is the first thing we could think of doing? So taking the raw average of the data point somehow doesn't seem right.

AUDIENCE: Weight them by the number of samples from each. That would be the best next step.

ESTHER DUFLO: Or even better than the number of sample?

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: By what?

AUDIENCE: By their means and [INAUDIBLE].

ESTHER DUFLO: Exactly. So the key thing would be the precision. So the key thing you could say-- next step to sample size-- would be precision. Maybe the precision vary for other things. It's going to be quite correlated with sample size because sample size is a big, big factor in precision of a study.

But if you-- so one thing you could do is to say, well, I'm not just going to average them. I'm going to take a weighted average of the study, giving more weight to the most precise study. So here, it's like, these are the confidence interval in the lines. So you add more weight to the ones that are small confidence interval. So for example, this study would have enormous weight because it has a much smaller confidence interval. It's very precise.

So one thing that it does do is that it's going to give more weight to the more precise observation. One thing that it does not do in the-- and what does it not take into account? In the factor that you talked about-- you said different countries, different sample sizes.

AUDIENCE: I think they might not be representative of the same population.

ESTHER DUFLO: Exactly-- is that, this would be a perfectly fine thing to do if you were persuaded that the treatment effect is actually constant. There is one treatment effect of vitamin A-- that, wherever you live, whoever you are, that's going to help you survive.

One could think-- it is not completely implausible way to think that there is in fact some biological phenomenon that is fundamental. So there is one real effect. And all of the variations we observe is purely due to sampling variation in the different samples.

Here, here, and here. Go ahead. Sorry, I don't know your first name.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Yeah, you.

AUDIENCE: Oh.

ESTHER DUFLO: You seemed--

AUDIENCE: Philippe. I was--

ESTHER DUFLO: Yeah, Philippe. Yes.

AUDIENCE: [INAUDIBLE] my hand. But the-- it's the-- when you-- can you repeat the question? Sorry, I was thinking through some--

ESTHER DUFLO: So here--

AUDIENCE: [CHUCKLES]

ESTHER DUFLO: Here, we-- excuse me. It was not really a question. But, well, people had their hands up anyway. But it was just a statement, which is, if we take the weighted average, weighting by precision, we are-- it's going to certainly work very well if we assume that the treatment effect is constant.

But what we are not taking into account is that maybe the treatment effect is not constant and that the difference in studies does not reflect just the fact that we have different samples and the underlying variation is what we end up doing for the potential outcome but also the fact that the treatment effect might vary from a place to another. Yeah?

AUDIENCE: I have a question. Would the differences in control variables used for each study sort of undermine that effort to average out across precision?

ESTHER DUFLO: If the treatment effect was in fact constant, no. Because you would-- if you had more control variables in one study, that would make it more precise. But basically, in medicine, they don't use that many control variables. So you can think of those as being straight control versus treatment. Yes, sir?

AUDIENCE: I was wondering [INAUDIBLE] do you ever have trouble in terms of timing? So for example, what if the country just looked very different in 1986 compared to how it does

ESTHER DUFLO: Yes. So that's another reason why the treatment effect might vary, is that the timings are very different. So for the reason that it's different countries, different setup-- maybe the nutrition level is different. Maybe the countries become richer. So kids have access to vitamin A in their diet or the like. It would also explain that there is variation in the treatment effect. Yes?

AUDIENCE: I had a question about this specific study. Are all of these studies only testing vitamin A? Or are some of them vitamin A plus or--

ESTHER DUFLO: It's vitamin A, except DEVTA, which has de-worming, too. But I think it's in a crosscut sample. So you can think of them as just vitamin A. Yeah?

AUDIENCE: You mean the way the treatment was [INAUDIBLE]

ESTHER DUFLO: Correctly.

AUDIENCE: [INAUDIBLE]

ESTHER DUFLO: Correctly. Maybe in some study, we have very good administration of the treatments, or 100% of the treated group was treated. Maybe in some study, the treatment was not very well administered. So we still can compare treatment and control. But in effect, it's an effect that's less efficient and less strong in some study because kids didn't actually take it or--

AUDIENCE: But even the way it was given. Like, was it in supplements, or was it--

ESTHER DUFLO: Yeah, it could have been supplement versus in pills, injected. I don't know how you give vitamin A. So I don't know if it can be injected in any way. But that's absolutely right.

So for all these reasons, we have reason to think that the heterogeneity that we see in the world is really a combination of two things. It's a combination of the fact that the treatment effect is going to vary from place to place and a combination of the fact that we also happen to have one particular sample and then-- one particular assignment in this sample. And as we discussed last time, that creates its own noise.

So the first approach is this one. You assume that the treatment effect is constant. You assume it's some tau. You just take the weighted average of all the studies that you have-- this is a tau k for each site-- weighted by the precision of each site. And this is what the DEVTA people did.

Of course, it was slightly self-serving because the DEVTA people aggregated the study, giving the most weight to their own study. So not surprisingly, they find that the average across all sites really reflects what they did.

So what-- the alternative would be to say, well, let's not assume that. But then you need to make another assumption to-- because you can't just say, well, the treatment effect could be just about anything. Otherwise, you don't-- like, what can you say? You can say nothing. You can just show this and say, well, maybe it has an effect somewhere, maybe not.

What you can say-- and this is another Rubin paper-- is to say, well, let's assume that the treatment effect are drawn from a normal distribution with an average-- so now think of the population as the population of potential treatment effect across sites. There is a-- you draw your treatment effect in your site k from this normal distribution, which has some mean, tau, and some standard deviation, sigma-- sigma tau. So assume that they are drawn from that distribution. Those are the real treatment effect for a site.

And then, in addition, the team goes out and performs its-- so in these different sigma comes the way it's administered-- the different country, the different time periods, et cetera. And in addition, once you're-- a site has a particular treatment effect. Great. [SNAPS FINGERS] You're good to go. You're running your study. Then on top, you're going to have a particular sample drawn from the population.

Plus, you're going to have a particular assignment. That's going to generate some noise. So you're going to estimate a tau-hat k, which is tau k drawn from this distribution and with some standard error. And the name of the game now is to estimate tau, tau k and to separate the source of the noise between the heterogeneity of effect and the fact that there is sampling variation in the way that you estimate each effect.

So I'm not going to get into the detail of how it is done. This is a method called Bayesian hierarchical modeling. Once you have that, you can just estimate that thing. If you're interested in doing something like that for your project-- so we're starting in a data set of estimates, which would be fine as long-- that's-- and we're going to post the code that Rachel used for this work.

You can just study to it as an R code. You can study it and adapt it for your purposes. And then she's super friendly. And I'm sure we'll be able to answer questions as you want. Incidentally, that's also true for Gabriel. He told me that, if they have questions, if they find a typo, if they find ways to make the code more effective, please let them know that they can contact me. So both of them would be happy to hear from you.

Anyway, when you do that, you're getting the results. So this is the tau and the standard error-- the confidence interval are based on sigma tau-- so the variability that comes from the treatment effect themselves. And you can see that when you do the Bayesian estimate to this, you find something-- so the fixed effect reported in the DEVTA study is the green one.

It's very sensitive to include it. I'm not including DEVTA, obviously, because it gets so much weight. But the Bayesian hierarchical model is much less sensitive because it does give less weight to this one. Because it takes into account the fact that there are many other studies, actually.

So on balance, what she finds is something-- when you're doing-- so her preferred analysis is this red one over here. She finds the effect-- she finds that, on balance, taking all of the data together, including DEVTA, you don't really change your view very much that it's an effective intervention, with a risk ratio of 0.77 or something like that-- and significantly different from 0.

AUDIENCE: So a quick question.

ESTHER DUFLO: Yep.

AUDIENCE: To build this Bayesian hierarchical model, she didn't need to go into each paper and look at--

ESTHER DUFLO: No, all you need-- if you-- No, if you're just interested in that, you don't even need the data. The nice thing is that, if you are interested to look into subgroups, you might need the micro data, et cetera. But if you're just interested in doing the average, all you need is the reported point estimate and the reported standard error. And sometimes you need the number-- the count of observation, which are typically reported in all papers.

So for meta analyses at that level, all you need is the data that's in the paper. Of course, you'll understand better the-- then you can see, well, there is a lot of variation in the effect. It doesn't want to pull. It's kind of-- it suggests that there is in fact a lot of variation in the actual treatment effect from site to site. And then you can get interested in why some sites are more interesting than-- have lower effects than others, et cetera.

OK? Questions on this one?

AUDIENCE: So just to clarify, to deal with the noise that variations in treatment effects are present, what we do is we pull the treatment effect from a normal distribution?

ESTHER DUFLO: Yes, you assume that it's coming from a normal distribution with some variance and some mean. And you're trying to estimate-- from your data, you're trying to estimate the variance and the mean of the underlying distribution of treatment effect and the variance of the mean for each site.

AUDIENCE: OK.

ESTHER DUFLO: And it's identified because you-- think of it just in term of, count number of parameters you want to identify. You have four things you're trying to identify, tau-- wait, no. You have two things, tau and sigma k. And then for each site, you have a difference. And you have six sites. So it's enough.

In practice, how you do it is by-- basically, you start with some priors, which means you tell, this is what I think. You're trying to put the prior very diffuse, with very large sigma tau, so that it doesn't-- it's not too-- it doesn't pull your data too much in one direction. You say, given this prior, how likely it is that this is the observation that I'm observing.

And then change the prior to make it more likely that you observe this observation. And then go back again, and again, and again. So it's a series of loops, where you start from your prior. You look at your data. You adjust your prior. You look at your data again. You adjust your prior. Look at your data again. You adjust your prior.

So that's the underlying mechanics of the-- it's-- some people couch this as having a lot of philosophical underpinning. But in truth, here, it's really just-- you could set up a maximum likelihood and estimate it. But doing it this way, starting with a prior and adjusting, is actually a much easier computation.

OK. So I do want to use the next 10-15 minutes or so to talk about nonparametric regression. I'm going to skip coming off Smirnoff, but I'll be very happy to have it in my sleeve for when I don't have enough things to say. You might not believe it, but my one fear in life is to not have enough things to say.

But I do want to start with nonparametric regression because this is going to be a nice introduction to looking at linear regression, which is going to start today. So now shift gears. Forget RCTs.

And think we have-- you observe a sample. And in this sample, there are-- you have two random variables observed for your sample, a Y and an X. And you are interested in expressing the conditional expectation of Y given X-- not X given Y, Y given X-- as expectation of Y given X is some function g of x.

So what does it mean? It means that you think that the expectation of Y can be expressed as a function of X, but you just don't know what it is. One way to represent it is that, you think there is a function that, for any possible realization of the random variable-- would give you-- this is Y. And this is X.

You don't have to have any-- and you're willing to say that, for any realization little y, little x, you would say that, y equals g of x plus epsilon-- some epsilon. So there is a function. And then y is here. So this is g of x. So y is g of x plus epsilon.

And you're trying to-- you might think that this relationship is causal-- you might think it's not. It's just a way of expressing the conditional expectation. It doesn't matter. It's causal, or it's not causal. But you're just interested in estimating this shape. A lot of a nice way of describing how two variables might be related. And then you might want to go one step further or not.

So the problem that I want you to think about is how to estimate this guy without imposing a specific functional form. How do we-- what is the best-- so another way to think about it sort of informally is, I have lots of points. My sample is really a bunch of points. Maybe I have one over here also. What's the best fit to all these bunch of points?

So let's think about how to do it. One common way to do this-- and that's not-- so there are various methods to address this problem, estimate this function g of x, once we have specified the model in this way. And a common way and very intuitive way is the kernel regression.

What does the kernel regression do is say, well, we know that E of Y given X-- the E of Y given that X is little x is simply z-- integral of yf of y given x dy. That's just a definition.

And then by Bayes' rule, we also know that f of y given x is simply f of x and y divided by f of x. The integral is first-- concerns both sides, but the integral of f of x dy doesn't depend on y. So we can remove the integral. And so we have on the nominator side integral of y f of x and y dy divided by f of x.

So what this kernel regression does-- it's going to replace f of x and this object by their empirical counterparts. We've already done the part below. And in fact, I taught it to you when we-- this is-- basically, we're trying to estimate the kernel density estimation, right? So we've already done that. We take-- we know what this is. We-- the bottom guy is-- we take a kernel function.

Do you remember what a kernel is? It's some function which-- where the density-- k is some density. So that's a positive function that integrate to 1. So typically, it's sort of bell-shaped-looking. So it's like this little kernel.

So if you remember how we estimate a kernel density, it's a weighted average of the fraction of point that I'm finding within the given bandwidth of x, where h is the size of the bandwidth. I don't know if you remember. I had drawn a graph to express it.

So for example, if I'm interested in estimating-- let me draw a bunch of bandwidth, a bunch of kernels. And then I'm interested in estimating the density at this point. It's just this point. For a while, it's just this point. And after a while, if I'm interested in estimating the density at this point, I'm summing those two.

If I'm interested in estimating-- and then they might-- at this point, there are three that I'm summing. So it goes like this. And I'm estimating the function this way. OK? So this is just a weighted proportion of the observation that are within distance h of some point x. That's the kernel density estimation.

Now, on the top, you could prove-- and I'm not going to try and prove it. But you could prove in about three lines or four lines of calculations and a change in variable that the integral of yf of x and y dy is-- you can take the y out of that integral. So it is simply the same kernel that we had before as a function of x but multiplied by y at every point.

So to estimate the top, if you're interested in estimating the top separately, you would take a weighted average of the-- you don't take just a count of observation. You count the count of observation weighted by yi. But you don't need to do that because you're going to combine them. And what is this now? When we combine them, what does it look like?

I'm not going to answer the question. It's pretty clear what it looks like-- unless there is a typo, in which case you can tell me. But I don't think so.

What is this? We just saw one, except it was another-- I just saw another thing it looks a little bit like. And I told you what it was. That's the thing that I'm not doing. I know. It was in Rachel's stuff. What was this?

AUDIENCE: Weighted precision.

ESTHER DUFLO: A weighted average, yes, where the weight was a precision. Now, what is this?

AUDIENCE: A weighted average.

ESTHER DUFLO: It's a weighted average, yes.

[LAUGHTER]

So at each point x where I'm interested in estimating this guy, it's even simpler than for a density, where I have to sum them up. And it's like-- of course it's R that does it, but it's a bit difficult to draw. A kernel regression is really simple. What I do with a kernel regression is I draw a bandwidth of size h.

And then if I'm interested in estimating and evaluating the function at this point, I draw a bandwidth around this point. And I'm just calculating the weighted average of the y points, of the y's, where the weights are given by the size of the bandwidth-- very simple.

So this guy is going to be somewhere here. It's a weighted average of all the value of y's that falls into this integral-- so in this case, I have these four points-- where I give more weight to the points that are close by. It's very easy. It's-- so in a sense, I could have started from that, saying, this is the kernel. It's just a weighted average of the y point for each x. But the justification is of course where I started from, which is, this is actually what comes from writing the Bayes' rule when I try to estimate e of y given x

So that's kernel. So basically, for each point, you estimate this. And then you can draw it-- you can do this, like, at 100 points, 1,000 points, et cetera-- wherever you're interested in evaluating your kernel. And you draw your graph.

So that's it. That's kernel regression for you. It couldn't be easier. It's really asking at each point what's the average.

AUDIENCE: So it's not linear. And it's not-- you can't express it as a function. It's just a line that goes through your points.

ESTHER DUFLO: It's a line that goes through your points, exactly.

AUDIENCE: It's like a--

ESTHER DUFLO: That's what it is.

AUDIENCE: A moving average or something.

ESTHER DUFLO: Exactly. It's exactly a moving average, where it moves-- it's a moving weighted average. Exactly. That's a kernel regression.

Now, what are the properties of a kernel regression when the-- so first of all, it is biased because the true function might not look like that. But the bias becomes smaller and smaller as the bandwidth goes to 0. So as you reduce the bandwidth size to 0, the bias goes to 0.

Unfortunately, there is another thing that doesn't-- so you could say, well, I'm going to take a 0 bandwidth, and I'm done. But the other problem that you have is that the smaller the bandwidth, the bigger the variance. So what happens with the variance is that it's when n and n, n, n times h goes to infinity that the variance goes to 0.

So what do you promise in a kernel? What's your promise? That's the way Josh Andres talks about it. In a kernel, your promise is that, as your sample size increases, you will also increase your-- you will decrease your bandwidth size. Because as your sample size increases, your n increases. That makes the precision increase. But-- and that allows you to reduce your bandwidth such that the bias will become smaller and smaller. So that's the big choice to be made.

Concretely, how do you do it? You do it by doing something called cross validation. So cross validation tries to minimize mean squared error. So let's first define the prediction error. So the prediction error for an observation is the value of the observation minus the value of g of x at that point.

Now, the cross validation does not use the prediction error itself because it's sort of recursive. It use the prediction error based from using all of the points except x so that it's not self-- it's not circular. So this is the leave-out prediction error as yi minus g minus ixi, where you don't use the observation i to compute the kernel.

So you compute the kernel leaving out observation i. OK? Make sense? For your sample, you just recompute your kernel at the value x without using the observation i itself. And then that's your-- you try to minimize that guy.

And in fact, it's not exactly that. But if you minimized that guy, you would have some issues at the boundary. Because when you're trying to estimate the kernel at the boundary, the bias is very large because you only have half of it. So when you're computing your average over here, you don't have any observation here. So it's going to be tilted towards the right.

So the kernel is biased at this end and that end, which by the way makes it not always the most intuitive nonparametric method. It's not the best for this reason-- that, at the boundary, you're screwed when you're trying to estimate the weighted average.

But suppose you're going with kernel. Then you need to give less weight in your cross validation to observations that are at the boundary. You can even remove them completely. So that's, like, the trimming function that does that. So this is a trimming function.

So basically, the cross validation is the weighted average of the prediction error, where, for each observation, you compute the prediction error by taking the true value of y minus the predicted value for x not using that observation. And then you downplay the observation at the boundary. That gives you your cross validation. Concretely speaking, R will do that for you. So it's not that you need to do it by hand. But this is what R does.

Here is a code in case you-- in case in some problem set that might come up in the very near future there is some nonparametric regression to do, here is a code, where I showed you-- you can-- this chooses the bandwidth to be too large, actually. This is-- I revert the thing. Sorry.

So the bandwidth is too large. So this is oversmoothed. The bandwidth is too small. So this is too squiggly. And this is the optimal bandwidth. And by the way, what was this? This is the relationship of earnings with age-- US male earning-age profile.

And you can actually add standard error. I'm not going to put the regression, but there are formulas for standard errors, which are very, very intuitive and are on the slides. So you can also-- this data will also compute pointwise 95% confidence interval that are plotted here on this graph. But that's it-- easy to understand, trivial to implement.