[CREAKING]

[RUSTLING]

[CLICKING]

SIGI ZHENG: There's a thinking approach that you need to grasp. So I know I'm not as funny and as smart as Juan and Azizi. So I need to make sure you like me. So I brought some food from Trader Joe's for today. And it's a way to make this very, very technical lecture a little bit funny. So that's about today. So today, we will talk about-- we already talked about green buildings and healthy buildings. Now we're going to talk about how to quantitatively measure the benefit of the green building or the or the green city and how to do a model.

But don't be so worried. And this model is so simple, and I'm sure you can understand. So just to quickly recap-- the quickly recap is the following. So for the entire this course, we talk about the business case. We talk about business case. Our point is you cannot [INAUDIBLE] for people, for the planet with three bottom lines-- and very, very important bottom line, the circle, is profit. So we need to understand whether there is a profit in the market that can incentivize the different stakeholders to build and supply, to own and to occupy the green and healthy buildings. So that's the cash flow-- I don't need to go through.

But I want to run through this-- in my first and second green building lecture, I just showed you some papers-- scholarly work, some literature. That's a literature to show there's a green premium. And here is the tenants-- they are willing to pay higher rent or the buyers, they are willing to pay higher price.

And then I show you this. Remember the rental premium of the green buildings? So I say these are all the studies, and Siqi's study is here-- and from all the studies, I give you a takeaway message, which is a significant rent premium of the green buildings, about 6%. And there's some distribution. There's a range.

So how did these scholars-- how did these researchers got to this conclusion of the 6% of the premium? That will be our today's focus. And we also saw some this price premium. This price premium say, oh, these-- all the papers. And the price premium overall is some negative, some positive, some significant, some not significant, but overall is 7.6% So that's basically today's topic is how to understand this premium and how to quantify.

So our outlines are following today-- first, I will still talk about the green building as a starting point, but since this is the methodology, this can apply later to all kinds of things-- healthy building and all climate resilience, all kind of things you want-- how to quantify the premium. And then talk about some cities-- if you understand the green building part, the green city is very easy. And I will talk about some little bit how to apply that to business and policy decision-making if we can estimate that premium.

First, has anyone heard of this hedonic pricing model? Raise your hand if you have heard of this terminology, "hedonic." Have you ever used that in your research-- or not research, like p-set exercise. OK, Wilson, what kind of p-sets you did in Wisconsin-Madison?

AUDIENCE: No, I [INAUDIBLE].

SIGI ZHENG: From your research assistant work with me. OK, good to know. So this is my-- I use hedonic pricing as a way to make a living, so my papers-- so that's why. And so here, after today's lecture, you need to understand this, and you should be able to use that. So that's the point.

But, first of all, I want to give you a big picture of this. So basically, when we want to say, OK, now we want to understand there's some benefit. This benefit is not market good. If this is a market good, like a coffee, like a computer, we can observe its price easily to form the market. But if this is nonmarket goods, like clean air-- nonmarket goods-- so this carbon, until it's priced, is nonmarket goods. Global warming-- nonmarket. You cannot really observe the price of those nonmarket goods. So there are two ways. There are two ways to derive, indirectly, the people's willingness to pay. So that's the first terminology I want you to understand. This is willing to pay-- WPP. We will see a lot of these words today. That's why I need the blackboard here. Consumers', no matter is buyers or tenants, willingness to pay for something.

So there are two ways to derive this if you cannot observe from the market. One is called stated preference. The other is revealed preference. To help you memorize this-- stated preference is basically talking. You talk. You speak. And reveal is doing. One is talking. The other is doing.

So today, we will focus on the doing part. But I want to give you a sense of the talking part. So basically, this short video gives you a sense of this willingness to pay by talking. So I have done this for many times for my own research. For example, I send students to survey these residents about their willingness to pay for a green building. So that's kind of-- I think, basically in your thesis later, maybe you can also apply this method if you want to do a survey and ask.

But today, this is not our focus because I just mentioned in that short video, this talking is not reliable for many, many reasons. If the respondent says that, oh, you ask me-- willing to pay for any public policy to reduce air pollution or all the things, and later, you will use as evidence to increase tax, then they have a tendency to lower their willingness to pay if they think they can just be free riders. So they will increase their willingness to pay as a pressure to others to get this done.

Now housing market or real estate market really provide us a much better opportunity for this just by doing. This is hedonic. So it's not about talking, but we observe. We observe from the real estate market of the real transactions-- price, rent. And you start to back out people's willingness to pay for market or nonmarket attributes. So that's a point.

So why housing market provide us a better opportunity? That's because I think all of you understand that housing market, when you buy a house is not just the structure itself. It's not the concrete. It's not just the columns and the walls and the windows. No, you are buying a place to live. And you are buying that neighborhood that you want to access to many, many opportunities. So that's the point of housing. Housing is so important. That's location based. And you buy or rent a house for so many other things. It's not just the rooms or the kitchen or the walls or the space.

So now give us a way to decompose-- using the housing transaction price to decompose why people are buying this house-- particularly location. So I give you a very simple example for you to understand this. For example, there is a single family house, $1 million. And actually, why this buyer is buying this house by $1 million? Because, of course, we have the household has two kids, and then they need three rooms. They need three rooms. They cannot live in a very small place. So that's three rooms. That's a component. That's very tangible. And then the household also wants to live in a place very convenient. So this household wants to pay for this location because there are two subway stops nearby, for example.

That's a hypothetical example. And then this is so convenient, then they are willing to pay 0.1 million for this accessibility to subway stops. And then there's such a nice park nearby. This household really likes parks. So the [INAUDIBLE] want to buy this location because $1.1 million to pay for the park and plus all this. Another thing is this is a green building. Suppose this household really want to pay for the green building. So because of the green building, they also want to get $1.1 million. So, of course, this is a few examples-- four examples of the real space of the house, transportation accessibility, some environmental park, and the green building stuff. So that's the four things I give-- four examples.

However, there are so many other things. There's a thousands of things related to this house. We cannot do all this. Then we always have this last point, last component called residual. It's a residual. It's like a trash can. It's all the things you cannot observe and you cannot measure, but you also must value those things go into this [INAUDIBLE].

So then you clearly can see this is hedonic model. Hedonic model is decomposition. And you decompose a very complex thing, which is a house-- it's a complex commodity-- into many pieces, a bundle of pieces that you are using money to buy. Now I just randomly give you some numbers, but there is a scientific way to estimate this. So I just throw all these numbers. Of course, cannot be doing this-- scientific way is to estimate.

So how to estimate? That's the hedonic model. So that's the following. 1 million on the left-hand side-- that's the total price, total price, total willingness to pay. Then decompose to now four components plus a trashcan. These four components for each components, there is a unit price times quantity. When we buy apples you have unit price and you have quantity. How many apples? So that's the same thing. So basically, for three rooms-- each room, $0.2 million. Two subway stop-- each subway stop $0.05 million. Then the LEED is a kind of one thing, one tangible thing. And then one nice park-- if you have two nice parks, that will be two parks, but not this one park. So that's a way to decompose the price-- total value to 4 components plus a trash can and each component you have price and quantity. So that's clear. So that's the way.

And how you get these numbers This is a model. This alpha is the unit price. X is the quantity [INAUDIBLE]. How to estimate-- you need data. You cannot just randomly come up with numbers without any data. You have input and output. So the way to do this is to do a regression. [INAUDIBLE] will tell you more. Yes, they give you cover regression? No-- next recitation.

Regression-- for example, we cannot just get one house. If you can only get one house, you cannot do this. You need thousands, hundreds of thousands of observations.

Now we have our big data. We have Zillow. After, we will share with you some Zillow data. So the Zillow data-- we will get the housing transactions in seaport area and the downtown Boston area.

Suppose you get 1,000 observations of the transactions. Then the blue boxes are the input. For each transaction, you will have the total value transaction price. You will have those quantities. You'll know each house is three bedrooms, two bedrooms, or one subway stop, two subway stop, or zero subway stop, and whether it's LEED or not. So you have the input of the quantity and the total price. Then you throw the data into software, which is R, and then the software will report this coefficient. This red things are reported from the software where you fit all the transactions into that. So that's a regression.

Then the point is from all those [INAUDIBLE] regressions, you get p, X1, X2, X3, X4. Then you're wrong. Then you get alpha 1, alpha 2, alpha 3, alpha 4. And it also reported a R-squared-- what's this? You know? OK, please.

AUDIENCE: [INAUDIBLE] the higher the correlation of the variables [INAUDIBLE] dependent variables [INAUDIBLE]

SIGI ZHENG: Good, yes. This is called this explanation power of this equation. So we want this to be as high as possible, but we are not so-- cannot go to 100%. So this 83% means these four variables can explain 83% of the variation among all those observations. 17% is in the trash can. So this, basically, 17% is here because other things you cannot observe. So that's the point of the hedonic-- easy. So that's basically today's point.

And then, of course, there's many, many complexities behind this. But this case is here. And when you do-- the real regression won't be 4, maybe 5, 6, no matter what-- and always, when you do this, you are not without purpose. When you do this, you always have a purpose. Purpose can, in general, you want to do an exercise or you are mostly interested in one coefficient. For example, if I was a researcher to do this, I want to understand alpha 3. I want to understand alpha 3 is people's willingness to pay for a LEED building.

So this willingness to pay-- and it can be many willing to pay. Now we have 4-- 1, 2, 3, 4. I'm most interested in 3, which is the LEED. Of course, if I want to get alpha 3, I must first run this and get all these others. Then I focus on alpha 3. Others are byproducts. For example, there is a transportation researcher-- Mackenzie is a transportation researcher. Then she is not interested in alpha 3. She is interested in alpha 2. The accessibility to subway stop means what? So then that later can relate to other policies.

A little bit of algebra, but not so difficult-- so when just now I see a very intuitive way to decompose-- when you read papers or any research piece, you will see such a model in this way that vectors-- always they take log because it's nonlinear. Log price as a function of some components-- x1 now is a vector. This is the physical attributes as to maybe locational attributes-- maybe each vector has several variables-- number of bedrooms, number of living rooms, and whether these are detached or attached, lots of physical things. So you have a physical basket and some locational basket like that.

Then [INAUDIBLE] this log, how to interpret the log? And just now, I show this level-- is no log thing. Then, you have the absolute value of this is-- that's easy to understand. When you take log on the left-hand side, how you interpret the coefficient-- as a small test of your math?

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: Yes, percentage change. So, for example, you can tell me-- look at this. This is a hypothetical. No price is like this. And I will ask if, let's say, living rooms, number of living rooms, if there is an additional one living room that will increase housing price by how much? 5%.

AUDIENCE: [INAUDIBLE] 5%.

SIGI ZHENG: Yes. So this one will become percentages because we all know that log y basically does this-- not so basic thing. Then you take the derivative. Then this part will become the percentage. So that's basically start of y divided by y. That's a percentage. So this thing will be become percentage, which means incremental one unit increase. One actual living room will lead to a 5% increase in housing price. And this one green park-- then it will increase by 5%. So that's an explanation of this when we take a log because when we look at all these empirical studies, that's all in log. It's not in level.

Here, this is the real paper. So you can see the real paper-- that this paper is by this 2009. And then they say, OK, all other variables, all controlled. There's so many other things. And one-- they say, all other variables controlled-- then one variable called LEED and the other is called Energy Star. So remember, if a number of rooms, not a numerical variable, like one room, two rooms, three rooms-- sometimes, it's a dummy. It's a dummy variable is dummy-- it takes only two values, 0 or 1.

And then, for example, in this example, subway is a dummy. I say dummy means within one mile of subway buffer. If this house-- this is a subway, and this is one mile. The house is here. Then it takes a value of 1. If the house is here, it takes a value of 0. So that's a dummy. So if the dummy, then the coefficient means from 0 to 1 change-- that will incur the housing price increase by 8%. So that's a dummy.

So, based on this principle, how to Interpret the LEED coefficient of 0.06? LEED is the dummy, right? OK.

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: Good. So that's why, from all the empirical studies, we got all those coefficients. Then Energy Star similarly say also 6% price premium. So that's tell you how those papers estimate the price premium by this hedonic regressions. The point is decompose-- they decompose all the housing values. Then they have all other attributes and the coefficient of LEED is 6%. So that's the price premium. It's not difficult at all.

Then a little bit go back to the Winthrop Center that-- this is a teaching case. Remember that? And, at that time-- let me see. You cannot see-- this is a building, right? The lower part is an office. The upper part is a condo. We talk about this a lot. And we have our guest speaker come in, and we understood that for the lower part of office, they chose very, very green, a much higher level of energy efficiency-- Passive House. For the upper level, they didn't because they need to keep the view, and they don't think that's important. Of course, I don't think they run the hedonic when they made this decision. But we can think through this in our way.

For example, Winthrop Center's team brand team can collect the data of nearby-- that's Boston area. If they can collect data from nearby-- office market and condo market because Winthrop center is still under construction. They don't have their own transaction price, but they can collect nearby comparable-- these comparable properties.

For example, they collect it and they run this regression like the following. I show this. Of course, there are so many other things. And for the upper part, the condo market nearby, they want this-- they got this. This I made up-- just want to show the logic-- and the [INAUDIBLE] is like this.

If they say for the upper part, these condo buyers value view so much compared to the lower part because lower part no view-- remember, ugly buildings nearby. So this coefficient 0.3, this coefficient 0.01, which means the people's willing to pay for view is so high compared to the lower.

However, the lower part of the green and the house, they have much higher willingness to pay compared to the condo. This is the relative magnitude of the coefficient can tell you what is the comparative advantage of that part of those attributes. Then this can justify why they chose the upper, the view as the comparative advantage because much higher willingness to pay, but for lower, they chose green and the house as the important attributes. So that's an example of how the market analysis can use hedonic model to assist their quantitative judgment. OK, not challenges-- Wilson, you have a question?

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: Yeah, the thing is that you don't have enough observations because, at that time, Passive House was still very new and maybe you couldn't get enough observation. That's a very real challenge. But besides this practical challenge, suppose we can get a lot of observations. And do you think this kind of thing to derive the coefficient has a challenge? Other challenges?

AUDIENCE: I think it also depends on how you define your variables. For example, you're not controlling for the level of quality of the building. You're just saying it's LEED or not. Maybe you're measuring the quality of the building with the LEED certification-- it's getting confused. So it's how you define it.

SIGI ZHENG: Yeah, you are saying although they call themselves LEED, maybe they don't have that LEED-specific [INAUDIBLE] features.

AUDIENCE: No, maybe if you're controlling for, let's say, three variables-- LEED, view, and health-- maybe the LEED variable is not giving you the value of the LEED building-- it's only giving you giving you the value of a new prime building.

SIGI ZHENG: Prime building, like class A, like that-- exactly. How to call this problem? How to call this problem, this just saying that it's a challenge that may LEED not just to LEED, but also other things bundled into LEED. The LEED building is newer. The LEED building is-- there's famous developers and those kind of things. How to call this?

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: Good. So I call it a very intuitive way. It's called this is not an apples-to-apples comparison. These are apples-to-oranges comparison. So that's why I got this Trader Joe thing because I want you to understand this apple-to-orange comparison. And if you answer my question correctly, I will give this to you.

So now I put here-- apples-to-oranges comparison, I look at this paper. Look at this paper. This is a very famous paper of one of the very early paper of the green building thing. And they collect data of the United States. And then, of course, they collect some green building data. And then they collect nearby building data because they need to-- they need observations, they need data, and then they give this table. It's too small. But I want to show you that they first compare-- treated buildings means green buildings, LEED buildings. Control buildings means nearby non-LEED buildings. That's a terminology. I need to put the terminology here so that you can remember control. Treated means the thing that we are studying, which is green. Control means others that are not green. Then they compare.

Of course, they say, oh, the rent of treated so high-- much higher. However, we also observe class A percentage. The treated green buildings-- 75% of them, they are class A. Only 25% of the control are class A. Age-- the treated buildings are average 24 years old and control-- 53 years old. So that gives you a sense. They are not the same thing. They are not the same animals, this apple and orange. So that's an apple and orange comparison.

So, in our sense, think about this. We observe, oh, this is a green building. This is a nongreen building. The price difference is, like, $1 million. It's a lot. However, how can we so confident? How can we be confident that's because of the greenness. Because the green building also located in such a beautiful neighborhood-- trees, green park, and all the things. This building is such an ugly place and a factory and pollution, all the things nearby-- not compared.

And so then that's basically the problem. How to better get from this apple to orange to get apple to apple? Do you have any answer for that?

AUDIENCE: You would want to compare buildings that were built [INAUDIBLE] close by, and then, also, you would get the percentage of class A buildings [INAUDIBLE].

SIGI ZHENG: You said close by. That's basically location difference. You won't get this-- one is a nice neighborhood, the other in a factory neighborhood. But you say even nearby may be different. It's the old and the new one-- how to do this? Even nearby, two adjacent buildings-- one is new. The other is old.

But this one's for you because I have several. This one.-- yeah, please.

AUDIENCE: Wouldn't there be a way to normalize it so you can compare buildings that are old [INAUDIBLE] so that you're not comparing the same figures, but you're comparing [INAUDIBLE] and you can still use an older building?

SIGI ZHENG: OK, your point-- could you repeat how--

AUDIENCE: I don't know exactly--

SIGI ZHENG: You just want to make this normalized way. Yeah, normalized is a point. And then this one is for you. And I ran out of-- please.

AUDIENCE: You want to add as many variables as you can to make sure that you're comparing the product, so you're going to add in the room count like you were saying, the relative age, condition of the exterior [INAUDIBLE] or the interior, et cetera, so you have a bunch of data points, and then you can look at the relative value of each of those individual data points and fits into the larger formula.

SIGI ZHENG: Good. So we should add as much as possible. The controls-- we call it controls. We can call this control variables, but as we know that always something in the trash can because there's so much difference and how can you control. So now I'm moving on a little bit, and I still need apple to orange. We have two.

Now this is called causal inference. This I think you really need to understand, although maybe you don't want to get technical details, but the terminology and the thinking. Causal inference, whether it's just a correlation or it's causal-- we want causal. We don't want to just play with the correlations.

And the causal inference-- the challenge is we cannot observe the counterfactual. This is a new terminology, "counterfactual." Counterfactual-- something in your dream you really want, but you cannot. That's a counterfactual. And I will try to show this. Let me see. Compare LEED building-- for example, this is a green building-- and another brown building. That's a treatment. This is control.

Then we know the building's in a place so good that I put a small green tree to show this nice neighborhood. And the green building-- the building is in an ugly place-- rocks and random things. So now if we know-- if we compare these two, we know these are different. This is an orange. This is an apple-- cannot-- two different animals.

And then how can we do? Ideally, we want a counterfactual. This is called counterfactual. This is a dream, not real. This counterfactual is all other things are the same as this real one. Let's say a tree is here. All the same as this one-- but the only difference is whether you are green or not. The nearby things, the size and all things-- not counterfactual, but of course-- let's see. I have this one. This is-- and then we know we want this one. All else equal, the effect of LEED-- is this real LEED minus counterfactual? So that's the real thing we want.

However, we don't want this. This is called bias. Bias means this brown building and that brown building-- they have different other things. That's called bias. Bias is disturbing us from reaching the real result. Algebra-- difficult, and so small you cannot see. But just want to tell you, but then, to make sure you understand, I use my fruits from Trader Joe's for you to understand. And later, after many years, we will think about the causal inference-- think about Sigi and the fruits. Then you understand. You remember.

So now the following is a lot. This green building-- say this is green apple-- LEED. Green apple-- green. This is an orange-- totally different thing. Now we need to create a counterfactual which is also apple-- same animal, same fruit-- apple. But this is a red apple or brown apple. So this is a counterfactual, So these are two real things. This is in dream-- cannot-- this dream. We observe these two. But we know these two are different.

Now we create-- make up one thing is a counterfactual. The same animal, but the only difference is the color. So that's the point. The difference between these two are the real effect of LEED. Understand? And these two-- bias. And you need algebra to understand. If you understand this-- econometrics master.

So what we observe is this green apple and an orange-- what we observe. And how to make sure you understand this? We make some real arrangement-- minus red apple, plus a red apple. The same-- cancel out. But we just plus a red apple minus a red apple-- minus and plus.

Then this part, called treatment effect, that's a green apple minus red apple-- same apples. And this one bothers us-- bias. The red apple-- counterfactual. What's the counterfactual difference between this orange? Understand? So this is from the algebra, you artificially add minus that counterfactual will add the counterfactual. Then you have two terms. The upper is the real treatment. This is control. This is bias. So that's the algebra in this graphical way is suddenly your dream come to this. This is not real. And we need this. We don't want this. But in reality, we only have this.

So of course, this is such a challenge. Of course, this is a challenge. Everyone knows it's a challenge. And then very smart people came up with techniques to make sure we can achieve.

And so, basically, as I mentioned, this is the counterfactual. The point of this is to make apples-to-oranges comparison to an apples-to-apples comparison. So that's the key. Without doing so, cannot argue your causal inference.

And then now because some of you, it seems that you understand this, and then what are the things-- what are the things why there is such a bother as the bias? What are the challenges to this? We call it a terminology called endogeneity. Endogeneity is a challenge to causal inference. If we can really observe two apples, no endogeneity, it means we estimate. In reality, we cannot. And then what are the real challenges? Anyone know some examples? Think about the green buildings. Let's use a real example to say why we cannot. Why there are some things that bother us-- not make this bias exist? This is a nongreen building. This is a green building. Both are nongreen building, but why there is a difference? Think about this. Go ahead.

AUDIENCE: [INAUDIBLE] about relevant comparator data, so there would be other factors that [INAUDIBLE] not relevant to what we have--

SIGI ZHENG: Can you-- this is too abstract. Can you make this very intuitive? Green building-- LEED building.

AUDIENCE: If we are observing a building in Boston and another building in New York, the market factors [INAUDIBLE] different. So it's not really an apples-to-apples comparison because the location is a different component.

SIGI ZHENG: Yeah, but always, if we really do this green building premium estimation, we don't go to two markets-- gather some data from New York and Boston. For example, we're all from Boston. And then is there any problem for that? Still different neighborhoods-- still neighborhoods because maybe some rich people live in this area and some poor people live in that area. And for poor people, the green is nothing. They want space. They don't care about the environmental things first. OK, this give you this green building. And other things? Any possible other way because there's some difference you cannot observe. That's called omitted variables. And what other things? Besides [INAUDIBLE] variables, anything you can imagine?

AUDIENCE: [INAUDIBLE] building dependent on the [INAUDIBLE] the brown building has--

SIGI ZHENG: Low rent.

AUDIENCE: [INAUDIBLE] the office tenants [INAUDIBLE] they will prefer to-- if they are preferred to the high rent, maybe surrounded this building [INAUDIBLE]

SIGI ZHENG: Yeah, but that's not the bias. That's what we want to see.

AUDIENCE: Actually, if the [INAUDIBLE] bias [INAUDIBLE]

SIGI ZHENG: OK, [INAUDIBLE] the company-- yes, good. So not just the location, but also the company. Some big companies-- they really value their image. So this is for you. Now I'm going to say this very complicated terminologies, but don't be scared. So in econometrics, we run our model y as a function of some other things. We call this treatment. The thing we are most interested in is D equals 1 is treated-- otherwise, D equals 0. So this is LEED or not. All these other variables are these controls [INAUDIBLE] X as a vector.

So this endogeneity means there's something in the trash can. This is a trash can. Remember, always you have a trash can, and the trash can and the treated variable, they are correlated. They are correlated. They are not independent. If the trash can has so many things in the trash can, but the trash can has nothing to do with the treated, then you won't need to worry. But if they are correlated-- very.

And three times-- just now we already talked about omitted variables, either from the location or from the company. So basically, there's so many things you cannot observe in a big company. You must build this-- a small company investing in this. And this is a nice [INAUDIBLE] neighborhood, all the things.

And second one is called reverse causality. Reverse causality is the story we want to tell-- is from the right-hand side to the left-hand side. That's a regression point. If it's treated, if this green was a premium-- so that's from the right-hand side to left-hand side. Unfortunately, there is another channel from left-hand side to the right-hand side. For example, for those very expensive buildings because of the location or because of anything-- so expensive and charged so high rental price not because of the energy-related things, but they are so-- they collect so much revenue. And the board of the company say we have so much revenue-- let's renovate. If you cannot charge enough, then you have money-- you need to cover daily things.

Now for the very rich, for the very expensive buildings, they have a lot of revenue. Then they decide to treat, not because they treated them become higher, but it's because they collect more revenue, then they decide to treat. That's opposite to what we want to say-- well, the estimate. So that costs reverse causality.

The third thing is called selection bias. That's called selective sample. So the sample you selected, you collect, is not representative of the entire market. For example, if we say let's estimate the green premium, the green price premium for green buildings, then you go to Zillow and you collect. However, we know the transactions for the new houses-- you always transacted. For the old houses, it's not always-- low frequency. Then for the transaction sample you collect it's not the representative of the entire housing stock. You always have more new ones transacted, less old ones transacted.

So then you have a selected buyers because you use a selected sample to estimate this. I'm not sure whether you-- so, basically, this is not underlying population. So that's called selection bias. All three reasons will always cause this problem of called endogeneity. So this is conceptual level, and there are so many technologies. And then we need to do apple-to-apple comparison. So the apple-to-apple comparison-- there are some techniques the economists and the econometrics scholars already developed. Anyone knows any names of them? This kind of how to get to apple-to-apple-- using any terminologies? I feel like you understand. Yeah, please?

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: OK, good. That's one example. Do you have one thing I gave you? No, OK, this one is for you. Mackenzie said regression discontinuity, yes, but maybe you don't know what's this. And other terminologies? Yes-- from your RA work for me, you know something-- please.

AUDIENCE: Instrumental variable.

SIGI ZHENG: Instrumental variable, yes. Since I already paid you, so no fruit-- OK, fine-- because you learn the knowledge from me. But it's so many things-- so many technologies. It's not for you to understand. If you understand this, graduate from the econ department, master's degree or PhD. So, yes, even after PhD, I still cannot do this.

So the thing is, it's so small, but I give you a sense of this. This is fascinating. There's so many good technologies. One, the big category is the following. At first, we come from some naive ways. Naive ways will be accurate responsibility-- not naive. Basic-- not naive. Accuracy is very high level, so we will teach you OLS, which is a basic regression. Even for basic regression, we have ways to control, for example, so-called fixed effect. But I don't want to go to that.

But anyway, if you do our basic-- we call the OLS, Ordinary Least Square. So that's the thing. Then at least you can try your best to put all enough controls-- all the data. If you have so many controls [INAUDIBLE] and have other ways.

And then move on-- more advanced. More advanced to solve this apple to apple comparison thing to achieve this apple-to-apple-- I have some more-- the thing is the following. The golden way, so-called RCT-- that's called Randomized Controlled Trial, RCT. That's really you do an experiment. You do experiment in a lab. Then you really observe-- it's not like you just observe from the real world. You do a lab thing, RCT.

And if you cannot do an experiment-- experiment is very hard to run-- and then you have quasi-experiment, like half, [INAUDIBLE]. Then you have some things-- today, I will talk a little bit about one advanced, which is called matching. You match. And then just now mechanism and this regression discontinuity-- that's later, maybe in our later sections of the climate we will have done. And Wilson said the instrumental variable, and another is called difference-in-differences. So this one, DID-- always use. I'm a DID master. I always use DID, and I will teach you how to do DID for climate-- the impact of climate risk on housing prices. That's for later session. But today, matching is very easy, and not that difficult one to understand.

But the golden way is to do experiment. So experiment-- let me give you a simple idea of the experiment. Remember, once-- lecture on healthy buildings. I think I mentioned it will be the healthy buildings. So, for example, we have a very healthy building-- CO2 concentration's 800 PPM is rather healthy. This is unhealthy-- all the CO2 concentration. Then you have students. You have students-- you. And then we are not going to tell you which are which-- two classrooms. One is healthy. The other is unhealthy. We randomly allocate half of you to one classroom, the other half of you to the other classroom. You don't know, and you don't understand why.

Then we do our final exam, [INAUDIBLE] exam or final exam. Then you do-- and then we compare. We come up-- because we randomized all of you and your other attributes should not have a systematic difference. Of course, each of you are different, but when randomized-- nondifference. And only difference will be this healthy and unhealthy-- then we measure these biological things and your heart beat and all the cognitive performance and your test score.

So this kind of way was used very widely if the researchers have enough resources. It's very hard. How can you get permission from MIT institute to randomize you put half of you to a very unhealthy place-- ethical problem. But if these are manageable, this is a golden way to get apple-to-apple because it's totally randomized apple-to-apple. Then you compare.

For example, primary elementary school-- then you put some air filters. For example, it's a developing country like China-- very polluted. And then for elementary school, it's many classrooms. You put some very good air filters in some classrooms, help clean the air inside. But others, we call it placebo-- also same machine, and the kids, they don't understand the difference. But it's not a real filter, but it's placebo. And then you run them, put the students in those classrooms. After some days, you observe. So that's called the experiment. That's so hard. You must have the collaboration with the elementary school, ministry of education, [INAUDIBLE] center and the parents and all the things.

So that's why-- well, if you do experiment, you publish very well. That's all our scholars' dream to do is experiment and publish well, but it's so hard. For example, we did one in pandemic together-- my lab's team, including several researchers. And we did one publishing a very top journal, *PNAS.* That's basically you randomize people, give them different information, and observe their behaviors. I don't want to go to the details, but I just want to let you know. Do you know two very famous professors at MIT-- econ department? They got Nobel Prize because of this randomized controlled trial in developing countries. Who knows their name?

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: Yes.

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: And her husband. Yes, very good. How did you know them?

AUDIENCE: [INAUDIBLE]

SIGI ZHENG: Oh, yes, they are doing product in India. Yeah so there's a couple-- power couple of two professors. They are the husband and wife, Duflo and Banerjee. They won Nobel Prize for this. They did all kinds of things in India and other countries-- so golden. We are common people. We cannot do that every day. So we need to rely on other technologies.

Now I'm going to tell you, as a common people, student or researcher, we don't need to do all those perfect things, but at least we can do something. The thing I want you to learn from today's class is we need to try. We cannot say I don't care. I just observed correlation and-- no, that's not rigorous. As an MIT student, rigorous-- and say we understand. If we just do a naive regression, get a coefficient of a variable green-- not real. There are some omitted variables. Endogeneity problem is not a causal inference. How to do?

Then I'm going to talk matching. I have so many. And then apple-to-apple-- twins, that's perfect. Some scholars, they did do that. They tracked the twins. The twins-- they were born. Same parents, same environment like that, and later, different outcomes because of the environment-- when a boy went to this city, the girl to that city, and the environment-- then they are long-term, lifetime outcomes [INAUDIBLE]

One thing is geographic matching. So matching-- twins are also matching, but the natural matching-- and a more advanced called propensity score matching. I'm going to skip this, but I will talk about a little bit. PSM is advanced matching, but the thing is geographic matching that's very well-- widely used in the real estate market because we are location things.

So I think some of you already mentioned this. If there's a green building, you are not going to compare a green building in this location to a green building in totally a different location. Like this is Boston downtown. This is a green building. And then you find another place in a suburban someplace-- no. Compare this green building with nearby nongreen building-- very close nearby. That's called geographic matching. Say, 0.2 square miles-- this circle. So that was what they did.

Now [INAUDIBLE] paper-- remember, the [INAUDIBLE] paper? They did this. After this match, they found-- ooh, become so similar. The class A both are 75%. One is 75%. The other is 71%. Of course, not perfect matching, but if you get the green building and there's two brown building so nearby, so same location, and will be similar, then they create [INAUDIBLE] not going to do this, not suburban area to compare-- no. And nearby buildings-- then this is treated this is control. Then they compare. So that's basically matching. So this is not difficult.

After matching, they say, oh, I'm more confident. They also say we are more confident. We are estimating some apple-to-apple thing. This is green rating [INAUDIBLE] 2.6% premium and sale price 13.3% premium. In fact, [INAUDIBLE] is 77.6%. So after matching, they are more confident on this difference instead of a random other place. So that's the matching thing.

Now I'm going to do more of this as-- think about this. So now we understand there's so much difficulties. There's causal inference. And we need some techniques to do apple-to-apple. I want to further understand the truth, the root of why we have problems. I will go back to this naive example. This [INAUDIBLE] come up when I introduce hedonic-- decompose. Remember, $1 million house-- 1, 2, 3, 4 components plus a trash can. Then this hedonic thing-- how to think about this biasing hedonic setting? That's the point. Let me see-- just now, I show this. This is all the same.

Now suppose, in the trash can, although we cannot measure, but we understand from our intuition, we understand the trash can is something called fancy. Fancy, of course, is so hard to measure how fancy-- like this, like that. But we know some buildings are fancier. Some other buildings are less fancy. And then we cannot measure. But we know there's something here. And maybe the fancy thing is correlated with LEED, like the LEED building, the famous designer, and the fancier. Some of you went to New York City, and together with you, we saw some very fancy buildings. So these are correlated.

If these are correlated and you cannot observe fancy in that trash can problem-- suppose if we could observe fancy. That fancy is $0.05 million. Suppose we could. That's a hypothetical thing. Actually, we couldn't. But suppose. And then, let me see if we can-- so fortunately, we can have a matter of fancy-- then no problem. We decompose the trash can to a real fancy and the leftover trash can. Everything is done. Good. Then this one is good.

However, in reality, we cannot observe. And to the extreme case, suppose the fancy-- basically, only the buildings are fancier. Or not only the building-- are nonfancy. And we couldn't-- then we wrongly. overestimate because all the buildings are fancier. So you will see all the buildings are $0.15 million more. But the real LEED is $1.1. And because the fancier-- they are so correlated, 100% perfectly correlated, then all the LEED buildings are fancy buildings. So you bundle them together. You cannot separate and you cannot observe.

Then you overestimate the LEED-- become $0.15 million. Is that clear? Lots of problem. So all the things you cannot observe if the residual-- they are correlated with your interest of the variable-- variable of interest. Then you overestimate or underestimate. This case is overestimate. So that's a problem. Any questions here?

So I want to use this to show you because just now I know it's very abstract in this way. Let me see. Yeah, remember this way? Algebra so abstract. At that time, I said [INAUDIBLE] the trash can got correlated with D. The D is a green. But from algebra, maybe you feel so abstract, so I'm going to just not use this fancy way to show you. The trash can-- something in the trash can got correlated, then problem. You cannot observe. You bundle them together. You estimate a bigger thing, which is not real.

So the thing-- again, think about this: apple to orange. Well, as we observe a LEED building and a brown building-- LEED fancy, brown building not fancy-- the difference is $0.15 million-- naive. This is very naive [INAUDIBLE].

We need to get this another apple, which is an apple-to-apple. And then at first, because this-- first, this [INAUDIBLE] is like this-- fancier green, nonfancy, nongreen. Not right. And we need to get the counterfactual, which is also fancy. And the numbering-- then we calculate real. And then that's an issue. It's not that. Then you match. Suppose that neighborhood are all fancy-- that's our assumption. Suppose geography matching-- we suppose this neighborhood are all fancy buildings. So we compare them. We rule out the fancy factor, and we get the green factor.

I hope you understand. Now move on a little bit-- given all your-- if you understand all this, so easy to extend to say this. That's why I spend so much time [INAUDIBLE] the same thing-- hedonic. And then nothing-- remember, just now, we had all the attributes in X1, which is building attributes-- green or not, how many rooms, like that. There's another factor we call the locational things. That's also important.

Then you decompose locational things. You buy a house not because you only need the space. You also want to be a good school district. That's why I bought my house in Lexington in the first place. At that time, I knew nothing about this city. I asked Azizi for advice. Then he told me several good towns with good schools-- one is Lexington. Then that's why I got Lexington. I said, my first priority is to send my son to a good school so that he later can get at least to a college. So that's school.

Subway-- important transportation. Green park-- air pollution. No one wants to live close to a factory, I think, unless you don't know that's a problem. So if you can decompose a similar things as what we did-- green space, clean air, walkability, other things-- estimate. These slide can summarize many, many past faculty research-- our urban studies and planning.

We have a transportation [INAUDIBLE] two professors, [INAUDIBLE] if they have the data, they estimate subway is important-- public transit. They need to come to Siqi, they don't know how to do-- they come to Siqi and say, can you give me some housing transaction data? No problem. I give them data. Then they use housing transactions to run the models, and they estimate the price premium of the subway station or bus stop, no matter what-- premium. Good. Then we publish a paper.

Next day, a professor called [INAUDIBLE] came-- Siqi, I want to work with you. Housing transaction estimate-- the green parks. That's a housing effect. Why people want to live close to park or [INAUDIBLE] easier to exercise, all the air pollution-- good for their health. They have willingness to pay. Because this is bad thing-- if [INAUDIBLE] bad thing, negative. It's not like all the coefficient are positive that matters-- depends on how you measure that thing.

The positive thing-- we call it positive amenities. Negative thing, we call it negative amenities. Air pollution is negative amenity. So that's why it's negative.

Then we use this to do-- and have a publish a paper-- published house. So that's kind of-- myself are so interested in school, not just because of my son, but I just want to understand the school premium housing price. I did several papers using China's data. This is real data. I estimated a good school district-- 8% price premium. So that's my real paper. So that's all cities.

I just want to showcase my paper. But you don't need to get into this. So I just want to tell a story. At first, I didn't know how to do this. When I came here, 2005, as a visiting scholar, I randomly run into a professor. Now professor is my coauthor. My coauthor told me you cannot just look at real estate as a-- inside the real estate. Real estate is very useful for other things. Study other things using real estate as a channel, as a tool. Real estate is a tool. Housing is a tool-- instead of your destination because people choose where to live, where to work, and where to play through real estate. And many things you cannot observe. You use real estate to break out out people's willingness to pay for so many things. I think, good, let's work together.

Then we produce this paper. And that paper, using Beijing transaction data-- that's many years ago, when I was just graduate student. So I collect all the data [INAUDIBLE] with 920 transactions. At that time, I learned the GIS, and I didn't know. Now I learned GIS. These are things I saw at that time. And land and the subway and bus stop and [INAUDIBLE] subway stops. Crime-- high crime areas. Good schools. Good schools-- where is concentrated [INAUDIBLE] and major universities. Universities are also good. Everyone want to live close-- then you have all this human capital and your mental-- I mean, it's like here.

This air pollution-- bad air pollution, good air pollution, and the park-- big parks. I collect all this. I work day and night. So that's something I want to tell another small story, when I was young. Now I'm not that old, but when I was relatively young, I worked so hard on these things, and I stayed up very late to collect data [INAUDIBLE] run regression. And I understand the hedonic. And later, when I become more senior as a professor, I have my students-- I ask my students to do. But I understand the rationale.

And why I tell this story-- because many of you may say, oh, this is too technical. Oh, I don't need to run regression, I don't want to do the p-set. But if you do the pset right now, later, you become a senior manager or boss, and you know how your employees are doing. You can point out, you didn't consider this variable. Did you consider fancy? Did you consider this and that? This estimate is not right. I know because back to MIT, I learned that from Siqi's class, is apples-to-apples comparison. You are not doing apples-to-apples.

Now don't think that later you will be so senior you don't need to understand this-- today, you understand. Later, you can advise your employees, high level. So that's why I'm-- now, I'm doing it. Now I don't want those kind of things. But because I run this, and I stayed up late running this 20 years ago, now I understand very well. Nobody, no my student, can cheat me. Everything point out. So that's why I want you to do your p sets.

OK, at that time, I was so junior, the other professor, so senior, he said, you run this, you do this, you do this. I did. And then, I did this. That's right. You must go through that process. Then I collect all this. At that time, not that easy to get this data from China. So I got all the data. Then GISA, [INAUDIBLE] I run regression. I learned at that time, Stata, not R. So I did this.

OK, finally, I'm not going to talk hard work. I'm going to talk some fun stuff. Many of us, we went to New York City with Carlos. And we did so many projects. And I went there one night before because I didn't want to take the bus early morning. So I went there early before-- the day before. I enjoyed some Broadway show and other things. And the other morning, I walked around. I walked through this park. I walked by this park because I knew I'm going to study this. And the other one is High Line. So some of you told me, where is the High Line Park?

So I want to use this example. So this Bryant Park is very close by to our first visit, Rockefeller Center, I walked through. And it's not always fancy. Not always fancy. I looked through this. Now, it's fancy. But back to 1980s, 1970s, it was very poor, poorly managed. The city ran out of money. New York city ran out of money of maintaining all the parks in New York City. They almost gave up this park. At that time, crime and the-- drug, drug trade, and crime, and all this trash all over the place. So that suffered so much.

And then, the city ran out of money. No money. No public money. They're going to give up. Then the people around communities, those buildings, those buildings, the companies, and some communities nearby, they self organized together. They went to the government. They negotiate with government saying, we can do this. OK, you don't have money? We are going to do.

They organize a restoration association of this park and they collect money, each one contribute a little bit into a pool. And they get a 15-year long-term lease from the government-- from the government. And then later, they extend to another five years. And then they immediately shut down the park and they started the restoration, the redevelopment of that.

And after five years, they opened and become very good-- very good place. And people stay here, winter shops, and the summer all this concert. And then the nearby housing got appreciation. So that's a way? Why, this nearby. There is a company called green-- Grace Building Company that's a office building nearby. At that time, they put in how many million dollars, I forgot the number, into that pool with the expectation later become from a form of negative amenity, ugly,

to a positive amenity, increase the appreciation value of the office building. They get money back. So that's why they are willing to put money first in. So that's a rationale. Not that all our philosophy [INAUDIBLE] but also, this price appreciation.

And then they-- in during the recession, during the recession, the financial crisis, all office buildings in Manhattan suffered. But they kept gradually higher occupancy rate, and higher rent at that time. So that's example of the private-public partnership. Then they went to the state, they say, we continue to run this park. And we have our civic group to run instead of using money.

And New York City learned this lesson from this park. And then from that time, they didn't put all their own money. They always partner with a private sector group, or the community group to do this maintenance. I think many of us, we went to New York City and Kairos brought us to five mixed-use projects, remember. And we know it's mixed-use projects always is a collaboration between the government and the city, between the city and the private developer, and the owner.

So that way. And then, of course, this is very clear. And I think it was also mentioned a lot about this in [INAUDIBLE] if you are there. But from my side, I want you to understand scientific. When you think about future, in your mind, that no matter whether quantitative or quantitative, in your mind, if you later become a decision maker, whether there's some group come to you, can you give us money to the pool, and we will do this and that, a nearby straight beauty, beauty, find a nearby neighborhood? Then how much money you are going to give? You estimate.

In your mind, you run a quick conceptual-level hedonic model. Then, OK. The park here, then the park is green park, better park. That will increase my property's value by how much? Then, I compare. Even conceptual level, that's important. Instead of say, oh, even the park has nothing to do with me. Then, that's basically in the business decision making.

I did this. This is a consultancy project when I was in China. A city is going to build a subway network. They are going to partner with the private sector. It's a Hong Kong subway company to build the subway. They want to understand how much money they are going to invest, and how much money they want that Hong Kong company to invest.

They promise to the Hong Kong company, if they build the subway, invest, they will get some land around subway stations. That's called TOD. If you invest, the Hong Kong company invest money to build a subway, they get some land parcels around. And they can build the real estate later. So that's always the Hong Kong model of the subway building. Subway construction.

Then I say, no problem. Siqi, hedonic master. I love hedonic. And I tell them, OK, after the subway, the premium 0.5 kilometer circle, 14%. 0.5 to 1, 7.5%, we estimate benefit this. And I estimate if the Hong Kong company can get land parcels around, they can get a benefit of what? And then they say, oh, I understand. The Hong Kong company, you need to invest this money in this so that I can give you real estate. So that's a final way.

OK, High Line Park. At first I learned this from Google. Then I went to New York city. You showed me High Line Park. So that's a thing. At first, it was relatively now [INAUDIBLE] area in the west of Chelsea, right? That's Chelsea, Manhattan. Some low-end shops around, but some relatively middle to low-income people, they have small houses nearby. The city upgrade to High Line Park, and then think about WPP, finally, WPP. We go back to WPP.

At first, like this. This blue line. Right? This is High Line Park or any green space. The blue line is willing to pay. Closer, pay higher. The far one is lower. But because at that time, it's just an abandoned railway, not that good. So it's very flat. Then, the city invests money, build this High Line Park, and then become so fancy, and then become the richer household. They're willing to live nearby and some hotels and restaurants nearby. Then willing to [INAUDIBLE] pay this distance to the High Line Park become like this.

So that's the rationale. But think about the final angle is environmental gentrification. So the good thing is the city become beautiful. High Line Park. However, inevitably, the red line, rich people's willingness to pay to be close, much higher than the original residents'. They bought in. They squeezed out. They pushed out others. So it's called gentrification. So it's always no free lunch. If there are green, amenities, gentrification. Rich people who have higher willingness to pay occupy.

Poor people move out from small shops into fancy restaurant and hotel, five-star hotels. So that's a debate. Later, we can discuss this. But that's a debate, and not like, oh, everything is so good, we invest, then housing price go up. If you go to this urban planning department, you need to think about this. Gentrification. OK, so I think that's all I might finish. And I'm going to give this. Thank you so much. And I will give it to three people on the front line because I want you to sit front. OK, see you next time.