The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or to view additional materials from hundreds of MIT courses, visit MITOpenCourseWare@OCW.MIT.edu.

PHILIPPE RIGOLLET: It's because if I was not, this would be basically the last topic we would ever see. And this is arguably, probably the most important topic in statistics, or at least that's probably the reason why most of you are taking this class. Because regression implies prediction, and prediction is what people are after to now, right? You don't need to understand what the model for the financial market is if you actually have a formula to predict what the stock prices are going to be tomorrow.

And regression, in a way, allows us to do that. And we'll start with a very simple version of regression, which is linear regression, which is the most standard one. And then we'll move on to slightly more advanced notions such as nonparametric regression. At least, we're going to see the principles behind it. And I'll touch upon a little bit of high dimensional regression, which is what people are doing today.

So the goal of regression is to try to predict one variable based on another variable. All right, so here the notation is very important. It's extremely standard. It goes everywhere essentially, and essentially you're trying to explain why as a function of x, which is the usual y equals f of x question-- except that, you know, if you look at a calculus class, people tell you y equals f of x, and they give you a specific form for f, and then you do something. Here, we're just going to try to estimate what this length function is. And this is why we often call y the explained variable and x the explanatory variable.

All right, so we're statisticians, so we start with data. All right, then what does our data look like? Well, it looks like a bunch of input, output to this relationship. All right, so we have a bunch of xi, yi. Those are pairs, and I can do a scatterplot of those guys. So each point here has a x-coordinate, which is xi, and a y-coordinate, which is yi, and here, I have a bunch of endpoints. And I just draw them like that.

Now, the functions we're going to be interested in are often function of the form y equals a plus b times x, OK. And that means that this function looks like this. So if I do x and y, this function looks exactly like a line, and clearly those points are not on the line. And it will basically never happen that those points are on a line. There's a famous T-shirt from, I think, U.C. Berkeley's staff department, that shows this picture and put a line between them like we're going to see it. And it says, oh, statisticians, so many points, and you still managed to miss all of them.

And so essentially, we don't believe that this relationship y is equal to a plus bx is true, but maybe up to some noise. And that's where the statistics is going to come into play. There's going to be some random noise that's going to play out, and hopefully the noise is going to be spread out evenly, so that we can average it if we have enough points. Average it out, OK.

And so this epsilon here is not necessarily due to randomness. But again, just like we did modeling in the first place, it essentially accounts for everything we don't understand about this relationship. All right, so for example-- so here, I'm not going to be-- give me one second, so we'll see an example in a second. But the idea here is that if you have data, and if you believe that it's of the form, a plus b x plus some noise, you're trying to find the line that will explain your data the best, right? In the terminology we've been using before, this would be the most likely line that explains the data. So we can see that it's slightly-- we've just added another dimension to our statistical problem. We don't have just x's, but we have y's, and we're trying to find the most likely explanation of the relationship between y and x.

All right, and so in practice, the way it's going to look like is that we're going to have basically two parameters to find the slope b and the intercept a, and given data, the goal is going to be to try to find the best possible line. All right? So what we're going to find is not exactly a and b, the ones that actually generate the data, but some estimators of those parameters, a hat and b hat constructed from the data.

All right, so we'll see that more generally, but we're not going to go too much in the details of this. There's actually quite a bit that you can understand if you do what's called univariate regression when x is actually a real valued random variable. So when this happens, this is called univariate regression. And when x is in rp for p larger than or equal to 2, this is called multivariate regression.

OK, and so here we're just trying to explain y is a plus bx plus epsilon. And here we're going to have something more complicated. We're going to have y, which is equal to a plus b1, x1 plus b2, x2 plus bp, xp plus epsilon-- where x is equal to-- the coordinates of x are given by x1, 2xp, rp. OK, so it's still linear. Right, they still add all the coordinates of x with a coefficient in front of them, but it's a bit more complicated than just one coefficient for one coordinate of x, OK?

So we'll come back to multivariate regression. Of course, you can write this as x transpose b, right? So this entire thing here, this linear combination is of the form x transpose b, where b is the vector that has coordinates b1 to bp. OK? Sorry, here, it's in [? rd, ?] p is the natural notation.

All right, so our goal here, in the univariate one, is to try to write the model, make sense of this little twiddle here-- essentially, from a statistical modeling question, the question is going to be, what distributional assumptions do you want to put on epsilon? Are you going to say they're Gaussian? Are you going to say they're binomial? OK, are you going to say they're binomial? Are you going to say they're Bernoulli?

So that's going to be what we we're going to make sense of, and then we're going to try to find a method to estimate a and b. And then maybe we're going to try to do some inference about a and b-- maybe test if a and b take certain values, if they're less than something, maybe find some confidence regions for a and b, all right? So why would you want to do this? Well, I'm sure all of you have an application, if I give you some x, you're trying to predict what y is.

Machine learning is all about doing this, right? Without maybe trying to even understand the physics behind this, they're saying, well, you give me a bag of words, I want to understand whether it's going to be a spam or not. You give me a bunch of economic indicators, I want you to tell me how much I should be selling my car for. You give me a bunch of measurements on some patient, I want you to predict how this person is going to respond to my drug-- and things like this.

All right, and often we actually don't have much modeling intuition about what the relationship between x and y is, and this linear thing is basically the simplest function we can think of. Arguably, linear functions are the simplest functions that are not trivial. Otherwise, we would just say, well, let's just predict x of y to be a constant, meaning it does not depend on x. But if you want it to depend on x, then your functions are basically as simple as it gets. It turns out, amazingly, this does the trick quite often.

So for example, if you look at economics, you might want to assume that the demand is a linear function of the price. So if your price is zero, there's going to be a certain demand. And as the price increases, the demand is going to move. Do you think b is going to be positive or negative here?

What? Typically, it's negative unless we're talking about maybe luxury goods, where you know, the more expensive, the more people actually want it. I mean, if we're talking about actual economic demand, that's probably definitely negative. It doesn't have to be, you know, clearly linear, so that you can actually make it linear, transform it into something linear.

So for example, you have this like multiplicative relationship, PV equals nRT, which is the Ideal gas law. If you want to actually write this relationship, if you want to predict what the pressure is going to be as a function of the volume and the temperature-- and well, let's assume that n is the Avogadro constant, and let's assume that the radius is actually fixed. Then you take the log on each side, so you get PV equals nRT.

So what that means is that log PV is equal to log nRT. So that means log P plus log V is equal to the log nR plus log T. So we said that R is constant, so this is actually your constant. I'm going to call it a. And then that means that log P is equal to minus log V. That log P is equal to a minus log V plus log T. OK? And so in particular, if I write b equal to negative 1 and c equal to plus 1, this gives me the formula that I have here.

Now again, it might be the case that this is the ideal gas law. So in practice, if I start recording pressure, and temperature, and volume, I might make measurement errors, there might be slightly different conditions in such a way that I'm not going to get exactly those. And I'm just going to put this little twiddle to account for the fact that the points that I'm going to be recording for log pressure, log volume, and log temperature are not going to be exactly on one line.

OK, they're going to be close. Actually, in those physics experiments, usually, they're very close because the conditions are controlled under lab experiments. So it means that the noise is very small. But for other cases, like demand and prices, it's not a law of physics, and so this must change. Even the linear structure is probably not clear, right. At some points, there's probably going to be some weird curvature happening.

All right, so this slide is just to tell you maybe you don't have, obviously, a linear relationship, but maybe you do if you start taking logs exponentials, squares. You can sometimes take the product of two variables, things like this, right. So this is variable transformation, and it's mostly domain-specific, so we're not going to go into more details of this. Any questions?

All right, so now I'm going to be giving-- so if we start thinking a little more about what these coefficients should be, well, remember-- so everybody's clear why I don't put the little i here? Right, I don't put the little i because I'm just talking about a generic x and a generic y, but the observations are x1, y1, right. So typically, on the blackboard I'm often going to write only xy, but the data really is x1, y1, all the way to xn, yn. So those are those points in this two dimensional plot. But I think of those as being independent copies of the pair xy. They have to have-- to contain their relationship. And so when I talk about distribution of those random variables, I talk about the distribution of xy, and that's the same.

All right, so the first thing you might want to ask is, well, if I have an infinite amount of data, what can I hope to get for a and b? If my simple size goes to infinity, then I should actually know exactly what the distribution of xy is. And so there should be an a and a b that captures this linear relationship between y and x. And so in particular, we're going to try to ask the population, or theoretic, values of a and b, and you can see that you can actually compute them explicitly. So let's just try to find how.

So as I said, we have a bunch of points on this line close to a line, and I'm trying to find the best fit. All right, so this guy is not a good fit. This guy is not a good fit. And we know that this guy is a good fit somehow. So we need to mathematically formulate the fact that this line here is better than this line here or better than this line here. So what we're trying to do is to create a function that has values that are smaller for this curve and larger for these two curves.

And the way we do it is by measuring the fit, and the fit is essentially the aggregate distance of all the points to the curve. And there's many ways I can measure the distance to a curve. So if I want to find so-- let's just open a parenthesis. If I have a point here-- so we're going to do it for one point at a time. So if I have a point, there's many ways I can measure its distance to the curve, right?

I can measure it like that. That is one distance to the curve. I can measure it like that by having a right angle here that is one distance to the curve. Or I can measure it like that. That is another distance to the curve, right. There's many ways I can go for it.

It turns out that one is actually going to be fairly convenient for us, and that's the one that says, let's look at the square of the value of x on the curve. So if this is the curve, y is equal to a plus bx. Now, I'm going to think of this point as a random point, capital X, capital Y, so that means that it's going to be x1, y1 or x2, y2, et cetera. Now, I want to measure the distance. Can somebody tell me which of the three-- the first one, the second one, or the third one-- this formula, expectation of y minus a minus bx squared is-- which of the three is it representing?

AUDIENCE: The second one.

PHILIPPE RIGOLLET: The second one where I have the right angle? OK, everybody agrees with this? Anybody wants to vote for something else? Yeah?

AUDIENCE: The third one?

PHILIPPE RIGOLLET: The third one? Everybody agrees with the third one? So by default, everybody's on the first one? Yeah, it is the vertical distance actually.

And the reason is if it was the one with the straight angle, with the right angle, it would actually be a very complicated mathematical formula, so let's just see y, right? And by y, I mean y. OK, so this means that this is my x, and this is my y. All right, so that means that this point is xy. So what I'm measuring is the difference between y minus a plus b times x. This is the thing I'm going to take the expectation off-- the square and then the expectation-- so a plus b times x, if this is this line, this is this point. So that's this value here. This value here is a plus bx, right?

So what I'm really measuring is the difference between y and N plus bx, which is this distance here. And since I like things like Pythagoras theorem, I'm actually going to put a square here before I take the expectation. So now this is a random variable. This is this random variable. And so I want a number, so I'm going to turn it into a deterministic number. And the way I do this is by taking expectation.

And if you think expectations should be close to average, this is the same thing as saying, I want that in average, the y's are close to the a plus bx, right? So we're doing it in expectation, but that's going to translate into doing it in average for all the points. All right, so this is the thing I want to measure. So that's this vertical distance. Yeah? OK. This is my fault actually. Maybe we should close those shades. OK, I cannot do just one at a time, sorry.

All right, so now that I do those vertical distances, I can ask-- well, now, I have this function, right-- to have a function that takes two parameters a and b, maps it to the expectation of y minus a plus bx squared. Sorry, the square is here. And I could ask, well, this is a function that measures the fit of the parameters a and b, right? This function should be small. The value of this function here, function of a and b that measures how close the point xy is to the line a plus b times x while y is equal to a plus b times x in expectation.

OK, agreed? This is what we just said. Again, if you're not comfortable with the reason why you get expectations, just think about having data points and taking the average value for this guy. So it's basically an aggregate distance of the points to their line. OK, everybody agrees this is a legitimate measure?

If all my points were on the line-- if my distribution-- if y was actually equal to a plus bx for some a and b then this function would be equal to 0 for the correct a and b, right? If they are far-- well, it's going to depend on how much noise I'm getting, but it's still going to be minimized for the best one. So let's minimize this thing.

So here, I don't make any-- again, sorry. I don't make an assumption on the distribution of x or y. Here, I assume, somehow, that the variance of x is not equal to 0. Can somebody tell me why? Yeah?

AUDIENCE: Not really a question-- the slides, you have y minus a minus bx quantity squared expectation of that, and here you've written square of the expectation.

PHILIPPE RIGOLLET: No, here I'm actually in the expectation of the square. If I wanted to write the square of the expectation, I would just do this. So let's just make it clear. Right? Do you want me to put an extra set of parenthesis? That's what you want me to do?

AUDIENCE: Yeah, it's just confusing with the [INAUDIBLE]

PHILIPPE RIGOLLET: OK, that's the one that makes sense, so the square of the expectation?

AUDIENCE: Yeah.

PHILIPPE RIGOLLET: Oh, the expectation of the square, sorry. Yeah, dyslexia. All right, any question? Yeah?

AUDIENCE: Does this assume that the error is Gaussian?

PHILIPPE RIGOLLET: No.

AUDIENCE: I mean, in the sense that like, if we knew that the error was, like, even the minus followed like-- so even the minus x to the fourth distribution, would we want to minimise the expectation of what the fourth power of y minus a equals bx in order to get [? what the ?] [? best is? ?]

PHILIPPE RIGOLLET: Why? So you know the answers to your question, so I just want you to use the words that-- right, so why would you want to use the fourth power?

AUDIENCE: Well, because, like, we want to more strongly penalize deviations because we'd expect very large deviations to be very rare, or more rare, than it would with the Gaussian [INAUDIBLE] power.

PHILIPPE RIGOLLET: Yeah so, that would be the maximum likely estimator that you're describing to me, right? I can actually write the likelihood of a pair of numbers ab. And if I know this, that's actually what's going to come into it because I know that the density is going to come into play when I talk about there.

But here, I'm just talking about-- this is a mechanical tool. I'm just saying, let's minimize the distance to the curve. Another thing I could have done is take the absolute value of this thing, for example. I just decided to take the square root before I did it. OK, so regardless of what I'm doing, I'm just taking the squares because that's just going to be convenient for me to do my computations for now. But we don't have any statistical model at this point. I didn't say anything-- that y follows this. X follows this. I'm just doing minimal assumptions as we go, all right?

So the variance of x is not equal to 0? Could somebody tell me why? What would my cloud point look like if the variance of x was equal to 0? Yeah, they would all be at the same point. So it's going to be hard for me to start fitting in a line, right?

I mean, best case scenario, I have this x. It has variance, zero, so this is the expectation of x. And all my points have the same expectation, and so, yes, I could probably fit that line. But that wouldn't help very much for other x's.

So I need a bit of variance so that things spread out a little bit. OK, I'm going to have to do this. I think it's just my-- All right, so I'm going to put a little bit of variance. And the other thing is here, I don't want to do much more, but I'm actually going to think of x as having means zero.

And the way I do this is as follows. Let's define x tilde, which is x minus the expectation of x. OK, so definitely the expectation of x tilde is what? Zero, OK. And so now I want to minimize in ab, expectation of y minus a plus b, x squared. And the way I'm going to do this is by turning x into x tilde and stuffing the extra-- and putting the extra expectation of x into the a. So I'm going to write this as an expectation of y minus a plus b expectation of x-- which I'm going to a tilde-- and plus b x tilde. OK? And everybody agrees with this?

So now I have two parameters, a tilde and b, and I'm going to pretend that now x tilde-- so now the role of x is played by x tilde, which is now a centered random variable. OK, so I'm going to call this guy a tilde, but for my computations I'm going to call it a. So how do I find the minimum of this thing?

Derivative equal to zero, right? So here it's a quadratic thing. It's going to be like that. I take the derivative, set it to zero. So I'm first going to take the derivative with respect to a and set it equal to zero, so that's equivalent to saying that the expectation of-- well, here, I'm going to pick up a 2-- y minus a plus bx tilde is equal to zero. And then I also have that the derivative with respect to b is equal to zero, which is equivalent to the expectation of-- well, I have a negative sign somewhere, so let me put it here-- minus 2x tilde, y minus a plus bx tilde. OK, see that's why I don't want to put too many parenthesis. OK.

So I just took the derivative with respect to a, which is just basically the square, and then I have a negative 1 that comes out from inside. And then I take the derivative with respect to b, and since b has x tilde. In [? factor, ?] it comes out as well. All right, so the minus 2's really won't matter for me. And so now I have two equations.

The first equation, while it's pretty simple, it's just telling me that the expectation of y minus a is equal to zero. So what I know is that a is equal to the expectation of y. And really that was a tilde, which implies that the a I want is actually equal to the expectation of y minus b times the expectation of x. OK? Just because a tilde is a plus b times the expectation of x. So that's for my a. And then for my b, I use the second one.

So the second one tells me that the expectation of x tilde of y is equal to a plus b times the expectation of x tilde which is zero, right? OK? But this a is actually a tilde in this problem, so it's actually a plus b expectation of x. Now, this is the expectation of the product of two random variables, but x tilde is centered, right? It's x minus expectation of x, so this thing is actually equal to the covariance between x and y by definition of covariance.

So now I have everything I need, right. How do I just-- I'm sorry about that. So I have everything I need. Now, I now have two equations with two unknowns, and all I have to do is to basically plug it in. So it's essentially telling me that the covariance of xy-- so the first equation tells me that the covariance of xy is equal to a plus b expectation of x, but a is expectation of y minus b expectation of x. So it's-- well, actually, maybe I should start with b.

Oh, sorry. OK, I forgot one thing. This is not true, right. I forgot this term. x tilde multiplies x tilde here, so what I'm left with is x tilde-- it's minus b times the expectation of x tilde squared. So that's actually minus b times the variance of x tilde because x tilde is already centered, which is actually the variance of x.

So now I have that this thing is actually a plus b expectation of x minus b variance of x. And I also have that a is equal to expectation of y minus b expectation of x. So if I sum the two, those guys are going to cancel. Those guys are going to cancel. And so what I'm going to be left with is covariance of xy is equal to expectation of x, expectation of y, and then I'm left with this term here, minus b times the variance of x. And so that tells me that b-- why do I still have the variance there?

AUDIENCE: So is the covariance really the expectation of x tilde times y minus expectation of y? Because y is not centered, correct?

PHILIPPE RIGOLLET: Yeah.

AUDIENCE: OK, but x is still the center.

PHILIPPE RIGOLLET: But x is still the center, right. So you just need to have one that's centered for this to work. Right, I mean, you can check it. But basically when you're going to have the product of the expectations, you only need one of the two in the product to be zero. So the product is zero.

OK, why do I keep my-- so I get a, a, and then the b expectation. OK, so that's probably earlier that I made a mistake. So I get-- so this was a tilde. Let's just be clear about the--

So that tells me that a tilde-- maybe it's not super fair of me to-- yeah, OK, I think I know where I made a mistake. I should not have centered. I wanted to make my life easier, and I should not have done that. And the reason is a tilde depends on b, so when I take the derivative with respect to b, what I'm left with here-- since a tilde depends on b, when I take the derivative of this guy, I actually don't get a tilde here, but I really get-- so again, this was not-- so that's the first one.

This is actually x here-- because when I take the derivative with respect to b. And so now, what I'm left with is that the expectation-- so yeah, I'm basically left with nothing that helps. So I'm sorry about. Let's start from the beginning because this is not getting us anywhere, and a fix is not going to help. So let's just do it again. Sorry about that. So let's not center anything and just do brute force because we're going to-- b x squared.

All right. Partial, with respect to a, is giving equal zero is equivalent, so my minus 2 is going to cancel, right. So I'm going to actually forget about this. So it's actually telling me that the expectation of y minus a plus bx is equal to zero, which is equivalent to a plus b expectation of x, is equal to the expectation of y. Now, if I take the derivative with respect to b and set it equal to zero, this is telling me that the expectation of-- well, it's the same thing except that this time I'm going to pull out an x. This guy is equal to zero-- this guy is not here-- and so that implies that the expectation of xy is equal to a times the expectation of x, plus b times the expectation of x square. OK?

All right, so the first one is actually not giving me much, so I need to actually work with the two of those guys. So I'm going to take the first-- so let me rewrite those two inequalities that I have. I have a plus b, e of x is equal to e of y. And then I have e of xy. OK, and now what I do is that I multiply this guy. So I want to cancel one of those things, right? So what I'm going to-- so I'm going to take this guy, and I'm going to multiply it by e of x and take the difference.

So I do times e of x, and then I take the sum of those two, and then those two terms are going to cancel. So then that tells me that b times e of x squared, plus the expectation of xy is equal to-- so this guy is the one that cancelled. Then I get this guy here, expectation of x times the expectation of y, plus the guy that remains here-- which is b times the expectation of x square.

So here I have b expectation of x, the whole thing squared. And here I have b expectation of x square. So if I pull this guy here, what do I get? b times the variance of x, OK? So I'm going to move here. And this guy here, when I move this guy here, I get the expectation of x times y, minus the expectation of x times the expectation of y. So this is actually telling me that the covariance of x and y is equal to b times the variance of x.

And so then that tells me that b is equal to covariance of xy divided by the variance of x. And that's why I actually need the variance of x to be non-zero because I couldn't do that otherwise. And because if it was, it would mean that b should be plus infinity, which is what the limit of this guy is when the variance goes to zero or negative infinity. I can not sort them out.

All right, so I'm sorry about the mess, but that should be more clear. Then a, of course, you can write it by plugging in the value of b, so you know it's only a function of your distribution, right? So what are the characteristics of the distribution-- so distribution can have a bunch of things. It can have movements of order 4, of order 26. It can have heavy tails or light tails.

But when you compute least squares, the only thing that matters are the variance of x, the expectation of the individual ones-- and really what captures how y changes when you change x, is captured in the covariance. The rest is really just normalization. It's just telling you, I want things to cross the y-axis at the right place. I want things to cross the x-axis at the right place. But the slope is really captured by how much more covariance you have relative to the variance of x. So this is essentially setting the scale for the x-axis, and this is telling you for a unit scale, this is the unit of y that you're changing.

OK, so we have explicit forms. And what I could do, if I wanted to estimate those things, is just say, well again, we have expectations, right? The expectation of xy minus the product of the expectations, I could replace expectations by averages and get an empirical covariance just like we can replace the expectations for the variance and get a sample covariance. And this is basically what we're going to be doing. All right, this is essentially what you want.

The problem is that if you view it that way, you sort of prevent yourself from being able to solve the multivariate problem. Because it's only in the univariate problem that you have closed form solutions for your problem. But if you actually go to multivariate, this is not where you want to replace expectations by averages. You actually want to replace expectation by averages here. And once you do it here, then you can actually just solve the minimisation problem.

OK, so one thing that arises from this guy is that this is an interesting formula. All right, think about it. If I have that y is a plus bx plus some noise. Things are no longer on something. I have that y is equal to a bx plus some noise, which is usually denoted by epsilon. So that's the distribution, right? If I tell you the distribution of x, and I say y is a plus b epsilon-- I tell you the distribution of y, and if [? they mean ?] that those two are independent, you have a distribution on y.

So what happens is that I can actually always say-- well, you know, this is equivalent to saying that epsilon is equal to y minus a plus bx, right? I can always write this as just-- I mean, as tautology. But here, for those guys-- this is not for any guy, right. This is really for the best fit, a and b, those ones that satisfy this gradient is equal to zero thing. Then what we had is that the expectation of epsilon was equal to expectation of y minus a plus b expectation of x by linearity of the expectation, which was equal to zero. So for this best fit we have zero.

Now, the covariance between x and y-- Between, sorry, x and epsilon, is what? Well, it's the covariance between x-- and well, epsilon was y minus a plus bx. Now, the covariance is bilinear, so what I have is that the covariance of this is the covariance of xn times y-- sorry, of x and y, minus the variance-- well, minus a plus b, covariance of x and x, which is the variance of x?

Covariance of xy minus a plus b variance of x. OK, I didn't write it. So here I have covariance of xy is equal to b variance of x, right?

Covariance of xy. Yeah, that's because they cannot do that with the covariance. Yeah, I have those averages again. No, because this is centered, right? Sorry, this is centered, so this is actually equal to the expectation of x times y minus a plus bx. The covariance is equal to the product just because this insight is actually centered. So this is the expectation of x times y minus the expectation of a times the expectation of x, plus b minus b times the expectation of x squared.

Well, actually maybe I should not really go too far. So this is actually the one that I need. But if I stop here, this is actually equal to zero, right. Those are the same equations. OK? Yeah?

AUDIENCE: What are we doing right now?

PHILIPPE RIGOLLET: So we're just saying that if I actually believe that this best fit was the one that gave me the right parameters, what would that imply on the noise itself, on this epsilon? So here we're actually just trying to find some necessary condition for the noise to hold-- for the noise. And so those conditions are, that first, the expectation is zero. That's what we've got here. And then, that the covariance between the noise and x has to be zero as well. OK, so those are actually conditions that the noise must satisfy.

But the noise was just not really defined as noise itself. We were just saying, OK, if we're going to put some assumptions on the epsilon, what do we better have? So the first one is that it's centered, which is good, because otherwise, the noise would shift everything. So now when you look at a linear regression model-- typically, if you open a book, it doesn't start by saying, let the noise be the difference between y and what I actually want y to be. It says let y be a plus bx plus epsilon.

So conversely, if we assume that this is the model that we have, then we're going to have to assume that epsilon-- we're going to assume that epsilon is centered, and that the covariance between x and epsilon is zero. Actually, often, we're going to assume much more. And one way to ensure that those two things are satisfied is to assume that x is independent of epsilon, for example. If you assume that x is independent of epsilon, of course the covariance is going to be zero. Or we might assume that the conditional expectation of epsilon, given x, is equal to zero, then that implies that. OK, now the fact that it's centered is one thing.

So if we make this assumption, the only thing it's telling us is that those ab's that come-- right, we started from there. y is equal to a plus bx plus some epsilon for some a, for some b. What it turns out is that those a's and b's are actually the ones that you would get by solving this expectation of square thing. All right, so when you asked-- back when you were following-- so when you asked, you know, why don't we take the square, for example, or the power 4, or something like this-- then here, I'm saying, well, if I have y is equal to a plus bx, I don't actually need to put too much assumptions on epsilon. If epsilon is actually satisfying those two things, expectation is equal to zero and the covariance with x is equal to zero, then the right a and b that I'm looking for are actually the ones that come with the square-- not with power 4 or power 25.

So those are actually pretty weak assumptions. If we want to do inference, we're going to have to assume slightly more. If we want to use T-distributions at some point, for example, and we will, we're going to have to assume that epsilon has a Gaussian distribution. So if you want to start doing more statistics beyond just like doing this least square thing, which is minimizing the square of criterion, you're actually going to have to put more assumptions. But right now, we did not need them. We only need that epsilon as mean zero and covariant zero with x.

OK, so that was basically probabilistic, right. If I were to do probability and I were trying to model the relationship between two random variables, x and y, in the form y is a plus bx plus some noise, this is what would come out. Everything was expectations. There was no data involved.

So now let's go to the data problem, which is now, I do not know what those expectations are. In particular, I don't know what the covariance of x and y is, and I don't know with the expectation of x and the expectation of y r. So I have data to do that. So how am I going to do this?

Well, I'm just going to say, well, if I want x1, y1, xn, yn, and I'm going to assume that they're [? iid. ?] And I'm actually going to assume that they have some model, right. So I'm going to assume that I have that a-- so that Yi follows the same model. So epsilon i [? rad, ?] and I won't say that expectation of epsilon i is zero and covariance of xi, epsilon i is equal to zero. So I'm going to put the same model on all the data. So you can see that a is not ai, and b is not bi. It's the same. So as my data increases, I should be able to recover the correct things-- as the size of my data increases. OK, so this is what the statistical problem look like.

You're given the points. There is a true line from which this point was generated, right. There was this line. There was a true ab that I use to draw this plot, and that was the line. So first I picked an x, say uniformly at on this intervals, 0 to 2. I said that was this one. Then I said well, I want y to be a plus bx, so it should be here, but then I'm going to add some noise epsilon to go away again back from this line. And that's actually me, here, we actually got two points correct on this line. So there's basically two epsilons that were small enough that the dots actually look like they're on the line. Everybody's clear about what I'm drawing?

So now of course if you're a statistician, you don't see this. You only see this. And you have to recover this guy, and it's going to look like this. You're going to have an estimated line, which is the red one. And the blue line, which is the true one, the one that actually generated the data. And your question is, while this line corresponds to some parameters a hat and b hat, how could I make sure that those two lines-- how far those two lines are? And one to address this question is to say how far is a from a hat, and how far is b from b hat? OK? Another question, of course, that you may ask is, how do you find a hat and b hat?

And as you can see, it's basically the same thing. Remember, what was a-- so b was the covariance between x and y divided by the variance of x, right? We check and rewrite this. The expectation of xy minus expectation of x times the expectation of y, divided by expectation of x squared minus expectation of x. The whole thing's-- OK?

If you look at the expression for b hat, I basically replaced all the expectations by bars. So I said, well, this guy I'm going to estimate by an average. So that's the xy bar, and is 1 over n, [? sum ?] from [? i co ?] 1, to n of Xi, times Yi. x bar, of course, is just the one that we're used to. And same for y bar. X squared bar, the one that's here, is the average of the squares. And x bar square is the square of the average.

OK, so you just basically replace this guy by x bar, this guy by y bar, this guy by x square bar, and this guy by x bar and no square. OK, so that's basically one way to do it. Everywhere you see an expectation, you replace it by an average. That's the usual statistical hammer. You can actually be slightly more subtle about this.

And as an exercise, I invite you-- just to make sure that you know how to do this competition, it's going to be exactly the same kind of competitions that we've done. But as an exercise, you can check that if you actually look at say, well, what I wanted to minimize here, I had an expectation, right? And I said, let's minimize this thing. Well, let's replace this by an average first. And now minimize.

OK, so if I do this, it turns out I'm going to actually get the same result. The minimum of the average is basically-- when I replace the average by-- sorry, when I replace the expectation by the average and then minimize, it's the same thing as first minimizing and then replacing expectation by averages in this case. Again, this is a much more general principle because if you don't have a closed form for the minimum like for some, say, likelihood problems, well, you might not actually have a possibility to just look at what the formula looks like-- see where the expectations show up-- and then just plug in the averages instead. So this is the one you want to keep in mind. And again, as an exercise. OK, so here, and then you do expectation replaced by averages. And then that's the same answer, and I encourage you to solve the exercise. OK, everybody's clear that this is actually the same expression for a hat and b hat that we had before that we had for a and b when we replaced the expectations by averages? Here, by the way, I minimize the sum rather than the average. It's clear to everyone that this is the same thing, right? Yep?

AUDIENCE: [INAUDIBLE] sum replacing it [INAUDIBLE] minimize the expectation, I'm assuming it's switched with the derivative on the expectation [INAUDIBLE].

PHILIPPE RIGOLLET: So we did switch the derivative and the expectation before you came, I think. All right, so indeed, the picture was the one that we said, so visually, this is what we're doing. We're looking among all the lines. For each line, we compute this distance. So if I give you another line there would be another set of arrows. You look at their length. You square it. And then you sum it all, and you find the line that has the minimum sum of squared lengths of the arrows.

All right, and those are the arrows that we're looking at. But again, you could actually think of other distances, and you would actually get different-- you could actually get different solutions, right. So there's something called, mean absolute deviation, which rather than minimizing this thing, is actually minimizing the sum from i to co 1 to n of the absolute value of y minus a plus bXi. And that's not something for which you're going to have a closed form, as you can imagine. You might have something that's sort of implicit, but you can actually still solve it numerically. And this is something that people also like to use but way, way less than the least squares one.

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: What did I just what?

AUDIENCE: [INAUDIBLE]

The sum of the absolute values of Yi minus a plus bXi. So it's the same except I don't square here. OK? So arguably, you know, predicting a demand based on price is a fairly naive problem. Typically, what we have is a bunch of data that we've collected, and we're hoping that, together, they can help us do a better prediction. All right, so maybe I don't have only the price, but maybe I have a bunch of other social indicators. Maybe I know the competition, the price of the competition. And maybe I know a bunch of other things that are actually relevant.

And so I'm trying to find a way to combine a bunch of points, a bunch of measures. There's a nice example that I like, which is people were trying to measure something related to your body mass index, so basically the volume of your-- the density of your body. And the way you can do this is by just, really, weighing someone and also putting them in some cubic meter of water and see how much overflows. And then you have both the volume and the mass of this person, and you can start computing density.

But as you can imagine, you know, I would not personally like to go to a gym when the first thing they ask me is to just go in a bucket of water, and so people try to find ways to measure this based on other indicators that are much easier to measure. For example, I don't know, the length of my forearm, and the circumference of my head, and maybe my belly would probably be more appropriate here. And so you know, they just try to find something that actually makes sense.

And so there's actually a nice example where you can show that if you measure-- I think one of the most significant was with the circumference of your wrist. This is actually a very good indicator of your body density. And it turns out that if you stuff all the bunch of things together, you might actually get a very good formula that explains things.

All right, so what we're going to do is rather than saying we have only one x to explain y's, let's say we have 20 x's that we're trying to combine to explain y. And again, just like assuming something of the form, y is a plus b times x was the simplest thing we could do, here we're just going to assume that we have y is a plus b1, x1 plus b2, x2, plus b3, x3.

And we can write it in a vector form by writing that Yi is Xi transposed b, which is now a vector plus epsilon i. OK, and here, on the board, I'm going to have a hard time doing boldface, but all these things are vectors except for y, which is a number. Yi is a number. It's always the value of my y-axis. So even if my x-axis lives on-- this is x1, and this is x2, y is really just the real valued function. And so I'm going to get a bunch of points, x1,y1, and I'm going to see how much they respond. So for example, my body density is y, and then all the x's are a bunch of other things. Agreed with that? So this is an equation that holds on the real line, but this guy here is an r p, and this guy's an rp.

It's actually common to talk to call b, beta, when it's a vector, and that's the usual linear regression notation. Y is x beta plus epsilon. So x's are called explanatory variables. y is called explained variable, or dependent variable, or response variable. It has a bunch of names. You can use whatever you feel more comfortable with. It should actually be explicit, right, so that's all you care about.

Now, what we typically do is that rather-- so you notice here, that there's actually no intercept. If I actually fold that back down to one dimension, there's actually a is equal to zero, right? If I go back to p is equal to 1, that would imply that Yi is, well, say, beta times x plus epsilon i. And that's not good, I want to have an intercept. And the way I do this, rather than writing a plus this, and you know, just have like an overload of notation, what I am actually doing is that I fold back. I fold my intercept back into my x.

And so if I measure 20 variables, I'm going to create a 21st variable, which is always equal to 1. OK, so you should need to think of x as being 1. And then x1 xp. And sorry, xp minus 1, I guess. OK, and now this is an rp. I'm always going to assume that the first one is 1. I can always do that.

If I have a table of data-- if my data is given to me in an Excel spreadsheet-- and here I have the density that I measured on my data, and then maybe here I have the height, and here I have the wrist circumference. And I have all these things. All I have to do is to create another column here of ones, and I just put 1-1-1-1-1. OK, that's all I have to do to create this guy. Agreed? And now my x is going to be just one of those rows. So that's this is Xi, this entire row. And this entry here is Yi.

So now, for my noise coefficients, I'm still going to ask for the same thing except that here, the covariance is not between x-- between one random variable and another random variable. It's between a random vector and a random variable. OK, how do I measure the covariance between a vector and a random variable?

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, so basically--

AUDIENCE: [INAUDIBLE]

PHILIPPE RIGOLLET: Yeah, I mean, the covariance vector is equal to 0 is the same thing as [INAUDIBLE] equal to zero, but yeah, this is basically thought of entry-wise. For each coordinate of x, I want that the covariance between epsilon and this coordinate of x is equal to 0. So I'm just asking this for all coordinates. Again, in most instances, we're going to think that epsilon is independent of x, and that's something we can understand without thinking about coordinates. Yep?

AUDIENCE: [INAUDIBLE] like what if beta equals alpha [INAUDIBLE]?

PHILIPPE RIGOLLET: I'm sorry, can you repeat the question? I didn't hear.

AUDIENCE: Is this the parameter of beta, a parameter?

PHILIPPE RIGOLLET: Yeah, beta is the parameter we're looking for, right. Just like it was the pair ab has become the whole vector of beta now.

AUDIENCE: And what's [INAUDIBLE]?

PHILIPPE RIGOLLET: Well, can you think of an intercept of a function that take-- I mean, there is one actually. There's the one for which betas-- all the betas that don't correspond to the vector of all ones, so the intercept is really the weight that I put on this guy. That's the beta that's going to come to this guy, but we don't really talk about intercept.

So if x lives in two dimensions, the way you want to think about this is you take a sheet of paper like that, so now I have points that live in three dimensions. So let's say one direction here is x1. This direction is x2, and this direction is y. And so what's going to happen is that I'm going to have my points that live in this three dimensional space.

And what I'm trying to do when I'm trying to do a linear model for those guys-- when I assume a linear model. What I assume is that there's a plane in those three dimensions. So think of this guy as going everywhere, and there's a plane close to which all my points should be. That's what's happening in two dimensional. If you see higher dimensions then congratulations to you, but I can't. But you know, you can definitely formalize that fairly easily mathematically and just talk about vectors.

So now here, if I talk about the least square error estimator, or just the least squares estimator of beta, it's simply the same thing as before. Just like we said-- so remember, you should think of as beta as being both the pair a b generalized. So we said, oh, we wanted to minimize the expectation of y minus a plus bx squared, right? Now, so that's in-- for p is equal to 1. Now for p lower than or equal to 2, we're just going to write it as y minus x transpose beta squared.

OK, so I'm just trying to minimize this quantity. Of course, I don't have access to this, so what I'm going to do with them going to replace my expectation by an average. So here I'm using the notation t because beta is the true one, and I don't want you to just-- so here, I have a variable t that's just moving around. And so now I'm going to take the square of this thing. And when I minimize this over all t in rp, the arc min, the minimum is attained at beta hat, which is my estimator. OK? So if I want to actually compute-- yeah?

AUDIENCE: I'm sorry, on the last slide did we require the expectation of [INAUDIBLE] to be zero?

PHILIPPE RIGOLLET: You mean the previous slide?

AUDIENCE: Yes. [INAUDIBLE]

PHILIPPE RIGOLLET: So again, I'm just defining an estimator just like I would tell you, just take the estimator that has coordinates for everywhere.

AUDIENCE: So I'm saying like [? in that sign, ?] we'll say the noise [? terms ?] we want to satisfy the covariance of that [? side. ?] We also want them to satisfy expectation of each [? noise turn ?] zero?

PHILIPPE RIGOLLET: And so the answer is yes. I was just trying to think if this was captured. So it is not captured in this guy because this is just telling me that the expectation of epsilon i minus expectation of some i is equal to zero. OK, so yes I need to have that epsilon has mean zero-- let's assume that expectation of epsilon is zero for this problem.

And we're going to need something about some sort of question about the variance being not equal to zero, right, but this is going to come up later. So let's think for one second about doing the same approach as we did before. Take the partial derivative with respect to the first coordinate of t, with respect to the second coordinate of t, with respect to the third coordinate of t, et cetera.

So that's what we did before. We had two equations, and we reconciled them because it was fairly easy to solve, right? But in general, what's going to happen is we're going to have a system of equations. We're going to have a system of p equations, one for each of the coordinates of t. And we're going to have p unknowns, each coordinate of t.

And so we're going to have the system to solve-- actually, i turns out it's going to be a linear system. But it's not going to be something that we're going to be able to solve coordinate by coordinate. It's going to be annoying to solve. You know, you can guess that what's going to happen, right. Here, it involved the covariance between x and epsilon, right. That's what it involved to understand-- sorry, the correlation between x and y to understand how the solution of this problem was.

In this case, there's going to be only the covariance between x1 and y, x2 and y, x3, et cetera, all the way to xp and y. There's also going to be all the cross covariances between xj and xk. And so this is going to be a nightmare to solve, like, in this system. And what we do is that we go on to using a matrix notation, so that when we take derivatives, we talk about gradients, and then we can invert matrices and solve linear systems in a somewhat formal manner by just saying that, if I want to solve the system ax equals b-- rather than actually solving this for each coordinate of x individually, I just say that x is equal to a inverse times. So that's really why we're going to the equation one, because we have a formalism to write that x is the solution of the system. I'm not telling you that this is going to be easy to solve numerically, but at least I can write it.

And so here's how it goes. I have a bunch of vectors. So what are my vectors, right? So I have x1-- oh, by the way, I didn't actually mention that when I put the lowercase, when I put the subscript, I'm talking about the observation. And when I put the superscript, I'm talking about the coordinates, right? So I have x1, which is equal to x1, x1 [? 1, ?] x 1p, x2, which is 1. x2, 1, x2 p, all the way to xn, which is 1, xn 1, x np. All right, so those are n observed x's, and then I have another y1, y2, yn, that comes paired with those guys. OK?

So the first thing is that I'm going to stack those guys into some vector that I'm going to call y. So maybe I should put an arrow for the purpose of the blackboard, and it's just y1 to yn. OK, so this is a vector in rn. Now, if I want to stack those guys together, I can either create a long vector of size n times p, but the problem is that I lose the role of who's a coordinate and who's an observation. And so it's actually nicer for me to just put those guys next to each other and create one new variable.

And so the way I'm going to do this is-- rather than actually stacking those guys like that, I'm getting their transpose and stack them as rows of a matrix. OK, so I'm going to create a matrix, which here is denoted typically by-- I'm going to write x double bar. And here, I'm going to actually just-- so since I'm taking those guys like this, the first column is going to be only ones. And then I'm going to have-- well, x1, 1, [? 1, ?] x1, p. And here, I'm going to have x n1, x np. OK, so here the number of rows is n, and the number of columns is p. One row per observation, one column per coordinate. And again, I make your life miserable because this really should be p minus 1 because I already used the first one for this guy. I'm sorry about that. It's a bit painful. So usually we don't even write what's in there. So we don't have to think about it. Those are just vectors of size p. OK?

So now that I created this thing, I can actually just basically stack up all my models. So Yi equals Xi transpose beta plus epsilon i for all i equal 1 to n. This transforms into-- this is equivalent to saying that the vector y is equal to the matrix x times beta plus a matrix, plus a vector epsilon, where epsilon is just epsilon 1 to epsilon n, right. So I have just this system, which I write as a matrix, which really just consists in stacking up all these equations next to each other.

So now that I have this model-- this is the usual least squares model. And here, when I want to write my least squares criterion in terms of matrices, right? My least squares criterion, remember, was sum from i equal 1 to n of Yi minus Xi transposed beta squared. Well, here it's really just the sum of the square of the coefficients of the vector y minus x beta. So this is actually equal to the norm squared of y minus x beta square.

That's just the square. Norm is, by definition, the sum of the square of the coordinates. And so now I can actually talk about minimizing a norm squared, and here it's going to be easier for me to take derivatives. All right, so we'll do that next time.