SARA ELLISON: OK, so last time right at the end of the lecture, I had introduced a more general linear model, the multivariate linear model. And I had just gone through the first couple of these slides, saying let's analyze this model using a different notation, in particular matrix notation, because the summation notation was just too clunky. It wasn't up for the job.

And so let me just go through quickly. Let's see. This was, I think, the next to last slide I had up. So if we let y be the column vector of all of the observations on the dependent variable, then let epsilon be the column vector of all of the errors, and then let x be the matrix, where across the rows of the matrix, we have first the column of ones, and then a column of each of the explanatory variables. And then sort of down the matrix, we have observations on each of the-- well, we have each of the observations.

Each observation corresponds to a row. So if we define this matrix and vectors this way, then we can write our multivariate linear model in the following very parsimonious fashion. y equals x beta plus epsilon. OK, so now, I'm basically going to go through the same assumptions, but in a slightly more general way from when I discussed assumptions in the bivariate model.

The assumptions have to be discussed in a more general way now because actually, they're a little bit more complicated with the multivariate model. So I'm going to condense all of the assumptions into two basic categories. One is the identification assumptions. And two is the assumptions on the error behavior.

OK, so in the multivariate linear model, in order to have identification, in order to be able to estimate our model, we have to have n greater than k plus 1. That just means we have to have more observations than we have explanatory variables, plus 1. And x has to have full column rank of k plus 1.

And what does that mean? In other words, means that the regressors have to be linearly independent. Or another way of saying this is that the matrix x prime, or x transpose x, is invertible. I'm going to go through this in some more detail in just a second.

And then the second main assumption or main category of assumptions are on the error behavior. And these actually are exactly like the assumptions we saw before. I'm just using matrix notation to express them.

So here we have the expectation of epsilon here. Epsilon is a vector. And that's the vector of zeros. The expectation of epsilon, epsilon transpose is equal to sigma squared times the n by n identity matrix. And this is in fact-- this matrix here is, in fact, just the matrix that we denote covariance of epsilon, which I'll show you a picture of it in a second.

It's just a matrix that has the variances of epsilon along the diagonal and the covariances on the off diagonal. And so what we're saying here is that the diagonal is equal to sigma squared. And the off diagonals are zeros.

In a stronger version of this is the epsilon vector has this multivariate normal distribution with this variance/covariance matrix. I'll go into more detail in both of these in just a second. So let's do the-- oh, n by n identity matrix. OK, so let's take a closer look at these assumptions.

So assumption one, the identification assumption, what exactly does this mean? Well, we need to have more observations than regressors. That shouldn't come as a surprise, especially if we think about the bivariate model. We have one regressor. You have to have at least two observations or else you can't draw a line. So this is just sort of generalizes that to higher dimensions.

We can't have any regressors that do not have positive sample variation. So we saw this assumption in the bivariate case. Remember I had a picture that looked like-- I had a picture that looked like this. And I said if all of our observations are on the same value of x, we can't identify how the conditional mean of y changes with x.

Well, again, in the multivariate regression, or the multivariate case, we can't identify a particular parameter if we have a regressor that doesn't have positive sample variation. So all of our regressors have to have positive sample variation.

And then the third one-- and this is the one that actually trips people up sometimes-- is that we can't have any regressors that are linear functions of one or more other regressors. And in matrix notation, that's another way to say that is the regressors are linearly independent. And that turns out to be equivalent to x prime x being invertible. Yep.

STUDENT: Can you give an example of that?

SARA ELLISON: I will give two examples in fact. OK, so here is one example. Let's imagine a case where we want to estimate the effect of schooling, work experience, and age on salary. And we have individual level data. So we have sort of a data set. And we have a bunch of different salaries.

And then we also have each person's years of schooling, each person's years of work experience, each person's age, and maybe some other stuff too. Doesn't matter. Well, it could be in our particular sample, it's quite possible that everyone in our sample started school at age six, went to school until he or she finished school, and then started working.

Wouldn't be crazy if that happened. Well, if in fact that was the case, then the years of schooling plus the years of work experience plus 6 is equal to the age. So if that's the case, we can't estimate this regression equation. And it sort of makes logical sense too in the sense that there's nothing that helps us separately identify what the effect of schooling, and work experience, and age are.

There's no variation that allows us to separately figure out the effects of all three of those, if, in fact, they're collinear in our sample. So we can't estimate such a model. Is that clear?

STUDENT: So you would drop a regressor.

SARA ELLISON: Exactly, you have to drop one of the regressors. Does this make sense? Yes.

STUDENT: If you drop age it still wouldn't work, like if you those?

SARA ELLISON: If you drop age, it still wouldn't work? Yes.

STUDENT: It still wouldn't work or--

SARA ELLISON: It would. It wouldn't work if everyone went to school until age 18. So we would still need to have-- we couldn't have a perfect linear relationship between number of years of schooling-- well, actually, that would just be no sample variation in years of schooling. But if some people went to school until age 18, and some people went to age 20, and some went to age 25, then if we dropped age from this regression, then we could estimate it.

STUDENT: You say why this doesn't hold is that if we took x1, x2, and x3, we could get values of beta 1, beta 2, and all of these, which would essentially make y 0. So 1, 1, and minus 1, for example. So then this equation would go [INAUDIBLE] and then y would be 0 in that case.

SARA ELLISON: So that's not the intuition I have. That may be correct in some way. But that's certainly not the intuition I have. I would just say that my intuition is just that we don't have any variation to separately identify the effects of these three things if they're perfectly linearly associated with one another. Yes.

STUDENT: So what if the two regressor are somewhat relating, but they are not perfectly relating?

SARA ELLISON: So that's an excellent question. So the question was, what if they are closely related? So maybe this doesn't quite hold in our sample. But it comes close to holding. Like we had a few people who went to school at age five instead of age six. And we had like a couple people who took a year off and didn't work. So this linear relationship is close to holding, but not quite.

That's something that Esther might be able to talk about next time. I'm not sure. So basically, in that case, you can-- maybe, maybe not. But anyhow, I'll tell you the answer. In that case, you can estimate this equation. But you end up sort of having trouble separately identifying the coefficients on these variables, on these three variables.

And in fact, they're going to be coefficients that have very-- that your estimator is going to be of very high variance estimator. And so in that case, what you might want to do is still drop one of them. That's going to give you much lower variance estimators for the remaining two. It's going to introduce a little bit of bias, if that regressor belongs in there.

It's going to introduce a little bit of bias. But you might be willing to accept that bias to have much lower variance estimators. Yes.

STUDENT: [INAUDIBLE] that digression of-- what if you had a large data set with a lot of variables. And you don't know that there are regressions in there that do have [INAUDIBLE]? What are the things that can insinuate from this, could be causing problems in my analysis or--

SARA ELLISON: Yeah, so first of all, if I tried to run this regression and this linear relationship existed in my data set, R would throw up its hands and say, you can't do that. So I can't even do it. So you would find that out. If this relationship didn't quite exist, it was close to existing in the data set, but not quite, R would go ahead and give you the results of this.

But one thing that you could do is you could compute the correlation coefficients for all of your regressors before you run the regression and see if any are really highly correlated. That wouldn't necessarily pick up a linear relationship like this. But the other thing you could do is after you run the regression, if you have really large standard errors, that could be a signal to you that you could have this situation that's close to perfect collinearity. Yes.

STUDENT: So if we have two regressors, x1 and x2, could we use the ratio in some ways to [INAUDIBLE] regression, as a third regressor? Or would that also be [INAUDIBLE]?

SARA ELLISON: So that wouldn't induce this sort of perfect collinearity problem. You could. You might not want to do it for other reasons. But yeah.

ESTHER DUFLO: There's one more. Continuing the question on if you had many [INAUDIBLE] and you didn't know which one to pick, [INAUDIBLE] fall away from traditional econometrics, it becomes then this [INAUDIBLE] we're going to talk about when we talk about machine learning. If you really don't want-- traditional econometrics assumes that you have a model that you are trying to test so you don't go on a giant fishing expedition. If you want to go on a giant fishing expedition, there are techniques for that. And that's [INAUDIBLE]. That's what we are going to introduce.

SARA ELLISON: Good. OK, so that's one example. A second example of this perfect multi-colinearity and one that sort of researchers run afoul of all the time is when they use dummy variables to indicate, say, observations falling into exhaustive and mutually exclusive set of classes. So here's an example.

Let's suppose I have a data set. Let's say I go talk to all my friends in the dorm to collect my data set for 1431. And I ask them what pets they have at home. And let's say all of them have pets. We could have a category for no pet as well. But anyhow, let's say all of them have pets.

But they either have a cat, a dog, or a fish. And so then I create three different dummy variables. One is equal to 1 if they have a cat, and 0 otherwise. 1 is equal to 1 if they have a dog, and 0 otherwise. And 1 is equal to 1 if they have a fish, and 0 otherwise. And everyone has exactly one of those pets.

I cannot include all three of those dummy variables in the regression because if we add up those three dummy variables, we get a column of ones. And that's perfectly co-linear with our column of ones that allows us to estimate the intercept. Now, there are other ways. You can, in fact, decide to include all three dummy variables and not include an intercept, not estimate an intercept in this regression.

It's entirely equivalent. It would be a little troubling if it wasn't equivalent. But it is in fact entirely equivalent. It's just have to interpret the coefficient estimates a different way. But you can't have both an intercept in your regression and a set of dummy variables that are a full set of exhaustive and mutually exclusive classes.

And like I said, R will not let you do this anyhow. OK, so now the second assumption or second-- yes.

STUDENT: [INAUDIBLE] data?

SARA ELLISON: It only changes the interpretation. So basically, I think I'll leave that question for Esther because she will be giving examples of how to use dummy variables in regressions and how to interpret the coefficients. So we'll leave that till later. OK, so then the second assumption was about the error behavior.

And as I said before, these assumptions aren't different for the multivariate model. It's just that I've expressed them using matrix notation. So let me just go through these exactly the same assumptions we saw in the bivariate model. Here I'm using matrix notation. So I'll just go through and show you what they mean.

So first of all, expectation of epsilon is equal to 0. Epsilon is a vector. And so it's just equal to a vector of zeros. And then for some reason, we often write instead of writing the assumption as the covariance matrix of epsilon equals sigma squared times the identity, the n by n identity matrix, we write it as the expectation of epsilon epsilon transpose is equal to that.

Well, it turns out because the expectation of epsilon is identically equal to 0, this matrix is equal to this matrix. You can do the calculations. It's just two lines to convince yourself. But I could have expressed this assumption by just saying that this matrix is equal to this.

But for whatever reason, we often see it written as this matrix is equal to this, same thing. Does everyone understand why this matrix equaling this is exactly the same assumptions we saw before? So this matrix is just simply a matrix containing the variances of the epsilons on the diagonal. So remember before we said each epsilon had variance sigma squared? That was our homoscedasticity assumption.

So each variance is sigma squared. And then all the covariances were 0. That was our no serial correlation assumption. All of the off diagonals here are 0. So it's the same thing, just in matrix form.

Yeah, I think I said this verbally. But this thing is denoted covariance of epsilon. And it's called the variance-covariance matrix of epsilon. OK, fine. We've got this linear model. We've got these assumptions. Just like before, we're going to now ask the question, how do we get beta hat? And what distribution does beta hat have?

And the answers are not going to be surprising. But they're going to be more beautiful than they were last time. So what is beta hat? Well, it's a vector that minimizes the sum of squared errors. So we've got a vector of residuals transpose times a vector of residuals. And that's sort of expanded out what it looks like.

OK, so we want to choose the beta hat that minimizes that thing. So what we do is we take the derivative with respect to beta, set it equal to 0, and obtain this. If you're not used to doing calculus with vectors and matrices, in the notes that I posted online, I write this out in more detail. And you can take a look at that if you want.

But basically, we get this sort of equation set equal to 0. And this is going to tell us what the beta hat that minimizes the sum of squared residuals is. Then we solve for beta hat. The negative 2 we can just divide both sides by negative 2. And so that goes away. Then we write this equation as x prime y, or x transpose y equals x transpose x times beta hat.

And then if this is invertible-- and remember, that was one of our assumptions. That was our identification assumption, that that thing was invertible. If that's invertible, then we get that beta hat is just equal to x prime x inverse x prime y. Beautiful. I mean, that was literally the derivation of the least squares estimators in matrix notation.

If you look at my notes that I've posted online, I mean, it's just pages of algebra using that summation notation. So this is why we love doing it in matrix notation. What do we want to know about beta hat? What do we always want to know about an estimator, so we can do inference? It's distribution.

Oh, it's right up there. OK, fine. So the expectation of beta hat is equal to beta. Again, in matrix notation, it's very simple. I haven't included the four lines or something like that. But if you treat the x's as fixed, then that makes the sort of proof very simple. They come outside the expectation operator and basically just falls out.

So it is unbiased. And the covariance of beta hat, remember this is the variance-covariance matrix. So it's the matrix that has along the diagonal the variances of each of the beta hats and on the off diagonals, the covariances between them. That is just equal to sigma squared times x prime x inverse. So again, very elegant, very beautiful, and not too hard to show if you treat the x's as fixed. And you can look on the website if you're interested.

And finally, we often don't know what sigma squared is. For inference, we need to know what sigma squared is or we need an estimate for sigma squared. So this is our unbiased estimate for sigma squared.

And as Esther anticipated, here we have to subtract off a k instead of a 2 because instead of it being a bivariate model, it's a multivariate model. And then finally, if we're willing to impose the more strict assumption on the error distribution, the errors are normally distributed, then the beta hats are also normally distributed.

So sometimes we want to impose that. Sometimes we want to be less proscriptive. OK, so now, finally we get to inference. So typically, in the linear model, we're going to want to test hypotheses involving the betas. That's where the real action is. I mean, I can dream up hypotheses involving sigma squared and things like that.

And that's fine. And occasionally, you might want to test a hypothesis involving sigma squared. But really we care about the betas because the betas are the parameters in our conditional mean function of our outcome variable.

And the questions that we usually want to answer using linear regression are about the nature of this conditional mean function. So sometimes we might only be interested in one of the betas. Other times we might want to simultaneously test hypotheses about a whole bunch of them.

And as we saw in the output that I showed you last lecture, statistical packages typically perform some standard tests on the betas for free and just report them with the output. And that's fine. And we can use those. And they're often quite handy.

But there may be other ones that we need to do ourselves. So they don't perform every conceivable test we might be interested in. OK, so let's start with a pretty general framework for testing hypotheses about beta. And it's not only quite general and flexible. It's also super intuitive. It's one of my favorite tests. I really like it.

OK, so let's consider hypotheses of the following form. A matrix r times beta is equal to a vector c. That's the null hypothesis. The alternative is that it's not equal to the vector c. So what is this matrix r? It's a matrix of restrictions. And its dimensions are r by k plus 1.

So it has the number of-- so the number of columns equal to the number of parameters, the number of betas that we're estimating in the linear model. And then the number of rows is the number of restrictions that we want to impose in our null hypothesis, the number of restrictions we want to test.

So we could have a matrix, where r is equal to 1. And then we're just testing one restriction. So that would correspond to something like beta 1 is equal to 0. Oh, so let me just say this. I'll get to some examples in a minute.

So almost any hypothesis involving beta you can dream up in the context of a linear model can be captured in this framework, not quite any, but most of them. You can test whether individual parameters are equal to 0. You can test whether individual parameters are equal to something other than 0.

You can test multiple hypotheses simultaneously. You can test hypotheses about linear combinations of parameters. The world is your oyster. So let me show you a few examples of these and exactly what the r matrix looks like and what the c vector looks like in these examples.

OK, so let's say, for instance, that we set up the matrix r to be just a row vector with a 0 in the first spot, and then a 1, and then the rest 0s. So what that matrix is doing is it's picking out beta 1. Remember, this spot corresponds to beta 0. So being in the second spot, it's picking out beta 1.

And c is just what beta 1 is equal to under the null. So that r and this c corresponds to the hypothesis that beta 1 is equal to 0. Let's suppose instead that we want to test a whole bunch of hypotheses simultaneously, that beta 1 is equal to 0, and beta 2 is equal to 0, and beta 3 is equal to 0, et cetera.

Well, then this is what our matrix would look like. So it would basically be an identity matrix with a column of 0's tacked on the front. And the reason why the column of 0's is tacked on the front is because that corresponds to the intercept. And we're not interested at least here in testing a hypothesis about the intercept.

And then the c vector is just a vector of 0's. So I do want to emphasize, even though I've sort of written this as like one equation, this is actually we're testing k hypotheses simultaneously here. So we have k equal signs.

OK, so here's a more complicated example. If our r matrix has in the first row a 1 and a negative 1, and then the rest 0's, sorry, 0, and then 1, negative 1, the rest 0's. And then the second row there's a 1 in the fourth spot, et cetera. And then the c vector looks like this. What does this correspond to in terms of a hypothesis we might want a test or a series of hypotheses?

Well, here, the first row gives us the hypothesis that beta 1 minus beta 2 is equal to 0. So I could just write that as beta 1 is equal to beta 2. The second row corresponds to beta 3-- this is beta 3 here-- being equal to 5. And the third row corresponds to beta k being equal to negative 2. Yes.

STUDENT: Can you explain the beta 1 equals beta 2?

SARA ELLISON: OK, so if I just multiply the matrix, the r matrix, by beta and sort of wrote these out as equations, I would get beta 1 minus-- so this is beta 0 here. So here's beta 1 minus beta 2 is equal to 0. And then I just rewrote that as beta 1 is equal to beta 2. That's all. Yep.

STUDENT: How often are we [INAUDIBLE] specific value rather than the range? And is there a [INAUDIBLE] against that?

SARA ELLISON: So yes and no. So basically, if we're not interested in-- if we're interested in whether beta is in a range, then what we might want to do is instead of doing a hypothesis test, where the null was a single value and the alternative was everything else, we might want to do, say, a one-sided test, where the null is that beta is less than some value and the alternative is that it's greater than some value.

We can do those. We can't do them in this framework. So I'll talk about that in a second. The other thing that you might be suggesting is instead of doing hypothesis testing, we might want to just report confidence intervals as well.

So remember that really hypothesis testing and constructing confidence intervals are kind of the same thing. It's just reporting the same information in different forms. And so it can just be a matter of style or preference. Instead of reporting hypothesis tests, you report confidence intervals. And that's perfectly fine. Yeah.

ESTHER DUFLO: So confidence interval, it can be harder to say whether between the [INAUDIBLE] minus [? 4 ?] is [INAUDIBLE]. I mean, it's kind of hard to see. They don't really add up.

SARA ELLISON: Yeah, and I guess the other more fundamental answer to your question is that sometimes we actually do-- there might be a theory that says beta should be equal to this number. And in order to test that theory, we want to perform a hypothesis test that beta is equal to that number. So that does come up, not every case. But yeah, it is relevant. Other questions? No.

STUDENT: You could also use it to if somebody came out with a paper today describing the treatment of malaria [INAUDIBLE] wanted to see if that was true or not, just take that beta and test for it and do hypothesis testing?

SARA ELLISON: Yeah.

STUDENT: OK.

SARA ELLISON: OK, oh, here's part of the answer to your question. One thing you can't do in this framework is test one-sided hypotheses. We'll get back to those. So now we have this framework. I mean, it's not really a framework, just sort of a notation in some sense to deal with hypotheses of all of the forms we just talked about.

And within the regression framework, we have a super intuitive and cool way to test these hypotheses. So first of all, let's think of the null as describing a set of restrictions on the model. So let me just go back for a second. So in this case, this null has three different restrictions, that beta 1 is equal to beta 2, that beta 3 is equal to 5, and that beta k is equal to minus 2.

And we think of the null as imposing restrictions on the model. Then here's how we perform the test. We estimate the unrestricted model. We impose the restrictions of the null and estimate that model. And then we compare the goodness of fit of those two models.

So that's why I love this test. It seems really intuitive to me that if you have a set of restrictions and they really bind, and they really sort of affect how good your fit is, then that tells you, well, maybe those restrictions are not true. If the restrictions on the other hand don't really bind that much, if your model fits almost as well with the restricted model as it does with the unrestricted model, then that tells you maybe these restrictions are true or close to true.

And we don't want to reject them. So that's the whole intuition and the idea behind this test. OK, so a couple of details before we get to the distribution, the test statistic. Estimating the unrestricted model is simple. Just run the regression. But how do we estimate the restricted model?

Well, it depends on what form the restrictions take. So let's say we're testing hypothesis, where just a bunch of the betas are equal to 0. How do we run the restricted model? Yes.

STUDENT: We just think the [INAUDIBLE] is kind of on the diagonal 1 so that [INAUDIBLE].

SARA ELLISON: Yes, exactly. So practically speaking, what we do is we just run the regression leaving out all of those x's. So that's the way we constrain the coefficients to be equal to 0. So we have the unrestricted. The unrestricted regression is just all of the x's are in there.

If we want to restrict that certain betas are equal to 0, we just run another regression, where we leave out the x's associated with the betas that we want to have equal to 0. So that's our restricted model. Then let's say the restriction is that the two betas are equal. We have beta 1 equals beta 2, or something like that. That's our null restriction.

Then how do we impose that restriction on a linear model? Well, actually, it might help if I write the linear model. So we have y sub i equals beta 0 plus beta 1 x1. I hope I'm using the same notation. Do I have my subscripts in the same order? I hope so.

OK, so let's suppose this is our unrestricted model. We want to restrict beta 1 to be equal to beta 2. Well, what do we do? We just create a new variable that's the sum of these two. So this is called x1i plus x2i, just a new variable. And then we only estimate one coefficient on that.

So our restricted model is just that we don't include this variable as a regressor or this variable as a regressor. We include their sum as a regressor. And that's how we're imposing the null restriction because when we include their sum, we're making their two coefficients equal. We're forcing their two coefficients to be equal. Yeah.

STUDENT: So are we testing the first beta 1 and the second beta 1 are the same?

SARA ELLISON: Exactly, yes. This is testing the hypothesis that beta 1 is equal to beta 2. Yep.

STUDENT: Is that different from testing the beta 2 [? at 0? ?]

SARA ELLISON: Yeah, it's definitely different. So here, these betas could be anything. They could be a million. We're just testing the hypothesis that they're equal. Yes.

STUDENT: But if they're equal, wouldn't it be like a linear combination of the ther [INAUDIBLE]? You know how they cannot be like a linear sum?

SARA ELLISON: I'm not sure if I understand your question. So basically, what I'm trying to do here is impose just this hypothesis, but not impose anything else about what the betas might be equal to.

STUDENT: The identification restriction is not on the betas. It's on the x's. Beta can have whatever medium conditions.

SARA ELLISON: Ah, you were confused about the identification assumption. Yep, yep, yep, that's right.

ESTHER DUFLO: You can reask your question saying, if you could post the sum on just x and it turned out that in fact they were equal, you can guess what the beta [INAUDIBLE] following x.

SARA ELLISON: OK, what if the restriction is that some beta is equal to a constant c? How would we impose that restriction and then re-estimate the restricted model? Well, let's suppose this is just a constant. So we impose that this one is equal to a constant. So then here, there's no parameter in this term that we need to estimate under the null.

So we just subtract the constant times x1 from the dependent variable and rerun that regression. And that's our restricted regression. Does that make sense? OK, so going back, we estimate the unrestricted model. We impose restrictions and estimate that model. And then we compare the goodness of fit.

And if the goodness of fit is not very different, we don't reject the null. If it's very different, we reject the null. In particular, this test statistic, which is basically the numerator has the difference in the sum of the restricted and unrestricted sum of squares and then in the denominator has the unrestricted sum of squares.

This is how we form the test statistic. And it turns out that has an F distribution under the null. And we reject the null for large values of this test statistic.