[SQUEAKING]

[RUSTLING]

[CLICKING]

SARA ELLISON: OK, so a little bit of a review. I want to put what Esther has been doing the last few lectures into a little bit broader context. And that's going to be a sort of a nice segue into talking about linear regression. So what did we do for the first sort of half of the semester? We established a foundation in probability, and we proceeded to talk about how to estimate unknown parameters, OK.

Most, if not all of that discussion, was focused on estimating parameters of a univariate distribution, OK. So we were talking about estimating the mean or estimating the variance or something like that. Some other parameter that characterizes sort of a univariate distribution.

But so much of what we care about in social science and in lots of other settings, I mean, maybe most other settings, involves joint distributions, though. And so we really have to be concerned with how to estimate parameters characterizing either joint distributions or conditional distributions or something like that, OK.

So Esther's discussion of causality was really the beginning of this. And it was also we could think of it as a special case, OK. So you can think of much of what she did as considering the joint distribution of two variables where one was simply a coin flip. So if you flip a coin, and it comes up heads, that's a treatment, that observation has a treatment. And if it comes up tails, that observation is assigned to control.

And then the other random variable that was defined on this sample, was the outcome of interest, say, infant mortality or website effectiveness or something like that, OK. So we didn't always talk about-- oh, and I should say, in fact, we were mostly concerned with the conditional distribution of the outcome variable conditional on this coin flip, OK.

So we can and we did think of the treatment and the control group sometimes as being sort of two separate populations. And so then we asked the question, well, how could we test whether they had a common mean? But we can also think of them as having one population and a joint distribution of two random variables on that population. And what we were really doing is estimating conditional means, conditional on this coin flip random variable.

So these are two different ways to think about randomized controlled trials and the things that Esther was talking about the last couple of lectures. OK, and they're sort of equivalent ways to think about them.

Now, I want to pose a question. What if, instead of a coin flip, the second random variable is continuous instead? So let's suppose we have a random variable. It can take on a whole range of values. We can still think of it as being sort of levels of treatment or something like that. We don't necessarily have to think of it that way, but you can think of it that way.

But instead of just being a coin flip treatment control, it can take on many different values. How do we analyze the conditional distribution of our outcome variable conditional on something like a continuous random variable? And furthermore, how do we estimate the parameters of that conditional distribution, and that's what we're going to be talking about today.

The workhorse model we use is called the linear model. And the way that we estimate the parameters in a linear model is linear regression, or that's typically the way that we estimate them. So why do we care about joint distributions in estimating the parameters associated with them? Well, we've seen a bunch of examples, but let me just throw out a couple more.

So we might be interested in prediction. We might be interested in determining causality, and we might just be interested in understanding the world better. So let me give you examples of each of these. OK, so let's imagine that I'm the type of person who reads XKCD comics. And I'm also the type of person who is likely to click on an ad for a t-shirt bearing the Russian cover design of *Moby Dick.*

So who might be interested in whether that statement is true?

AUDIENCE: Amazon.

SARA ELLISON: Amazon or whoever's selling the t-shirts with the cover of the Russian*Moby Dick* on them. So basically anyone who wants to sell me that product or anyone who wants to serve me an ad that might get me to click on the ad and buy the product, cares about whether the person who reads XKCD comics is also the kind of person who buys such t-shirts.

It turns out I am such a person. I do both of those things. And I thought I'll just show you my favorite XKCD comic, and let you read it for a second. And I did just order this t-shirt from Out of Print, which I think is a pretty nice looking t-shirt. That's for all you course six majors, I guess.

OK, can I go ahead. OK, so anyhow, there's nothing about causality in this example. I didn't read the XKCD comic and think, oh, now, I have to go by this sort of t-shirt. It's just that I'm the kind of person who might do both of those things. So someone wanting to sell me that t-shirt, would like to know that, so they can serve me the ad when I'm reading XKCD.

We might want to determine causality. Esther talked about a number of examples in this vein. So let's suppose I want to know the answer to the following question. If I give my dog a treat every time he does not bark at another dog walking by our house, will he stop barking at other dogs?

So that's the kind of causal question I might want a answer to. The answer, by the way, is no. And the only reason I brought up this example is so I would have an excuse to show you my sweet little dog. But I have not carried out such a statistical analysis. I'm pretty sure the answer is no, though.

And then maybe we just care about joint distributions because we want to understand the world better. So if we were behavioral economists, we might be interested in the following question. Are people only influenced by price, quality, characteristics, and the expected weather that they have in the area of the country where they live when they're deciding whether to purchase a convertible? Or are they also influenced by the weather on that particular day?

So behavioral economists might be interested in a question like this because they want to understand what influences economic decisions that agents make. Do they make rational decisions, or are they deciding to buy a convertible because they know what the weather is going to be like over the next five years, and that's the period of time, they're going to own the car? Or are they likely to make these sort of decisions that are hard to justify by any rational economic model, like, gosh, it's really sunny today. Maybe I'll buy a convertible that I'm going to now have for the next five years?

Well, it turns out that they are. And this was a recent paper in the*QJE,* "The Psychological Effect of Weather and Car Purchases" that fairly convincingly established that people are, in fact, influenced by the weather on the day that they make a purchase. When the weather's bad, snowy they're more likely to make a purchase of an SUV. And this is controlling for the expected weather going forward. And likewise, if it's sunny, they're more likely to buy a convertible controlling for the expected weather.

So anyhow, I went to a talk. I saw this paper a few years ago. I was convinced by the results, but I also sort of chuckled. I thought, oh, that's so ridiculous. How could someone be so influenced?

Well, turns out that last summer, I decided I was going to buy this car.

[LAUGHTER]

And so I was waiting for the 2016 model to come out. And it was a really hot model of car. I couldn't even test drive it. It was like really hard to-- and I thought, OK, I thought about the *QJE* paper, and I said, all I have to do is wait until the weather gets bad, and then the models of this car will build up on the dealer's lots. And I'll just waltz in and be able to buy one, and maybe I'll even get a deal on it.

Well, it turns out, as soon as the weather got bad, I kind of lost my interest in buying the convertible, and I just ended up buying a sedan, so.

[LAUGHTER]

OK, so in each of these examples, we had two or more random variables that had a joint distribution. And we wanted to know the characteristics of their joint distribution in order to answer the questions that we were interested in. OK, so how do we do this?

Well, I told you already, the workhorse model to do this is the linear model. So let me introduce the linear model, bivariate style. We'll generalize that later. And let me just go through and define what the things in this linear model are. And then we'll talk about properties of it, and we'll talk about how to estimate the parameters in it and so forth.

OK, well, the linear model is a model that establishes a relationship between two random variables on which we have repeated observations. So these are y and x. So y is what we call the dependent variable. Sometimes you will hear that referred to as the explained variable or the regressand. I would say hardly ever do I hear those terms, but those are both legitimate terms, I guess, for the dependent variable. And then the x variable is typically called the regressor or explanatory variable, or sometimes called the independent variable.

So even though regressor and regressand seem to go together, or independent variable and dependent variable seem to go together, that's not the way these terms are used. Typically, we refer to the y variable as the dependent variable, and typically, you'll refer to the x variable as either the regressor or the explanatory variable.

And there are some reasons for that. But this is just basically what people do, just so you're familiar with the terms. And then we have a third random variable thrown into the mix, an unobserved random variable, and we call that the error.

We also have two parameters in this model, or at least two obvious parameters. We have another one that we'll talk about later. And these are parameters we want to estimate, and we call these the regression coefficients.

OK, so what does this model do? It allows us to consider the mean of a random variable y as a function of another random variable x And if we obtain estimates for beta 0 and beta 1, then we have an estimated conditional mean function for y. So let's add some basic assumptions, and then we will get to something called the classical linear-regression model when we sort of pile on a few more assumptions.

OK, first of all, let's assume that our two random variables on the right side are uncorrelated, so x sub i and epsilon sub i are uncorrelated random variables. Let's assume what this second assumption called identification means is just basically, that we have some variation in our x variable. I'm going to show you a picture illustrating this in a second. But basically, that just means that the repeated observations on the x variable are not all on the same point.

We're going to assume that the expectation of our error is equal to 0. We'll assume a homoskedasticity, which means that the expectation of our error squared is equal to sigma squared for all i. And we're going to also assume that we do not have serial correlation. So in particular, the expectation of epsilon i epsilon j is equal to 0 for all i and j.

So let me talk about each one of these in turn and show you some pictures, so you understand exactly what the meat of these assumptions is. Oh, a couple notes, we sometimes impose an alternate assumption to 1 for our convenience. And that's sometimes we assume that instead of x being a random variable, xi being a random variable, we assume that xi's are fixed and repeated samples or nonstochastic.

This makes some of the proofs easier. Sometimes we do it for convenience. The results essentially go through without that assumption. But we'll go back and forth between them. And then the other note is that assumptions-- we had the three different assumptions on the error term. And all of those assumptions could be subsumed under one stronger assumption that the epsilons are iid normal 0 sigma squared.

So sometimes, especially when we talk about inference, we'll impose that assumption. Again, that assumption is not necessary for almost anything that we do in linear regression, but sometimes it's sort of convenient. It makes the proofs easier and more elegant, I guess.

So let's look at this assumption number two called identification. So I have this kind of strange looking equation up at the top, but all that means is that we're ruling out a case where all of our observations on x are on a particular value x0.

And the reason we rule out this case is, because this doesn't give us any variation in x that we need to identify the mean of y as a function of x. If we're trying to fit a sort of a regression line, we just have a bunch of observations on a particular point, and we can't do it.

The third assumption is that the error has a 0 mean. Well, so basically, what we rule out-- so here I've drawn a picture of an error having a positive mean. And we just rule this out because we don't have any information that would help us separately sort out whether our error just had a 0 mean, and the regression line was shifted up, or rather the error had a positive mean, and our regression line was shifted down. We don't have any information to separately to sort those two possibilities out. So we just assume that the error has 0 expectation.

And I should say, well, another way to say this is that those two-- the expectation of the error and the intercept of the regression line are not separately identified. That's another piece of terminology. That just means we don't have any way to them out with the data.

Our fourth assumption is homoscedasticity. So what is homoscedasticity? It just means that we're going to assume that the variance for all of our errors that the variances is equal to sigma squared for all-- oh, that's supposed to be all little i, but my autocorrect made it capital I, I think.

So this is a picture of what the opposite of homoskedasticity might look like. This is called heteroskedasticity. And here I've drawn it so that it looks like the error variance for low values of x is much higher than the error variance for higher values of x. And so we're going to rule that out for now. That's actually not a fatal problem. We can deal with it, but at least for now, we're going to assume that we have homoskedasticity and not heteroskedasticity.

So right about now, you're probably thinking, what is the etymology of homo and heteroskedasticity and is she even spelling it right? Because in fact, my autocorrect kept trying to replace k with c whenever I was typing homoskedasticity. Turns out that the autocorrect in PowerPoint has not read this article in *Econometrica* on the etymology and the correct spelling of homoskedasticity and heteroskedasticity. So anyhow, I've put this up on the website. You can peruse this at your leisure. But this will answer all your questions about the origin of these words.

And then the fifth assumption is that we have no serial correlation. So that means, in other words, that the errors are not correlated if they're associated with different observations. So a stronger way of saying this, is that the errors are independent across observations. We don't need to impose that. We're just going to say that they're uncorrelated.

Here is a picture. I drew a picture of what positive serial correlation might look like. So here we have some positive errors over on the left side. And then if you have one positive error, the sort of close-by errors are also likely to be positive.

So you have sort of an area where errors are mostly positive, and then you get a negative error. And then most of the errors close by are likely to be negative. So this is what positive serial correlation might look like. The errors are kind of clumped spatially like this. Again, this is not fatal. We can deal with this, but we're going to assume, for now, that we don't have it.

And then as I said, assumptions three through five could be subsumed under a stronger assumption that epsilon i are iid normal 0 sigma squared. So we don't typically need this assumption. Sometimes we impose it. That just means that basically, the error just has a little normal distribution that looks like that.

So what are the things that we can say about this model first of all? What are some of the properties? Well, we can calculate what the expectation of y is. And I should say we can think of this-- so right now, during this calculation, I'm making the assumption, just for my convenience, that the x's are nonstochastic.

And so I haven't said that the expectation of y is conditional on x, but we can think of this as basically the conditional distribution of y given x. And so, basically, if I'm assuming that the x's are nonstochastic, it makes my life a little bit easier because I take the expectation of beta 0 plus beta 1 x sub i plus epsilon sub i, and the x's are nonstochastic, so they're just constants. They come outside the expectation, and it just makes the calculations easy. And so we get, in particular, that the expectation of y sub i is equal to beta 0 plus beta 1 x sub i.

We can do a similar calculation with variance of y sub i, and we get that that is just equal to sigma squared. So it's the same variance as the error. And then we can also do a similar calculation showing that for any i and j, where i is not equal to j, the covariance between y sub i and y sub j is equal to zero. Yeah, oh, and so I just want to emphasize, the betas that are parameters in the conditional mean function.

So we've got this linear function. We've got a set of assumptions that we're willing to impose for now. So our next question is, how do we find estimates for beta 0 and beta 1? Well, I'm just going to throw up three different options. And I'll very quickly dismiss two of them but just so you know that there are other things that you can consider other than least squares estimates.

So one thing that we could do is, we could find the least squares estimates. In other words, the values of beta 0 hat and beta 1 hat, such that this expression here is minimized. And I'll show you pictures of what these look like in a second.

We could, instead of minimizing the sum of squared errors, which is what-- so each one of these in the parentheses is a-- well, it's an error if beta 0 and beta 1 are the true parameters. It's a residual if they're the estimates. But what we do is, if we plug-in the estimates, we call that a residual. And then if we square the residuals and sum them up and find the parameters that minimize that sum of squares, that's called a least squares estimator.

We, instead of squaring them, we could just take their absolute values instead and come up with something called the least absolute deviations estimator. That's the second possibility. Or a third possibility I'll throw out, is something called the reverse least-squares estimator. Let me just show you pictures of each of these in turn.

So actually, the least-squares estimator and the least-absolute deviations estimator have the same picture because what we're doing, is for the least-squares estimators, we're just taking this quantity here and squaring it. And for the least-absolute deviations, we're not squaring it. We're just taking the absolute value.

So they have the same pictures. But what we do is, for the least-squares estimator, we compute all of these deviations. And we choose the line that minimizes those squared deviations. Yes?

AUDIENCE: So, I guess, the end is how did these two different [INAUDIBLE].

SARA ELLISON: So I'm going to talk just a little bit, I'm going to have a slide, maybe two slides or something, about the properties of least squares and why. We almost always use least squares, instead of least-absolute deviations. I'm not going to actually say too much about them. But when we get to that slide, if you have any other questions, let me know.

AUDIENCE: OK.

SARA ELLISON: So these are the first two possibilities. And then the picture for the third possibility just looks like this. We're minimizing the sum of squares of those deviations instead, in a reverse least-squares estimator. So why do we always just focus on least-squares estimators? Why is that sort of what we always do?

Well, under the assumptions of the classical linear-regression model that I had up a few minutes ago, OLS, which stands for Ordinary Least Squares, provides the minimum variance, in other words, the most efficient, unbiased estimator of beta 0 and beta 1. It is also the maximum likelihood estimate under normality of errors. And the estimates are consistent and asymptotically normal. And these things are not true of the other estimators. So that's why we typically always use the least-squares estimators.

Are there cases when you might want to use a least-absolute deviations estimator? Maybe, I mean maybe if you're particularly worried about the credibility of your data, and you think you might have some outliers that you don't want to take out of your data set, but you don't want them to have undue influence on your parameter estimates. You might be able to justify using a least-absolute deviations estimator.

But typically, people are going to go for the least-squares estimators because its properties are so good. Yeah, oh, sorry, yeah, go ahead.

AUDIENCE: And so is the reason that that OLS is favorable, even though they're both based on the same distance, I guess, the sensitivity in the OLS is higher because it's just squared?

SARA ELLISON: Yeah, it's more sensitive to observations that are kind of out in the tails because it's squared. But even given that, it still has all of these favorable properties. So you have a question?

AUDIENCE: I guess not really anymore. I was sort of thinking the first characteristic is a little bit circular because we define efficiency in terms of variance. So if you had a [INAUDIBLE] permutations here [INAUDIBLE] necessary [INAUDIBLE]. I guess what you're really saying, is that it provides the minimal barrier.

SARA ELLISON: Exactly, yeah, so I just put in parentheses "most efficient" to just remind you that that's how we defined-- yeah. That's the terminology we used basically, but yeah. Yep.

AUDIENCE: Are theses properties more or less-- homoskedasticity goes out the window?

SARA ELLISON: So good question. I would have to think for just a second. So basically, it's no longer the most efficient. It's still unbiased. It's no longer the most efficient. The inference has to be-- under heteroskedasticity, the sort of standard inference that we'll see over the next couple of lectures, has to be modified because then the standard errors, if you use the standard inference, are biased. But it's still going to have good properties. And we have very good ways of fixing, of recapturing the best properties under the case of heteroskedasticity.

OK, so fine, maybe, I hope, I've convinced you that least-squares estimators are the way to go and the way that we're typically going to estimate this linear model. So we've got all of these sum of squared residuals. And we're trying to choose the line that's minimizing those.

Does that mean we have to do a numerical minimization every time we want to solve for our least-squares estimators? No, we don't. We have these lovely closed form solutions. So all we have to do is plug into these formulas and realistically speaking, all we have to do is type the command into R.

So anyhow, if you went on to do more advanced econometrics, you would, in fact, encounter estimators where you have to do complicated numerical minimizations every time to get parameter estimates. OLS is not one of them. We have these nice closed-form solutions. How did I get these? Pages of tedious calculations that are up on the website for your viewing pleasure.

So I have a couple things to say about this. First of all, I can't remember what I put it under. I think I put in under derivation. There's a page of notes under-- what's it under? Resources, maybe, and it's called derivation of OLS estimators. And so you're welcome to look on those notes and see the calculations whereby I arrived at the least-squares estimators.

The notation is slightly different, I believe, in those notes. Instead of having the intercept denoted by beta 0, I think I used an alpha instead. And then a beta for the slope. And so the notation is slightly different, but anyhow, you can look at those if you want.

However, I don't want you to get the idea that OLS estimators are horrible complicated things. They're really very elegant, and they're very intuitive. But the thing that's holding us back is this notation.

So Ashley, next time, we will derive the least-squares estimators using matrix notation, and you will see how beautiful they are. I'm not going to drag you through the derivation using the summation notation because it's just it's painful. Yeah.

AUDIENCE: When we go 1 over nc epsilon, why do you have [INAUDIBLE].

SARA ELLISON: Doo, doo, doo, doo, doo, oh, they would-- I think it's just because-- oh, yeah. I think it's just because the-- well, so that thing on the bottom, is like a sample variance. And so you might compute that separately, and then just-- yeah.

And as I just said, they could be lovely or still if we weren't too afraid of using matrix notation, which we're not. So we'll do it next time. A couple of important definitions. So I already told you just sort of verbally, what a residual is. But here's a picture of residual, and you have the formula here.

So residual is basically, the deviation between an ordered pair of a particular x and y and the fitted regression line. And we denote that as epsilon sub i hat, typically. Sometimes we'll see other notation, as well.

Here, I've drawn in the regression line, also known as the fitted line. So that's just beta 0 hat plus beta 1 hat times x. And then it's also convenient to define something called a fitted value. And that's basically, just for any particular value of x, the value of y on the fitted line associated with that value of x on the estimated regression line. So we'll come back. We'll use this notation, the y sub i hat and the epsilon sub i hat notation, and we'll come back to these quantities in the next couple of lectures.

What do we always ask when we learn about a new estimator? And why do we ask it? What do we always want to know about an estimator?

AUDIENCE: [INAUDIBLE]

SARA ELLISON: Hmm?

AUDIENCE: It's [INAUDIBLE].

SARA ELLISON: Well, that's a specific case. So we want to know its distribution. We always want to know its distribution. And I already told you these are unbiased estimators, but we want to know its distribution. And why do we want to know its distribution? Because we're not going to be able to perform inference. We're not going to be able to create a hypothesis test or a confidence interval if we don't know the distribution, or we don't know, at least, the variance of our estimators.

So let me just define a couple of things. Let's let x bar just equal what it typically is equal, just 1 over n times the sum of the xi's. And then let's define this thing called sigma squared hat sub x, which is 1 over n times the sum over i of xi minus x bar squared.

And if we use that notation, then we can figure out what the mean and the variance and the covariance of beta hat, beta 0 hat, and beta 1 hat are. And again, how did I get these? Pages of tedious calculations up on the website for your viewing pleasure. But you can skip them if you want because as I said, we'll see a much nicer way to come up with these formulas next time.

And then how did I get the mean actually? I think the calculation of the mean of these is also up on the website. But in fact, I didn't. I already told you they were unbiased estimators. So we already knew that the mean of beta hat was equal to-- beta hat 0 was equal to beta 0 and same thing with beta 1.

So let me take a moment to talk about a little bit of comparative statics. I think that this discussion will give you a little bit more of a feel for the mechanics of linear regression and how this thing works. So I list them here, and I'll go through each one separately.

If we have a larger sigma squared, so imagine having two different data sets, they're identical except for one of them has a larger error variance, how is that going to affect our estimates? Well, you can just look at the formulas for the variance and the covariance and so forth of beta of our estimates, and you can see that if we have a larger sigma squared, that's going to mean that the variance of beta hat is larger. I'll show you a picture of that in a second.

If we have a larger variance of x, so in other words, our x's are more spread out, that's going to lead to a smaller variance in beta hat, and a larger n also means a smaller variance of beta hat. And then furthermore, if the mean of x is positive, then we're going to have a negative covariance, a negative relationship between beta 0 hat and beta 1 hat.

And I didn't write it down here, but if x bar is negative, then that flips. So let me show you pictures of these, and that will make the comparative statics a little clearer, I think. Oh, and I should say, oh, just a point of notation, I'm often going to just start sliding into matrix and vector notation. And when I say beta hat, I'm talking about the vector beta 0 hat and beta 1 hat.

So what's the first comparative static? What does that suggest about the mechanics of regression? So we have two identical data sets. The only difference is that the error variance is greater than 1. What does that mean? That in this case, we're going to be less sure of our estimates. The variance of beta hat is greater. We're just going to have less confidence in our estimates. That makes sense, right. If we have very small error variances, we're going to be able to estimate that linear relationship very precisely.

How about if we have the same data sets, but in this case, the x's on the left are very spread out, we have a lot of variation in the x's and in the other case, the x's are all smushed together into one little area? Well, in the case where they're all smushed together, remember, in the limit, we can't even estimate our linear regression coefficients.

Remember the picture I showed you in the beginning where all of the observations on x were on a single point. We couldn't even estimate the linear regression coefficients there. Here, we can estimate them, but the variance is going to be pretty big. We're going to be less sure of our estimates in this case because we don't have a lot of variation of x to help us identify the effect we're interested in. A larger n means a smaller variance of beta hat. Well, I don't need to draw a picture of this. We'll just note that this follows from the fact that beta hat are consistent estimators.

And then finally, there is this mechanical relationship between the two estimates. So if we have for the mean of our x's is positive, then basically, if we overestimate the intercept of the line, we're going to underestimate the slope, or an overestimate of the intercept, is associated with an underestimate of the slope. An underestimate of the intercept, is associated with an overestimate of the slope. So it's just this mechanical relationship. And then if x bar is negative, then the mechanical relationship shifts, it'll flips.

So one step further, if we use the stronger assumption that the errors are iid normal 0 sigma squared, then we obtain-- in addition, so we already have the mean and the variance and the covariance of these two estimates. But we also have that they have normal distributions if the errors have normal distributions. So if we're willing or if we feel like imposing that extra assumption of normality of errors, then what we get out of it is normality of our estimators.

One loose end, note that the distributions of beta 0 hat and beta 1 hat are functions of sigma squared. That is the error variance. We often don't know the error variance. Why would we know the error variance? So we estimate it.

And we're going to use a different estimator than estimators we've seen before. So we're going to use an estimator that's equal to 1 over n minus 2 times the sum of the residual squared. And this estimator, it turns out, is unbiased for sigma squared in the linear model. That's why we use it.

Why the minus 2 in the denominator? Well, it's because we're estimating two parameters. Do you remember our unbiased estimator for the variance when we were just estimating the mean had an n minus 1 in the denominator? Well here we need the n minus 2 because we're estimating two parameters, a beta 0 and a beta 1. And it turns out that that's what we need for sigma squared to be unbiased. Yep.

AUDIENCE: Can you remind us why that is, or why we both need it?

SARA ELLISON: So I didn't do the proof. And to be honest, I don't feel like I have a useful intuition to tell you. I'm happy to give you the proof, and the proof of this isn't much more complicated than in the univariate case. But yeah, it just turns out that that's what you need to do to have an unbiased estimator of the variance.

Do you have a useful intuition? I don't know. No, OK.

AUDIENCE: That applies if you had to twist me three things.

[LAUGHTER]

SARA ELLISON: Yes, yes, exactly.

AUDIENCE: I don't know why, but well--

SARA ELLISON: Mm-hmm, it does. So now I want to ask you to think back to when we were doing univariate inference, and we replaced an unknown variance with an estimate of the variance. So we had a sample mean, and the sample mean had a normal distribution, but we didn't know the variance-- sorry, the standardized sample mean had a standard normal distribution. But we actually didn't know what the variance was, and we replaced an estimate for the variance.

What happened? You guys remember? That's what happened, T distribution. Well, same thing is going to happen here. I won't talk about t-tests in the context of linear regression today. I should be able to get to them next time. But basically, just to give you a little bit of a preview, the same thing is going to happen here.

We have, under the assumption of normal errors, we have that beta 0 hat and beta 1 hat have normal distributions, but we don't know what their variance is. We're going to have to use an estimated variance instead, typically. And so that's where the t-test in the linear regression context comes from.

So now we have most of the pieces the model, the estimators, information about the distribution of the estimators. We could proceed with inference. But I'm going to put that off for a little while. And what we're going to do now is take a quick detour into something called analysis of variance. And then I'm also going to talk about some sort of practical issues in estimating linear regressions before we get to inference. We'll do inference next time.

So we want some way, typically, to indicate how closely associated x and y are. Or how much of y's variation is explained by x's variation. And what we do is we perform an analysis of variance that I'll do in just a second. And that's going to lead us to a measure of goodness of fit for our linear regression. And that's, typically, a useful gauge of how the regression is doing in some sense.

So let's start by defining the sum of squared residuals. So we've actually already seen the sum of squared residuals. That's what we are minimizing choosing beta 0 hat and beta 1 hat to minimize. Let's just define this quantity here as the sum of squared residuals.

And here's a picture of it. So the picture can't indicate-- I mean, each one of these-- so what we do is, we calculate each one of these, we square it, and we add them together. That's the sum of squared residuals. And this is, in some sense, a measure of goodness of fit.

But it's not a very useful measure of goodness of fit because it's not unit free, which is inconvenient. So let's say we have a data set and both of the variables are measured in pounds, British pounds. And then we translate both of the variables to dollars. Then the sum of squared residuals will actually change. And that's not a very convenient feature for a measure of fit to have.

But if we divide the sum of squared residuals by some other sum of squares, the total sum of squares, for instance, then that gives us a unit-free measure because the units cancel. So what's the total sum of squares? Well, the total sum of squares is we don't even fit a regression line. We just take the sample mean of y and just calculate all of the deviations from the sample mean and square those up.

So graphically, what does that look like? So let's say this line is y bar. So what we've done is we've just taken all of the deviations from y bar, squared them, and sum them up. That's the total sum of squares.

So we've got the residual sum of squares, and we've got the total sum of squares. And note that that is going to be a unit-free measure of goodness of fit that's always going to be between 0 and 1. Why is it always between 0 and 1? Yep.

AUDIENCE: [INAUDIBLE] these are always-- right, so I guess the question is, actually, why is the SSR always less than SSD? Because they're both positive.

SARA ELLISON: Yeah.

AUDIENCE: Yeah, OK.

SARA ELLISON: They're both positive, yes. So you've got half of it right. OK, Yep.

AUDIENCE: Is this like it's the maximum error you can have?

SARA ELLISON: No, well, no, I could draw a line that gave me more error. But basically, you're on the right track. The SSR is the least error you can have. Why is that? That's how we chose the regression line. To minimize that quantity.

So any other line, it's going to give you higher sum-of-squared errors. Anything else you do is going to give a higher sum-of-squared errors than the SSR. And so yeah, so you got half of it. They're both non-negative by construction. And the fact that the regression line is the least-squares line ensures that SSR is always less than or equal to SST.

So that could be our goodness-of-fit measure. Actually, I guess we wanted a goodness-of-fit measure that had larger values when the fit was better or explained more. So instead what we decided is we define our goodness-of-fit measure r squared to be equal to 1 minus SSR over SST. It's an arbitrary decision, but that's what it is.

And it turns out that the total sum of squares can be decomposed into two terms, the residual sum of squares and something called the model sum of squares. And so then r squared, it could also be written as just the model sum of squares over the total sum of squares. Here's the model sum of squares, and here's a picture of it, just the fitted value, the difference between the fitted value and y bar. Oh, that's just something about the decomposition.

So we've got this thing called r squared, and in a bivariate regression, so in the regressions we've seen so far, they just have one explanatory variable, this r squared is actually equal to the sample correlation coefficient for x and y. So we have, in fact, seen this measure before. Why do we feel like we have to define this new goodness-of-fit measure for linear regression? Well, it is in fact, a more general formulation, and it's defined for linear models with more than one explanatory variable.

So when we get to multiple regression, we can have the same. We can define the r squared exactly the same way as we just defined it. And in addition to using r squared as a basic measure of goodness of fit of our regression, we can also use it as the basis of a test of the hypothesis that our slope is equal to zero.

And when we get to multiple regression, it's going to be a test of the hypothesis that beta 1-- that's sort of the slopes on all of our explanatory variables are equal to zero if we have k explanatory variables. And so what exactly is this test?

Well, we create this test statistic, which is just equal to n minus 2 times r squared over 1 minus r squared. And it turns out that this thing, under assumption that the errors are normally distributed, has an f distribution under the null. And it's going to be large. So we're going to reject the null if it's large.

So why does that make sense? Is anyone willing to take a stab at explaining the intuition behind this? No, OK. [LAUGHS] I'll take a stab at it then. So basically, the idea is if our r squared is large, that means that the variation in x is explaining a lot of the variation in y. We have a high measure for our goodness of fit.

And so this thing is going to be large. This quantity is going to be large. And so we want to reject the null that the coefficient on x, this beta 1 is equal to zero. If the x does a good job of explaining-- the variation in x does a good job of explaining the variation we're seeing in y, then beta 1 isn't equal to 0, or with high probability, we can reject the hypothesis that beta 1 is equal to 0 in that case.

And so then we want to-- when this is large, that's when we want to reject this hypothesis. Does that make sense? Somewhat? We'll see the f test again.

AUDIENCE: If k represents number of variables or observations.

SARA ELLISON: Sorry.

AUDIENCE: And [INAUDIBLE] represents number of variables.

SARA ELLISON: No, number of observations, number of observations, yeah. So let's talk about a few practical issues, then we'll introduce multiple regression, although, I don't think we'll get to that this time. And then we'll return to inference after that.

So what does regression output look like, and how do we interpret it? So first, let me show you some regression output from Stata. So I know that you guys are using R, and I'll show you some R output in a second. But most of the regression output I have sitting on my computer happens to be in Stata, and it's probably good for you to see the elements in different statistical packages and be able to figure out what all of these numbers mean.

So here in this data output, we have beta 1 hat is just listed under this column that says coefficient. And then beta 0 hat is at the bottom. Stata always puts the estimated constant beta 0 hat at the bottom.

Right after the estimated coefficients for beta 1 and beta 0, we have the standard errors listed. So those are the estimated standard deviations of the distributions of those estimators.

AUDIENCE: Where [INAUDIBLE]?

SARA ELLISON: So I was the one who gave lhd3rev its name. And that's just a variable. I could tell you what's in this regression. This is basically a regression taken from one of my papers where we were trying to explain the ratio of detail, advertising, and pharmaceuticals to sales trying to explain that ratio with characteristics of the market, like revenue. So that's one of our measures of revenue.

So that's a crazy name for a variable. We called it something different in the paper. But in my sort of crazy state of code, that's what it ended up being. The _cons, that's a name given by Stata. So that's basically just saying, that's the estimator for the constant, or the intercept in the regression.

So the standard errors are listed there. And then we'll get to the-- I'll point out the t-test in a second. But here are the results for the f test that I mentioned. Let me just say something. So typically, statistical packages, when they give you the output for a regression, give you a bunch of stuff for free. And one of the things they give you for free is this standard f test I just mentioned.

So they'll just run it for you for free, and they'll report the results. And here are the results. Let me take this question, and then I'll explain it. Yeah.

AUDIENCE: So when you're describing [INAUDIBLE], what's the convention for how much of this you share? What if--

SARA ELLISON: Oh, I'm so glad you asked, and I will talk about that next time. So I'll have examples of tables that I want your-- well, they're examples of tables from my papers. But they're also examples of tables that I want the tables in your empirical project to look like. And yeah, it will be clear next time so good question.

So here are the results of the standard f test. And what this particular f test says, remember, we want to reject for large values of the test statistic. We want to reject the null that beta 1 is equal to zero if we get large values of that f test. So they perform the f test, and they also gave us the p value.

So last time, Esther defined with a p value is that's basically the probability that-- I hope I don't screw up this verbal definition. It's the probability that it's the point at which if you were doing a size p test, you would reject the null.

So what size test do we usually do? We do 5% tests. We do 1% tests, things like that. Here, we have to do a test as large as basically an 18% test in order to reject the null. We're never going to do an 18% test, so we don't reject the null here.

So basically, if you looked up 1.9, which is the value of our f statistic for this particular regression, if you looked up the critical values in an f table, you would conclude that you don't want to reject the null under this case. And that's exactly what the results here tell you. They say, unless you're choosing a ridiculously high alpha for your test, you don't want to reject the null.

How about this one, though? For this one, we would reject the null that beta 1 is equal to zero if we were doing a 5% test. If we were doing a 1% test, we would not reject the null. So basically, what this f test is telling us, is that if we want to do a three-- this is the p value. The p value tells us the boundary of the size of the test at which we would reject. Is that clear?

And here-- we also get this for free every time we run a regression --they perform a t-test for basically, the hypothesis that that particular coefficient is equal to zero. And we'll talk about-- I'll make more precise next time exactly what these tests are. But this is just a t-test for each coefficient being equal to zero. And in this case, we don't reject the null for either one. I guess, down here, we reject the null for this one at 5%.

So here are some R output courtesy of Angela, who was up late last night running some regressions. So it looks a little different. It has most of the same elements in it. So here, instead of calling it the constant, they call it the intercept. And they put it first, instead of second.

This is the variable name that Angela came up with. So again, that's sort of your choice when you're programming. You call your variables what you want. And you also get the standard errors, and you get the t-tests for free. The f test here, instead of being listed in the upper-right corner, like it is in this data output, it's listed at the bottom, but it's the same information. It's basically gives you the value of the f statistic, and then tells you the p value at which-- the size of the test at which you would reject the null. And here, this is a very small p value, so any traditional test, you would reject the null that beta 1 is equal to 0 here.

And then you can even-- well, let me just go back for a second and note that the r- squared of this regression is 0.41. And if you're wondering what a 0.41 r squared might look like in a scatter plot, so this is a scatter plot of the variables and the fitted-regression line through them.

And if you're curious about the particular variables, this is just these are variables from the General Social Survey about attitudes on abortion and how they have been evolving over the past couple of decades in the US. OK, questions?

So how do we interpret these parameter estimates, and in particular, how do we interpret the estimated coefficient on our x variable, this sort of beta 1 hat? Well, beta 1 hat is the estimated effect of y of a one unit increase in x. So the precise nuances of the interpretation, are going to depend a little bit on whether we think that we've estimated a causal relationship or whether we think we've estimated something else.

So the language you might use could be a little bit different, but this is the basic idea. This is how we interpret what beta 1 hat means. So let me show you-- these are some data that I downloaded from, let's see, the ESPN website. I did this several years ago. So this is data involving the 2005 Major League Baseball season.

And what I did is I downloaded the number of wins of each team and the attendance, the complete season-long attendance for each team. And wanted to see if there was some relationship between attendance and number of wins.

So what did I do? I regressed attendance, that's my dependent variable, my y variable, on the number of wins that each team had. And we see here based on this regression, that one additional win is associated-- oh, and I should say, the attendance was in thousands. That was the unit I used. So one additional win is associated with an additional 31,000 fans in attendance over the course of the season.

So we might want to exercise a little caution to think that this is a causal relationship. We'll get to that more later in the semester. But at least, sort of this regression suggests that an additional win is associated with 31,000 fans in attendance, an additional 31,000 fans in attendance.

Another practical issue, what if x only takes on two values, 0 and 1? Well, we have a special name for that type of variable. We call it a dummy variable, or sometimes we call it an indicator variable. And there's no problem at all in that. There's nothing in our assumptions that rule this out. So that is not a problem for estimating a linear regression. And in fact, dummy variables can be quite useful.

The pictures will look a little different. I'll show you a picture in a second. Here's what I mean. If our x variable only takes on values 0 and 1, then we don't have this nice cloud of data. We just have these two. But nothing wrong with that.

And then we estimate then, sort of beta 0 hat is interpreted as the mean of the group of observations that have the dummy variable equaling 0. And beta 1 hat is the increment in the mean or the effect of having the dummy variable equal 1.

So dummy variables serve a number of important roles in linear models. And we've actually seen one already, randomized controlled trials. So when Esther was talking about having two groups, a treatment and control group, we could just simply have done the whole analysis in the linear regression framework assigning a 0 to one of the groups and a 1 to the other group.

So suppose we have some treatment whose effect we're interested. We randomly assign a treatment to half of the observations. Leave the other half untreated. We assign the treated observations x equals 1 and the untreated x equals 0. And then if we estimate the regression above, beta 1 hat will be the estimated effect of the treatment.

And by the way, x need not be randomly assigned half zeros and half ones or anything like that to be a dummy variable. So we do use dummy variables or can use dummy variables if we do have this sort of random assignment variable. But we can use it for lots of other things too. Any characteristic of the observations that exists on some but not all observations, can be represented with a dummy variable. So there's nothing about a dummy variable that we're assuming is randomly assigned or anything like that.

And we'll see lots of other uses of dummy variables when we get to multiple regression because we can use them to interact with other explanatory variables and shift slopes around and trace out non-linear relationships. We can do all kinds of tricky things with dummy variables, and we'll see that later.

Another practical issue that we'll deal with in more detail in multiple regression, you might be thinking, well, fine, the linear model seems pretty useful. We can use it for all kinds of stuff. But isn't it really restrictive? I mean, maybe the relationship between these variables, there is a relationship, but it's just not linear. And so we're imposing linearity on here. And that seems like something we wouldn't often want to do.

Well, there are a couple of things I can say about this. One, is that the linear model is actually super flexible. So for instance, we can take non-linear functions of both of the variables, x variable and y variable, and analyze the relationships between those non-linear functions of x and y within a linear framework.

We can also create interaction variables, which is the product of two variables. And that allows for other types of non-linearity. And we can do all of this. And like I said, we can use dummy variables to allow for non-linear shapes by tricky multiplication. We'll show you examples of that.

But the point I want to make now, is just that this is actually-- it seems like it might be a restrictive framework. It's actually a super flexible framework. And we'll see all kinds of examples about how flexible the linear framework actually is.

One additional comment I want to make about this, is that we can do the non-parametric version of a linear regression that Esther talked about last time, a kernel regression. And researchers do perform. They do estimate kernel regressions, and they're useful in a lot of situations.

But there are serious trade offs. And there are very good reasons why the linear regression is the workhorse that it is. And the trade-offs mostly come in the form of efficiency. A linear regression is going to be a much, much more efficient way to estimate a relationship than a kernel regression. So the kernel regression is sort of infinitely flexible in some sense. But you pay for it because it's not an efficient estimator at all, really.

So questions about those practical issues concerning linear regression? So let me just get started. I have a few minutes now. And we'll get started on a generalization of the linear model, the sort of multivariate linear model. And this is what it looks like.

So why might we be interested in this more general linear model? I mean, maybe the answer is kind of obvious. But I could talk about a lot of-- we could talk about a lot of examples where we're primarily interested in the relationship between one particular x variable and our dependent or y variable, but there are other factors that might come into play that we need to account for in our analysis.

So that's one very good reason to consider this multivariate linear model, as opposed to just the bivariate linear model. It could also be that we actually have many variables. We don't have any a priori notion about what x variables might be important predictors of y.

So think back to the example of a firm trying to decide who to serve ads on that *Moby Dick* T-shirt to. Well, they might not have any sort of a priori reason to think, well, it's people who read XKCD who might buy this t-shirt.

And what they would like to do is get the web browsing behavior of lots and lots of people and see all kinds of websites that they visit, figure out which ones actually buy this t-shirt, and then use the results of that analysis to figure out who to serve ads to in the future. So that's another reason why we want to consider models that have lots of-- potentially, lots of regressors in them, not just one.

So as I said before, the notation we were using for the bivariate regression it was just sort of not up to the task. It was kind of clunky. It was a little awkward. And it turns out that all of the things that I can say about linear regression can be said much more elegantly using matrix notation.

So what we're going to do-- by the way, maybe I should take a show of hands. How many people are familiar with some basic matrix notation and linear algebra facts? Yes, OK, so there might be a couple who aren't, but it seems like the vast majority of the class is.

So I have put up on the website, also under Resources, a separate handout. And I forget what it's called. Maybe it's called matrix notation et cetera or something like that. And basically, it goes through in greater detail than I'm going to do in lecture.

It gives you both a primer for matrix notation and reminds you definitions, matrix algebra, linear algebra definitions. Reminds you of those. But then also goes through the linear regression stuff in more detail than I'm going to do in lecture. So you can look at those notes if you want. So this is a job for matrix notation. We don't have too much time, but I'll just get started here.

Let's replace this whole parade of x's with a vector, a row vector. And let's say, let's define a new x that's identically equal to 1. And that's called x sub 0 1. And the reason we do this is because, implicitly in the bivariate regression, we had beta 0 times 1.

We never had the 1 there, but we sort of implicitly there was a 1 multiplying beta 0. And then we had beta 1 times x sub i. Well here, we want to explicitly have the 1 in that first spot. So we can do everything in matrix notation. And so let's replace this whole parade of x's with a row vector.

And so for each observation, we have a vector of different measures x0 up through x sub k for each observation. And then let's also replace all of the betas that multiply all of these regressors by a vector. In this case, a column vector, a k plus 1 by 1 column vector of parameters beta.

And then that lets us use a much more condensed notation for our multivariate linear model. And it's just y sub i equals x sub i beta plus epsilon sub i. So this is the model for each individual observation. And so we can go even further. And we can basically put all the observations into matrices or vectors and compress our notation even further.

So now let's let y the entire set of observations on our outcome variable across our different observations. Let's put that into an n by 1 column vector. And then let's put our errors also into an n by 1 column vector.

And then let's take each one of these row vectors that I defined on the previous slide-- so remember, for each observation, we had a row vector that was the measures of all of the x's for each observation. Let's stack those up and put those into an n by k plus 1 matrix.

So remember this column here is identically equal to 1. And then the second column on, are all measures of our explanatory variables x1 through xk. Yes.

AUDIENCE: Is there an in there or is that just [INAUDIBLE]?

SARA ELLISON: Yes, it is supposed to be x. So yes, the i subscript snuck in there, and I apologize for it.

AUDIENCE: So each row is an x.

SARA ELLISON: Yes.

AUDIENCE: [INAUDIBLE]

SARA ELLISON: So each row is x sub i.

AUDIENCE: OK.

SARA ELLISON: Yes, so each row vector, we'll call x sub i. This entire matrix we'll call it x. So you're absolutely right. Is everyone on board with this structure? So one more slide and then we'll call it a day.

Then when we have defined everything and created that structure, we now can write the multivariate linear regression model or linear model as y equals x beta plus epsilon, with all of those dimensions up there.

So we'll start here next time. And see you guys in a couple of days.