PROFESSOR: Today's topic is regression analysis. And this subject is one that we're going to cover it today covering the mathematical and statistical foundations of regression and focus particularly on linear regression. This methodology is perhaps the most powerful method in statistical modeling. And the foundations of it, I think, are very, very important to understand and master, and they'll help you in any kind of statistical modeling exercise you might entertain during or after this course.

And its popularity in finance is very, very high, but it's also a very popular methodology in all other disciplines that do applied statistics. So let's begin with setting up the multiple linear regression problem. So we begin with a data set that consists of data observations on different cases, a number of cases. So we have n cases indexed by i.

And there's a single variable, a dependent variable or response variable, which is the variable of focus. And we'll denote that y sub i. And together with that for each of the cases, there are explanatory variables that we might observe. So the yi's, the dependent variables, could be returns on stocks. The explanatory variables could be underlying characteristics of those stocks over a given period. The dependent variable could be the change in value of an index, the S&P 500 index or the yield rate, and the explanatory variables can be various macroeconomic factors or other factors that might be used to explain how the response variable changes and takes on its value.

Let's go through various goals of regression analysis. OK, first it can be to extract or exploit the relationship between the dependent variable and the independent variable. And examples of this are prediction. Indeed, in finance that's where I used regression analysis most. We want to predict what's going to happen and take actions to take advantage of that.

One can also use regression analysis to talk about causal inference. What factors are really driving a dependent variable? And so one can actually test hypotheses about what are true causal factors underlying the relationships between the variables.

Another application is for just simple approximation. As mathematicians, you're all very familiar with how smooth functions can be that are smooth in the sense of being differentiable and bounded. Those can be approximated well by a Taylor series if you have a function of a single variable or even a multivariable function. So one can use regression analysis to actually approximate functions nicely.

And one can also use regression analysis to uncover functional relationships and validate functional relationships amongst the variables. So let's set up the general linear model from a mathematical standpoint to begin with. In this lecture, OK, we're going to start off with discussing ordinary least squares, which is a purely mathematical criterion for how you specify regression models. And then we're going to turn to the Gauss Markov theorem which incorporates some statistical modeling principles there. They're essentially weak principles.

And then we will turn to formal models with normal linear regression models, and then consider extensions of those to broader classes. Now we're in the mathematical context. And a linear model is basically attempting to model the conditional distribution of the response variable yi given the independent variables xi.

And the conditional distribution of the response variable is modeled simply as a linear function of the independent variables. So the xi's, xi one through xip, are the key explanatory variables that relate to the response variables, possibly. And the beta-- one beta, two beta, or beta p-- are the regression parameters which would be used defining that linear relationship.

So this relationship has residuals, epsilon i, basically where there's uncertainty in the data-- whether it's either due to a measurement error or modeling error or underlying stochastic processes that are driving the error. This epsilon i is a residual error variable that will indicate how this linear relationship varies across the different n cases.

So OK, how broad are the models? Well, the models really are very broad. First of all, polynomial approximation is indicated here. It corresponds, essentially, to a truncated Taylor series approximation to a functional form. With variables that exhibit cyclical behavior, Fourier series, can be applied in a linear regression context.

How many people in here are familiar with Fourier series? Almost everybody. So Fourier series basically provide a set of basis functions that allow you to closely approximate most functions. And certainly with bounded functions that possibly have a cyclical structure to them, it provides a complete description. So we could apply Fourier series here.

Finally, time series regressions where the cases i one through n are really indexes of different time points can be applied. And so the independent variables can be variables that are observable at a given time point or known at a given time. So those can include lags of the response variables. So we'll see actually when we talk about time series that there's auto regressive time series models that can be specified. And those are very broadly applied in finance.

All right, so let's go through what the steps are for fitting a regression model. First, one wants to propose a model in terms of what is it that we have to identify or be interested in a particular response variable. And critical here is specifying the scale of that response variable. Choongbum was discussing problems of modeling stock prices. If, say, y is the stock price? Well, it may be that it's more appropriate to consider modeling it on a logarithmic scale than on a linear scale.

Who can tell me why that would be a good idea?

AUDIENCE: Because the changes might become more perception of the price rather than absolute changes in price.

PROFESSOR: Very good, yeah. So price changes basically on the percentage scale which log changes would be maybe much better predicted by knowing factors than the absolute price level. OK, and so we have to have a collection of independent variables, which to include in the model. And it's important to think about how general this set up is. I mean, the independent variables can be functions, lag values of the response variable. They can be different functional forms of other independent variables.

So the fact that we're talking about a linear regression model here is it's not so when the team in terms of the linearity. We can really capture lot of nonlinear behavior in this framework. So then third, we need to address the assumptions about the distribution of the residuals, epsilon, over the cases. So that has to be specified.

Once we've set up the model in terms of identifying the response of the explanatory variables and the assumptions underlying the distribution of the residuals, we need to specify a criterion for judging different estimators. So given a particular setup, what we want to do is be able to define a methodology for specifying the regression parameters so that we can then use this regression model for prediction or whatever our purpose is.

So the second thing we want to do is define a criterion for how we might judge different estimators of the progression parameters. We're going to go through several of those. And you'll see those-- least squares is the first one, but there are actually more general ones. In fact, the last section of this lecture on generalized estimators will cover those as well.

Third, we need to characterize the best estimator and apply it to the given data. So once we choose a criterion for how good an estimate of regression parameters is, then we have to have a technology for solving for that. And then fourth, we need to check our assumptions.

Now, it's very often the case that at this fourth step, where you're checking the assumptions that you've made, you'll discover features of your data or the process that it's modeling that make you want to expand upon your assumptions or change your assumptions. And so checking the assumptions is a critical part of any modeling process. And then if necessary, modify the model and assumptions and repeat this process.

What I can tell you is that this sort of protocol for how you fit models is what I've applied many, many times. And if you are lucky in a particular problem area, the very simple models will work well with small changes in assumptions. But when you get challenging problems, then this item five of modify the model and/or assumptions is critical.

And in statistical modeling, my philosophy is you really want to, as much as possible, tailor the model to the process you're modeling. You don't want to fit a square peg in a round hole and just apply, say, simple linear regression to everything. You want to apply it when the assumptions are valid. If the assumptions aren't valid, maybe you can change the specification of the problem so a linear model is still applicable in a changed framework.

But if not, then you'll want to extend to other kinds of models. But what we'll be doing-- or what you will be doing if you do that-- is basically applying all the same principles that are developed in the linear modeling framework. OK, now let's see. I wanted to make some comments here about specifying assumptions for the residual distribution.

What kind of assumptions might we make? OK, would anyone like to suggest some assumptions you might make in a linear regression model for the residuals? Yes? What's your name, by the way?

AUDIENCE: My name is Will.

PROFESSOR: Will, OK. Will what?

[? AUDIENCE: Ossler. ?]

PROFESSOR: [? Ossler, ?] great. OK, thank you, Will.

AUDIENCE: It might be-- or we might want to say that the residual might be normally distributed and it might not depend too much on what value of the input variable we'd use.

PROFESSOR: OK. Anyone else? OK. Well, that certainly is an excellent place to start in terms of starting with a distribution that's familiar. Familiar is always good. Although it's not something that should be necessary, but we know from some of Choongbum's lecture areas that Gaussian and normal distributions arise in many settings where we're taking basically sums of independent, random variables. And so it may be that these residuals are like that.

Anyway, a slightly simpler or weaker condition is to use the Gauss-- what are called in statistics the Gauss Markov assumptions. And these are assumptions where we're only concerned with the means or averages, statistically, and the variances of the residuals. And so we assume that there's zero mean. So on average, they're not adding a bias up or down to the dependent variable.

And those have a constant variance. So the level of uncertainty in our model doesn't depend on the case. And so indeed, if errors on the percentage scale are more appropriate, then one could look at, say, a time series of prices that you're trying to model. And it may be that on the log scale, that constant variance looks much more appropriate than on the original scale, which would happen.

And then a third attribute of the Gauss Markov assumptions is that the residuals are uncorrelated. So now uncorrelated does not mean independent or statistically independent. So this is a somewhat weak condition, or weaker condition, than independence of the residuals. But in the Gauss Markov setting, we're just setting up basically a reduced set of assumptions that we might apply to fit the model.

If we extend upon that, we can then consider normal linear regression models, which Will just suggested. And in this case, those could be assumed to be independent and identically distributed-- IID is that notation for that-- with Gaussian or normal. With mean 0 and variance sigma squared. We can extend upon that to consider generalized Gauss Markov assumptions where we maintain still the 0 mean for the residuals, but the general-- we might have a covariance matrix which does not correspond to independent and identically distributed random variables.

Now, let's see. In the discussion of probability theory, we really haven't talked yet about matrix valued random variables, right? But how many people in the class have covered matrix value or vector valued random variables before? OK, just a handful.

Well, a vector valued random variable, we think of the values of these n cases for the dependent variable to be an n valued and n vector of random variables. And so we can generalize the variance of individual random variables to the variance covariance matrix of the collection. And so you have a covariance matrix characterizing the variance of the n vector which gives us the-- the ij element gives us the value of the covariance.

All right, let me put the screen up and just write that on the board so that you're familiar with that. All right, so we have y1, y2 down to yn, our m values of our response variable. And we can basically talk about the expectation of that being equal to mu 1, mu 2, down to mu n.

And the covariance matrix of y1, y2, down to yn is equal to a matrix with the variance of y1 in the upper 1, 1 element, and the variance of y2 in the 2, 2 element, and the variance of yn in the nth column and nth row. And in the i j-th row, ij, we have the covariance between yi and yj.

So we're going to use matrices to represent covariances. And that's something which I want everyone to get very familiar with because we're going to assume that we are comfortable with those, and apply matrix algebra with these kinds of constructs. So the generalized Gauss Markov theorem assumes a general covariance matrix where you can have nonzero covariances between the independent variables or the dependent variables and the residuals. And those can be correlated.

Now, who can come up with an example of why the residuals might be correlated in a regression model? Dan? OK. That's a really good example because it's nonlinear. If you imagine sort of a simple nonlinear curve and you try to fit a straight line to it, then the residuals from that linear fit are going to be consistently above or below the line depending on where you are in the nonlinearity how it might be fitting. So that's one example where that could arise. Any other possibilities?

Well, next week we'll be talking about some time series models. And there can be time dependence amongst variables where there are some underlying factors maybe that are driving the process. And those ongoing factors can persist in making the linear relationship over or under gauge the dependent variable. So that can happen as well.

All right, yes?

AUDIENCE: The Gauss Markov is just the diagonal case?

PROFESSOR: Yes, the Gauss Markov is simply the diagonal case. And explicitly if we replace y's here by the residuals, epsilon, one through epsilon n, then that diagonal matrix with a constant diagonal is the simple Gauss Markov assumption, yeah. Now, I'm sure it comes as no surprise that Gaussian distributions don't always fit everything. And so one needs to get clever with extending the models to other cases.

And there are-- I know-- Laplace distributions, Pareto distributions, contaminated normal distributions, which can be used to fit regression models. And these general cases really extend the applicability of regression models to many interesting settings. So let's turn to specify the estimator criterion in two.

So how do we judge what's a good estimate of the regression parameters? Well, we're going to cover least squares, maximum likelihood, robust methods, which are contamination resistant. And other methods exist that we will mention but not get into really in the lectures are Bayes methods and accommodating incomplete or missing data.

Essentially, as your approach to modeling a problem gets more and more realistic, you start adding more and more complexity as it's needed. And certainly issues of-- well, robust methods is where you assume most of the data arrives under normal conditions, but once in a while there may be some problem with the data. And you don't want your methodology just to break down if there happens to be some outliers in the data or contamination.

Bayes methodologies are the technology for incorporating subjective beliefs into statistical models. And I think it's fair to say that probably all statistical modeling is essentially subjective. And so if you're going to be good at statistical modeling, you want to be sure that you're effectively incorporating subjective information in that. And so Bayes methodologies are very, very useful, and indeed pretty much required to engage in appropriate modeling.

And then finally, accommodate incomplete or missing data. The world is always sort of cruel in terms of you often are missing what you think is critical information to do your analysis. And so how do you deal with situations where you have some holes in your data? Statistical models provide good methods and tools for dealing with that situation.

OK. Then let's see. In case analyses for checking assumptions, let me go through this. Basically when you fit a regression model, you check assumptions by looking at the residuals, which are the basically estimates of the epsilons, the deviations of the dependent variable from their predictions. And what one wants to do is analyze these to determine whether our assumptions are appropriate.

OK, but the Gauss Markov assumptions would be, do these appear to have constant variance? And it may be that their variance depends on time, if the i is indexing time. Residuals might depend on the other variables as well, and one wants to determine that that isn't the case. There are also influence diagnostics identifying cases which are highly influential.

It turns out that when you are building a regression model with data, you treat all the cases as if they're equally important. Well, it may be that certain cases are really critical to estimated certain factors. And it may be that much of the inference about how important a certain factor is is determined by very small number of points. So even though you have a massive data set that you're using to fit a model, it could be that some of the structure is driven by a very small number of cases. So influence diagnostics give you a way of analyzing that.

In the problem set for this lecture, you'll be deriving some influence diagnostics for linear regression models and seeing how they're mathematically defined. And I'll be distributing a case study which illustrates fitting linear regression models for asset prices. And you can see how those play out with some practical examples.

OK, finally there's outlier detection. With outliers, it's interesting. The exceptions in data are often the most interesting. It's important in modeling to understand whether certain cases are unusual. And sometimes their degree of idiosyncrasy can be explained away so that one essentially discards those outliers. But other times, those idiosyncrasies lead to extensions of the model. And so outlier detection can be very important for validating a model.

OK, so with that introduction to regression, linear regression, let's talk about ordinary least squares. Ah. OK, the least squares criterion is for a given a regression parameter, beta, which is considered to be a column vector-- so I'm taking the transpose of a row vector. The least squares criterion is to basically take the sum of squared deviations from the actual value of the response variable from its linear prediction.

So yi minus y hat i, we're just plugging in for y hat i the linear function of the independent variables and the squaring that. In the ordinary least squares estimate, beta hat, minimizes this function. So in order to solve for this, we're going to use matrices. And so we're going to take the y vector, the vector of n values of the dependent variable, the response variable, and x, the matrix of values of the independent variable.

It's important in this set up to keep straight that cases go by rows and columns go by values of the independent variable. Boy, this thing is ultra sensitive. Excuse me. Do I turn off the touchpad here? OK. So we can now define our fitted value, y hat, to be equal to the matrix x times beta.

And with matrix multiplication, that results in the y hat 1 through y hat n. And q of beta can basically be written as y minus x beta transposed y minus x beta. So this term here is an n vector minus the product of the x matrix times beta, which is another n vector. And we're just taking the cross product of that.

And the ordinary least squares estimate for beta solves the derivative of this criterion equaling 0. Now, that's in fact true, but who can tell me why that's true? Say again?

AUDIENCE: Is that minimum?

PROFESSOR: OK. So your name?

AUDIENCE: Seth.

PROFESSOR: Seth? Seth. Very good, Seth. Thanks, Seth. So if we want to find a minimum of q, then that minimum will have, if it's a smooth function, will have a minimum slope equals 0. Now, how do we know whether it's a minimum or not? It could be a maximum.

AUDIENCE: [INAUDIBLE]?

PROFESSOR: OK, right. So in fact, this is a-- q of beta is a convex function of beta. And so its second derivative is positive. And if you basically think about the set-- basically, this is the first derivative of q with respect to beta equaling 0. If you were to solve for the second derivative of q with respect to beta, well, beta is a p vector. So the second derivative is actually a second derivative matrix, and that matrix, you can solve for it. It will be x transpose x, which is a positive, definite or semi-definite matrix. So it basically had a positive derivative there.

So anyway, this ordinary least squares estimates will solve this dq of beta by beta equals 0. What does dq beta by d beta j? Well, you just take the derivative of this sum. So we're taking the sum of all these elements. And if you take the derivative--

well, OK, the derivative is a linear operator. So the derivative of a sum is the sum of the derivatives.

So we take the summation out and we take the derivative of each term, so we get 2 minus xij, then the thing in square brackets, yi minus that. And what is that? Well, in matrix notation, if we let this sort of bold x sub squared j denote the j-th column of the independent variables, then this is minus 2. Basically, the j-th column of x transpose times y minus x beta.

So this j-th equation for ordinary least squares has that representation in terms in matrix notation. Now if we put that all together, we basically can define this derivative of q with respect to the different regression parameters as basically the minus twice the j-th column stacked times y minus x beta, which is simply minus 2x transpose, y minus x beta.

And this has to equal 0. And if we just simplify, taking out the two, we get this set of equations. It must be satisfied by the ordinary least squares estimate, beta. And that's called the normal equation books on regression modeling. So let's consider how we solve that. Well, we can reexpress that by multiplying through the x transpose on each of the terms. And then beta hat basically solves this equation.

And if x transpose x inverse exists, we get beta hat is equal to x transpose x inverse x transpose y. So with matrix algebra, we can actually solve this. And matrix algebra is going to be very important to this lecture and other lectures. So if this stuff is-- if you're a bit rusty on this, do brush up. This particular solution for beta hat assumes that x transpose x inverse exists.

Who can tell me what assumptions do need to make for x transpose x to have an inverse? I'll call you in a second if no one else does. Somebody just said something. Someone else. No? All right. OK, Will.

AUDIENCE: So x transpose x inverse needs to have full rank, which means that each of the submatrices needs to have a smaller range.

PROFESSOR: OK, so Will said, basically, the matrix x needs to have full rank. And so if x has full rank, then-- well, let's see. If x has full rank, then the singular value decomposition which was in the very first class can exist. And you have basically p singular values that are all non-zero. And x transpose x can be expressed as sort of a, from the singular value decomposition, as one of the orthogonal matrices times the square of the singular values times that same matrix transpose, if you recall that definition.

So that actually is-- it basically provides a solution for x transpose x inverse, indeed, from the singular value decomposition of x. But what's required is that you have a full rank in x. And what that means is that you can't have independent variables that are explained by other independent variables. So different columns of x have to be linear, or they can't linearly dependent on any other columns of x. Otherwise, you would have reduced rank.

So now if beta hat doesn't have full rank, then our least squares estimate of beta might be non-unique. And in fact, it is the case that if you are really interested in just predicting values of a dependent variable, then having non-unique unique least squares estimates isn't as much of a problem, because you still get estimates out of that. But for now, we want to assume that there's full column rank in the independent variables.

All right. Now, if we plug in the value of the solution for the least squares estimate, we get fitted values for the response variable, which are simply the matrix x times beta hat. And this expression for the fitted values is basically x times x transpose x inverse x transpose y, which we can represent as hy.

Basically, this h matrix in linear models and statistics is called the hat matrix. It's basically a projection matrix that takes the linear vector, or the vector of values of the response variable, into the fitted values. So this hat matrix is quite important. The problem set's going to cover some features, go into some properties of the hat matrix.

Does anyone want to make any comments about this hat matrix? It's actually a very special type of matrix. Does anyone want to point out what that special type is? It's a projection matrix, OK. Yeah. And in linear algebra, projection matrices have some very special properties. And it's actually an orthogonal projection matrix.

And so if you're interested in that feature, you should look into that. But it's really a very rich set of properties associated with this hat matrix. It's an orthogonal projection, and it's-- let's see. What's it projecting? It's projecting from n space into what? Go ahead. What's your name?

AUDIENCE: Ethan.

PROFESSOR: Ethan, OK.

AUDIENCE: Into space called x.

PROFESSOR: Basically, yeah. It's projecting into the column space of x. So that's what linear regression is doing. So in focusing and understanding linear regression, you can think of, how do we get estimates of this p vector? That's all very good and useful, and we'll do a lot of that. But you can also think of it as, what's happening in the n dimensional space? So you basically are representing this n dimensional vector y by its projection onto the column space.

Now, the residuals are basically the difference between the response value and the fitted value. And this can be expressed as y minus y hat, or in minus h times y. And it turns out that in minus h is also a projection matrix, and it's projecting the data onto the space orthogonal to the column space of x.

And to show that that's true, if we consider the normal equations-- which are x transpose y minus x beta-- hat equaling 0, that basically is x transpose epsilon hat equals 0. And so from the normal equations, we can see of what they mean is they mean that the residual vector epsilon hat is orthogonal to each of the columns of x. You can take any column in x, multiply that by the residual vector, and get 0 coming out. So that's a feature of the residuals as they relate to the independent variables.

OK, all right. So at this point, we've gone through really not talking about any statistical properties to specify the betas. All we've done is talked. We've introduced the least squares criterion and said, what value of the beta vector minimizes that least squares criterion? Let's turn to the Gauss Markov theorem and start introducing some statistical properties, probability properties.

So with our data, yx-- yes? Yes.

AUDIENCE: [INAUDIBLE]?

PROFESSOR: That epsilon--

AUDIENCE: [INAUDIBLE]?

PROFESSOR: OK. Let me go back to that. It's that x, the columns of x, and the column vector of the residual are orthogonal to each other. So we're not doing a projection onto a null space. This is just a statement that those values, or those column vectors, are orthogonal to each other.

And just to recap, the epsilon is a projection of y onto the space orthogonal to the column space. And y hat is a projection onto the column space of y. And these projections are all orthogonal projections, and so they happen to result in the projected value epsilon hat must be orthogonal to the column space of x if you project it out. OK? All right.

So the Gauss Markov theorem, we have data, y and x again. And now we're going to think of the observed data, little y, 1 through yn, is actually an observation of the random vector capital Y composed of random variables y1 up to yn. And the expectation of this vector conditional on the values of the independent variables and their regression parameters given by x beta, so the dependent variable vector has expectation given by the product of the independent variables matrix times the regression parameters.

And the covariance matrix of y given x and beta is sigma squared times the identity matrix, the n dimensional identity matrix. So the identity matrix has 1's along the diagonal n dimensional and 0's off the diagonal. So the variances of the y's are the diagonal entries, those are all the same, sigma squared. And the covariance between any two are equal to 0 conditional.

OK, now the Gauss Markov theorem. This is a terrific result in linear models theory. And it's terrific in terms of the mathematical content of it. I think it's-- for a math class, it's really a nice theorem to introduce you to and highlight the power of, I guess, results that can arise from applying the theory.

And so to set this theorem up, we want to think about trying to estimate some function of the regression parameters. And so OK, our problem is with ordinary least squares-- it was, how do we specify the regression parameters, beta 1 through beta p? Let's consider a general target of interest, which is a linear combination of the betas.

So we want to predict a parameter, theta, which is some linear combination of the regression parameters. And because that linear combination of the regression parameters corresponds to the expectation of the response variable corresponding to a given row of the independent variables matrix, this is just a generalization of trying to estimate the means of the regression model at different points in the space, or to be estimating other quantities that might arise.

So this is really a very general kind of thing to want to estimate. It certainly is appropriate for predictions. And if we consider the least squares estimate by just plugging in theta hat one through beta hat p, solve by the least squares-- well, it turns out that those are an unbiased estimator of the parameter theta. So if we're trying to estimate this combination of these unknown parameters, you plug in the least squares estimate. You're going to get an estimator that's unbiased.

Who can tell me what unbiased is? It's probably going to be a new concept for some people here. Anyone? OK, well it's a basic property of estimators in statistics where the expectation of this statistic is the true parameter. So it doesn't, on average, probabilistically, it doesn't over or underestimate the value. So that's what unbiased means.

Now, it's also a linear estimator of theta in terms of this theta hat being a particular linear combination of the dependent variables. So with our original response variable y, in the case of y1 through yn, this theta hat is simply a linear combination of all the y's. And now why is that true?

Well, we know that beta hat, from the normal equations, is solved by x transpose x inverse x transpose y. So it's a linear transform of the y vector. So if we take a linear combination of those components, it's also another linear combination of the y vector. So this is a linear function of the underlying-- of the response variables.

Now, the Gauss Markov theorem says that, if the Gauss Markov assumptions apply, then the estimator theta that has the smallest variance amongst all linear unbiased estimators of theta. So it actually is like the optimal one, as long as this is our criteria. And this is really a very powerful result. And to prove it, it's very easy. Let's see.

Actually, these notes are going to be distributed. So I'm going to go through this very, very quickly and come back to it later if we have more time. But you basically-- the argument for the proof here is you consider another linear estimate, which is also an unbiased estimate. So let's consider a competitor to the least squares value and then look at the difference between that estimator and theta hat.

And so that can be characterized as basically this vector, f transpose y. And this difference in the estimates must have expectation 0. So basically, if we look at-- if theta tilde is unbiased, then this expression here is going to be equal to zero-- which means that f, the difference in these two estimators, f defines the difference in the two estimators, has to be orthogonal to the column space of x.

And with this result, one then uses this orthogonality of f and d to evaluate the variance of theta tilde. And in this proof, the mathematical argument here is really something-- I should put some asterisks on a few lines here. This expression here is actually very important. We're basically looking at the decomposition of the variance to be the variance of B transpose y, which is the variance of the sum of these two random variables.

So the page before basically defined d and f such that this is true. Now when you consider the variance of a sum, it's not the sum of the variances. It's the sum of the variances plus twice the sum of the covariances. And so when you are calculating variances of sums of random variables, you have to really keep track of the covariance terms.

In this case, this argument shows that the covariance terms are, in fact, 0, and you get the result popping out. But that's really a-- in an econometrics class, they'll talk about blue estimates of regression, or the blue property of the least squares estimates. That's where that comes from. All right, so let's now consider generalizing from Gauss Markov to allow for unequal variances and possibly correlated nonzero covariances between the components. And in this case, the regression model has the same linear set up. The only difference is the expectation of the residual vector is still 0. But the covariance matrix of the residual vector is sigma squared, a single parameter times, let's say capital sigma.

And we'll assume here that this capital sigma matrix is a known n by n positive definite matrix specifying relative variances and correlations between the observations. OK. Well, in order to solve for regression estimates under these generalized Gauss Markov assumptions, we can transform the data yx to y star equals sigma to the minus 1/2 y and x to x star, which is sigma to the minus 1/2 x.

And this model then becomes a model, a linear regression model, in terms of y star and x star. We're basically multiplying this regression model by sigma to the minus 1/2 across. And epsilon star actually has a covariance matrix equal to sigma squared times the identity. So if we just take a linear transformation of the original data, we get a representation of the regression model that satisfies the original Gauss Markov assumptions.

And what we had to do was basically do a linear transformation that makes the response variables all have constant variance and be uncorrelated. So with that, we then have the least squares estimate of beta is the least squares, the ordinary least squares, in terms of y star and x star. And so plugging that in, we then have x star transpose x star inverse x star transpose y star. And if you multiply through, that's how the formula changes.

So this formula characterizing the least squares estimate under this generalized set of assumptions highlights what you need to do to be able to apply that theorem. So with response values that have very large variances, you basically want to discount those by the sigma inverse. And that's part of the way in which these generalized least squares work. All right.

So now let's turn to distribution theory for normal regression models. Let's assume that the residuals are normals with mean 0 and variance sigma squared. OK, the conditioning on the values of the independent variable, the y's, the response variables, are going to be independent over the index i. They're not going to be identically distributed because they have different meanings, mu i, for the dependent variable yi. But they will have a constant variance.

And what we can do is basically condition on x beta and sigma squared and then represent this model in terms of the distribution of the epsilons. So if we're conditioning on x and beta, this x beta is a constant known. we've conditioned on it. And the remaining uncertainty is in the residual vector, which is assumed to be all independent and identically distributed normal random variables.

Now, this is the first time you'll see this notation, capital N sub little n, for a random vector. It's a multivariate normal random variable where you consider an n vector where each component is normally distributed, with mean given by some corresponding mean vector, and a covariance matrix given by a covariance matrix. In terms of independent and identically distributed values, the probability structure here is totally well-defined.

Anyone here who's taken a beginning probability class knows what the density function is for this multivariate normal distribution because it's the product of the independent density functions for the independent components, because they're all independent random variables. So this multivariate normal random vector has a density function which you can write down, given this is your first probability class.

OK, here I'm just highlighting or defining the new vector for the means of the cases of the data. And the covariance matrix sigma is this diagonal matrix. And so basically sigma ij is equal to sigma squared times the Kronecker delta for the ij element. Now what we want to do is, under the assumptions of normally distributed residuals, to solve for the distribution of the least squares estimators. We want to know, basically, what kind of distribution does it have?

Because what we want to be able to do is to determine whether estimates are particularly large or not. And maybe there's no structure at all and the regression parameters are 0 so that there's no dependence on a given factor. And we need to be able to judge how significant that is. So we need to know what the distribution is of our least squares estimate.

So what we're going to do is apply moment generating functions to derive the joint distribution of y and the joint distribution of beta hat. And so Choongbum introduced the moment generating function for individual random variables for single variate random variables. For n variate random variables, we can define the moment generating function of the y vector to be the expectation of e' to the t transpose y.

So t is an argument of the moment generating function. It's another n vector. And it's equal to the expectation of e to the t1 y1 plus t2 y2 up to tn yn. So this is a very simple definition.

Because of independence, the expectation of the products, or this exponential sum is the product of the exponentials. And so this moment generating function is simply the product of the moment generating functions for y1 up through yn. And I think-- I don't know if it was in the first problem set or in the first lecture, but e to the ti mu i plus a 1/2 ti squared sigma squared was the moment generating function for the single univariate normal random variable, meaning ui invariance sigma squared.

And so if we have n of these, we take their product. And the moment generating function for y is simply e to the t transpose me plus 1/2 t transpose sigma of t. And so for this multivariate normal distribution, this is its moment generating function. And this happens to be the distribution of y is a multivariate normal with mean u and covariance matrix sigma.

So a fact that we're going to use is that if we're working with multivariate normal random variables, this is the structure of their moment generating functions. And so if we solve for the moment generation function of some other item of interest and recognize that it has the same form, we can conclude that it's also a multivariate normal random variable.

So let's do that. Let's solve for the moment generation function of the least squares estimate, beta hat. Now rather than dealing with an n vector, we're dealing with a p vector of the betas, beta hats. And this is simply the definition of the moment generating function.

If we plug-in for basically what the functional form is for the ordinary least squares estimates and how they depend on the underlying y, then we basically-- OK, we have A equal to, essentially, the linear projection of y. That gives us the least squares estimate. And then we can say that this moment generating function for beta hat is equal to the expectation of E to the t transpose y, where little t is a transpose tau.

Well, we know what this is. This is the moment generating function of x-- sorry, of y-- evaluated at the vector little t. So we just need to plug in little t, that expression of A transpose tau. So let's do that. And you do that and it turns out to be E to the t transpose mu plus that. And we go through a number of calculations.

And at the end of the day, we get that the moment generating function is just E to the tau transpose beta plus a 1/2 tau transpose this matrix tau. And that is the moment generation function of a multivariate normal. So these few lines that you can go through after class basically solve for the moment generating function of beta hat. And because we can recognize this as the MGF of a multivariate normal, we know that that's-- beta hat is a multivariate normal, with mean, the true beta, and covariance matrix given by the object in square brackets there.

OK, so this is essentially the conclusion of that previous analysis. The marginal distribution of each of the beta hats is given by beta hat-- by a univariate normal distribution with mean beta j and variance equal to the diagonal. Now at this point, saying that is like an assertion. But one can actually prove that very easily, given this sequence of argument. And can anyone tell me why this is true?

Let me tell you. If you consider plugging in the moment generating function, the value tau, where only the j-th entry is non-zero, then you have the monitoring function of the j-th component of beta hat. And that's a Gaussian moment generating function. So the marginal distribution of the j-th component is normal. So you get that almost for free from this multivariate analysis.

And so there's no hand waving going on in having that result. This actually follows directly from the moment generating functions. OK, let's now turn to another topic. Related, but it's the qr decomposition of x.

So we have what are independent variables. x, we want to express this as a product of an orthonormal matrix q which is n by p, and an upper triangular matrix, r. So it turns out that any matrix, n by p matrix, can be expressed in this form. And we'll quickly show you how that can be accomplished. We can accomplish that by conducting a Gram-Schmidt orthonormalization of the independent variables matrix x.

And let's see. So if we define r, the upper triangular matrix in the qr decomposition, to have 0's off the diagonal below and then possibly nonzero value along the diagonal into the right, we're just going to solve for q and r through this Gram-Schmidt process. So the first column of x is equal to the first column of q times the first element, the top left corner of the matrix r.

And if we take the cross product of that vector with itself, then we get this expression for r when 1 squared-- we can basically solve for r 1,1 as the square root of this dot product. And Q1 is simply the first column of x divided by that square root. So this first element of the queue matrix and the first element r, this can be solved for right away.

Then let's solve for the second column of Q and the second column of the R matrix. Well, x2, the second column of the x matrix, is the first column of q times R1, 2, plus the second column of Q times r 2,2. And if we multiply this expression by Q1 transpose, then we basically get this expression for r 1,2. So we actually have just solved for r 1,2.

And Q2 is solved for by the arguments given here. So basically, we successively are orthogonalizing columns of x to the previous columns of x through this Gram-Schmidt process. And it basically can be repeated through all the columns.

Now with this QR decomposition, what we get is a really nice form for the least squares estimate. Basically, it simplifies to the inverse of R times Q transpose y. And this basically means that you can solve for least squares estimates by calculating the QR decomposition, which is a very simple linear algebra operation, and then just do a couple of matrix products to get the-- well, you do have to do a matrix inverse with R to get that out.

And the covariance matrix of beta hat is equal to sigma squared x transpose x inverse. And in terms of the covariance matrix, what is implicit here but you should make explicit in your study, is if you consider taking a matrix, R inverse Q transpose times y, the only thing that's random there is that y vector, OK? The covariance of a matrix times a random vector is that matrix times the covariance of the vector times the transpose of the matrix.

So if you take a matrix transformation of a random vector, then the covariance of that transformation has that form. So that's where this covariance matrix is coming into play. And from the MGF, the moment generating function, for the least squares estimate, this basically comes out of the moment generating function definition as well.

And if we take x transpose x, plugging in the QR decomposition, only the R's play out, giving you that. Now, this also gives us a very nice form for the hat matrix, which turns out to just be Q times Q transpose. So that's a very simple form.

So now with the distribution theory, this next section is going to actually prove what's really a fundamental result about normal linear regression models. And I'm going to go through this somewhat quickly just so that we cover what the main ideas are of the theorem. But the details, I think, are very straightforward.

And these course notes that will be posted online will go through the various steps of the analysis. OK, so there's an important theorem here which is for any matrix A, m by n, you consider transforming the random vector y by this matrix A. It is also a random normal vector. And its distribution is going to have a mean and covariance matrix given by mu z and sigma z, which have this simple expression in terms of the matrix A and the underlying means and covariances of y.

OK, earlier we actually applied this theorem with A corresponding to the matrix that generates the least squares estimates. So with A equal to x transpose x inverse, we actually previously went through the solution for what's the distribution of beta hat. And with any other matrix A, we can go through the same analysis and get the distribution.

So if we do that here, well, we can actually prove this important theorem, which says that with least squares estimates of normal linear regression models, our least squares estimate beta hat and our residual vector epsilon hat are independent random variables. So when we construct these statistics, they are statistically independent of each other.

And the distribution of beta hat is multivariate normal. The sum of the squared residuals is, in fact, a multiple of a chi squared random variable. Now who in here can tell me what a chi squared random variable is? Anyone?

AUDIENCE: [INAUDIBLE]?

PROFESSOR: Yes, that's right. So a chi squared random variable with one degree of freedom is a squared normal zero one random variable. A chi squared with two degrees of freedom is the sum of two independent normals 0 1 squared. And so the sum of n squared residuals is, in fact, an n minus p chi squared random variable scale it by sigma squared.

And for each component j, if we take the difference between the least squares estimate beta hat j and beta j and divide through by this estimate of the standard deviation of that, then that will, in fact, have a t distribution on n minus p degrees of freedom. And let's see, a t distribution in probability theory is the ratio of a normal random variable to an independent chi squared random variable, or the root of an independent chi squared random variable.

So basically these properties characterize our regression parameter estimates and t statistics for those estimates. Now, OK, in the course notes, there's a moderately long proof. But all the details are given, and I'll be happy to go through any of those details with people during office hours.

Let me just push on to-- let's see. We have maybe two minutes left in the class. Let me just talk about maximum likelihood estimation. And in fitting models and statistics, maximum likelihood estimation comes up again and again. And with normal linear regression models, it turns out that ordinary least squares estimate are, in fact, our maximum likelihood estimates.

And what we want to do with a maximum likelihood is to maximize. We want to define the likelihood function, which is the density function for the data given the unknown parameters. And this density function is simply the density function for a multivariate normal random variable. And the maximum likelihood estimates are the estimates of the underlying parameters that basically maximize the density function. So it's the values of the underlying parameters that make the data that was observed the most likely.

And if you plug in the values of the density function, basically we have these independent random variables, yi, whose product is the joint density. The likelihood function turns out to be basically a function of the least squares criterion. So if you fit models by least squares, you're consistent with doing something decent in at least applying the maximum likelihood principle if you had a normal linear regression model.

And it's useful to know when your statistical estimation algorithms are consistent with certain principles like maximum likelihood estimation or others. So let me, I guess, finish there. And next time, I will just talk a little bit about generalized M estimators. Those provide a class of estimators that can be used for finding robust estimates and also quantile estimates of regression parameters which are very interesting.