[SQUEAKING]

[RUSTLING]

[CLICKING]

PETER KEMPTHORNE: All right. Well, let's continue with our discussion of linear regression modeling. And as we introduced last time, there are ways one can formalize fitting a model, which is to propose a model, specify criteria for judging different fits or estimators of model parameters, then find the best estimators, and then check our assumptions underlying the specification of the criterion. And possibly, if necessary, modify the model because the assumptions that we made aren't satisfied. So we either want to add additional assumptions or consider transformations of the model, perhaps.

So with ordinary least squares regression, we have this criterion for specifying regression parameters. And we specify these in terms of a y vector. And this has values of a dependent variable across cases in the data, an x matrix, which will have n rows and p columns.

And we'll have general elements Xij. And then our model will be that y is equal to x, the matrix, times a beta vector. So we can have a beta vector that has p components, and an error vector, which has n terms corresponding to the n cases.

So in specifying the regression model, we have that our n vector y is going to be a linear combination of the columns of x. And there will be an error vector characterizing the discrepancy from that.

So how should we specify this beta vector if we have a least squares criterion? We can look at sum 1 to n of yi minus y hat i, where, say, y hat is equal to x beta hat. And this criterion is basically a sum-of-squares criterion. So it's like a sum of squared errors.

Now minimizing this can be done just with simple calculus. We can plug in for-- in the formula and have this sum-of-squares criterion, which is simply y minus x beta transpose y minus x beta. So we have Q of beta is basically y minus x beta squared.

So it's the distance of our n vector from the prediction, or fitted value, I guess. Maybe I'll put the beta hat here to make it explicit. And we basically want our fitted values to be as close as possible to the actual values.

Now because this criterion is a quadratic-- it's a quadratic in the regression coefficients, we can simply take derivatives of Q and solve for those derivatives equal to 0. So that's the first-order conditions. And let's see. How do we know whether the solution to this equation actually minimizes the sum of squared errors criterion?

A first-order condition is one where the derivative of the function is equal to 0. That could be at a minimum, but it also could be at a maximum. Yes?

AUDIENCE: You also want to check the second order derivative and make sure that it's positive.

PETER KEMPTHORNE: Yes. We want to take the second-order derivative. And if it's positive, then we have a-- I guess it's a concave-- or, I'm sorry, a convex function. And so the first derivative will solve that.

Now this is the first derivative equation. If we were to look at dQ of beta by d beta j-- let's see. Is that beta j? Right.

If we were to look at the second order derivative by d beta d beta transpose, then analytically, we'll basically get this term times minus x. And so that will end up being a positive term, in terms of products of columns of x and the full x matrix. And so with that, we have this set of equations.

Now if you read up on regression theory, linear model theory, this equation of the derivative of Q with respect to beta equaling 0 is equivalent to this equation here, which is just obtained by eliminating the minus 2 factor and solving for beta. And this is called the normal equations.

What is familiar to some people is that in statistics theory, the names of important terms is often not very sophisticated, in terms of here we have what we call the normal equations. There's nothing really normal about it, but it's a way of specifying a set of equations that we need to solve.

And if the X transpose X matrix is invertible, then we have a solution by multiplying both sides of this equation by the inverse. And in order for this process to work, we need to have that X must have full column rank.

All right. So here, if we plug in that least squares estimate-- so we have beta hat equal to x transpose x inverse x transpose y. If we plug that into our formula for the fitted values, then this becomes this formula times y. And I'll put square brackets around that. And this is called the hat matrix by some statisticians.

Now this matrix, H, is actually a very special matrix. It has special properties. Does anyone know what those special properties might be? Anybody in the back know? Kai.

AUDIENCE: From memory, it's a projection onto the column space of x.

PETER KEMPTHORNE: Yes, it is. Exactly. So H is a projection matrix and it's onto the column space of the X matrix. Now what are properties of a projection matrix?

If we multiply the projection matrix by itself, we actually get the projection matrix. And what's relevant is that if we were to take Hy hat, which is then equal to H times Hy, if this first term is in the column space of x as a projection, then projecting it onto that same column space will have no difference. So this is logically what it should be. And if we multiply out the projection matrix times itself, we get this property.

OK. Well, Importantly in modeling is to look at residuals, which are model errors. And so if we have epsilon hat is equal to y minus y hat, this is y minus-- or y times the identity matrix minus Hy. And this matrix here, multiplying y to give us our residuals, it is also a projection matrix.

So this is a projection as well. And if we multiply I minus H times I minus H and expand it out, we get I squared minus 2H plus H squared. And that equals I minus H. So we can mathematically verify that it's a projection.

And a very important property of this residual vector is that it is orthogonal to the fitted values. So in the normal equations, we have this equation must equal zero. This term here is the residual vector. And if we multiply that n vector by the transpose of x, and it's equal to 0, then we must have that the error vector, epsilon hat, is orthogonal to x.

Now let's see. I'm not sure if it's on the next slide or not. I think something to keep in mind that's really quite useful when we use linear algebra with least squares is that we have y, being in n-dimensional space, the-- we have an X matrix n by p, which corresponds to a column space of x, which is a subspace of R to the n.

And let's see. If we have y hat and epsilon hat, these are orthogonal to each other. So if we think in n dimensions of having a y vector and have a y hat n vector, then y minus y-- sorry. Not y hat, y minus y hat. So if we take the vector from here to here, that's y minus y hat.

OK, this is going to be orthogonal to y hat. So we actually have right angles here in between the y hat vector and the y vector. Sorry, we have-- the y hat vector and the error vector.

And so from this, we have that the squared length of y is equal to the squared length of y hat plus the squared length of the residual vector. And then if we expand this squared length out, we actually get 0 because of the orthogonality.

So with least squares, we basically have a Pythagorean theorem that generalizes from two dimensions to n dimensions. And that generalization is quite convenient and useful when we think about not the mathematical exercise of least squares, but we think about having probability distributions on, say, the errors.

So that leads us to trying to extend the mathematical regression model to a probability model. And so with this, we have constants assumptions for the independent variables in the x matrix, constants for the regression parameters. But the error terms, those will be assumed to follow some probability model.

What's useful to think about is, well, what would be the simplest possible probability model to apply to a regression? Well, we could assume that the errors are iid, which means independent and identically distributed. And we could also, for convenience, assume that they are normally distributed.

And so this is a normal linear regression model. And when we write out the model equation here, the model equation basically is constants plus the error vector. And the error vector, epsilon 1 to epsilon n being iid normal, this, in fact, is consistent, or is equivalent to a multivariate normal distribution of dimension n with a 0 vector for the mean and a covariance matrix that's diagonal with constant variance.

So this is going to be a convenient model for us to work with. And what we'll see is that different inferences about how important different independent variables are in a regression model can be judged using this model as a baseline model for that evaluation. So importantly, the distribution of the residuals leads us to a distribution of the dependent variable vector y.

And so in this slide, we use the notation of a mu vector for the conditional expectation of the y vector given x and beta. And we'll define this just to be a mu vector. That's an n vector.

And we'll define the covariance matrix of the y vector, or the conditional covariance. OK. This is going to be equal to the product of y minus its mean times the transpose of y minus its mean. And so this is an n by 1, 1 by n term. So this is an n by n matrix. And we'll call this capital sigma as a covariance matrix.

Now for the special case of iid errors, this covariance matrix will be diagonal with constant sigma squared. And it will turn out that this conditional distribution of y given x and beta will have a multinomial distribution with a mean vector mu and a covariance matrix sigma. So an n-dimensional multivariate normal distribution.

OK. Well, what we're going to do is show results of what distribution the y vector is, as well as show what the distribution is of the least squares estimate is using moment-generating functions. So this example of the application of moment-generating functions is really quite neat because the arguments are very simple and accessible.

So if we have an invariate random vector y and a constant vector t, then the moment-generating function of the Y vector is equal to the expected value of the exponential of t transpose Y. So this moment-generating function has-- is a function of the t vector. And we just plug into this formula.

Now because the Y's, in fact, are independent of each other, then we have this expected product of terms is the product of the expectations. So we're going to use that independence property where expectations of products are products of the expectations.

And then for each of the Y's, we can plug in the moment-generating function for each Yi. And that's e to the ti of mu i plus a half ti squared sigma squared. We went through this moment-generating function in our section on probability theory.

Now what's important from this example is if we take this product of all the moment-generating functions, and then try to express it in terms of the t vector, the mu vector, and the sigma matrix, we get this formula. And this formula is, in fact, the moment-generating function for a multivariate normal with a given mean vector, mu, and a given covariance matrix, sigma. So that result is trivial. Not surprising.

What's less surprising is that we can use moment-generating functions to solve for the distribution of the least squares estimate. So this first formula here is the moment-generating function for beta hat. And in writing that out, M sum beta hat, I'm going to use a different argument, tau, which will be a p vector. And this is going to be the expected value of E to the tau transpose beta hat. So this is the definition of the moment-generating function.

And if we define the matrix A to be basically the premultiple factor of A that gives us beta hat, then our moment-generating function for beta hat can be written as this expression, just plugging in AY for beta hat. But then we can express tau transpose A equal to t transposed.

So this moment-generating function of the least squares estimates beta hat is actually the value of the moment-generating function of the multinomial vector evaluated at t equal to A transpose tau. And so we can simply plug in this value of t and get our result.

And so t transpose mu ends up being tau transpose beta. And t transpose sigma t is equal to this product. And if we simplify it, we basically have a mean vector, beta, for the least squares estimate, and a covariance matrix for the least squares estimate that's given by sigma squared x transpose x inverse.

So this is our moment-generating function of beta hat. And by the uniqueness of moment-generating functions, we can recognize this as a multinomial with the given mean vector and the given covariance matrix. So we have multivariate normality of the least squares estimates, which is very convenient and useful.

Now, from this, we end up getting marginal distributions that are normal for each of the least squares estimates. So we're just thinking in this slide of the j-th component of beta hat. Well, it's univariate normal with mean beta J and variance given by sigma squared times the j-th diagonal element of x transpose x inverse. And so we get this nice distribution result.

Now there's more distribution theory here to go through. But let's see. Something just to think about that's, I think, very powerful is in some examples where we have this model, we may have control over what our independent variables matrix is. And in a scientific experiment, we might vary different conditions of the experiment, yielding y for different conditions of x.

And with these models, we might be interested in a particular column of the X matrix and want to estimate the beta j, the regression parameter relationship for that. And so if we had control over. The X matrix, we might try to make the j-th diagonal the smallest we can so that the precision of that regression parameter estimate is as small as possible.

And so suppose, I guess, for simplicity that the X matrix-- suppose that x transpose x is a diagonal matrix. And here's Cj. The diagonal values of x transpose x inverse will simply be-- so if we do a minus 1 there, we can do a minus 1 here.

Our diagonal entries will be smallest-- smaller the more large the j-th column is. And so when you think of regression models and experiments, we basically would perhaps want to have the spread in an independent variable be as big as possible in order to specify the line.

So if we think of a regression model of y on X and we have-- if we have points of x and y, basically the-- if we look at the sample mean of the X's and we consider concentrating our x values as far away as possible, that, presumably, will yield more accurate estimates of the slope parameter in the relationship.

And so you can think of-- basically, if we observe points that are far away from the mean, then those points will pin down the regression line the greatest. And this issue arises in experimental design. And in experimental design, one can think of how one can construct X matrices that have these properties.

And as you might expect, the properties relate to the eigenvalues of X transpose X and looking at methods for maximizing those eigenvalues.

All right. Well, we can do some more distribution theory, which is to look at what the distribution is of our error vector y vector. But we can also look at any matrix A times y and look at the distribution of that transformation. It'll be a linear combination of the y's for each component of z.

And it will actually have an m-dimensional multinomial distribution with mean value given by A times the expected value of y, and covariance matrix equal to A times the covariance of yA transposed.

And so this matrix, A, will give us the distribution of beta hat, as we just did. But we can also use different definitions of and A and z. And the normal distribution theory leads to these key properties of normal linear model terms. We have-- the beta hat least squares estimate is multinomial.

We have that the error vector, epsilon hat, is also multinomial. But it's multinomial in n dimensions with mean vector 0. So epsilon hat, which is equal to I minus H y is distributed as a multinomial in n dimensions with a covariance matrix given by sigma squared times the projection.

Now what's significant is that this covariance matrix here is not a full rank. So this is nonsingular. Actually, is that the right word? No. It's singular, I guess. But it's singular. It's not invertible.

And what does that mean?

Well, that means that there are linear combinations of the error vector that have zero variance. There's a linear dependence in the error vectors. And so if we consider x hat and x transpose this, and this is equal to the 0 vector. This is the normal equations.

We basically have that fixed linear combinations of these residuals. The residual vectors are identically equal to 0. So we don't have n independent error terms. And so because of that, when we look at, for example, estimating the error variance sigma squared, well, if we take the sum of the squared residuals, that's the trace of the covariance matrix of the residual vector.

And that's equal to the trace of this projection matrix I n minus H. And so the trace of a sum is the sum of the traces. And the trace of A matrix is-- product is also the trace of the product with terms reversed. So this gives us sigma squared n minus trace of H.

And we get n minus p sigma squared for the expected value of sigma squared. And in regression models with normal linear models, we use this relationship to estimate sigma squared. Dividing both sides by n minus p, we get an unbiased estimate of the error variance.

Now an additional property that's from part C of this theorem is that the error vector, epsilon hat, and the regression parameter, beta hat, those happen to be independent of each other. And the independence follows by looking at the joint distribution, the joint moment-generating function of Ay, and showing that that is equal to the moment-generating function of beta hat times the moment-generating function of epsilon hat. So the product of moment-generating functions allows us to determine the independence there.

And when we have independence of the error vector and the regression parameter vector, then we can compute t-statistics for regression parameters. And we end up, as stated in the notes there, that tj equal to beta hat j minus the true beta j divided by sigma hat times Cjj.

This distribution will have a t-distribution. Now when you studied statistics, if you took a statistics course, you, I'm sure, came across the t-distribution. And the t-distribution is rather remarkable. This turns out to be equal in distribution to a normal 0 distribution, mean 0 and some variance divided by the square root of a chi squared distribution divided by its degrees of freedom, where these two things are independent of each other. And so the t-distribution-- well, who can comment on what properties a t-distribution has compared to a normal?

The t-distribution is symmetrical. It's somewhat bell-shaped. Well, if this numerator is a normal, and this denominator is random, it's random near 1, then we'll have a heavier-tailed distribution with the t. And

So we need to be able to quantify how significantly different from 0 the t-statistic is. So if we have an H0 that beta J is equal to 0, then this distribution turns out to be-- basically have a t-distribution-- a t-distribution with n minus p degrees of freedom.

And so we can judge whether our data provides evidence against this null hypothesis or not. And we can test other hypotheses as well, not only that the beta j is equal to 0. We'll be covering this in a little more detail with one of the lecture notes that was passed out for today.

But the-- let's see. Does anyone know the history of the t-distribution, how it was discovered?

AUDIENCE: Through Guinness, the Guinness Factory?

PETER KEMPTHORNE: Yes. There was a statistician who was working at Guinness with quality control. And he would look at-- he would look at samples, very small samples of size four, from a-- so we'll just write here "Guinness."

He would look at sample, say, x1, x2, x3, x4 of a measurement. I'm not sure quite what the quality measure was, but he would calculate x bar and the standard deviation of x, which is the square root of the sum of xi minus x bar squared over 3.

Well, actually, I think he did it-- well I'll write it in three. And what he discovered was that if we look at Z equal to x bar over sX, these outcomes of the sample means, which are suitably rescaled, should follow a normal distribution because we're looking at a z-score. This was not distributed as a normal 0, 1. There was much greater variability.

And so the person who did this work, they wouldn't allow him to publish under his real name. And so he published a paper under the pseudonym "Student." And so we call this the Student's t-distribution now.

But what was really quite remarkable was to see that in small samples, the variability of these kinds of statistics, in theory, should be close to normal, but in fact are systematically different. And with-- let's see. With other questions that come up in regression, one can construct an F test of whether all the regression parameters beyond the first P are equal to 0 or not.

And one can look at the residual sum of squares from fitting the full model, and then the residual sum of squares from fitting a submodel using just the first k. Sorry, the first k. I misspoke. And then this F statistic is a ratio of normalized sums of squares and differences thereof.

And this F-test statistic comes up as an analysis of variance statistic where one can estimate variability using the residual sum of squares from the full model. And one can then also look at the residual sum of squares from the submodel. And if both models are true, meaning beta k plus 1 to p are equal to 0, then these formulas here are estimating the same variance and they're independent of each other. And so we get the F distribution as the ratio of two chi squared distributions that have different degrees of freedom.

Well let's take a look at some real data. The example data set I want to introduce is actually not a financial data set. This is a medical study. But this example comes up in some work by Efron and Hastie, Brad Efron and Trevor Hastie. And they have some books. There's *The Elements of Statistical Learning.* Actually, this is Hastie. And then there's-- with Efron, there's an advanced book on computational statistics, a text that I use in 18655.

And in this data set, one basically has a response variable and independent variables. And what's really key to know is experience we have with one regression problem extends to other regression problems. Just the names of variables change, but the same issues arise. And so with this data set, we're trying to predict the lpsa variable as it varies with other independent variables for subjects in a data set.

This data set happened to be a data set on patients, or subjects with prostate cancer. And so with any data set, one can compute summary statistics, which help us detect whether there are issues with the data or not. There's this pairs function in R, which is quite convenient for just displaying all of the data. And if you look at this Pairs Plot, every pair of variables in the data set are plotted in a scatter plot with the other variables.

And so along the diagonal, we have the names of the variables, and then we have different relationships of those. Now if lpsa is the dependent variable of interest, if we look of across this row, we're looking at the scatter plot of different independent variables with that. And it looks like the lca vol is probably a very strong relationship. Some of the variables, in fact, are discrete, and we're able to see that as well.

And if we just fit a simple regression, we end up getting output from the regression, which consists of an estimate. We have a coefficient table with an estimate column, a standard error column, and then a t value, and then finally what's called a p value column. This is the probability that we have a larger value of the t-statistic being observed. And what's being tabled here is results of hypothesis testing for each of the regression coefficients equaling 0 or not in the true model.

So we have estimates of these different regression parameters. We can get standard errors of those. The standard errors correspond to sigma hat Cjj. Actually, I think it's the square root of sigma squared Cjj.

And then the t-value is just the signal estimate to noise ratio, which scales in magnitude how important those factors are. And we can calculate the p value. Let's see. So if we have, say, a beta hat j value, and it has this t-distribution, which is centered at the true beta j, and the spread is given by C hat Cj square root, we can calculate-- we can basically test whether beta j is equal to 0.

And so if we set this equal to 0, if we observe a value of beta hat j, we can calculate how likely is it that we get a beta hat j minus beta minus 0 over sigma hat Cjj. This is our t-statistic. We can look at calculating this probability that the t-statistic-- or of getting a greater than-- a larger value of that t-statistic.

And with this, let's see. If we consider changing the scale here to just be beta hat j divided by sigma hat Cjj square root, this is our t-statistic, then we basically have a t-distribution for the outcome of the least squares estimate. And we're calculating how likely is it we would get as large or larger a t-statistic if the true regression parameters were 0 or not.

Now, in looking at these estimates, one of the challenges in part is that the scale of the parameter estimates varies depending on the units of the independent variables. And so in some problems, we have very diverse independent variables in the data set. And the units are a property of the data set, but they're not really an important part of the problem. And so what we can do is standardize the covariates by normalizing them to have mean 0 and standard deviation 1.

So if we have our X matrix equaling-- let's see. How do I want to do this? So if we have P columns in our X matrix, we can basically take X1 and transform it to X1 minus X bar 1 times the vector of ones, and divide through by the standard deviation of X1. And so we'll call this a random vector, Z1, of standardized values.

And if we do this, we're basically shifting the X's and rescaling the shifted values, it turns out that our regression model on the Z-scores or standardized scores are the same as the regression coefficients in the original units. So it's obvious, I guess.

But if we have y is equal to beta 1 X1 plus beta 2 X2 up to beta PXP plus an error vector. And we consider Zj is just simply equal to Xj minus X bar j times the vector of ones. And 1 over, say, sj, where sj is equal to the sample variance of the Xj's.

We basically can substitute in for each of the X's the formula here. So X bar J is equal to sjzj plus X bar J plus X bar J. So this times this is X bar minus little x bar. And this is times the vector of ones. So if we plug-in to our Xj values here, this, then we end up getting this equation with the same regression parameters coming into play.

Well, when we fit the scaled regression independent variables model, then notice that the t-values and the p values for these two regressions are identical. So if we pick, say, the largest t value, svi-- or that's not quite the largest lca vol. But let's do svi, 2.949.

If we scroll back to the original units, the t-statistics are identical across all the variables. And so are the p-values. So in terms of interpreting how important different variables are, we get the same judgments of what's important in terms of t-values and p-values. But what's particularly useful though, when we use the standardized covariates is that the magnitude of the coefficients, when we have the standardized scale, that corresponds to the impact of one standard deviation move of the independent variable.

So if we have an age value that's one standard deviation above the mean, it doesn't have much of an impact on the dependent variable. If we have the svi, a one standard deviation move above a z-score of 1, then it has a larger impact on the dependent variable. So standard deviation units it can be a very convenient way to rescale our data.

All right. And judging the quality of regression models, we can calculate fitted values and pair those with the observed values. And look at the scatter plot of observed versus fitted values. And this simple scatter plot will have a correlation statistic. And the square of that correlation statistic is actually called the multiple R-squared, or coefficient of determination.

If you're familiar with simple linear regression where you calculate correlation statistics, you can think of the squared correlation as being useful. In multiple regression, we generalize a single correlation statistic with the multiple R-squared coefficient. And so it basically tells us how predictable our dependent variable is, given the independent variables.

Now in terms of evaluating the assumptions of our regression model, there are regression diagnostics that we can apply to fitted models and there are a number of different important measures that we can apply. And in the stats package, there's this influence measures general function, which gives a table of different statistics. There's this R student for computing studentized residuals.

If we have epsilon hat, it is distributed as multivariate normal with mean 0 and covariance matrix I minus H times sigma squared. Our residuals may have very different variances because of this I minus H factor in the covariance. And so studentized residuals basically will divide the residuals by the square root of that variance and an estimate of that.

And that estimate of the variance that's in the divisor leads to non-normal but t-distributions for the epsilon hats. So we'll see that play out in a moment. And then there are other statistics, like how much does the regression parameter change if we include or exclude different data points, different cases in the data.

So let's just see some of the results. So here's the student residuals for this regression model. And we have a histogram, which is unimodal, symmetrical. Here is what's called a quantile plot of the residuals that-- where the quantile plot will follow a straight line if the data are consistent with the normal distribution model and the t-distribution for normalizing the residuals by estimates of the variance.

What's rather useful when you use these methods for different regression problems is that you'll see that there's sampling variability due to just the data set we're working with. If you were to collect a new data set under identical conditions, you would get different results. These red bands correspond to bands you might expect to see with variation from one example to another.

Now this here is a plot called the influencePlot in the car package, and it plots hat values versus studentized residuals. The hat values are the diagonals of the H matrix. And what's important about the hat values is that-- let's see.

If the hat values were equal to 1, so if, say, H sub ii equals 1, then we would have y hat i is actually equal to yi. And so the i-th case would be the only case in the data set that allows you to estimate that value. So that case would have very high influence.

So hat values generally are-- well, are close to p over n. And lower values correspond to lower influence. And studentized residuals are given by these magnitudes. And I believe what is drawn in this slide is-- yeah, it is the circle size is proportional to Cook's distance. And so let's see.

With Cook's distance, if we have beta hat equal to x transpose x inverse x transpose y, we can think of beta hat sub minus i, which is the least squares estimate if we were to exclude the i-th case. So this corresponds to excluding the i-th case.

Then what we have that this beta hat is distributed multinomial with a mean vector, The True beta and covariance matrix given by this. We can actually look at beta hat minus-- or beta hat i minus beta hat. And then consider the distance of this or the magnitude of the change in the beta. And Cook's distance will actually correspond closely to a Chi squared distribution because of the relationship.

So we're basically looking at the distance of beta hat i from the true beta and normalizing that by the covariance matrix. Now let's see. There's basically a whole slew of different diagnostics that we can graph. And I think it's very useful to have graphical methods to highlight what's perhaps important.

In looking at this plotting from the car package, or plot.lm, we can see how the residuals compare with the fitted values. And we don't like to see any systematic pattern in that plot. We'd like that just to be flat. How that will change sometimes is that as the fitted values increase, the residuals might increase in magnitude as well. So that would suggest that the residuals are dependent on the magnitude of the fitted values.

We can also look at the normal Q-Q plot to see whether the data are consistent with normal or not. And then one can look at the scale of the-- the scale of-- measures of the scale of the residuals.

So we can take the square root of the standardized residuals, or the magnitude, and see whether those have variation, depending upon whether the fitted values are small or large. Here it looks like there may be a nonlinear relationship there. When we discover relationships, then that leads us to refine our model assumptions.

Let's see. This residual plot's function is actually a function that looks at the sensitivity of our linear regression model to potential nonlinearity in the model dependence on independent factors. And so it actually tries to fit curvature terms into the residuals in the model. And if there's residual-- if there's curvature present in the residuals, then we might have a nonlinear model and need to expand with the nonlinear terms.

Let's see. Well, with this normal linear model, we can consider the assumptions underlying that minus normality. And those are called the Gauss-Markov assumptions. So if we have our regression model for our y vector and X matrix, we set that up as the conditional expectation is X beta and the covariance matrix is equal to sigma squared times the identity.

And with these Gauss-Markov assumptions, there's a really important theorem in linear models, which is that the least squares estimates of the regression parameters give us the best estimates of any linear combination of the true regression parameters.

So if we want to estimate a parameter theta, which is a linear combination of the regression parameters, then if the Gauss-Markov assumptions are satisfied, then plugging in the least squares estimates provides an unbiased estimator that is also the unbiased estimator with the smallest variability.

And these are called best linear unbiased estimates. And what's important-- well, I mean as it's framed here, it's quite general. But think about this. If we have the constants c1 through cp correspond to specific values of the explanatory three variables corresponding to a different case, then we can be estimating the true mean value for a given case of the x variables.

If we have two x values, x vectors for different cases, and we're interested in what's the difference in mean values for those two cases, then we can have these c's represent the difference in the x's. And we're estimating the difference in y values, the difference in mean values for different cases. So it really is a very, very general theorem.

And while it's a very wonderful theorem, it depends on the Gauss-Markov assumptions. So what we can do is consider generalizing the Gauss-Markov assumptions to, again, have 0 mean errors, but have the covariance matrix of the errors be given by sigma squared times a capital sigma matrix.

And this kind of covariance structure in error terms actually arises frequently in time series modeling, where we're ordering our data by time and maybe near values of the error that are observed at nearby times are more correlated than those at faraway times. And the covariant structure of those errors might be systematically represented by a multiple of a known sigma matrix.

Well, if this is the case, we can transform our original data by premultiplying by the inverse square root of sigma and do the same for our x matrix. And if we do that, we're basically transforming our model to the starred case where the Y star and X star, as we've represented here, the epsilon star will have mean 0, and the covariance matrix of the epsilon stars will actually be diagonal with constant variance.

So in this transformed case, we have the same regression parameters. We have the epsilon stars with the Gauss-Markov assumptions being satisfied. So by the Gauss-Markov theorem, we just can write directly our least squares estimate in terms of the X stars and Y stars. And when we write that out, it ends up giving us this formula for the generalized least squares estimate.

And so if we have variable errors where we know the relative variation of those errors, we can accommodate that and get best estimates with this generalized least squares formula. And importantly, the generalized least squares estimate basically does a weighted regression where we weight cases proportional to the inverse of the variance of the random variable.

So if this sigma matrix, in fact, were diagonal with different variances, we would be downweighting those cases with high variance and performing the least squares computations in a weighted way. So that's the generalized least squares estimate. All right. Well, we'll finish there for today and cover more regression next time.