[SQUEAKING]

[RUSTLING]

[CLICKING]

PETER KEMPTHORNE: All right. We'll get started. What I wanted to do today was to make sure we thoroughly cover the theory and application of regression modeling.

And so to begin, there's a recap, which is the theory for, if we have linear regression, and we have, in particular, a normal linear regression model, then we have these properties of beta hat being distributed as a multinormal random variable with mean the true regression parameter and covariance matrix given by the X transpose X inverse multiplied by sigma squared.

And we also have that epsilon hat is a multinormal distribution-- it's actually not of full rank, but it's in n dimensions-- with mean 0, a 0 vector, and the covariance matrix is sigma squared times the identity minus the hat matrix. And with these properties, we can show that they are independent random variables, if we have the normal model assumption.

So these are independent if the error vector is multinormal. And with that, we're able to get an estimate of the error variance very easily from the residual vector. The sum of squares of the residual vector can be solved for having expectation equal to n minus p sigma squared, and that leads to dividing the sum of squared residuals by n minus p to get an unbiased estimate.

Now, what's interesting mathematically is that because this estimate of the error variance depends on the residuals, which are independent of beta hat, we can then use this information to construct t-statistics for the least squares estimates.

So we have t-statistics for the j-th component of the least squares estimate, so beta hat sub j minus beta j divided by sigma hat Cjj where Cjj is the j-th diagonal entry of the variance-- inverse variance.

And actually, the Cjj should be equal to the square root of X transpose X inverse sub jj. So there's a square root missing there. And this distribution, under the normal linear model, has-- or this formula has a t-distribution. So this is distributed as a t-distribution with n-minus-p degrees of freedom, degrees of freedom equal to n minus p.

And this is true if we plug in the true beta j. So centering our least squares estimate at beta j allows us to write that out. So if we look at our beta hat value and consider it around the true beta j, we basically have a t-distribution for the outcome.

And why this becomes very useful is that if we want to test the hypothesis that beta j is equal to 0, then if this is 0, we can basically define an interval around 0 and consider rejecting the null hypothesis if our t hat with beta j equaling 0 in magnitude is greater than some critical value-- say, some constant C.

And so this t-statistic is commonly applied in judging different regression models. Of particular interest is whether factors can be excluded from the regression. So if a beta j component is equal to 0, that means we can just exclude the variable.

Now, interestingly, if we want to test whether multiple beta j's are equal to 0, a special case of that is just to say, let's assume that the first k beta j's are nonzero. We'll just assume that that's true. But let's test whether the beta j's for j above k are equal to 0.

And what we can do is calculate residuals under the two hypotheses. First, assuming that they're all nonzero, we get an RSS 1. But if we assume the submodel is true, we can get a second set of residuals.

And because the full model is more-- sort of, it minimizes the sum of squares further, we can look at the difference in residual sum of squares, normalize that by p minus k, and compare that with the residual sum of the squares under the full model divided by n minus p. So this is an estimate of sigma squared using the full model. This numerator is also an estimate of sigma squared based on the null hypothesis being true.

Now, what's interesting to highlight is that the F-test when k, the number of nonzero coefficients, is p minus 1, then this F-test corresponds to testing whether the p-th beta j is equal to 0. And so it turns out that if we have t hat p-- sorry-- t hat, yeah, p, of beta-- of the true beta, the beta with beta p equal to 0- this squared is actually equal to this F-statistic.

[LAUGHS] I'm just going to-- [LAUGHS] there's a huge rumbling in the room right now. And if anyone finds out that it's something serious we should be thinking about, please let me know, [LAUGHS] or let the whole class know.

It sort of sounds like an earthquake, which I experienced when I was at UC Berkeley. [LAUGHS] Anyway, all right, well, we'll continue, unless I get warnings from you all.

Anyway, it turns out that this t-squared statistic is actually equal to the F-test formalized here. So F-tests are very useful. In special cases, they're equivalent to the square of a t-test statistic.

Now, let's see, with the notes from last time, we went through these. We highlighted how, with generalized least squares-- this is another key takeaway in regression modeling. So, so generalized least squares, with generalized least squares, we have an error vector, which is not a diagonal identity times sigma squared, but it's some base matrix sigma times little sigma squared.

And with this setup, we can actually transform this regression model into one which satisfies our Gauss-Markov assumptions. And then, so with x star and y star being the model equation multiplied by the inverse square root of sigma, then this model satisfies the Gauss-Markov assumptions, zero-mean constant variance, and so this estimator, with the X stars and Y stars but the same beta, gives us our best estimate of the beta hat-- best in terms of minimum being unbiased and having smallest variance.

OK. Well, a next important aspect of regression modeling is looking at Maximum Likelihood Estimation, MLE. And for maximum likelihood estimation, we use the probability model of the regression-- which, in this case, is normal linear regression models, so if we have normal epsilon vector.

Then we compute the density of our data Y, given the explanatory variables in X, and the beta, regression parameter beta. That should be a beta there. And we simply identify those parameter values that maximize this probability. So we're looking at the parameter values that make the data we observed the most likely.

And intuitively, that's maybe a reasonable thing. If you take Mathematical Statistics next term, you'll learn that that, in fact, is an optimal estimator, in terms of having smallest variance in large samples. As the sample size, the number of cases n increases, this will lead to the best estimator.

And this MLE for the best-- or the best estimator being the MLE will be true for different distributions as well. We're going to show what it is for the Gaussian distribution. And what we can see on this slide is that the log of the likelihood is the log of this density on the top line. It can be equal to minus n over 2 log sigma squared minus 1/2 sigma squared q of beta, where q of beta is our least squares criterion.

So maximum likelihood estimation for normal linear regression models can proceed in two steps. First, we can see that the beta regression coefficient enters in only through a factor of Q of beta. So minimizing Q of beta gives us the MLE for beta.

And then once we have the MLE for beta, we can plug that in to the likelihood, the log likelihood-- or the likelihood, for varying sigma squared. And when we do this, we actually can just maximize this sort of partial likelihood with beta fixed at beta hat.

Maximize that over sigma squared, and that turns out to be a very easy computation, calculating the first-order equation for the log likelihood. And we get the maximum likelihood estimate is simply the sum of squared residuals divided by n.

Now, with maximum likelihood, in this case, we have, actually, a biased estimate of the error variance. And it's important just to note that that's the case.

All right. Well, what's interesting to generalize is to find estimators, which minimize a Q function, which may be different from least squares-- or sum of squares. And so we can consider a Q function which minimizes the sum of a function h evaluated at case i.

So let's see here. Let's pull this down. So, so with generalized M-estimators, we think of Q of beta being the sum 1 to n of h of y i, the x i beta, and sigma squared, I guess. And the choice of h, well, if we choose h to just be the square of the linear estimate of y given x i, that's least squares.

There's no reason, in principle, why we should use squares. We could use absolute deviations and sum those. And if we wanted maximum likelihood estimates, we could choose h to be minus the log of the density.

And there are robust estimators in statistics. And these robust estimators consider variations, basically, of least squares and mean absolute deviations as special cases, that yield estimators that are robust.

Now, what's, to me, very interesting is that if we were to know what the density is of the errors, then maximum likelihood estimation would say use that. When we don't know the distribution, then these others may be very useful as alternatives. And robust estimators were originally proposed for models where we may have a normal distribution most of the time, but some contamination distribution with large errors a small percentage of the time.

Now, and another interesting example of this generalized M-estimator is the quantile estimator, where we consider sort of mean absolute deviation, but we scale that for positive residuals by a factor tau. And for negative values of the residual, we scale the absolute deviation by 1 minus tau.

This h function basically is equivalent to the mean absolute deviation when tau is equal to 0.5. But when tau is-- if we have h-- let's see.

So h of y, x, when this is equal to tau y i minus x i beta when this is positive. And it's equal to 1 minus tau when this is-- actually, it's equal to minus this when this is less than 0.

Then, for example, when tau is equal to 0.9, this turns out to be a quantile estimate that is working to estimate the 90th percentile of outcomes. So when we have a regression model and we are sort of just-- if we have a bunch of points, and we consider a linear regression model, this would be y equal to-- or y hat equal to sort of the mean absolute deviation estimate, we actually can get estimates of the 90th percentile of outcomes of the regression model.

Now, what's interesting mathematically is, How would one calculate sort of the mean absolute deviation estimate, or this quantile estimator? Well, we would want to minimize the sum of these h functions. And these h functions actually are convex.

So as James Sheppard said, this is, like, a bowl-shaped problem. But it's a bowl with sort of square-- or with straight edges around it. And so if we take this Q function with this h, we can't minimize it by solving the first-order equations.

So if we have Q equal to the sum 1 to n of h of the arguments, we can't solve this-- we can't solve for this equaling 0. But we can solve for what the derivative is at any guesses, and we can move our guess to a lower value of the Q function with that. And that's, in fact, how we learn that the median of a sample is the best estimate of the center of the data if we consider mean absolute deviations.

OK. Well, let's consider another extension of regression. It's famous enough to have its own name, ridge regression. And so with ridge regression, we are estimating the regression parameters where we have-- basically, it's sort of least squares plus a penalty, for our criterion.

So the left argument of this expression is simply the sum of squared residuals, but then we're adding to that a penalty depending upon the squared length of the beta vector.

And so in order for this to be a reasonable thing to do, we actually need to-- or, we don't need to, but it's useful, to consider standardizing our independent variables x, making them have mean 0, and also centering y, and then consider the regression where we rescale the columns of the predictor matrix as centered with the inverse of the covariance matrix between those.

So in this formulation, this is the statement of how the ridge regression parameter estimate is specified. Now, the centering of the independent variables and scaling them to have unit variance is important because it places all of the variables, all of the predictor variables, on an equal footing.

If we are looking at the sum of squared beta j's, then if we just use different units, we could make any j beta, beta j, as large as we want. But with this standardization, we basically eliminate that aspect, so there's no dependence on the original units of X.

And let's see. With this ridge regression, the notes go through how we compute these. I want to highlight, though, how this ridge regression is actually related to Bayesian models, where we would assume a prior distribution on our regression parameters.

And so the-- if we consider a Bayes prior on beta, and that Bayes prior is basically multinormal in p dimensions, with a mean vector of 0, and a covariance matrix that is-- let's see, I'll call it omega, and omega is proportional to the identity matrix in p dimensions. Then what we've done is we have a criterion that corresponds to the exponential part of the log likelihood for-- plus the log likelihood of the prior density.

So this ridge regression is interpretable as putting a prior distribution on the betas that have this form, mean 0, and covariance proportional to the identity matrix. And what that suggests a priori is that any direction in that p-dimensional space is equally likely as any other for characterizing the regression parameters.

And this beta hat, which minimizes this argument, ends up being the mode of the posterior distribution. So that's a neat connection to be aware of.

The notes here go through, with ridge regression, how we compute them. The minimization basically leads to formulas for the ridge regression estimate that look almost like least squares, except we have this additional factor, lambda times the identity, inside that braces, before we take the inverse.

And with ridge regression, we can look at the singular value decomposition of our independent variables that have been centered and standardized. And when we do that we basically have this final formula for the ridge regression.

And we end up basically having the fitted values y hat just being equal to a sum of factors times cj uj, where these factors are 1 when lambda is 0. So when lambda is 0, ridge is equal to least squares.

And so the least squares fitted values are given here. But with ridge regression, we have shrinkage. And what's curious is that there's less shrinkage for the larger squared singular values, and there's more shrinkage for the smaller ones.

So if we think of the independent variables, or predictor variables, being in p-dimensional space, then ridge regression will not shrink much those regression coefficients that correspond to regressor variables that are along sort of the first principal component axis.

And there's much more shrinkage in those dimensions where there's very little variation. So these are actually connected to principal components variables, which is actually covered sort of in the next section.

So, yeah, it was-- we introduced ridge regression using the singular value decomposition of X. But we can motivate principal components regression using the same singular value decomposition.

So with our data predictor matrix X, we consider the sample covariance matrix of those row vectors. And we can compute the eigenvalue eigenvector decomposition of that. And with these eigenvectors, which are normalized to have length 1, we simply multiply our matrix X by the j-th eigenvector, and we get our j-th principal component variable.

So this is familiar to us in, terms of how principal components sort of computations work. And with principal components regression, one has principal component variables which are orthogonal to each other, and so the computation of their coefficients is very simple. We can calculate them separately, and we can consider using maybe just the first m principal component variables in our regression.

And so if we do this, we can see some easy formulas for the estimates. Ultimately, we get fitted values of y, from principal components regression, which are a projection of the Y vector onto the space spanned by the first-- in this case, the first m principal component directions.

And what's curious is just how these three methods-- least squares, ridge, and principal components regression-- are estimating the response variable, or dependent variable, using essentially the same fundamental pieces, the cj and uj's, but are either skipping the last p minus m for principal components, shrinking those based on the size of the eigenvalues for ridge, or using all of them.

Now, as you might expect, researchers have extended estimation methods beyond these. The LASSO regression model, or regression estimate, is one where, instead of having the penalty be the sum of squared beta j's or the length of the beta j, it's the sum of the absolute values.

And this penalty is a really interesting one. And some of-- let's see. I'm sure some people here have perhaps used lasso regression. Has anyone here used lasso regression?

Nobody? That's fine. That means this isn't sort of repeat information.

But it turns out that-- well, let me just draw this. Suppose we have beta 1 versus beta 2, and we're considering penalties. So the sum of the beta j squareds equal to a constant, this is the sort of ridge penalty.

So, so if we have a beta hat, say, a beta hat here, a point that we observe, it basically is sort of shrunk towards zero. Or we have beta hat, and the beta-hat ridge will be shrunk towards this prior mean some amount.

With the lasso penalty, we have the sum of the magnitudes of the beta j's is equal to a constant. Then we basically have diamonds where, along each of these diamond edges, the penalty is the same magnitude.

And so with lasso regression, we basically consider finding an estimate that minimizes the least squares but also is constrained to have sort of sum of absolute magnitudes that's smaller. And so this kind of penalty ends up actually leading to shrinkage, which tends to concentrate at the vertices of this penalty function.

So what we end up getting, often, is a lasso estimate that has some null parameters, zero-value parameters. So it's a way of excluding variables from the regression, depending on what lambda is.

All right. Well, let's see. There's this ETF case study that I posted which-- and it's sort of a long document to sift through, but I guess what's important is just to look at the table of contents to see what's of interest to check out in this. And what I set up as a problem was to consider collecting prices on exchange-traded funds that correspond to different sectors in the US market.

And so in the US market, there are about 10 sectors-- consumer staples, energy, financials, health, and so forth. And then there are also exchange-traded funds that correspond to market indexes, like the S&P 500 or the Dow Jones Industrial or the NASDAQ market.

And I consider regressing a sector ETF on the index ETFs. Now, why would we be interested in regressing a sector index on these market-index ETFs?

AUDIENCE: [INAUDIBLE].

PETER KEMPTHORNE: Well, let me just go through here. OK. Or, this document goes through, listing the symbols for all of these different ETFs, plotting them, and actually then converting the prices to weekly returns, and then looking at how they're correlated with each other.

So in this example, I believe it was the Consumer Staples Sector Fund, that I'm thinking of regressing on the SPY, which corresponds to the S&P 500; MDY, which is a mid-cap index; QQQ; and the Diamonds. And if we look at these sort of cumulative returns of these different ETFs, here's the cumulative return over a 14-year period or something. Well, any other-- any more thoughts on why we might want to regress this, on these indexes? AJ?

AUDIENCE: The index return should be-- the entire variance should be explained by their sectors. Assuming everything in each index is contained within the market index, it should explain perfectly the variance in returns.

PETER KEMPTHORNE: Right. OK, so that's looking at it almost the other way around. If we were looking at the index ETF as a function of the sectors, we should have almost a perfect explanation of that in a regression model. What can be useful with this regression is if we wanted to invest in consumer staples but we wanted to have our investment be hedged against the market risk, that is large.

And so if we could eliminate the dependence on the S&P 500, eliminate the dependence on the mid-cap and the QQQs, then our hedged investment, we'd be less concerned about major drops in the market. So we'd be reducing the market risk. So this is where it can be very useful.

Now, this kind of approach has also been used to try and mimic hedge fund strategies. And so there's some papers out there on hedge fund replication, where hedge funds basically maybe have most of their returns attributable to liquid market instruments that one can trade in a portfolio. And so hedge fund replication methods have actually done that kind of regression as well.

Anyway, with this, with this setup, we can just regress this XLP on the index ETFs, and we get this regression coefficient table here. What's important to see is whether the regression coefficients are significantly different from zero or not. And the t-value and p-value columns basically tell us that all of these explanatory factors are useful here.

And let's see, when we study the-- let's see. When we study the regression model-- OK, we have the R-squared of this multiple regression.

The correlation of the fitted and actual values is given by-- or the square of that is given by the R-squared. And we can do various regression diagnostics which are used here.

What-- I'll let you go through these details. It's a bit repetitive from what we talked about in the last lecture. But let's see. We can also model the independent variables, or do a principal components analysis of those.

And if we compute a principal component analysis of our independent variables-- in this case, the index sector exchange-traded funds-- we then have a sort of decomposition of variability, which indicates that the first principal component variable explains 90% of the variability in these data, the first two, 97%. So we're actually explaining a lot of variability with just the first two.

Now, in looking at the regression on the PC variables, let's see, we can consider the-- well, each of these PC variables, Principal Component variables, are orthogonal to each other. So in estimating the regression coefficients, we can do so with simple linear regressions of the ETF fund, y on that, on each variable.

And so the first principal component variable has an R-squared of 0.545. And the second principal component variable has almost no correlation. And so it's-- well, it's almost none. Its p-value is still 0.0518, which is almost statistically significant.

With very large samples, we're able to judge small correlations as being statistically significant, but maybe it's not practically significant. And then the third has a small R-squared as well, and the fourth, again, rather small.

But what's of interest is whether the-- or, it's basically just, How statistically significant are the principal component variables? And in the sort of updated example, the-- let's see here. Well, if we consider the regression parameter based on only the first three variables, we basically get these results.

Now, when I did this a few years ago, it ended up being the case that the p-values for-- I think it was the second principal component variable, was not significant at all.

And so in choosing what factors to include in the model, the principal components regression approach of basically including successive high-eigenvalue principal component variables in successively so long as they're significant, that actually sort of doesn't include, really, necessarily, the important regressors in the model. So we can have statistically significant high-order principal component variables that don't have much variability in them, but they actually are explaining significant factors affecting the response variable.

So anyway, this note goes through and additionally considers ridge regression and lasso regression. And so with the ridge regression results, the display of regression parameters is traced out as we consider different constraints on the L1-- or on the-- well, on the L2 norm.

And if we look at the L1 norm of these estimates, one can see, basically, how as one sort of decreases the penalty for the complexity of the regression parameters, we get final estimates given by these ending values. But if we increase the penalty on those, then they basically are shrinking towards zero. Although, the shrinkage isn't necessarily in a monotone manner.

And then if we look at the lasso regression, with a same L1-norm sort of scaling, one can see how, as that lambda parameter associated with the penalty is increased, we basically have the L1 norm getting smaller, of the resulting estimates. And this shows you how-- I guess the third variable is the first to be excluded, then the second, and then, finally, the fourth.

So, all right. So anyway, so this note goes through an illustration of lasso and principal components regression. And it also provided a comparison of different coefficients by method. And what this final chart shows is the comparison of beta estimates by method. And one can see that most of the methods give similar results, but using just the first three principal component variables gives us very different estimates of regression parameters.

So there's sensitivity of the estimators to the data. And it's perhaps useful to-- let's see. Well, it's useful to consider these alternatives. With very large sample sizes, you generally don't get much difference in the ridge or the Lasso results. It's usually in smaller sample sizes that you get differences.

With the principal components regression, in this example, the first three principal component variables end up giving estimates, that are in the yellow. But if we chose which principal component variables to include based on their statistical significance, then the more significant principal component variables ended up leading to this green values, which are comparable to the others.

So this suggests that-- well, actually in this case, the green values are equal to least squares, just because all of the principal component variables were statistically significant.

All right. Well, the next thing I wanted to go through is this linear regression modeling for the capital asset pricing model. And in this note, we first review what the capital asset pricing model is, how it relates to returns on stocks and returns on market index.

And what we're going to do is go through the computation, computations, involved with an empirical analysis of this capital asset pricing model, and then consider testing hypotheses about individual coefficients, and then conclude with actually fitting this capital asset pricing model to all stocks in the S&P 500 index, and comparing the results of the capital asset pricing model parameters for that.

So with this, we begin with basically a description of what the capital asset pricing model is. We went through this, I think, in our third lecture, which says that in an efficient market, the expected return of a stock, asset j, should equal the risk-free return plus a beta factor times the excess return of the market.

And so beta j is sort of the market risk factor. And the excess return of the market portfolio is simply how much bigger that market return is compared with the risk-free rate.

Now, with these parameters, if we have data consisting of empirical returns R for stocks j at different times t, different values over t of the risk-free rate, and we consider the excess return of each asset and the excess return of the market, then this data can be used to specify a linear regression model where we have an intercept alpha j and a slope beta j for the regression.

So we're looking here at R star market, and this is R star j. And we'll plot values over time t of these excess returns. And we can basically fit a regression model where we have y hat-- or R star jt is equal to alpha j plus beta j R star mt.

And if the capital asset pricing model holds, then the alpha j's should be 0, because the alpha j's-- if we take the expectation in this equation with residuals that are mean 0 constant variance, then we get the capital asset pricing model equation when the alpha j is equal to 0.

So what we can do with real data is fit this model and then test whether alpha j is 0 or not. And if it is consistent with a 0 value, then the pricing of the asset can be judged to be sort of in an equilibrium described by the capital asset pricing model.

And so let's take a look. I think what's perhaps useful is just to see the code for doing these computations. It's included in the note here.

But if we pick a particular stock, GE stock, and fit the regression model, we basically-- in this code, the only thing we need to do is specify the symbol for the stock, and then extract from our data object the excess market return and the excess asset return. And then fit that regression, and we get these coefficients table here.

Now, in this example, the intercept coefficient is the estimate of alpha. And if GE is being priced consistent with the capital asset pricing model, then that estimate should be consistent with the null value. And the p-value for testing whether that intercept is 0 or not gives us evidence supporting the capital asset pricing model.

Now, the market, the excess return of the market has an estimate that's about 1, with some standard error. It has a huge t-value and a huge p-value. Those values being large, or a huge t-value and very small p-value, are not surprising at all, just because they're correlated. And a 0 value for the slope is sort of meaningless here.

What could be useful is testing whether our beta coefficient is consistent with sort of the average market risk of beta equal to 1, or not. But here's basically a plot of the linear regression model for GE, and a discussion just of getting comfortable with R-squared values and calibrating our views of scatter plots to those values.

And then there's residuals analysis, that is very important. As I introduced at the beginning of our regression study, we make assumptions about our model. And we then, once we fit the model, we check our assumptions to see whether those are satisfied or not.

And in this case, we might assume that the residuals are normally distributed. And here's a histogram of the residuals from this GE regression. And there actually are two bell-shaped curves here.

There's one, that's in green, that corresponds to the maximum likelihood estimate of the mean and variance of that normal. But there's also a blue curve, which corresponds to a robust estimator of the normal distribution parameters.

So, and we just talked a bit about normal linear regressions and their being optimal. Well, they're optimal so long as the normal distribution is applied. And one can calculate what's called a QQ plot, of the residuals.

Where the theoretical distribution is normal, we can think of ordered samples of normals, from smallest to largest, and pair those with our ordered residuals, from smallest to largest. And this plot will look like a straight line if the residuals are Gaussian.

And the green line corresponds to the MLE fit of the Gaussian. Basically, the slope will equal the standard deviation of the distribution, and the median roughly will correspond to the mean value. But what we have is a robust estimate of the standard deviation of the residuals, gives us this blue line.

And the robust estimate of the residual variance is based on the interquartile range of the residuals. So the middle 50%, what kind of standard deviation would match the middle 50% range of the residuals? And there's an argument that this robust estimate might in fact be better.

And another way of evaluating the residuals is to compute fitted percentiles under the residual model. So if we observe an epsilon hat with mean mu hat and standard deviation sigma hat, we can do the inverse of the probability integral transform and get the percentile of that outcome.

And so it turns out that these percentiles should be uniformly distributed, from, like, 1 to 99, if this normal model is correct. And so we can actually consider different estimates of sigma hat, from robust to MLE.

And with the MLE, here's the histogram of the fitted percentiles. It's basically underestimating percentiles that are close to the mean-- or the median. And so it really isn't fitting the residuals very well across the whole range.

And if instead we consider fitted percentiles from the robust estimate of the variance, then the fitted percentiles of the residuals is consistent with uniform in the middle of the data, except in the top 1% and the bottom 1%. So it's very possible that our data has a non-normal residual distribution or error distribution. And perhaps that non-normal distribution could be well modeled with a mixture model.

Now, OK, there's regression diagnostics, which I'll let you read through. And then there's this section on testing hypotheses about individual model coefficients. Let's see. If you have this on your computer, you can probably see it a little better than seeing this screen here.

But with the normal-- or with our testing of individual coefficients, we basically use the fact that our least squares estimator is multivariate normal with mean equal to the true beta vector and covariance matrix, given here. So, so we can make various tests of different regression parameters.

So we have the t-statistics for testing whether a parameter value is 0 or not, and use the t-distribution to judge how statistically significant that is. In our capital asset pricing model case, we actually are interested in testing whether the alphas are zero or not. And one could also consider testing for whether the beta factor is equal to 1 or not.

1, for the beta, is actually what the average beta is across all stocks in the market. And so one can consider judging whether assets sort of have significantly high risk, high betas, or low betas, using a test of whether the estimated regression coefficient is significantly different from 1 or not.

And so with this, I've basically computed this for GE. And it turns out that, I guess, the t-stat for GE, which corresponds to a beta of 1.083, that's fairly close to 1, in terms of absolute magnitude, but it's actually a bit-- it's statistically significant, according to the arguments given here.

Now, there are many packages in R, but one of them is the CAR package, which allows one to conduct tests of linear hypotheses like these. And so it goes through and performs the same kinds of tests for whether the intercept is 0 or not. And in implementing these hypothesis tests, this CAR package uses the sort of F-test approach. So one has results that give us different residual sum of squares and F-statistics, and p-values for the F-statistics corresponding to these.

Now, let's see. With-- let's see. Before showing the results for all the stocks in the S&P 500, there's another note on regression analysis hypothesis testing.

And what's interesting to think about is we can consider testing whether a submodel, a model without as many factors in it, is suitable. But one can also test for whether there's a change in model parameters with a regime, so that maybe there's an initial period with a given set of model parameters, and then a regime shift following that, where the parameter values have changed.

And so the mathematical theory for this is covered in section 3, and one can consider testing-- with the capital asset pricing model, maybe-- whether there's a change in regression parameters over an entire period. And so what I want to just go to is the results. Actually, the linear algebra for these hypothesis testing methods is very straightforward. And I encourage you to read through these, just to see how they play out.

But let's see. What's interesting, I guess, is, with this change between two periods, if we consider sort of an original regression model, y equals X beta plus epsilon for a very long period, we might consider splitting it up into the first part, period A, and the second part, period B, and have two equations for the respective sets of data.

And in thinking about the two periods, we basically can test whether the regression parameters for the second period are the same as the regression parameters for the first period. And we can write out the model equations with appropriate matrices. And then we basically have an F-test statistic for seeing whether these regression parameters are different or not.

And one can also-- instead of considering separate parameters for each period, one can consider a beta value here for the entire period, and then use that beta value together with the change in parameter values in period B from period A. And so one can set up the regression model where one is interested in whether this change parameter delta is equal to 0 or not. And so, mathematically, we get F-test from that.

Now, with the GE stock I subjectively chose this period split for A from B. If we look at the cumulative sum of standardized residuals from the regression model, then we basically have residuals that are consistently negative, cumulating to a lower value around 2019, and then later expanding beyond that. So this sort of time series of accumulated residuals sort of highlights that there's strong time dependence in the residuals.

But if we fit the regression models to these-- separate regression models to these two periods, we end up getting highly significant results for the change in the two periods. There's basically an intercept, which is the alpha, which changes from negative to positive, and the market risk basically changes from 0.773 beta in the first period, to 0.773 plus 0.3537 in the second period.

So these capital asset pricing models are not expected, perhaps, to hold over really, really long periods of time. And if we wanted to apply them, we'd likely sort of choose a rolling window approach perhaps, which then leads to questions of, how do we decide on what rolling window is appropriate, and opens up other issues to study.

So anyway, so with that background for hypothesis testing, let's go back to fitting the capital asset pricing model to all stocks in the S&P 500. Let's see. The code here basically fits this capital asset pricing model regression for all the stocks that have data going back several years.

So there were, I think, around 380 or 400 stocks. And they're grouped by sector. And we can see how the R-squareds of the regressions vary from 0 to 1. They typically are above 0.2.

So there's some reasonable R-squared relationships that are being explained. And if we look at the alphas by sector-- here's sort of parallel box plots of the alphas, the alpha estimates within each of the sectors. And so this highlights how some-- let's see-- some sectors, like, I guess, construction, sort of had positive alphas, as did computer technology.

Let's see. The conglomerates tended to have negative alphas there. And we can calculate the p-values for testing whether the alphas are consistent with 0 or not. And this graph here shows how the p-values for almost all the stocks exceed the p-value threshold of 0.05.

So if we can't reject the null hypothesis of alpha equaling 0, the p-value is above 0.05. And one can see that there's basically sort of a handful, or a couple of handful, of stocks that have a significant alpha. And if we use a 5% threshold, well, we'd expect 5% of the results to be statistically significant, when there's no statistical significance. And so maybe these, in fact, are consistent with the capital asset pricing model pretty well.

Now, let's see. One can plot the alpha versus the beta across all the stocks and get this result, which suggests that when you have higher beta, you may get higher alpha as well. And when you have very low beta, maybe the low beta is partly due to a lower alpha potential.

But this chart doesn't show how alpha varies by sector. And so if we try to look at alpha by sector we get different fits for different sectors, which is too complex, really, to interpret as graphed here. But it serves as a first step at trying to understand how alphas might vary within different sectors.

Perhaps interestingly, here's the beta coefficient by sector. And here, one can see that certain sectors, like consumer staples, have much lower beta values. And other sectors, like computer and technology, have very high betas. So in terms of identifying those sectors that will have low market risk or high market risk, these properties play out quite well.

And let's see. In terms of looking at the top and bottom stocks by beta, well, we have consumer discretionary, computer technology sector stocks are the sectors where the top 10 betas are. And if we look at the lowest betas, ranging from 0.37 down to 0.18, most of these are in consumer staples.

Finally, if we look at the top 10 stocks with significant alphas, OK, here is the table of alpha values that were highest, and one was in oils and energy sector ENPH, with the alpha value of 0.003.

Let's see. Here's-- so if we have an alpha value of 0.003, it's actually-- yeah, so 3-- whatever. I won't have the extra.

But this is a daily alpha. If we wanted to annualize this daily alpha, what would that-- how would we compute that annualized alpha?

If you're earning this alpha every day, it sort of is an increment of return that you're expecting to get every day. And so if we have 1 plus alpha hat to the number of days-- there's roughly 252 days per year-- and then subtract 1, then we end up getting-- let's see. I computed this.

AUDIENCE: [INAUDIBLE].

PETER KEMPTHORNE: It's 2.12 minus 1, basically. So this is, like, a hundred percent. So this stock happened to just deliver a huge amount of returns.

But what's interesting to see, though, is that this relatively simple model, the capital asset pricing model, appears to be consistent with the stocks over a long period of time. That being said, the more complex models are warranted. And the data collection phase of this, our project, included importing Fama-French factors that are extensions of sort of the market risk that can be used in these models, so.