The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: OK. Well, last time I was lecturing, we were talking about regression analysis. And we finished up talking about estimation methods for fitting regression models. I want to recap the method of maximum likelihood, because this is really the primary estimation method in statistical modeling that you start with. And so let me just review where we were.

We have a normal linear regression model. A dependent variable y is explained by a linear combination of independent variables given by a regression parameter, beta. And we assume that there are errors about all the cases which are independent identically distributed normal random variables.

So because of that relationship, the dependent variable vector, y, which is an n vector for n cases, is a multivariate normal random variable. Now, the likelihood function is equal to the density function for the data. And there's some ambiguity really about how one manipulates the likelihood function.

The likelihood function becomes defined once we've observed a sample of data. So in this expression for the likelihood function as a function of beta and sigma squared, we're considering evaluating the probability density function for the data conditional on the unknown parameters.

So if this were simply a univariate normal distribution with some unknown mean and variance, then what we would have is just a bell curve for mu centered around a single observation, y, if you look at the likelihood function and how it varies with the underlying mean of the normal distribution. So this likelihood function is-- well, the challenge really in maximum estimation is really calculating and computing the likelihood function.

And with normal linear regression models, it's very easy. Now, the maximum likelihood estimates are those values that maximize this function. And the question is, why are those good estimates of the underlying parameters? Well, what those estimates do is they are the parameter values for which the observed data is most likely.

So we're able to scale the unknown parameters by how likely those parameters could have generated these data values. So let's look at the likelihood function for this normal linear regression model. These first two lines here are highlighting-- the first line is highlighting that our response variable values are independent. They're conditionally independent given the unknown parameters.

And so the density of the full vector of y's is simply the product of the density functions for those components. And because this is a normal linear regression model, each of the y i's is normally distributed. So what's in there is simply the density function of a normal random variable with mean given by the beta sum of independent variables for each i, case i, given by the regression parameters.

And that expression basically can be expressed in matrix form this way. And what we have is the likelihood function ends up being a function of our q of beta, which was our least squares criteria. So the least squares estimation is equivalent to maximum likelihood estimation for the regression parameters if we have a normal linear regression model.

And there's this extra term, minus n. Well, actually, if we're going to maximize the likelihood function, we can also maximize the log of the likelihood function, because that's just a monotone function of the likelihood. And it's easier to maximize the log of the likelihood function which is expressed here.

And so we're able to maximize over beta by minimizing q of beta. And then we can maximize over sigma squared given our estimate for beta. And that's achieved by taking the derivative of the log likelihood with respect to sigma squared.

So we basically have this forced order condition that finds the maximum because things are appropriately convex. And taking that derivative and solving for zero, we basically get this expression. So this is just taking the derivative of the log likelihood with respect to sigma squared.

And you'll notice here I'm taking the derivative with respect to sigma squared is a parameter, not sigma. And that gives us that the maximum likelihood estimate of the error variance is q of beta hat over n. So this is the sum of the squared residuals divided by n. Now, I emphasize here that that's biased. Who can tell me why that's biases or why it ought to be biased?

AUDIENCE: [INAUDIBLE].

PROFESSOR: OK. Well, it should be n minus 1 if we're actually estimating one parameter. So if the independent variables were, say, a constant, 1, so we're just estimating a sample from a normal with mean beta 1 corresponding to the unit's vector of the x, then we would have a 1 degree of freedom correction to the residuals to get an unbiased estimator.

But what if we have p parameters? Well, let me ask you this. What if we had n parameters in our regression model? What would happen if we had a full rank in independent variable matrix and n independent observations?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Yes, you'd have an exact fit to the data. So this estimate would be 0. And so clearly, if the data do arise from a normal linear regression model, 0 is not unbiased. And you need to have some correction. Turns out you need to divide by n minus the rank of the x matrix, the degrees of freedom in the model, to get a biased estimate.

So this is an important issue, highlights how the more parameters you add in the model, the more precise your fitted values are. In a sense, there's dangers of curve fitting which you want to avoid. But the maximum likelihood estimates, in fact, are biased. You just have to be aware of that. And when you're using different software, fitting different models, you need to know whether there are various corrections be made for biasness or not.

So this solves the estimation problem for normal linear regression models. And when we have normal linear regression models, the theorem we went through last time-- this is very important. Let me just go back and highlight that for you.

This theorem right here. This is really a very important theorem indicating what is the distribution of the least squares, now the maximum likelihood estimates. Of our regression model? They are normally distributed. And the residuals, sum of squares, have a chi squared distribution with degrees of freedom given by n minus p.

And we can look at how much signal to noise there is in estimating our regression parameters by calculating a t statistic, which is take away from an estimate Its expected value, its mean, and divide through by an estimate of the variability in standard deviation units. And that will have a t distribution.

So that's a critical way to assess the relevance of different explanatory variables in our model. And this approach will apply with maximum likelihood estimation in all kinds of models apart from normal linear regression models. It turns out maximum likelihood estimates generally are asymptotically normally distributed. And so these properties here will apply for those models as well.

So let's finish up these nodes on estimation by talking about generalized m estimation. So what we want to consider is estimating unknown parameters by minimizing some function, q of beta, which is a sum of evaluations of another function, h, evaluated for each of the individual cases. And choosing h to take on different functional forms will define different kinds of estimators.

We've seen how when h is simply the square of the case minus its regression prediction, that leads to least squares, and in fact, maximum likelihood estimation, as we saw before. Rather than taking the square of the residual, the fitted residual, we could take simply the modulus of that. And so that would be the mean absolute deviation.

So rather than summing the squared deviations from the mean, we could sum the absolute deviations from the mean. Now, from a mathematical standpoint, if we want to solve for those estimates, how would you go about doing that? What methodology would you use to maximize this function?

Well, we try and apply basicaly the same principles of if this is a convex function, then we just want to take derivatives of that and solve for that being equal to 0. So what happens when you take the derivative of the modulus of y minus xi beta with respect to beta?

AUDIENCE: [INAUDIBLE].

PROFESSOR: What did you say? What did you say?

AUDIENCE: Yeah, it's not [INAUDIBLE]. The first [INAUDIBLE] derivative is not continuous.

PROFESSOR: OK. Well, this is not a smooth function. But let me just plot xi beta here, and yi minus that. Basically, this is going to be a function that has slope 1 when it's positive and slope minus 1 when it's negative. And so that will be true, component wise, or for the y. So what we end up wanting to do is find the value of the regression estimate that minimizes the sum of predictions that are below the estimate plus the sum of the predictions that are above the estimate given by the regression line.

And that solves the problem. Now, with the maximum likelihood estimation, one can plug in minus log the density of yi given beta x and sigma i squared. And that function simply sums to the log of the joint density for all the data. So that works as well.

With robust m estimators, we can consider another function, chi, which can be defined to have good properties with estimates. And there's a whole theory of robust estimation-- it's very rich-- which talks about how best to specify this chi function. Now, one of the problems with least squares estimation is that the squares of very large values are very, very large in magnitude.

So there's perhaps an undue influence of very large values, very large residuals under least squares estimation and maximum [INAUDIBLE] estimation. So robust estimators allow you to control that by defining the function differently. Finally, there are quantile estimators, which extend the mean absolute deviation criterion.

And so if we consider the h function to be basically a multiple of the deviation if the residual is positive and a different multiple, a complimentary multiple if the derivation, the residual, is less than 0, then by varying tau, you end up getting quantile estimators, where what you're doing is minimizing the estimate of the tau quantile.

So this general class of m estimators encompasses most estimators that we will encounter in fitting models. So that finishes the technical or the mathematical discussion of regression analysis. Let me highlight for you-- there's a case study that I dragged to the desktop here. And I wanted to find that.

Let me find that. There's a case study that's been added to the course website. And this first one is on linear regression models for asset pricing. And I want you to read through that just to see how it applies to fitting various simple linear regression models. And enter full screen.

This case study begins by introducing the capital asset pricing model, which basically suggests that if you look at the returns on any stocks in an efficient market, then those should depend on the return of the overall market but scaled by how risky the stock is. And so if one looks at basically what the return is on the stock on the right scale, you should have a simple linear regression model.

So here, we just look at a time series for GE stock in the S&P 500. And the case study guide through how you can actually collect this data on the web using r. And so the case notes provide those details. There's also the three-month treasury rate which is collected. And so if you're thinking about return on the stock versus return on the index, well, what's really of interest is the excess return over a risk free rate.

And the efficient markets models, basically the excess return of a stock is related to the excess return of the market as given by a linear regression model. So we can fit this model. And here's a plot of the excess returns on a daily basis for GE stock versus the market.

So that looks like a nice sort of point cloud for which a linear model might fit well. And it does. Well, there are regression diagnostics, which I'll get to-- well, there are regression diagnostics which are detailed in the problem set, where we're looking at how influential are individual observations, what's their impact on regression parameters.

This display here basically highlights with a very simple linear regression model what are the influential data points. And so I've highlighted in red those values which are influential. Now, if you look at the definition of leverage in a linear model, it's very simple. A simple linear model is just those observations that are very far from the mean have large leverage.

And so you can confirm that with your answers to the problem set. This x indicates a significantly influential point in terms of the regression parameters given by Cook's distance. And that definition is also given in the case notes.

AUDIENCE: [INAUDIBLE].

PROFESSOR: By computing the individual leverages with a function that's given here, and by selecting out those that exceed a given magnitude. Now, with this very, very simple model of stocks depending on one unknown factor, risk factor given the market, in modeling equity returns, there are many different factors that can have an impact on returns.

So what I've done in the case study is to look at adding another factor which is just the return on crude oil. And so-- I need to go down here. So let me highlight something for you here With GE's stock, what would you expect the impact of, say, a high return on crude oil to be on the return of GE stock? Would you expect it to be positively related or negatively related?

OK. Well, GE is a stock that's just a broad stock invested in many different industries. And it really reflects the overall market, to some extent. Many years ago, 10, 15 years ago, GE represented maybe 3% of the GNP of the US market. So it was really highly related to how well the market does.

Now, crude oil is a commodity. And oil is used to drive cars, to fuel energy production. So if you have an increase in oil prices, then the cost of essentially doing business goes up. So it is associated with an inflation factor. Prices are rising.

So if you can see here, the regression estimate, if we add in the factor of the return on crude oil, it's negative 0.03. And it has a t value of minus 3.561. So in fact, the market, in a sense, over this period, for this analysis, was not efficient in explaining the return on GE crude oil is another independent factor that helps explain returns.

So that's useful to know. And if you are clever about defining and identifying and evaluating different factors, you can build factor asset pricing models that are very, very useful for investing and trading. Now, as a comparison to this case study, also applied the same analysis to Exxon Mobil. Now, Exxon Mobil is an oil company. So let me highlight this here.

We basically are fitting this model. Now let's highlight it. Here, if we consider this two factor model, the regression parameter corresponding to the crude oil factor is plus 0.13 with a t value of 16. So crude oil definitely has an impact on the return of Exxon Mobil, because it goes up and down with oil prices.

This case study closes with a scatter plot of the independent variables and highlighting where the influential values are. And so just in the same way that with a simple linear regression it was those that were far away from the mean of the data were influential, in a multivariate setting here, it's bivariate, the influential observations are those that are very far away from the centroid.

And if you look at one of the problems in the problem set, it actually goes through and you can see where these leveraged values are and how it indicates influences associated with the Mahalanobis distance of cases from the centroid of the independent variables. So if you're a visual type mathematician as opposed to an algebraic type mathematician, I think these kinds of graphs are very helpful in understanding what is really going on.

And the degree of influence is associated with the fact that we're basically taking least squares estimates, so we have the quadratic form associated with the overall process. There's another case study that I'll be happy to discuss after class or during office hours. I don't think we have time today during the lecture. But it concerns exchange rate regimes.

And the second case study looks at the Chinese Yuan, which was basically pegged to the dollar for many years. And then I guess through political influence from other countries, they started to let the Yuan vary from the Dollar, but perhaps pegged it to some basket of securities of currencies. And so how would you determine what that basket of currencies is?

Well, there are regression methods that have been developed by economists that help you do that. And that case study goes through the analysis of that. So check that out to see how you can get immediate access to currency data and be fitting these regression models and looking at the different results and trying to evaluate those.

So let's turn now to the main topic-- let's see here-- which is time series analysis. Today in the rest of the lecture, I want to talk about univariate time series analysis. And so we're thinking of basically a random variable that is observed over time and its discrete time process. And we'll introduce you to the Wold representation theorem and definitions of stationarity and its relationship there.

Then, look at the classic models of how autoregressive moving average models. And then extending those to nonstationarity with integrated autoregressive moving average models. And then finally, talk about estimating stationary models and how we test for stationarity.

So let's begin from basically first principles. We have a stochastic process, a discrete time stochastic process, x, which consists of random variables indexed by time. And we're thinking now discrete time. The stochastic behavior of this sequence is determined by specifying the density or probability mass functions for all finite collections of time indexes.

And so if we could specify all finite dimensional distributions of this process, we would specify this probability model for the stochastic process. Now, this stochastic process is strictly stationary if the density function for any collection of times, t1 through tm, is equal to the density function for a tau translation of that.

So the density function for any finite dimensional distribution is stationary is constant under arbitrary translations. So that's a very strong property. But it's a reasonable property to ask for if you're doing statistical modeling. And what do you want to do when you're estimating models? You want to estimate things that are constant. Constants are nice things to estimate.

And parameters [INAUDIBLE] are constant. So we really want the underlying structure of the distributions to be the same. That was strict stationarity, which requires knowledge of the entire distribution of the stochastic process. We're now going to introduce a weaker definition, which is covariance stationarity. And a covariant stationary process has a constant mean, mu, a constant variance, sigma squared, and a covariance over increments tau, given by a function gamma of tau that is also constant.

Gamma isn't a constant function, but basically for all t, covariance of xt, xt plus tau is this gamma of tau function. And we also can introduce the autocorrelation function of the stochastic process, rho of tau. And so the correlation of two random variables is the covariance of those random variables divided by the square root of the product of the variances.

And [INAUDIBLE] I think introduced that a bit. in one of his lectures, where we were talking about the correlation function. But essentially, the correlation function is if you standardize the data or the random variables to have mean 0-- so subtract off the means and then divide through by their standard deviations. So those translated variables have mean 0 and variance 1. Then the correlation coefficient is the covariance between those standardized random variables.

So this is going to come up again and again in time series analysis. Now, the Wold representation theorem is a very, very powerful theorem about covariant stationary processes. It basically states that if we have a zero mean covariance stationary time series, then it can be decomposed into two components at a very nice structure.

Basically, xt can be decomposed into vt plus st. vt is going to be a linearly deterministic process, meaning that past values of vt perfectly predict what vt is going to be. So this could be like a linear trend or some fixed function of past values. It's basically a deterministic process. So there's nothing random in vt. It's something that's fixed without randomness.

And st is a sum of coefficients, psi i times eta t minus i, where the eta t's are linearly unpredictable white noise. So what we have is st is a weighted average of white noise with coefficients given by the psi i. And the coefficients psi i such that psi 0 is 1. And the sum of the squared psi i's is finite.

And the white noise eta t-- what's white noise? It has expectation zero. It has variance given by sigma squared that's constant. And it has covariance across different white noise elements that's 0 for all t and s. So eta t's are uncorrelated with themselves, and of course, they are uncorrelated with the deterministic process. So this is really a very, very powerful concept.

If you are modeling a process and it has covariant stationarity, then there exists a representation like this of the function. So it's a very compelling structure, which we'll see how it applies in different circumstances. Now, before getting into the definition of autoregressive moving average models, I just want to give you an intuitive understanding of what's going on with the Wold decomposition.

And this, I think, will help motivate why the Wold decomposition should exist from a mathematical standpoint. So consider just some univariate stochastic process, some time series xt that we want to model. And we believe that it's covariant stationary. And so we want to specify essentially the Wold decomposition of that.

Well, what we could do is initialize a parameter p, the number of past observations, in the linearly deterministic term. And then estimate the linear projection of xt on the last p lag values. And so what I want to do is consider estimating that relationship using a sample of size n with some ending point, t0, less than or equal to t.

And so we can consider y values like a response variable being given by the successive values of our time series. And so our response variables, yj, can be considered to be x t0 minus n plus j. And define a y vector and a z matrix as follows.

So we have values of our stochastic process in y. And then our z matrix, which is essentially a matrix of independent variables, is just the lagged values of this process. So let's apply ordinary least squares to specify the projection. This projection matrix should be familiar now. And that basically gives us a prediction of y hat depending on p lags.

And we can compute the projection residual from that fit. Well, we can conduct time series methods to analyze these residuals, which we'll be introducing here in a few minutes, to specify a moving average model. We can then have estimates of the underlying coefficients psi and estimates of these residuals, eta t. And then we can evaluate whether this is a good model or not.

What does it mean to be an appropriate model? Well, the residual should be orthogonal to longer lags than t minus s, or longer lags than p. So we basically shouldn't have any dependence of our residuals on lags of the stochastic process that weren't included in the model. Those should be orthogonal.

And the eta t hats should be consistent with white noise. So those issues can be evaluated. And if there's evidence otherwise, then we can change the specification of the model. We can add additional lags. We can add additional deterministic variables if we can identify what those might be. And proceed with this process.

But essentially that is how the Wold decomposition could be implemented. And theoretically, as our sample gets large, if we're observing this time series for a long time, then well certainly the limit of the projections as p, the number of lags we include, gets large, should be essentially the projection of our data on its history. And that, in fact, is the projection corresponding to defining the coefficient's psi i.

And so in the limit, that projection will converge and it will converge in the sense that the coefficients of the projection definition correspond to the psi i. And now if p goes to infinity is required, now p means that there's basically a long term dependence in the process. Basically, it doesn't stop at a given lag. The dependence persists over time. Then we may require that p goes to infinity.

Now, what happens when p goes to infinity? Well, if you let p go to infinity too quickly, you run out of degrees of freedom to estimate your models. And so from an implementation standpoint, you need to let p/n go to 0 so that you have essentially more data than parameters that you're estimating. And so that is required.

And in time series modeling, what we look for are models where finite values of p are required. So we're only estimating a finite number of parameters. Or if we have a moving average model which has coefficients that are infinite number, perhaps those can be defined by a small number of parameters. So we'll be looking for that kind of feature in different models.

Let's turn to talking about the lag operator. The lag operator is a fundamental tool in time series models. We consider the operator, L, that shifts a time series back by one time increment. And applying this operator recursively, we get, if it's operating 0 times, there's no lag, one time, there's one lag, two times, two lags-- doing that iteratively.

And in thinking of these, what we're dealing with is like a transformation on infinite dimensional space, where it's like the identity matrix sort of shifted by one element-- or not the identity, but an element. It's like the identity matrix shifted by one column or two columns. So anyway, inverses of these operators are well defined in terms of what we get from them.

So we can represent the Wold representation in terms of these lag operators by saying that our stochastic process, xt, is equal to vt plus this psi of L function, basically a functional of the lag operator, which is a potentially infinite order polynomial of the lags.

So this notation is something that you need to get very familiar with if you're going to be comfortable with the different models that are introduced with ARMA and ARIMA models. Any questions about that?

Now relating to this-- let me just introduce now, because this will come up somewhat later. But there's the impulse response function of the covariant stationary process. If we have a stochastic process, xt, which is given by this Wold representation, then you can ask yourself what happens to the innovation at time t, which is eta a, how does that affect the process over time?

And so, OK, pretend that you are chairman of the Federal Reserve Bank. And you're interested in the GNP or basically economic growth. And you're considering changing interest rates to help the economy. Well, you'd like to know what an impact is of your change in this factor, how that's going to affect the variable of interest, perhaps GNP.

Now, in this case, we're thinking of just a simple covariant stationary stochastic process. It's basically a process that is a random awaited sum, a moving average of innovations, eta t. But the question is, basically an covariant stationary process could be represented in this form. And the impulse response function relates to what is the impact of eta t. What's its impact over time?

Basically, it affects the process at time t. That, because of the moving average process, it affects it at t plus 1, affects it at t plus 2. And so this impulse response is basically the derivative of the value of the process with the j previous innovation is given by psi j. So the different innovations have an impact on the current value given by this impulse response function.

So looking backward, that definition is pretty well defined. But you can also think about how does an impact of the innovation affect the process going forward. And the long run cumulative response is essentially what is the impact of that innovation in the process ultimately? And eventually, it's not going to change the value of the process.

But what is the value to which the process is moving because of that one innovation? And so the long run cumulative response is given by basically the sum of these individual ones. And it's given by the sum of the psi i's. So that's the polynomial of psi with lag operator, where we replace the lag operator by 1. We'll see this again when we talk about vector autoregressive processes with multivariate time series.

Now, the Wold representation, which is a infinite order moving average, possibly infinite order, can have an autoregressive representation. Suppose that there is another polynomial psi i star of the lags, which we're going to call psi inverse of L, which satisfies the fact if you multiply that with psi of L, you get the identity lag 0.

Then this psi inverse, if that exists, is basically the inverse of the psi of L. So if we start with psi of L, if that's invertible, then there exists a psi inverse of L, with coefficients psi i star. And one can basically take our original expression for the stochstic process, which is as this moving average of the eta's, and express it as this essentially moving averages of the x's.

And so we've essentially inverted the process and shown that the stochastic process can be expressed as an infinite order autoregressive representation. And so this infinite order autoregressive representation corresponds to that intuitive understanding of how the Wold representation exists. And it actually works with the regression coefficients in that projection several slides back corresponds to this inverse operator.

So let's turn to some specific time series models that are widely used. The class of autoregressive moving average processes has this mathematical definition. We define the xt to be equal to a linear combination of lags of x, going back p lags, with coefficients phi 1 through phi p. And then there are residuals which are expressed in terms of a qth order moving average.

So in this framework, the eta t's are white noise. And white noise, to reiterate, has mean 0, constant variance, zero covariance between those. In this representation, I've simplified things a little bit by subtracting off the mean from all of the x's. And that just makes the formulas a little bit more simpler.

Now, with lag operators, we can write this ARMA model as phi of L of pth order polynomial of lag L given with coefficients 1 phi 1 up to phi p, and theta of L given by 1 theta 1 theta 2 up to theta q.

This is basically a representation of the ARMA time series model. Basically, we're taking a set of lags of the values of the stochastic process up to order p. And that's equal to a weighted average of the eta t's. If we multiply by the inverse of phi of L, if that exists, then we get this representation here, which is simply the Wold decomposition.

So the ARMA models basically have a Wold decomposition if this phi of L is invertible. And we'll explore these by looking at simpler cases of the ARMA models by just focusing on autoregressive models first and then moving average processes second so that you'll get a better feel for how these things are manipulated and interpreted. So let's move on to the pth order autoregressive process. So we're going to consider ARMA models that just have autoregressive terms in them.

So we have phi of L xt minus mu is equal to eta t, which is white noise. So a linear combination of the series is white noise. And xt follows then a linear regression model on explanatory variables, which are lags of the process x. And this could be expressed as xt equal to c plus the sum from 1 to p of phi j xt minus j, which is a linear regression model with regression parameters phi j.

And c, the constant term, is equal to mu times phi of 1. Now, if you basically take expectations of the process, you basically have coefficients of mu coming in from all the terms. And phi of 1 times mu is the regression coefficient there. So with this autoregressive model, we now want to go over what are the stationarity conditions.

Certainly, this autoregressive model is one where, well, a simple random walk follows an autoregressive model but is not stationary. We'll highlight that in a minute as well. But if you think it, that's true.

And so stationarity is something to be understood and evaluated. This polynomial function, phi, where if we replace the lag operator, L, by z, a complex variable, the equation phi of z equal to 0 is the characteristic equation associated with this autoregressive model.

And it turns out that we'll be interested in the roots of this characteristic equation . Now, if we consider writing phi of L as a function of the roots of the equation, we get this expression where you'll notice if you multiply all those terms out, the 1's all multiply out together, and you get 1. And with the lag operator, L, to the pth power, that would be the product of 1 over lambda 1 times 1 over lambda 2, or actually negative 1 over lambda 1 times negative 1 over lambda 2, and so forth-- negative 1 over lambda p.

Basically, if there are p roots to this equation, this is how it would be written out. And the process xt is covariant and stationary if and only if all the roots of this characteristic equation lie outside the unit circle. So what does that mean? That means that the norm modulus of the complex z is greater than 1. So they're outside the unit circle where it's less than or equal to 1.

And the roots, if they are outside the unit circle, then the modulus of the lambda j's is greater than 1. And if we then consider taking a complex number, lambda, basically the root, and have an expression for 1 minus 1 over lambda L inverse, we can get this series expression for that inverse. And that series will exist and be bounded if the lambdi are greater than 1 in magnitude.

So we can actually compute an inverse of phi of L by taking the inverse of each of the component products in that polynomial. So an introductory time series, of courses, as they talk about stationarity and unit roots, but they don't really get into it, because people don't know complex math, don't know about root. So anyway, but this is just very simply how that framework is applied.

So we have a polynomial equation for the characteristic equation whose roots we're looking for. Those roots have to be outside the unit circle for stationarity of the process. Well, it's basically conditions for invertibility of the process of the autoregressive process. And that invertibility renders the process in infinite order moving average process.

So let's go through these results for the auto regressiveprocess of order one, where things-- always start with the simplest cases to understand things. The characteristic equation for this model is just 1 minus phi z. The root is 1/phi. So lambda is greater than 1-- if the modulus of lambda is greater than 1, meaning the root is outside the unit circle, then phi is less than 1.

So for covariant stationarity of this autoregressive process, we need the magnitude of phi to be less than 1 in magnitude. The expected value of x is mu. The variance of x is sigma squared x. This has this form, sigma squared over 1 minus phi. That expression is basically obtained by looking at the infinite order moving average representation.

But notice that if phi is positive, then the variance of x is actually greater than the variance of the innovations. And if phi is less than 0, then it's going to be smaller. So the innovation variance basically is scaled up a bit in the autoregressive process. The covariance matrix is phi times sigma squared x. You'll be going through this in the problem set.

And the covariance of x is phi to the j power sigma squared x. And these expressions can all be easily evaluated by simply writing out the definition of these covariances in terms of the original model and looking at what terms are independent, cancel out, and that proceeds.

Let's just go through these cases. Let's show it all here. So we have if phi is between 0 and 1, then the process experiences exponential mean reversion to mu. So an autoregressive process with phi between 0 on 1 corresponds to a mean reverting process. This process is actually one that has been used theoretically for interest rate models and a lot of theoretical work in finance.

The Vasicek model is actually an example of the Ornstein-Uhlenbeck process, which is basically a mean reverting Brownian motion. And any variables that exhibit or could be thought of as exhibiting mean reversion, this model can be applied to those processes, such as interest rate spreads or real exchange rates, variables where one can expect that things never get too large or too small. They come back to some mean.

Now, the challenge is, that usually may be true over short periods of time. But over very long periods of time, the point to which youre reverting to changes. So these models tend to not have broad application over long time ranges. You need to adapt.

Anyway, with the AR process, we can also have negative values of phi, which results in exponential mean reversion that's oscillating in time, because the autoregressive coefficient basically is a negative value. And for phi equal to 1, the Wold decomposition doesn't exist. And the process is the simple random walk.

So basically, if phi is equal to 1, that means that basically just changes in value of the process are independent and identically distributed white noise. And that's the random walk process. And that process was covered in earlier lectures as non stationary.

If phi is greater than 1, then you have an explosive process, because basically the values are scaling up every time increment. So those are features of the AR 1 model. For a general autoregressive process of order p, there's a method-- well, we can look at the second order moments of that process, which have a very nice structure, and then use those to solve for estimates of the ARMA parameters, or autoregressive parameters.

And those happen to be specified by what are called the Yule-Walker equations. So the Yule-Walker equations is a standard topic in time series analysis. What is it? What does it correspond to?

Well, we take our original autoregressive process of order p. And we write out the formulas for the covariance at lag j between two observations. So what's the covariance between xt and xt minus j?

And that expression is given by this equation. And so this equation for gamma of j is determined simply by evaluating the expectations where we're taking the expectation of xt in the autoregressive process times the fix xt minus j minus mu.

So just evaluating those terms, you can validate that this is the equation. If we look at the equations corresponding to j equals 1-- so lag 1 up through lag p-- this is what those equations look like. Basically, the left hand side is gamma 1 through gamma p. The covariance to lag 1 up to lag p is equal to basically linear functions given by the phi of the other covariances.

Who can tell me what the structure is of this matrix? It's not a diagonal matrix? What kind of matrix is this? Math trivia question here. It has a special name. Anyone?

It's a Toeplitz matrix. The off diagonals are all the same value. And in fact, because of the symmetry of the covariance, basically the gamma of 1 is equal to gamma of minus 1. Gamma of minus 2 is equal to gamma plus 2. Because of the covariant stationarity, it's actually also symmetric.

So these equations allow us to solve for the phis so long as we have estimates of these covariances. So if we have a system of estimates, we can plug these in in an attempt to solve this. If they're consistent estimates of the covariances, then there will be a solution.

And then the 0th equation, which was not part of the series of equations-- if you go back and look at the 0th equation, that allows you to get an estimate for the sigma squared. So these Yule-Walker equations are the way in which many ARMA models are specified in different statistics packages and in terms of what principles are being applied. Well, if we're using unbiased estimates of these parameters, then this is applying what's called the method of moments principle for statistical estimation.

And with complicated models, where sometimes the likelihood functions are very hard to specify and compute, and then to do optimization over those is even harder. It can turn out that there are relationships between the moments of the random variables, which are functions of the unknown parameters. And you can solve for basically the sample moments equalling the theoretical moments and you apply the method of moments estimation method.

Econometrics is rich with many applications of that principle. The next section goes through the moving average model. Let me highlight this. So with an order q moving average, we basically have a polynomial in lag operator, L, which is operated upon the eta t's. And if you write out the expectations of xt, you get mu.

The variance of xt, which is gamma 0, is sigma squared times 1 plus the squares of the coefficients in the polynomial. And so this feature, this property here is due to the fact that we have uncorrelated innovations in the eta t's. The eta t's are white noise.

So the only thing that comes through in the square of xt and the expectation of that is the squared powers of the etas, which have coefficients given by the theta i squared. So these properties are left-- I'll leave you just to verify very straightforward.

But let's now turn to the final minutes of the lecture today to accommodating non stationary behavior in time series. The original approaches with time series was to focus on estimation methodologies for covariant stationary process. So if the series is not covariant stationary, then we would want to do some transformation of the data, of the series, into a stationary so that the resulting process is stationary.

And with the difference in operators, delta, Box and Jenkins advocated moving non-stationary trending behavior, which is exhibited often in economic time series, by using a first difference, maybe a second difference, or a kth order difference. So these operators are defined in this way.

Basically with the kth order operator having this expression here, this is the binomial expansion of a kth power, which can be useful. It comes up all the time in probability theory. And if a process has a linear time trend, then delta xt is going to have no time trend at all, because you're basically taking out that linear component by taking successive differences.

Sometimes, if you have a real series and you look at the difference, it appears non-stationary, you look at first differences, that can still not appear to be growing over time, in which case sometimes the second difference will result in a process with no trend. So these are sort of convenient tricks, techniques to render the series stationary.

And let's see. There's examples here of linear trend reversion models which are rendered covariant stationary under first differencing. In this case, this is an example where you have a deterministic time trend. But then you have reversion to the time trend over time.

So we basically have eta t, the error about the deterministic trend, is a first order autoregressive process. And the moments here can be derived this way. Leave that as an exercise.

One could also consider the pure integrated process and talk about stochastic trends. And basically, random walk processes are you are often referred to in econometrics as stochastic trends. And you may want to try and remove those from the data, or accommodate them.

And so the stochastic trend process is basically given by the first difference, xt, is just equal to eta t. And so we have essentially this random walk from a given starting point. And it's easy to verify it if you knew the 0th point, then the variance of the t'th time point would be t sigma squared, because we're summing t independent innovations.

And the covariance between t and lag t minus j is simply t minus j sigma squared. And the correlation between those has this form. What you can see is that this definitely depends on time. So it's not a stationary process. So this first differencing results in stationarity. And the end difference process has those features.

Let's see where we are. Final topic for today is just how you incorporate non-stationary process into ARMA processes. Well, if you take first differences or second differences and the resulting process is covariant stationary, then we can just incorporate that differencing into the model specification itself, and define ARIMA models, Autoregressive Integrated Moving Average Processes.

And so to specify these models, we need to determine the order of the differencing required to move trends, deterministic or stochastic, and then estimating the unknown parameters, and then applying model selection criteria. So let me go very quickly through this and come back to it the beginning of next time.

But in specifying the parameters of these models, we can apply maximum likelihood, again, if we assume normality of these innovations eta t. And we can express the ARMA model in state space form, which results in a form for the likelihood function, which we'll see a few lectures ahead.

But then we can apply limited information, maximum likelihood, where we just condition on the first observations of the data and maximize the likelihood. Or not condition on the first few observations, but also use their information as well, and look at their density functions, incorporating those into the likelihood relative to the stationary distribution for their values.

And then the issue becomes, how do we choose amongst different models? Now, last time we talked about linear regression models, how you'd specify a given model, here, we're talking about autoregressive, moving average, and even integrated moving average processes and how do we specify those, well, with the method of maximum likelihood, there are procedures which there are measures of how effectively a fitted model is, given by an information criterion that you would want to minimize for a given fitted model.

So we can consider different sets of models, different numbers of explanatory variables, different orders of autoregressive parameters, moving average parameters, and compute, say, the Akaike information criterion or the Bayes information criterion or the Hannan-Quinn criterion as different ways of judging how good different models are. And let me just finish today by pointing out that what these information criteria are is basically a function of the log likelihood function, which is something we're trying to maximize with maximum likelihood estimates.

And then adding some penalty for how many parameters we're estimating. And so what I'd like you to think about for next time is what kind of a penalty is appropriate for adding an extra parameter. Like, what evidence is required to incorporate extra parameters, extra variables, in the model. Would it be t statistics that exceeds some threshold or some other criteria.

Turns out that these are all related to those issues. And it's very interesting how those play out. And I'll say that for those of you who have actually seen these before, the Bayes information criterion corresponds to an assumption that there is some finite number of variables in the model. And you know what those are.

The Hannon-Quinn criterion says maybe there's an infinite number of variables in the model, but you want to be able to identify those. And so anyway, it's a very challenging problem with model selection. And these criteria can be used to specify those. So we'll go through that next time.