

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:

All right. Let's see. We're going to start today with a wrap up of our discussion of univariate time series analysis. And last time we went through the world representation theorem, which applies to covariant stationary processes, a very powerful theorem. And implementations of the covariant stationary processor with ARMA models. And we discussed estimation of those models with maximum likelihood.

And here in this slide i just wanted to highlight how when we estimate models with maximum likelihood we need to have an assumption of a probability distribution for what's random, and in the ARMA structure we consider the simple case where the innovations, the η_t , are normally distributed white noise. So they're independent and identically distributed normal random variables.

And the likelihood function can be maximized at the maximum likelihood parameters. And it's simple to implement the limited information maximum likelihood where one conditions on the first few observations in the time series. If you look at the likelihood structure for ARMA models, the density of an outcome at a given time point depends on lags of that dependent variable.

So if those are unavailable, then that can be a problem. One can implement limited information maximum likelihood where you're just conditioning on those initial values, or there are full information maximum likelihood methods that you can apply as well. Generally though the limited information case is what's applied.

Then the issue is model selection. And with model selection the issues that arise with time series are issues that arise in fitting any kind of statistical model. Ordinarily one will have multiple candidates for the model you want to fit to data. And the issue

is how do you judge which ones are better than others. Why would you prefer one over the other?

And if we're considering a collection of different ARMA models then we could say, fit all ARMA models of order pq with p and q varying over some range. p from 0 up to p_{\max} , q from q up to q_{\max} . And evaluate those pq different models.

And if we consider σ^2 of pq being the mle of the error variance. Then there are these model selection criteria that are very popular. Akaike information criterion, and Bayes information criterion, and Hannan-Quinn. Now these criteria all have the same term, log of the mle of the error variance. So these criteria don't vary at all with that. They just vary with this second term, but let's focus first on the AIC criterion.

A given model is going to be better if the log of the mle for the error variance is smaller. Now is that a good thing? Meaning, what is the interpretation of that practically when you're fitting different models? Well, the practical interpretation is the variability of the model about where you're predicting things, our estimate of the error variance is smaller. So we have essentially a model with a smaller error variance is better.

So we're trying to minimize the log of that variance. Minimizing that is a good thing. Now what happens when you have many sort of independent variables to include in a model? Well, if you were doing a Taylor series approximation of a continuous function, eventually you'd sort of get to probably the smooth function with enough terms, but suppose that the actual model, it does have a finite number of parameters. And you're considering new factors, new lags of independent variables in the auto-regressions.

As you add more and more variables, well, there really should be a penalty for adding extra variables that aren't adding real value to the model in terms of reducing the error variance. So the Akaike information criterion is penalizing different models by a factor that depends on the size of the model in terms of the dimensionality of the model parameters. So $p + q$ is the dimensionality of the

auto regression model.

So let's see. With the BIC criterion the difference between that and the AIC criterion is that this factor too is replaced by $\log n$. So rather than having a sort of unit increment of penalty for adding an extra parameter, the Bayes information criterion is adding a \log in penalty times the number of parameters.

And so as the sample size gets larger and larger, that penalty gets higher and higher. Now the practical interpretation of the Akaike information criterion is that it is very similar to applying a rule which says, we're going to include variables in our model if the square of the t statistic for estimating the additional parameter in the model is greater than 2 or not. So in terms of when does the Akaike information criterion become lower from adding additional terms to a model?

If you're considering two models that differ by just one factor, it's basically if the t statistic for the model coefficient on that factor is a squared value greater than two or not. Now many of you who have seen regression models before and applied them, in particular applications would probably say, I really don't believe in the value of an additional factor unless the t statistic is greater than 1.96, or two or something. But the Akaike information criterion says the t statistic should be greater than the square root of 2.

So it's sort of a weaker constraint for adding variables into the model. And now why is it called an information criterion? I won't go into this in the lecture. I am happy to go into it during office hours, but there's notions of information theory and [INAUDIBLE] information of the model versus the true model, and trying to basically maximize the closeness of our fitted model to that.

Now the Hannan-Quinn criterion, let's just look at how that differs. Well, that basically has a penalty midway between the $\log n$ and two. It's two, \log , $\log n$. So this has a penalty that's increasing with size n , but not as fast as $\log n$. This becomes relevant when we have models that get to be very large because we have a lot of data. Basically the more data you have, the more parameters you should be able to incorporate in the model if they're sort of statistically valid factors, important

factors.

And the Hannan-Quinn criterion basically allows for modeling processes where really an infinite number of variables might be appropriate, but you need larger and larger sample sizes to effectively estimate those. So those are the criteria that can be applied with time series models.

And I should point out that, let's see, if you took sort of this factor $2 \log \sigma^2$ over n and inverted it to n over $2 \log \sigma^2$, that term is basically one of the terms in the likelihood function of the [INAUDIBLE] model. So you can see how this criterion is basically manipulating the maximum likelihood value by adjusting it for a penalty for extra parameters.

Let's see. OK. Next topic is just test for stationarity and non-stationarity. There's a famous test called the Dickey-Fuller test, which is essentially to evaluate the time series to see if it's consistent with a random walk. We know that we've been discussing sort of lecture after lecture how simple random walks are non-stationary.

And the simple random walk is given by the model up here, $x_t = \phi x_{t-1} + \eta_t$, if ϕ is equal to 1. Right? That is a non-stationary process. Well, in the Dickey-Fuller test we want to test whether ϕ equals 1 or not. And so we can fit the AR(1) model by least squares and define the test statistic to be the estimate of $\phi - 1$ over its standard error where $\hat{\phi}$ is the least squares estimate and the standard error is the least squares estimate, the standard error of that.

If our coefficient ϕ is less than 1 in modulus, so this really is a stationary series, then the estimate $\hat{\phi}$ converges in distribution to a normal zero $(1 - \phi)^2$. And let's see. But if ϕ is equal to one, OK, so just to recap that second to last bullet point is basically the property that when $|\phi| < 1$, then our least squares estimates are asymptotically normally distributed with mean 0 if we normalize by the true value, and $(1 - \phi)^2$.

If ϕ is equal to 1, then it turns out that $\hat{\phi}$ is super consistent with rate $1/t$. Now this super consistency is related to statistics converging to some value. And

what is the rate of convergence of those statistics to different values. So in normal samples we can estimate sort of the mean by the sample mean. And that will converge to the true mean at rate of $1/\sqrt{n}$.

When we have a non-stationary random walk, the independent variables matrix is such that $x^T x$ over n grows without bound. So if we have y is equal to $x\beta + \epsilon$, and $\hat{\beta}$ is equal to $(x^T x)^{-1} x^T y$, the problem is-- well and $\hat{\beta}$ is distributed as ultimately normal with mean β and variance $\sigma^2 (x^T x)^{-1}$.

This $(x^T x)^{-1}$ matrix, when the process is non-stationary, a random walk, it grows infinitely. $x^T x$ over n actually grows to infinity in magnitude just because it becomes unbounded. Whereas $x^T x$ over n , when it's stationary is bounded. So anyway, so that leads to the super consistency, meaning that it converges to the value much faster and so this normal distribution isn't appropriate. And it turns out there's Dickey-Fuller distribution for this test statistic, which is based on integrals of diffusions and one can read about that in the literature on unit roots and test for non-stationarity.

So there's a very rich literature on this problem. If you're into two econometrics, basically a lot of time's been spent in that field on this topic. And the mathematics gets very, very involved, but good results are available.

So let's see an application on some of these time series methods. Let me go to the desktop here if I can. In this supplemental material that'll be on the website, I just wanted you to be able to work with time series, real time series and implement these autoregressive moving average fits and understand basically how things work. So in this it introduces loading the R libraries and Federal Reserve data into R, basically collecting it off the web.

Creating weekly and monthly time series from a daily series, and it's a trivial thing to do, but when you sit down and try to do it gets involved. So there's some nice tools that are available. There's the ACF and the PACF, which is the auto-correlation function and the partial auto-correlation function, which are used for interpreting

series. Then we conduct Dickey-Fuller test for unit roots and determine, evaluate stationarity, non-stationarity of the 10-year yield.

And then we evaluate stationarity and cyclicity in the fitted auto regressive model of order 2 to monthly data. And actually 1.7 there, that cyclicity issue relates to one of the problems on the problem set for time series, which is looking at with second order auto regressive models. Is there cyclicity in the process?

And then finally looking at identifying the best auto regressive model using the AIC criterion. So let me just page through and show you a couple of plots here. OK. Well, there's the original 10 year yield collected directly from the Federal Reserve website over a 10 year period. And, oh, here we go. This is nice. OK. OK.

Let's see this section 1.4 conducts the Dickey-Fuller test. And it basically determines that the p value for non-stationarity is not rejected. And so, with the augmented Dickey-Fuller test, the test statistic is computed. Its significance is evaluated by the distribution for that this statistic. And the p value tells you how extreme the value of the statistic is, meaning how unusual is it. The smaller the p value, the more unlikely the value is.

The p value is what's the likelihood of getting as extreme or more extreme a value of the test statistic, and the test statistic is evidence against the null hypothesis. So in this case the p values range basically 0.2726 for the monthly data, which says that basically there is evidence of a unit root in the process.

Let's see. OK. There's a section on understanding partial auto-correlation coefficients. And let me just state what the partial correlation coefficients are. You have the auto-correlation functions, which are simply the correlations of the time series with lags of its values. The partial auto-correlation coefficient is the correlation that's between the time series and say, it's a p'th lag that is not explained by all lags lower than p. So it's basically the incremental correlation of the time series variable with the p'th lag after controlling for the others.

And then let's see. With this, in section eight here there's a function in r called ar, for

auto regressive, which basically will fit all auto regressive models up to a given order and provide diagnostic statistics for that. And here is a plot of the relative AIC statistic for models of the monthly data. And you can see that basically it takes all the AIC statistics and subtracts the smallest one from all the others. So one can see that according to the AIC statistic a model of order seven is suggested for this treasure yield data. OK.

Then finally because these auto regressive models are implemented with regression models, one can apply Regression Diagnostics that we had introduced earlier to look at those data as well. All right. So let's go down now. [INAUDIBLE]. OK. [INAUDIBLE]. Full screen. Here we go. All right.

So let's move on to the topic of volatility model. The discussion in this section is going to begin with just defining volatility. So we know what we're talking about. And then measuring volatility with historical data where we don't really apply sort of statistical models so much, but we're concerned with just historical measures of volatility and their prediction.

Then there are formal models. We'll introduce geometric Brownian motion, of course. That's one of the standard models in finance. But also Poisson jump-diffusion, which as an extension of geometric Brownian motion to allow for discontinuities. And then there's a property of these Brownian motion and jump-diffusion models which is models with independent increments. Basically you have disjoint increments of the process, basically are independent of each other, which is a key property when there's time dependence in the models.

There can be time dependence actually in the volatility. And arch models were introduced initially to try and capture that. And were extended to GARCH models, and these are the sort of simplest cases of time dependent volatility models that we can work with and introduce. And in all of these the sort of mathematical framework for defining these models and the statistical framework for estimating their parameters is going to be highlighted. And while it's a very simple setting in terms of what these models are, these issues that we'll be covering relate to virtually all

statistical modeling as well.

So let's define volatility. OK. In finance it's defined as the annualized standard deviation of the change in price or value of a financial security, or an index. So we're interested in the variability of this process, a price process or a value process. And we consider it on an annualized time scale. Now because of that, when you talk about volatility it really is meaningful to communicate, levels of 10%. If you think of, at what level do sort of absolute bond yields vary over a year? It's probably less than 5%. Bond yields don't.

When you think of currencies, how much do those vary over a year. Maybe 10%. With equity markets, how do those vary? Well, maybe 30%, 40% or more. With the estimation and prediction approaches, OK, these are what we'll be discussing. There's different cases. So let's go onto historical volatility. In terms of computing the historical volatility we'll be considering basically a price series of t plus 1 points. And then we can get t period returns for corresponding to those prices, which is the difference in the logs of the prices, or the log of the price relatives.

So r_t is going to be the return for the asset. And one could use other definitions, like sort of the absolute return, not take logs. It's convenient in much empirical analysis, I guess, to work with the logs because if you sum logs you get sort of log of the product. And so total cumulative returns can be computed easily with sums of logs. But anyway, we'll work with that scale for now.

OK. Now the process r_t , the return series process is going to be assumed to be covariant stationary, meaning that it does have a finite variance. And the sample estimate of that is just given by the square root of the sample variance. And we're also considering an unbiased estimate of that. And if we want to basically convert these to annualized values so that we're dealing with a volatility, then if we have daily prices of which in financial markets they're usually-- in the US they're open roughly 252 days a year on average.

We multiply that $\hat{\sigma}$ by 252 root. And for weekly, root 52, and root 12 for monthly data. So regardless of the periodicity of our original data we can get them

onto that volatility scale.

Now in terms of prediction methods that one can make with historical volatility, and there's a lot of work done in finance by people who aren't sort of trained as a econometricians or statisticians, they basically just work with the data. And there's a standard for risk analysis called the risk metrics approach, where the approach defines volatility and volatility estimates historical estimates just using simple methodologies.

And so that's just go through what those are here. One can basically for any period t can define the sample volatility, just to be the sample standard deviation of the period t returns. And so with daily data that might just be the square of that daily return. With monthly data it could be the sample standard deviation of the returns over the month and it would be the sample over the year.

Also with intraday data, it could be the sample standard deviation over intraday periods of say, half hours or hours. And the historical average is simply the mean of those estimates, which uses all the available data. One can consider the simple moving average of these realized volatilities. And so that basically is using the last m for some finite m values to average.

And one could also consider an exponential moving average of these sample volatilities where we have our estimate of the volatility is $1 - \beta$ times the current period volatility plus β times the previous estimate. And these exponential moving averages are really very nice ways to estimate processes that change over time. And they're able to track the changes quite well and they will tend to come up again and again.

This exponential moving average actually uses all available data. And there can be discrete versions of those where you say, well let's use not an equal weighted average like the simple moving average, but let's use a geometric average of the last m values in an exponential way. And that's the exponential weighted moving average that uses the last m . OK. There we go. OK.

Well, with these different measures of sample volatility, one can basically build models to estimate them with regression models and evaluate. And in terms of the risk metrics benchmark, they consider a variety of different methodologies for estimating volatility. And sort of determine what methods are best for different kinds of financial instruments. And different financial indexes.

And there are different performance measures one can apply. Sort of mean squared error of prediction, mean absolute error of prediction. Mean absolute prediction error, and so forth to evaluate different methodologies. And on the web you can actually look at the technical documents for risk metrics and they go through these analyses and if your interest is in a particular area of finance, whether it's fixed income or equities, commodities, or currencies, reviewing their work there is very interesting because it does highlight different aspects of those markets.

And it turns out that basically the exponential moving average is generally a very good method for many instruments. And the sort of discounting of the values over time corresponds to having roughly between, I guess, a 45 and a 90 day period in estimating your volatility. And in these approaches which are, I guess, they're a bit ad hoc. There's the formalism. And defining them is basically just empirically what has worked in the past. Let's see.

While these things are ad hoc, they actually have been very, very effective. So let's move on to formal statistical models of volatility. And the first class's model is the Geometric Brownian Motion. So here we have basically a stochastic differential equation defining the model for Geometric Brownian Motion. And [INAUDIBLE] will be going in some detail about stochastic differential equations. And stochastic calculus for representing different processes, continuous processes.

And the formulation is basically looking at increments of the price process s is equal to basically a μs of t . Sort of a drift term, plus a σs of t , a multiple of dw of t , where σ is the volatility of the security price. μ is the mean return per unit time. dw of t is the increment of a standard Brownian motion processor, Wiener process. And this w process is such that it's increments, basically the change in

value of the process over between two time points is normally distributed, with mean 0 and variance equal to the length of the interval. And increments on disjoint time intervals are independent.

And well, if you divide both sides of that equation by s of t then you have ds of t over s of t is equal to μdt plus σdw of t . And so the increments ds of t normalized by s of t are a standard Brownian Motion with drift μ and volatility σ .

Now with sample data from this process, now suppose we have prices observed at times t_0 up to t_n . And for now we're not going to make any assumptions about what those time increments are, what those times are. They could be equally spaced. They could be unequally spaced. The returns, the log of the price relative price change from time t_{j-1} to t_j are independent random variables. And they are independent. Their distribution is normally distributed with mean given by μ times the length of the time increment. And variance σ^2 times the length of the increment.

And these properties will be covered by [INAUDIBLE] in some later lectures. So for now what we can just know that this is true and apply this result. If we fix various time points for the observation and compute returns this way. If it's a Geometric Brownian Motion we know that this is the distribution of the returns.

Now knowing that distribution we can now engage in maximum likelihood estimation. OK. If the increments are all just equal to 1, so we're thinking of daily data, say. Then the maximum likelihood estimates are simple. It's basically the sample mean and the sample variance with 1 over n instead of 1 over $n-1$ to the [? mle's. ?] If Δ_j varies then, well, that's actually a case for in the exercises.

Now does anyone, in terms of, well, in the class exercise the issue that is important to think about is if you consider a given interval of time over which we're observing this Geometric Brownian Motion process, if we increase the sampling rate of prices over a given interval, how does that change the properties of our estimates? Basically, do we obtain more accurate estimates of the underlying parameters?

And as you increase the sampling frequency, it turns out that some parameters are estimated much, much better and you get basically much lower standard errors on those estimates. With other parameters you don't necessarily. And the exercise is to evaluate that.

Now another issue that's important is the issue of sort of what is the appropriate time scale for Geometric Brownian Motion. Right now we're thinking of, you collect data, whatever the periodicity is of the data is you think that's your period for your Brownian Motion. Let's evaluate that. Let me go to another example. Let's see here. Yep. OK. Let's go Control minus here. OK. All right.

Let's see. With this second case study there was data on exchange rates looking for regime changes in exchange rate relationships. And so we have data from that case study on different foreign exchange rates. And here in the top panel I've graphed the euro dollar exchange rate From the beginning of 1999 through just a few months ago.

And the second panel is a plot of the daily returns for that series. And here is a histogram of those daily returns. And a fit of the Gaussian distribution for the daily returns if our sort of time scale is correct. Basically daily returns are normally distributed. Days are disjoint in terms of the price change. And so they're independent and identically distributed under the model. And they all have the same normal distribution with mean μ and variance σ^2 . OK.

This analysis assumes basically that we're dealing with trading days for the appropriate time scale, the Geometric Brownian Motion. Let's see. One can ask, well, what if trading dates really isn't the right time scale, but it's more calendar time. The change in value over the weekends maybe correspond to price changes, or value changes over a longer period of time. And so this model really needs to be adjusted for that time scale.

The exercise that allows you to consider different Δt 's shows you what the maximum likelihood estimates. You'll be driving maximum likely estimates if we have different definitions of timescale there. But if you apply the calendar time scale to

this euro, let me just show you what the different estimates are of the annualized mean return and the annualized volatility.

So if we consider trading days for euro it's 10.25% or 0.1025. If you consider clock time, it actually turns out to be 12.2%. So depending on how you specify the model you get a different definition of volatility here. And it's important to basically understand sort of what the assumptions are of your model and whether perhaps things ought to be different.

In stochastic modeling, there's an area called subordinated stochastic processes. And basically the idea is if you have a stochastic process like Geometric Brownian Motion or simple Brownian motion, maybe you're observing that on the wrong time scale. You may fit the Geometric Brownian Motion model and it doesn't look right. But it could be that there's a different time scale that's appropriate. And it's really Brownian Motion on that time scale.

And so formally it's called a subordinated stochastic process. You have a different time function for how to model the stochastic process. And the evaluation of subordinating a stochastic process leads to consideration of different time scales. With, say, equity markets, in futures markets sort of the volume of trading, sort of cumulative volume of trading might be a really inappropriate measure of the real time scale. Because that's a measure of, in a sense, information flow coming into the market through the level of activity.

So anyway I wanted to highlight how with different time scales you can get different results. And so that's something to be evaluated. In looking at these different models, OK, these first few graphs here show the fit of the normal model with the trading day time scale. Let's see. Those of you who've ever taken a statistics class before, or an applied statistics, may know about normal q-q plots.

Basically if you want to evaluate the consistency of the returns here with a Gaussian distribution, what we can do is plot the observed ordered sorted returns against what we would expect the sorted returns to be if it were from a Gaussian sample. So under the Geometric Brownian Motion model the daily returns are a sample

independent and identically distributed random variable sample from a Gaussian distribution.

So the smallest return should be consistent with the smallest of the sample size n . And what's being plotted here is the theoretical quantiles or percentiles versus the actual ones. And one would expect that to lie along a straight line if the theoretical quantiles were well-predicting the actual extreme values.

What we see here is that as you the theoretical quantiles get high and it's in units of standard deviation units, the realized sample returns are in fact much higher than would be predicted by the Gaussian distribution. And similarly, on the low end side. So there's a normal q-q plot that's used often in the diagnostics of these models.

Then down here I've actually plotted a fitted percentile distribution. Now what's been done here is if we modeled the series as a series of Gaussian random variables then we can evaluate the percentile of the fitted Gaussian distribution that was realized by every point. So if we have a return of say negative 2%, what percentile is the normal fit of that? And you can evaluate the cumulative distribution function of the fitted model at that value to get that point.

And what should the distribution of percentiles be for fitted percentiles if we have a really good model? OK. Well, OK. Let's think. If you consider the 50th percentile you would expect, I guess, 50% of the data to lie above the 50th percentile and 50% to lie below the 50th percentile, right? OK. Let's consider, here I divided up into 100 bins between zero and one so this bin is the 99th percentile. How many observations would you expect to find in between the 99th and 100 percentile? This is an easy question.

AUDIENCE: 1%.

PROFESSOR: 1%. Right. And so in any of these bins we would expect to see 1% if the Gaussian model were fitting. And what we see is that, well, at the extremes they're more extreme values. And actually inside there are some fewer values. And actually this is exhibiting a Leptokurtic distribution for the actually realized samples, basically the

middle of the distribution is a little thinner and it's compensated for by fatter tails.

But with this particular model we can basically expect to see a uniform distribution of percentiles in this graph. If we compare this with a fit of the clock time we actually see that clock time does a bit of a better job at getting the extreme values closer to what we would expect them to be. So in terms of being a better model for the returns process, if we're concerned with these extreme values, we're actually getting a slightly better value with those.

So all right. Let's move on back to the notes. And talk about the Garman-Klass Estimator. So let me do this. All right. View full screen. OK. All right. So, OK.

The Garman-Klass Estimator is one where we consider the situation where we actually have much more information than simply sort of closing prices at different intervals. Basically all transaction data's collected in a financial market. So really we have virtually all of the data available if we want it, or can pay for it. But let's consider a case where we expand upon just having closing prices to having additional information over increments of time that include the open, high, and low price over the different periods.

So those of you who are familiar with bar data graphs that you see whenever you plot stock prices over periods of weeks or months you'll be familiar with having seen those. Now the Garman-Klass paper addressed how can we exploit this additional information to improve upon our estimates of the close to close. And so we'll just use this notation. Well, let's make some assumptions and notation.

So we'll assume that μ is equal to 0 in our Geometric Brownian Motion model. So we don't have to worry about the mean. We're just concerned with volatility. We'll assume that the increments are one for daily, corresponding to daily. And we'll let little f between zero and one correspond to the time of day at which the market opens.

So over a day, from day zero, day one at f we assume that the market opens and basically the Geometric Brownian Motion process might have closed on day zero

here. So this would be c_0 and it may have opened on day one at this value. So this would be O_1 . Might have gone up and down and then closed here with the Brownian Motion process. OK. This value here would correspond to the high value.

This value here would correspond to the low value on day one. And then the closing value here would be c_1 . So the model is we have this underlying Brownian Motion process is actually working over continuous time, but we just observe it over the time when the markets open. And so it can move between when the market closes and opens on any given day and we have the additional information. Instead of just the close, we also have the high and low.

So let's look at how we might exploit that information to estimate volatility. OK. Using data from the first period as we've graphed here, let's first just highlight what the close to close return is. And that basically is an estimate of the one period variance. And so $\hat{\sigma}_0^2$ is a single period squared return. $c_1 - c_0$ as a distribution which is normal with mean 0, and variance σ^2 .

And if we consider squaring that, what's the distribution of that? That's the square of a normal random variable, which is chi squared, but it's a multiple of a chi squared. It's σ^2 times chi squared one random variable. And with a chi squared random variable the expected value is 1. The variance of a chi squared random variable is equal to 2. So just knowing those facts gives us the fact we have an unbiased estimate of the volatility parameter σ and the variance is $2\sigma^4$ to the fourth.

So that's basically the precision of close to close returns. Let's look at two other estimates. Basically the open to close return, σ^2 normalized by f , the length of the interval f . So we have σ^2 is equal to $O_1 - c_0$ squared. OK.

Actually why don't I just do this? I'll just write down a few facts and then you can see that the results are clear. Basically $O_1 - c_0$ is distributed normal with mean 0 and variance $f\sigma^2$. And $c_1 - O_1$ is distributed normal with mean 0 in variance $1 - f\sigma^2$. OK.

This is simply using the properties of the diffusion process over difference periods of time. So if we normalize the squared values by the length of the interval we get estimates of the volatility. And what's particularly significant about these estimates one and two is that they're independent. So we actually have two estimates of the same underlying parameter, which are independent. And actually they both have the same mean and they both have the same variance.

So if we consider a new estimate, which is basically averaging those two. Then this new estimate has the same, is unbiased as well, but it's variance is basically the variance of this sum. So it's a $1/2$ squared times this variance plus a $1/2$ squared times this variance, which is a $1/2$ the variance of each of them. So this estimate has lower variance than our close to close. And we can define the efficiency of this particular estimate relative to the close to close estimate as 2.

Basically we get double the precision. Suppose you had the open high close for one day. How many days of close to close data would you need to have the same variance as this estimate? No.

AUDIENCE: [INAUDIBLE]. Because of the three data points [INAUDIBLE].

PROFESSOR: No. No. Anyone else? One more. Four. OK. Basically is if the variance is a $1/2$ basically to get the standard deviation, or the variance to be-- I'm sorry. The ratio of the variance is two. So no. So it actually is close to two.

Let's see. Our $1/n$ is-- so it actually is two. OK. I was thinking standard deviation units instead of squared units. So I was trying to be clever there. So it actually is basically two days. So sampling this with this information gives you as much as two days worth of information. So what does that mean? Well, if you want something that's as efficient as daily estimates you'll need to look back one day instead of two days to get the same efficiency with the estimate.

All right. The motivation for the Garman-Klass paper was actually a paper written by Parkinson in 1976, which dealt with using the extremes of a Brownian Motion to estimate the underlying parameters. And when [INAUDIBLE] talks about Brownian

Motion a bit later I don't know if you'll derive this result, but in courses on stochastic processes one does derive properties of the maximum of a Brownian Motion over a given interval and the minimum. And it turns out that if you look at the difference between the high and low squared divided by 4, $\log 2$, this is an estimate of the volatility of the process.

And the efficiency of this estimate turns out to be 5.2, which is better yet. Well, Garman and Klass were excited by that and wanted to find even better ones. So they wrote a paper that evaluated all different kinds of estimates. And I encourage you to Google that paper and read it because it's very accessible. And it sort of highlights the statistical and probability issues associated with these problems. But what they did was they derived the best analytics scale invariant estimator, which has this sort of bizarre combination of different terms, but basically we're using normalized values of the high low close normalized by the open.

And they're able to get an efficiency of 7.4 with this combination. Now scale [? invariant ?] estimates in statistical theory, they're different principles that guide the development of different methodologies. And one kind of principle is issues of scale invariance. If you're estimating a scale parameter, and in this case volatility is telling you essentially how large is the variability of this process, if you were to say multiply your original data all by a given constant, then a scale invariant estimator should be such that your estimator changes in that case, only by that same scale factor. So sort of the estimator doesn't depend on how you scale the data. So that's the notion of scale invariance.

The Garman and Klass paper actually goes to the nth degree and actually finds a particular estimator that has an [? efficiency ?] of 8.4, which is really highly significant. So if you are working with a modeling process where you believe that the underlying parameters may be reasonably assumed to be constant over short periods of time, well, over those short periods of time if you use these extended estimators like this, you'll get much more precise measures of the underlying parameters than from just using simple close to close data. All right.

Let's introduce Poisson Jump Diffusions. With Poisson Jump Diffusions we have basically a stochastic differential equation for representing this model. And it's just like the Geometric Brownian Motion model, except we have this additional term, $\Gamma \sigma dz + \pi dt$.

Now that's a lot of different variables, but essentially what we're thinking about is over time a Brownian Motion process is fully continuous. There are basically no jumps in this Brownian Motion process. In order to allow for jumps, we assume that there's some process π of t , which is a Poisson process. It's counting process that counts when jumps occur, how many jumps have occurred. So that might start at 0 at the value 0.

Then if there's a jump here it goes up by one. And then if there's another jump here, it goes up by one, and so forth. And so the Poisson Jump Diffusion model says, this diffusion process is actually going to experience some shocks to it. Those shocks are going to be arriving according to a Poisson process. If you've taken stochastic modeling you know that that's a sort of a purely random process.

Basically exponential arrival rate of shocks occur. You can't predict them. And when those occur, $d\pi$ of t is going to change from 0 up to the unit increments. So $d\pi$ of t is 1. And then we'll realize $\Gamma \sigma z$ of t . So at this point we're going to have shocks. Here this is going to be $\Gamma \sigma z = 1$. And at this point, maybe it's a negative shock, $\Gamma \sigma z = -1$ of this is 0.

And so with this overall process we basically have a shift in the diffusion, up or down according to these values. And so this model allows for the arrival of these processes to be random according to the Poisson distribution, and for the magnitude of the shocks to be random as well.

Now like the Geometric Brownian Motion model this process sort of has independent increments, which helps with this estimation. One could estimate this model by maximum likelihood, but it does get tricky in that basically over different increments of time the change in the process corresponds to the diffusion increment, plus the sum of the jumps that have occurred over that same increment.

And so the model ultimately is a Poisson mixture of Gaussian distributions. And in order to evaluate this model, model's properties moment generating functions can be computed rather directly with that. And so we're one can understand how the moments of the process vary with these different model parameters. The likelihood function is a product of Poisson sums. And there's a closed form for the EM algorithm, which can be used to implement the estimation of the unknown parameters.

And if you think about observing a Poisson Jump Diffusion process, if you knew where the jumps occurred, so you knew where the jumps occurred and how many there were per increment in your data, then the maximum likelihood estimation would all be very, very simple. And because this sort of is a separation of the estimation of the Gaussian parameters from the Poisson parameters.

When you haven't observed those values then you need to deal with methods appropriate for missing data. And the EM algorithm is a very famous algorithm developed by the people up at Harvard, Rubin, Laird, and Dempster, which deals with basically if the problem is much simpler, if you could posit there being unobserved variables that you would observe, then you sort of expand the problem to include your observed data, plus the missing data, in this case where the jumps have occurred.

And you then do conditional expectations of estimating those jumps. And then assuming that those jumps had those occurring with those frequencies, estimating the underlying parameters. So the EM algorithm is very powerful and has extensive applications in all kinds of different models.

I'll put up on the website a paper that I wrote with David [? Packard ?] and his student, [INAUDIBLE], which goes through the Maximum Likelihood methodology for this. But looking at that, you can see how with an extended model, how Maximum Likelihood gets implemented and I think that's useful to see. All right.

So let's turn next to ARCH models. And OK. Just as a bit of motivation the

Geometric Brownian Motion model and also the Poisson Jump Diffusion model are models which assume that volatility over time is essentially stationary. And with the sort of independent increments of those processes, the volatility over different increments is essentially the same. So the ARCH models were introduced to accommodate the possibility that there's time dependence in volatility.

And so let's see. Let's see. Let me go. OK. At the very end, I'll go through an example should showing that time dependence with our euro dollar exchange rates. So the set up for this model is basically we look at the log of the price relatives y_t and we model the residuals to not be of constant volatility, but to be multiples of sort of white noise with mean 0 and variance 1, where σ_t is given by this essentially ARCH function, which says that the volatility at a given period t is a weighted average of the squared residuals over the last p lags.

And so if there's a large residual then that could persist and make the next observation have a large variance. And so this accommodates some time dependence. Now this model actually has parameter constraints, which are never a nice thing to have when you're fitting models.

In this case the parameters α_1 through α_p all have to be positive. And why do they have to be positive?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right. Variance is positive. So if any of these alphas were negative, then there would be a possibility that under this model that you could have negative volatility, which you can't. So if we estimate this model to estimate them with the constraint that all these parameter values are non-negative. So that does complicate the estimation a bit.

In terms of understanding how this process works one can actually see how the ARCH model implies an autoregressive model for the squared residuals, which turns out to be useful. So the top line there is the ARCH model saying that the variance of the t period return is this weighted average of the past residuals. And

then if we simply add a new variable u_t , which is our squared residual minus its variance, to both sides we get the next line, which says that ϵ_t^2 follows an auto-regression on itself with the u_t value being the disturbance in that auto-regression.

Now u_t , which is $\epsilon_t^2 - \sigma^2$, what is the mean of that? The mean is 0. So it's almost white noise. But its variance is maybe going to change over time. So it's not sort of standard white noise, but it basically has expectation 0. It's also the conditional independence, but there's some possible variability there.

But what this implies is that there basically is an autoregressive model where we just have time varying variances in the underlying process. Now because of that one can sort of quickly evaluate whether there's ARCH structure in data by simply fitting an autoregressive model to the squared residuals. And testing whether that regression is significant or not. And that formally is a Lagrange multiplier test. Some of the original papers by [? Engle ?] go through that analysis. And the test statistic turns out to just be the multiple of the r^2 for that regression fit.

And basically under, say, a null hypothesis, that there isn't any ARCH structure, then this regression model should have no predictability. This ARCH model in the residuals, basically if there's no time dependence in those residuals, that's evidence of there being an absence of ARCH structure. And so under the null hypothesis of no ARCH structure that r^2 statistic should be small. It turns out that sort of n times the r^2 statistic with p variables is asymptotically a chi-square distribution with p degrees of freedom. So that's where that test statistic comes into play.

And in implementing this the fact that we were applying essentially least squares with the outer-regression to implement this Lagrange multiplier test, but we were assuming, well, we're not assuming, we're implicitly assuming the assumptions of Gauss-Markov in fitting that. This corresponds to the notion of quasi-Maximum Likelihood estimates for unknown parameters. And quasi-Maximum Likelihood

estimates $\{\sigma_t\}$ are extensively used in some stochastic volatility models.

And so essentially situations where you sort of use the normal approximation, or the second order approximation to get your estimates, and they turn out to be consistent and decent.

All right. Let's go to Maximum Likelihood Estimation. OK Maximum Likelihood Estimation basically involves-- the hard part is defining the likelihood function, which is the density of the data given the unknown parameters. In this case, the data are conditionally independent.

The joint density is the product of the density of y_t given the information at $t-1$. So basically the joint probability density is the density at each time point conditional on the past, and then the density times the density of the next time point conditional on the past. And those are all normal random variables. So these are the normal PDFs coming into play here.

And so what we want to do is basically maximize this likelihood function subject to these constraints. And we already went through the fact that the $\{\alpha_i\}$ have to be greater than zero, And it turns out you also have to have that the sum of the alphas is less than one. Now what would happen if the sum of the alphas was not less than one?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right. And you basically could have the process start diverging. Basically these auto-regressions can explode. So let's go through and see. Let's see. Actually, we're going to go to GARCH models next. OK. Let's see. Let me just go back here a second. OK. Very good.

OK. In the remaining few minutes let me just introduce you to the GARCH models. The GARCH model is basically a series of past values of the squared volatilities, basically the q sum of past squared volatilities for the equation for the volatility σ_t^2 .

And so it may be that very high order ARCH models are actually important. Or very high order ARCH terms are found to be significant when you fit ARCH models. It could be that much of that need is explained by adding these GARCH terms. And so let's just consider a simple GARCH model where we have only a first order ARCH term and a first order GARCH term.

So we're basically saying that this is a weighted average of the previous volatility, the new squared residual. And this is a very parsimonious representation that actually ends up fitting data quite, quite well. And there are various properties of this GARCH model which we'll go through next time, but I want to just close this lecture by showing you fits so the ARCH models and of this GARCH model to the euro dollar exchange rate process. So let's just look at that here. OK. OK.

With the euro dollar exchange rate, actually there's the graph here which shows the auto-correlation function and the partial auto-correlation function of the squared returns. So is there dependence in these daily volatilities? And basically these blue lines are plus or minus two standard deviations of the correlation coefficient. Basically we have highly significant auto-correlations and very highly significant partial auto-correlations, which suggests if you're familiar with ARMA process that you would need a very high order ARMA process to fit this squared residuals.

But this highlights how with the statistical tools you can actually identify this time dependence quite quickly. And here's a plot of the ARCH order one model and the ARCH order two model. And on each of these I've actually drawn a solid line where the sort of constant variance model would be. So ARCH is saying that we have a lot of variability about that constant mean.

And a property, I guess, of these ARCH models is that they all have sort of a minimum value for the volatility that they're estimating. If you look at the ARCH function, that α_0 now is the constant term is basically the minimum value, which that can be. So there's a constraint sort of on the lower value.

Then here's an ARCH 10 fit which, it doesn't look like it sort of has quite as much of a uniform lower bound, but one could go on and on with higher order ARCH terms,

but rather than doing that one can also fit just a GARCH (1, 1) model. And this is what it looks like.

So basically the time varying volatility in this process is captured really, really well with just this two parameter GARCH model as compared with a high order autoregressive model. And it sort of highlights the issues with the Wold decomposition where a potentially infinite order autoregressive model will effectively fit most time series. Well, that's nice to know, but it's nice to have a parsimonious way of defining that infinite collection of parameters and with the GARCH model a couple of parameters do a good job.

And then finally here's just a simultaneous plot of all those volatility estimates on the same graph. And so one can see the increased flexibility ability basically of the GARCH models compared to the ARCH models for capturing time varying volatility. So all right. I'll stop there for today. And let's see. Next Tuesday is a presentation from Morgan Stanley so. And today's the last day to sign up for a field trip.