Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 04

Lecture 18: Diagnostic Checking -1

Thank you. Hello all, welcome to this course on time series modeling and forecasting using R. So, in the current week, we are exploring the idea of model identification, as in how do you identify the correct orders of an underlying model. And just to give you a brief review of where we are now. So, we started this week by looking at model identification essentially from a So, let us say drawing some ACF plots or PACF plots and so on and so forth.

Now, again, just to remind you of the problem. So, let us say if you have an ARMA model which needs to be fitted on some practical data or some real data. So, how do you actually identify the optimal orders for that ARMA model? So, precisely P and Q, right? So, this is the first part.

So, using visual checks, and then the second idea was using some information criteria. So, AIC, SBC, which is also called BIC, or Hannan and Quinn criteria, or HQIC, etc. So, model identification was done there. So, once you identify the orders, then we talked about estimation. And then, within estimation, we covered broadly two ideas.

So, method of moments and MLE. So, method of moments means that depending on how many parameters you want to estimate based on the underlying model, you create those many equations based on the population parameters and then the sample versions of that. So, let us say you equate population mean to the sample mean, you equate the population variance with the sample variance, or one can actually create or equate, let us say, population correlation with the sample correlation. So, depending on how many parameters you want to estimate, you create those many equations and then solve for the underlying parameters. So, this was the method of moments, and then towards the end of the last lecture, we talked about MLE or maximum likelihood estimation.

But generally speaking, in a time series framework, since our model is not really easy, be it ARMA, ARIMA, SARIMA, whatever. So, we require some numerical approximation technique to solve for the parameters or to maximize that likelihood. So, how do you handle that problem? You write down the likelihood structure or rather the log-likelihood function and then try to maximize that log-likelihood function. So, and then this is exactly where you require the idea of numerical approximation because one cannot maximize it on pen and paper that easily. And right towards the end, we saw how you can kind of find out the MLE based on, let us say, the conditional least squares idea.

So, you find out the least squares, which are of course conditional because you have to feed the algorithm with some starting values and then based on whatever estimator they throw at you. So, those would be the conditional least squares estimators based on maximizing the likelihood. All right. So, assuming that you fix the orders or perform the model identification and the estimation, then the next step in analyzing a time series is diagnostic checking. So, what exactly do you mean by this term diagnostic checking?

So, let us say after identifying and estimating the model parameters, one has to assess the goodness of fit of the model, right, which is one, and the validity of all the assumptions that you are making initially. So, this idea of diagnostic checking is twofold. So, how good the fit is, and then, are all the assumptions valid or not. So, firstly, again, just a quick revision of what assumptions we have made. So, firstly, all these assumptions are based on the errors of the model and not on the actual time series.

And what exactly are the assumptions? So, the error should follow a normal distribution was one. The error should be independent. All the errors should have a constant variance. So, the variance should not change.

So, all these were the basic assumptions. So, if you are estimating the model and you have got an estimated model. So, before forecasting, you have to actually ensure that all the assumptions are being met or not. Before you move ahead with forecasting, okay, so this in-between step is called diagnostic checking, okay. So, if the model fit is perfect, essentially you have performed all the diagnostic checks and you have kind of passed that level, then one can actually go ahead and do forecasting, okay.

Now, how can you actually perform all these diagnostic checks? So, how can you check for normality? How can you check for the independence of the errors? How can you check for constant variance? So, these are some of the points that we will discuss in today's lecture and the one after this.

Alright. So, now the first aspect is the normality of errors. So, again, if you remember, one of the key assumptions that we have made earlier is that all the errors in the underlying model should follow a normal distribution of some sort. And then, whatever mean, whatever variance, but the distribution should be normal. So, we actually have multiple ways of checking this hypothesis or verifying this claim.

So, one can actually draw a very basic histogram of the standardized residuals. Now, firstly, understand what you mean by standardized residuals. So, a standardized residual is nothing but something like Et hat. So, Et hat is the estimated residual divided by its variance. So, we are basically applying some standardization technique here.

So, essentially, standardization is when you subtract the mean and then divide by the standard deviation. But then, if you assume that, let us say, the mean is 0 for simplicity. So, assume the mean to be 0; then you do not have to subtract anything from here, but you still divide by the underlying standard deviation. So, and then, once you get hold of the standardized residuals, you can create a histogram of the standardized residuals. Or one can actually get hold of something called a normal QQ plot.

So, QQ stands for quantile-quantile. So, a normal quantile-quantile plot again of the standardized residuals. Or one can look at something called Tukey's 5-number summary. So, what does this 5-number summary stand for? It gives you several other measures along with, let us say, skewness and kurtosis.

Now, all of you probably know that if the distribution has to be normal, then the skewness has to be 0 because you have a symmetric curve. So, since the curve is symmetric, pretty much something like this, then the skewness has to be close to 0. And then, the kurtosis for a normal distribution should be 3. Or, if you talk about excess kurtosis, so excess kurtosis means the difference between the kurtosis and that of a normal distribution, which has to be basically 0. So, excess kurtosis should be 0, or kurtosis should be 3, basically.

So, if all these ideas are kind of seen in the underlying data or the underlying fit, then probably we can actually see or probably we can actually say that the underlying distribution of the errors is indeed normal. And apart from that, one can actually test out using some hypothesis testing, which is much more formal, right? And then these are some of the tests. So, let us say Shapiro-Wilk or Jarque-Bera, etc. Okay.

So, these two pairs of people, they kind of provided some hypothesis testing to find out whether the underlying distribution is normal or not. Okay. So, this is all about the normality of error. So, we will again go into a bit more detail and then try to explain some of these points here. Okay.

So, for this, what we will do is we will create a simulated data initially. So, if you see here, what we are doing is we are simulating a normal white noise data. So, firstly, white noise means what? So, we are assuming the mean to be 0, basically. So, the mean is 0, and then here we can actually fix some variance, which is a constant.

So, variance could be 1, or variance has to be fixed to some number. And then you look at the residuals, and then this is nothing but a histogram of the residuals. So, this histogram is of a simulated sample from a normal distribution of size 100. So, the sample size in this case is 100. So, my n is 100.

So, what do you do? So, using some software, you simulate a set of random observations or a set of random variables from a normal distribution. Let us say having mean 0 and with some fixed variance, and then the sample size is 100. And then this is the underlying histogram. Now the question is, why does this histogram not look perfectly symmetric?

So, why do you think? So, if you observe this histogram here, you can actually see that you have a few more observations on the left-hand side of the average than on the right-hand side. So, probably there might be a skew. Something like that, isn't it? So, probably there might be a skew which is given by something like that, right.

So, on the left-hand side, you have a lot of observations of the average, but then on the right-hand side, you do not have as many observations. So, why do you see something like this? Because the underlying distribution is normal, right? So, you have simulated from a normal distribution. So, why is it not showing you a perfectly symmetric kind of graph, right?

And the answer to this question is that the sample size is relatively small. So, we are only looking at 100 observations here. So now, in the next slide, we will see what would happen if you increase this n to, let us say, 1000. Then what? So, do we get some improved kind of histogram and so on and so forth?

And, by the way, there is one more thing. So, using or utilizing the same sample containing 100 observations, let us say if you simply plot the data. Now, again, our assumption is that since it is white noise data, the mean is assumed to be 0, and then here you see two colors, by the way. So, the red line gives you a line at 0, which is our mean, and then the blue line gives you the actual mean of the data.

Now, again, the same thing. So, why is the mean not exactly equal to 0? Because there is a small difference between the red line and the blue line. So, if you are simulating from a normal variable with a mean of 0, then the mean of the underlying data should also be close to 0. But why aren't you seeing something like that?

Because, again, the sample size is small. So, n is relatively small because you only have 100 observations. But what would happen again, as we discussed, if you increase this n to, let us say, something like 1000 or 5000? So, would we get better approximations? The answer is, hopefully, yes.

So now we will see one more histogram and the subsequent plot, but this time, the histogram and the plot, if you compare the means, are much better than the earlier case, and why? Because now we are creating a normal sample using not 100 but 1000 observations. So, on the left-hand side, you have a histogram of a simulated sample from a normal distribution of size 1000, and on the right-hand side, you are basically plotting all those observations, x-axis versus y-axis, and here you can clearly see that the red line and the blue line are almost superimposed. So, the mean is very, very close to 0, which we want, and here you can see a perfect kind of symmetry when it comes to capturing the data, okay. So, here the idea is that if the sample size is large, right? So, the underlying distribution would tend to a normal distribution, okay.

And then here, I think many of you might have studied this idea about CLT. So, CLT is the central limit theorem, right? So, what does the CLT tell us? So, CLT tells us that the sample average kind of converges to the population average if you have a large sample. So, by large sample, we mean that n should be at least 30, right?

And then if you have an even larger sample, then the sample average kind of converges to the population average. So, my y-bar would actually converge to the corresponding population mean. So, here we have a small comment that it is much closer to normality. But then again, the idea is that if you want to check for normality, probably drawing a histogram or confirming what the mean is would be a good idea to start with. So, if the histogram is showing you some sort of symmetric nature, then we can actually gauge that the underlying distribution might be close to a normal distribution.

So, this is one idea. Then the other idea is using some formal tests like we discussed earlier. So, let us say checking for normality using a couple of tests. So, the first one is Jarque–Bera. So, by the way, Jarque–Bera are again two people, and they proposed this test of normality.

And essentially, if you see this equation here, this JB stands for the underlying test statistic, right. So, now you should understand that any hypothesis testing problem always contains or always relies on some test statistic, right. So, here this capital JB that you see is the test statistic underlying the Jarque–Bera test, okay. And what goes on or what are the ingredients in this Jarque–Bera test are nothing but skewness and kurtosis. So, skewness and kurtosis are used to construct this test statistic, which is my Jarque–Bera.

So essentially, what is being tested here is that you are kind of testing whether the skewness or excess kurtosis are collectively equal to 0 or not. So, just a short while back, we discussed that if the underlying distribution has to be normal, then both skewness as well as excess kurtosis should be close to 0. So, this Jarque–Bera kind of proposes a unique kind of a test where collectively you can actually test whether the underlying sample skewness and the sample excess kurtosis are collectively close to 0 or not. And then in this period, they came out with this JB test statistic and so on. And by the way, this JB test statistic is based on two things: skewness and kurtosis.

So, let us say beta 1 and beta 2 represent skewness and kurtosis respectively. Then what exactly is the test statistic? So, this JB is n by 2 and then you have a bracket and then beta 1 hat square plus beta 2 hat minus 3 whole square divided by 4. And by the way, any test statistic has an underlying distribution. So, the distribution here is a chi-square distribution with degrees of freedom 2.

$$JB = \frac{n}{2}\left[\hat{\beta}_1^2 + \frac{(\hat{\beta}_2 - 3)^2}{4}\right] \sim \chi_2^2$$

If $JB > \chi_{\alpha,2}^2$ then we reject the null and conclude non-normality

So, this JB test statistic kind of follows a chi-square distribution with degrees of freedom 2. Now, how do you test for something like that? So, for example, if the underlying JB test statistic is bigger than the chi-square critical value, then one can actually reject the null. So, by the way, this idea is a pretty standard idea when it comes to applying a hypothesis testing kind of structure, right. So, you look at the test statistic, and if the underlying value of the test statistic happens to be bigger than the tabulated chi-square value, then one can actually reject the null hypothesis, OK.

Or otherwise, one can actually conclude based on the p-value also. So, if you are implementing the Jarque–Bera test in a software, let us say R, then R would actually give you a p-value, right? And then again, the same thing. So, if the p-value happens to be less than the corresponding alpha that you have chosen, let us say 5 percent or 1 percent or 10 percent, whatever.

So, if you are seeing something like this, then we can actually reject the null hypothesis. And by the way, rejecting the null, the conclusion is non-normality because for a Jarque–Bera test, H0 is the distribution is normal or normality holds, and then the alternative hypothesis is normality does not hold. So, if you are rejecting the null hypothesis, you have to go with the alternative, which kind of means that the distribution of the residuals is not normal. So, in this period, one can actually apply some formal hypothesis tests. So, let us say Jarque–Bera or one can actually apply the second one, which is Shapiro-Wilk.

So, the Shapiro-Wilk test is kind of similar. But then the idea is slightly different. So, the Shapiro-Wilk test statistic is a combination of ordered sample values and a particular constant generated from the means, variances, and covariances of the ordered statistics of a normal distribution. So, I think this definition is slightly difficult to understand, but this capital W is the Shapiro-Wilk test statistic. And then here you see a lot of ingredients, right?

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum (x_i - \bar{x})^2}$$

I mean, you see the summation of a_i. So, what is a_i? So, a_i's are some constants. Firstly, a_i's are some constants, and then x, and then within bracket i. So, x within bracket i denotes the order statistics. So, again, if somebody is not very familiar with order statistics. So, what do you mean by that? So, let us say you have a sample, right?

So, x_1, x_2 up to x_n. Now, let us say you order this sample. So, order means you keep all the observations in some order, let us say increasing or decreasing. And then you come up with this notation. So, x bracket 1, x bracket 2, so on, and then x bracket n.

So, this is my actual sample, and then these are my order statistics for that sample. So, what do you mean by x bracket 1? So, x bracket 1 means nothing but the minimum, right? Because since you ordered them, x bracket 1 is the minimum observation, x bracket 2 means the second minimum, and so on, and then x bracket n is the maximum. So, minimum, second minimum, third minimum, fourth minimum, all the way up to maximum, are my order statistics for that particular sample.

So, again coming back to this W. So, we actually order the normal observations and then we get hold of X bracket I. And then we square those, and then in the denominator, we have summation x i minus x bar square, which is kind of similar to variance, right. So, in a way, what you are doing is you are basically standardizing again, is it not? So, you are basically standardizing again, okay. Now, again, so this test statistic is a combination, like we discussed here, of ordered sample values, which are x i x x bracket i, and some constants generated from the means, variances, and covariances of particular order statistics of a normal distribution, okay.

And again, the same thing, that one can actually reject the null if the underlying values of the test statistic are too small, okay. So, if the values are too small, then you can actually reject the null hypothesis of normality. So, again, for a Shapiro-Wilk, the hypotheses do not change, so the null hypothesis is normality. Whereas, the alternative hypothesis is the distribution is not normal. So, non-normality, similar to Jarque–Bera, okay. So, in this period, one can actually choose a formal testing procedure to kind of check if the underlying residuals are indeed normal or not, okay.

And one can actually have one more visual check. So, this plot is kind of called a normal QQ plot. So, this plot is called a normal QQ plot. Now, what exactly goes on in this normal QQ plot or normal quantile-quantile plot? So firstly, one should understand what you mean by quantiles.

So, this term quantile. So, quantile is kind of similar to percentiles. So, I am sure that many of you have heard about CAT scores. So, in CAT, you get percentiles. So 90th percentile or 95th percentile and so on.

So, by the way, what do you mean by the 95th percentile? It means that 95% of the candidates who have given that exam have scored less than you. So, if your CAT score happens to be 95th percentile or let us say 98th percentile, this means that 98% of all the candidates who appeared for CAT have scored less than what you have. And then quantile is kind of similar to percentile. So, the idea is clear, hopefully. So, what gets plotted in this particular normal QQ plot?

So, we will try to understand that. So, any normal QQ plot plots two sets of quantiles against each other. So, you have a plot like this. So, x-axis and y-axis. So, on both axes, there should be some quantiles which are plotted.

So, quantiles here and then quantiles there. So, why do you see two different quantiles is that on one of the axes, we plot the theoretical quantiles of a standard normal distribution, which are plotted on the x-axis. So, let us say these are the theoretical quantiles. Quantiles and on the y-axis, we plot the sample quantile.

So, sample quantiles are generated from the sample that you have or the data that you have, and then theoretical quantiles are generated by assuming a standard normal distribution. So, for a standard normal distribution, all the quantiles are fixed; you have like fixed numbers, for example, the 95th quantile is 1.64 and so on. So, for every possible quantile, let us say 95th, 98th, or 99th, or let us say 88th, whatever. So, for all possible quantiles, you have a fixed number because the distribution is known, right? Standard normal, right. So, on the y-axis, we plot the sample quantile.

So, sample quantiles come from the data that you are analyzing, and then on the x-axis, you actually plot the theoretical quantiles assuming some distribution. And essentially, if this graph gives you a straight line, right, if this graph shows you a straight line, then normality holds. And why is that? Because both these quantiles would match, isn't it? So, since you are plotting these two quantiles, one theoretical, the other one is sample, and if both those quantiles fall on a straight line or the graph gives you kind of a straight line, then normality holds true, right, because essentially the sample quantiles and the theoretical quantiles, they match.

So, which means that the underlying assumption, which is standard normal distribution when it comes to theoretical quantiles, holds true. So, this is the idea behind a normal QQ plot or a normal quantile-quantile plot. So, for example, this is a very typical kind of a normal QQ plot where normality holds because almost all the points here fall on a straight line. Which means that the theoretical quantiles are kind of equal to the corresponding sample quantiles. And hence, which also means that the underlying distribution, which kind of produces these theoretical quantiles, is equal to the distribution of your data.

And generally, here, we are taking a standard normal, right. So, we are assuming a standard normal distribution, and then these are the quantiles for a standard normal distribution. So, if the theoretical quantiles and the sample quantiles match, which they are doing here, then the underlying distribution holds true. So, this is the idea about a normal QQ plot; then probably in the next slide, I will show you one plot where normality does not hold. So, let us say something like this.

So, here even if you try to draw a straight line, then probably the straight line would look something like that. But then you can see that you have a lot of departures away from the straight line. So, for example, here or even here, right? So, all the points are not at all close to the straight line. So, this is a clear case where normality does not hold true, basically, right?

So, normality fails. So, hopefully, the idea is clear. So, one can actually do a lot of different checks. So, let us say create some QQ plots, create some basic histograms, or let us say run some formal tests, Shapiro-Wilk, Jarque–Bera, etc. And then kind of check the assumption of normality based on any practical data.

Okay, so I think the first idea was checking for normality. Now, the second idea is how would you detect serial correlation. So, in time series, you have something called serial correlation. So, what exactly do you mean by serial correlation? So, generally, all the residuals are found to be correlated.

with some of their own lagged values. So, for example, let us say E t might be correlated with, let us say, E t minus 4, or you know, E t minus 1 might be correlated with E t minus 5, etc. So, let us say E t is correlated with one of its lagged values, or E t minus 1 is correlated with one of its lagged values. OK. And if you remember, the assumption is that ETs are independent.

So, there should not be any correlation, essentially. Right. But if they are displaying some correlation, then you have some problem with the assumption. OK. And this correlation is called serial autocorrelation.

So, this property is called serial autocorrelation of the residuals. So, here, the best way to detect this is to draw the ACF plot. So, the ACF plot is the autocorrelation function, and again, on the x-axis, you have all the lags. So, 0, 1, 2, 3, 4, up to, let us say, 40, and here, can you see a clear pattern when it comes to correlation? So, these correlations are all positive.

These correlations are all negative, then again positive, again negative, again positive, again negative, and so on. So, if you identify all the correlations which are plotted here in this plot, then you can see a clear pattern. Right. So, if you have a pattern among the correlations, then you have a serial autocorrelation problem, because what happens is, let us say, if you take any of these residuals, right, for that matter. So, let us say, if you talk about E t or rather not E t, but then let us say E t plus 10 at lag 10, something like that, and then let us say E t plus 11, or if you consider E t plus 11 and E t plus 12, okay.

So, if you look at any two consecutive correlations of errors, right? Then can't you see that both the correlations are on the same side of 0. So, for example here, if you take any consecutive correlations, these are positive, or for example here, if you take any consecutive correlations, these are negative, which means that the current error and the one error in history or in the future are correlated. So, the same thing is happening here also. If you see any two consecutive correlations, then those are on the same side of 0. So, either positive or negative.

And clearly, you see some pattern also. So, there is some underlying pattern among the correlations, which means that there is some serial autocorrelation problem in the errors. So, I will give you an example where you do not have any serial autocorrelation problem. So, if you draw the ACF of the errors, now let us say these are the hypothetical bands, then as discussed earlier, so probably after lag 1, Right.

All the other correlations should be inside the band or not significant. So, if something like this is happening, then we can safely say that the residuals do not possess a serial autocorrelation problem. But if you have any particular pattern among the correlations of errors, then we can actually say that the errors have a serial autocorrelation problem. All right. And then, by the way, you have some formal tests in this regard also.

So, one can actually go ahead and test it formally using, let us say, the Box-Pierce test or, let us say, the Ljung-Box test. So, by the way, Box and Pierce developed a test to detect serial autocorrelation back in 1970. And then this was actually modified by two other people. So, Ljung and Box in 1978. So, the Box-Pierce test or the Ljung-Box test are two tests for checking if serial correlations are there or not.

So, the underlying null hypothesis is that the data are independent. So, if the data is independent or if the residuals are independent, then there are no correlations. And of course, the alternative would be that you have a serial correlation problem. So, you have a serial correlation problem. Now, what exactly is the underlying test statistic?

So, again, since you are checking it formally using a hypothesis test, the underlying test statistic is given by the Box-Pierce test, or BP. And the test statistic is not that difficult. So, this is based on some correlation, or rather squared correlation values, if you see here. So, n multiplied by the summation of rho k hat squared, right, and then k goes from 1 to h. So, h is up to what lags you want to check, basically, right. I mean, of course, one cannot keep on checking for each and every lag, right.

$$Q_{BP} = n \sum_{k=1}^{h} \hat{\rho}_k^2$$

I mean, there should be some limit, right. So, h is the number of lags being tested in this Box-Pierce test, and n is nothing but the sample size. And again, the same criteria. So, if the value of the test statistic happens to be bigger than the tabulated chi-square value, then we can reject the null. Otherwise, we fail to reject the null.

And then again, if you reject the null, this means that there is autocorrelation, as discussed earlier on the previous slide. So, if you reject the null, you have to go with the alternative. So, which means that there is some serial autocorrelation among the residuals. And what could be a solution here? So, if serial autocorrelation persists, then the model needs to be checked again.

So the model fit itself needs to be checked again. And usually, it is better to add another lag in either the AR component or the MA component of the model. So, let us say initially you propose this ARMA 1, 2 model right using information criteria or whatever technique, then you estimate the model, and then you are coming into the diagnostic checking kind of framework, and then you apply this Box-Pierce test right, and then you find out that there is indeed some serial autocorrelation problem among the errors of this model. Then what should you do? So one solution is you should probably check by adding some orders to either an AR or MA component.

So probably try fitting, let us say, ARMA 2, 2 or probably let us say something like ARMA 1, 3 for example. And then again do the entire exercise one more time. Estimate the model. So, once you identify the orders, then apply some information criteria to choose between these two first. Then estimate the model and then proceed with diagnostic checking.

So, in a way, you can actually test out formally, much more formally, whether the distribution is normal or not, whether the errors are correlated or not, and so on and so forth. So, there is one last idea, which is how do you capture the changing variance problem, which we will discuss in the next lecture. Thank you.