Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 08

Lecture 40: Practical Session in R - 8

Hello all, welcome to this course on time series modeling and forecasting using R. Now, we are almost done with the current week, and just to summarize where we stand this week. So, the broad topic we studied this week was co-integration and, of course, causality towards the end, right? Now, again, we have a practical session, which is the last session of this week, and then, as you see, we have opened up a fresh new code, which is code 8 for the current week, which is week 8. So, just to tie all the things down, we will work with a few ideas about, let us say, cointegration using different scenarios, and then maybe how one can perform some cointegration tests on, let us say, a set of time series variables, and so on and so forth.

So, before we proceed with any of the practical stuff, or let's say the coding material, just to remind you very quickly as to what exactly one means by cointegration. So, let's say, again, if you have not really revised any of the sessions which have happened this week, then a strong suggestion would be just to go back and try to revise all the videos for this week, especially in the topic of cointegration, causality, pertaining to, let's say, the different tests of cointegration, and so on and so forth, okay? But the broad idea about cointegration is that even if you have two time series variables or time series processes, let us say XT and YT, and let us say even if individually both the time series processes, let us say Xt and Yt, are I (1). And again, what do you mean by I (1)?

So, I (1) mean integrated with a certain order of 1, right. So, integrated with an order of 1 means that both series are individually non-stationary, of course, since they are integrated. And integrated with a certain order of 1 means that if you difference each of the series individually once, right, or a single time, then one can actually get a stationary series, okay. So, the integration actually defines the number of times one has to difference the series in order to make it stationary or convert from a non-stationary series

to a stationary series. So, this is a broad idea about what you mean by integrated with a certain order.

Now, again going back to the definition of co-integration. So, let us say if you have two time series processes, Xt and Yt, which are, let us say, integrated with an order of 1. But if at all we can find a certain cointegration vector, let us say with a coefficient beta, right, such that the linear transformation of Xt and Yt, along with, of course, the coefficient beta, happens to be I (0). Or in other words, even if Xt and Yt are individually I (1) or, let us say, integrated with some order—let us say, in general, D—but if there exists some linear combination of Xt and Yt, so let us say Xt plus beta into Yt or something like that, such that this entire linear combination happens to be I (0). And again, what do you mean by I (0)?

So, I (0) mean integrated with an order of 0, which means that the underlying linear combination is stationary. So, if at all such a thing happens, then we will say that Xt and Yt are cointegrated, okay. So, again, just to summarize in one sentence: even if Xt and Yt are individually integrated with some order D, but if there exists some linear combination with that cointegrating factor beta such that the linear combination happens to be stationary or I (0), then one can say that Xt and Yt happen to be cointegrated, okay. So, hopefully, one again strong suggestion is that if you do not know or if you do not recollect any of the definitions we covered this week, then again, you have to go back, of course, and view all the videos again or maybe take some notes before you actually proceed to do the practical session, which we will cover today. Alright.

So, again, in front of you, you see a window, and then you have a couple of libraries which one requires. So, the first one is URCA, and then the second one is VARS or variables. And, of course, these are not new libraries. So, we have been studying these libraries even in the last session—of course, the last practical session—and so on and so forth. So, we will quickly incorporate these two packages inside the R environment by hitting run, okay.

So, the first one is URCA, and then the second one is VARS, right? Again, just a couple of things since you have missed any of the factors or any of the points from the last practical session: obviously, you will have to install these two packages first. Either by going here into Tools and then hitting Install Packages, or you can simply type in install.packages and then, within brackets, the name of the package. Okay. So, hopefully, once you install the packages successfully, it should show you a small message here that

the packages have been installed successfully—something like that. And once you install the packages, then you can actually load them into the R environment, okay?

And even after you load them into the R environment, it should actually throw you this message: 'Loading the required package,' let's say, sandwich, LM test. So, by the way, these two packages are inbuilt in one of these two packages, okay? So, you need not, by the way, install these two packages separately. So, if you are installing URCA and VARS, then obviously these two are hidden packages inside these two, okay? So, R just tells you what inbuilt packages one has in the packages you are basically loading now—something like that, okay?

So, hopefully, once you install the packages and then call them in the R environment, we will play around with the practical dataset upfront, which is the Canada dataset. So, data, and then Canada in brackets, okay? Now, again, if you're not very sure as to what this dataset means, you can actually type in something like a question mark and then Canada. Here it comes. Right.

And then, if you hit enter or simply run, on the right-hand side, it will show you a small description of the underlying dataset. So, this dataset's name is Canada. And then, this is a macroeconomic time-series dataset. So, as you can read the heading as well, it sort of gives you a small description. So, the original time series are published by the OECD, and the sample range is from the first quarter of 1980 until the fourth quarter of 2000.

So, this is just a small description of the timeline of the dataset. And then, the following series have been utilized in the construction of the series provided in the dataset named Canada. So, the underlying dataset contains all these main economic indicators, by the way. So, unemployment rate, manufacturing real wage, consumer price index, etc. And, by the way, each of these variables has its own code in the dataset as well.

And then, the later part of the dataset contains quarterly national accounts. So, let's say the Canadian nominal GDP. So, by the way, the entire dataset revolves around the GDP or the underlying macroeconomic factors of Canada. And probably the next thing is labor force statistics. So, one can actually go ahead and read all the variables which are involved.

So, here you see the short forms of all the variables involved in the data set Canada. For example, production, employment, etc., All right. By the way. Yeah. So, this is the description. So, PROD or prod is a measure of productivity or production.

E is used for employment. U is the unemployment rate. And then RW stands for real wage. So hopefully the data set is clear. So again. One small point I like to make is rather than jumping directly into all the coding and then getting bogged with all the coding, one should actually first get a feel of the data set first. I mean, what variables the data set contains, right? And then what do each and every variable means? So, in our case, what would be the focus variable?

So, all these points, one should actually take a note of that. So, rather than simply jumping directly into the coding one by one, right, because this could be really mind-boggling. I mean, once you load a data, but then if you do not know that what goes inside the data or what variables the data set contains, then the entire code is of no use basically, right. So, once you load any data set in the R environment, let us say Canada or for that matter any other data set, one should actually go variable by variable and try to understand that what is going on inside each and every variable in the data set. So, hopefully once you load the dataset Canada inside the R environment, then we will give a slightly different name.

So, Canada underscore TS which stands for time series and then this line which is line 7 just just converts the dataset in a time series framework. And here we can mention the starting year. So, 1980 and then frequency is 4. So, you have a quarterly dataset. Okay. So, one can run this. And then again, just to just to tell you or just to show you as to how the plot of some of the variables inside the Canada data set look look like, then we can actually plot them. And here we're actually only plotting the first four columns.

So, you can actually specify which columns the plot should contain. So, I'll run run this and then show you the plot. So, this is exactly how the plot looks like. Let me zoom in. All right. So, you see a heading which which tells you Canada microeconomic data. And on the left-hand side, you can see all the variables of interest.

So, unemployment rate, real wage, productivity and employment. And this is just a visual way of checking that how each and every variable should behave. So, for example, employment grows over all those years. Right. Same is true for real wage. Same is true for productivity.

But when you look at the unemployment rate, you can see some seasonality there. Can you see that? So, let us say unemployment is not that high initially, but it probably picks up in some of the in-between years, around 1983 or 1984. But then again, by the time

1990 arrives, the unemployment rate returns to the same levels and then picks up again. So, you can see some sort of seasonality or periodicity in the unemployment dataset.

So, a strong suggestion is that even before you try to analyze any data—and I have repeated this many times before— the very first thing one should do is try to plot the underlying data and variables just to get a feel for how they behave over time. I mean, are you seeing any trend in the data? Are you seeing any seasonality in the data, right? Are you seeing any noise or, say, changing variance patterns in the data, right?

So, all these things would be quite clear if you plot the data in the initial stages, right? So, I will close the plot now and go back to the code. And now, the very first thing we'll do is check whether each variable in the data is stationary or non-stationary. For that, I can use an augmented Dickey-Fuller test, or ADF test. Now, again, I'm not going into the details of the ADF test, as we've covered this before in one of the earlier practical sessions.

But this ADF test kind of tells you whether a series is stationary or not. And here you can actually specify a few things in this ur.df. So, ur.df is the name of the test, which is the ADF test, and we will give it a name like ADF_underscore_test_underscore_e. And then here, like I said, you can actually tweak some of the input values. For example, one can mention the type to be drift, right?

And then select lag. So, select lag is AIC. So, it sort of selects the lags automatically based on the least AIC values and so on and so forth, right? Okay. So, we will run this ADF test, and then I can get a summary of the ADF test, and then in the console, you can get a summary of the ADF test. So, initially it tells you the fitted regression line.

So, it tells you test regression drift, right? And by the way, this is the name of the test. So, the augmented Dickey-Fuller test for the unit root, which of course tells you whether the series is stationary or not, and then this is the linear model it is trying to fit—the first difference fitted on the first lag plus a constant, which is 1, plus the first lag of the difference, and this is the coefficient of the estimated regression line, and so on and so forth. And then down the line, it also throws you the R-squared value of the fitted regression line, which is, let us say, 0.52, etc. And of course, the p-value.

So, the p-value is really small. So, if the p-value is small, what do you do? So, you will reject the null, right? And then go with the alternative, right? And down the line, it tells you the value of the individual test statistics. So, let us say -0.29 for the first one and then

2.17 for the second one. And the very last thing it tells you is the critical values for observing or rather comparing the test statistic values.

For example, the first one is -0.29 or let us say -0.3 roughly for us. And then that -0.3 value should be compared to any of the first row that you see here, any of the critical values at 1 percent point, 5 percentage point, or 10 percentage point. And then, clearly, one can see that if you take the absolute value of this test statistic, it happens to be 0.3, and 0.3 is less than any of the absolute values that you see here, which means that one cannot reject the null. Similarly, if you go with the second one, the test statistic value is 2.17, and again, the same story. So, 2.17, since it is less than any of the critical values, again, one cannot reject the null.

And again, if you remember, if you go back to one of the earlier practical sessions we've covered, the null hypothesis in the ADF test is that the series is not stationary. Isn't it? So, the null hypothesis is that the series or the underlying series is not stationary. The alternative hypothesis in the ADF test is that the underlying series is stationary. So, here since you're not able to reject the null, we can actually say that both series individually are not stationary. Does this make sense so far?

So, this is exactly how one can apply the ADF test on. The only difference here is that we are applying ADF test on multiple series. And then since we are applying on more than one series, you can see a small change in the syntax here. Okay, so now the next thing we will do is we will cover an array of cointegration tests and again probably in the last session or the last-to-last session we have covered the entire theory behind how to develop the cointegration test and then what exactly do you mean by applying a test for checking for cointegration and then what all cointegration test one has, right. So, from that theory session if you remember one of the very famous tests for Johansen test right.

So, we will try to apply this Johansen test on some of the variables. So, the name is Johansen cointegration test and again quickly just to remind you that what exactly is a null hypothesis. So, null hypothesis is no cointegration of course and the alternative hypothesis is there exists some cointegration between the C's. So, we will give it a name again. So, let us say JO underscore test something like that and now the underlying function to be used is CA dot JO.

So, CA dot JO means Johansen test. So, JO stands for Johansen test right and now we will be applying this Johansen test on which variable. So, on the time series generated

variable of the data set Canada. So, Canada underscore TS. So, again remember the original data set was Canada.

And then we converted the dataset into a time series framework, right? So, and then that name was Canada_TS. So, whatever test you apply, always make sure that you are applying the tests on the time series variant of the dataset. So, for example, Canada_TS, etc. And again, inside this Johansen test, you need not pay much attention, but just to let you know that you can actually tweak a few values even here.

So, let us say type equal to trace or k equal to 2, right, and then ecdet and then spec, so transitory. So, all these are kind of part of the syntax themselves. So, we will run this Johansen test first and then again we can run the summary of the Johansen test. So, summary applied on JO_test again if you come down if you scroll down in the console. So, it sort of throws you a lot of things here, and probably I will try to explain a few ideas from the output here.

So, the name is Johansen procedure, of course. And the test type is trace statistic without linear trend and constant in cointegration. So, here you are kind of ignoring any linear trend and a constant in the cointegration values. Now, the next output thing that you see here is all the different eigenvalues that you have. So, and of course, all these eigenvalues are small.

If you see, the first one is 0.0539 because this is 5.39 into 10 to the power minus 1, which is roughly 0.0539. The next one is 0.026, right? The next one is 0.012, etc. And as you see, the eigenvalues kind of reduce in magnitude. So, the biggest eigenvalue is this one, which is 0.053.

Okay. And now, the most important part in the Johansen test is the values of the test statistic and the critical values of the test. So, again, you see a small table here. On the left-hand side, you see different values of R. So, again, remember what R means. So, R is the rank of the cointegration matrix.

Okay. Again, just to repeat, R stands for the rank of the cointegration matrix. In other words, R tells you how many variables are apt for taking, or rather for considering, cointegration. So, would they be 0, would they be 1, would they be 2, or would they be 3? So, this is exactly what the first column tells you, and then, of course, the values of the test statistic in the next column and all the corresponding critical values in the second, third, and fourth columns.

So, at different percentage points. So, 10, 5, and 1. So, this is exactly how one should read the Johansen test output. So, the first column that you see gives you the rank or the optimal rank of the cointegration matrix. The second column gives you the actual test statistic values, and all the three columns that follow give you the corresponding critical values for that particular test statistic value.

Now, again, how do you conclude anything from here? So, again, you start from the bottom up, let us say. So, the first value that you see is r equal to 0, let us say, and again, as you clearly see, for r equal to 0, the value of the test statistic is pretty high, which is 105.84. And again, clearly, this value happens to be bigger than any of the critical values at any percentage points. So, if the value of the underlying test statistic happens to be bigger than the critical values, then one can reject the null hypothesis safely, right?

And if you are rejecting the null, we can actually reject the claim that r is 0 and then go for the next level. This is hopefully clear that you go one by one. So, at the value of r being 0, we are able to reject the null. So, we can move on. Now, the next value is r is bigger than 0, but then less than or equal to 1.

And now, what happens? So again, the underlying value of the test statistic is 42.31, and again, clearly 42.31 is bigger than any of the critical values at 10 percent, 5 percent, or 1 percent. So again, we will reject the null, and then we will again ignore this level here and then move on. Okay. Now, the next one—so by this time, we know that r is not equal to 0, r is not equal to 1 also, and then r has to be bigger than 1. Okay, so the next value or the next level is r is bigger than 1 but then less than or equal to 2. Now, what happens? The underlying test statistic value is 17.39, and here you can clearly see that 17.39 is less than any of the critical values. All right.

So, the test statistic value happens to be less than the critical values. We accept the null hypothesis. Is it? So, the exact opposite. So, we accept the null hypothesis, and then if we accept the null or fail to reject the null, then we can actually go—or we have to go—with this level, which is r is less than or equal to 2. Now, once you accept the null hypothesis or fail to reject the null at any level, there is no point in going beyond that. Okay. Make sense?

So, from this one, we can actually conclude that r is bigger than 1 but less than or equal to 2. Again, just to summarize: r is bigger than 1 but less than or equal to 2. And why exactly? Because we stopped here at this level. So, from this level, we concluded that we can actually accept the null hypothesis or fail to reject the null, and hence we can actually

fix this value of r, which is less than or equal to 2, right? But r has to be obviously bigger than 1. Make sense so far?

And again, the next thing it throws you are the eigenvectors. So, both the eigenvalues as well as the eigenvectors for each and every variable in this series: employment, productivity, real wage, and then unemployment. So, this sort of tells you, or this gives you a matrix, because you have eigenvectors corresponding to each of the time series. And now, the last thing it tells you is the loading matrix. So, the loading matrix is nothing but the weights matrix.

So, again, if you remember, the Johansen test requires some underlying weight factor on each of the time series variables. So, this is the underlying weight matrix. So, this is a summary of the Johansen test. So, again, I have specifically written down in the small comments here that the underlying Johansen test applied on the Canada dataset suggests r has to be bigger than 1 but less than or equal to 2, which means that one needs a linear combination of two series for stationarity. So, since the conclusion is that r is bigger than 1 but less than or equal to 2.

So, we can actually take a linear combination of two series to achieve stationarity. So, this is the ultimate conclusion of the Johansen test. So, one has to identify the appropriate value of r, let us say either 0, or bigger than 1 but less than 2, or bigger than 2 but less than 3. And once you identify the particular value of r, then the value of the underlying r means that one requires the linear combination of how many time series to achieve stationarity. Make sense?

So, here again just to summarize one last time is that in this case one requires a linear combination of two different series to achieve stationary. Now, the next thing is just if you want to convert the Johansen test result to a VECM representation or the error correction method representation. Then one can do that and probably we are not going to detail here. So, let me just quickly run this and then summary of the underlying VECM model. But now the more important part is forecasting with ECM.

So, again I have written down forecasting with ECM or error correction model and then for this method you by the way require to convert the Johansen test result to a VECM. So, I will say that this small chunk is a requirement for forecasting. So, the first thing for forecasting is that we will try to fit a VAR representation if needed for forecasting. So, VAR stands for vector autoregressive and then we give a name let us say VAR

underscore model and then this is the underlying function or the inbuilt function. So, VEC to VAR applied on the Johansen test R is 1.

So, we will run this first and then forecast. So, once you fit the model or once you fit the VAR model, then forecast for 10-time steps ahead because n dot ahead is 10. So, again R has an inbuilt command called as predict. We have been using predict number of times even in a few last practical sessions as well. So, forecast and then predict of the fitted VAR model and then how many time steps ahead? So, 10.

So, we will run this, and now the last thing to do is plot the forecasts. Now, again, one point to mention here is that generally, what happens is if your plot is very large, right, then R can throw this error. It says that error in plot: new figure margins are too large. So, in that case, what you do is you basically run this code, okay. You basically run this code and then again run the plot command one more time, okay.

So, now, can you see the plot? So, let me zoom in, okay. So, this is the forecasted plot, and then it sort of tells you that after you apply the VAR model, this is exactly how the individual series would forecast. So, forecast for the employment rate, then productivity, real wage, and then unemployment. Make sense?

So, this is exactly how the forecasting exercise looks like. And now, the next thing we will do is we will play around with a few other cointegration tests apart from the Johansen test. So, again, we will incorporate the same dataset, which is Canada. So, we will again call the same dataset. And here, if you remember from the Johansen test, it kind of concluded that one requires the linear combination of two series.

So, here we will only focus on two series. So, the first one is productivity, and the second one is the employment rate. So, Y is productivity, and X is employment. Now, the first test we will cover is the Engel-Granger test, and then we will give it a name: eg_test. So, this tells you the Engel-Granger test and then the summary of the test.

So, again, just to remind you of one thing: the summary of the test—the only thing one should watch out for—is the value of the test statistic and the underlying critical value. So, again, here one can see that the test statistic value is 14.33, which is less than any of the critical values. So, one fails to reject the null. And if you fail to reject the null, then we can say that the residuals are non-stationary or the series are not co-integrated. And hopefully, the same conclusion can be drawn from the Johansen test as well.

So, now the only difference with the Johansen test is that this is only applied to two variables. So, X and Y, which were productivity and employment. So, summary of the Johansen test. So, again, the same conclusion: the residuals are not stationary, and the series are not co-integrated. And now, the last thing we can do is run the Granger causality test.

So, does x causes y or does y causes x, right? So, can we somehow speak anything about causality between x and y, ok? So, again on the same data set, so data set is Canada and then here we can difference the data set once to achieve stationarity because again if you remember Granger causality can only be applied if the series is stationery. Now, choosing the lag length. So, we can choose a particular lag length again based on the minimum AIC value.

Then we can fit a VAR model right on the different series and the last thing is we can perform the Granger causality test. And here we are performing if test if E Granger causes the other variable which is. So, here we are testing if the employment Granger causes the productivity ok. And for this, we have this Hopier's test, right? For checking this, we have this Hopier's test, right?

And then again, we will input Y and then X and then fit some Arima models, extract the residuals, right? And then calculate the cross correlation between the residuals and calculate the Hopier's test statistic, right? And then compare that with the chi-square critical values. And then this is the result. So, Q statistic and then the P value.

So, the p-value happens to be 0.32, which is bigger than the null hypothesis. So, we fail to reject the null, and if you fail to reject the null, it means that no causality exists between the series. So, I sort of rushed through the later part of the code, but again, take your time. So, since you already have the code, spend some more time exploring more about, let us say, the causality aspect. So, we spent a whole lot of time on co-integration, but not that much on causality.

So, again, a strong suggestion is to go back and try running the causality aspect of the code one more time. So, try to understand whether employment Granger-causes productivity or how you apply this Hopier's test, and so on and so forth.

Thank you.