Time Series Modeling and Forecasting **Spring 2020**

Midterm Exam Answers

1. (10%) What is meant by seasonality in a time series context? What is the simplest robust way to control for it?

- Seasonality refers to systematic patterns in the data related to its temporal frequency, e.g. hourly, daily, weekly, monthly, quarterly, annual.
- Two series may appear to be related to one another simply because they share similar seasonal patterns, for example sales of ice cream and swim suits may both be higher over summer, just because it is warm, not because one drives the other.
- An easy and robust way to control for seasonal effects is to include an indicator variable for the different "seasons" (one less in total than the number of seasons).

2. (10%) What is covariance stationarity? Why do we care?

- To be precise, given a time series x_t, $COV(x_t,x_{t+h})$ depends only on h, not t.
- Intuitively, the correlation between elements of the series may depend on how far apart in time they are, but cannot depend on when in time they are.
- This is important because we count on the relationships across time being stable to estimate parameters and standard errors, or make forecasts. If the relationships across time change unpredictably with time, we could never learn anything useful for application to future data.
- In practice, the longer the window of analysis, the more likely there have been structural breaks that cause a lack of covariance stationarity. But, the shorter the window of analysis, the less precisely we can estimate the relationships of interest. So, there is potentially a trade off between the degree of violation of covariance stationarity and the amount of data available to estimate the model.

3. (10%) We discussed the Wold representation theorem and AR←→ MA invertibility at length. Explain why these things are important to our understanding of how to estimate accurate approximations of time series processes with simple and robust statistical procedures.

- Approximately, and for our purposes, the Wold representation theorem says covariance stationary and weakly dependent time series may be exactly represented by the sum of a deterministic component and an infinite moving average process involving a white noise residual. This is all it says. The rest of the answer involves combining this with other things we know to get somewhere useful.
- Since regression returns the expected value of the response variable, the deterministic portion of the process can be approximated by a distributed lag model.
- While MA processes are harder to estimate robustly, AR processes are easily and robustly estimated using ordinary least squares.
- With these conditions, it is also true that any MA (AR) process can be converted to an infinite AR (MA) process.
- If the current effect of past shocks goes to zero relatively fast, which is usually true in relevant applications, the approximation will be accurate with a relatively small number of terms, rather than an infinite (and therefore inestimable) number.
- Putting all this together, most time series can be well approximated by an ARDL model with a limited number of lags (the best approximation may still have a large residual) that is readily estimable using OLS.

Midterm Exam Answers

4. (10%) What does dynamically complete mean? Why do we care? How do we check whether a model is dynamically complete?

- Dynamically complete means adding more lags of predictor or response variables will not improve prediction of the response variable. This means the residuals are not systematically related to unincluded lags.
- The first reason to care is that, for forecasting purposes, if we can improve predictive power by adding lags, we should.
- The second reason to care is that, for purposes of inference, standard errors are wrong in the presence of serial correlation. If the residuals are independent of unincluded lags, there is no serial correlation.
- One useful way of checking for dynamic completeness is to examine the partial autocorrelagram of the residuals.
- A more formal way is to conduct a Breusch-Godfrey test for serial correlation. If we cannot reject the null hypothesis of serial correlation, we can proceed on the assumption the model is dynamically complete.

5. (10%) Consider the ARDL model $y_t = \alpha + \beta x_t + r_t$ where x is an exogenous predictor variable and the residual r follows the AR(1) process $r_t = \rho r_{t-1} + \varepsilon_t$ in which ε is a white noise disturbance. Derive the dynamically complete version of the model.

$$\begin{aligned} y_t &= \alpha + \beta x_t + \rho r_{t-1} + \varepsilon_t \\ &= \alpha + \beta x_t + \rho\left(y_{t-1} - \alpha - \beta x_{t-1}\right) + \varepsilon_t \\ &= \left(1-\rho\right)\alpha + \beta x_t - \rho\beta x_{t-1} + \rho y_{t-1} + \varepsilon_t \end{aligned}$$

6. (10%) Interpret the output provided for question 6. Include an explanation of what that output implies for modeling relationships among these time series.

- The output is related to checking whether or not (log) retail sales are highly persistent.
- The AC for retail sales is consistent with a strong AR process.
- The PAC for retail sales shows the first order autocorrelation coefficient is near 1.
- After differencing, the AC and PAC show no indication of strong dependence.
- The dickey fuller test provides no evidence whatsoever against the null hypothesis that retail sales are I(1) (that the first order autocorrelation coefficient is not 1). (But, if it is near one, even if we know it is not exactly one, it would still be best practice to difference, anyway.)
- This all implies retail sales should be differenced before conducting further analysis.

Midterm Exam Answers

7. (20%) You are interested in whether the Warehousing and Storage industry or the Leisure and Hospitality industry contribute more to the area's economic base. You estimate Models 1 and 2 to shed light on this question, working under the assumption that the sector with the larger impact on the economic base will have a larger association with induced (or derived) retail employment. Note that variables are not in logarithmic form in these models.

a. What is the difference between models 1 and 2? Why does it matter?

Model 2 uses Newey-West standard errors, which are robust to heteroskedasticity and serial correlation. It matters because inference about the question of interest is not valid unless we either rule out these things or use standard errors that are robust to them.

b. What is the cumulative effect of a one unit increase in warehousing and storage employment on retail employment?

The initial increase is 1.58, followed by an additional 0.3 after one month, a reduction of 0.28 after 2 months, and an additional increase of 0.93 after 12 months, for a cumulative effect of 2.53.

c. What is the cumulative effect of a one unit increase in leisure and hospitality employment on retail employment?

The initial increase is 0.39, followed by a decrease of 0.9 at one month, a further decrease of 0.40 at 2 months, and an additional increase of 0.21 at 12 months, for a cumulative effect of 0.1.

d. Provide and defend an answer to the question of which sector is more important to Lakeland's economic base based on this output.

As is clear from the answers to (a) and (b), the estimated induced increase in retail employment is much larger for employment expansion in the warehousing and storage (WHS) sector than for the leisure and hospitality (LH) sector. Examining the standard errors of the individual effects in Model 2, the effects of lags 1 and 2 for WHS may not be different from zero. However, they largely cancel one another anyway, and so together contribute almost nothing to the total effect anyway. These effects are estimated with robust standard errors, so we can have confidence they represent statistically significant correlations. Based only on this analysis, WHS is more integral to the areas economic base than LH.

Midterm Exam Answers

8. (20%) Consider the output for models 3-6. Which is most promising as a basis for forecasting retail sector employment? Why? Thoroughly support your answer. Note that variables are in logarithmic form in these models.

Fit statistics for all four models, along with the results of a Breusch-Godfrey test for serial correlation, cumulative through lag 12, are shown in the table below. Based only on these, the models perform nearly identically, with a very slight edge in favor of Model 5.

Model	Akaike	Bayesian	N-Fold	10-Fold	K	Breusch-Godfrey
3	-2102.2	-1732.7	0.0123	0.0128	96	0.0016
4	-2119.6	-1842.4	0.0118	0.0123	72	0.0010
5	-2096.5	-2001.1	0.0107	0.0107	25	0.0026
6	-2048.7	-1983.3	0.0122	0.0128	17	<0.0001

Model 3 contains all variables, and all lags form 1-12. Joint hypothesis tests following Model 3 suggest warehousing and storage employment and building permits should be dropped from the model. However, the Breush-Godfrey test after Model 3 suggests it was not dynamically complete, and these tests were not robust, so the results should be taken with caution. Model 4 drops them and keeps lags 1-12, model 5 drops them and keeps only lags 1, 2, and 12, but adds lag 24 of the dependent variable. Model 6 is purely autoregressive with lags 1-4 and 12.

Model 6 is most parsimonious (K=17), but seems to ignore relevant information based on the (suggestive only) joint hypothesis test results. Model 5 is the most parsimonious of the others (K=25), by far (compared to K=96 and K=72), and is slightly better on the cross-validation measures and the BIC. So, on the combination of parsimony, joint hypothesis tests, and model selection measures, model 5 is arguably the best of these. But, it does not appear dynamically complete, so some other model might be better. [One could make an argument for Model 6, which is purely autoregressive, as it is simplest and does not perform much worse than the others. There is no argument for 3 or 4.]