Time Series Modeling and Forecasting **Final Exam Prep and Out of Class Work v1**

What to turn in

This is mostly to guide your study for the in-class final exam, but I will use it to support grading what you do on the in-class exam and I will also include some weight on the portions not reflected on the in-class exam—you can't do all of this in 2 hours. However, including the project, I think you have written enough formal reports for the semester, so don't worry about drafting this in terms of a formal report for a client. Instead just provide reasonably through yet concise answers to each question in the order they are asked. Make sure to provide any figures asked for in the appropriate order and appropriate place. As always, attach your do and log files. Upload it on canvas and bring a paper copy to the exam to make my life easier when grading.

Part A: Some Math, Theory, Representation, and Specification Questions

1) Derive the moving average representation of the AR(1) process $u_t = \rho u_{t-1} + \varepsilon_t$ where 0<ρ<1 and ε is a typical mean zero disturbance.

2) Derive the autoregressive representation of the MA(1) process $u_t = \varepsilon_t + \theta\varepsilon_{t-1}$ where 0<θ<1 and ε is a typical mean zero disturbance. Explain why for modeling purposes a limited number of AR terms can represent this process well even though there will be an arbitrarily large number of AR terms in the exact representation you derived.

3) Suppose you have data well modeled by $y_t = \alpha + y_{t-1} + \delta_1 q_{1t} + \lambda t + 0.5\gamma t^2 + \beta x_t + \varepsilon_t$ where q_{1t} is 1 t is in quarter 1 of the year and 0 otherwise, x is a typical predictor variable, and ε is a typical mean zero disturbance. Write out the first difference. Explain why the first difference of y, while trending, is trend stationary—that is it will return relatively quickly to trend after a shock, while the undifferenced y shows an accelerating time trend and is not stationary around that accelerating trend.

Data for Parts B and C

The file "fl and us monthly data.csv" contains nine monthly non-seasonally adjusted time series related to US and Florida economic performance. In addition to year and month, for the exam we will use the natural log of these:

- fl_nonfarmemp_m: Florida's total nonfarm employment.
- fl_amsparksarcadesemp_m: Florida employment in amusement parks and arcades.
- fl_bldpmt_m: Florida housing units authorized by building permit.

Tourism and retirement are important drivers of Florida's economy, and thus of total nonfarm employment. The retirement related sector shows up in a large construction industry—building homes for retirees and homes and places of businesses for those that serve them. These two drivers are somewhat captured by employment in amusement parks and building permit activity. We could get better fitting models by including other variables, but that is not the point here—the point is just to demonstrate your knowledge in a reasonably realistic context.

Part B: Time Series Modeling (Not Forecasting)

1) Estimate a model of nonfarm employment as a function of contemporaneous (lag 0) employment in amusement parks and arcades and building permit activity. Control for a linear time trend and monthly seasonal effects. Use OLS on undifferenced data. Is the model consistent with expectations about the role of tourism and construction?

2) Visually and statistically determine whether nonfarm employment should be differenced before running OLS. Explain what this means and why it matters.

3) Estimate the model using OLS on differenced data. Comment on changes worth noting.

4) Arguably the impact of construction is spread out over many months. Similarly, adding another employee in amusement parks may be associated with related hiring later. Estimate a distributed lag (DL) model that includes lags 0-1 for amusement park employment (log differences) and lags 0-12 for building permits (log differences). Do the results fit the story about the importance of construction and amusement parks? What insight do you get from the lags?

5) It is useful to think of the economy as continually adjusting toward equilibrium in response to a never ending series of shocks. In a partial adjustment model, today's value is an average of past values and today's equilibrium value because it takes time to converge. In other models if last period's value was a particularly large jump it is more likely to have represented a transitory event so a correction in the opposite direction might be expected. Such reasoning suggests including lagged values of the dependent variable as predictor variables. Since the periodicity of our data is a year, one might argue for including lags 1 and 12 of the (log difference of) nonfarm employment to capture adjustment processes. Estimate such an autoregressive distributed lag (ARDL) model. Interpret the autoregressive terms.

6) Conduct a Brush-Godfrey test for autocorrelation on the results of the previous model, examining lags 1-12, and interpret the results.

7) Estimate the model from (5) using the Newey estimator with a 12 lags. The rule of thumb we covered would suggest less than 12. How many? Why then do I suggest 12?

Part C: Forecasting

Due to the number of lags to search through for this problem, GSREG or similar automated searches would take more time than they are worth for our purposes. Instead you will compare a few reasonable models using LOOCV to choose the "best" one from that smaller set.

1) Estimate a model to forecast the log change of January 2018 nonfarm employment using lags 1-12 of all three variables and month dummies. Obtain the LOOCV RMSE.

2) Based on the results from (1), estimate a few other versions of the model you think might perform better and compare them using the LOOCV RMSE.

3) Use the best model from (1) and (2) to forecast the log difference of nonfarm employment for January 2018. Explain your choice. Obtain the 95% forecast interval.

4) Prepare a figure showing the point and interval forecast of the log difference from January 2007 on for visual evaluation of the model. Use the figure to discuss the model's fit.

For questions 5-8: Generate the two month difference (NOT the second difference) of nonfarm employment by subtracting the second lag of the log of Florida nonfarm employment from the current value. Call it d2mlnflnonfarm.

5) Estimate a model to forecast the two month log change of February 2018 nonfarm employment by regressing d2mlnflnonfarm on lags 2-12 of all three variables and monthly dummies. Obtain the LOOCV RMSE.

6) Based on the results from (5), estimate a few other versions you think might perform better and compare them using the LOOCV RMSE.

7) Use the best model from (5) and (6) to forecast the two month difference of nonfarm employment for February 2018. Explain your choice. Obtain the 95% forecast interval.

8) Prepare a figure showing the point and interval forecast of the log difference from January 2007 on for visual evaluation of the model. Use the figure to discuss the model's fit.

Combining horizons and converting back to nonfarm employment

9) Using the work from (4) convert the point forecast for the January 2018 log difference of nonfarm employment and the upper and lower bounds of the corresponding forecast interval to the corresponding values of nonfarm employment.

10) Using the work from (7) convert the point forecast for the February 2018 two month log difference of nonfarm employment and the upper and lower bounds of the corresponding forecast interval to the corresponding values of nonfarm employment.

11) Plot nonfarm employment from January 2017 through December 2017 along with the point forecast and forecast interval for the first two months of 2018. To do this create three variables: Predicted, Upper, and Lower. Leave them missing before December 2017. For December 2017 set them all equal to the actual. For January 2018, use the one step values from (9). For February the two step values form (10). Then use tsline to make a chart for Nonfarm Employment, Predicted, Upper, and Lower against time. This should give you something like a "cone of uncertainty", known in time series modeling as a fan chart or a river of blood chart.

12) Briefly discuss the forecast, justifying major modeling choices or assumptions and giving the proper interpretation of the cone of uncertainty.