Assignment #5

*Submit your **printed** solution at the start of class on Thursday, November 21.*

Short answers, please. *Explain your answer concisely. If you refer to a plot in your answer, include that plot as part of your answer.* Do not include extraneous plots that you do not refer to in your narrative. "Significance" implies statistical significance. Presume necessary conditions for inference hold unless the question addresses these.

The data for this analysis is in the file a5_furniture_sales.csv. The time series "sales" defined in this file gives the monthly sales at US furniture stores in millions of dollars from January 1992 through August 2024. You should plot the full time series before segmenting the data required in the following questions.

For these initial questions, use the data prior to Covid, January 1992 through December 2019 (28 years). Throughout, we will model the log of the sales.

1. Explain the need for the log transformation; it isn't so clear in this application. To support your answer, compare the relationship between the annual mean and SD of sales to the relationship between the annual mean and SD of the log(sales) data.

   To compute these summary statistics, you will find the R function `tapply` useful. For example, the code shown at the right computes annual means and standard deviations of the sales data.

```
year <- floor(time(sales))
m <- tapply(sales, year, mean)
s <- tapply(sales, year, sd)
```

2. Is the seasonal pattern in log(sales) consistent over 1992-2019?

   To form an answer, compare the seasonal variation in 3 periods: 1995-1996, 2005-2006, and 2015-2016. Your analysis should be graphical and support a visual comparison the seasonal variability in the log(sales) data.

3. Summarize a basic regression model that predicts log(sales) in 2020-2021 fit to the data in 2011-2019. As part of your summary, show and discuss a sequence plot of the these 9 years of sales with the fitted values from this model.

   For predictors, use a linear time trend, dummy variables for month, and an AR(1) residual predictor. (If you find this choice of time period puzzling, look back at the original sequence plot of log(sales).) The following R code does the heavy lifting. Make sure you understand what it's doing.

```
nYears <- 11
yt   <- window(log(sales), start=2011)
tt   <- 2011 + 0:(nYears*12 - 1)/12          # want two extra years  (11 year total)
mnth <- factor(rep(month.abb, nYears), levels=month.abb)
X    <- model.matrix( ~ tt + mnth - 1)[,-2]  # omit january
regr <- sarima(yt, p=1,d=0,q=0,   xreg=X[1:((nYears-2)*12),-13])
```

4. Use the model estimated in Q4 to predict sales in 2020 – 2021. Show a graphical summary of the prediction intervals of the fitted model. [Use `sarima.for` to produce these results. The command is *very* similar to that used in Q3 to fit the model.]

The following questions build an SARIMA model for the log(sales) time series using the full time period 1992-2019.

5. Build time series that give the number of weekdays and the total number of days for the 30 years 1992-2021. As a summary, show and discuss the ACF and PACF of the count of days in a month. [R code that does this calculation in a different problem is

given in Lecture_20.Rmd, but you don't have to use that approach. You need the extra 2 years to prepare for forecasting out 24 months.]

6. Identify an initial SARIMA model from the ACF/PACF for the month-to-month differences of log(sales) time series using the 28 year period 1992-2019.
7. Fit the proposed model, including as non-stochastic predictors the time series of weekdays and total days in a month. Summarize the coefficients in the estimated model and note any issues with lack of fit.
8. Revise your initial model to improve the fit. Use the estimated coefficients, estimated error variance, BIC statistic, and residual diagnostics to support your choice. [Don't get carried away as you're not going to reduce all of the residual autocorrelations. For example, you will probably find an "significant" residual autocorrelation at lag 33. Don't try to add parameters to fit this one.]
9. Show a graphical summary of forecasts with 95% prediction intervals from the model estimated in Q8 for 2020-2021. Include the actual data for the forecast period in this graph. How have the forecasts from the SARIMA model performed?
10. Compare forecasts from the SARIMA model to those produced in Q4 using the basic regression model. How well do the models perform?