Statistics 435/535/711 26 January 2022
P. Shaman

Assignment 1—due 9 February at 11:59 pm

You may work together for this assignment. However, perform calculations and the writeup by yourself. File sharing is not permitted. Submit via Canvas. Your submission should be either a Word or a pdf file. R Markdown documents will not be graded.

Your submission should include clear focus on discussion and interpretation of the statistical results in your presentation.

The data set for this assignment gives grocery store sales in the U.S., monthly in millions of dollars, for the years 1992 through 2021. The data are in the file GroceryStoreSales.txt.

To do this assignment, you need to convert the class for two variables given in the data frame. To do so, employ these steps at the outset:

Read in the data frame using this form of command:

```
sales<-read.csv("F:/Stat71122Spring/GroceryStoreSales.txt")
```

For example, give the name *sales* to the data frame. Next, give these commands:

```
attach(sales)
Time<-as.numeric(Time)
fMonth<-as.factor(Month)
```

The last two lines convert the variable *Time* to numeric class and the variable *Month* to factor class. In addition, augment the data frame using the following command:

```
sales<-data.frame(sales,fMonth)
```

1. During the years 1992 to 2021 the Business Cycle Dating Committee of the National Bureau of Economic Research defined three periods of economic contraction, 2001(4) to 2001(11), 2008(1) to 2009(6), and 2020(3) to 2020(4). Make separate time series plots for (i) Sales and (ii) log Sales, and mark the contraction periods on the plots. Discuss and compare the two plots. Comment on trend structure and volatility. Do the plots reveal any unusual features? If yes, describe what is notable and discuss the underlying causes. Do the two plots indicate whether an additive decomposition model or a multiplicative decomposition model should be fit to model Sales? Explain your answer.

2. Fit a multiplicative decomposition model to the Sales data. Include a polynomial trend, a seasonal component using the *fMonth* variable, and trigonometric variables to investigate calendar structure. If you find an outlier value which warrants use of a dummy for adjustment,

form the dummy and include it in your model.) Comment on and interpret the statistical results. (The calendar pairs are discussed on pages 38–39 of the 24 January notes.)
[R hint: To fit a fourth-degree polynomial trend, for example, include as explanatory variables in the `lm` command

```
Time + I(Time^2) + I(Time^3) + I(Time^4)
```

As an alternative, you can use `poly(Time,4)`

These two approaches give identical overall fits, but produce different coefficient estimates. The latter employs orthogonal polynomials, and the former does not. Either form can be used for this assignment—the overall results will be the same.]

If a calendar pair is not significant, remove it and refit.

(a) Tabulate and plot the estimated static seasonal indices and give a detailed interpretation of them in the context of the data collection.

(b) Save the residuals from the fit. Form a normal quantile plot of these residuals, test the residuals for normality, plot the residuals vs. time, and plot their autocorrelations. Describe each of these results. What conclusions do you draw from the residual analysis? In particular, what structures in the time series has the model failed to capture? Give a careful and detailed description in answering these questions.

3. Use the `decompose` function in R with a multiplicative formulation to construct seasonal index estimates for the Sales data. Compare these static seasonal index estimates to those from part 2(a) using (i) a table and (ii) a plot. Discuss what this reveals.

4. Calculate the lag 1, lag 2, and lag 3 residuals from the model in part 2 and add these three new variables to the model you fit in part 2. Be sure to include the calendar trigonometric pairs if they turn out to be significant now. Then perform a residual analysis for this new model. What improvements do you notice? [R hint: The following code will create these lagged variables (fill in the name of the model you fit in part 2):

```
lresid<-c(rep(NA,360))
lag1resid<-lresid;lag2resid<-lresid;lag3resid<-lresid
lag1resid[2]<-resid(model)[1];lag1resid[3]<-resid(model)[2]
lag2resid[3]<-resid(model)[1]
for(i in 4:360){
i1<-i-1;i2<-i-2;i3<-i-3
lag1resid[i]<-resid(model)[i1];lag2resid[i]<-resid(model)[i2]
lag3resid[i]<-resid(model)[i3]
}
```

5. Perform the following analysis to investigate the presence of dynamic seasonality.

(a) Analyze the data spanning the years 1992 through 2007 and calculate the static seasonal estimates.

[R hint: To exclude the data for the years 2008 through 2021, you can use the following command (assuming you decide to use a fourth-degree polynomial in time for the trend, e.g.):

```
model<-
lm(logSales~poly(Time,4)+fMonth+c348+s348+c432+s432,data=sales[1:192,])
```

In this command `sales` is the name of the augmented data frame.]

(b) Repeat part (a) for the years 2008 through 2019.

(c) Give a careful comparison, using a table and a plot, of the results in parts (a) and (b). Do you find evidence of dynamic seasonality?

6. Write a summary of the results obtained from parts 2 through 5.