Statistics 5350/7110 12 September 2023

P. Shaman

Assignment 1 solution

***You may work together for this assignment. However, perform calculations and the writeup by yourself. File sharing is not permitted. Submit a Word file or a pdf via Canvas.***

***This assignment is tightly scripted, and thus your submission should focus clearly and thoroughly on discussion and interpretation of the statistical results in your presentation.***

The data set for this assignment is NewCarRetailSales.txt. It gives monthly sales of new cars in millions of dollars in the U.S. for the period from 1992(1) to 2023(6).

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:

carsales<-read.csv("F:/Stat711023Fall/NewCarRetailSales.txt")

That is, give the name *carsales* to the data frame. Next, give these commands:

attach(carsales)

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:

carsales<-data.frame(carsales,fMonth)

> carsales<-read.csv("F:/Stat711023Fall/NewCarRetailSales.txt")

> attach(carsales)

> Time<-as.numeric(Time)

> fMonth<-as.factor(Month)

> carsales<-data.frame(carsales,fMonth)

> head(carsales)

Date Sales logSales dlogSales Month Time obs118 obs213 obs339 obs340

1 1992-01-01 24056 10.08814 NA 1 1 0 0 0 0

2 1992-02-01 25041 10.12827 0.040130036 2 2 0 0 0 0

3 1992-03-01 28018 10.24060 0.112332679 3 3 0 0 0 0

4 1992-04-01 27982 10.23932 -0.001285714 4 4 0 0 0 0

5 1992-05-01 28924 10.27243 0.033110254 5 5 0 0 0 0

6 1992-06-01 30591 10.32846 0.056034148 6 6 0 0 0 0

obs341 c348 s348 c432 s432 fMonth

1 0 -0.57757270 -0.6411532 -0.9101060 0.4143756 1

2 0 -0.33281954 -0.9840585 0.6565858 -0.7542514 2

3 0 0.96202767 -0.8692052 -0.2850193 0.9585218 3

4 0 -0.77846230 -0.3500202 -0.1377903 -0.9904614 4

5 0 -0.06279052 0.3319852 0.5358268 0.8443279 5

6 0 0.85099448 0.8595596 -0.8375280 -0.5463943 6

1. Make separate time series plots for (i) *Sales*, (ii) log of *Sales*, and (iii) the log return of *Sales*. List and mark the periods of economic downturn as determined by the Business Cycle Dating Committee. Discuss and compare the three 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 plots in (i) and (ii) indicate whether an additive decomposition model or a multiplicative decomposition model should be fit to model sales? Explain your answer.

For the span of the new car sales data there were three time periods judged to be contractions by the Business Cycle Dating Committee. They are

2001(4) to 2001(11), 2008(1) to 2009(6), and 2020(3) to 2020(4).

The first was caused by a drop in manufacturing, and perhaps was also a consequence of the 2000 dot com bubble. The second was the recession caused by a financial crisis which involved, among other problems, inflated real estate prices. The third, only two months in length, occurred at the onset of the COVID pandemic.

> Salescontraction<-c(rep(NA,111),Sales[112:119],rep(NA,73),Sales[193:210],rep(NA,128),Sales[339:340],rep(NA,38))

> plot(ts(Sales,start=c(1992,1),freq=12),ylab="Sales",main="Monthly US Car Sales, in Millions of Dollars",col="green",lwd=2)

> lines(ts(Salescontraction,start=c(1992,1),freq=12),col="red",lwd=2)

> legend("topleft",legend=c("expansion","contraction"),col=c("green","red"),lty=1,cex=0.8)

A graph showing a graph of a car sales

Description automatically generated with medium confidence

**The plot shows upward trending during the period from 1992 to 2001. There is a leveling in 2001, in response to the 2001 recession. This recession occurred in response to the bursting of the dot com bubble. The low sales figure in September 2001 and subsequent very high sales in October 2001 are clearly visible. From 2002 to 2007 sales are essentially level. During the 2008–2009 great recession there is a substantial drop. When the recession ends in mid-2009, sales begin a steady increase lasting until the onset of COVID in March 2020. The derivative does slow somewhat in 2016. When COVID appeared in March 2020, retail activity declined, consumer confidence was upset by the pandemic, and supply chain problems affected the auto industry and other sectors of the economy. Pent-up demand led to a very high level of new car sales in 2021, and the rise in sales has continued to mid-2023.**

**In addition to the evident trending, the time series exhibits a very strong seasonality, with sales relatively high in March and May through August, and relatively low from November to February (except that the December level is close to trend). Volatility of the time series does increase as the level rises, suggesting the need for a multiplicative model, rather than an additive one.**

**The most unusual features of the plot are seen during the marked recession periods, especially with sales being severely affected in response to the onset of the COVID pandemic and the uncertainties it generated.**

A graph showing a graph of a car sales

Description automatically generated with medium confidence

**This plot of the log of sales closely resembles the plot of sales. The log transformation appears to be somewhat of an overcorrection, though, as volatility in the early years is now greater than that in the later years. Although the above two plots are not conclusive, my preference is to fit a multiplicative decomposition model to the data.**

A graph showing a graph of a car sales

Description automatically generated with medium confidence

**The differencing operation used to give the log return series has eliminated the trend, except during the 2008–09 recession, but the remaining trend there is slight. There is a very strong seasonal structure, although it is a modification of the original seasonal. There are a few episodes of increased volatility, during the 2001 recession, following the 2008–09 recession, and after the onset of the COVID pandemic. However, these are brief in duration, and otherwise the volatility is rather stable.**

2. The data frame has five variables which address outliers. For each of these specify the month and year of the outlier and explain why the sales figure was unusual.

**118, October 2001. September 2001 was the month of the 11 September airplane attacks on the World Trade Center and the Pentagon. This resulted in a sudden decline in economic activity, and thus low car sales in September 2001. Recovery was quick, and October 2001 sales were especially high, partly because automakers offered substantial discounts.**

**213, September 2009. This month, after recovery from the 2008–09 recession was starting, showed relatively low sales for September. The federal government established the Consumer Assistance to Recycle and Save (CARS) program, also called “Cash for Clunkers,” beginning in July 2009. Consumers were offered a credit of $3,500 or $4,500 to trade in an old vehicle in exchange for a new more fuel-efficient one. Three billion dollars were allocated, and much of it was spent in July and August. Heavy sales in those months led to lower than usual sales in September 2009. Opinions about the success of the program varied. Some argued it stimulated the economy and reduced pollution. Others asserted it led to a shortage of used cars, thereby increasing used car prices and hurting low income persons.**

**339–341, March to May 2020. As noted above, the COVID pandemic hit the U. S. noticeably in mid-March 2020. March and April sales were depressed, and sales rebounded sharply in May.**

3. (a) Fit a multiplicative decomposition model to *Sales.* Include a polynomial trend, a seasonal component using the *fMonth* variable, trigonometric variables to investigate calendar structure, and outlier dummies. Retain only significant trigonometric pairs, and only significant outlier dummies.

[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.]

> model1<-lm(logSales~poly(Time,5)+fMonth+obs118+obs213+obs339+obs340+obs341);summary(model1)

Call:

lm(formula = logSales ~ poly(Time, 5) + fMonth + obs118 + obs213 +

obs339 + obs340 + obs341)

Residuals:

Min 1Q Median 3Q Max

-0.32243 -0.03908 0.01334 0.05766 0.21895

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.79825 0.01527 707.048 < 2e-16 \*\*\*

poly(Time, 5)1 5.24901 0.08712 60.247 < 2e-16 \*\*\*

poly(Time, 5)2 0.16394 0.08694 1.886 0.06016 .

poly(Time, 5)3 1.77618 0.08658 20.515 < 2e-16 \*\*\*

poly(Time, 5)4 -0.93842 0.08681 -10.810 < 2e-16 \*\*\*

poly(Time, 5)5 -0.62698 0.08726 -7.185 3.96e-12 \*\*\*

fMonth2 0.02894 0.02159 1.340 0.18097

fMonth3 0.20134 0.02177 9.249 < 2e-16 \*\*\*

fMonth4 0.14641 0.02177 6.725 7.02e-11 \*\*\*

fMonth5 0.19376 0.02177 8.899 < 2e-16 \*\*\*

fMonth6 0.16950 0.02160 7.847 5.03e-14 \*\*\*

fMonth7 0.17155 0.02177 7.880 4.02e-14 \*\*\*

fMonth8 0.19158 0.02177 8.800 < 2e-16 \*\*\*

fMonth9 0.10456 0.02196 4.762 2.80e-06 \*\*\*

fMonth10 0.08818 0.02196 4.016 7.23e-05 \*\*\*

fMonth11 0.03057 0.02178 1.404 0.16124

fMonth12 0.09144 0.02178 4.199 3.40e-05 \*\*\*

obs118 0.24585 0.08816 2.789 0.00558 \*\*

obs213 -0.31691 0.08814 -3.595 0.00037 \*\*\*

obs339 -0.50130 0.08841 -5.670 2.95e-08 \*\*\*

obs340 -0.62470 0.08841 -7.066 8.44e-12 \*\*\*

obs341 -0.23801 0.08841 -2.692 0.00743 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08636 on 356 degrees of freedom

Multiple R-squared: 0.926, Adjusted R-squared: 0.9216

F-statistic: 212.1 on 21 and 356 DF, p-value: < 2.2e-16

**The model includes a fifth-degree polynomial for estimation of trend, monthly dummies to capture static seasonal structure, and five dummies for outlier values. *R*-square is 0.926, and the residual standard error is 0.086. The calendar pairs with frequencies 0.348 and 0.432 are not included—they are not significant.**

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

> b1<-coef(model1)[1]

> b2<-coef(model1)[7:17]+b1

> b3<-c(b1,b2)

> seas3<-exp(b3-mean(b3))

> seas3

(Intercept) fMonth2 fMonth3 fMonth4 fMonth5 fMonth6

0.8885614 0.9146527 1.0867495 1.0286565 1.0785358 1.0526850

fMonth7 fMonth8 fMonth9 fMonth10 fMonth11 fMonth12

1.0548546 1.0761965 0.9864983 0.9704703 0.9161433 0.9736374

**Jan. 0.8886 Jul. 1.0549**

**Feb. 0.9147 Aug. 1.0762**

**Mar. 1.0867 Sep. 0.9865**

**Apr. 1.0287 Oct. 0.9705**

**May 1.0785 Nov. 0.9161**

**Jun. 1.0527 Dec. 0.9736**

> plot(ts(seas3),ylab="Seasonal Index",xlab="Month",main="Estimated Seasonal Indices, Model 1")

A graph with numbers and lines

Description automatically generated

**The seasonal indices show that new car sales peak in March, May, and August, when they are estimated to be 8.67 percent, 7.85 percent, and 7.62 percent, respectively, above trend level. Sales are also estimated to be relatively high in June and July, at 5.27 percent and 5.49 percent above trend. The months of low sales are estimated to be January, February, and November, at 11.14 percent, 8.53 percent, and 8.39 percent, respectively, below the level of the trend. Cold winter months tend to deter consumers from new car shopping, and perhaps March sales are high because inaction during January and February leads to greater demand as the weather warms. Many new model years cars become available for sale during the late summer, and this leads to high sales volume in August. The plot shows a small rise in sales in December, probably the result of holiday shopping.**

(c) 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. The model fails to capture trend structure fully. Where does it fail and what is the cause? What conclusions do you draw from the residual analysis? In particular, what structure in the time series has the model failed to capture?

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(model1))

Shapiro-Wilk normality test

data: resid(model1)

W = 0.93782, p-value = 1.795e-11

**The quantile plot is decidedly nonnormal. The lower tail of the distribution is long relative to normality, and the upper tail is too short. The model overestimates sales volume during the 2008–09 Great Recession, and this is largely responsible for the long lower tail.**

> residcontraction<-c(rep(NA,111),resid(model1)[112:119],rep(NA,73),resid(model1)[193:210],rep(NA,128),resid(model1)[339:340],rep(NA,38))

> plot(ts(resid(model1),start=c(1992,1),freq=12),xlab="Time",ylab="Residual",main="Residuals of Model1",col="green",lwd=2)

> lines(ts(residcontraction,start=c(1992,1),freq=12),col="red",lwd=2)

**The residual plot which follows shows that the model estimates trend very poorly, especially during the 2008–09 recession. Estimation of trend is good during the 2001 and 2020 recessions. There are also a few evident mild outliers not accounted for by the model.**

**Volatility remains relatively constant throughout the duration of the time series.**

A graph showing a number of green and red lines

Description automatically generated

**The acf plot of the residuals has a pattern of slow and smooth decline of the estimated correlations as the lag number increases. This is indicative of remaining trend structure not captured by the model. Also notable are the very small local peaks at lag 12 and near lag 24, which point to some remaining dynamic seasonality.**

A graph with lines and numbers

Description automatically generated

4. (a) Calculate the lag 1, lag 2, and lag 3 residuals from the model in part 3 and add these three new variables to the model you fit in part 3. Be sure to include the calendar trigonometric pairs and outlier dummies if they turn out to be significant now.

The part 3 model fails to reduce to white noise. The lag residuals help to capture added structure (in the irregular part) which the part 3 model fails to account for. These new variables model the irregular part as a third-order autoregressive process. Later in the course we’ll study autoregressive processes and understand why these three lagged variable are employed.

**Code to create the lagged residual variables follows.**

> lresid<-c(rep(NA,378))

> lag1resid<-lresid;lag2resid<-lresid;lag3resid<-lresid

> lag1resid[2]<-resid(model1)[1];lag1resid[3]<-resid(model1)[2]

> lag2resid[3]<-resid(model1)[1]

> for(i in 4:378){

+ i1<-i-1;i2<-i-2;i3<-i-3

+ lag1resid[i]<-resid(model1)[i1];lag2resid[i]<-resid(model1)[i2]

+ lag3resid[i]<-resid(model1)[i3]

+ }

> carsales<-data.frame(carsales,lag1resid,lag2resid,lag3resid)

> model2<-lm(logSales~poly(Time,5)+fMonth+c432+s432+obs118+obs213+obs339+obs340+obs341+lag1resid+lag2resid+lag3resid);summary(model2)

Call:

lm(formula = logSales ~ poly(Time, 5) + fMonth + c432 + s432 +

obs118 + obs213 + obs339 + obs340 + obs341 + lag1resid +

lag2resid + lag3resid)

Residuals:

Min 1Q Median 3Q Max

-0.168352 -0.029151 0.005665 0.028297 0.124648

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.797782 0.008386 1287.595 < 2e-16 \*\*\*

poly(Time, 5)1 5.246608 0.047774 109.822 < 2e-16 \*\*\*

poly(Time, 5)2 0.185592 0.048159 3.854 0.000139 \*\*\*

poly(Time, 5)3 1.786460 0.048357 36.943 < 2e-16 \*\*\*

poly(Time, 5)4 -0.912891 0.048733 -18.732 < 2e-16 \*\*\*

poly(Time, 5)5 -0.603506 0.049310 -12.239 < 2e-16 \*\*\*

fMonth2 0.029002 0.011856 2.446 0.014933 \*

fMonth3 0.201642 0.011953 16.869 < 2e-16 \*\*\*

fMonth4 0.145720 0.011862 12.284 < 2e-16 \*\*\*

fMonth5 0.193368 0.011859 16.305 < 2e-16 \*\*\*

fMonth6 0.169673 0.011763 14.424 < 2e-16 \*\*\*

fMonth7 0.174353 0.011861 14.700 < 2e-16 \*\*\*

fMonth8 0.193278 0.011857 16.301 < 2e-16 \*\*\*

fMonth9 0.101335 0.011958 8.474 6.78e-16 \*\*\*

fMonth10 0.087550 0.011958 7.321 1.72e-12 \*\*\*

fMonth11 0.030690 0.011853 2.589 0.010023 \*

fMonth12 0.092206 0.011856 7.777 8.51e-14 \*\*\*

c432 0.008334 0.003523 2.366 0.018528 \*

s432 0.007026 0.003433 2.047 0.041421 \*

obs118 0.298405 0.047969 6.221 1.42e-09 \*\*\*

obs213 -0.186188 0.048409 -3.846 0.000143 \*\*\*

obs339 -0.423151 0.048064 -8.804 < 2e-16 \*\*\*

obs340 -0.598183 0.048218 -12.406 < 2e-16 \*\*\*

obs341 -0.209782 0.048218 -4.351 1.79e-05 \*\*\*

lag1resid 0.636024 0.050974 12.477 < 2e-16 \*\*\*

lag2resid -0.028696 0.061144 -0.469 0.639134

lag3resid 0.300224 0.051186 5.865 1.04e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04666 on 348 degrees of freedom

(3 observations deleted due to missingness)

Multiple R-squared: 0.9778, Adjusted R-squared: 0.9761

F-statistic: 589.4 on 26 and 348 DF, p-value: < 2.2e-16

**Model 2 includes a fifth-degree polynomial to estimate trend, monthly dummies to track static seasonality, the calendar pair with frequency 0.432, dummies for five outliers, and the lag 1, lag 2, and lag 3 residuals from model 1 in part 3. Thus, model 2 includes the three lagged residual variables from model 1 and the frequency 0.432 calendar pair, variables which are not a part of model 1. Addition of these variables has increased *R*-square from 0.926 to 0.978 and decreased the residual standard error from 0.086 to 0.047. Specifically, model 2 accounts for a greater fraction of the variation of the data and produces a fit which is on balance closer to the data.**

**Although the lag 2 residual variable is not significant, it sits between the lag 1 and lag 3 residuals and I will retain it.**

(b) Perform a residual analysis for this new model. What improvements do you notice? Discuss in detail. Is there any additional structure this new model has failed to capture? Explain carefully.

A graph of a normal q-q plot

Description automatically generated

> qq<-qqnorm(resid(model2))

> qqline(resid(model2))

> identify(qq)

[1] 163 199 200 211 339

**Although the Shapiro-Wilk test shown below rejects the null hypothesis of normality, the model 2 residual quantile plot is much closer to normality than is the model 1 plot. The problem is in the lower tail—several data pointsare overestimated by model 2. Two of them are July and August 2008 during the Great Recession; one is July 2009, the month after the Great Recession as recover begins; and another is March 2020, the onset of COVID.**

> shapiro.test(resid(model2))

Shapiro-Wilk normality test

data: resid(model2)

W = 0.98538, p-value = 0.0007713

> residcontraction2<-c(rep(NA,111),resid(model2)[112:119],rep(NA,73),resid(model2)[193:210],rep(NA,128),resid(model2)[339:340],rep(NA,38))

> plot(ts(resid(model2),start=c(1992,1),freq=12),xlab="Time",ylab="Residual",main="Residuals of Model 2",col="green",lwd=2)

> lines(ts(residcontraction2,start=c(1992,1),freq=12),col="red",lwd=2)

A graph showing a number of green and red lines

Description automatically generated

**The model 2 residual plot is a big improvement over the model 1 plot. The trend is poorly estimated during the Great Recession and at the onset of the COVID pandemic. Otherwise the plot is quite good, with only minimal changes in volatility.**

A graph with lines and numbers

Description automatically generated

**Although the residual acf plot indicates failure to reduce to white noise, we see considerable improvement relative to the model 1 plot. The significant lag 12 residual correlation shows that some dynamic seasonality has not been accounted for by model 2. The other residual correlations beyond the dashed lines show very modest significance.**

(c) Form the seasonal estimates from your model and tabulate them. How do they compare to the estimates in part 3?

> b1<-coef(model2)[1]

> b2<-coef(model2)[7:17]+b1

> b3<-c(b1,b2)

> seas4<-exp(b3-mean(b3))

> seas4

(Intercept) fMonth2 fMonth3 fMonth4 fMonth5 fMonth6

0.8884875 0.9146327 1.0869849 1.0278660 1.0780277 1.0527847

fMonth7 fMonth8 fMonth9 fMonth10 fMonth11 fMonth12

1.0577228 1.0779306 0.9832420 0.9697815 0.9161780 0.9743066

> options(digits=5)

> cbind(seas3,seas4)

seas3 seas4

January 0.88856 0.88849

February 0.91465 0.91463

March 1.08675 1.08698

April 1.02866 1.02787

May 1.07854 1.07803

June 1.05268 1.05278

July 1.05485 1.05772

August 1.07620 1.07793

September 0.98650 0.98324

October 0.97047 0.96978

November 0.91614 0.91618

December 0.97364 0.97431

**The table compares static seasonal estimates from model 1 in part 3 and the estimates from model 2 in part 4. The two sets of estimates are virtually identical. Model 1 fails to reduce to white noise, largely because of inadequate trend estimation. Model 2 has much better trend estimation and comes close to producing residuals consistent with white. We conclude that the failure of model 1 to address trend adequately does not hamper its ability to produce accurate estimation of static seasonal structure.**

5. (a) As an alternative to the models in parts 3 and 4, fit a ~~multiplicative~~ decomposition model to the log return of sales, that is, to *dlogSales.* As a start, include a seasonal component using the *fMonth* variable, trigonometric variables to investigate calendar structure, and outlier dummies. Retain only significant trigonometric pairs, and only significant outlier dummies.

**Part 1 showed that the log return of sales is essentially devoid of trend. We start by fitting a model without trend to the log return data. It also contain the calendar pair with frequency 0.432 and the five outlier dummies. The calendar pair with frequency 0.348 is not significant and is omitted.**

> model3<-lm(dlogSales~fMonth+c432+s432+obs118+obs213+obs339+obs340+obs341);summary(model3)

Call:

lm(formula = dlogSales ~ fMonth + c432 + s432 + obs118 + obs213 +

obs339 + obs340 + obs341)

Residuals:

Min 1Q Median 3Q Max

-0.15852 -0.03262 0.00189 0.03277 0.17181

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.088828 0.009011 -9.858 < 2e-16 \*\*\*

fMonth2 0.120845 0.012643 9.558 < 2e-16 \*\*\*

fMonth3 0.261283 0.012741 20.507 < 2e-16 \*\*\*

fMonth4 0.036849 0.012744 2.891 0.00407 \*\*

fMonth5 0.139559 0.012742 10.953 < 2e-16 \*\*\*

fMonth6 0.074979 0.012642 5.931 7.10e-09 \*\*\*

fMonth7 0.097003 0.012746 7.611 2.44e-13 \*\*\*

fMonth8 0.112548 0.012742 8.833 < 2e-16 \*\*\*

fMonth9 0.001371 0.012850 0.107 0.91510

fMonth10 0.084603 0.012853 6.582 1.65e-10 \*\*\*

fMonth11 0.026047 0.012742 2.044 0.04167 \*

fMonth12 0.153089 0.012746 12.011 < 2e-16 \*\*\*

c432 0.008344 0.003697 2.257 0.02462 \*

s432 0.006962 0.003666 1.899 0.05838 .

obs118 0.301180 0.051145 5.889 8.97e-09 \*\*\*

obs213 -0.218350 0.051125 -4.271 2.50e-05 \*\*\*

obs339 -0.393755 0.051106 -7.705 1.30e-13 \*\*\*

obs340 -0.123531 0.051105 -2.417 0.01614 \*

obs341 0.384053 0.051104 7.515 4.60e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05016 on 358 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.7215, Adjusted R-squared: 0.7075

F-statistic: 51.52 on 18 and 358 DF, p-value: < 2.2e-16

**The model variables are monthly dummies to estimate static seasonality, the frequency 0.432 calendar pair, and five outlier dummies. *R*-square is 0.7215, and the residual standard error is 0.0502.**

(b) Perform a residual analysis for this model. Discuss the results in detail. Is there any additional structure this new model has failed to capture? Explain carefully.

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(model3))

Shapiro-Wilk normality test

data: resid(model3)

W = 0.99449, p-value = 0.194

**The residual distribution is consistent with normality, as the quantile plot and the Shapiro–Wilk test indicate.**

> residcontraction3<-c(rep(NA,111),resid(model3)[112:119],rep(NA,73),resid(model3)[193:210],rep(NA,128),resid(model3)[339:340],rep(NA,38))

> plot(ts(resid(model3),start=c(1992,1),freq=12),xlab="Time",ylab="Residual",main="Residuals of Model 3",col="green",lwd=2)

> lines(ts(residcontraction3,start=c(1992,1),freq=12),col="red",lwd=2)

A graph showing a graph of a model

Description automatically generated with medium confidence

**The model 3 residual plot is quite good, indicating little remaining trend (except for a small dip during the 2008–09 Great Recession) and relatively stable volatility. There are a few mild outlier data values. The plot is better than that provided by model 2, because the differencing operation to produce the log return data has eliminated most, but not all, of the trend structure in the log sales series.**

A graph with lines and numbers

Description automatically generated

**The residual correlation plot shows failure to reduce to white noise. There is evidence of remaining dynamic seasonality, and failure to capture additional structure. Part (e) below will attempt to remedy this. In the meantime, let’s construct static seasonal estimates from model 3.**

(c) Construct seasonal estimates from this model for the sales data (*not for the log return of sales*). [See the discussion on pages 35–36 and the required code on page 37 of the 7 September notes. Modify the R code at the top of page 37, and follow it with the R code several lines later, also modified.]

> b1<-coef(model3)[1]

> b2<-coef(model3)[2:12]+b1

> b3<-c(b1,b2)

> x<-b3-mean(b3)

> s12<-0

> for(j in 2:12){

+ xsub<-x[j:12]

+ s12<-s12+sum(xsub)

+ }

> s12<-s12/12

> s<-c(rep(0,times=12))

> s[12]<-s12

> for(j in 1:11){

+ xsub<-x[1:j]

+ s[j]<-s[12]+sum(xsub)

+ }

> seas5<-exp(s)

> seas5

[1] 0.8877071 0.9133678 1.0814673 1.0230826 1.0725418 1.0540734 1.0589916

[8] 1.0806008 0.9866302 0.9790183 0.9162131 0.9735895

**Here are the static seasonal estimates listed above:**

**Jan. 0.8877 Jul. 1.0590**

**Feb. 0.9134 Aug. 1.0806**

**Mar. 1.0815 Sep. 0.9866**

**Apr. 1.0231 Oct. 0.9790**

**May 1.0725 Nov. 0.9162**

**Jun. 1.0541 Dec. 0.9736**

(d) There are now three sets of seasonal estimates, from parts 3 and 4 and the present part 5. Compare them with a table and a plot (place all three sets of estimates on one plot). Discuss.

> cbind(seas3,seas4,seas5)

seas3 seas4 seas5

January 0.8885614 0.8884875 0.8877071

February 0.9146527 0.9146327 0.9133678

March 1.0867495 1.0869849 1.0814673

April 1.0286565 1.0278660 1.0230826

May 1.0785358 1.0780277 1.0725418

June 1.0526850 1.0527847 1.0540734

July 1.0548546 1.0577228 1.0589916

August 1.0761965 1.0779306 1.0806008

September 0.9864983 0.9832420 0.9866302

October 0.9704703 0.9697815 0.9790183

November 0.9161433 0.9161780 0.9162131

December 0.9736374 0.9743066 0.9735895

> plot(ts(seas3),ylab="Seasonal Index",xlab="Month",main="Estimated Seasonal Indices, Models 1, 2, and 3",col="red",lwd=2)

> lines(ts(seas4),col="blue",lwd=2)

> lines(ts(seas5),col="goldenrod",lwd=2)

A graph with blue and orange lines

Description automatically generated

**The three sets are estimates are very similar. There are some minor differences, noticeable in May and October and stemming from model 3. The results show that failure to estimate trend structure does not hamper estimation of static seasonal structure.**

**Let’s next attempt to improve model 3. We do so by adding the lag 1 and lag 2 residuals from model 3.**

(e) Construct the lag 1 and lag 2 residuals from the model in (a) and refit the model in (a) with these two variables added. Perform a residual analysis and comment on the results. [The required R code to obtain the lag 1 and lag 2 residuals for the model fit is tricky. Model 3 has 377 residuals, not 378. Form the lagged residuals with length 378, with NA values at the start of each. Use the following code:

lresid2<-c(rep(NA,378))

lag1resid2<-lresid2;lag2resid2<-lresid2

lag1resid2[3]<-resid(model3)[1]

for(i in 4:378){

i1<-i-2;i2<-i-3

lag1resid2[i]<-resid(model3)[i1];lag2resid2[i]<-resid(model3)[i2]

}

carsales<-data.frame(carsales,lag1resid2,lag2resid2)

> lresid2<-c(rep(NA,378))

> lag1resid2<-lresid2;lag2resid2<-lresid2

> lag1resid2[3]<-resid(model3)[1]

> for(i in 4:378){

+ i1<-i-2;i2<-i-3

+ lag1resid2[i]<-resid(model3)[i1];lag2resid2[i]<-resid(model3)[i2]

+ }

> carsales<-data.frame(carsales,lag1resid2,lag2resid2)

> model4<-lm(dlogSales~fMonth+c432+s432+obs118+obs213+obs339+obs340+obs341+lag1resid2+lag2resid2);summary(model4)

Call:

lm(formula = dlogSales ~ fMonth + c432 + s432 + obs118 + obs213 +

obs339 + obs340 + obs341 + lag1resid2 + lag2resid2)

Residuals:

Min 1Q Median 3Q Max

-0.179681 -0.028667 0.003488 0.028779 0.137102

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.088804 0.008547 -10.391 < 2e-16 \*\*\*

fMonth2 0.120544 0.012089 9.972 < 2e-16 \*\*\*

fMonth3 0.263846 0.012185 21.653 < 2e-16 \*\*\*

fMonth4 0.036914 0.012087 3.054 0.002429 \*\*

fMonth5 0.139545 0.012085 11.547 < 2e-16 \*\*\*

fMonth6 0.074964 0.011990 6.252 1.17e-09 \*\*\*

fMonth7 0.096893 0.012089 8.015 1.61e-14 \*\*\*

fMonth8 0.112581 0.012085 9.316 < 2e-16 \*\*\*

fMonth9 0.000192 0.012189 0.016 0.987440

fMonth10 0.084950 0.012192 6.968 1.58e-11 \*\*\*

fMonth11 0.026036 0.012085 2.154 0.031880 \*

fMonth12 0.153045 0.012089 12.660 < 2e-16 \*\*\*

c432 0.008029 0.003512 2.286 0.022825 \*

s432 0.007354 0.003492 2.106 0.035926 \*

obs118 0.289560 0.048703 5.945 6.61e-09 \*\*\*

obs213 -0.182725 0.048803 -3.744 0.000211 \*\*\*

obs339 -0.411038 0.048568 -8.463 6.99e-16 \*\*\*

obs340 -0.126300 0.048475 -2.605 0.009563 \*\*

obs341 0.383561 0.048471 7.913 3.25e-14 \*\*\*

lag1resid2 -0.263554 0.051658 -5.102 5.49e-07 \*\*\*

lag2resid2 -0.256745 0.051474 -4.988 9.59e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04758 on 354 degrees of freedom

(3 observations deleted due to missingness)

Multiple R-squared: 0.7513, Adjusted R-squared: 0.7372

F-statistic: 53.46 on 20 and 354 DF, p-value: < 2.2e-16

**The model 4 explanatory variables are all those in model 3, plus the lag 1 and lag 2 residuals from model 3. *R*-square is 0.7513, in comparison to 0.7215 for model 3, and the residual standard error is 0.0476, less than the value 0.0502 for model 3. The two lag residual variables are significant, and model 4 has improved the performance of model 3.**

**Next we examine the residual diagnostics.**

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(model4))

Shapiro-Wilk normality test

data: resid(model4)

W = 0.98681, p-value = 0.001736

**Normality is rejected for the model 4 residual distribution. The cause is overestimation of sales during the 2001 and 2008–09 recessions. The residual plot below illustrates this and shows a small amount of remaining trend structure, as with the model 3 residuals. The model 3 and model 4 residual time series are similar.**

A graph showing a graph of a graph

Description automatically generated with medium confidence

A graph with lines and text

Description automatically generated

**As the residual acf plot shows, model 4 does not provide reduction to white noise. However, this plot is better than the corresponding plot for model 3. Here the significant correlations at lags 12, 18, 24, and 30 indicate there is remaining dynamic seasonality. There are significant correlations at other lags, but they are only modestly significant. We conclude that addition of the two lagged residual variables has resulted in improved performance relative to model 3.**

6. This part does not require computation, only discussion. Analysis of the sales data has shown that dynamic seasonality is present. Discuss how one can perform analysis to illustrate and estimate how the seasonal structure has evolved over time.

**Divide the time span from 1992 to 2023(1) into intervals of approximately equal length. They can be overlapping or nonoverlapping. Build a multiplicative decomposition model for each interval and estimate the static seasonal structure. Then tabulate and plot the various seasonal estimates and discuss how they change as time progresses.**

**Later in the course we will fit a seasonal ARIMAX model to estimate and depict the dynamic seasonality.**