Statistics 5350/7110 Fall 2023

P. Shaman

Assignment 4 solution

1. The time series for this problem gives monthly U.S. advance retail sales for grocery stores for 1992(1) to 2023(10), in millions of dollars. The file is GrocerySales.txt.

The U.S. Bureau of the Census publishes its first estimate of retail sales nine working days after the end of each month. This estimate is reported by the press and is closely watched by the financial markets and economists.

The aim of this problem is to provide (i) static seasonal index estimates by two different methods, one using regression and a second employing seasonal ARIMAX modeling; and (ii) dynamic seasonal index estimates from the seasonal ARIMAX modeling.

> grocsales<-read.csv("F:/Stat711023Fall/GrocerySales.txt")

> attach(grocsales)

> head(grocsales)

Date Time Month Sales logSales c220 s220 c348

1 1992-01-01 1 1 27306 10.21486 0.1873813 0.9822873 -0.57757270

2 1992-02-01 2 2 26223 10.17439 -0.9297765 0.3681246 -0.33281954

3 1992-03-01 3 3 27235 10.21226 -0.5358268 -0.8443279 0.96202767

4 1992-04-01 4 4 27588 10.22514 0.7289686 -0.6845471 -0.77846230

5 1992-05-01 5 5 28883 10.27101 0.8090170 0.5877853 -0.06279052

6 1992-06-01 6 6 28039 10.24135 -0.4257793 0.9048271 0.85099448

s348 c432 s432

1 0.8163393 -0.9101060 0.4143756

2 -0.9429905 0.6565858 -0.7542514

3 0.2729519 -0.2850193 0.9585218

4 0.6276914 -0.1377903 -0.9904614

5 -0.9980267 0.5358268 0.8443279

6 0.5251746 -0.8375280 -0.5463943

(a) List the periods of economic downturn during 1992–2023, as defined by the Business Cycle Dating Committee of the National Bureau of Economic Research. Plot the sales series with the periods of economic downturn marked. Comment in detail on features of the time series revealed by the plot. Repeat for the log sales series.

**According to the Business Cycle Dating Committee, economic downturns occurred during April 2001 to November 2001, January 2008 to June 2009, and March 2020 to April 2020.**

> Salescontraction<-c(rep(NA,111),Sales[112:119],rep(NA,73),Sales[193:210],rep(NA,128),Sales[339:340],rep(NA,42))

> plot(ts(Sales,start=c(1992,1),freq=12),ylab="Sales",main="Monthly U.S. Grocery Sales, in Millions",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 green line

Description automatically generated

**The plot of monthly sales shows a steady increase with increasing derivative, much of this driven by increasing population. There is a noticeable seasonality present, and volatility increases as the level of sales rises. The year 2001 recession had no effect upon grocery sales, but the 2008–09 recession caused some disruption of the seasonal pattern, with somewhat increased sales during 2008. Given the economic hardship caused by the recession, many consumers had to forego some of their restaurant meals, and this led to greater grocery purchases to provide meals at home. The impact of the 2020 recession at the onset of the COVID pandemic was a dramatic increase in grocery purchases during spring 2020, as people were confined to home, sought comfort, and needed to cook their meals. This sharp rise was short-lived, as monthly grocery sales returned to the upward trend with the increasing slope established prior to 2020. One should note that the most grocery purchases are a necessity, rather than a luxury.**

A graph of a graph showing a green line

Description automatically generated with medium confidence

**The log transformation has done a decent job of stabilizing the variance. Comments about this log plot are otherwise essentially the same as those for sales.**

*For the remainder of this assignment restrict calculations to data for the years 1992 through 2019.*

(b) Construct a spectral plot for the log sales series. Mark the frequencies which determine seasonal structure and calendar features. Discuss carefully what the plots reveal, what guidance they provide for model construction. [To resolve spectral peaks at close frequencies (such as 0.333 and 0.348), choose a sufficiently small span.]

> spectrum(logSales[1:336],4)

> abline(v=c(1/12,2/12,3/12,4/12,5/12,6/12),lty=2,col=”red”)

> abline(v=c(0.22,0.348,0.432),lty=2,col=”blue”)

A graph of a graph

Description automatically generated

**The spectral plot, which uses a relatively small span, provides very useful information about the sales time series. The spectral peak at low frequency indicates there is a prominent trend in sales, and peaks at the seasonal frequencies marked by the red lines are a consequence of notable seasonal structure. In addition, calendar structure is present in the series, especially at frequency 0.348, and we may need to also include the calendar pairs at frequencies 0.220 and 0.432 in our regression model, as the plot shows.**

(c) Fit a regression model to the log sales data and use it to construct estimated static seasonal indices. Tabulate, plot, and interpret the estimated indices. Also, give a thorough residual analysis of the model fit, including a normal quantile plot of the residuals with the Shapiro-Wilk test, a plot of the residuals vs. time, residual acf and pacf plots, a residual spectral plot, and Bartlett’s Kolmogorov–Smirnov test. Discuss your findings.

> fMonth<-as.factor(Month)

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

> modelreg<-lm(logSales~poly(Time,3)+fMonth+c348+s348+c432+s432,data=grocsales[1:336,]);summary(modelreg)

Call:

lm(formula = logSales ~ poly(Time, 3) + fMonth + c348 + s348 +

c432 + s432, data = grocsales[1:336, ])

Residuals:

Min 1Q Median 3Q Max

-0.036787 -0.009130 -0.000511 0.008853 0.046729

Coefficients:

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

(Intercept) 10.5650450 0.0026854 3934.286 < 2e-16 \*\*\*

poly(Time, 3)1 4.0395237 0.0142118 284.238 < 2e-16 \*\*\*

poly(Time, 3)2 0.2126847 0.0142022 14.976 < 2e-16 \*\*\*

poly(Time, 3)3 -0.0631851 0.0142248 -4.442 1.23e-05 \*\*\*

fMonth2 -0.0707948 0.0037963 -18.648 < 2e-16 \*\*\*

fMonth3 0.0156135 0.0037960 4.113 4.98e-05 \*\*\*

fMonth4 -0.0095836 0.0037963 -2.524 0.012075 \*

fMonth5 0.0448994 0.0037965 11.826 < 2e-16 \*\*\*

fMonth6 0.0131941 0.0037966 3.475 0.000581 \*\*\*

fMonth7 0.0422880 0.0037972 11.137 < 2e-16 \*\*\*

fMonth8 0.0293493 0.0037972 7.729 1.44e-13 \*\*\*

fMonth9 -0.0115298 0.0037981 -3.036 0.002599 \*\*

fMonth10 0.0091522 0.0037980 2.410 0.016533 \*

fMonth11 0.0099149 0.0037990 2.610 0.009488 \*\*

fMonth12 0.0732441 0.0037994 19.278 < 2e-16 \*\*\*

c348 -0.0096148 0.0010958 -8.774 < 2e-16 \*\*\*

s348 -0.0001238 0.0010961 -0.113 0.910166

c432 -0.0011122 0.0010990 -1.012 0.312314

s432 0.0037775 0.0010936 3.454 0.000627 \*\*\*

---

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

Residual standard error: 0.0142 on 317 degrees of freedom

Multiple R-squared: 0.9962, Adjusted R-squared: 0.996

F-statistic: 4646 on 18 and 317 DF, p-value: < 2.2e-16

**The regression model uses a third-degree polynomial in time to estimate trend, monthly dummies to capture static seasonality, and trigonometric pairs at frequencies 0.348 and 0.432 to account for calendar structure. The calendar pair with frequency 0.220 is not significant. *R-*square for the model is very high, 0.9962, and the residual standard error is 0.0142.**

> b1<-coef(modelreg)[1]

> b2<-coef(modelreg)[5:15]+b1

> b3<-c(b1,b2)

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

> seas

(Intercept) fMonth2 fMonth3 fMonth4 fMonth5 fMonth6

0.9879278 0.9204060 1.0034739 0.9785052 1.0332961 1.0010490

fMonth7 fMonth8 fMonth9 fMonth10 fMonth11 fMonth12

1.0306013 1.0173526 0.9766027 0.9970111 0.9977718 1.0630036

A graph with numbers and lines

Description automatically generated

**December is the month for high sales, estimated at 6.30 percent above trend level, and February has the lowest sales, estimated to be 7.6 percent below the trend. There are small local peaks in May and July, each about 3 percent above trend level. All the other months are estimated to be within about 2 percent of the trend. The high values for May, July, and December are undoubtedly associated with holiday periods when families and friends get together to enjoy cooked meals. Perhaps it is curious that November does not show elevated sales for the Thanksgiving holiday.**

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(modelreg))

Shapiro-Wilk normality test

data: resid(modelreg)

W = 0.99453, p-value = 0.2754

**According to the Shapiro-Wilk test, the regression residuals conform to normality.**

A graph showing the results of a stock market

Description automatically generated

**The regression did not fully account for the trend of log sales. Poor trend estimation is mostly during the years 2002 to 2011. Otherwise volatility of the residuals is relatively constant, and the normal quantile plot of the residuals shows there are only two modest outlier residuals, At the outset of the 2008–09 recession the model underestimates log sales, and in 2009 it overestimates log sales**

A graph of a graph

Description automatically generated with medium confidence

**The residual acf and pacf plots show the regression model has failed to produce residuals consistent with white noise, and so do the following residual spectral plot and Bartlett’s test.**

> spectrum(resid(modelreg),5)

> abline(v=c(1/12,2/12,3/12,4/12,5/12,6/12),lty=2,col="red")

> abline(v=0.22,lty=2,col="blue")

A graph of a graph

Description automatically generated

**The spectral plot has a peak at low frequency, indicating failure to fully estimate trend structure. The peaks at frequencies 1/12, 2/12, 3/12, 4/12, and 5/12 provide evidence of remaining dynamic seasonality, and the peak at the calendar frequency 0.220 shows this calendar pair does require attention.**

> bartlettB.test(resid(modelreg))

Bartlett B Test for white noise

data:

= 4.8271, p-value < 2.2e-16

(d) Fit a seasonal ARIMAX model to the log sales data and use it to construct estimated static and dynamic seasonal indices. To do so, follow the steps used in the 16 November notes to analyze the U.S. beer production data:

(i) Fit a seasonal ARIMAX model to the logged data. Analyze the residuals from this ARIMAX fit. Exclude the 1992 residuals in performing the residual analysis. Your ARIMAX model should include both ordinary and seasonal differencing, and calendar variables.

**Look at the acf and pacf of the doubly differenced data. These plots do not adjust for calendar effects.**

> d1d12lsales.ts<-ts(diff(diff(logSales[1:336]),12))

> par(mfrow=c(1,2))

> acf(d1d12lsales.ts,37)

> pacf(d1d12lsales.ts,37)

A graph of a graph

Description automatically generated with medium confidence

**The plots are very busy—perhaps because there is no adjustment for the calendar variables. Let’s try an ARX(2,1,0)(3,1,0)12 with use of xreg to include the three calendar pairs. This model will address the partial correlations at lags 1 and 2 and the significant partial correlations in the neighborhoods of lags 12, 24, and 36. Ordinary differencing will account for the trending of log sales, and seasonal differencing will pick up static seasonality,**

> df<-data.frame(c220,s220,c348,s348,c432,s432)

> arxmodel1<-arima(logSales[1:336],order=c(2,1,0),seasonal=list(order=c(3,1,0),period=12),xreg=df[1:336,])

> arxmodel1

Call:

arima(x = logSales[1:336], order = c(2, 1, 0), seasonal = list(order = c(3,

1, 0), period = 12), xreg = df[1:336, ])

Coefficients:

ar1 ar2 sar1 sar2 sar3 c220 s220 c348

-0.8520 -0.3361 -0.6084 -0.4114 -0.2626 2e-03 -0.0013 -0.0095

s.e. 0.0528 0.0531 0.0573 0.0628 0.0578 4e-04 0.0004 0.0008

s348 c432 s432

-3e-04 -0.0012 0.0036

s.e. 8e-04 0.0009 0.0009

sigma^2 estimated as 0.0001346: log likelihood = 977.45, aic = -1930.91

> library("lmtest")

> coeftest(arxmodel1)

z test of coefficients:

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

ar1 -0.85201765 0.05281618 -16.1318 < 2.2e-16 \*\*\*

ar2 -0.33607061 0.05309856 -6.3292 2.465e-10 \*\*\*

sar1 -0.60840841 0.05727261 -10.6230 < 2.2e-16 \*\*\*

sar2 -0.41135648 0.06284888 -6.5452 5.943e-11 \*\*\*

sar3 -0.26255321 0.05775241 -4.5462 5.463e-06 \*\*\*

c220 0.00202513 0.00039832 5.0841 3.693e-07 \*\*\*

s220 -0.00129204 0.00039797 -3.2466 0.001168 \*\*

c348 -0.00949312 0.00079521 -11.9379 < 2.2e-16 \*\*\*

s348 -0.00026523 0.00079611 -0.3332 0.739015

c432 -0.00120620 0.00088728 -1.3594 0.174009

s432 0.00362590 0.00088329 4.1050 4.043e-05 \*\*\*

---

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

**All three calendar pairs are significant. Residual diagnostics follow.**

A graph with green and red lines

Description automatically generated

**Except for some outliers, primarily involving overestimation of log sales, and a small dip during the 2008–09 recession, the model has done a good job of capturing trend structure.**

**But the quantile plot of the residuals does show a somewhat long lower tail, and the residuals do not conform to normality. For calculations we omit the residuals for the year 1992.**

> sel<-1:12

> qqnorm(resid(arxmodel1)[-sel])

> qqline(resid(arxmodel1)[-sel])

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(arxmodel1)[-sel])

Shapiro-Wilk normality test

data: resid(arxmodel1)[-sel]

W = 0.98744, p-value = 0.006576

> par(mfrow=c(1,2))

> acf(resid(arxmodel1)[-sel],37)

> pacf(resid(arxmodel1)[-sel],37)

A graph of a model

Description automatically generated with medium confidence

**The residual acf plot shows reduction to white noise, but there are significant partial correlations at lags 14, 24, 32, and 36. However, these estimated partial correlations are barely significant and are not cause for concern.**

A graph of a graph

Description automatically generated

> bartlettB.test(resid(arxmodel1)[-sel])

Bartlett B Test for white noise

data:

= 0.52856, p-value = 0.9427

**The residual spectal plot is sufficiently flat to indicate reduction to white noise by the ARIMAX model. The big drop at low frequency is an artifact of the use of the log transformation on the vertical scale of the plot. Bartlett’s test clearly confirms reduction to white noise.**

(ii) With data and residuals for 1992 excluded, construct the ARIMAX predicted values. Thus, the predicted values will span the years 1993 to 2019, 324 data points. Plot these predicted values and discuss the appearance of the plot.

> sel2<-13:336

> arimapred<-logSales[sel2]-resid(arxmodel1)[-sel]

> arimapred.ts<-ts(arimapred,start=c(1993,1),freq=12)

> plot(arimapred.ts,xlab="Time",ylab="ARIMAX Predictions",main="ARIMAX Predictions from (2,1,0)(3,1,0)12 Model")

A graph showing the growth of a number of individuals

Description automatically generated with medium confidence

**The predicted values show a rather steady upward trend and a strong seasonal structure. The slope does decrease slightly during the 2008–09 Great Recession. Volatility does become less severe as time progresses. Next, we will use a regression model to estimate and remove trend and calendar structures from the predicted values. This will leave a series with seasonality and additive noise.**

(iii) Fit a regression model to estimate trend and calendar structure in the ARIMAX predicted values. Then form the residuals from this regression. [One has to be careful to remove both trend and calendar features from the ARIMAX predicted values. In the regression use a polynomial in time of degree five, and include all three calendar pairs. Some of these variables will not be significant in the regression, but, nonetheless, include them. All of the explanatory variables must have length 324; for example, use Time2<-as.numeric(1:324)and

c220a<-c220[1:324].]

> Time2<-as.numeric(1:324)

> c220a<-c220[1:324];s220a<-s220[1:324]

> c348a<-c348[1:324];s348a<-s348[1:324]

> c432a<-c432[1:324];s432a<-s432[1:324]

> modelpred<-lm(arimapred~poly(Time2,5)+c220a+s220a+c348a+s348a+c432a+s432a);summary(modelpred)

Call:

lm(formula = arimapred ~ poly(Time2, 5) + c220a + s220a + c348a +

s348a + c432a + s432a)

Residuals:

Min 1Q Median 3Q Max

-0.112411 -0.018101 0.000799 0.022501 0.099914

Coefficients:

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

(Intercept) 10.589314 0.002096 5051.741 < 2e-16 \*\*\*

poly(Time2, 5)1 3.881783 0.037733 102.874 < 2e-16 \*\*\*

poly(Time2, 5)2 0.175555 0.037734 4.652 4.85e-06 \*\*\*

poly(Time2, 5)3 -0.031021 0.037737 -0.822 0.41169

poly(Time2, 5)4 0.001088 0.037737 0.029 0.97701

poly(Time2, 5)5 0.039954 0.037741 1.059 0.29057

c220a -0.000745 0.002970 -0.251 0.80209

s220a 0.002352 0.002961 0.794 0.42759

c348a -0.005251 0.002969 -1.768 0.07798 .

s348a 0.008743 0.002961 2.953 0.00338 \*\*

c432a 0.002795 0.002963 0.943 0.34630

s432a 0.002642 0.002966 0.891 0.37378

---

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

Residual standard error: 0.03773 on 312 degrees of freedom

Multiple R-squared: 0.9715, Adjusted R-squared: 0.9705

F-statistic: 965.4 on 11 and 312 DF, p-value: < 2.2e-16

**The regression model uses a fifth-degree polynomial to estimate trend and includes the three calendar trigonometric pairs. Many individual coefficients are not significant because the residuals include seasonal structure, and this has inflated the standard errors and deflated the *t*-statistics. The residuals from this regression will be devoid of most of the trend and of the calendar effects.**

(iv) Plot the residuals from (iii) and discuss the appearance of the plot. Has the regression in (iii) been successful in removing trend structure from the ARIMAX predicted values?

> arimapred2<-resid(modelpred)

> arimapred2.ts<-ts(arimapred2,start=c(1993,1),freq=12)

> plot(arimapred2.ts,xlab="Time",ylab="Seasonal Predictions",main="Seasonal Predictions from (2,1,0)(3,1,0)12 ARIMAX Model")

**T**A graph showing the results of a forecast

Description automatically generated with medium confidence

**The adjusted ARIMAX predicted values have very little remaining trend—there is a drop in 2009 during the Great Recession. In addition, volatility becomes relatively low beginning in 2009. The result isn’t perfect, but hopefully will suffice to provide estimation of static and dynamic seasonality.**

(v) Using the detrended ARIMAX predicted values from (iv), form ARIMAX static seasonal estimates. Compare these estimates to those obtained from the regression in part (c), using a table and a plot. Discuss the results.

**We begin with calculation of static seasonal estimates.**

> monmeans<-tapply(arimapred2,Month[1:324],mean)

> seas2<-monmeans-mean(monmeans)

> seas2<-exp(seas2)

> seas2

1 2 3 4 5 6 7 8

0.9871876 0.9209788 1.0025057 0.9787444 1.0336335 1.0024135 1.0317565 1.0170235

9 10 11 12

0.9758344 0.9961167 0.9965283 1.0643804

> cbind(1:12,seas,seas2)

seas seas2

January 1 0.9879278 0.9871876

February 2 0.9204060 0.9209788

March 3 1.0034739 1.0025057

April 4 0.9785052 0.9787444

May 5 1.0332961 1.0336335

June 6 1.0010490 1.0024135

July 7 1.0306013 1.0317565

August 8 1.0173526 1.0170235

September 9 0.9766027 0.9758344

October 10 0.9970111 0.9961167

November 11 0.9977718 0.9965283

December 12 1.0630036 1.0643804

> plot(ts(seas),xlab="Month",ylab="Estimated Seasonal Indices",main="Estimated Seasonals from Regression and ARIMAX",lty=1,lwd=2,col="red")

> lines(ts(seas2),lty=1,lwd=2,col="blue")

> legend("topleft",legend=c("regression","ARIMAX"),col=c("red","blue"),lty=1,cex=0.8)

A graph with a line and numbers

Description automatically generated

**The static seasonal estimates from the regression and ARIMAX models are essentially the same.**

(vi) Form the ARIMA dynamic seasonal estimates and plot and discuss the results. [Use the following R code. It is the same as the code on page 7 of the 16 November notes, except for a few modifications to adapt to the present data set.]

y<-arimapred2

seasm<-matrix(rep(0,324),ncol=27)

j<--11

for(i in 1:27){

j<-j+12;j1<-j+11

seasm[,i]<-exp(y[j:j1]-mean(y[j:j1]))

}

year<-seq(1993,2019)

seas2m<-matrix(rep(seas2,27),ncol=27)

name<-c("January","February","March","April","May","June","July","August","September","October","November","December")

par(mfrow=c(3,1))

for(i in 1:3){

plot(year,seasm[i,],xlab="Year",ylab="Indices",main=name[i],type="l",lwd=2,col="red")

lines(year,seas2m[i,],lty=1,lwd=2,col="blue")

}

A graph of different numbers

Description automatically generated with medium confidence

A graph of different times

Description automatically generated with medium confidence

A graph of different numbers

Description automatically generated with medium confidence

A graph of different numbers

Description automatically generated with medium confidence

**The twelve monthly plots show dynamic seasonality is present for grocery sales during the years from 1993 to 2019. June, July, August, and December show decrease of the estimated seasonal index over time, and January, February, and November show some increase. However, the changes over time of the indices are rather modest. The increases in the summer months and December suggest a rise over the years in purchases for recreational use and holiday celebrations.**

2. Consider monthly simple returns for Disney common stock for the years 1971 to 2014. The data are in Disney.txt. During the first three decades of this period, Disney stock rose substantially in value. There were six stock splits throughout these 44 years:

3/01/1971 2 for 1

1/16/1973 2 for 1

3/06/1986 4 for 1

5/18/1992 4 for 1

7/10/1998 3 for 1

6/13/2007 1014 for 1000

> disney<-read.csv("F:/Stat711023Fall/Disney.txt")

> attach(disney)

> head(disney)

time date month disneyprice disneyreturn

1 1 19710129 1 158.875 0.123784

2 2 19710226 2 177.750 0.118804

3 3 19710331 3 101.000 0.136990

4 4 19710430 4 113.000 0.118812

5 5 19710528 5 121.375 0.074115

6 6 19710630 6 114.000 -0.060350

> fmonth<-as.factor(month)

> disney<-data.frame(disney,fmonth)

(a) Plot the time series and discuss its appearance. Identify evident outliers.

> plot(ts(disneyreturn,start=c(1971,1),freq=12),ylab="log Return",main="Monthly Disney log Returns, 1971 to 2014",lwd=2)

A graph showing a sound wave

Description automatically generated

**The returns display no trending. There is some varying volatility, and there are a few outlier data points. Let’s define dummies for November 1973, October 1974, May 1975, and October 1987. The first three of these are the most prominent.**

> obs35<-ifelse(time==35,1,0);obs45<-ifelse(time==45,1,0)

> obs49<-ifelse(time==49,1,0);obs159<-ifelse(time==159,1,0)

> disney<-data.frame(disney,obs35,obs45,obs49,obs159)

(b) Use data for the years 1971 to 1992.

(i) The Disney log returns include some seasonal structure. Fit a regression model to the log returns. Include dummies for outliers as needed. [I defined four dummies for outliers; don’t use a lot of outlier dummies.] Tabulate and plot the estimated static seasonal indices, and interpret and discuss them.

> model1<-lm(disneyreturn~fmonth+obs35+obs45+obs49+obs159,data=disney[1:264,]);summary(model1)

Call:

lm(formula = disneyreturn ~ fmonth + obs35 + obs45 + obs49 +

obs159, data = disney[1:264, ])

Residuals:

Min 1Q Median 3Q Max

-0.29072 -0.05393 0.00000 0.05873 0.26424

Coefficients:

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

(Intercept) 0.053144 0.019476 2.729 0.006815 \*\*

fmonth2 -0.011597 0.027229 -0.426 0.670535

fmonth3 -0.026290 0.027544 -0.954 0.340762

fmonth4 -0.036463 0.027229 -1.339 0.181755

fmonth5 -0.019289 0.027229 -0.708 0.479358

fmonth6 -0.053388 0.027229 -1.961 0.051032 .

fmonth7 -0.065409 0.027229 -2.402 0.017034 \*

fmonth8 -0.050442 0.027229 -1.853 0.065140 .

fmonth9 -0.076159 0.027544 -2.765 0.006120 \*\*

fmonth10 -0.049519 0.027229 -1.819 0.070173 .

fmonth11 -0.030181 0.027544 -1.096 0.274259

fmonth12 -0.004849 0.027229 -0.178 0.858794

obs35 -0.389313 0.091352 -4.262 2.88e-05 \*\*\*

obs45 -0.342406 0.091352 -3.748 0.000222 \*\*\*

obs49 0.362061 0.091352 3.963 9.67e-05 \*\*\*

obs159 0.316646 0.091352 3.466 0.000622 \*\*\*

---

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

Residual standard error: 0.08925 on 248 degrees of freedom

Multiple R-squared: 0.2608, Adjusted R-squared: 0.2161

F-statistic: 5.833 on 15 and 248 DF, p-value: 2.624e-10

**The model contains monthly dummies to estimate static seasonality, and the dummies for the four identified outliers. *R*-square is 0.2608, and the residual standard error is 0.0892, which corresponds to movement of the return equal to about 0.9 percent.**

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(model1))

Shapiro-Wilk normality test

data: resid(model1)

W = 0.995, p-value = 0.5

**The residuals show good agreement with normality, as indicated by the plot and the Shapiro–Wilk test.**

> b1<-coef(model1)[1]

> b2<-coef(model1)[2:12]+b1

> b3<-c(b1,b2)

> seas<-b3-mean(b3)

> options(digits=3)

> seas

(Intercept) fmonth2 fmonth3 fmonth4 fmonth5 fmonth6

0.03530 0.02370 0.00901 -0.00116 0.01601 -0.01809

fmonth7 fmonth8 fmonth9 fmonth10 fmonth11 fmonth12

-0.03011 -0.01514 -0.04086 -0.01422 0.00512 0.03045

> cbind(1:12,seas)

seas

January 1 0.03530

February 2 0.02370

March 3 0.00901

April 4 -0.00116

May 5 0.01601

June 6 -0.01809

July 7 -0.03011

August 8 -0.01514

September 9 -0.04086

October 10 -0.01422

November 11 0.00512

December 12 0.03045

A graph with numbers and lines

Description automatically generated

**Disney returns for 1971 to 1992 peak in January and December, estimated at 3.5 percent and 3.0 percent, respectively, above trend level. These months are associated with year-end selling and buying for income tax purposes. There is also a local peak in May, estimated to be 1.6 percent above trend. The months are September and July, estimated at 4.1 percent and 3.0 percent, respectively, below trend, and June and August returns are also estimated to be low.**

(ii) Fit an ARIMAX model to the regression residuals. Describe your model.

A screenshot of a graph

Description automatically generated

**Let’s use the pacf plot to select an ARIMA model. There are significant partial correlations at lags 18, 24, and 36, and the normal quantile plot shows there are no notable outliers among the regression residuals. We use an ARIMA(0,0,0)(6,0,0)6 model.**

> armodel1<-arima(resid(model1),order=c(0,0,0),seasonal=list(order=c(6,0,0),period=6))

> armodel1

Call:

arima(x = resid(model1), order = c(0, 0, 0), seasonal = list(order = c(6, 0,

0), period = 6))

Coefficients:

sar1 sar2 sar3 sar4 sar5 sar6 intercept

-0.017 -0.010 -0.236 -0.199 0.081 -0.201 -0.001

s.e. 0.060 0.061 0.060 0.059 0.062 0.062 0.003

sigma^2 estimated as 0.00652: log likelihood = 288.09, aic = -560.19

> coeftest(armodel1)

z test of coefficients:

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

sar1 -0.01674 0.06023 -0.28 0.7811

sar2 -0.00970 0.06089 -0.16 0.8734

sar3 -0.23606 0.06002 -3.93 8.4e-05 \*\*\*

sar4 -0.19882 0.05931 -3.35 0.0008 \*\*\*

sar5 0.08050 0.06157 1.31 0.1910

sar6 -0.20050 0.06213 -3.23 0.0012 \*\*

intercept -0.00123 0.00326 -0.38 0.7064

---

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

**As expected, the estimates at lags 18, 24, and 36 are significant.**

(iii) Perform a residual analysis for the fitted ARIMAX model. Discuss the results.

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(armodel1))

Shapiro-Wilk normality test

data: resid(armodel1)

W = 0.992, p-value = 0.19

**The quantile plot shows slightly long activity in both tails, but the Shapiro–Wilk test indicates the ARIMA residuals are acceptably consistent with normality.**

A comparison of a graph

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated

> bartlettB.test(resid(armodel1))

Bartlett B Test for white noise

data:

= 0.823, p-value = 0.51

**The residual acf and pacf plots, the residual spectral plot, and Bartlett’s test all confirm reduction to white noise by the ARIMA model.**

(c) Repeat part (b) using the years 1993–2014.

**There is no need to introduce outlier dummies, and we fit the regression model with only monthly dummies to capture static seasonal structure.**

> model2<-lm(disneyreturn~fmonth,data=disney[265:528,]);summary(model2)

Call:

lm(formula = disneyreturn ~ fmonth, data = disney[265:528, ])

Residuals:

Min 1Q Median 3Q Max

-0.25180 -0.03249 0.00123 0.03987 0.22993

Coefficients:

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

(Intercept) 0.044567 0.015083 2.955 0.003426 \*\*

fmonth2 -0.015401 0.021331 -0.722 0.470980

fmonth3 -0.043150 0.021331 -2.023 0.044143 \*

fmonth4 -0.012925 0.021331 -0.606 0.545110

fmonth5 -0.032477 0.021331 -1.523 0.129136

fmonth6 -0.072981 0.021331 -3.421 0.000727 \*\*\*

fmonth7 -0.055565 0.021331 -2.605 0.009737 \*\*

fmonth8 -0.068858 0.021331 -3.228 0.001412 \*\*

fmonth9 -0.060564 0.021331 -2.839 0.004891 \*\*

fmonth10 -0.010518 0.021331 -0.493 0.622399

fmonth11 -0.004719 0.021331 -0.221 0.825079

fmonth12 -0.028621 0.021331 -1.342 0.180881

---

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

Residual standard error: 0.07075 on 252 degrees of freedom

Multiple R-squared: 0.114, Adjusted R-squared: 0.0753

F-statistic: 2.947 on 11 and 252 DF, p-value: 0.001089

***R*-square is rather low at 0.1140, and the residual standard error is 0.0708, which corresponds to movement of the return equal to about 0.7 percent.**

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(model2))

Shapiro-Wilk normality test

data: resid(model2)

W = 0.97586, p-value = 0.0001877

**Both tails of the residual distribution are long, and normality does not prevail.**

> options(digits=3)

> b1<-coef(model2)[1]

> b2<-coef(model2)[2:12]+b1

> b3<-c(b1,b2)

> seas2<-b3-mean(b3)

> seas2

(Intercept) fmonth2 fmonth3 fmonth4 fmonth5 fmonth6

0.03381 0.01841 -0.00934 0.02089 0.00134 -0.03917

fmonth7 fmonth8 fmonth9 fmonth10 fmonth11 fmonth12

-0.02175 -0.03504 -0.02675 0.02330 0.02910 0.00519

> cbind(1:12,seas2)

seas2

January 1 0.03381

February 2 0.01841

March 3 -0.00934

April 4 0.02089

May 5 0.00134

June 6 -0.03917

July 7 -0.02175

August 8 -0.03504

September 9 -0.02675

October 10 0.02330

November 11 0.02910

December 12 0.00519

A graph with numbers and lines

Description automatically generated

**Disney returns for 1993 to 2014 peak in January, estimated at 3.4 percent above trend level. There are high value estimated indices for October and November, at 2.3 percent and 2.9 percent, respectively, above trend. The high reading in January is associated with buying at the beginning of the tax year. There is also a local peak in April, estimated to be 2.1 percent above trend, and February is comparable at 1.8 percent above trend. The low months are June through September, estimated to be between 2 and 4 percent below trend.**

**Next we fit an ARIMA model.**

A screenshot of a graph

Description automatically generated

**There are significant partial correlations at lags 6, 18, and 22. Even though there is significance at lag 22, rather than 24, we try an ARIMA(0,0,0)(4,0,0)6 model.**

> armodel2<-arima(resid(model2),order=c(0,0,0),seasonal=list(order=c(4,0,0),period=6))

> armodel2

Call:

arima(x = resid(model2), order = c(0, 0, 0), seasonal = list(order = c(4, 0,

0), period = 6))

Coefficients:

sar1 sar2 sar3 sar4 intercept

-0.1756 -0.0206 -0.1083 -0.1057 -1e-04

s.e. 0.0611 0.0626 0.0626 0.0619 3e-03

sigma^2 estimated as 0.004552: log likelihood = 336.87, aic = -661.75

> coeftest(armodel2)

z test of coefficients:

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

sar1 -1.7556e-01 6.1085e-02 -2.8741 0.004052 \*\*

sar2 -2.0569e-02 6.2557e-02 -0.3288 0.742300

sar3 -1.0832e-01 6.2574e-02 -1.7311 0.083437 .

sar4 -1.0571e-01 6.1901e-02 -1.7078 0.087675 .

intercept -8.6107e-05 2.9926e-03 -0.0288 0.977046

---

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

**The model gives significance for the estimate at lag 6 and marginal significance for estimates at lags 18 and 24. The signal is rather weak.**

**We perform residual analysis.**

A graph of a normal q-q plot

Description automatically generated

> shapiro.test(resid(armodel2))

Shapiro-Wilk normality test

data: resid(armodel2)

W = 0.97486, p-value = 0.0001306

**The lower tail of the distribution of the ARIMA residuals is long, and the distribution does not conform to normality.**

A comparison of a graph

Description automatically generated

A graph of a graph

Description automatically generated

> bartlettB.test(resid(armodel2))

Bartlett B Test for white noise

data:

= 0.80013, p-value = 0.5439

**The residual pacf plot shows very mild significance at lags 22 and 24, but the residual spectral plot and Bartlett’s test indicate adequate reduction to white noise.**

(d) Compare the static seasonal estimates for 1971–1992 and 1993–2014, and write a summary of your findings.

> cbind(1:12,seas,seas2)

1971–92 1993–2014

January 1 0.03530 0.03381

February 2 0.02370 0.01841

March 3 0.00901 -0.00934

April 4 -0.00116 0.02089

May 5 0.01601 0.00134

June 6 -0.01809 -0.03917

July 7 -0.03011 -0.02175

August 8 -0.01514 -0.03504

September 9 -0.04086 -0.02675

October 10 -0.01422 0.02330

November 11 0.00512 0.02910

December 12 0.03045 0.00519

A graph with blue and red lines

Description automatically generated

**There are differences in static seasonal estimation for the two time spans, indicating the presence of dynamic seasonality. January indices are the same, but the estimated index for December is lower and the estimated indices for October and November are higher for 1993–2014. In addition, the local peak in May for 1971–1992 has shifted to April for 1993–2014. Although there are some differences for the summer months, they give the lowest estimated static estimates for both time spans.**