Take Home Midterm (STAT 6555)

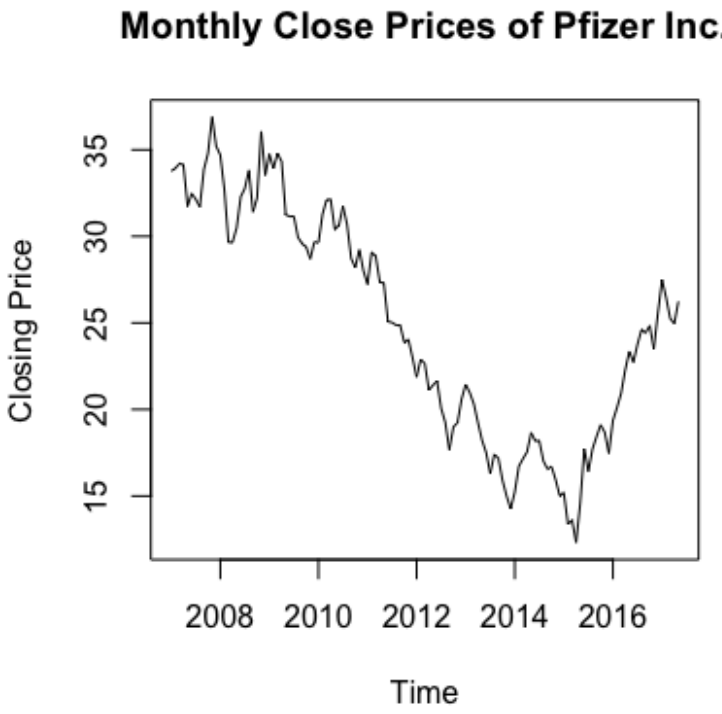
Spring 2017

Time Series Analysis of Closing Stock Prices using the ARIMA model

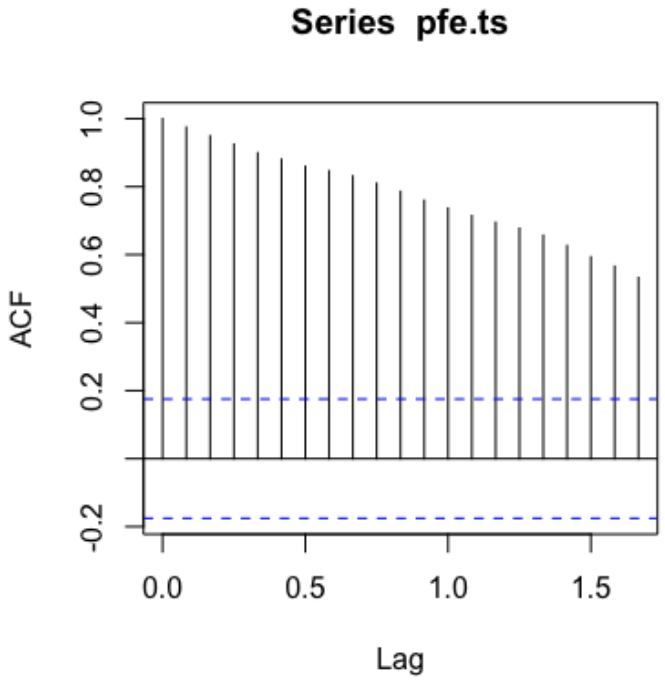
Submitted By: Bheeni Garg (uf4522)

The objective of the report is to analyze and forecast the closing stock prices of a company using ARIMA model. I have chosen the stocks of Pfizer Inc. for a period of 10 years.

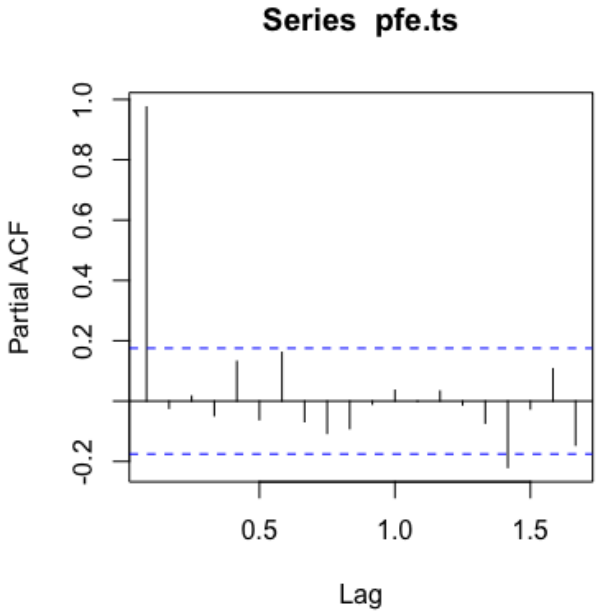
Closing prices (US dollars) for Pfizer Inc. stock for 125 trading months from January 1, 2007 up to May 1, 2017 are read into R using the Quandl library and the time series is plotted.



The series shows a clear decreasing trend until second half of 2015 and then shows an increasing trend. The ACF plot looks like below:



The ACF plot shows that the autocorrelations are significant and positive and gradually decreasing. The autocorrelations are significant for a large number of lags but perhaps the autocorrelations at lags 2 and above are merely due to the propagation of the autocorrelation at lag 1. This is confirmed by the PACF plot:

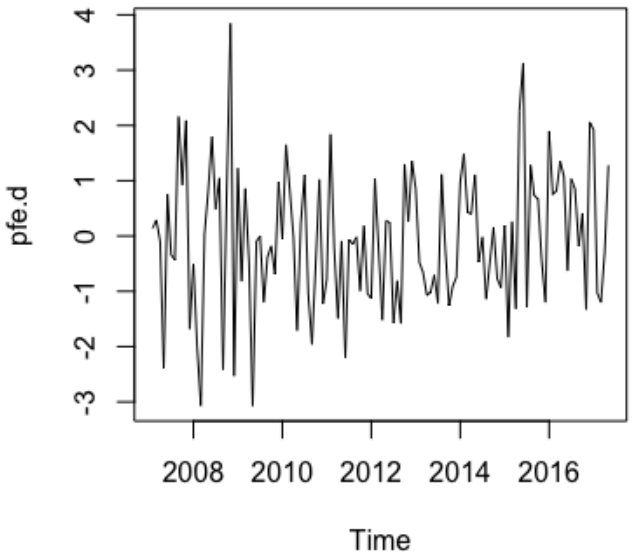


It can be noted that the PACF plot has a significant spike at lag 1, meaning that all the higher-order autocorrelations are effectively explained by the lag-1 autocorrelation. There is another mildly significant lag at month 17 which probably indicates that it provides some additional information about the stock price.

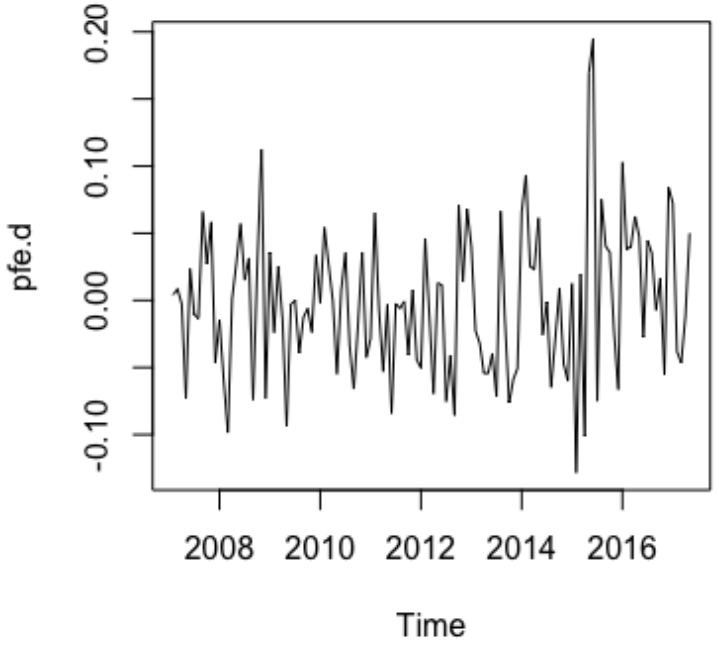
In order to model the time series, we need to take the following steps for greater precision and accuracy:

1) Apply a suitable transformation to decrease the variation in the values.
2) Remove any trend effect and make the series stationary. This is achieved by differencing the transformed data.
3) Identify the dependence orders of ARIMA and coefficient estimation
4) Run diagnostic checks to choose the best model
5) Forecast using the final model

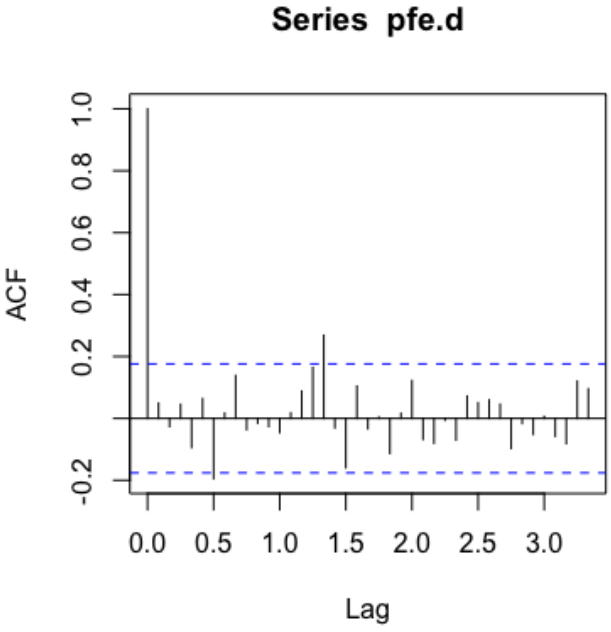
I start with checking the variance to see if any transformation is required. A strong trend, however, hides any effect so I first difference the original series.

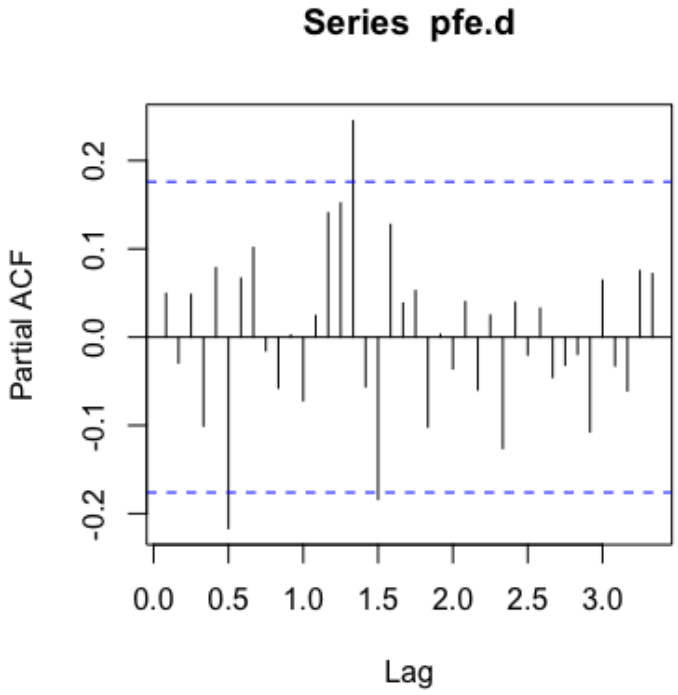


Now that the series is somewhat stationary (slight increasing trend can still be noted), the variance looks non-constant. So, we apply a log transformation and then difference it to get the following plot:



The variance is slightly improved and the series look stationary. We now plot the ACF and the PACF of the transformed and differenced series to determine whether AR or MA terms are needed and if yes, then also determine the order.

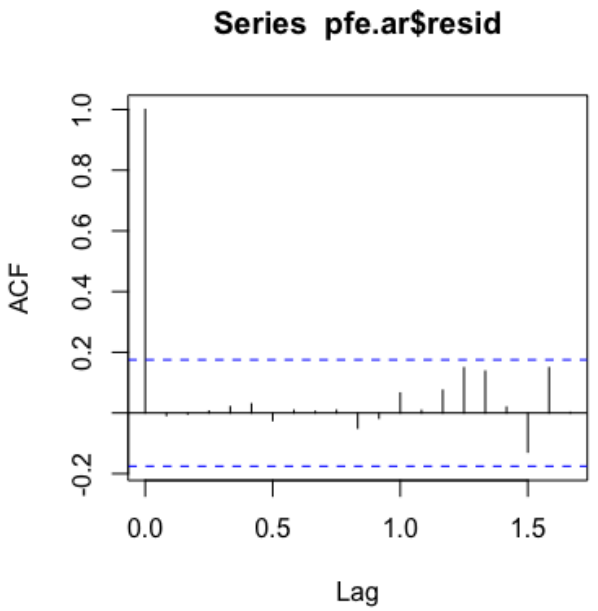


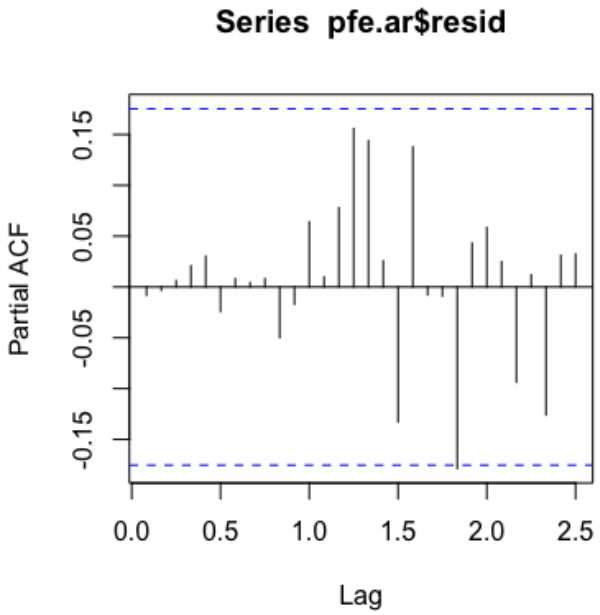


The ACF plot looks ok but a slightly significant lag at month 16 indicates a higher order MA process. There are significant lags at months 6 and 16 as seen in the PACF plot also indicating a higher order AR process. The ACF and PACF also indicate a slight seasonal trend in the series.

Moving to step 3, different AR and/or MA models are fitted and the autocorrelations of the residuals are checked to get further insights on the order. On fitting the lower order AR and MA models, the ACF look like white noise but the PACF continue to show highly significant lags for months 6 and 16. The first order differencing is also taken into account during modelling.

The model ARIMA (4,1,6) is the final model chosen with the ACF and PACF plots for the residuals that look fine. Apparently, the PACF improves as the order of the MA is increased which indicates a high correlation with the current and past errors (lag 18). The weakly significant lag at month 18 in the PACF plot is ignored in order to achieve a parsimonious model.





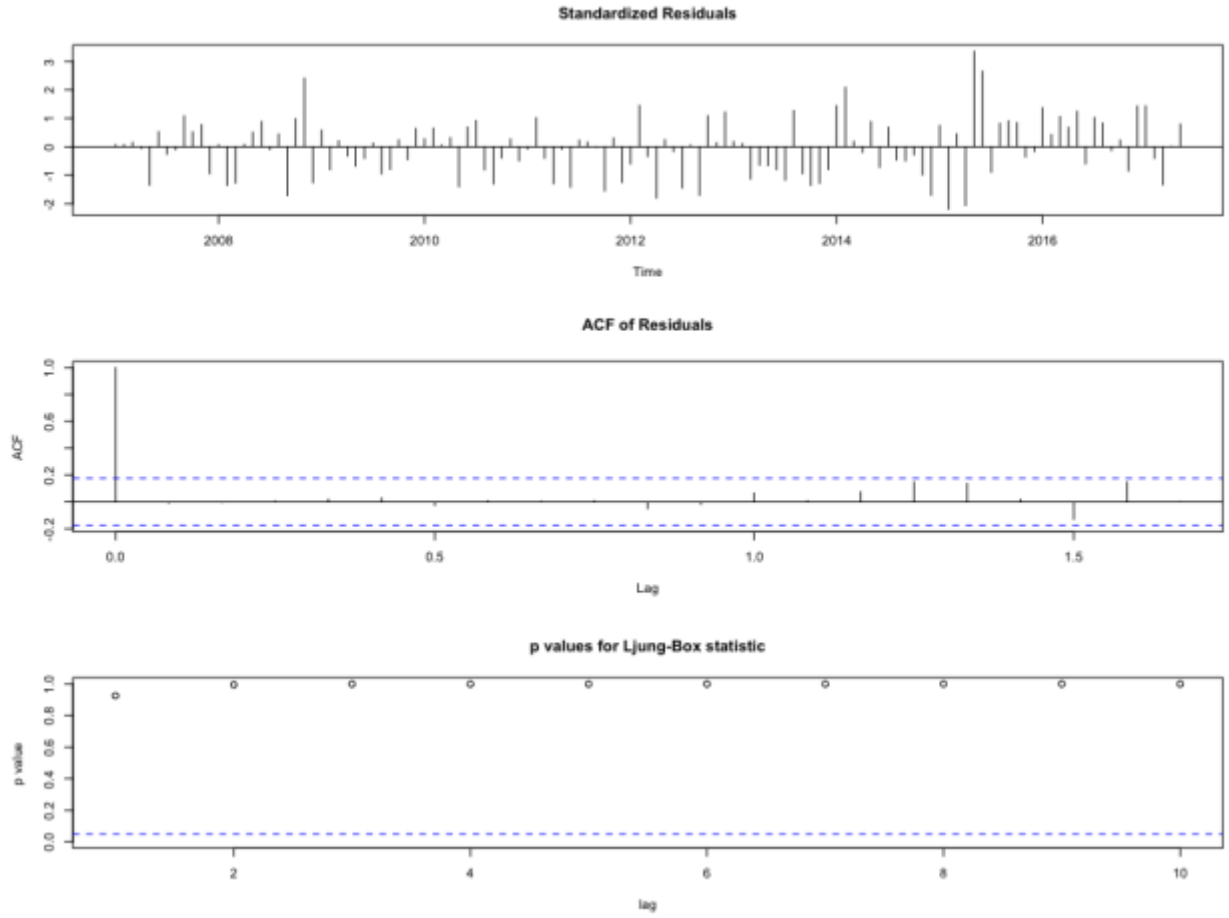
The following table gives the summary of the fitted model:

```
Call:
arima(x = pfe.tr, order = c(4, 1, 6))

Coefficients:
          ar1      ar2      ar3      ar4     ma1     ma2     ma3     ma4     ma5      ma6
      -0.1503  -0.3736  -0.0933  -0.6306  0.2390  0.3748  0.2066  0.5895  0.1028  -0.2868
s.e.   0.1831   0.1508   0.1708   0.2145  0.1864  0.1617  0.1693  0.2084  0.1047   0.1046

sigma^2 estimated as 0.002525:  log likelihood = 194.09,  aic = -366.17
```

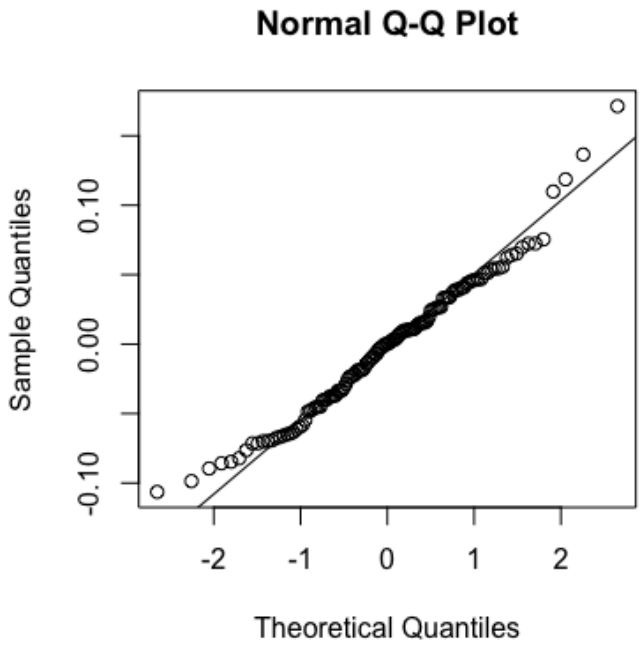
The autoregression coefficients ϕ_1, ϕ_2 , ϕ_3, ϕ_4 after reverse log transformation are 0.860, 0.688, 0.910, 0.532 and the moving average coefficients $\theta_1, \theta_2, \theta_3$, θ_4 , θ_5 , θ_6 are 1.27, 1.45, 1.23, 1.80, 1.10, 0.75 respectively.

Diagnostic Plots

Running the diagnostic check on the ARIMA(4,1,6) gives the following plots:



The standardized residual plot looks somewhat random and the ACF plot looks fine too.



The Q-Q plot shows that the residuals are somewhat normal for lower values but become highly non-normal as the values increase.

Forecasting

Finally, after the model is built, I now use it to forecast stock prices for the next two years. Although, the stock price prediction in not very accurate for long lags, I do it for the sake of better visibility in the plot.

The predicted values are from June 2017 to May 2019 are as below after the log transformation is reversed:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
2017						26.13994	25.24563	25.68197	26.25696	26.25297
2018	25.61682	25.78905	25.90230	26.06128	26.17972	25.98146	25.88070	25.85853	25.84381	25.98795
2019	25.97573	25.89967	25.89009	25.95035	25.96189					

	Nov	Dec
2017	26.26913	25.93224
2018	26.03755	25.99148
2019		

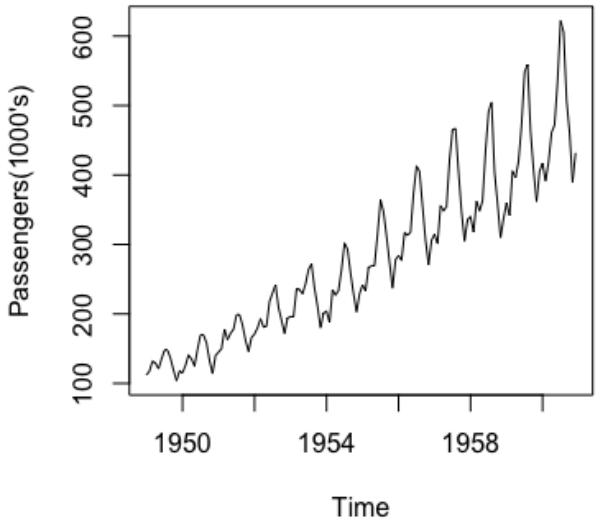
The following plot shows the forecasted trend in points with dashed lines indicating the upper and lower bounds.

As expected the variance is pretty high owing to the volatility of the stock markets. Also, the forecast bounds get exceptionally high moving further in time.

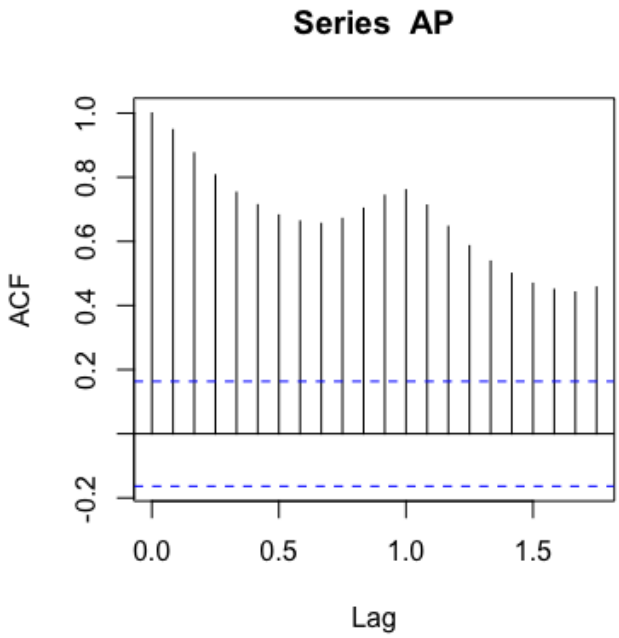
2) Air Passengers Dataset

Studying the data

Once the dataset is loaded into R, we plot the time series:



The ACF plot is created to understand the data.



The plot clearly shows an increasing trend indicating the increase in the number of air passengers with time. Also, a repeating pattern shows seasonal variation in the time series which again, appears to be increasing. The ACF plot shows that the lags are positively correlated and a spike at the end of 1 year shows seasonal effect.

The series is transformed to account for the non-constant variance. A regression model is then fitted to the series using R resulting in the following output :

```
Call:
lm(formula = log(AP) ~ Time + as.factor(Seas))

Residuals:
     Min       1Q   Median       3Q      Max
-0.156370 -0.041016  0.003677  0.044069  0.132324

Coefficients:
                   Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)      -2.308e+02  2.798e+00 -82.467  < 2e-16 ***
Time              1.208e-01  1.432e-03   84.399  < 2e-16 ***
as.factor(Seas)2 -2.206e-02  2.421e-02  -0.911   0.36400
as.factor(Seas)3  1.082e-01  2.421e-02   4.468  1.69e-05 ***
as.factor(Seas)4  7.690e-02  2.421e-02   3.176   0.00186 **
as.factor(Seas)5  7.453e-02  2.422e-02   3.078   0.00254 **
as.factor(Seas)6  1.967e-01  2.422e-02   8.121  2.98e-13 ***
as.factor(Seas)7  3.006e-01  2.422e-02  12.411  < 2e-16 ***
as.factor(Seas)8  2.913e-01  2.423e-02  12.026  < 2e-16 ***
as.factor(Seas)9  1.467e-01  2.423e-02   6.054  1.39e-08 ***
as.factor(Seas)10 8.532e-03  2.423e-02   0.352  0.72537
as.factor(Seas)11 -1.352e-01  2.424e-02  -5.577 1.34e-07 ***
as.factor(Seas)12 -2.132e-02  2.425e-02  -0.879  0.38082
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0593 on 131 degrees of freedom
Multiple R-squared:  0.9835,        Adjusted R-squared:  0.982
F-statistic: 649.4 on 12 and 131 DF,  p-value: < 2.2e-16
```

Since dummy variables for a characteristic (seasons here) are always compared to the missing value (month 1 here), the baseline for the model is the first month. It can be noted that the seasonal effects (month) for 2, 10 and 12 are not statistically significant (or different) than month 1.

The full estimated model is :

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_{s2} X_{s2,t} + \hat{\beta}_{s3} X_{s3,t} + \hat{\beta}_{s4} X_{s4,t} + \hat{\beta}_{s5} X_{s5,t} + \hat{\beta}_{s6} X_{s6,t} + \hat{\beta}_{s7} X_{s7,t} + \hat{\beta}_{s8} X_{s8,t} + \hat{\beta}_{s9} X_{s9,t} + \hat{\beta}_{s10} X_{s10\ t} + \hat{\beta}_{s11} X_{s11,t} + \hat{\beta}_{s12} X_{s12,t}$$

$\widehat{Bookings}_t$ = -0.02308 + 0.1208*Time - 0.02.206*(Seas)2 + 0.108*(Seas)3 + 0.0769*(Seas)4 + 0.0745*(Seas)5 + 0.1967*(Seas)6 + 0.3*(Seas)7 + 0.2913*(Seas)8 + 0.1467*(Seas)9 + 0.0085*(Seas)10 – 0.135*(Seas)11 - 0.0213*(Seas)12

Interpreting Trend and Seasonality

Time, representing successive months, is interpreted as the effect of the linear trend in bookings over time, holding the effect of the seasons constant. The values need to be reverse transformed before interpreting.

$\hat{\beta}_1$: Each additional month sees an estimated increase of $\hat{\beta}_1$bookings, after adjusting for the season.

$\hat{\beta}_{s2}$: After accounting for the trend, estimated bookings in the second month are about $\hat{\beta}_{s2}$ more than first month bookings although not significant in the model

$\hat{\beta}_{s3}$: After accounting for the trend, estimated bookings in the third month are about $\hat{\beta}_{s3}$ more than first month bookings.

Likewise, all the other coefficients are interpreted.

1 Year Forecast with Prediction Intervals

Using the predict function in R, I get the following predicted values:

```
1        2        3       4        5          6        7        8        9       10       11
486.27 480.47  552.84  541.24  545.4238  622.5214 697.7000 698.2402 610.3285 536.9516 469.7774
    12
531.7603
```

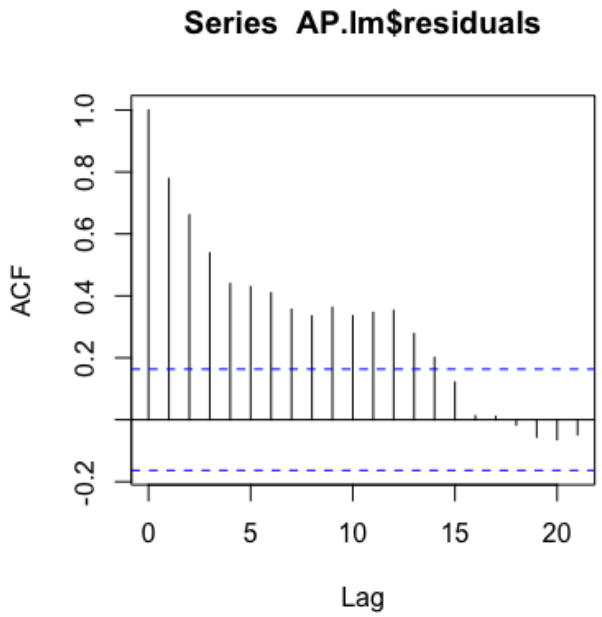
The 95% prediction intervals can be calculated using the formula:

$\hat{Y}_t \pm 1.96 * s.e.$

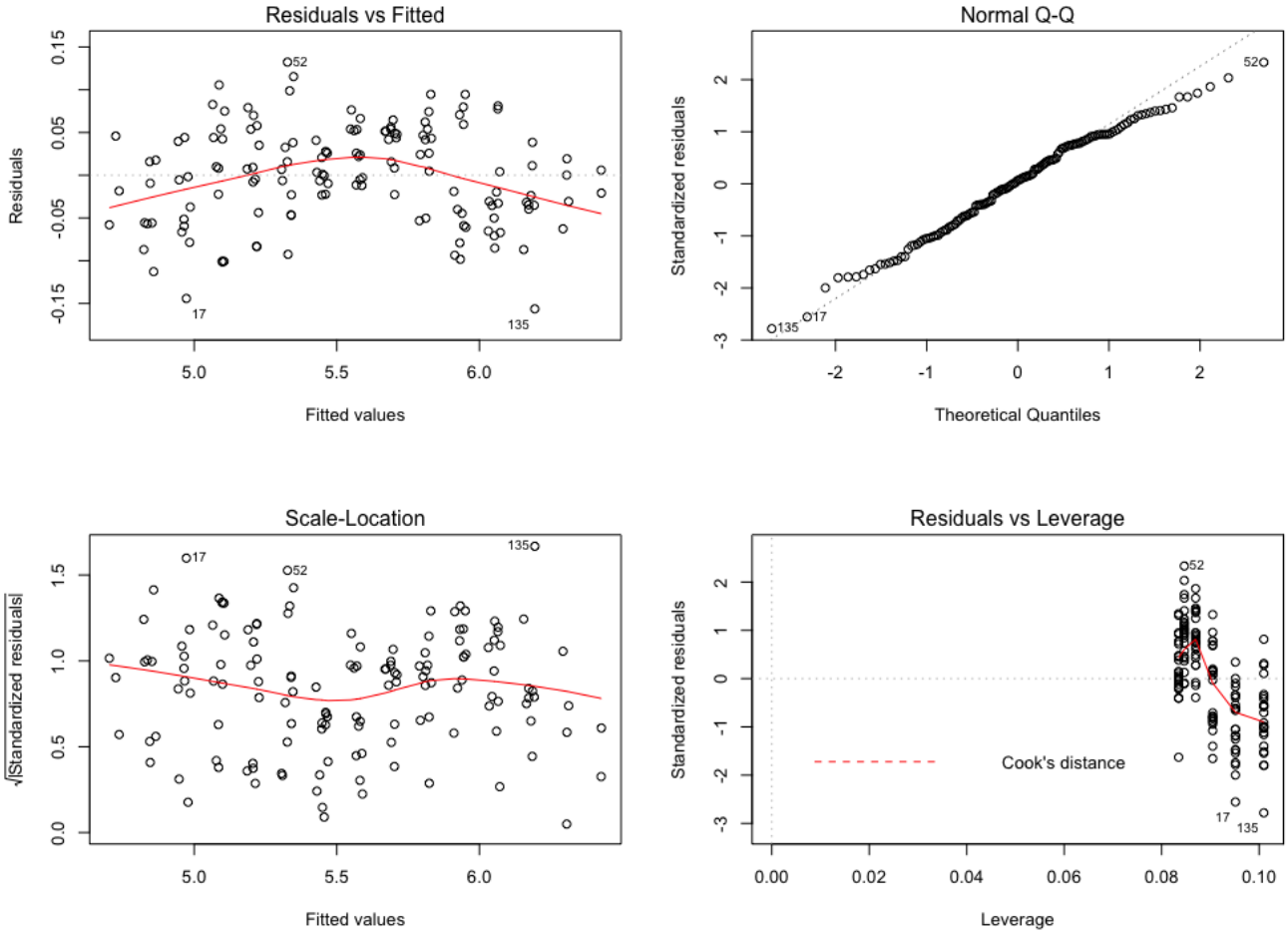
$\widehat{Bookings}_{Jan}$ = (454.1030, 518.4302)
$\widehat{Bookings}_{Feb}$ = (478.513 482.433)
$\widehat{Bookings}_{march}$ = (550.8399, 554.8383)
$\widehat{Bookings}_{April}$ = (539.2430, 543.2414)
$\widehat{Bookings}_{May}$ = (543.4246, 547.4230)
$\widehat{Bookings}_{June}$ = (620.5222 ,624.5206)
$\widehat{Bookings}_{July}$ = (695.7008, 699.6992)
$\widehat{Bookings}_{Aug}$ = (696.2410 ,700.2394)
$\widehat{Bookings}_{Sep}$ = (608.3293, 612.3277)
$\widehat{Bookings}_{Oct}$ = (534.9524, 538.9508)
$\widehat{Bookings}_{Nov}$ = (467.7782, 471.7766)
$\widehat{Bookings}_{Dec}$ = (529.761, 533.7595)

Checking Assumptions and reporting Issues

The autocorrelation plot of the residuals show that the lags are still highly correlated and looks far from white noise. A seasonal pattern is still prevalent so a better model is needed.



Diagnostics Plots



The residuals vs fitted plot shows a convex trend again indicates a better model. The Q-Q plot shows that the residuals are fairly normal. The Shapiro Wilk p- value is 0.3408 which confirms normality.

R Code

```
library(Quandl)
library(forecast)

## Analysis of Stock Price data
## Loading the data
pfe <- Quandl("EOD/PFE", api_key="p_rCezUsxM28xLb3i942", start_date="2007-01-01", end_date="2017-05-1", collapse = "monthly")
pfe.ts <- ts(pfe[,5], start = 2007, frequency = 12)

## plot the series
plot(pfe.ts, type = "l", ylab = "Closing Price", main = "Monthly Close Prices of Pfizer Inc.")
str(pfe.ts)

## ACF
acf(pfe.ts)
pacf(pfe.ts)

## differencing
pfe.d <- diff(pfe.ts)
plot(pfe.d, type ="l")

## transformation and differencing
pfe.tr <- log(pfe.ts)
pfe.d <- diff(pfe.tr, 1)
plot(pfe.d, type ="l")
acf(pfe.d, lag.max = 40)
pacf(pfe.d, lag.max = 40)

## ARIMA model building
pfe.ar <- arima(pfe.tr, order=c(4,1,6))
acf(pfe.ar$resid)
pacf(pfe.ar$resid, lag.max = 30)
summary(pfe.ar)

# diagnostic plots
tsdiag(pfe.ar)
qqnorm(pfe.ar$residuals)
qqline(pfe.ar$residuals)

## Forecasting
pfe.pr <- predict(pfe.ar, n.ahead= 2*12)
pfe.pr$pred
pfe.p <- exp(pfe.pr$pred)
pfe.p
```

```
## creating forecast bounds
U <- exp(pfe.pr$pred + pfe.pr$se)
L <- exp(pfe.pr$pred - pfe.pr$se)
ts.plot(pfe.ts, pfe.p, col=1:2, ylab = 'Close Price($)')
lines(pfe.p, col=2, type='p')
lines(U, col='blue', lty='dashed', lwd = 2)
lines(L, col='blue', lty='dashed', lwd = 2)

# Air Passengers dataset
# load the data
AP <- AirPassengers
str(AP)
AP

# basic time series plot
plot(AP, type = "l", ylab = "Passengers(1000's)")
acf(AP)

# Seasonal component
Seas <- cycle(AP)

# Trend Component
Time <- time(AP)
Time

# Time Series Regression model
AP.lm <- lm(log(AP) ~ Time + as.factor(Seas))
summary(AP.lm)
acf(AP)
par(mfrow=c(2,2))
plot(AP.lm)

new.t <- seq(1961, len = 1 *12, by=1/12)
new.t
new.dat <- data.frame(Time = new.t, Seas = rep(1:12, 1))
pr <- predict(AP.lm, new.dat)[1:12]
exp(pr)

## 95% prediction intervals
486.2666 + c(-1.96,1.96)*16.41
480.4730 + c(-1.96,1.96)*1
552.8391 + c(-1.96,1.96)*1.02
541.2422 + c(-1.96,1.96)*1.02
545.4238 + c(-1.96,1.96)*1.02
622.5214 + c(-1.96,1.96)*1.02
```

```
697.7000 + c(-1.96,1.96)*1.02
698.2402 + c(-1.96,1.96)*1.02
610.3285 + c(-1.96,1.96)*1.02
536.9516 + c(-1.96,1.96)*1.02
469.7774 + c(-1.96,1.96)*1.02
531.7603 + c(-1.96,1.96)*1.02

# Diagnostics
par(mfrow=c(2,2))
plot(AP.lm)
acf(AP.lm$residuals)
par(mfrow=c(2,2))
plot(AP.lm)
shapiro.test(AP.lm$residuals)
```
