Assignment1_Seshadri

Sri Seshadri

4/7/2018

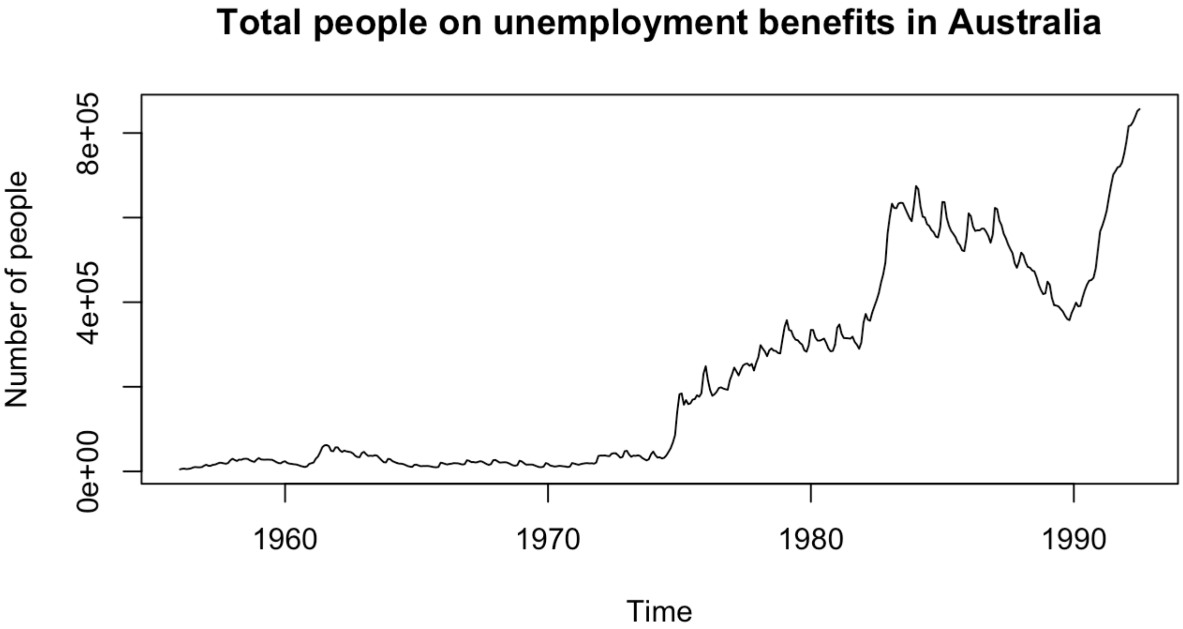
Section 2.8, Question 2.1 page 59 - Effect of transformation on time series data.

a) Monthly total of peole on unemployment benefits in Australia (Jan 1956 - July 1992)

Figure 1 shows the time series of number of people in unemplyment benefits in Australia by month. The time series is affected by

```
    *  Population growth over time
    *  External factors like:
        * state of the economy
        * the benefit provided by the government.
```

It would be useful to normalize the data by population, to get the percent unemployed of the total population. Then if need be a transformation on the normalized data can be made.



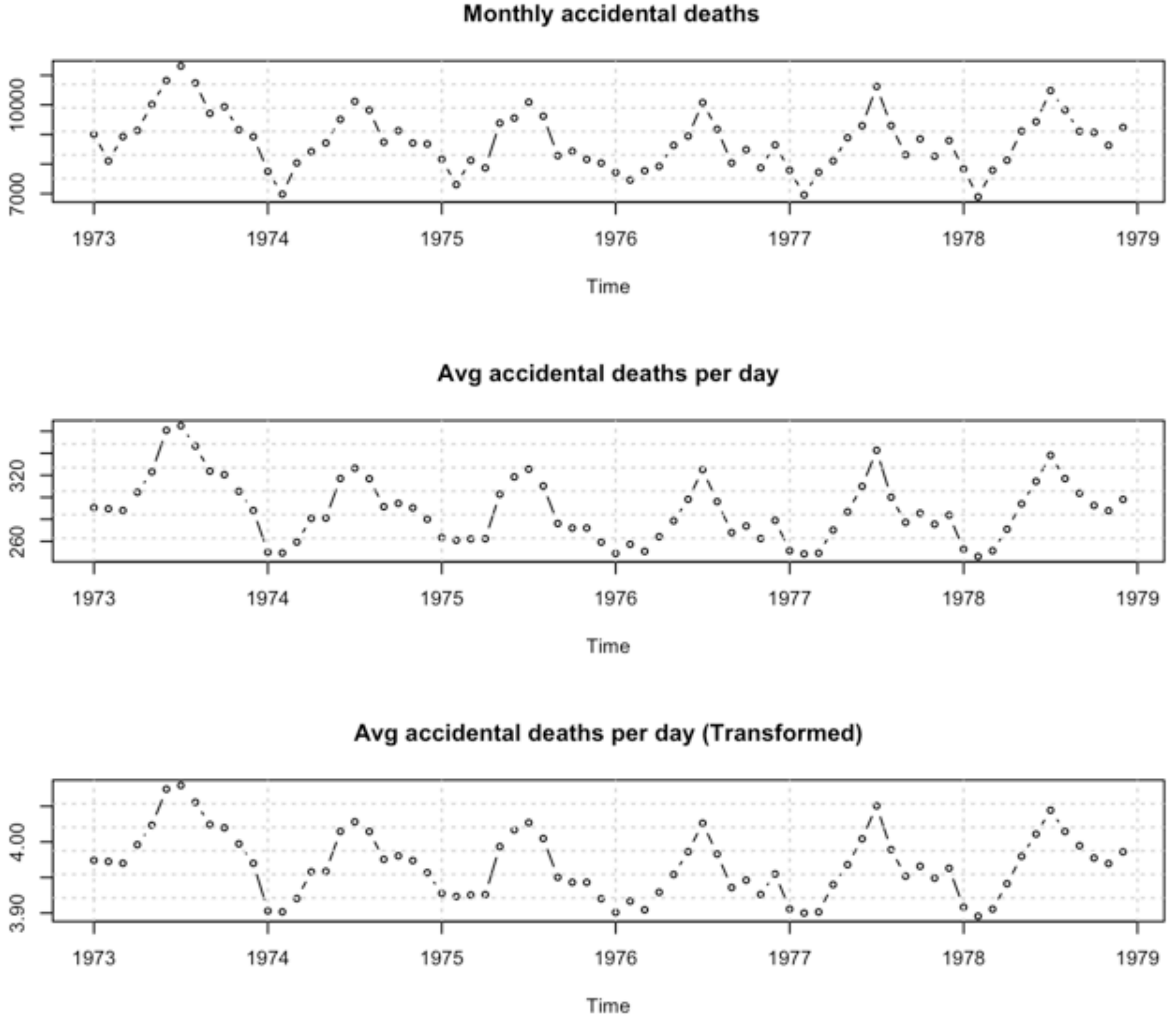
Monthly number of people on unemployment benefits in Australia

b) Monthly total accidental deaths in the United States

The top plot in figure 2 shows the total accidental deaths by month in the US. There is seasonality in the data, where the total accidents peaking at July. The variation in the seasonality may be mitigated by normalizing the totals by dividing by the number of days in the month.

The middle plot in figure 2 shows the normalized total by days in the months; i.e. Monthly average accidental deaths per day. We see that there is some smoothing of the raw data. While there is not much variation in seasonality, there is interest in further making the size of the seasonal variation equal across seasons.

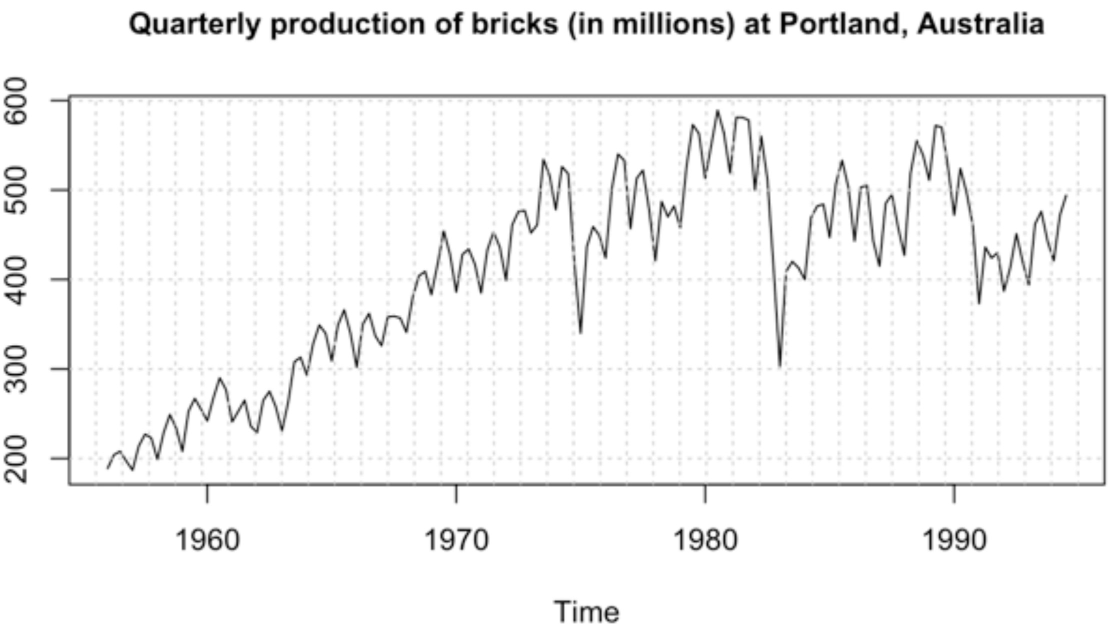
The bottom chart of figure 2 shows a box-cox transformation of the average accidental deaths. Its seen that there isn’t much affect as expected.

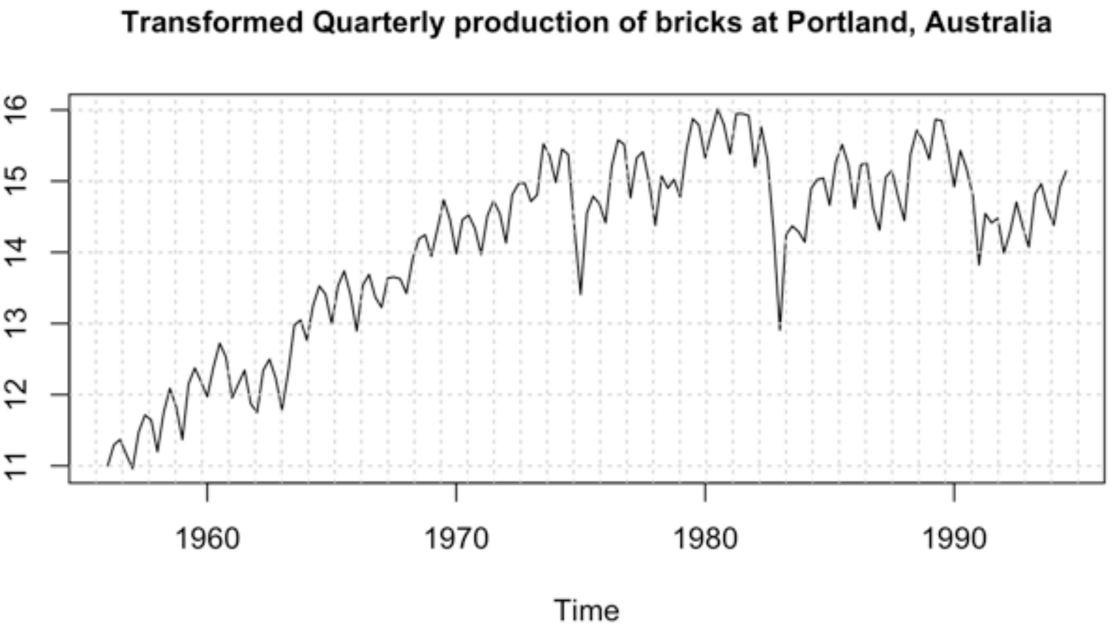


Top: Total accidental deaths by month in the US, Middle: Monthly Average Accidental deaths per day, Bottom: Transformed monthly average accidental deaths per day

c) Quarterly production of bricks (in millions) at Portland, Australia

The top chart in Figure 3 shows the time series plot of quarterly brick production at Portland Australia. The time series exhibits an increasing trend and seasonality. The variation increases with time / levels. Box-Cox transformation is appropriate for this case. The ideal lambda for the data was 0.255. The bottom chart of figure 3 shows the transformed data. The variation is better across time, of course the huge downward spikes is not fully mitigated.



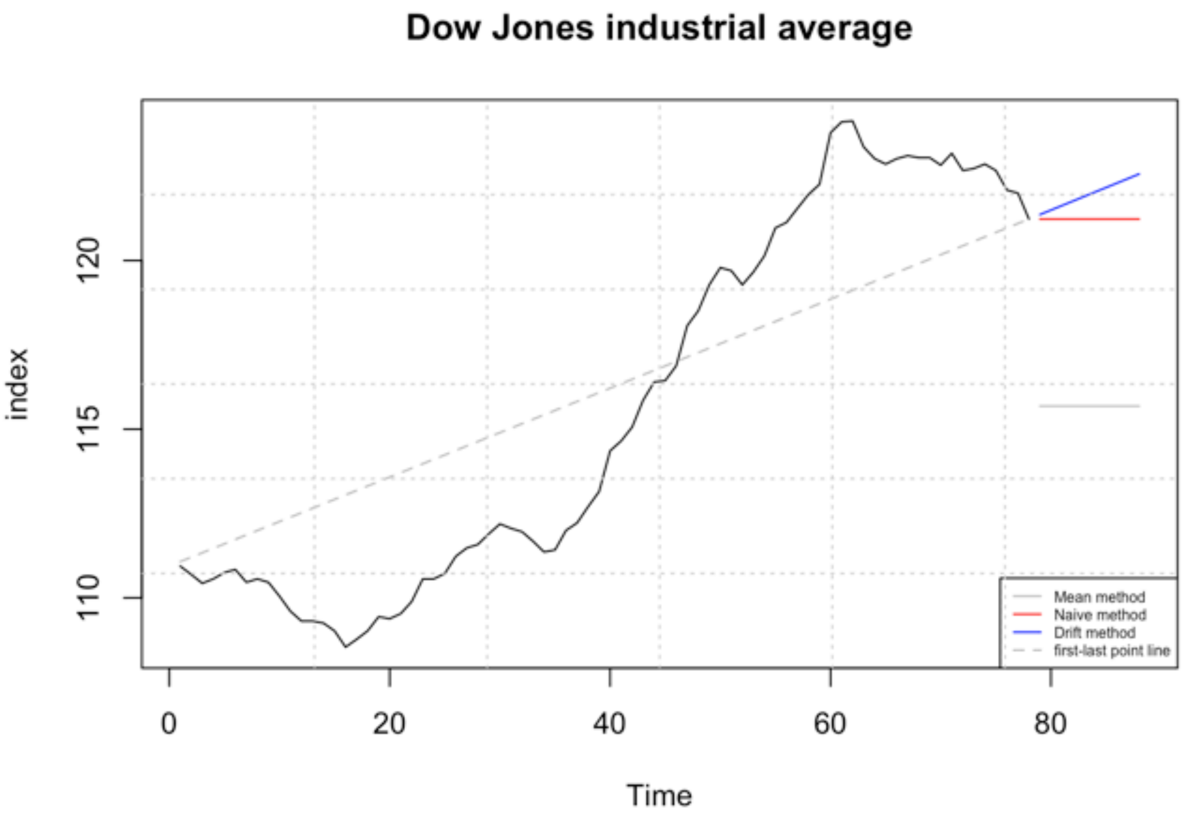


Top: Quarterly production of bricks (in millions) at Portland, Australia, Bottom : BoxCox transformed Quarterly production of bricks

Question 2.2 Page 60

Time series modeling of Dow Jones index

In this section the Dow jones industrial average is modeled as a time series. The drift method is used to fit the time series and forecasted for the next 10 periods. Figure 4 shows the time series plot with forecast from drift,mean and naive mean methods. Table 1 shows the forecasts from drift method and the blue line in fig 4 plots the forecasts. The drift method forecast is nothing but the extension of the line that joins the first and the last point. The slope of the line is the difference between the last and the first point divided by the number of data points. The grey dashed line in fig 4 shows the fitted drift for the Dow Jones data.



Time series modeling of Dow Jones industrial average

Drift method forecast for 10 periods

Time	index
79	121.3636
80	121.4973
81	121.6309
82	121.7645
83	121.8982
84	122.0318
85	122.1655
86	122.2991

87	122.4327
88	122.5664

Comparison between models

The drift method is compared with other models like the Mean of the time series, the naive method and the seasonal naive method. The accuracy methods are compared in table 2. Since the time series is set up as a daily closing index. The seasonal forecast for the next point is identical as the previous point (same season as the last point). Hence the Naive and Seasonal Naive forecasts are the same for this scenario.

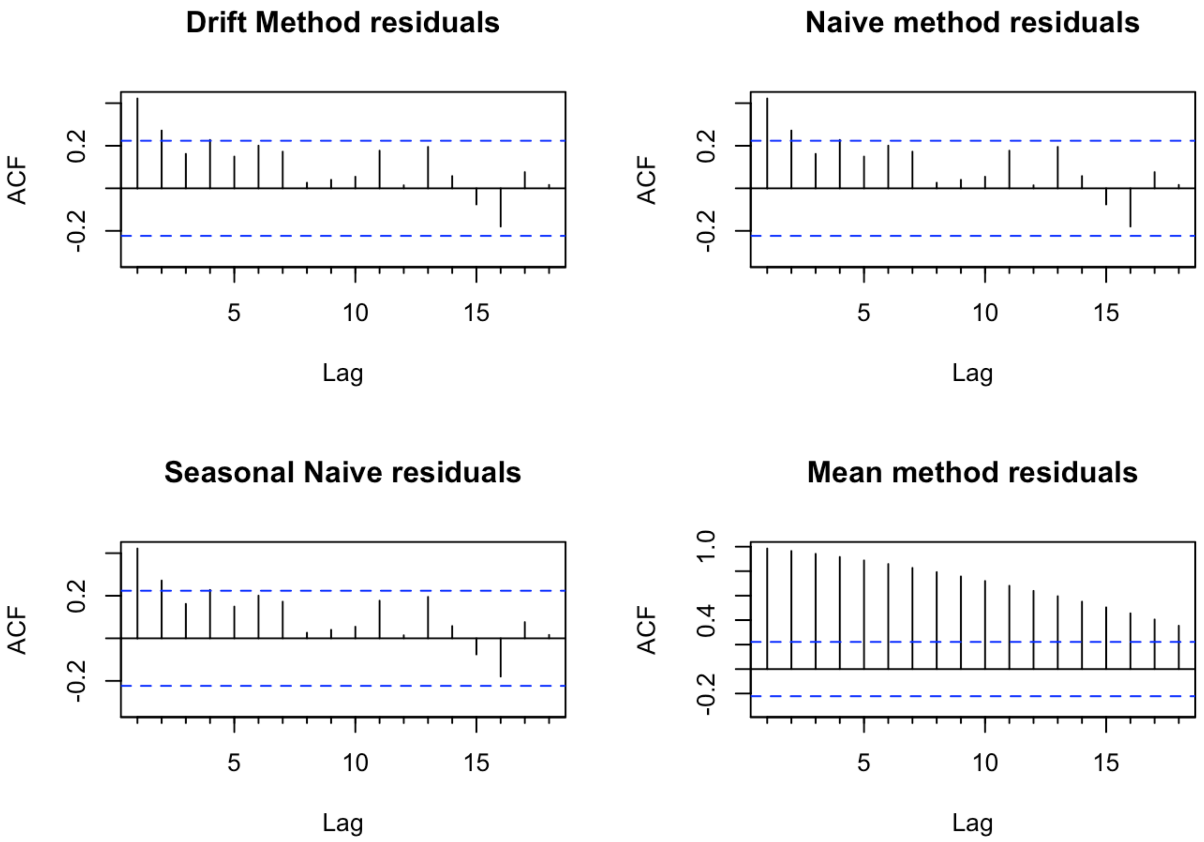
The Drift method is the best performing model amongt the competing naive models, based on MAE / MAPE. This is not surprising as we see an increasing trend in the Dow Jones index. The other methods use the mean or the last value of the time series as the forecast. The naive models do not account for the increasing trend. The Drift method does.

Accuracy metrics of benchmark models

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Drift	0.000	0.424	0.325	-0.001	0.280	0.952	0.422
Mean	0.000	5.471	5.104	-0.222	4.400	14.938	0.985
Naive	0.134	0.445	0.342	0.114	0.294	1.000	0.422
SeasonalNaive	0.134	0.445	0.342	0.114	0.294	1.000	0.422

Residual Analysis

Figure 5 and 6 shows the residual analysis of the time series model fits. The mean model’s residuals show that the means are significantly different from the other models and show it’s poor performance.



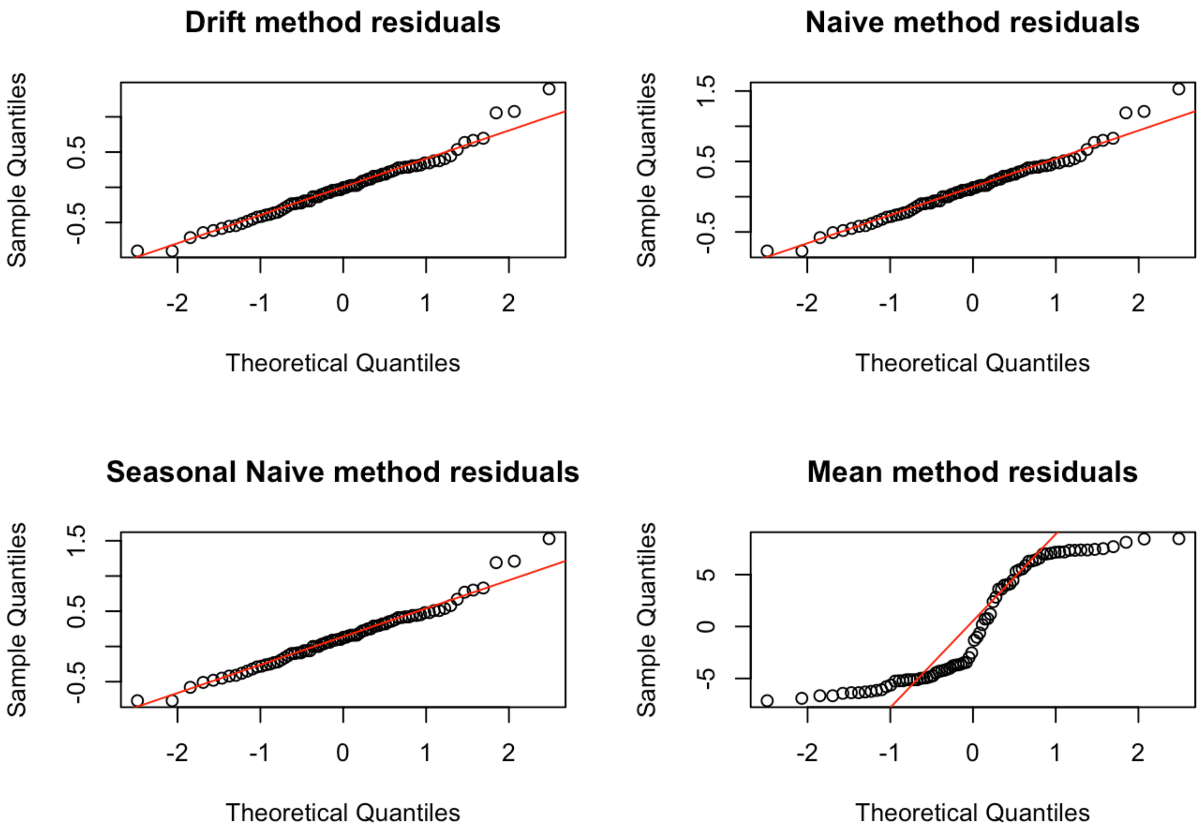
Drift Method residuals

Naive method residuals

Seasonal Naive residuals

Mean method residuals

Autocorrelation plot of residuals



Drift method residuals

Naive method residuals

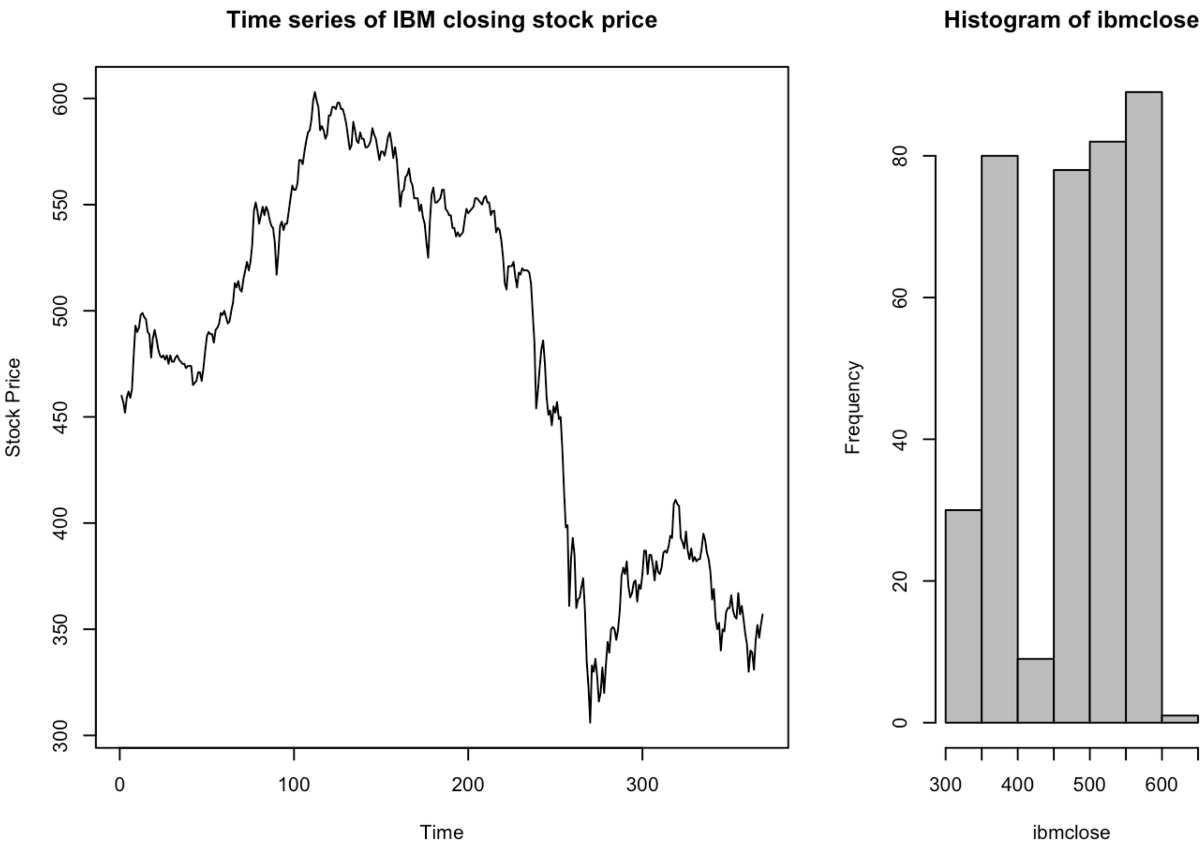
Seasonal Naive method residuals

Mean method residuals

Residual Analysis

Question 2.3 IBM Stock prices

Figure 7 shows the time series plot of IBM stock price at close. The data is split into training and test set. The first 300 data points are retained as training set and the rest of the 69 data points are held out as test set.



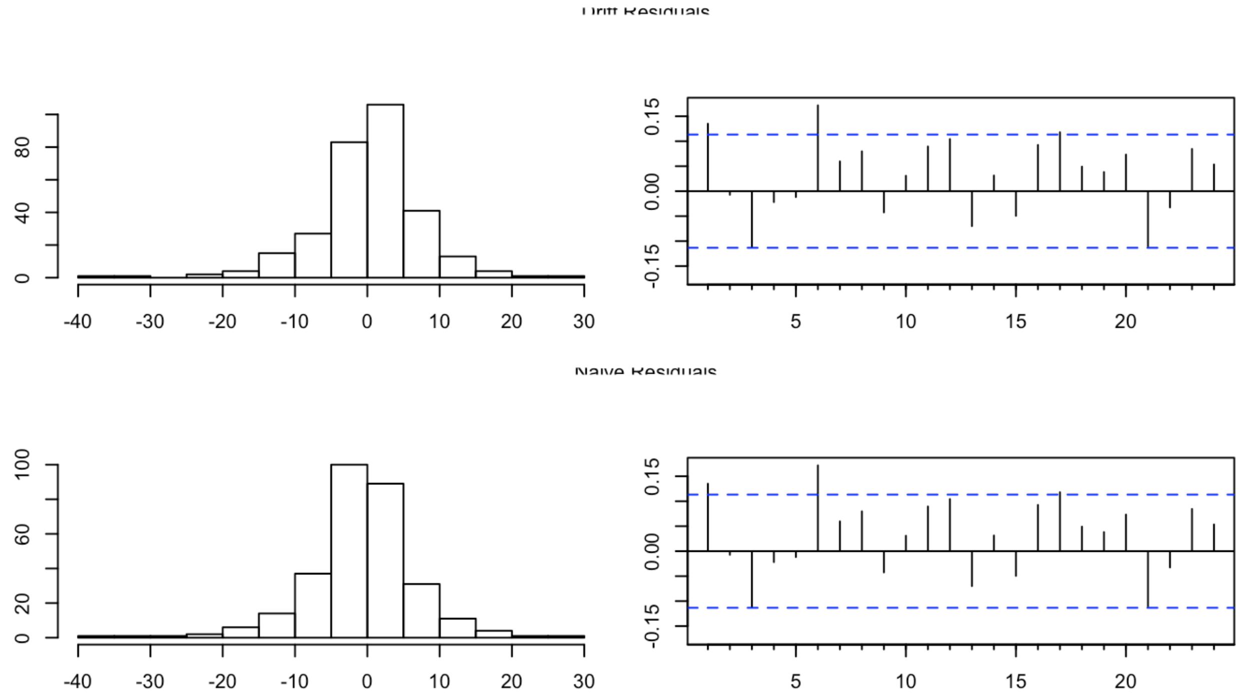
IBM Stock price at close

Time series modeling of the training set

The Drift, naive and mean method modeling was trained on the training set and the accuracy statistics are shown in table 3. Since the distribution of the data is bimodal, BoxCox transformation was deemed not appropriate. Figure 9 shows the residual analysis. Both the naive and the drift methods are very close in their forecasts.

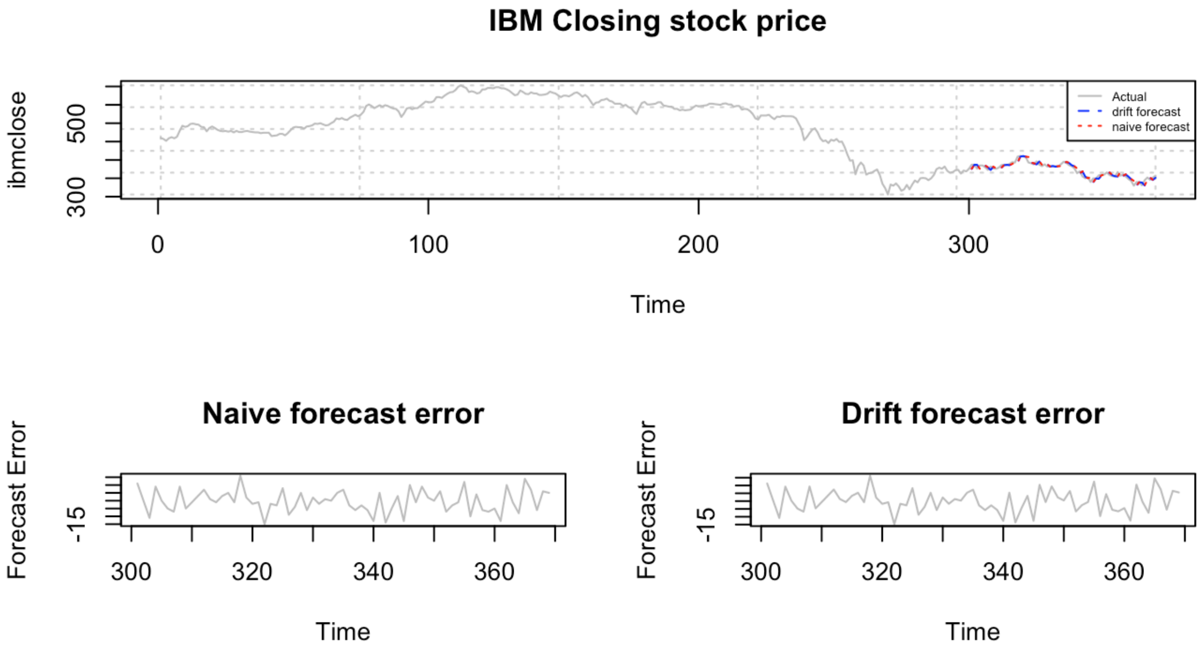
Forecast Accuracy

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Drift	0.00	7.30	5.13	-0.03	1.12	1.01	0.14
Naive	-0.28	7.30	5.10	-0.08	1.12	1.00	0.14
Mean	0.00	73.62	58.72	-2.64	13.03	11.52	0.99



Residual Analysis

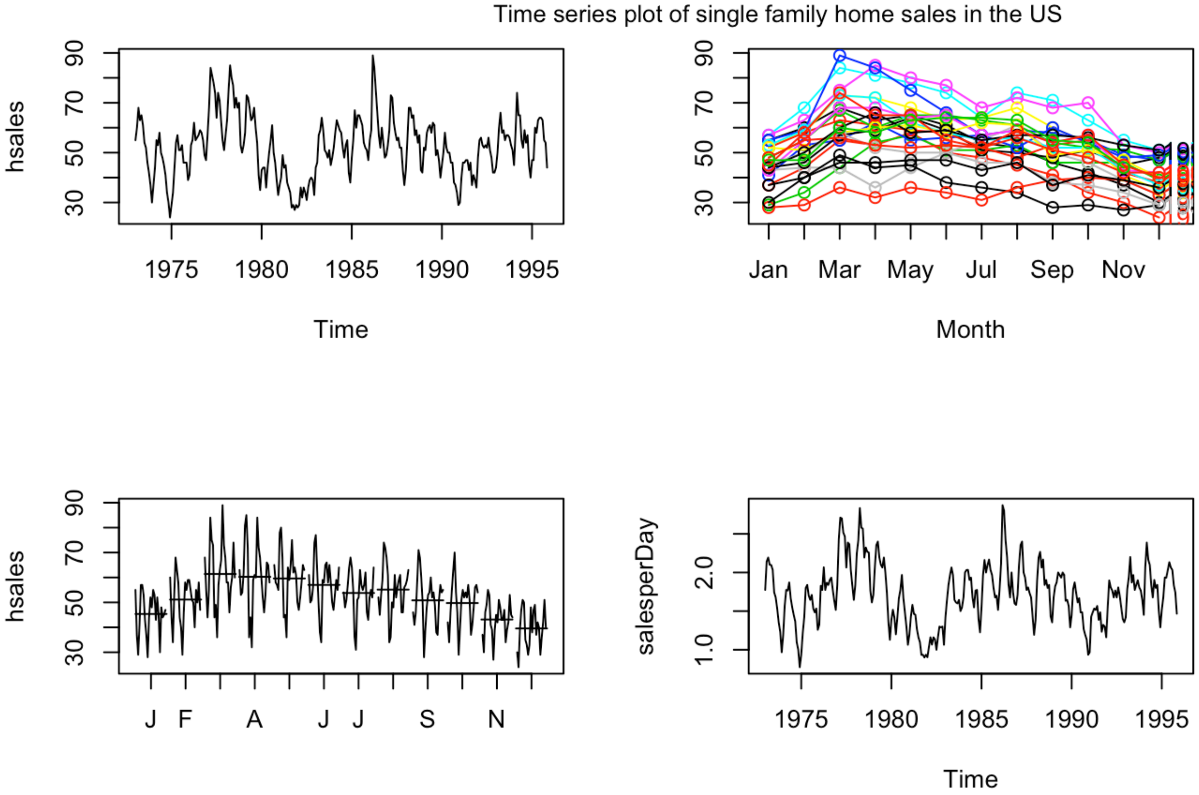
Figure 9 shows the actuals and the forecasts along with the time series of the forecast errors. Both the naive and drift methods are very identical.



IBM stock price forecast

2.4 Single family house sales in the US

Figure 10 shows the time series plot of home sales in the US. The top right panel shows that there is seasonality in the time series. The bottom left panel shows the variation of sales by month over years. The data is attempted to be smoothed by plotting the sales per day over years.



Time series plot of single family home sales in the US

Model training

The time series data is split into training and test set. On the training set the naive, seasonal naive and the drift methods are fit. The drift and Naive models have identical error statistics.

Accuracy statistics on training data

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Naive	-0.05	6.36	5.05	-0.88	10.04	0.59	0.18
SeasonalNaive	-0.11	10.73	8.60	-2.64	17.96	1.00	0.84
Drift	0.00	6.36	5.05	-0.77	10.03	0.59	0.18

The same methods are fit on the sales normalized by the number of days in a month i.e Sales per day. The accuracy of the training set is seen in the table below. Again the Naive and the drift methods are identical.

Accuracy statistics on Sales per data- training data

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1

Naive	0	0.21	0.17	-0.91	10.09	0.59	0.25
Seasonal	0	0.35	0.28	-2.65	17.93	1.00	0.85
Drift	0	0.21	0.17	-0.80	10.08	0.59	0.25

The models are fit on a tranformed data set. While transformed data points of values zero make the MAPE; infinite. The MAE and other statistics convey the same message of drift and Naive being equal.

Accuracy on Transformed training data

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Naive	0	0.13	0.10	-Inf	Inf	0.57	0.23
Seasonal	0	0.23	0.18	-Inf	Inf	1.00	0.86
Drift	0	0.13	0.10	-Inf	Inf	0.57	0.23

Test errors

The test errors of Naive is very subtly better than the drift method as seen from the below tables. Hence Naive method is chosen.

Forecast error on test data

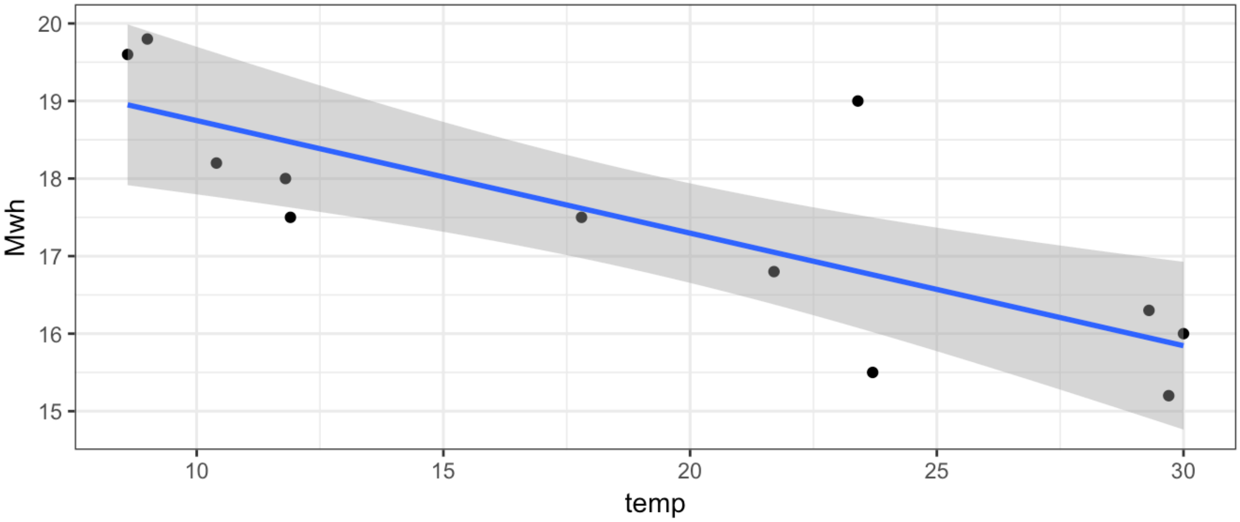
	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Naive	-0.17	8.97	6.61	-1.90	12.06	0.39	1.17
Drift	-0.20	9.01	6.65	-1.96	12.14	0.39	1.17

Forecast error on test data - Sales per day per month

	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Naive	0.4	0.48	0.41	19.71	20.67	0.5	1.83
Drift	0.4	0.48	0.41	19.79	20.73	0.5	1.84

Section 4.10 Question 4.1

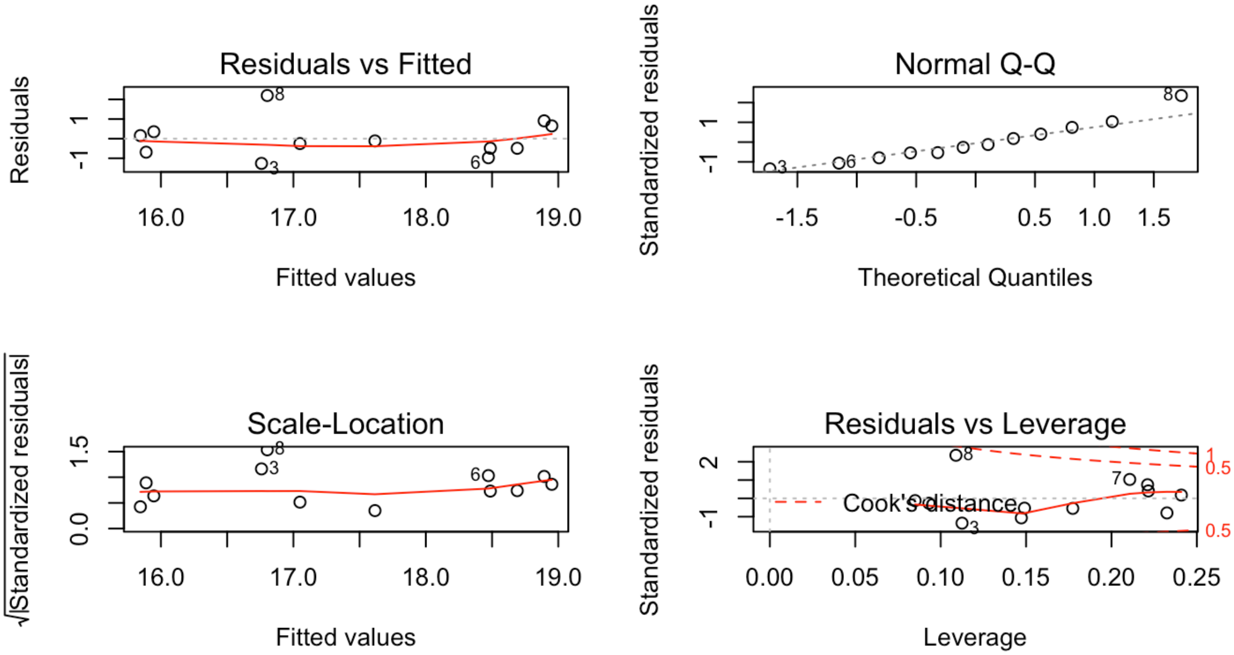
a. Figure 11 shows rthe relationship between temperature and electricity consumption. The relationship is negative probabaly because the usage is for heating purposes.



Electricity consumption by temperature

b. The model though is statistically significant has an R-Squared of only 59%. Figure 12 shows the residual plots. The residuals are fairly random. Observation 8 has the maximum residual error.

```
##
## Call:
## lm(formula = Mwh ~ temp, data = econsumption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2593 -0.5395 -0.1827  0.4274  2.1972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.19952    0.73040   27.66 8.86e-11 ***
## temp        -0.14516    0.03549   -4.09  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9888 on 10 degrees of freedom
## Multiple R-squared:  0.6258, Adjusted R-squared:  0.5884
## F-statistic: 16.73 on 1 and 10 DF,  p-value: 0.00218
```



Residual plots

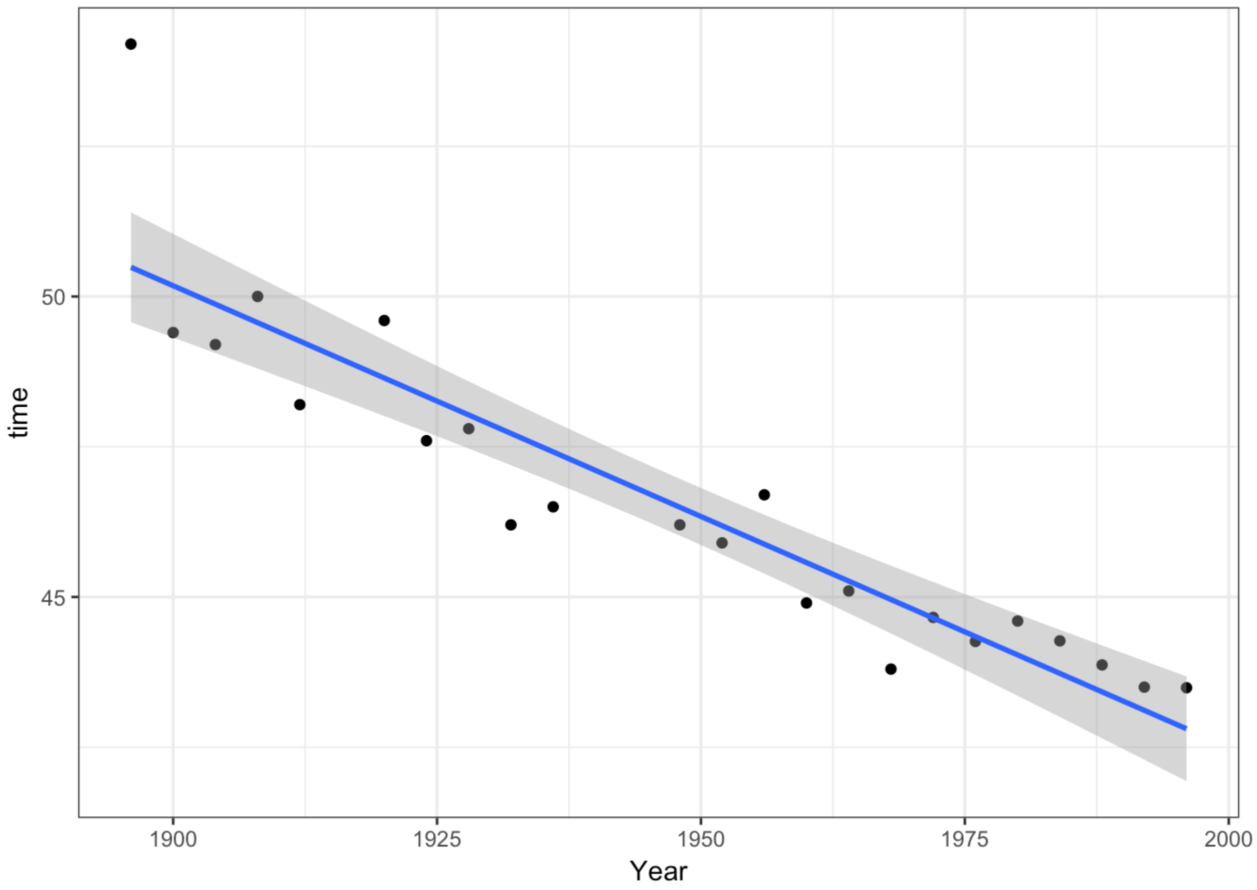
c & d) The forecast for consumption for 10 and 35 degree days and their prediction interval is provided below. The model only explain 59% of the variation in consumption. The predictions need to be used carefully.

```
##   Point Forecast     Lo 80    Hi 80    Lo 95    Hi 95
## 1       18.74795 17.27010 20.22579 16.34824 21.14766
## 2       15.11902 13.50469 16.73335 12.49768 17.74035
```

\pagebreak

Section 4.10 Question 4.2

a & b) Figure 13 shows the winning times of 400m dash in olympic finals between 1900 to 2000. There is a decreasing trend over years.



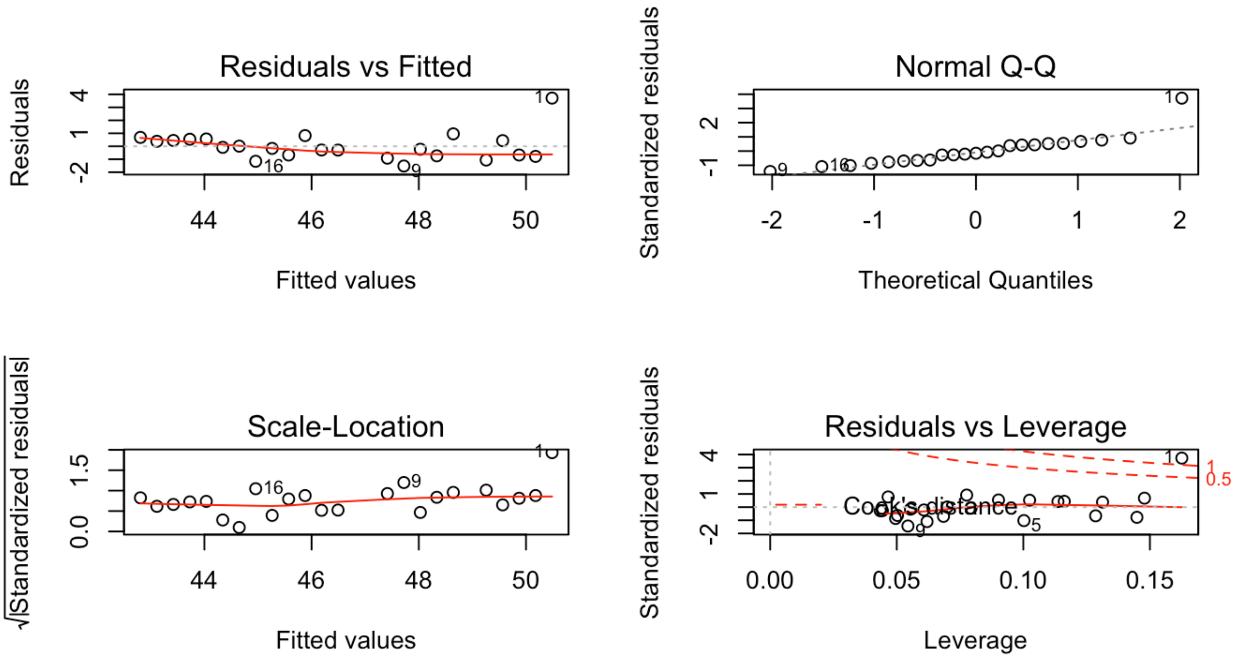
Winning times of 400 meters dash

c. The decrease is at the rate of 7.6%

d. The residual point shows there is one influential point - observation 1. This point can lead to bias in the model fit. However otherwise the model has a 83% R-Squared. e and f) The table below shows the prediction and the actual times. The assumption made in using this model is :

- There is no human limitation for running;
- the location of the race does not contribute to performance of the athletes.
- The residuals follow a normal distribution.

```
##
## Call:
## lm(formula = time ~ Year, data = olympic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5215 -0.7037 -0.1642  0.4952  3.7141
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 196.079876  14.177031   13.83 5.09e-12 ***
## Year         -0.076790   0.007278  -10.55 7.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.089 on 21 degrees of freedom
## Multiple R-squared:  0.8413, Adjusted R-squared:  0.8337
## F-statistic: 111.3 on 1 and 21 DF,  p-value: 7.515e-10
```



Residual Analysis

Prediction of run times

Year	Prediction	Actuals
2000	42.49977	43.84
2004	42.19261	44.00
2008	41.88545	43.75
2012	41.57829	43.94

Section 4.10 Question 3

Elasticity is defined as $(dy/dx)(x/y)$. The Regression model is as follows:

$$logy = \beta_0 + log\beta_1 x + \varepsilon$$

$$y = e^{\beta_0 + log\beta_1 x + \varepsilon}$$

Differentiating y by x, we get …

$$dy/dx = (e^{\beta_0 + log\beta_1 x + \varepsilon})\frac{\beta_1}{x}$$

$$(\frac{dy}{dx})(\frac{x}{y}) = (e^{\beta_0 + log\beta_1 x + \varepsilon})(\frac{\beta_1}{x})(\frac{x}{y})$$

$$(\frac{dy}{dx})(\frac{x}{y}) = (y)(\frac{\beta_1}{x})(\frac{x}{y})$$

$$(\frac{dy}{dx})(\frac{x}{y}) = \beta_1$$

Section 6.7 Question 6.1

Prove 3 X 5 MA = weighted 7 MA with (0.067,0.133,0.2,0.2,0.2,0.133,0.067)

The weights can be converted into fraction (1/15, 2/15, 1/5, 1/5,1/5, 2/15, 1/15)

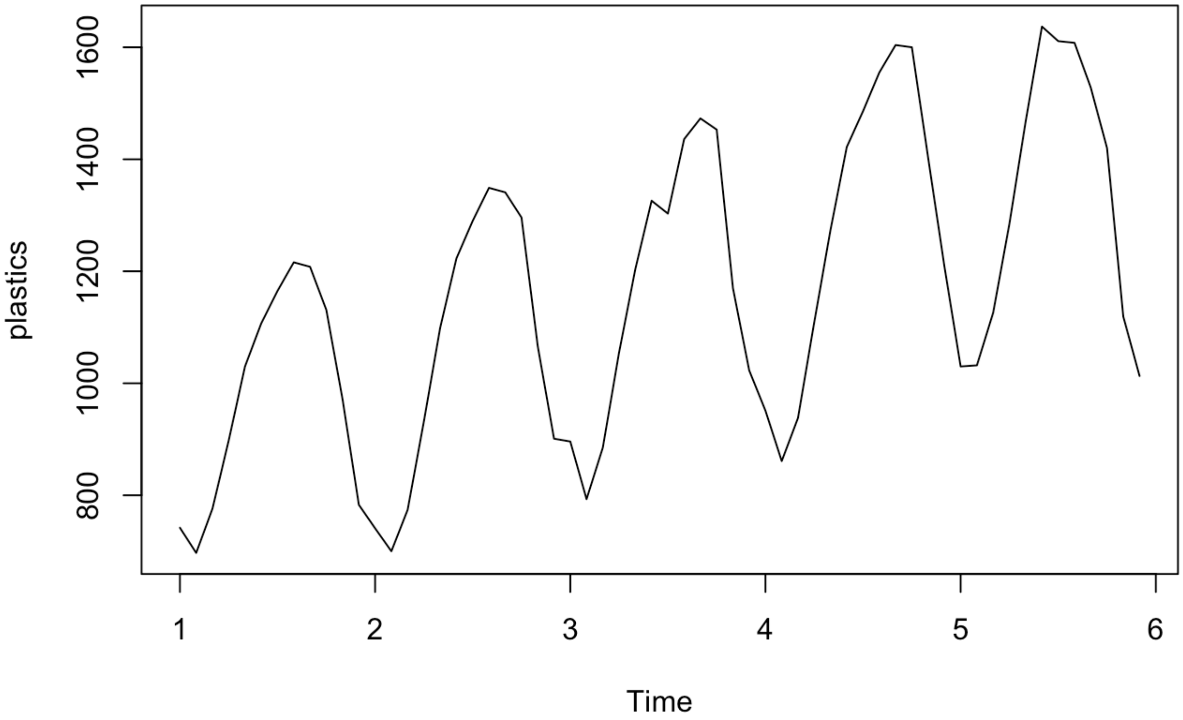
$$3X5MA = \frac{1}{15}[y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1}] + \frac{1}{15}[y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}] + \frac{1}{15}[y_{t-1} + y_t + y_{t+1} + y_{t+2} + y_{t+3}]$$

$$= \frac{1}{15}y_{t-3} + \frac{2}{15}y_{t-2} + \frac{1}{5}y_{t-1} + \frac{1}{5}y_t + \frac{1}{5}y_{t+1} + \frac{2}{15}y_{t+2} + \frac{1}{15}y_{t+3}$$

The above equation is identical to that of 7 MA with weights (1/15, 2/15, 1/5, 1/5,1/5, 2/15, 1/15)

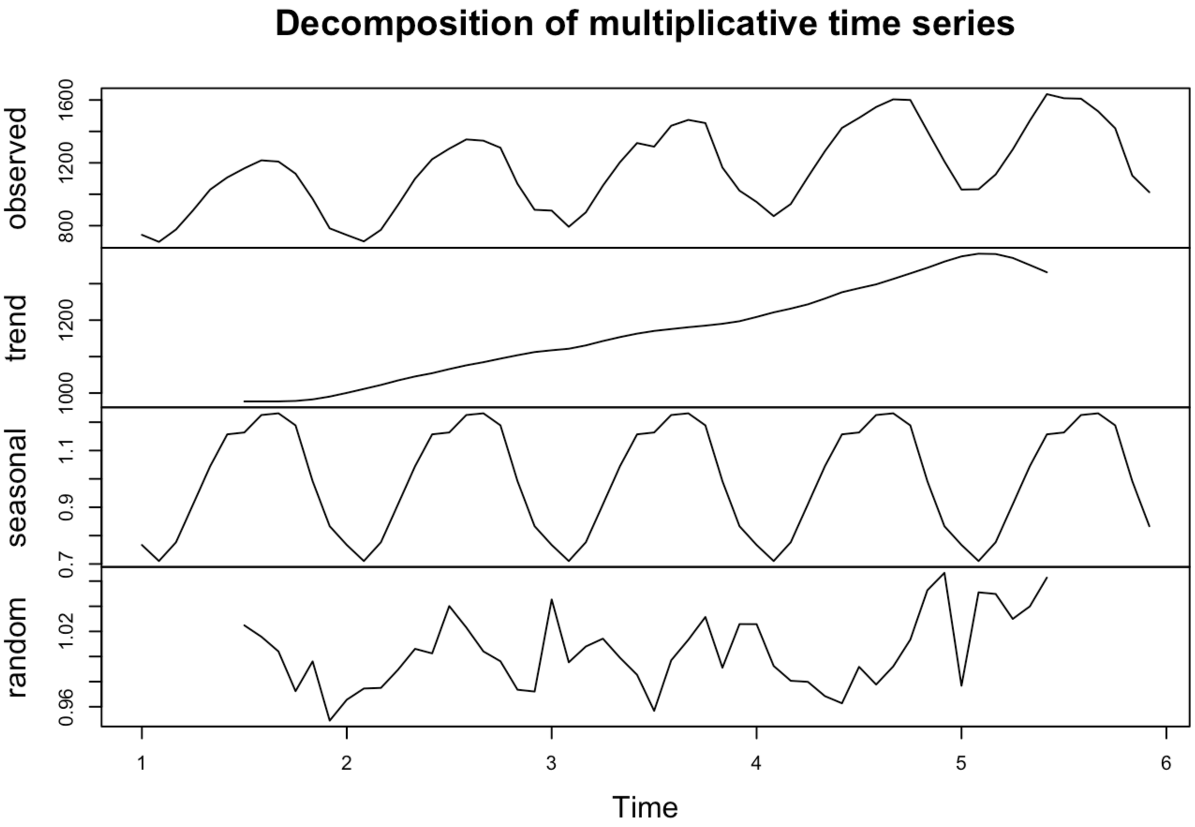
Section 6.7 Question 6.2

a. Figure 15 shows the time series of sales of manufacturer A. It has a seasonality of period 1 year and an increasing trend.



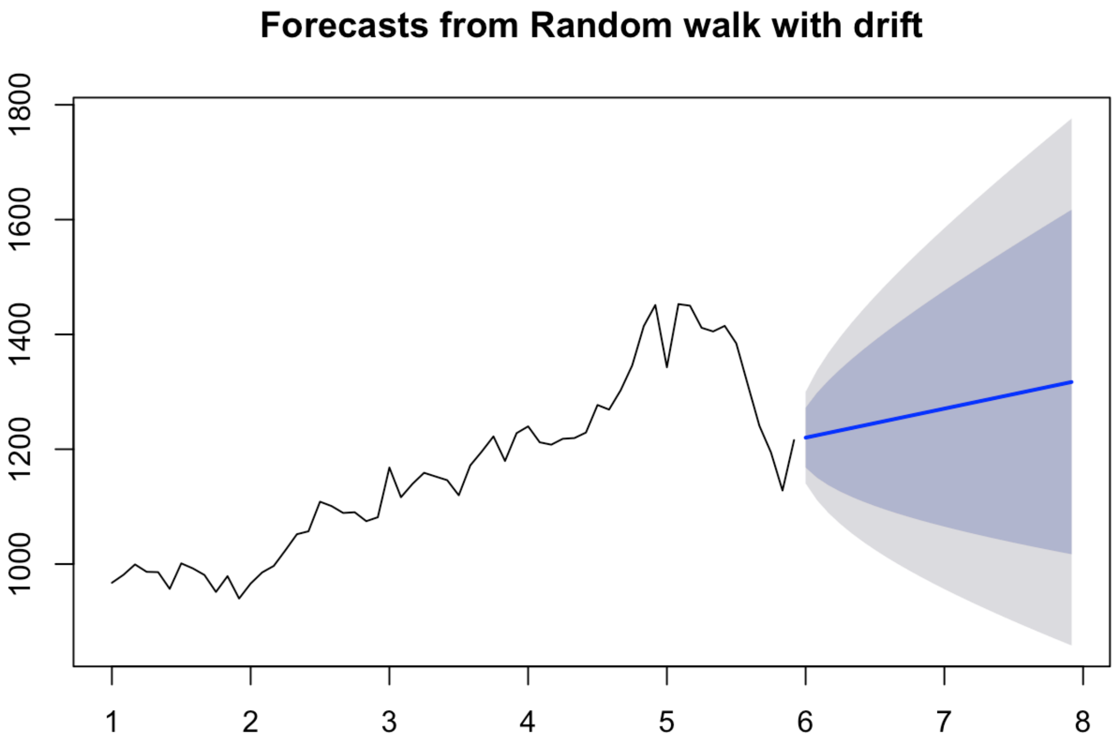
Time series plot of sales

b & c) The classical multiplicative decomposing supports the interpretation of figure 15.



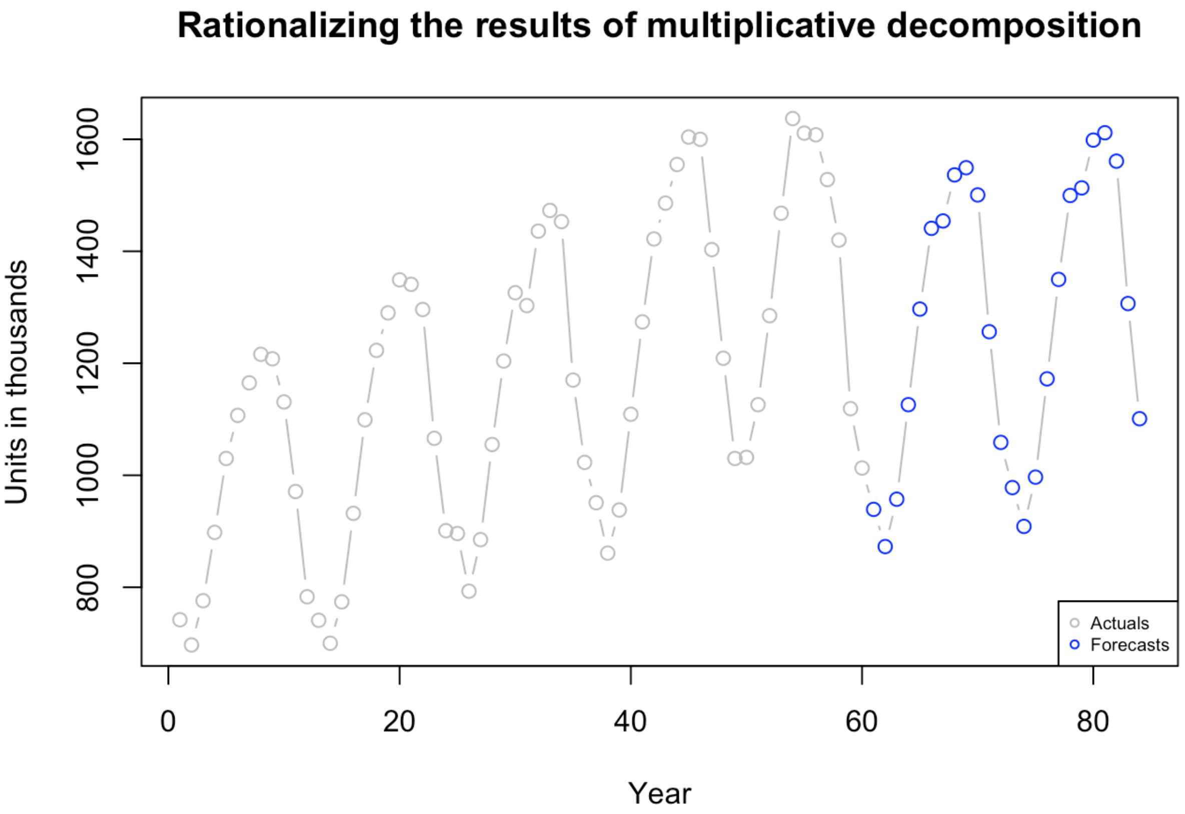
Classical multiplicative decomposing

d,g and h) Figure 16 shows the time series plot of the seasonally adjusted data.



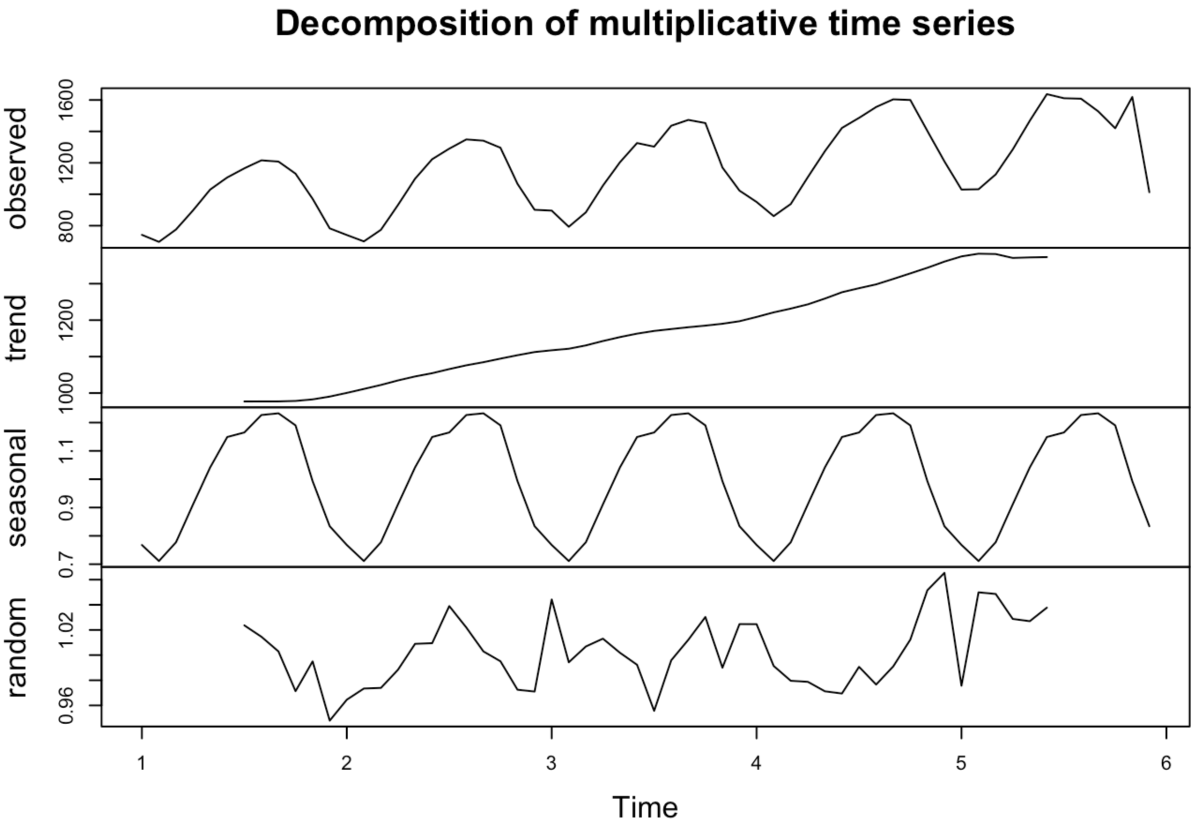
Seasonally Adj Data

The drift method forecasts from the seasonally adjusted data is multiplied with the seasonal components and the mean of the random error to produce the forecasts for the next 2 years. The plot of the data is shownn figure 17.

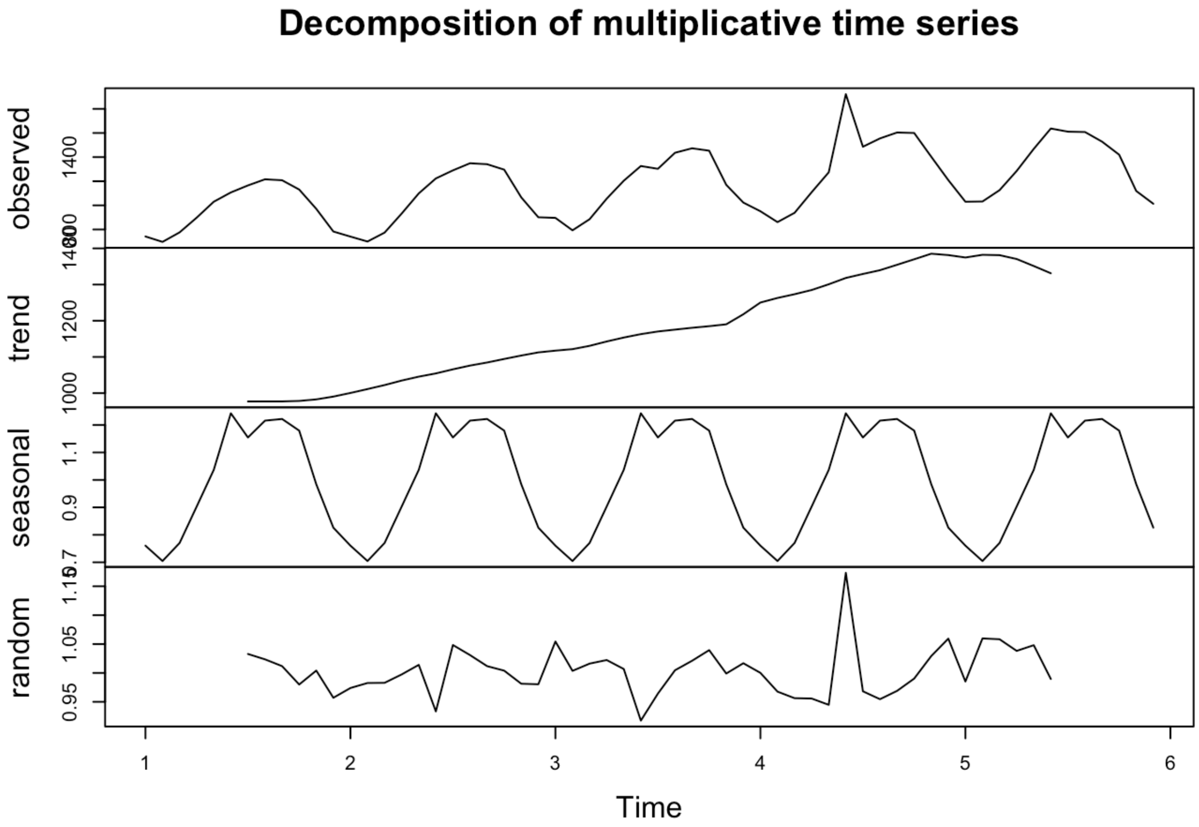


Rationalizing results

f. Figure 18 shows the decomposition when 500 units is added to Nov of year 5. The effect is none, this is because of the classical decomposition that has a disadvantage of losing predictions in the begining and ending of the time series (loss of half year each). Figure 19 Shows the decomposition when 500 units is added to June of year 4. This has a slight effect on the predictions and it can be seen in the random error chart as a spike. The moving average smoothes over the spike to minimize its effect.



Added 500 to december of year 5



Added 500 to Jun of Year 4

Section 6.7 Question 6.3

a. The trend component is the major source of variation in the data. There is a steady increase (constant rate) for the most part from 1978 to 1995. The seasonal component and the random error each take up only about 8% (16% in total) of the varaition in the data. The month of November and Feb has not changed year over year. The month of March has seen much volatility. Likely because of the recession in 91’ & 92’.

b. the recession is slightly visible in the trend, but not much. However the random error spikes in the negative direction shows that the model did not pick up the recession.