Assignment 1 - Problem 2

Shivakanth Thudi

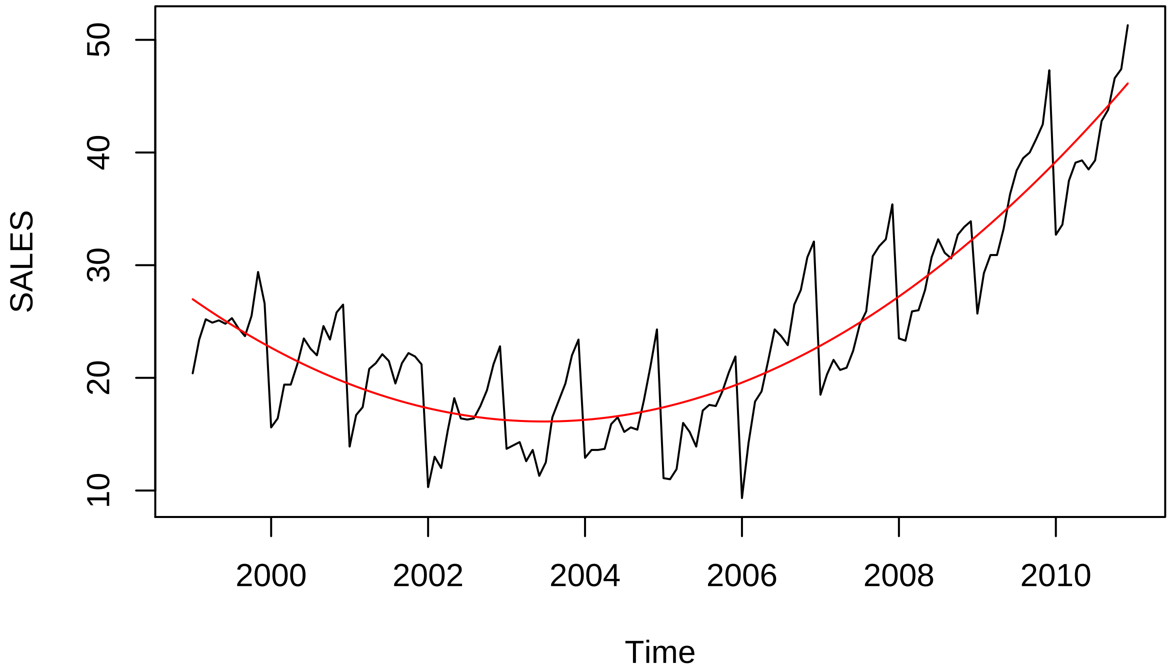
10/30/2016

Part a - Modeling the trend component

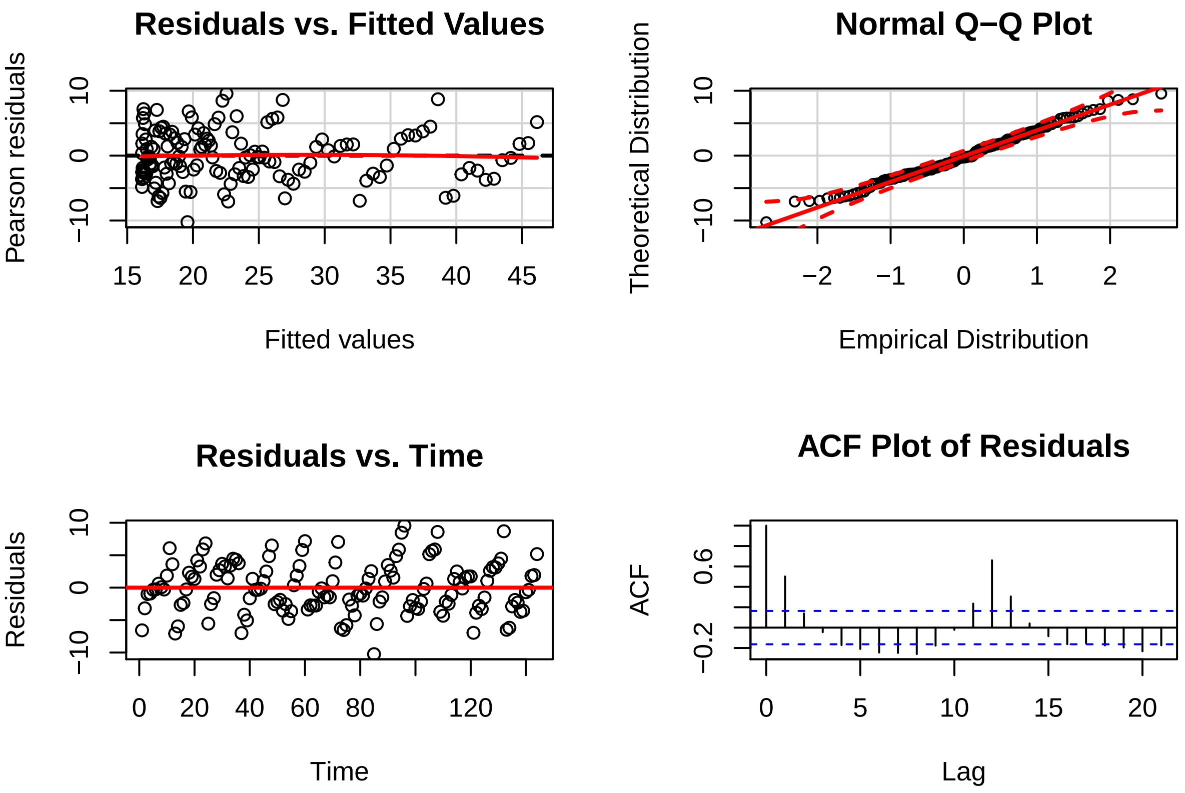
```
plot(sales)
t <- time(sales)  # Extracting time as the explanatory variate
month <- as.factor(cycle(sales))  # Introducing month as the season

# Model the trend only, as a quadratic
t2 <- t^2
reg1 <- lm(sales ~ t + t2)
summary(reg1)
```

```

Call:
lm(formula = sales ~ t + t2)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.2493  -2.7326  -0.2823   2.6100   9.5576 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.175e+06  1.211e+05   17.95   <2e-16 ***
t           -2.171e+03  1.208e+02  -17.97   <2e-16 ***
t2           5.419e-01  3.014e-02   17.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.881 on 141 degrees of freedom
Multiple R-squared:  0.8146,    Adjusted R-squared:  0.812 
F-statistic: 309.8 on 2 and 141 DF,  p-value: < 2.2e-16
```

```
plot(sales)
# superimpose the fit of model reg1 on the plot of the data
points(t, predict.lm(reg1), type = "l", col = "red")
```



```r
# Diagnostic plots for reg1 model

# Dividing the plotting page into 4 panels
par(mfrow = c(2, 2))
# Plot of fitted values vs residuals
residualPlot(reg1, main = "Residuals vs. Fitted Values")
# qq-plot of residuals
qqPlot(reg1$residuals, distribution = "norm", main = "Normal Q-Q Plot",
    xlab = "Empirical Distribution", ylab = "Theoretical Distribution")
# Plotting the residuals vs time
plot(reg1$residuals, main = "Residuals vs. Time", ylab = "Residuals",
    xlab = "Time")
abline(h = 0, col = "red", lwd = 2)  # plotting a horizontal line at 0
# Sample acf plot of residuals
acf(reg1$residuals, main = "ACF Plot of Residuals")
```



Model Fit

The residuals look homoscedastic - we also confirm this with a Breusch-Pagan Test. From the qqplot, we see that the residuals are normally distributed, which we confirm with a Shapiro Wilk test.

Plotting the residuals against time, we observe that the error terms seem to follow a pattern and this is verified in the ACF plot where we see some significant lag h correlations. This indicates that the OLS assumption of independence of error terms does not hold, and consequently casts doubt on the validity of predictions and prediction intervals that can be made by this model.

Despite the fact that OLS assumptions were not met, the model has a good fit, with an R^2 value of 0.81.

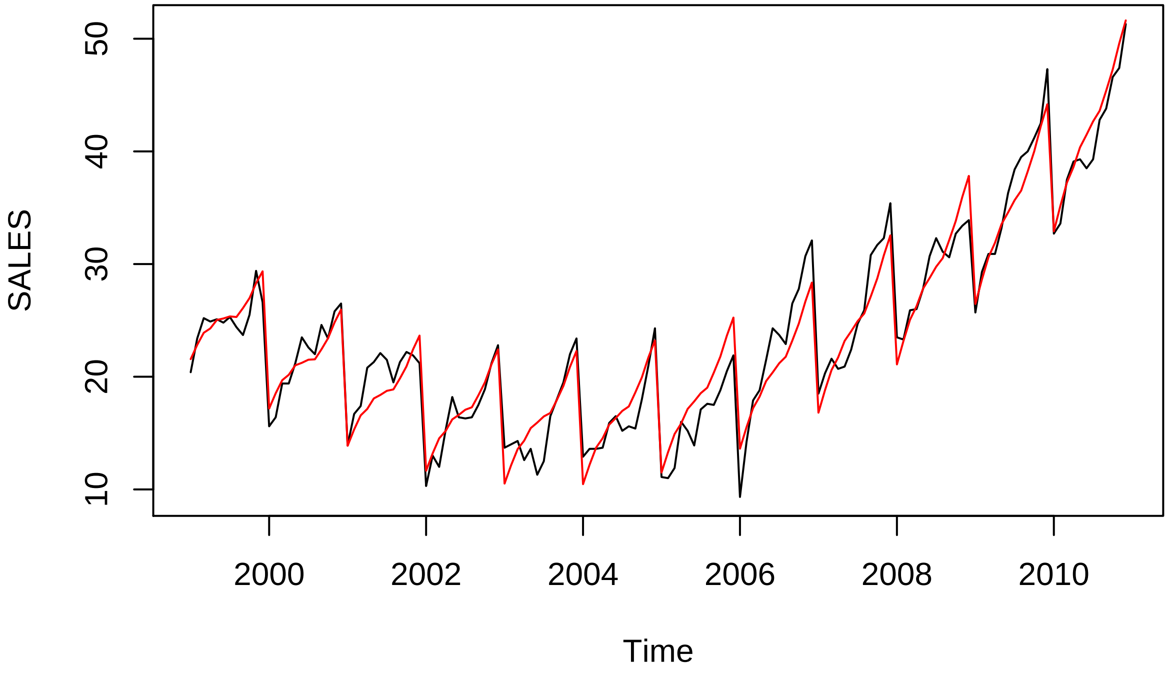
Part b - Classical Decomposition

```r
# Model the trend and seasonal components
reg2 <- lm(sales ~ t + t2 + month)
summary(reg2)
```

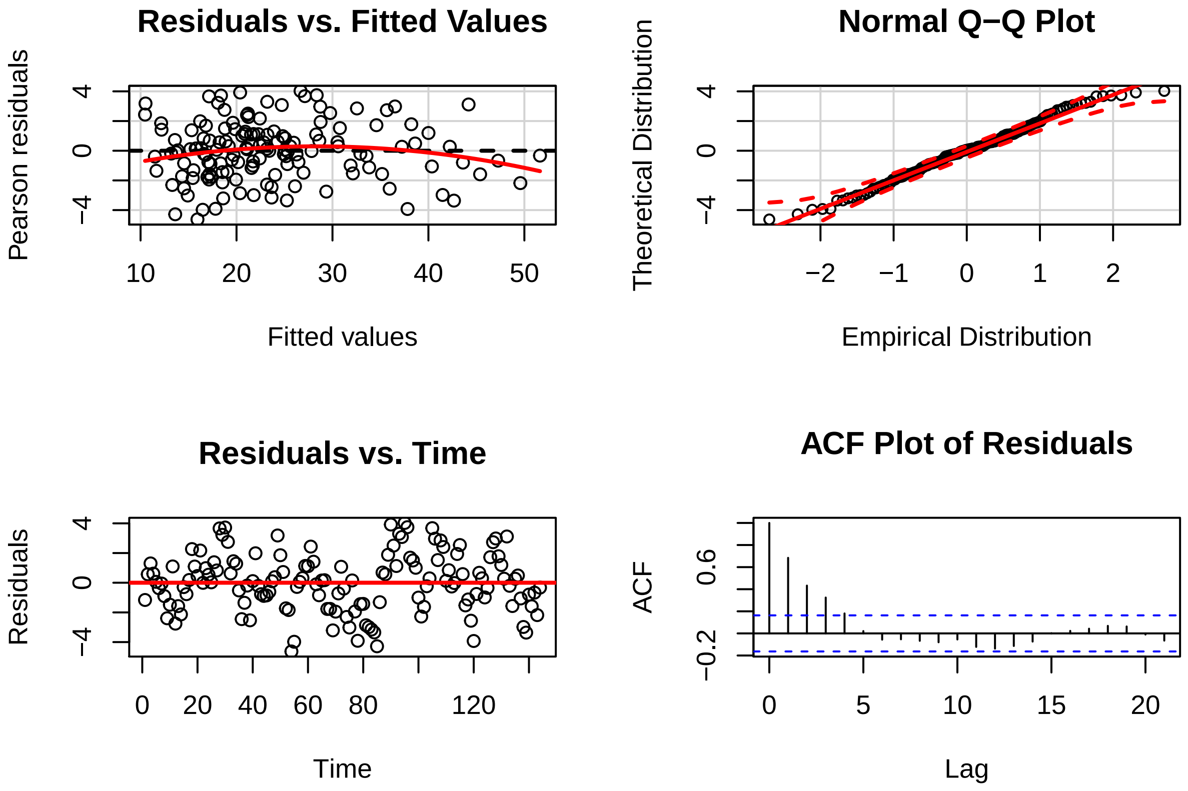
```
Call:
lm(formula = sales ~ t + t2 + month)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6296 -1.3720  0.0598  1.2164  4.0276 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.175e+06  6.356e+04  34.219  < 2e-16 ***
t           -2.171e+03  6.340e+01 -34.243  < 2e-16 ***
t2           5.418e-01  1.581e-02  34.267  < 2e-16 ***
month2       1.674e+00  8.313e-01   2.014 0.046047 *
month3       3.144e+00  8.313e-01   3.782 0.000236 ***
month4       3.922e+00  8.314e-01   4.718 6.06e-06 ***
month5       5.060e+00  8.315e-01   6.086 1.21e-08 ***
month6       5.565e+00  8.316e-01   6.693 5.93e-10 ***
month7       6.121e+00  8.317e-01   7.360 1.86e-11 ***
month8       6.428e+00  8.318e-01   7.728 2.61e-12 ***
month9       7.577e+00  8.319e-01   9.108 1.29e-15 ***
month10      8.811e+00  8.321e-01  10.589  < 2e-16 ***
month11      1.047e+01  8.323e-01  12.580  < 2e-16 ***
month12      1.186e+01  8.325e-01  14.240  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.036 on 130 degrees of freedom
Multiple R-squared:  0.9529,	Adjusted R-squared:  0.9482 
F-statistic: 202.5 on 13 and 130 DF,  p-value: < 2.2e-16
```

```r
plot(sales)
# superimpose the fit of model reg2 on the plot of the data
points(t, predict.lm(reg2), type = "l", col = "red")
```



```r
# Diagnostic plots for reg2 model

# Dividing the plotting page into 4 panels
par(mfrow = c(2, 2))
# Plot of fitted values vs residuals
residualPlot(reg2, main = "Residuals vs. Fitted Values")
# qq-plot of residuals
qqPlot(reg2$residuals, distribution = "norm", main = "Normal Q-Q Plot",
    xlab = "Empirical Distribution", ylab = "Theoretical Distribution")
# Plotting the residuals vs time
plot(reg2$residuals, main = "Residuals vs. Time", ylab = "Residuals",
    xlab = "Time")
abline(h = 0, col = "red", lwd = 2)  # plotting a horizontal line at 0
# Sample acf plot of residuals
acf(reg2$residuals, main = "ACF Plot of Residuals")
```



Model Fit

The model fits much better when we model both the seasonal term and the trend term. The R^2 value has increased from 0.81 to 0.95.

We look at the residual diagnostics and observe that the OLS assumptions of homoscedasticity and normality are met, but the OLS assumption of independence of error terms is not. Consequently, this casts doubt on the model's ability to make predictions and estimate prediction intervals.

Part c - Comparison of models

We see that classical decomposition into a trend and seasonal term has significantly improved the model's fit, with an increase in R^2 from 0.81 to 0.95. Thus, adding a seasonal term greatly improved the model's fit. The ACF plot in (a) indicated that there was seasonality not being captured by the model, but after classical decomposition we see in the ACF plot in (b) that this is no longer the case. However, we see significant lag h spikes at h = 1, 2 and 3, which imply that the model still cannot be used for forecasting.

Part d - OLS Assumptions

Both models satisifed the OLS assumptions of homoscedasticity and normality, but **failed to satisfy the OLS assumption of independence of the error terms**. We include the BP-tests for homoskedasticity and the Shapiro tests for normality for reference.

Model	Breusch-Pagan Test for homoscedasticity	Shapiro-Wilk test for normality
Model (a) - Trend only, quadratic term	p-value = 0.5786578	p-value = 0.55355
Model (b) - Trend and Seasonal, quadratic term and indicator variables	p-value = 0.0038492	p-value = 0.3871075

Since the OLS assumption of independence of error terms was not met, we conclude that predictions and prediction intervals made by these models are not valid.

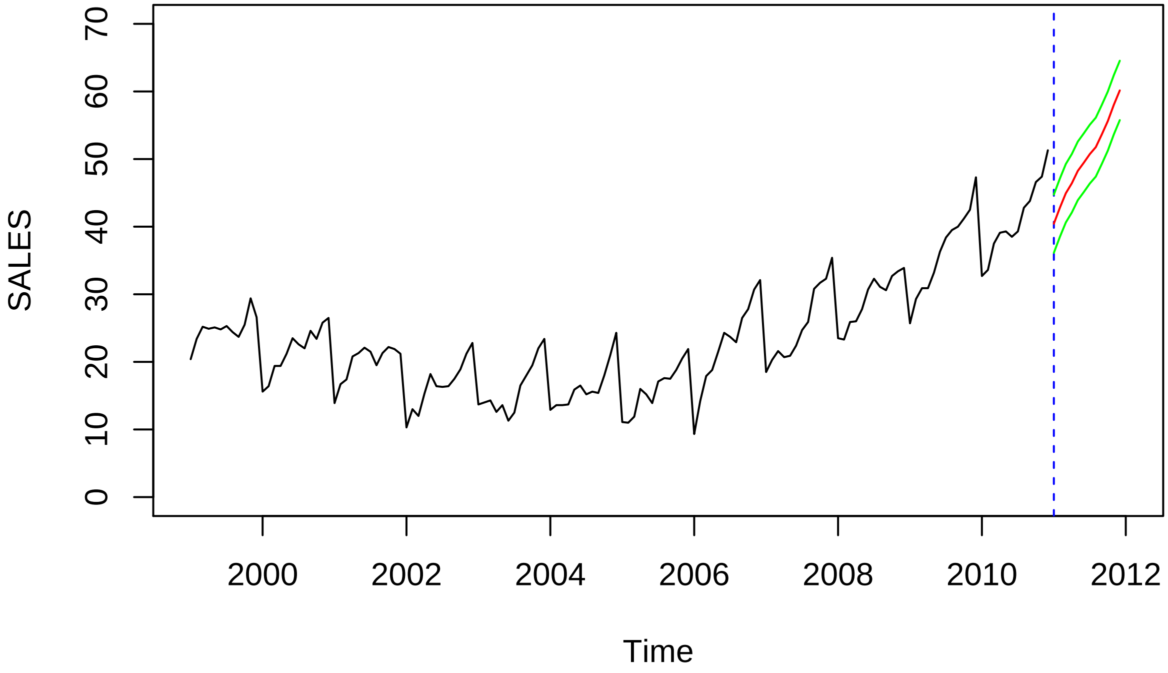
Part e - Predictions

```
# Prediction in Sales data

# Forecasting for Year 2011
t.new <- seq(2011, 2012, length = 13)[1:12]
t2.new <- t.new^2
month.new <- factor(rep(1:12, 1))  # Introducing the seasonal value for forecasting

# Putting the values for forecasting into a dataframe
new <- data.frame(t = t.new, t2 = t2.new, month = month.new)

# Computing the prediction as well as the prediction interval
pred <- predict.lm(reg2, new, interval = "prediction")

pred
```

```
        fit      lwr      upr
1  40.44214 36.13380 44.75048
2  42.79328 38.47881 47.10774
3  44.94692 40.62606 49.26777
4  46.41722 42.08973 50.74471
5  48.25420 43.91981 52.58858
6  49.46617 45.12463 53.80770
7  50.73647 46.38754 55.08541
8  51.76511 47.40853 56.12170
9  53.64375 49.27927 58.00824
10 55.61406 51.24142 59.98670
11 58.01770 53.63667 62.39873
12 60.15467 55.76500 64.54435
```

```
par(mfrow = c(1, 1))
# Plotting the data
plot(sales, xlim = c(1999, 2012), ylim = c(0, 70))
```

```r
# Adding a vertical line at the point where prediction starts
abline(v = 2011, col = "blue", lty = 2)
# Plotting the predictions
lines(pred[, 1] ~ t.new, type = "l", col = "red")
# Plotting lower limit of the prediction interval
lines(pred[, 2] ~ t.new, col = "green")
# Plotting upper limit of the prediction interval
lines(pred[, 3] ~ t.new, col = "green")
```


SALES
Time