Setup:

Traditionally, the quality of the Bordeaux vintage (wine) is first evaluated by experts in March of the next year. However, it turns out that these first ratings are rather unreliable. What about using a statistical approach that predicts the price of the vintage as a function of its age, rainfall, and temperature?

Consider data on average temperature during the growing season (April through September, in degrees centigrade), rain during the harvest season (rain in August and September, in total millimeters), rain prior to the growing season (October of the previous year to March of the current year, in total millimeters), age of the vintage, and average price for the Bordeaux vintage relative to the year 1961 (the response variable). Data for the years 1952 through 1980 are provided (the data for 1954 and 1956 are missing; relative prices for these two vintages could not be established because they were poor vintages and very little wine is now sold from those two years).

Problem:

Construct a regression model that describes the price of the vintage as a function of its age, rainfall, previous rainfall, and temperature. Interpret the data through appropriate scatter plots (such as plots of price against the various explanatory variables), summary statistics, and the output from the regression analysis. Check the adequacy of the model by looking at residual plots.

Dataset Column Descriptions:

- Year
- Temp: Temperature (in degrees C)
- Rain: (rain during harvest season, in total millimeters)
- PrevRain: (rain prior to the growing season, in total millimeters)
- Age: (age of the vintage, where 1983 = 0)
- Price: (average price relative to the year 1961) – response variable

The contents of the Bordeaux Vintage Wine dataset are first imported as a data frame in R using the function read.csv() and the structure of the dataset is obtained using str() as shown below. Also, a summary with various statistical parameters of all the variables in the dataset is shown below using summary() command. We see that the mean is usually greater than the median for all variables.

```
## Import Bordeaux Vintage Data
  dataBV <- read.csv("Proj3_Dataset.csv", header = TRUE)
  str(dataBV)
  summary(dataBV)

## Attach dataset
  attach(dataBV)
```

```
> str(dataBV)
'data.frame':   27 obs. of  6 variables:
 $ Year    : int  1952 1953 1955 1957 1958 1959 1960 1961 1962 1963 ...
 $ Temp    : num  17.1 16.7 17.1 16.1 16.4 ...
 $ Rain    : int  160 80 130 110 187 187 290 38 52 155 ...
 $ PrevRain: int  600 690 502 420 582 485 763 830 697 608 ...
 $ Age     : int  31 30 28 26 25 24 23 22 21 20 ...
 $ Price   : num  0.368 0.635 0.446 0.221 0.18 0.658 0.139 1 0.331 0.168 ...
```

```
> summary(dataBV)
      Year          Temp            Rain         PrevRain        Age             Price       
 Min.   :1952   Min.   :14.98   Min.   : 38.0   Min.   :376.0   Min.   : 3.00   Min.   :0.1010  
 1st Qu.:1960   1st Qu.:16.15   1st Qu.: 88.0   1st Qu.:543.5   1st Qu.: 9.50   1st Qu.:0.1375  
 Median :1967   Median :16.42   Median :123.0   Median :600.0   Median :16.00   Median :0.2210  
 Mean   :1967   Mean   :16.48   Mean   :144.8   Mean   :608.4   Mean   :16.19   Mean   :0.2877  
 3rd Qu.:1974   3rd Qu.:17.01   3rd Qu.:185.5   3rd Qu.:705.5   3rd Qu.:22.50   3rd Qu.:0.3495  
 Max.   :1980   Max.   :17.65   Max.   :292.0   Max.   :830.0   Max.   :31.00   Max.   :1.0000  
```

We can also look at the spread of the predictor variables from the boxplots as shown in Figure 1. We see that all the variables have somewhat a regular distribution, although not symmetric. Also, none of these variables have any outliers. These observations are supported by the summary statistics shown above.

```
## Data Description
  ## Price = average price relative to 1961, y
  ## Temp = temperature(in °C), x1
  ## Rain = rain during harvest season (in mm), x2
  ## PrevRain = rain pior to growing season (in mm), x3
  ## Age = age of vintage, x4

## Boxplot of predictor variables
  par(mfrow = c(1,4))
  boxplot(Temp, main="Temperature (in °C)")
  boxplot(Rain, main="Rain (in mm)")
  boxplot(PrevRain, main="PrevRain (in mm)")
  boxplot(Age, main="Age (in years)")
```

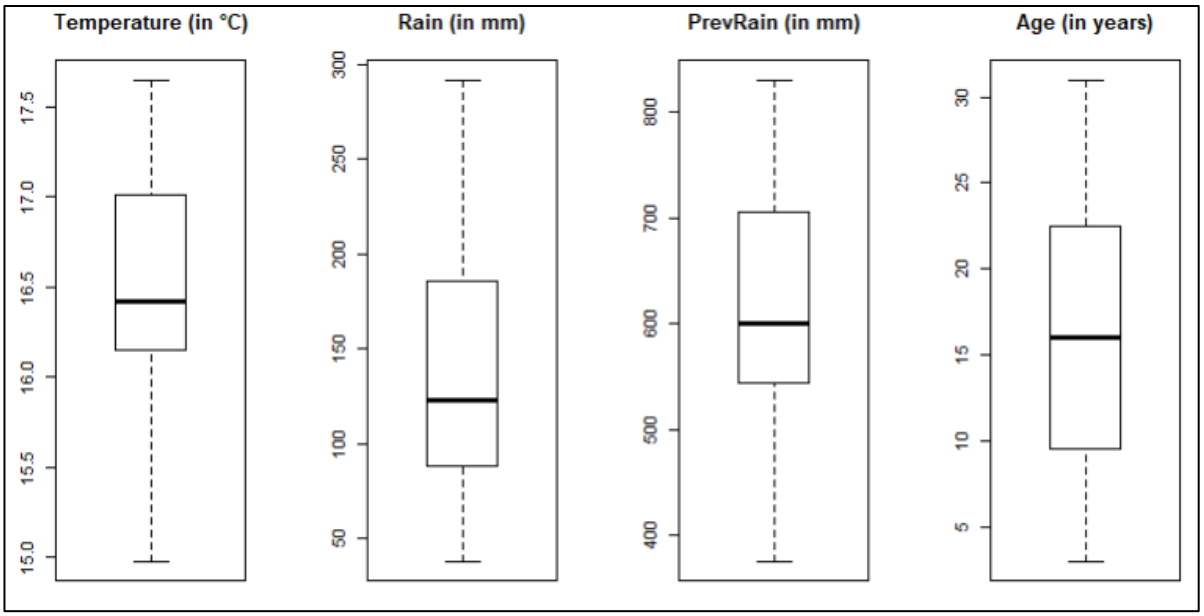

Figure 1 Boxplot for each of the Predictor Variables

Next, the scatter plot of the response variable, *Price*, versus all the predictor variables, *Temp*, *Rain*, *PrevRain* and *Age* is shown below. As seen in Figure 2, the price of vintage generally shows an increasing trend with respect to the temperature during growing season and the age of vintage. Also, a decreasing trend in price of the vintage is seen with respect to rain during the harvest season. Further, there is no significant trend visible between the price and the rain prior to growing season.

We also see that the variance in price either increases or decreases as we move along the x-axis in these plots. Thus, this suggests that we might have a problem of heteroskedasticity in the data. This problem can be corrected by using suitable transformation on the response variable or the predictor variables, which we will explore later.

```
## Plot Relationship between Price and predictor variables
  par(mfrow = c(2,2))
  plot(Price~Temp, xlab = "Temperature (in °C)", ylab = "Price (relative to 1961)")
  plot(Price~Rain, xlab = "Rain (in mm)", ylab = "Price (relative to 1961)")
  plot(Price~PrevRain, xlab = "PrevRain (in mm)", ylab = "Price (relative to 1961)")
  plot(Price~Age, xlab = "Age (in years)", ylab = "Price (relative to 1961)")
```

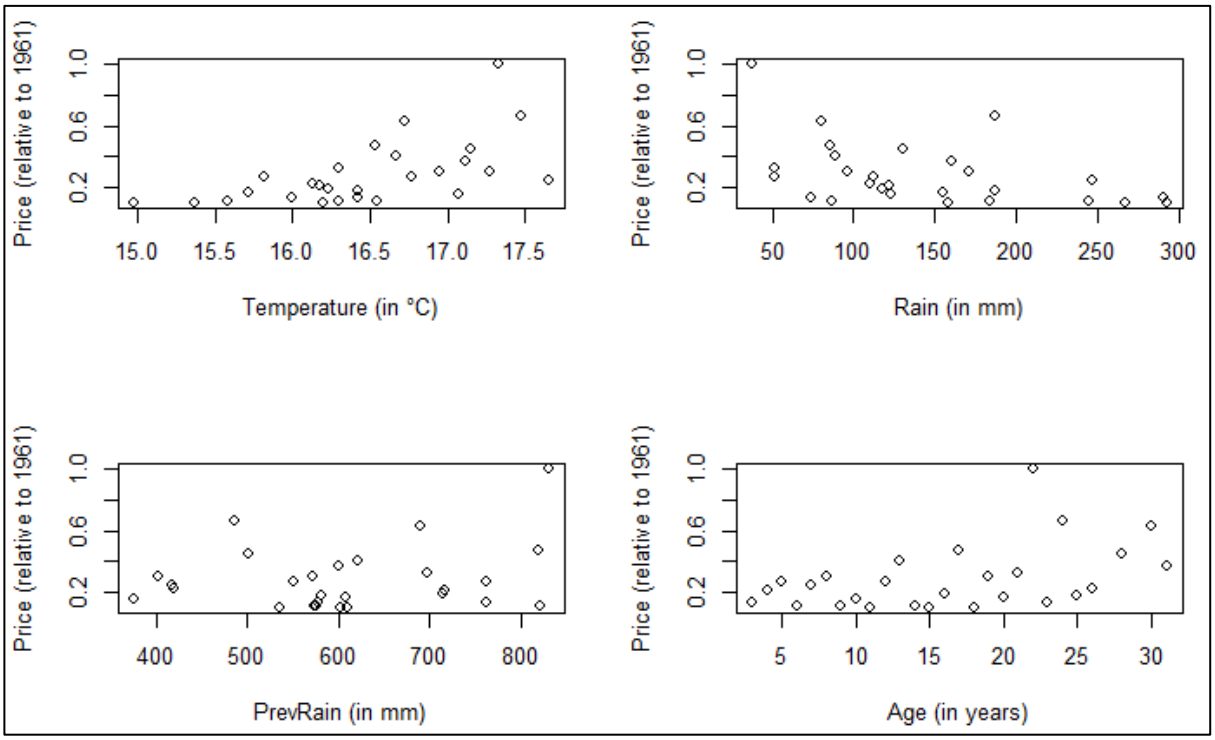

Figure 2 Scatter Plot of Price v/s all Predictor Variables

Before fitting the regression model on the response variable, *Price*, let us also investigate how the predictor variables, *Temp*, *Rain*, *PrevRain* and *Age* are related to one another. A correlation matrix can be obtained using the command cor() in R as shown below. We see that the coefficient of correlation between each pair of the predictor variables is not close to either -1 or +1. This suggest that the variables are not highly correlated and thus the wine dataset doesn't suffer from multicollinearity problem.

We can also do this graphically by constructing scatter plots of all pair-wise combinations of predictor variables in the data frame. The plot is shown in Figure 3 where we see that there is no significant pattern seen in any of the pair-wise plots and thus there is no multicollinearity.

```
## Correlation Matrix of Predictor variables
  dataBV2 <- data.frame(Temp, Rain, PrevRain, Age)
  cor(dataBV2)
## Scatter plot of predictor variables
  plot(dataBV2)
```

```
> cor(dataBV2)
                Temp        Rain    PrevRain         Age
Temp       1.00000000 -0.02614913 -0.32192500  0.29437203
Rain      -0.02614913  1.00000000 -0.26798907  0.05884976
PrevRain -0.32192500 -0.26798907  1.00000000 -0.05118354
Age        0.29437203  0.05884976 -0.05118354  1.00000000
```

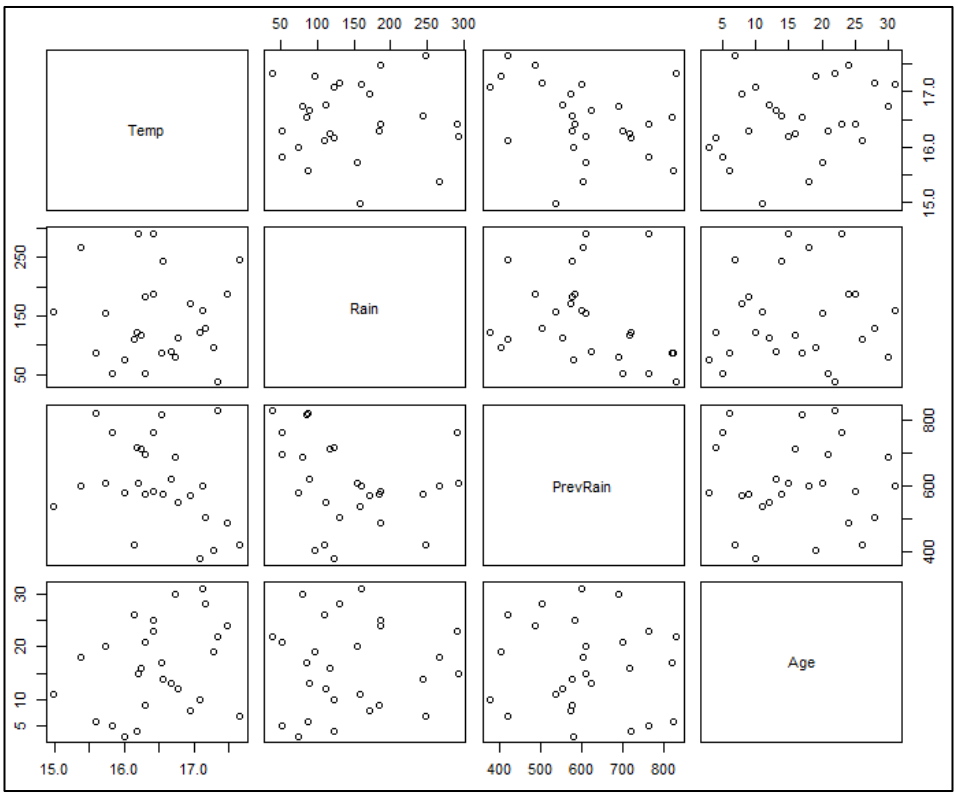

Figure 3 Pair-wise Scatter Plot of Predictor Variables

Let us now develop a regression model for the response variable, *Price*, using the predictor variables, *Temp*, *Rain*, *PrevRain* and *Age*. Let *Y* represent the response variable, *Price*, and let X_1, X_2, X_3 and X_4 represent the predictor variables, *Temp*, *Rain*, *PrevRain* and *Age*, respectively. Thus, a multiple linear regression that may be suitable to predict *Y* can be represented as,

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \varepsilon$$

where, ε is a random error term, and β_0, β_1, β_2, β_3 and β_4 are model parameters that we need to estimate. To estimate these parameters we use a linear model function lm() in R that fits a multiple linear regression model, and to obtain a summary we use the command summary() as shown below.

```
## Run Regression Model
  lm1 <- lm(Price~Temp + Rain + PrevRain + Age)
  summary(lm1)
```

```
> summary(lm1)

Call:
lm(formula = Price ~ Temp + Rain + PrevRain + Age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14072 -0.08770 -0.01074  0.03410  0.26783

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.1716289  0.6928899  -4.577 0.000147 ***
Temp         0.1903096  0.0390606   4.872 7.18e-05 ***
Rain        -0.0010351  0.0003314  -3.123 0.004947 **
PrevRain     0.0005638  0.0001979   2.849 0.009338 **
Age          0.0080519  0.0029410   2.738 0.012013 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1176 on 22 degrees of freedom
Multiple R-squared:  0.7356,    Adjusted R-squared:  0.6875
F-statistic:  15.3 on 4 and 22 DF,  p-value: 4.017e-06
```

Thus, the resulting multiple regression model is,

$$Y = -3.172 + 0.190 \cdot X_1 - 0.001 \cdot X_2 + 0.001 \cdot X_3 + 0.008 \cdot X_4$$

The estimated coefficients of the model indicate the direction of the relationship between the price of vintage and each of the corresponding predictors. The coefficients of temperature and rain prior to the growing season are positive i.e. higher the temperature and the rain prior to the growing season, higher is the price of the vintage wine. Next, the coefficient of harvest rain is negative i.e. rain during harvest season decreases the price of the wine. Further, the positive coefficient for age reflects the fact that the older wines are more expensive.

From the above summary output, we also observe that R^2 value achieved is 0.7356 i.e. about 73.56% of the variation in *Price* is explained by a multivariate regression on the explanatory variables, *Temp*, *Rain*, *PrevRain* and *Age*.

To test the importance (significance) of each predictor variables in developing the above regression model, we can perform a partial t-test on the model parameters. This can be done by testing the null hypothesis $H_0 : \beta_i = 0$ against an alternative hypothesis $H_1 : \beta_i \neq 0$ for $i = 1, 2, 3, 4$. If the *p*-value from this test is lower than the significance level, α, then we can reject the null hypothesis and conclude that the predictor variable corresponding to the parameter β_i is important in explaining the response variable, *Y*. Subsequently, we can obtain the *p*-value for each of the predictor variables from the above summary output. We see that all the predictor variables have a very small *p*-value and thus we can conclude that they contribute significantly to the regression at a significance level of 0.05.

Let us look at various residual plots to check the adequacy of the model. The residual versus fitted value plot in Figure 4 shows a funnel effect and shows that the large residuals are all positive. This indicates that the residuals have unequal variances. The curve in the normal plot of the residuals shows that the residuals are right skewed. The Shapiro-Wilk Normality Test

results in a small *p*-value, thereby, also indicating non-normality of residuals as shown below. Further, in Figure 5, we see that the residuals vs predictor variables plot shows reasonably a good scatter of residuals except in the residuals vs *Temp* plot. This suggests that a linear model on *Price* is suitable, however, polynomial terms of the predictor variables can also be inspected to improve the regression fit.

```
## Residual Plots
  par(mfrow = c(1,2))
  plot(lm1, which = 1:2)

## Residual Plots e_i vs x_i
  par(mfrow = c(2,2))
  plot(resid(lm1)~Temp)
  abline(h = 0)
  plot(resid(lm1)~Rain)
  abline(h = 0)
  plot(resid(lm1)~PrevRain)
  abline(h = 0)
  plot(resid(lm1)~Age)
  abline(h = 0)
```

```
> shapiro.test(resid(lm1))

        Shapiro-Wilk normality test

data:  resid(lm1)
W = 0.90843, p-value = 0.02096
```

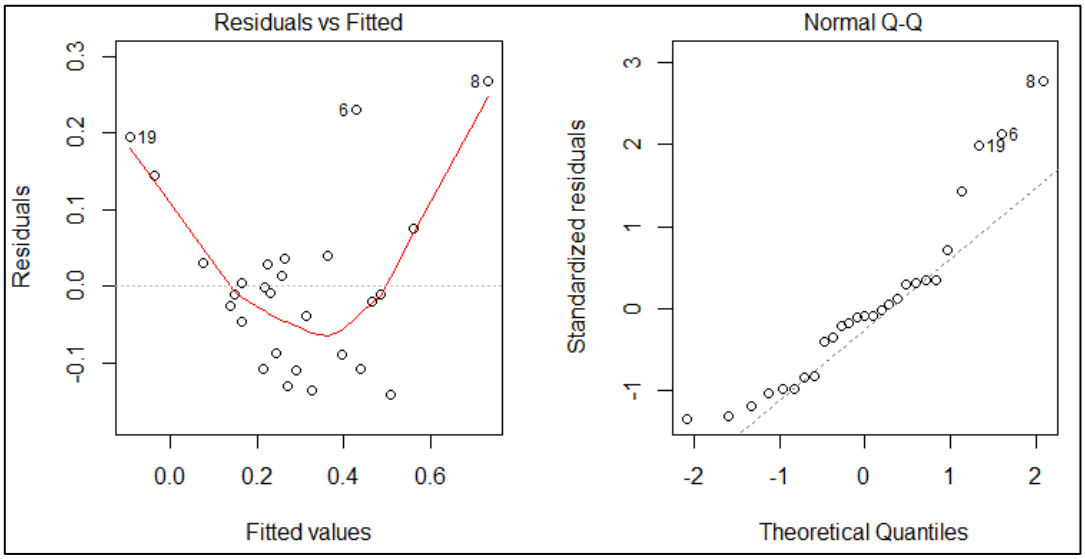

Figure 4 Residual Plots (Equal Variance and Normality Check)

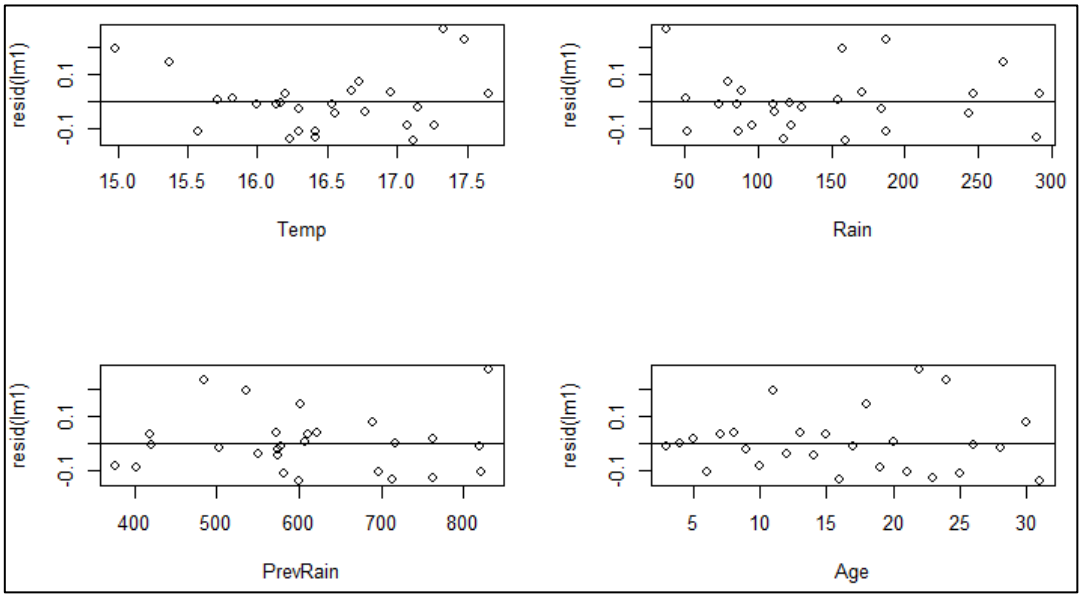

Figure 5 Residual vs Predictor Variables Plot (Linearity Check)

The independence of residuals can also be tested by calculating the Lag 1 autocorrelation of the residuals as shown in the code below. The Lag 1 autocorrelation of the residuals is, $r_1 = -0.588$. Also, $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{27}} = 0.385$. Since $|-0.588| > 0.385$, therefore, the independence of errors is not reasonable.

```
> ## Find r1: Lag 1 Autocorrelation
> acf(resid(lm1), lag.max = 1, plot = FALSE)

Autocorrelations of series 'resid(lm1)', by lag

    0     1
1.000 -0.588
```

From the above diagnosis of residuals, we realize that there are certain aspects of the fit that are not very satisfactory. This suggests that a transformation of the response variable, *Price*, might be effective. To choose a suitable power, we can use the Box-Cox procedure as shown in the code below. The maximum value of log-likelihood occurs at a power (λ) of about -0.3, which is reasonably close to a log (i.e. $\lambda = 0$). Also, the plot in Figure 6 includes 0 in the 95% CI for λ, which suggests that a log transformation of the response variable is reasonable.

```
## BoxCox Plot
library("MASS")
bc = boxcox(Price ~ Temp + Rain + PrevRain + Age, data=dataBV)
lambda = bc$x[which.max(bc$y)] ## Optimal lambda
```

```
> lambda
[1] -0.3
```

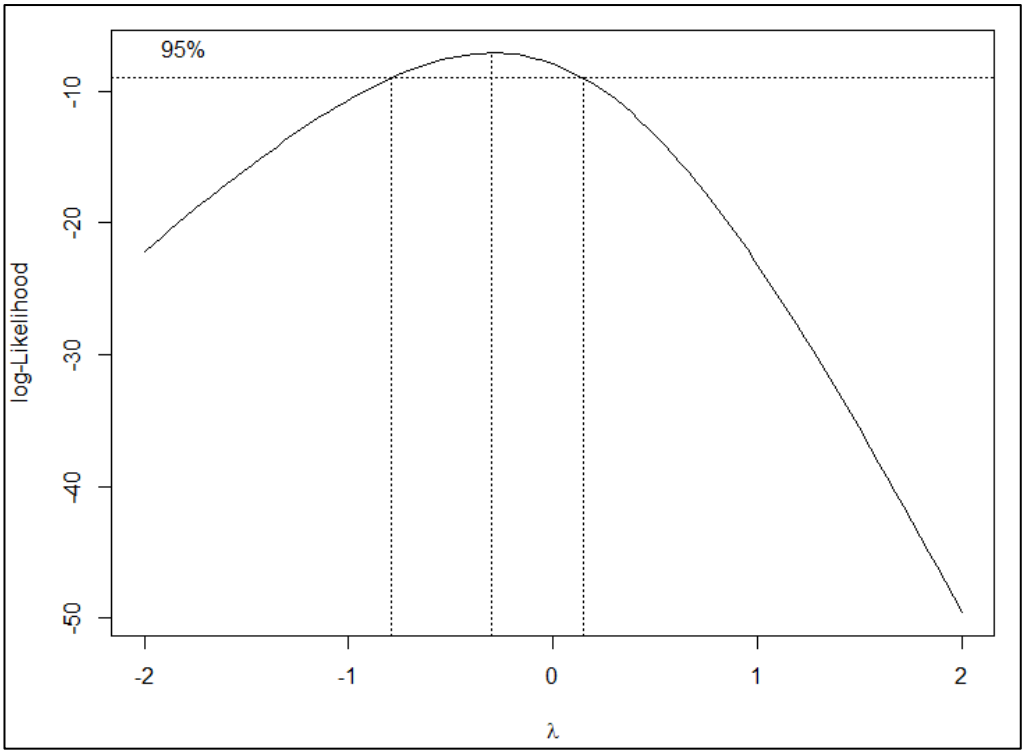

Figure 6 Box-Cox Plot for Bordeaux Vintage Wine Dataset

We now transform the response variable, *Price*, using log (natural log) as shown below. Also, a scatter of Log(*Price*) versus all the predictor variables, *Temp*, *Rain*, *PrevRain* and *Age* is obtained as shown in Figure 7. We see that Log(*Price*) follows a very similar trend with respect to various predictor variables as we had seen in Figure 2. However, we see that the variance in Log(*Price*) is relatively uniform to what we had seen in Figure 2. This suggests that the transformed data is expected to not suffer from the heteroskedasticity problem, and thus might result in a better regression fit.

```
## Log transformation of response variable
  logPrice <- log(Price)
## Plot Relationship between LogPrice and predictor variables
  par(mfrow = c(2,2))
  plot(logPrice~Temp, xlab = "Temperature (in °C)", ylab = "Log(Price)")
  plot(logPrice~Rain, xlab = "Rain (in mm)", ylab = "Log(Price)")
  plot(logPrice~PrevRain, xlab = "PrevRain (in mm)", ylab = "Log(Price)")
  plot(logPrice~Age, xlab = "Age (in years)", ylab = "Log(Price)")
```

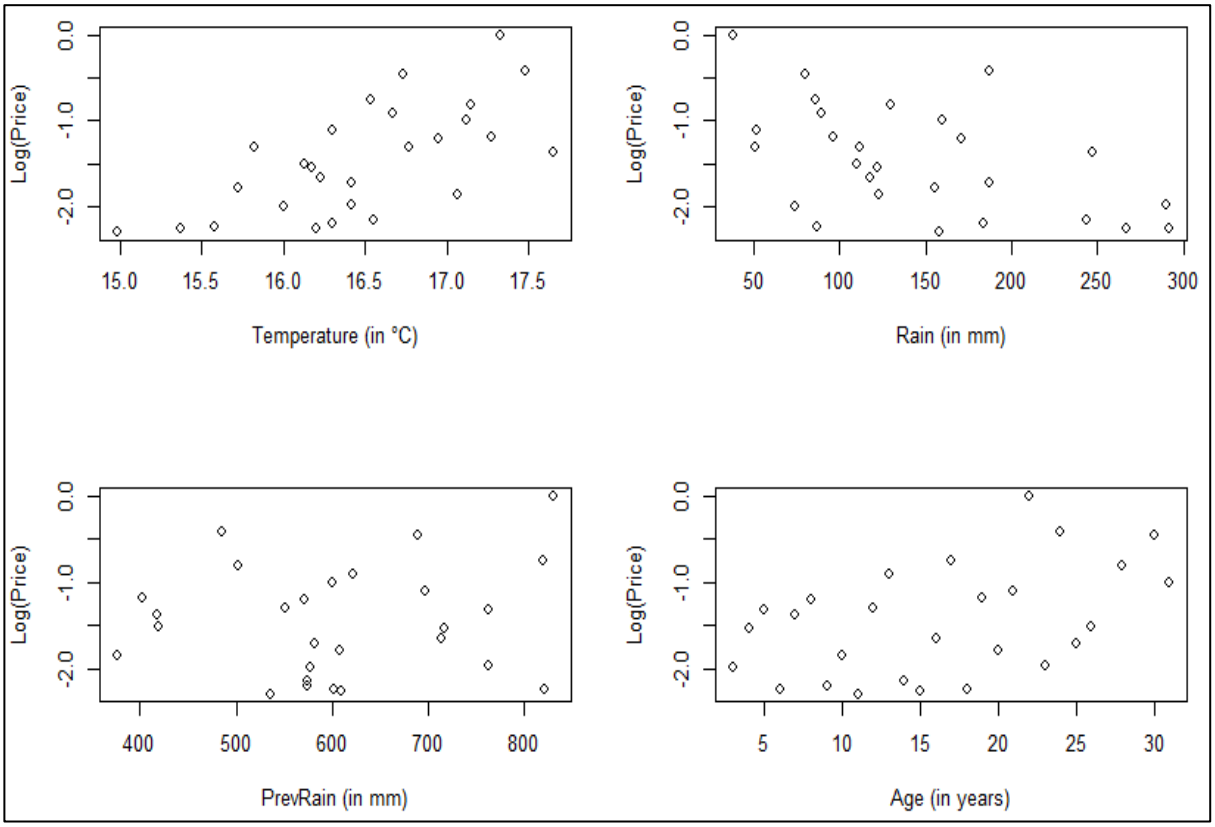

Figure 7 Scatter plot of Log(Price) v/s all Predictor Variables

Let us now refit and obtain a new multiple linear regression model with Log(*Price*) as the response variable and *Temp*, *Rain*, *PrevRain* and *Age* as the predictor variables as shown below. The summary of the fit with the estimated model parameters is also shown.

```
## Run Regression Model on logPrice
  lm2 <- lm(logPrice~Temp + Rain + PrevRain + Age)
  summary(lm2)
```

```
> summary(lm2)

Call:
lm(formula = logPrice ~ Temp + Rain + PrevRain + Age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45748 -0.23902  0.01067  0.18533  0.53642

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.216e+01  1.686e+00  -7.213 3.15e-07 ***
Temp         6.170e-01  9.502e-02   6.493 1.57e-06 ***
Rain        -3.866e-03  8.062e-04  -4.795 8.66e-05 ***
PrevRain     1.171e-03  4.814e-04   2.432  0.02359 *
Age          2.390e-02  7.155e-03   3.341  0.00296 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2861 on 22 degrees of freedom
Multiple R-squared:  0.8282,    Adjusted R-squared:  0.797
F-statistic: 26.51 on 4 and 22 DF,  p-value: 3.89e-08
```

Thus, the resulting multiple regression model is,

$$\text{Log}(Y) = -12.159 + 0.617 \cdot X_1 - 0.004 \cdot X_2 + 0.001 \cdot X_3 + 0.024 \cdot X_4$$

The estimated coefficients of the model indicate a similar direction of the relationship between the Log(*Price*) and each of the corresponding predictors as we had seen for *Price* before. The coefficients of temperature, rain prior to the growing season and age are positive. And, the coefficient of harvest rain is negative.

From the above summary output, we also observe that R^2 value now achieved is 0.8282, which is a lot better than the previous R^2 value of 0.7356. Thus, about 82.82% of the variation in Log(*Price*) is explained by a multivariate regression on the explanatory variables, *Temp*, *Rain*, *PrevRain* and *Age*.

To test the importance (significance) of each predictor variables in developing the above regression model, we can perform a partial t-test on the model parameters. We can obtain the *p*-value for each of the predictor variables from the above summary output. We see that all the predictor variables have a very small *p*-value and thus we conclude that they contribute significantly in explaining the regression at a significance level of 0.05.

Let us now look at various residual plots to check the adequacy of the new regression model. The residual versus fitted value plot in Figure 8 now doesn't show a funnel effect and residuals appear to be reasonably well scattered. Thus, the residuals can be considered to have equal variances. The curve in the normal plot of the residuals shows relatively higher degree of normality. The Shapiro-Wilk Normality Test results in a large *p*-value, thereby, also indicating normality of residuals to be reasonable. Further, in Figure 9, we see that the residuals vs predictor variables plot shows reasonably a good scatter of residuals. This suggests that a linear model on Log(*Price*) is suitable.

```
## Residual Plots
  par(mfrow = c(1,2))
  plot(lm2, which = 1:2)

## Residual Plots e_i vs x_i
  par(mfrow = c(2,2))
  plot(resid(lm2)~Temp)
  abline(h = 0)
  plot(resid(lm2)~Rain)
  abline(h = 0)
  plot(resid(lm2)~PrevRain)
  abline(h = 0)
  plot(resid(lm2)~Age)
  abline(h = 0)
```

```
> shapiro.test(resid(lm2))

        Shapiro-Wilk normality test

data:  resid(lm2)
W = 0.95865, p-value = 0.3441
```

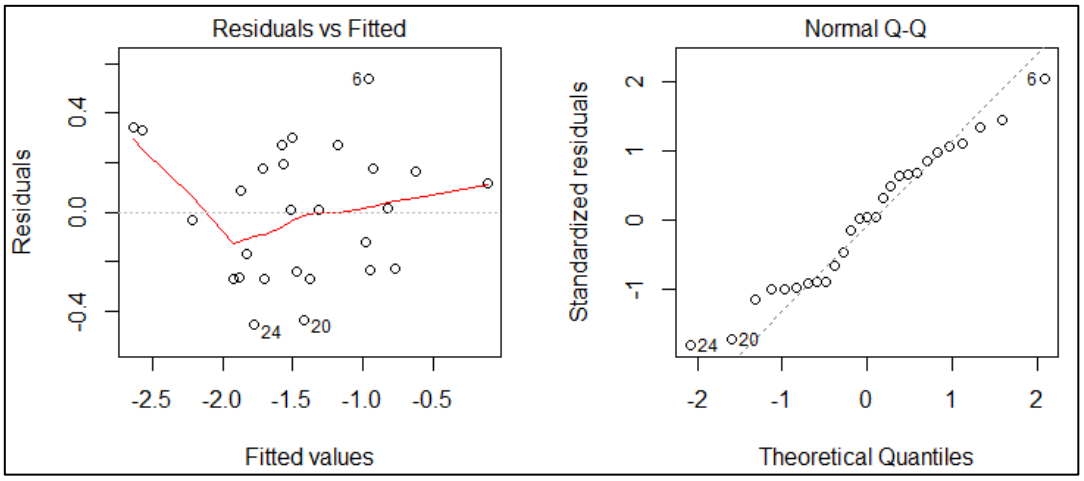

Figure 8 Residual Plots (Equal Variance and Normality Check)

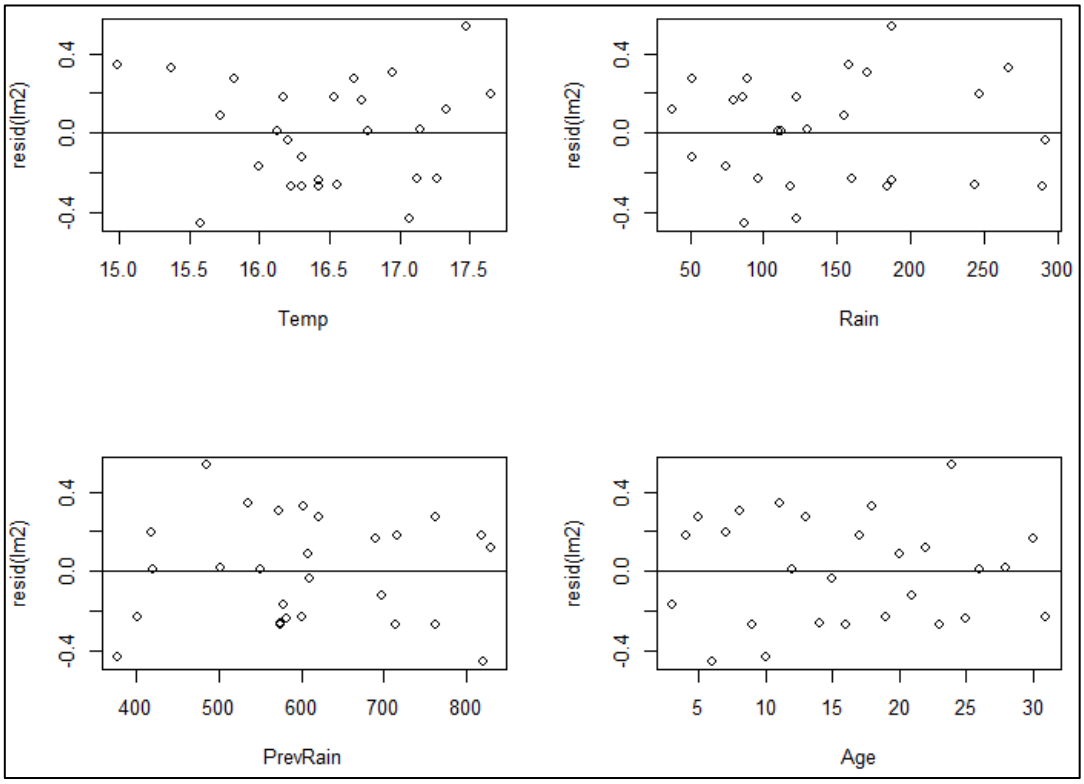

Figure 9 Residual vs Predictor Variables Plot (Linearity Check)

The independence of residuals can also be tested by calculating the Lag 1 autocorrelation of the residuals as shown in the code below. The Lag 1 autocorrelation of the residuals is, $r_1 = -0.417$. Also, $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{27}} = 0.385$. Even though the value of Lag 1 autocorrelation is better than the previous model, but since $|-0.417| > 0.385$, therefore, the independence of errors is still not reasonable.

```
> ## Find r1: Lag 1 Autocorrelation
> acf(resid(lm2), lag.max = 1, plot = FALSE)

Autocorrelations of series 'resid(lm2)', by lag

    0     1
1.000 -0.417
```

Let us now investigate if we can further improve the fit of the model and meet all the residual assumptions. In order to do this, let us check if adding the quadratic relationships (i.e. squared terms) of the significant predictor variables fits the model better. The quadratic terms are added one at a time and their significance is tested as shown below. The result in Table 1 shows that only $Temp^2$ is significant with small p-value, and the corresponding linear model also achieves a highest R^2 value of 86.21%. As a result, we can consider adding a quadratic relationship of temperature variable to the previous model to improve its fit.

```
## Quadratic terms
  Temp2 = Temp^2
  Rain2 = Rain^2
  PrevRain2 = PrevRain^2
  Age2 = Age^2

## Run Regression Model on logPrice with quadratic
## predictors added one at a time
  lm3 <- lm(logPrice~Temp + Rain + PrevRain + Age + Temp2)
  lm4 <- lm(logPrice~Temp + Rain + PrevRain + Age + Rain2)
  lm5 <- lm(logPrice~Temp + Rain + PrevRain + Age + PrevRain2)
  lm6 <- lm(logPrice~Temp + Rain + PrevRain + Age + Age2)
```

Table 1 Checking Significance of Quadratic Predictor Variables

Term Added	*p*-value	Term Significant?	Model R^2
$Temp^2$	0.03359	Yes	86.21%
$Rain^2$	0.98413	No	82.82%
$PrevRain^2$	0.31700	No	83.64%
Age^2	0.89113	No	82.84%

Based on the above conclusion, let us obtain a multiple linear regression model with Log(*Price*) as the response variable and *Temp*, *Rain*, *PrevRain*, *Age* and $Temp^2$ as the predictor variables as shown below. The summary of the fit with the estimated model parameters is also shown.

```
## Run Regression Model on logPrice
  lm3 <- lm(logPrice~Temp + Rain + PrevRain + Age + Temp2)
  summary(lm3)
```

```
> summary(lm3)

Call:
lm(formula = logPrice ~ Temp + Rain + PrevRain + Age + Temp2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56447 -0.14823 -0.00744  0.19657  0.38077

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.8983511 27.7751374   1.833  0.08109 .
Temp        -7.1239013  3.4055189  -2.092  0.04878 *
Rain        -0.0039643  0.0007405  -5.354 2.61e-05 ***
PrevRain     0.0014836  0.0004623   3.209  0.00421 **
Age          0.0248916  0.0065743   3.786  0.00108 **
Temp2        0.2364667  0.1039969   2.274  0.03359 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2623 on 21 degrees of freedom
Multiple R-squared:  0.8621,    Adjusted R-squared:  0.8293
F-statistic: 26.27 on 5 and 21 DF,  p-value: 2.278e-08
```

Thus, the resulting multiple regression model is,

$$\text{Log}(Y) = 50.898 - 7.124 \cdot X_1 - 0.004 \cdot X_2 + 0.001 \cdot X_3 + 0.025 \cdot X_4 + 0.236 \cdot X_1^2$$

The estimated coefficient of temperature is now negative; however, the quadratic term of temperature has a positive coefficient. The coefficients of rain prior to the growing season and age are positive. And, the coefficient of harvest rain is negative.

From the above summary output, we observe further improvement in R^2 value which is now 0.8621 compared to the value of 0.8282 obtained in the previous model without the quadratic term. Thus, about 86.21% of the variation in Log(*Price*) is explained by a multivariate regression on the explanatory variables, *Temp*, *Rain*, *PrevRain*, *Age* and $Temp^2$.

To test the importance (significance) of each predictor variables in developing the above regression model, we can perform a partial t-test on the model parameters. We can obtain the *p*-value for each of the predictor variables from the above summary output. We see that almost all the predictor variables have a very small *p*-value and thus we conclude that they contribute significantly in explaining the regression at a significance level of 0.05.

Let us now look at various residual plots to check the adequacy of the new regression model. The residual versus fitted value plot in Figure 10 now shows even better scatter of residuals. Thus, the residuals can be considered to have equal variances. The curve in the normal plot of the residuals shows increasing degree of normality. The Shapiro-Wilk Normality Test results in a very large *p*-value, thereby, also indicating normality of residuals to be even more reasonable. Further, in Figure 11, we see that the residuals vs predictor variables plot shows

reasonably a good scatter of residuals. This suggests that the new linear model developed on Log(*Price*) is suitable.

```
## Residual Plots
  par(mfrow = c(1,2))
  plot(lm3, which = 1:2)

## Residual Plots e_i vs x_i
  par(mfrow = c(2,3))
  plot(resid(lm3)~Temp)
  abline(h = 0)
  plot(resid(lm3)~Rain)
  abline(h = 0)
  plot(resid(lm3)~PrevRain)
  abline(h = 0)
  plot(resid(lm3)~Age)
  abline(h = 0)
  plot(resid(lm3)~Temp2)
  abline(h = 0)
```

```
> shapiro.test(resid(lm3))

        Shapiro-Wilk normality test

data:  resid(lm3)
W = 0.97638, p-value = 0.7732
```

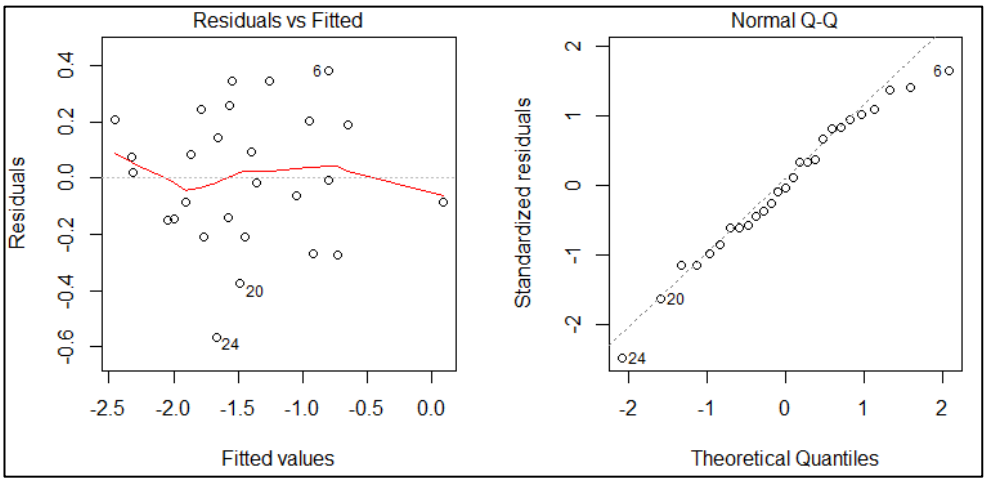

Figure 10 Residual Plots (Equal Variance and Normality Check)

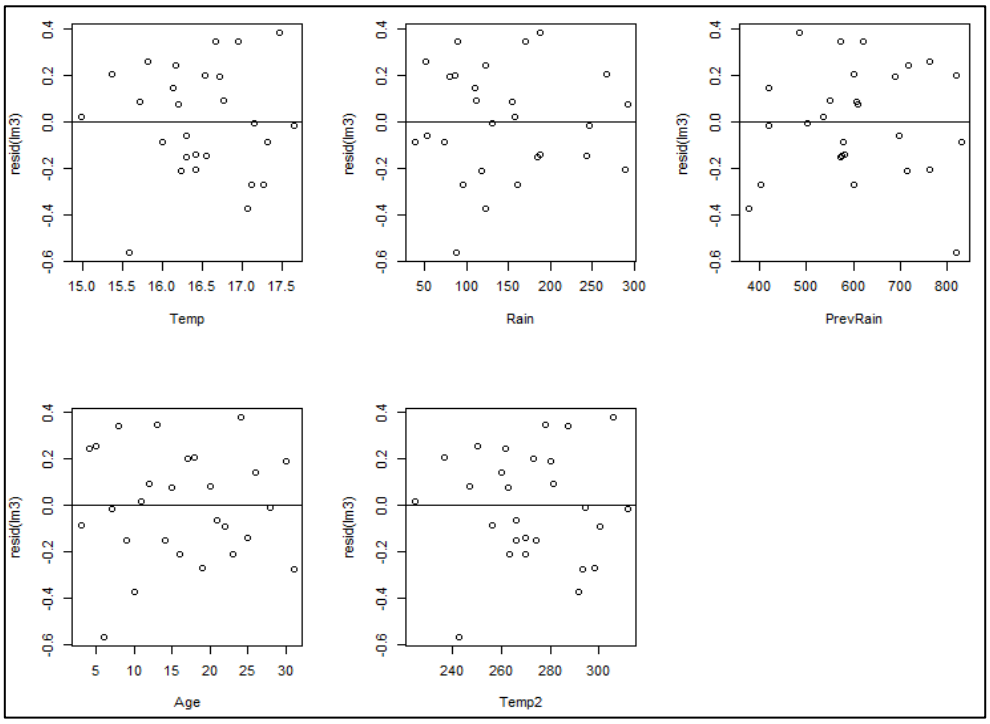

Figure 11 Residual vs Predictor Variables Plot (Linearity Check)

The independence of residuals can also be tested by calculating the Lag 1 autocorrelation of the residuals as shown in the code below. The Lag 1 autocorrelation of the residuals is, $r_1 = -0.286$. Also, $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{27}} = 0.385$. Since $|-0.286| < 0.385$, therefore, the independence of errors is now reasonable.

```
> ## Find r1: Lag 1 Autocorrelation
> acf(resid(lm3), lag.max = 1, plot = FALSE)

Autocorrelations of series 'resid(lm3)', by lag

    0     1
1.000 -0.286
```

Conclusions

- Model using Log(*Price*) as response fits reasonable well.
- All the predictor variables *Temp*, *Rain*, *PrevRain* and *Age* are significant in explaining the regression.
- Including the quadratic term of the temperature variable, $Temp^2$, further improves the fit of the regression model by achieving the largest R^2 value of 86.21%.
- The diagnosis of residuals indicates that the final model is adequate.
- The best multiple linear regression model obtained is:

$$\text{Log}(Y) = 50.898 - 7.124 \cdot X_1 - 0.004 \cdot X_2 + 0.001 \cdot X_3 + 0.025 \cdot X_4 + 0.236 \cdot X_1^2$$

(OR)

$$\text{Log}(Price) = 50.898 - 7.124 \cdot (Temp) - 0.004 \cdot (Rain) + 0.001 \cdot (PrevRain) + 0.025 \cdot (Age) + 0.236 \cdot (Temp)^2$$

APPENDIX

The *backward elimination* and *forward selection* methods were also employed to check the significance of predictor variables *Temp*, *Rain*, *PrevRain*, *Age*, and $Temp^2$ in explaining the variation in Log(*Price*). Both these methods ended up giving the same results as shown in the final model above, thereby, indicating significance of all the predictor variables. The implementation of *backward elimination* and *forward selection* methods is shown below.

```
## Backward Elimination
  dataBV2 <- data.frame(logPrice, Temp, Rain, PrevRain, Age, Temp2)
  full = lm(logPrice ~ ., data = dataBV2) ## Saves the full model
  step(full, data = dataBV2, direction = "backward")
```

```
Start:  AIC=-67.05
logPrice ~ Temp + Rain + PrevRain + Age + Temp2

           Df Sum of Sq    RSS     AIC
<none>                  1.4447 -67.054
- Temp      1   0.30105 1.7458 -63.943
- Temp2     1   0.35569 1.8004 -63.111
- PrevRain  1   0.70848 2.1532 -58.280
- Age       1   0.98622 2.4310 -55.004
- Rain      1   1.97191 3.4167 -45.814

Call:
lm(formula = logPrice ~ Temp + Rain + PrevRain + Age + Temp2,
    data = dataBV2)

Coefficients:
(Intercept)         Temp         Rain     PrevRain          Age        Temp2
  50.898351    -7.123901    -0.003964     0.001484     0.024892     0.236467
```

```
## Forward Selection
  null = lm(logPrice ~ 1, data = dataBV2) ## ~1 = intercept only
  step(null, scope = list(lower = null, upper = full),
    direction = "forward")
```

```
Start:  AIC=-23.55
logPrice ~ 1

           Df Sum of Sq     RSS     AIC
+ Temp2     1    4.6657  5.8142 -37.459
+ Temp      1    4.6636  5.8163 -37.450
+ Rain      1    2.6957  7.7842 -29.581
+ Age       1    2.2242  8.2557 -27.993
<none>                  10.4799 -23.552
+ PrevRain  1    0.1912 10.2887 -22.049

Step:  AIC=-37.46
logPrice ~ Temp2

           Df Sum of Sq    RSS     AIC
+ Rain      1   2.53861 3.2756 -50.952
+ PrevRain  1   1.48008 4.3342 -43.391
+ Age       1   0.80770 5.0065 -39.497
<none>                  5.8142 -37.459
+ Temp      1   0.00005 5.8142 -35.459
```

```
Step:  AIC=-50.95
logPrice ~ Temp2 + Rain

           Df Sum of Sq    RSS     AIC
+ Age       1   1.02056 2.2551 -59.032
+ PrevRain  1   0.61757 2.6581 -54.592
<none>                  3.2756 -50.952
+ Temp      1   0.05055 3.2251 -49.372

Step:  AIC=-59.03
logPrice ~ Temp2 + Rain + Age

           Df Sum of Sq    RSS     AIC
+ PrevRain  1   0.50929 1.7458 -63.943
<none>                  2.2551 -59.032
+ Temp      1   0.10186 2.1532 -58.280

Step:  AIC=-63.94
logPrice ~ Temp2 + Rain + Age + PrevRain

       Df Sum of Sq    RSS     AIC
+ Temp  1   0.30105 1.4447 -67.054
<none>              1.7458 -63.943

Step:  AIC=-67.05
logPrice ~ Temp2 + Rain + Age + PrevRain + Temp


Call:
lm(formula = logPrice ~ Temp2 + Rain + Age + PrevRain + Temp,
    data = dataBV2)

Coefficients:
(Intercept)        Temp2         Rain          Age     PrevRain         Temp
  50.898351     0.236467    -0.003964     0.024892     0.001484    -7.123901
```