**Studying the relationship between Hollywood’s total quarterly gross and movie with highest gross that quarter**

In this analysis I am trying to understand the relationship between Hollywood’s total quarterly gross(sum of gross of all movies released in that quarter) and total number of movies released in that quarter and with performance of movie with highest gross for that quarter.The data is collected from from boxofficemojo.com.The data is from first quarter of year 2000 to first quarter of year 2015,Having a total of 61 records.

We have three predictor variables,

1. Total number of movies released in that quarter
2. Gross of highest earning movie in that quarter
3. Total number of theaters in which the highest earning movie of the quarter was screened.

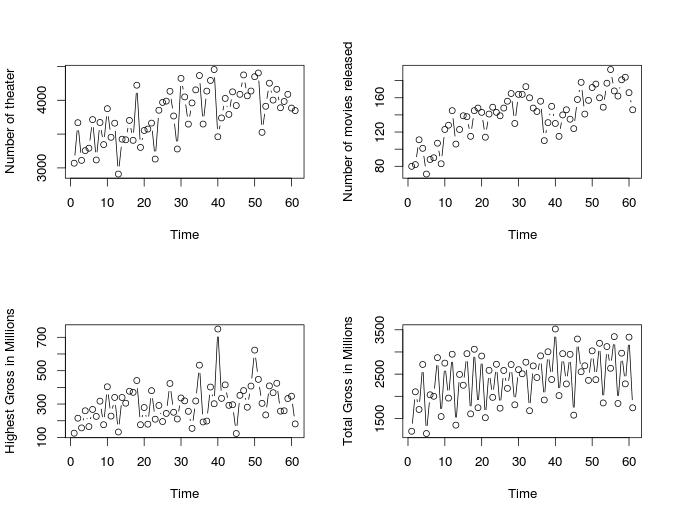
The Total Gross of all movies released in that quarter is the target variable.The Total Gross for this analysis is considered in Million US Dollars.If we do a scatter plot between the Total Gross and the prediction variables.We get the below plots





We can see some relationship between Total Gross and highest Gross.and a moderate to weak relationship between Total Gross and movies released or total gross and number of theaters.

Since the data is quarterly data,I want to check the time order plots of each variable.We get the below plots if we plot them against time



We can clearly see a cyclic pattern followed by each variable .We can clearly see the Gross of first quarter is always least and is highest for the fourth or second quarter in general.A similar seasonal effect can be even seen for total movies released and number of theaters.

If we perform a Multi linear regression with Total gross as target variable and Highest gross,theaters and total movies as prediction variables.We get the below results

*Call:  
lm(formula = Gross ~ TopMovie + TotalMovies + Theaters)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-819.0 -352.8 -57.7 328.9 716.8   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 691.9115 548.6775 1.261 0.2124   
TopMovie 2.7778 0.5243 5.298 1.96e-06 \*\*\*  
TotalMovies 6.5494 2.4062 2.722 0.0086 \*\*   
Theaters -0.0146 0.1860 -0.078 0.9377   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 414.3 on 57 degrees of freedom  
Multiple R-squared: 0.5291, Adjusted R-squared: 0.5043   
F-statistic: 21.35 on 3 and 57 DF, p-value: 2.148e-09*

The R-squared value of 53% shows that there is a weak relationship between the target variables and prediction variables.We can reject the null hypothesis that all slopes are zero from the obtained p-value.The intercept of 619 does not give out any information in this case.The coefficient of Theaters is negative and its p-value of .93 is statistically significant suggests that the slope of this variable is equals to zero.

The standard error signifies that the total Gross of the movies during the quarter can be estimated with in [-828,+828]Million dollars with 95% confidence interval.

As the total theaters does not provide any prediction power,I decide to go with only two prediction variables and perform regression on the data.We get the below output

*Call:  
lm(formula = Gross ~ TopMovie + TotalMovies)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-814.77 -352.29 -47.47 323.80 719.03   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 654.5530 270.4067 2.421 0.0186 \*   
TopMovie 2.7638 0.4886 5.657 4.98e-07 \*\*\*  
TotalMovies 6.4504 2.0316 3.175 0.0024 \*\*   
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 410.8 on 58 degrees of freedom  
Multiple R-squared: 0.5291, Adjusted R-squared: 0.5128   
F-statistic: 32.58 on 2 and 58 DF, p-value: 3.282e-10*

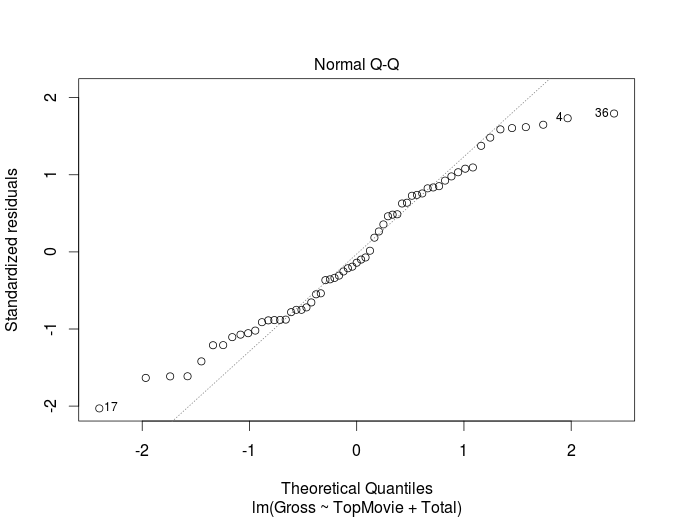
The R-squared value and standard errors in this case have slightly improved but they haven't considerably changed.The Top Movie coefficient of 2.7 says that keeping the other variable constant a Million Dollar increase in Gross of Highest earning movie is associated with 2.7 Million Dollars rise in total gross for the quarter.The Total movies coefficient of 6.4 indicate that a new movie released in the quarter is associated with 6.4Million Dollars rise in Total Gross.

When I check for the collinearity in this model,We get the below result

*> vif(boxoffice)  
TopMovie TotalMovies   
1.168951 1.168951*

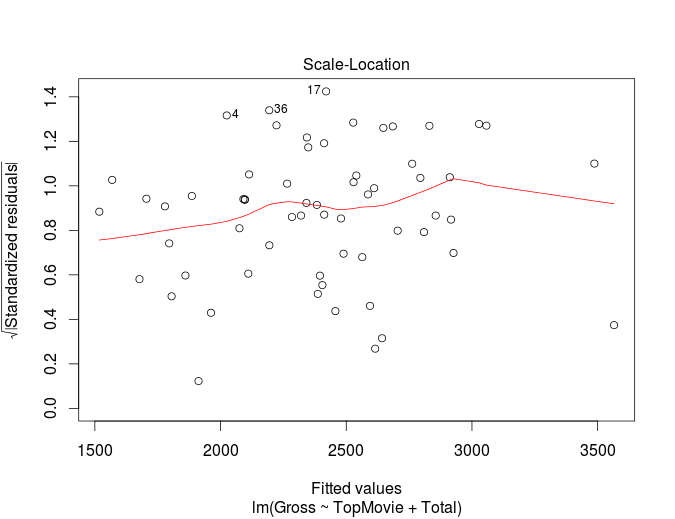
Since both values are below 10.It shows that both variables are independent of each other.

Moving on to diagnosis ,by performing the Normality check.We get the below plot



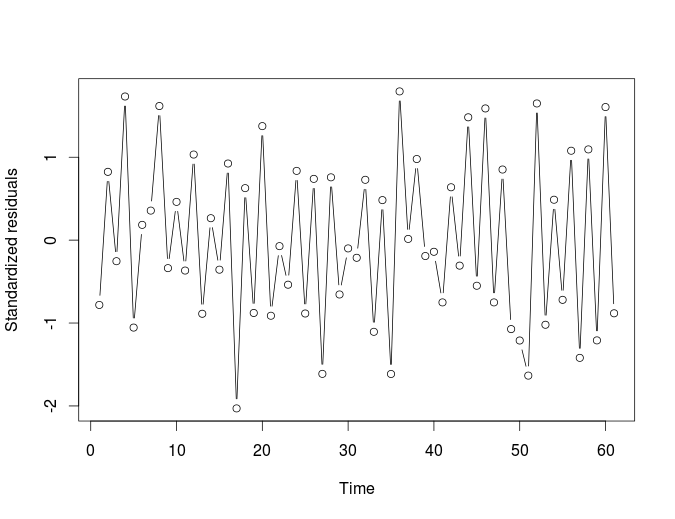
We can see that the residuals are not normally distributed.We can see a bimodal type of distribution in residuals.

Checking for non-constant variance we get the below plot



We don't really see any non-constant variance(the hill shape in the middle can be due to auto-correlation).

When we check the plot for auto-correlation among the residuals we get the below plot



We can see cyclic pattern in the standardized residuals.There can be auto correlation among the residuals due to this.

We perform the test to identify the auto-correlation between the residuals,starting with Durbin-Watson test.The assumption of normality and constant variance have to be satisfied for this test.The normality is a problem with this data,but still I am performing durbin-watson test.

*> durbinWatsonTest(boxoffice)  
 lag Autocorrelation D-W Statistic p-value  
 1 -0.6203892 3.218459 0  
 Alternative hypothesis: rho != 0*

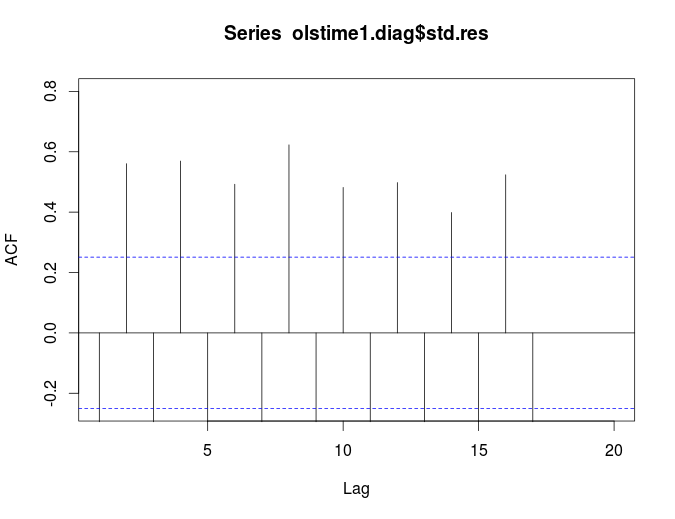
The higher value of durbin watson test(3.2) shows that there is negative correlation for lag 1.The Autocorrelation is -.62.

4-3.2=.78,

.78<Ql and Qu .Hence we can reject the null hypothesis of the residuals being normally distributed with mean zero

The durbin-watson test can only get the autocorrelation for lag 1.To get the autocorrelations for lag>1.We need to check the ACF plot.

Below is the ACF plot for the model.



The ACF plot shows that there is an autocorrelation between each and every lag of this data.The autocorrelation is negative for all the odd lags and positive for all even lags.This clearly indicates that this is not a AR(1) time series as the autocorrelations are not dropping down.

The autocorrelation values are shown below.

*[1] -0.6160696 0.5605579 -0.6057601 0.5693187 -0.5175045 0.4923658 -0.5575970  
 [8] 0.6228791 -0.5639137 0.4819131 -0.5199608 0.4978784 -0.4971935 0.3988714  
[15] -0.4997684 0.5236973 -0.4283383 NA NA NA*

The Gross of Hollywood Revenues each quarter are negatively correlated.The Gross is low in first and third quarter and it increases in second and fourth quarter.We don't see many movies released during the first quarter as it is Oscar’s season.and the revenue’s go up during the Summer they dip again in the fall,and rise back to its peak during the thanksgiving and christmas break.This explains why there is a autocorrelation between each lag of this data.

Performing Runs-Test we get the below result

*Runs Test  
  
data: olstime1.diag$std.res  
Standard Normal = 4.4735, p-value = 7.696e-06*

This test does not have any assumptions.We can reject the Null hypothesis of having no autocorrelation in the residuals by the p-value obtained in this test.

We can clearly confirm there is autocorrelation between the residuals from the above three test.The ACF also shows that there is a clear seasonal effect in the data.

Now we need to deseasonalize the data and try to remove the autocorrelation between the residuals.

I introduce 3 new indicator variable to deseasonalize the data(Q1,Q2,Q3). These variables take the value of one if data belongs to that quarter else they have the value of zero.

Doing a Multi Linear Regression with these 3 Indicator variables.we get the below result.

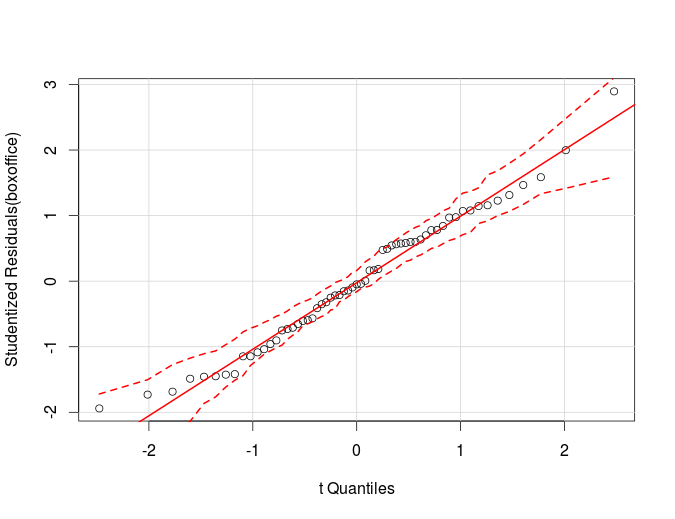
*Call:  
lm(formula = Gross ~ TopMovie + TotalMovies + Q1 + Q2 + Q3)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-302.88 -124.19 -8.34 119.85 461.40   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1658.3994 139.6375 11.876 < 2e-16 \*\*\*  
TopMovie 1.5518 0.2301 6.744 9.91e-09 \*\*\*  
TotalMovies 5.5174 0.9100 6.063 1.27e-07 \*\*\*  
Q1 -986.4385 69.5184 -14.190 < 2e-16 \*\*\*  
Q2 -245.8969 64.9556 -3.786 0.000381 \*\*\*  
Q3 -757.8474 64.8697 -11.683 < 2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 176.5 on 55 degrees of freedom  
Multiple R-squared: 0.9176, Adjusted R-squared: 0.9101   
F-statistic: 122.4 on 5 and 55 DF, p-value: < 2.2e-16*

We can see the R-squared value has increased from 52% to 92% by deseasonalizing,which strengthens the model.The Standard error has also come down from 400 to 176.5 by introducing the three new indicator variables.

The Coefficient of Q variables indicate that,The associated total gross will be reduced 986Million Dollars if it is Quarter-1 and 245 Million Dollars if it is Quarter-2 and 757 Million Dollars if it is Quarter-3.

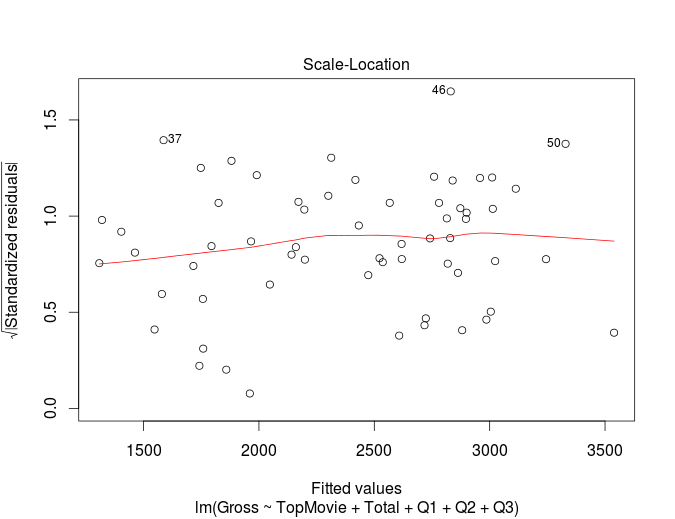
This shows the Total Gross of Hollywood is highest in 4th Quarter followed by 2nd and 3rd Quarters and is least in 1st Quarter.

Checking for the normality of the residuals,we get the below plot



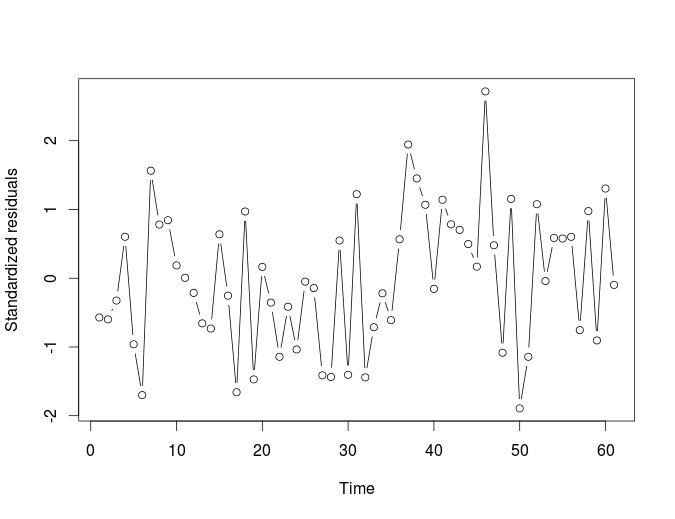
From the above graph we can say the residuals are following the line and are normally distributed with 95% confidence interval.

Check for non-constant variance we get the below graph



There is no non-constant variance seen.But there seem to appear a few outliers in the data.

Looking at the time series plot for the standardized residuals we get the below graph



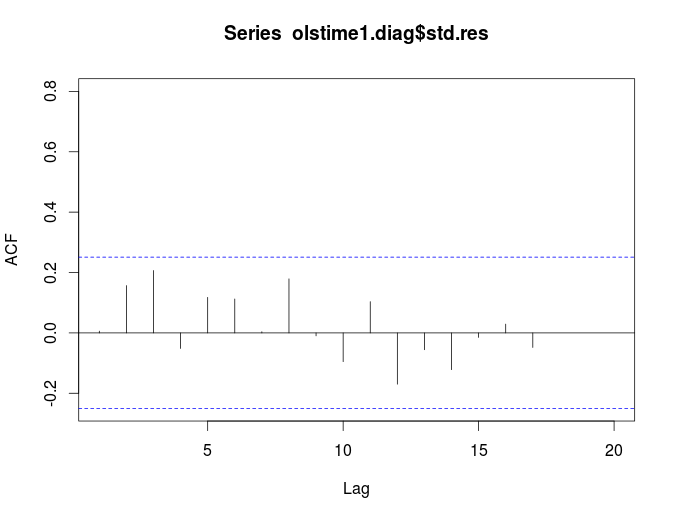
We can't clearly see any cycles as we have seen in the previous plot.We will again perform the autocorrelation test to see,if there is any autocorrelation between the variables.

Both the normality and constant variance condition satisfies for performing the Durbin-Watson test.Below is the result of performing durbin-watson

*durbinWatsonTest(boxoffice)  
 lag Autocorrelation D-W Statistic p-value  
 1 0.01063319 1.973314 0.94  
 Alternative hypothesis: rho != 0*

The autocorrelation for lag-1 is shown as .01,with a statistically significant p-value of .94.Hence we accept the Null hypothesis that there is no autocorrelation been shown for lag-1 using Durbin-Watson

We use ACF plot to identify if there is is any autocorrelation for lag>1.We get the below plot



We don't see any significant autocorrelation between the residuals for any lag.

Below are the autocorrelation values obtain.

*[1] 0.006033095 0.156710287 0.206633291 -0.051418706 0.117713010 0.112376495  
 [7] 0.003967640 0.179128546 -0.009849852 -0.095298884 0.103574958 -0.169953528  
[13] -0.055769728 -0.121655189 -0.014942306 0.029066600 -0.048199596 NA  
[19] NA NA*

Clearly we see that there is no significant autocorrelation between the residuals using the Autocorrelation function.

Performing Runs test to identify if there is any autocorrelation between the residuals,we get the below result

*Runs Test  
  
data: olstime1.diag$std.res  
Standard Normal = -0.4986, p-value = 0.618*

The significant p-value of .618 ,makes us to accept the the Null Hypothesis that there is no autocorrelation between the errors.

By performing the above 3 test,we can show that we have removed the autocorrelation between the residuals by deseasonalizing using the three indicator variables.