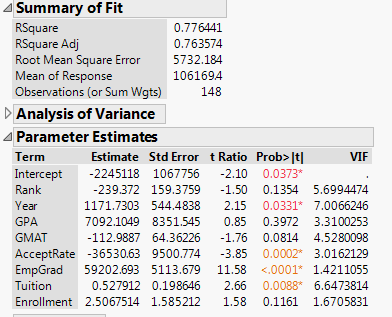
I.

A. For this first part, the goal is to predict the Salary variable.

a. Run a standard OLS regression using the variables just as they are (without School name),  
with no modifications at all, to predict the Salary variable. What is your fitted equation?

**ANS:**

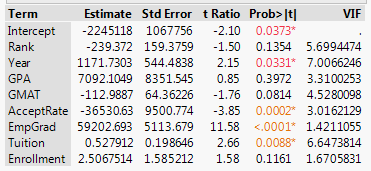


b. Are there any issues or violations with your output results from part a?

**ANS:**

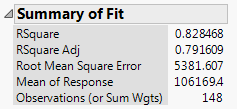
Yes,

* Some independent variables are not statistically significant (at the usual 1%/5% threshold).



c. Attempt to run an FEM (Fixed Effects Model) with this data. (at this point, still don’t do  
anything else and do NOT bother eliminating insignificant variables yet) Is the group of  
dummy variables statistically significant? Prove your answer statistically.

**ANS:**



To check statistically, calculate restricted F-test:

F(18,148-27) = (0.82 – 0.77)/18 / (1-0.82)/(148-27)

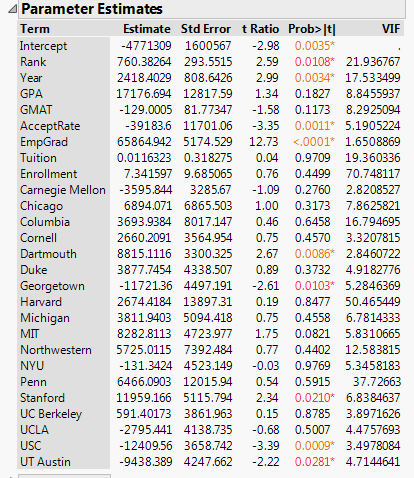
F(18,121) = (0.05/18) / (0.18/121) = 1.867

The F-Statistics at 5% level of significant, F(18,121), is 1.69 .

So, we can reject the null Hypothesis that all coefficients of dummy variables are 0 and conclude that the group of dummy variables is statistically significant.

d. Are there any other issues remaining in the output of the model from part c?

**ANS**:



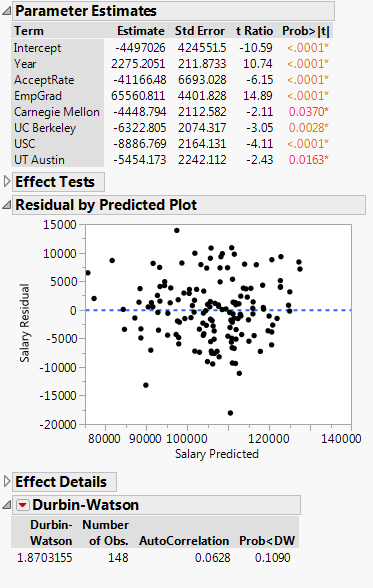
* Some independent variables are highly correlated with others (high VIF) or not statistically significant (at the usual 1%/5% threshold).

e. Work with the data to correct for the problem you noted in part d. (You shouldn’t need to  
spend TOO much time doing this!) Again, predict the Salary variable. Include a brief report  
of any models you tried and why you discarded them. (maybe a short table) What is your  
final model, including the output and residual plot?

**ANS:**

|  |  |
| --- | --- |
| Step | Explanation |
| Drop Enrollment | Too high VIF(70) |
| Drop Rank | Too high VIF(22) |
| Drop Tuition | Not significant at 5% |
| Drop UCLA | Not significant at 5% |
| Drop NYU | Not significant at 5% |
| Drop GPA | Not significant at 5% |
| Drop Cornell | Not significant at 5% |
| Drop Michigan | Not significant at 5% |
| Drop Duke | Not significant at 5% |
| Drop Northwestern | Not significant at 5% |
| Drop Dartmouth | Not significant at 5% |
| Drop MIT | Not significant at 5% |
| Drop Columbia | Not significant at 5% |
| Drop Chicago | Not significant at 5% |
| Drop Harvard | Not significant at 5% |
| Drop Stanford | Not significant at 5% |
| Drop GMAT | Not significant at 5% |
| Drop Georgetown | Not significant at 5% |
| Drop Penn | Not significant at 5% |

This is the output of my final model:

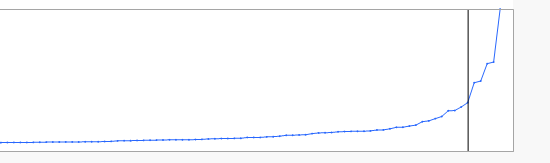


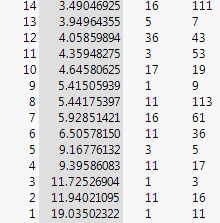
The final model is:

B. For this second part, you are to use your knowledge of Clustering Analysis.

a. Using a combination of *Hierarchical* and *K-Means* clustering methods, what number of  
clusters would you suggest using all the data except the *School* names? What is your  
reason for choosing that number?

**ANS:**

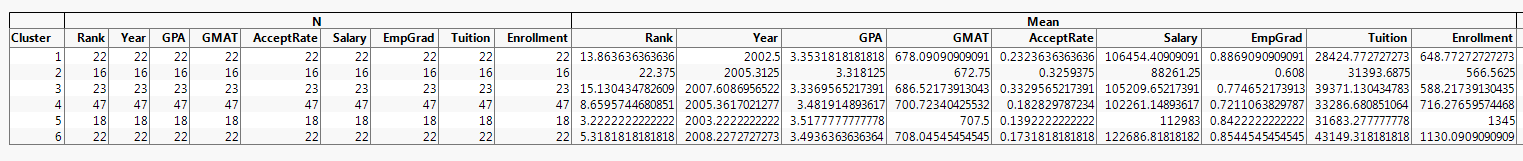


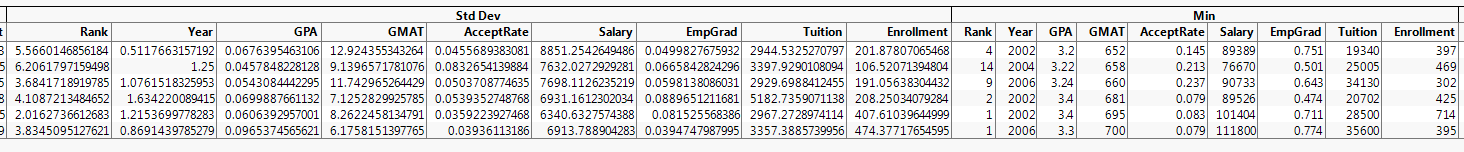


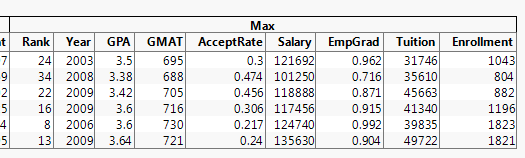
As the number of observations is not too big, I use Hierarchical clustering for determining the optimal number of group. The scree plot shows that the Distance slowly drops after splitting into 6 clusters. Based on this observation, the data should be split into 6 groups.

b. Regardless of your answer to the previous question, form 6 clusters of the data and save  
the cluster numbers to your worksheet (under the red triangle, choose *Save Clusters*).  
Create a summary table of the clusters’ statistics and copy it to your answer sheet.  
(*Analyze – Tabulate*, drag summary statistics of *N, Mean, Std Dev, Min,* and *Max* to the  
top to be the column headings. Then drag the *Cluster Number* variable to the left side to  
be the row headings, and finally select the 9 numerical variables and drag them just to the  
right of the *Cluster Number* variable. You can save this to its own table by using the red  
triangle next to *Tabulate*.)

**ANS:**

****

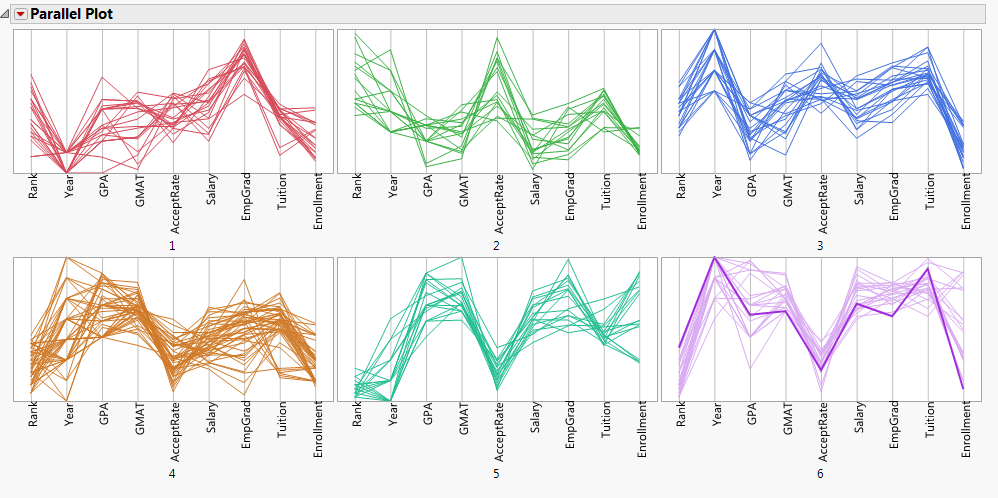
****

****

c. How would you characterize each cluster? Give just a brief description of each of the 6 in  
a summary table.

**ANS:**

The parallel coordinate plot of each cluster is displayed below:



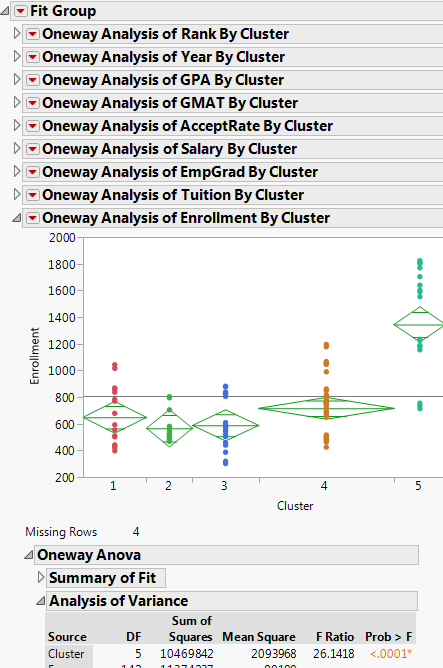
Using data from both the table and the plot, each group can be summarized as follows:

|  |  |
| --- | --- |
| Group | Summary |
| Group 1 | This group consisted of mostly middle-tier schools. It has relative lower tuition rate and not too hard to get accepted. Since salary is comparable to other groups, this group probably has the best ROI. |
| Group 2 | This group comprised of lower rank schools. It is easier to get in but a salary is not as high as in the other groups. |
| Group 3 | School ranks in this group is quite higher than group 2. It is still easier to get in than group 2 but the downside is a higher tuition rate. |
| Group 4 | This group spreads all the years from the survey. % of employment upon graduation fluctuates heavily in this group. |
| Group 5 | This cluster consisted of the top-tier schools. Undergrad GPA and GMAT of incoming students are comparably higher than other groups. It has relatively low acceptance rate. Salary upon graduation is quite high but surprisingly not as high as the last group (this is likely because salary reported in the last group come from the more recent years). |
| Group 6 | This group come from the more recent years. The group’s overall rank is not as high as group 5’s but it is easier to get accepted and still earn a comparable salary. |

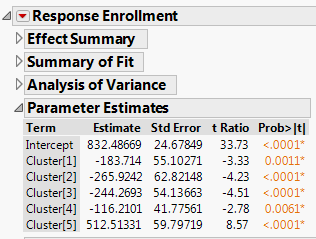
d. Which, if any, of the quantitative variables are significantly predictable using only the  
Cluster Number? Justify your answer statistically. What method did you use to answer  
this?

**ANS:**

I use ANOVA analysis and investigate the F-statistics of fitting each variable using Cluster Number. The F-statistics of all 9 variables are significant and this indicates that the overall means of each variable among 6 groups are not equal.

****

I run OLS by fitting each variable by the dummy variable Cluster Number. For each case, at least one of the dummy variable Cluster Number is significant. Statistically, this indicates that Cluster Number can be used to predict the value of each variable.



II.   
a. What type of analysis is appropriate to analyze this? What are the null hypotheses for these  
two goals?

**ANS:**

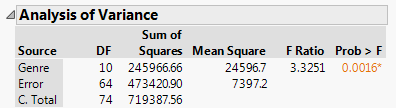
ANOVA Analysis

* There is no different in the means of US Domestic Revenues between different movie Genres
* There is no different in the means of US Domestic Revenues between different MPAA Ratings

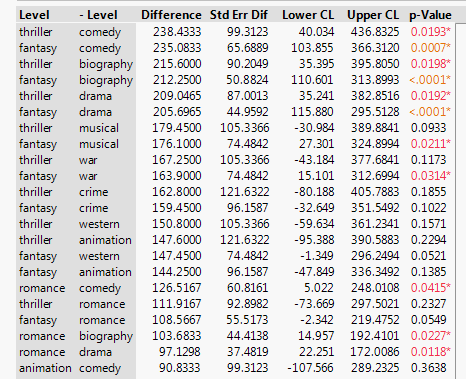
b. Determine whether there is a statistically significant difference overall in US DomesticRevenues between different Genres. Which Genres, if any, are statistically different from  
each other in terms of US Domestic Revenues?

**ANS:**

Calculate F-statistics from ANOVA:



At 5% significant level, we can conclude that there is a significant difference overall in US Domestic Revenues between different Genres.



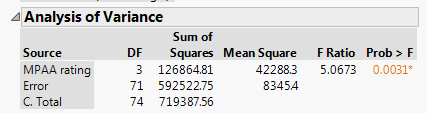
From the above data, significantly difference pairs of Genres are:

|  |  |
| --- | --- |
| Genre1 | Genre2 |
| Thriller | Comedy |
| Fantasy | Comedy |
| Thriller | Biography |
| Fantasy | Biography |
| Thriller | Drama |
| Fantasy | Drama |
| Fantasy | Musical |
| Fantasy | War |
| Romance | Comedy |
| Romance | Biography |
| Romance | Drama |

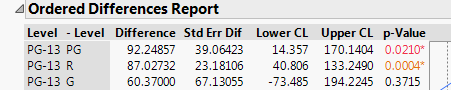
c. Determine whether there is a statistically significant difference overall in US DomesticRevenues between different MPAA Ratings. Which Ratings, if any, are statistically different  
from each other in terms of US Domestic Revenues?

**ANS:**

Calculate F-statistics from ANOVA:



At 5% significant level, we can conclude that there is a significant difference overall in US Domestic Revenues between different MPAA Ratings.



From the above data, significantly difference pairs of MPAA Ratings are:

|  |  |
| --- | --- |
| MPAA Rating1 | MPAA Rating2 |
| PG-13 | PG |
| PG-13 | R |

d. Give a general interpretation of your final analysis.

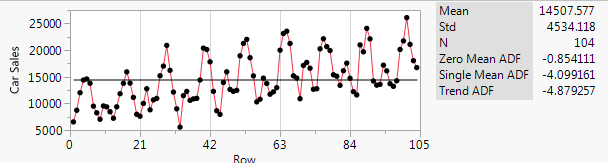
**ANS:**

* In general, there is a significant difference overall in US Domestic Revenues between different Genres. However, a different in some pairs of Genres is not conclusive and all Genres do not have completely different US Domestic Revenues.
* US Domestic Revenues of PG-13 films are significantly different from PG and R films’ US Domestic Revenues. The different in US Domestic Revenues for other pairs of MPAA Rating are not so apparent.

III.

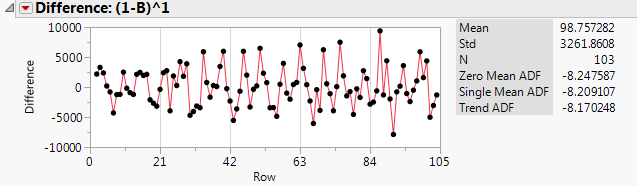
a. Is the *Car Sales* variable stationary? How do you know (statistically)? If it is *not* stationary,  
what needs to be done to make it stationary? Justify your answer

**ANS:**



Car Sales is not stationary, as shown by the Dickey-Fuller test of no constant term (Zero Mean ADF). It does not pass at either 1% or 5% significant threshold.

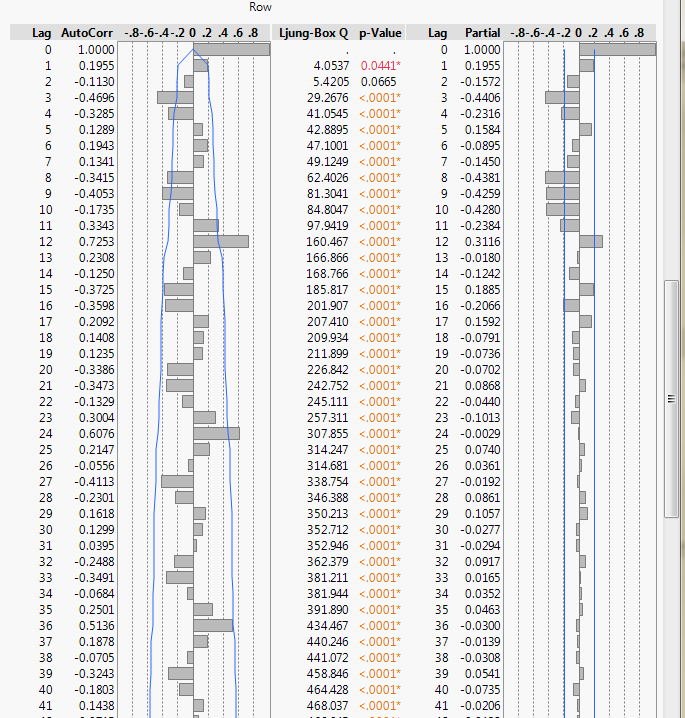
To make it stationary, take a difference of 1 lag of the series. Now, all Dickey-Fuller tests indicate that the series is stationary at 1% significant level.



b. Does there seem to be a seasonal, or cyclical, pattern to the data? Justify your answer

**ANS:**

Yes, continue from a., since the data come from monthly observation, we look at the significant spikes for ACF and PACF plots for every 12 lags. ACF plot suggests that up to 3 periods of seasonal MA term should be included. PACF plot suggests that up to 1 period of seasonal AR term should be included.

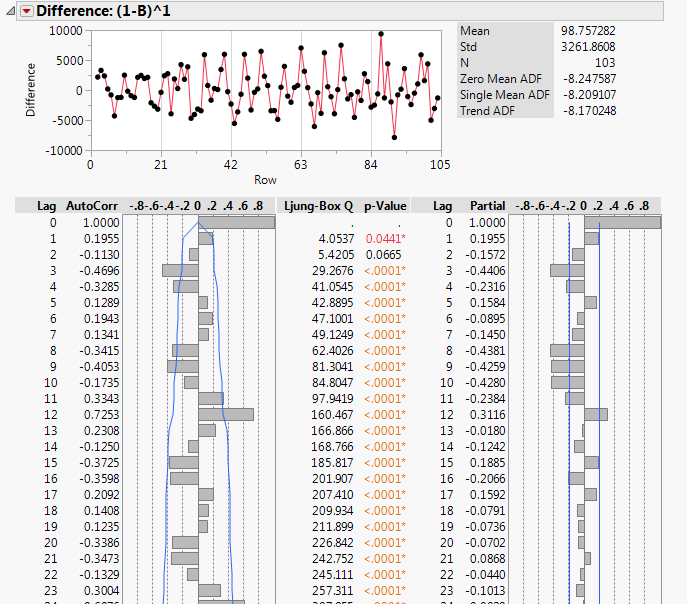


c. Using your vast knowledge of *Time Series* analysis, create a solid model to predict *Car  
Sales* for the next 4 months (I actually have the true values and can compare them to your  
estimates!), including 95% intervals.

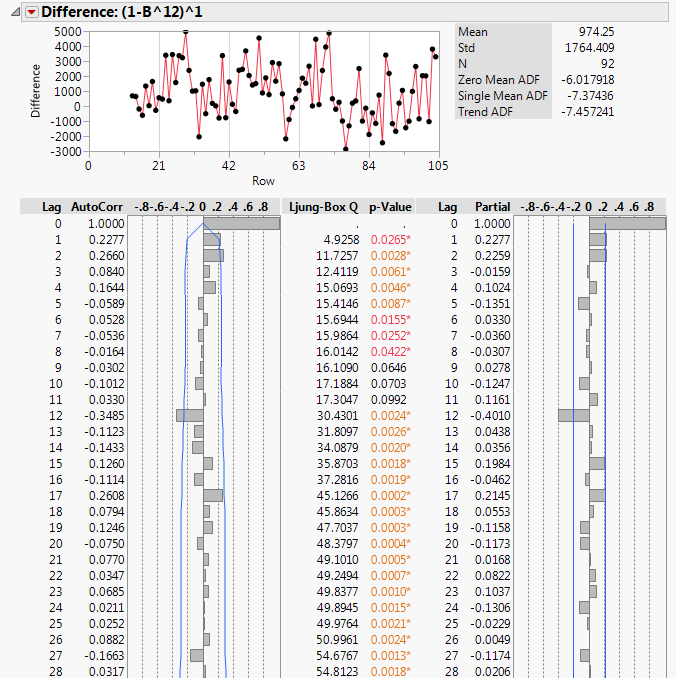
**ANS:**

I look at the 3 versions of differences:

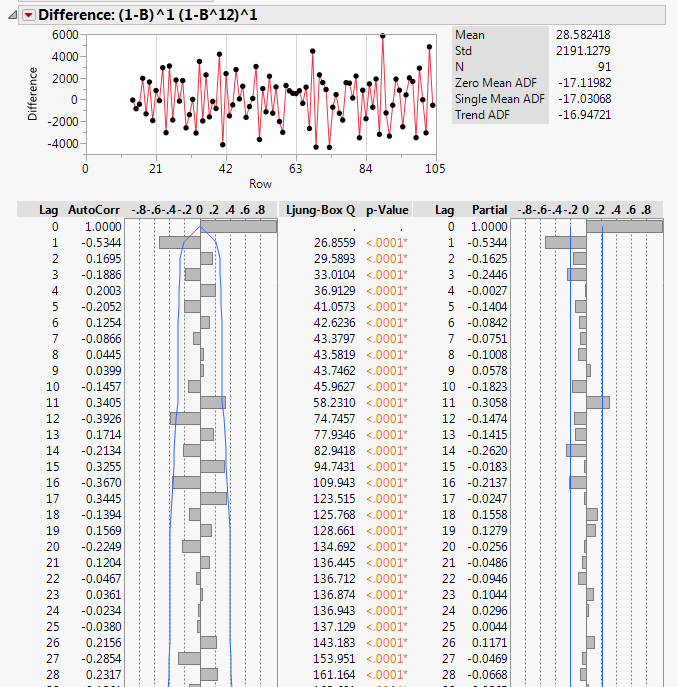
1. Non-Seasonal 1st difference



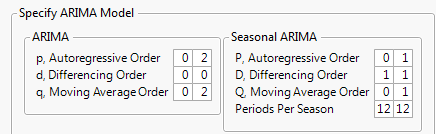
1. Seasonal 1st difference



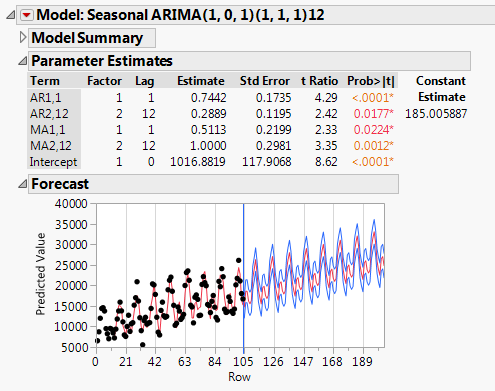
1. Both non-seasonal and seasonal 1st difference



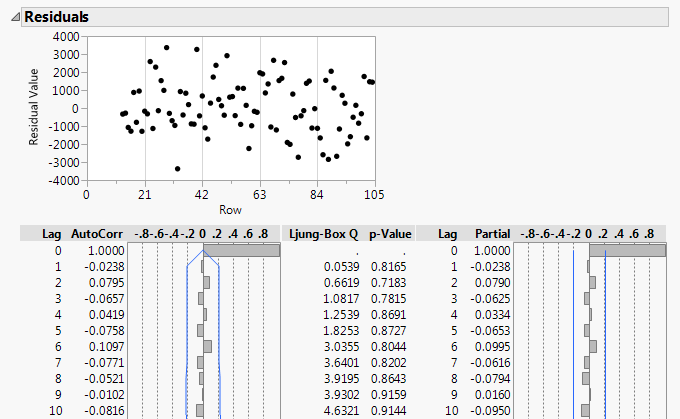
All are statistically stationary but it is the 2nd version (1st difference on just seasonal part) that yields the most compelling plot (ACF and PACF plot die down pretty fast and not too much random spikes). So, I decided to run the ARIMA group model based on the suggestion of the 2nd version:



Among the top models, I select the 2nd model, ARIMA(1,0,1)x(1,1,1), because all parameters are significant(at 5% threshold), prediction error, r-square and AIC are not so much different from the top model.



I don’t see any problems in the residual plot and it appears to be stationary.



This is the prediction of Car Sales for the next 4 months:

|  |  |  |
| --- | --- | --- |
| Prediction | Upper 95% | Lower 95% |
| 14753.447665 | 17445.223906 | 12061.671424 |
| 18576.585669 | 21340.341807 | 15812.82953 |
| 18579.219469 | 21382.040037 | 15776.398902 |
| 16254.336129 | 19078.556556 | 13430.115702 |