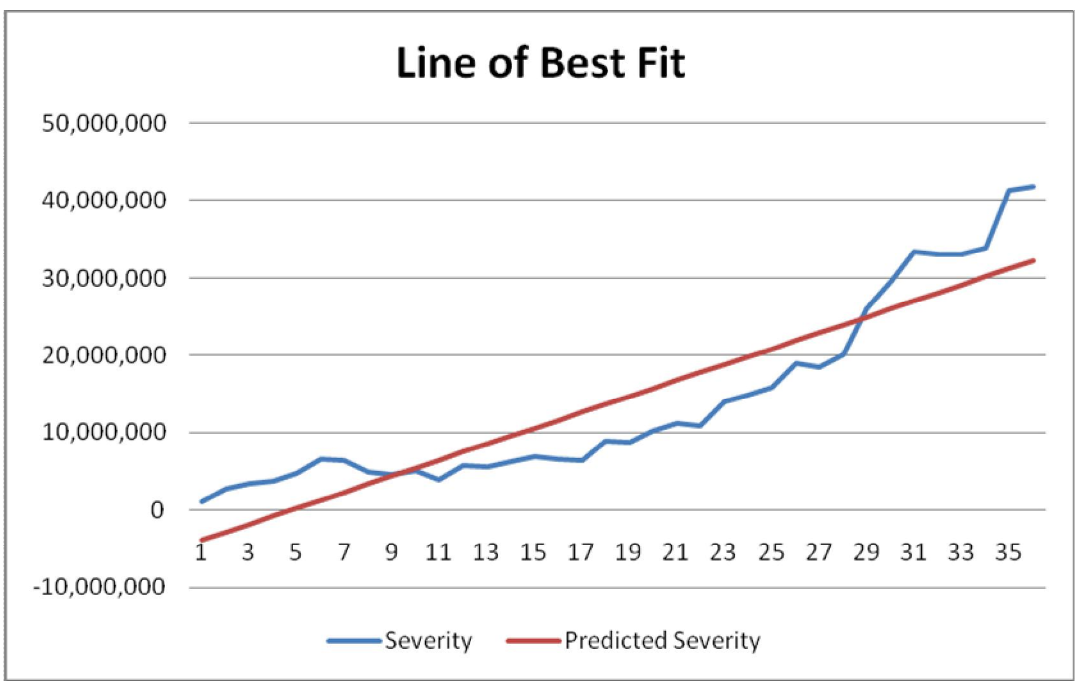
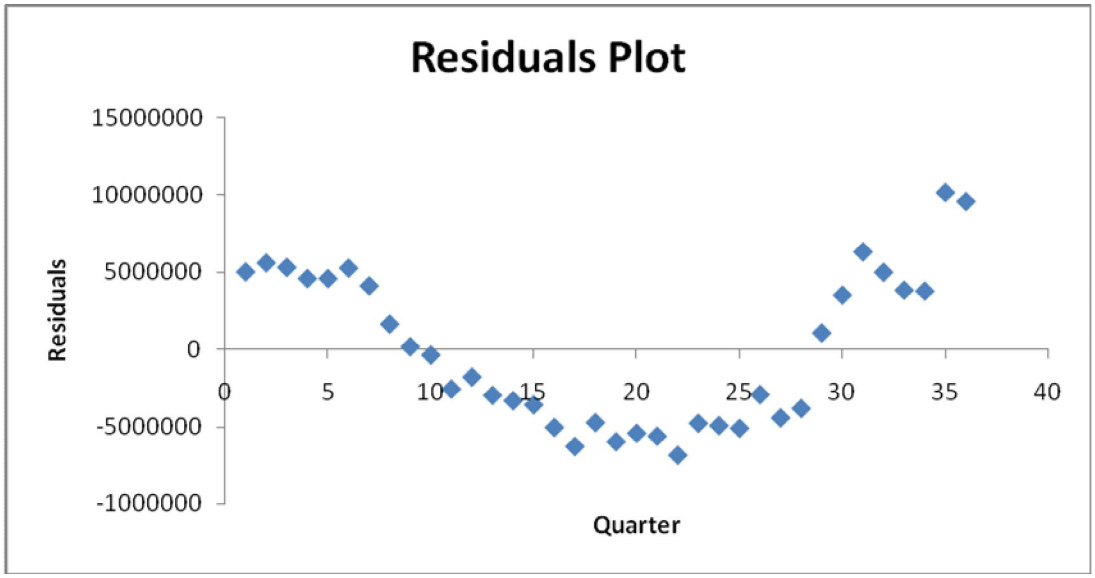
REGRESSION ANALYSIS – STUDENT PROJECT

Claim Severity Model

Regression analysis is a very useful tool and has widespread applications. In this project I applied regression analysis to NJI Life countrywide claim severity data of their health insurance business. They are a relatively new company and their data is available only for the last nine years (i.e. 36 quarters). I take time in quarters as an independent variable (X) and claims severity (Y) as a dependent variable.

Using the regression analysis tools in Excel, I fit a line of best fit with other statistical data (R^2, t-test, residuals and f-test). The fitted regression equation is Y = -4906720.88 + 1031947.86*X + error. As f-test and t-test have the same interpretation for the two variable regression analysis, so I first checked the f-test, to see if there is any significant statistical relationship between my two variables (X: Time in quarters, Y: Claim Severity). An f- value of 162.59 is very significant for my data set and t-values of -2.85 for intercept and 12.75 for x are also significant at 95% level of confidence. All of the p-values related to f and t statistics are also significant, reiterating that line of best fit is suitably fitted. The R^2 value of 0.827 and an adjusted R^2 of 0.821 are quite acceptable. Line of best fit plot and residual plot for my regression equation are given below:

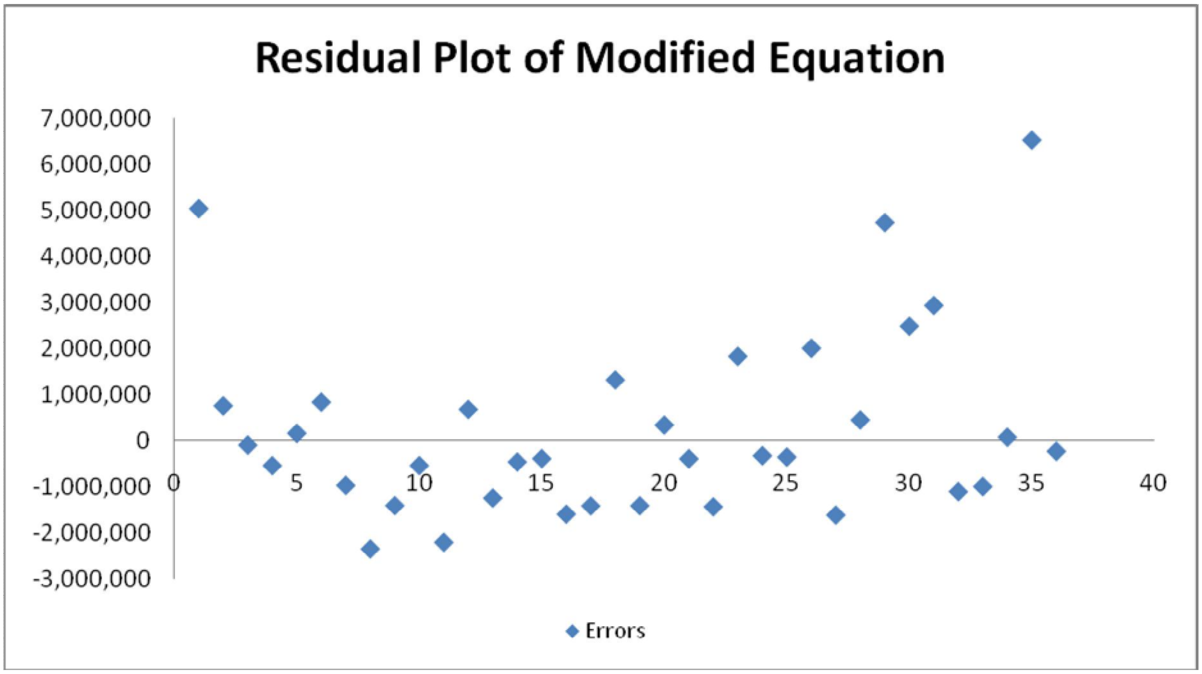


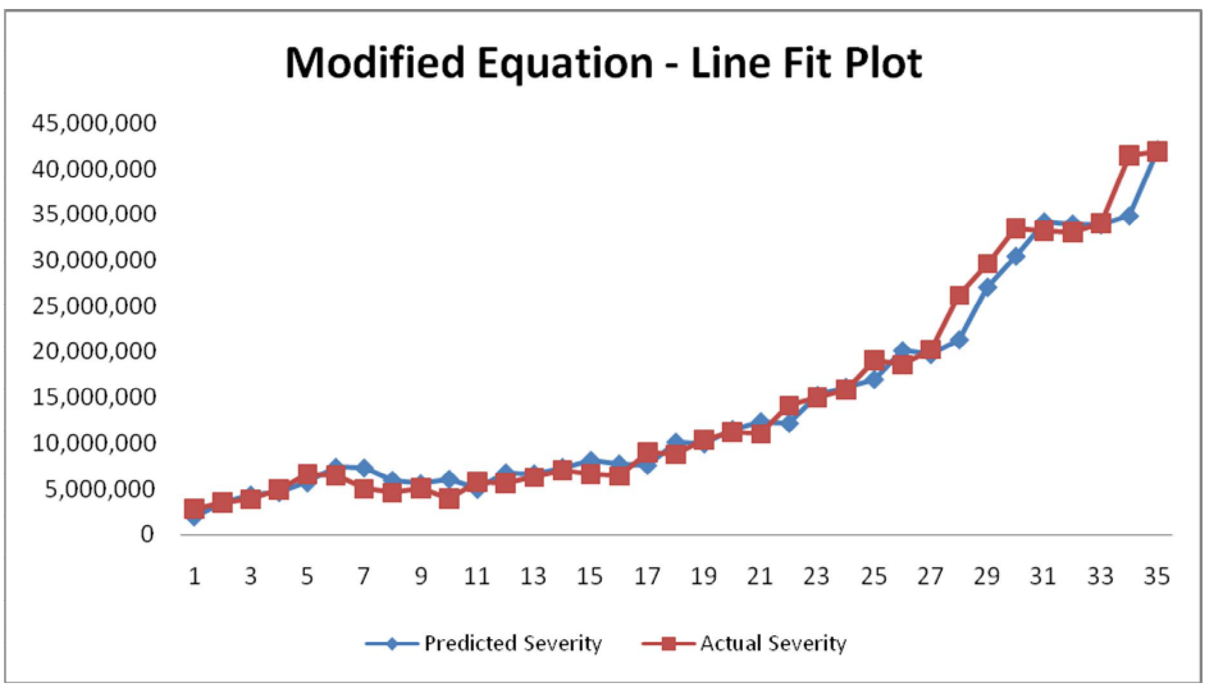


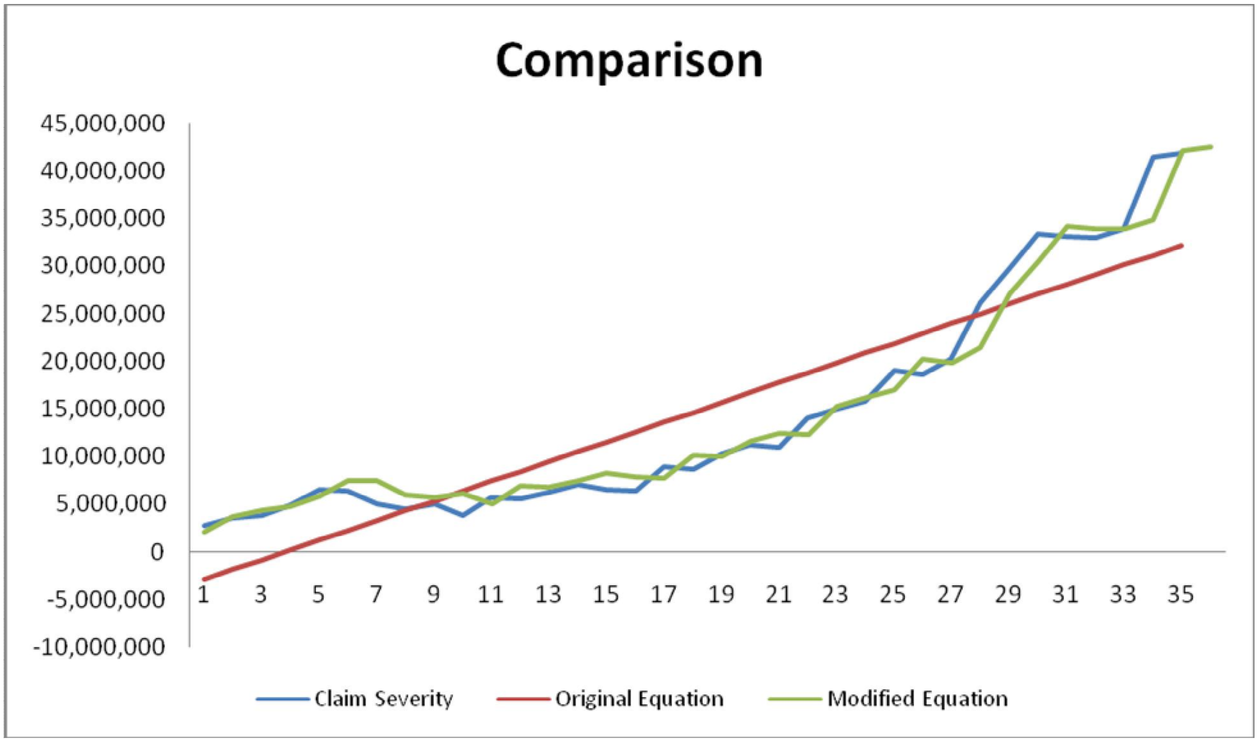
Residuals plot is clearly indicating the presence of serial correlation as the series of positive errors then negative errors and then positive errors is an indication of serial correlation. To test serial correlation, I performed the Durbin Watson test on the data (see Cell D24:E63) and the resultant value of 0.1404 strongly indicates the presence of positive serial correlation.

The presence of serial correlation usually indicates that we may have missed an important explanatory variable. To eliminate serial correlation I decided to add one more explanatory variable. I thought that addition of the previous error term would make sense, as if last data point has positive error, then this one should also have a positive error and vice versa. My modified regression equation may then be expressed as Y = -4906720.88 + 1031947.86*X + T*(Previous error term) + error.

Numerical value of 'T' has to be between 0 and 1, as it is the weight given to the previous error term, I calculate T by setting up a grid and plugging the values 0.1, 0.2, 0.3 …. 0.9, to find which value of 'T' would give us the least Error Sum of Squared (see Orange section in sheet Final Model L3:AC43), T= 0.9 gives us the lowest ESS i.e. 149,165,541,074,966, substantially less than the ESS of original equation i.e. 865,139,948,726,934. Then I thought that Excel's Solver Add in may give me the exact value of T with the lowest ESS. Using the solver Add in I got the value of ***T=0.964653238*** (see cells G3:H43). My final regression equation is **Y = -4906720.88 + 1031947.86*X + 0.964653238*(Previous error term) + error**. Now I perform some statistical tests to check the validity of my model. My main issue is serial correlation. I performed the Durbin Watson test again on my modified model and got the value of 1.57 (see cells I4:J46), much better than my original equation and within the indicating range (1.5 – 2.5) for no serial correlation. R^2 of 97.08% and an Adjusted R^2 of 97% strongly suggest that most of the variation is explained by my model and is a much better fit. An F test value of 549.09 also suggests that my model is a very good fit. Residual plot also shows no sign of serial correlation and is given below:


Residual Plot of Modified Equation
7,000,000
6,000,000
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
0
-1,000,000
-2,000,000
-3,000,000
0
5
10
15
20
25
30
35
40
Errors


Modified Equation - Line Fit Plot
45,000,000
40,000,000
35,000,000
30,000,000
25,000,000
20,000,000
15,000,000
10,000,000
5,000,000
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
Predicted Severity
Actual Severity



My predicted and forecasted results from the modified equation are

Claim Paid Quarter	Claim Severity	Predicted Severity	Percentage Difference
29	26,101,169	21,365,334.36	18.14%
30	29,583,677	27,094,892.82	8.41%
31	33,430,046	30,490,781.45	8.79%
32	33,137,795	34,237,669.78	-3.32%
33	33,002,409	33,992,224.92	-3.00%
34	33,977,790	33,898,100.39	0.23%
35	41,399,784	34,875,480.85	15.76%
36	41,848,772	42,071,607.41	-0.53%

Forecasted for 1st Quarter of 2009	
37	**42,541,201**