Rokhsona Pervin

Project: Predictive model of Employee performance Evaluation Score

Introduction:

In our Project of "Performance Evaluations", the employee's performance scores are calculated based on some subjective and some objective scores. The employee assessment criteria include adaptability, accountability, communication, decision-making skills, training and development, attendance, sensitivity to cost effectiveness etc. However, in the employee records, 10 variables, of which some are numerical and some categorical, are collected.

In this our goals are to select appropriate predictor variables to predict the evaluations scores. In regression analysis course throughout the semester the rules and techniques for regression analysis we have learned will be used to find an appropriate regression model for predicting the evaluation scores of employees.

Description of variables:

A short description of the variables is presented below:

The **response variable** is Score, which is numerical and continuous.

The Predictor variables are as follows:

Name	Type
Lag	Numerical and continuous
Emp@eoy	Numerical and continuous
Pos@eoy	Numerical and continuous
Type	Categorical: 0= 3 month evaluation 1=annual evaluation
Grade	Categorical : 1 =Clerical employee 2=Skilled employee 3= Supervisory 4= Management
Educ	Numerical and continuous
Emp@eval	Numerical and continuous
Pos@eval	Numerical and continuous
Whether employed or not	Categorical: 0=Unemployed 1=Employed

In the original dataset, grade was given in the following way: 1-7 scale are clerical employee, 8-9 scale are skilled employee, 10-12 are supervisory, and 13-17 are management. We converted this variable as a categorical, as is descripted in the above variable description table. Since this variable categorical with four levels, we have created three nominal variables using four levels. Also, we have created another variable" whether employed or not" depending on information given the performance Evaluations report.

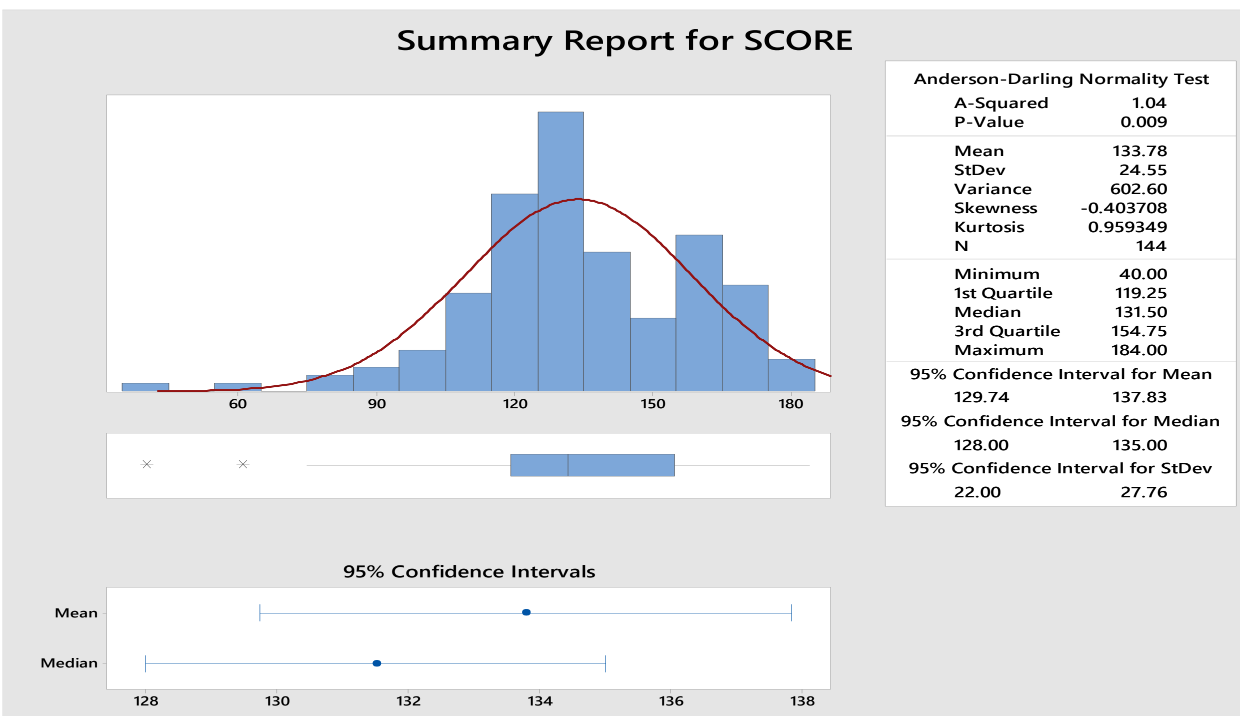
Part 1

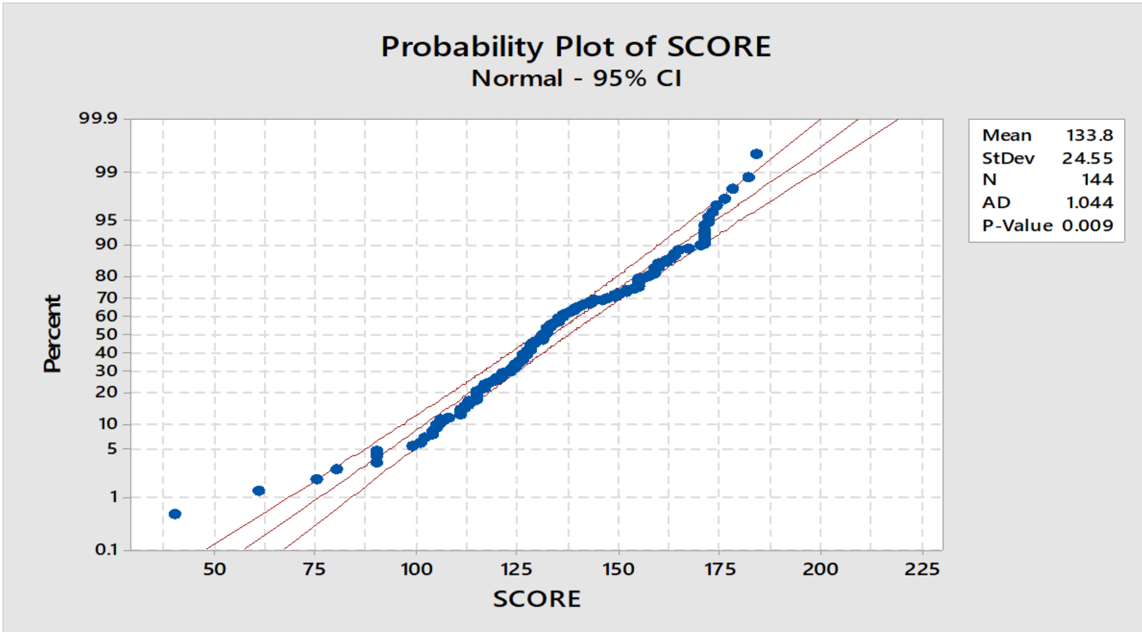
In the analysis part, first we analyzed the response variable- score in order to find some important features of it. The summary statistic along with the graphical presentation of this variable shows that the minimum score is 40 and maximum score 184. The range of data is 144. The data are widely spread. The histogram shows that the score is approximately symmetrical with slightly longer left tail, unimodal and highly spread.

The boxplot also shows the same results. The data between first quartile and median is less spread compare to within the data between median and third quartile.

The mean score for employee performance is 133.78 and standard deviation of 24.55.

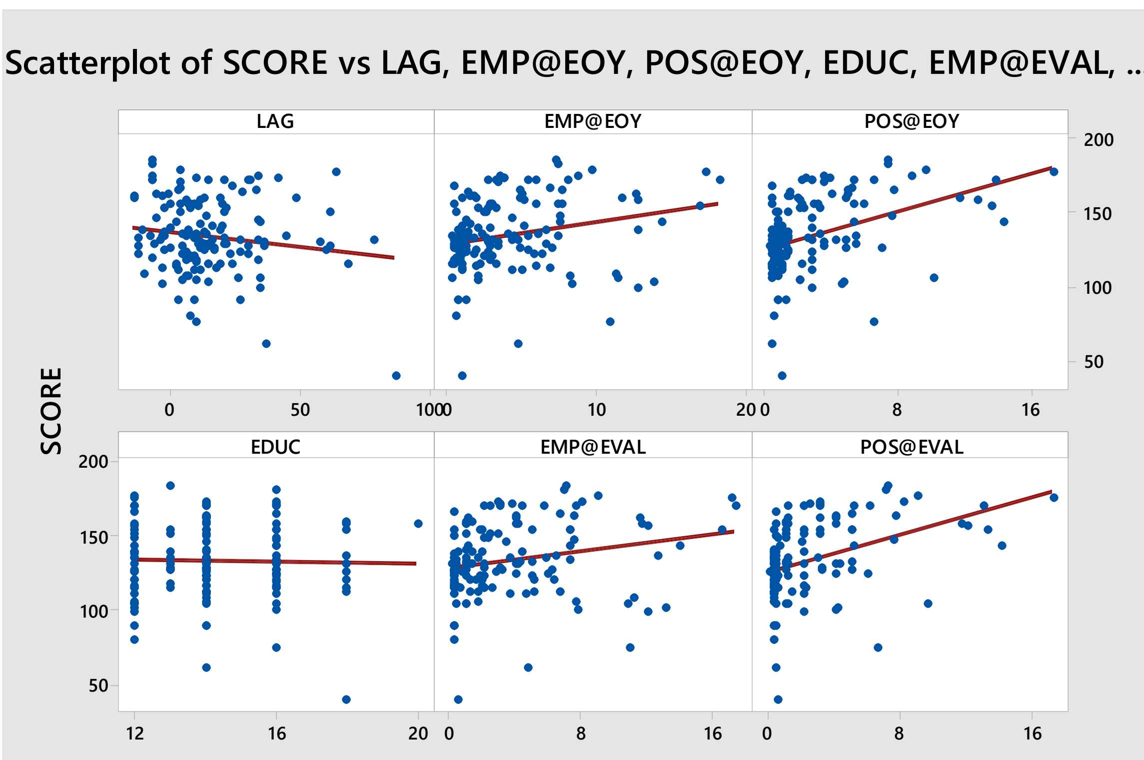
The Anderson darling test shows that the score variable does not follow normal distribution. The normal probability plot also shows that the score variable is not normal. There are some low scores that are considered to be outliers.





Part 2

In order to find if the response variable-score and the potential independent variable are related, we created scatter plot.



The scatter plot shows that the evaluation score is negatively related with the lag. The more gap between the evaluation date and the date on which the results are discussed with employee, the score tends to be low. The relationship appears to be very low.

On the other hands, the score variable is positively correlated with the variable em@eoy, pos@woy, emp@eval and pos@eval. However, the education variable seems not to be related with the score variable.

Correlation: SCORE, LAG, EMP@EOY, POS@EOY, EDUC, EMP@EVAL, POS@EVAL

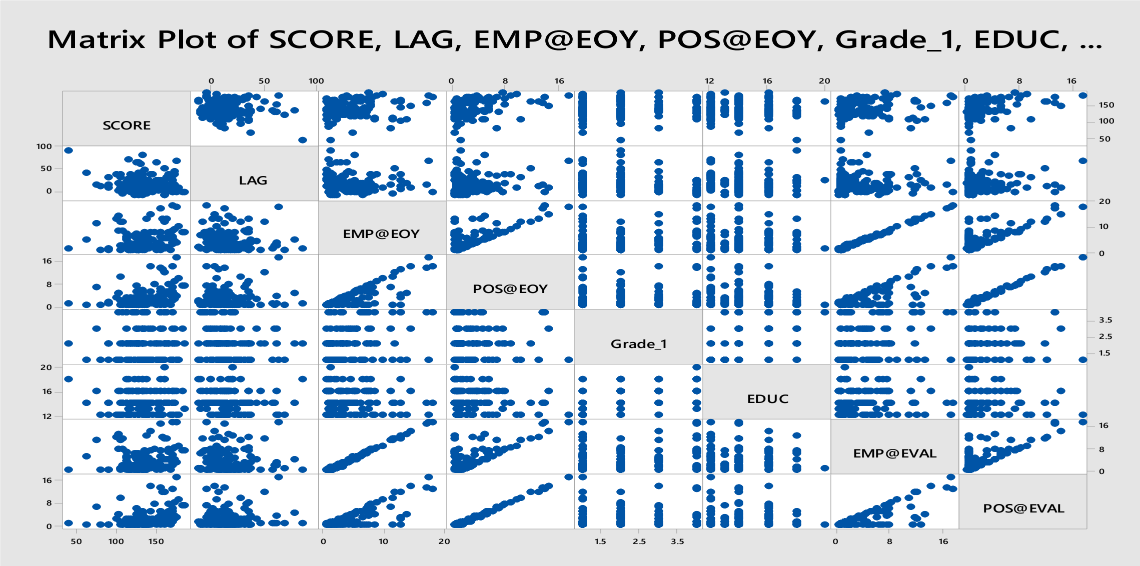
	SCORE	LAG	EMP@EOY	POS@EOY	EDUC	EMP@EVAL
LAG	-0.151					
	0.071					
EMP@EOY	0.237	-0.059				
	0.004	0.479				
POS@EOY	0.414	0.048	0.786			
	0.000	0.567	0.000			
EDUC	-0.013	-0.120	-0.038	-0.080		
	0.875	0.152	0.654	0.340		
EMP@EVAL	0.228	-0.038	0.996	0.776	-0.022	
	0.006	0.648	0.000	0.000	0.797	
POS@EVAL	0.406	0.074	0.787	0.995	-0.061	0.786
	0.000	0.379	0.000	0.000	0.466	0.000

Cell Contents: Pearson correlation
P-Value

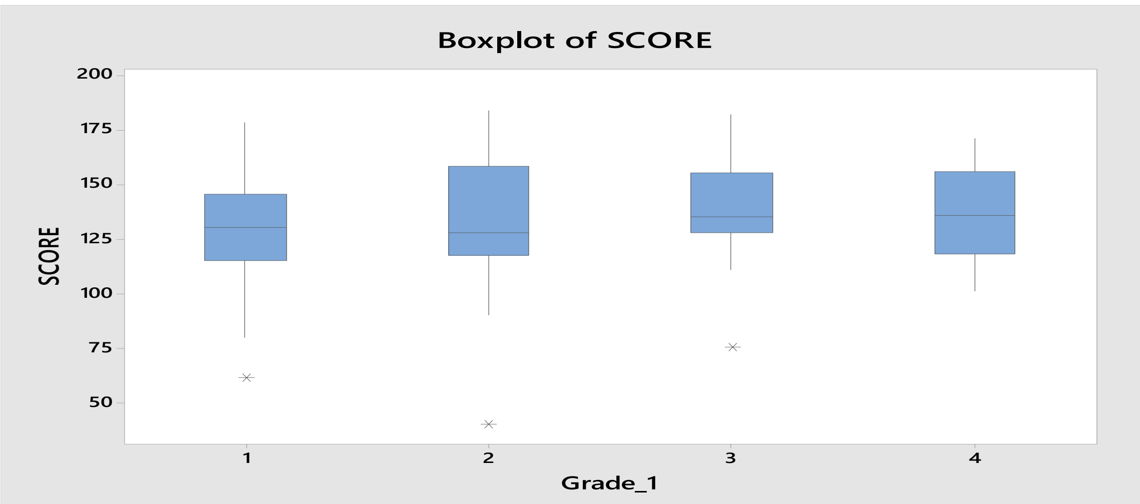
The correlation matrix shows that the score variable is not statistically related with lag and education variable. However, the other independent variables such as emp@eoy, emp@eval, pos@eoy, pos@eval are statistically positively related . The significant correlations are below 0.5.

Therefore, we can say that the important independent variable -emp@eoy, emp@eval, pos@eoy, pos@eval can be used to predict the score variable.

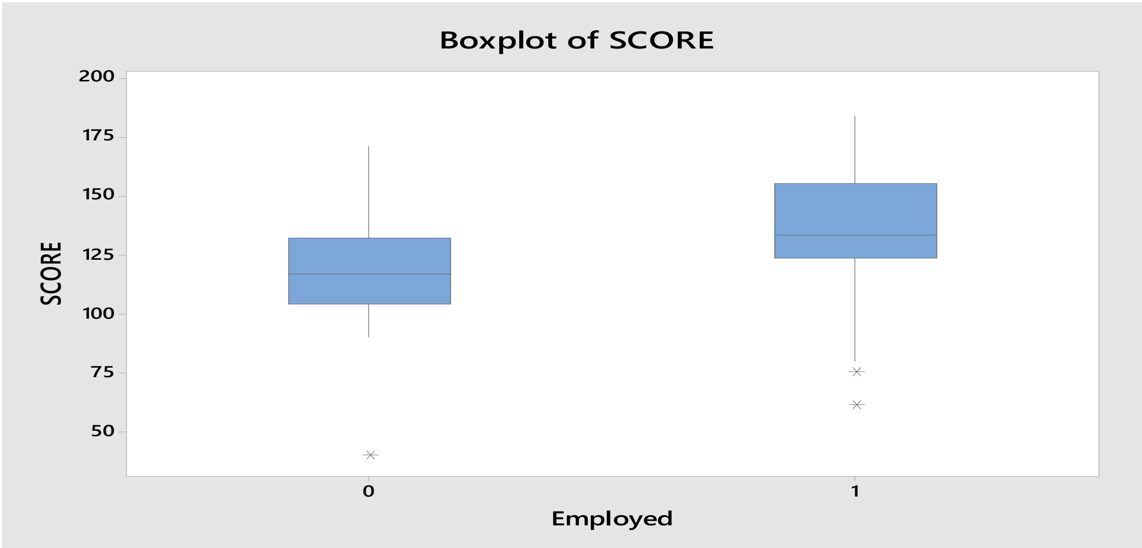
In addition, the correlation matrix reveals that there is a strong multicolinearity among the following predictor variables: Emp@eoy, Pos@eoy, emp@eval, Pos@eval, emp@eval. As is seen in the matrix plot.



The following graph shows how the mean score changes as the employee grade moves from low to high. It is clear that there is a positive relationship between score and grade.



The most interesting feature is that the employee left the company has the low score compared to the employees who are still working at the company. The boxplot shows the clear difference in mean score of two different type of employee.



Best Subsets Regression: SCORE versus LAG, EMP@EOY, ...

Response is SCORE

```
                                             S
                                             k
                                             i
                                             l
                                             l
                                             e
                                             d S
                                               u
                                             e p
                                         E   C m e E
                                       E P M   l p r m
                                       M O P   a l v p
                                       P S @   r o i l
                                       @ @ E T i y s o E
                                     L E E V Y c e o y D
          R-Sq    R-Sq  Mallows        A O O A P a s r e U
Vars  R-Sq  (adj)  (pred)      Cp      S  G Y Y L E l s y d C
   1  21.0   20.5    18.8    27.8  21.891        X
   1  17.1   16.6    14.9    36.1  22.423      X
   2  26.5   25.5    22.8    18.1  21.188        X       X
   2  23.9   22.8    20.9    23.8  21.573      X   X
   3  29.7   28.2    25.2    13.3  20.797      X   X       X
   3  27.9   26.3    22.3    17.3  21.069  X       X       X
   4  32.0   30.0    25.8    10.6  20.535      X X X       X
   4  31.8   29.9    25.6    10.9  20.557    X X   X       X
   5  34.7   32.3    26.9     6.9  20.199  X X X   X       X
   5  34.6   32.3    26.9     6.9  20.203  X   X X X       X
   6  35.2   32.4    26.4     7.6  20.182  X   X X X X     X
   6  35.2   32.4    26.4     7.7  20.187  X X X   X X     X
   7  35.9   32.6    26.3     8.2  20.148  X   X X X X X   X
   7  35.9   32.6    26.3     8.2  20.149  X X X   X X X   X
   8  36.6   32.9    26.1     8.6  20.110  X   X X X X X X X
   8  36.6   32.9    26.1     8.7  20.113  X X X   X X X X X
   9  37.4   33.2    25.2     9.0  20.062  X X X   X X X X X X
   9  37.3   33.1    25.0     9.2  20.077  X   X X X X X X X X
```

10 37.4 32.7 23.8 11.0 20.137 X X X X X X X X X X

After examining R^2, $R^2_{adj.}$, S and Mallow's C_p, it seems that variable Lag, pos@eoy, Emp@eval, Type, Clarical, Skilled employess, and employed seem good.

Other three method to select variables (Summary)

Approach	α	Variables in model
Stepwise	0.15	LAG, EMP@EVAL, POS@EOY, TYPE, EMPLOYED
Forward	0.25	LAG, EMP@EVAL, POS@EOY, TYPE, EMPLOYED
Backward	0.1	LAG, EMP@EOY, POS@EOY, TYPE, EMPLOYED

Minitab output for variable selection:

Method

Categorical predictor coding (1, 0)

Backward Elimination of Terms

α to remove = 0.1

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	29867	5973.4	14.64	0.000
LAG	1	2435	2435.0	5.97	0.016
EMP@EOY	1	2567	2567.2	6.29	0.013
POS@EOY	1	5766	5766.2	14.13	0.000
TYPE	1	3394	3394.0	8.32	0.005
Employed	1	5982	5981.8	14.66	0.000
Error	138	56305	408.0		
Total	143	86172			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
20.1992	34.66%	32.29%	26.95%

Method

Categorical predictor coding (1, 0)

Forward Selection of Terms

α to enter = 0.25

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	29848	5969.6	14.63	0.000
LAG	1	2293	2292.8	5.62	0.019
POS@EOY	1	5755	5755.2	14.10	0.000
EMP@EVAL	1	2548	2548.2	6.24	0.014
TYPE	1	3244	3244.2	7.95	0.006
Employed	1	6139	6138.7	15.04	0.000
Error	138	56324	408.1		
Total	143	86172			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
20.2026	34.64%	32.27%	26.89%

Method

Categorical predictor coding (1, 0)

Stepwise Selection of Terms

α to enter = 0.15, α to remove = 0.15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	29848	5969.6	14.63	0.000
LAG	1	2293	2292.8	5.62	0.019
POS@EOY	1	5755	5755.2	14.10	0.000
EMP@EVAL	1	2548	2548.2	6.24	0.014
TYPE	1	3244	3244.2	7.95	0.006
Employed	1	6139	6138.7	15.04	0.000
Error	138	56324	408.1		
Total	143	86172			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
20.2026	34.64%	32.27%	26.89%

Interpretation:

After examining R^2, $R^2_{adj.}$, S and Mallow's C_p, it seems that variable Lag, pos@eoy, Emp@eval, Type, Clarical, Skilled employess, and employed seem good. However.

We selected our independent variables Lag, Emp@eoy, Pos@eoy,Type and Employed From the backward selection method because its R sq is 34.66% more than the other selection procedure.

the R^2 value always increases as one adds more variables into the model. The important thing is to select a model with a reasonably relatively high R^2 value and with a reasonable number of predictor variables.

Part 3

Results for: Performance Evaluations.MTW

Regression Analysis: SCORE versus LAG, EMP@EOY, POS@EOY, TYPE, Employed

Method

Categorical predictor coding (1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	29867	5973.4	14.64	0.000
LAG	1	2435	2435.0	5.97	0.016
EMP@EOY	1	2567	2567.2	6.29	0.013
POS@EOY	1	5766	5766.2	14.13	0.000
TYPE	1	3394	3394.0	8.32	0.005
Employed	1	5982	5981.8	14.66	0.000
Error	138	56305	408.0		
Total	143	86172			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
20.1992	34.66%	32.29%	26.95%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	111.26	5.20	21.40	0.000	
LAG	-0.2281	0.0934	-2.44	0.016	1.06
EMP@EOY	-1.812	0.722	-2.51	0.013	2.76
POS@EOY	3.79	1.01	3.76	0.000	3.67
TYPE					
1	12.57	4.36	2.88	0.005	1.67
Employed					
1	19.20	5.02	3.83	0.000	1.02

Interpretation:

$$\mathrm{H}_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

vs. H_a: Not all β_j's are 0.

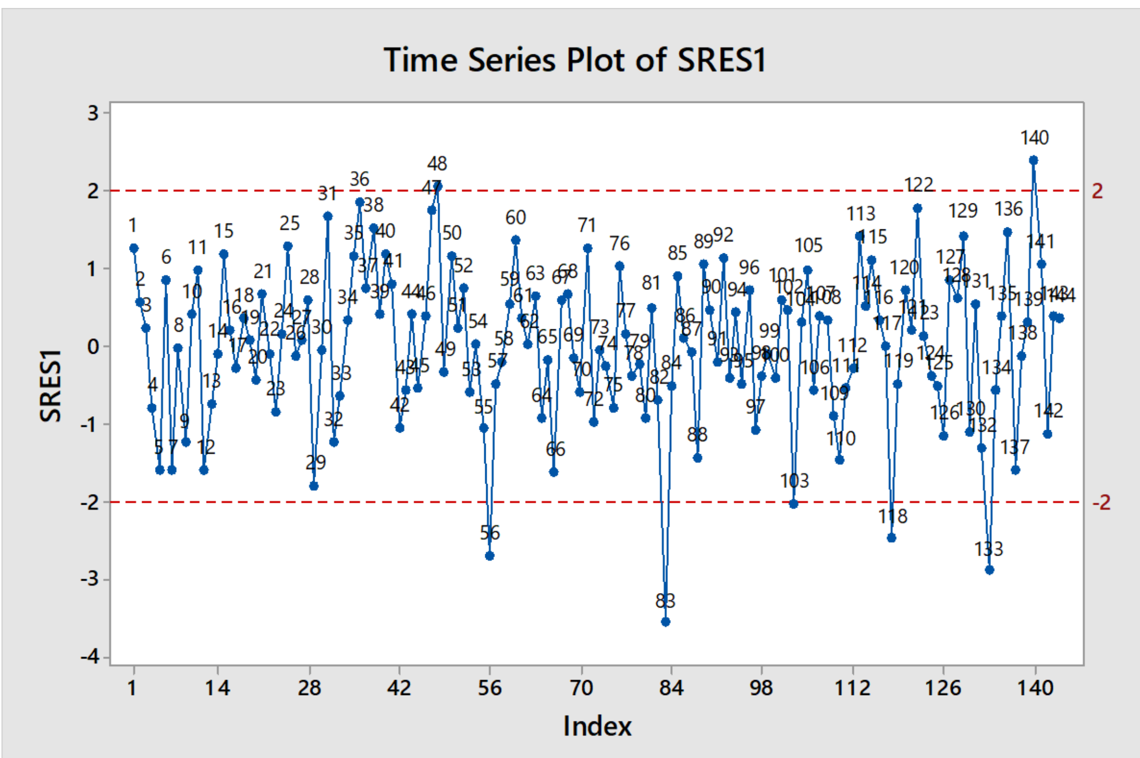
Since p-value is less than α=0.05, we reject the null hypothesis. We have sufficient evidence to conclude that at least one of the β_j's is not equal to zero.

Detecting Outliers:

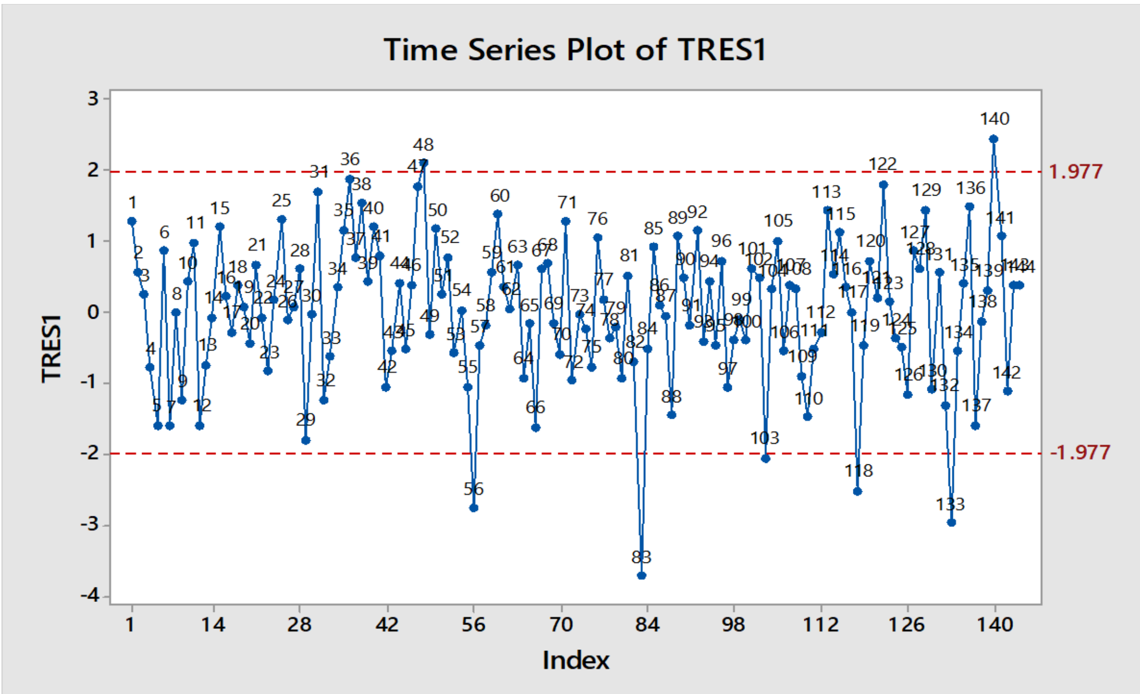
Potential outliers suggested by 5 different approaches

Approaches	Potential Outliers
Studentized Residual	# 140#56#88 #118 #133
Deleted Residual	# 140#56#88 #118 #133
Leverages	#5,#9,#16,#52,#,65,#84,#110,#112,#122,#124,#129,#133,#137
DFFITS	#5#9 #47 #56 #83 #100 #122#129#133##137,#140
COOK	#122#129#83#133#140#110

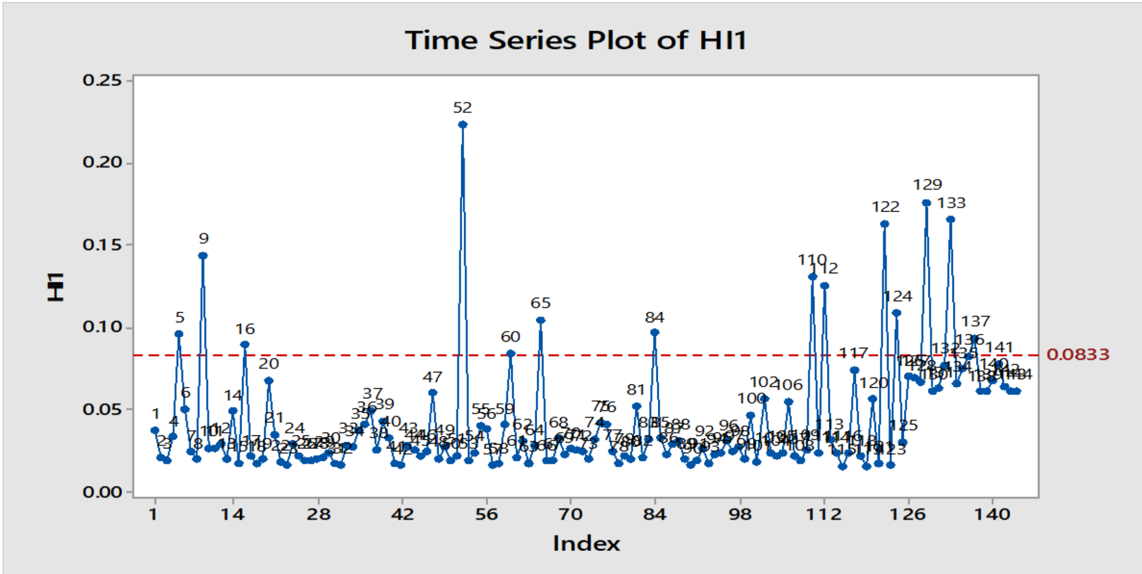
Studentized residual: If the normality assumption holds, the residuals (e_i's) should distribute approximately as a standard normal distribution. As a result, we would expect to see a studentized residual with an absolute value exceeding 2 only about 5% of the time. When we look at the graph of the studentized residuals we see that there are approximately 5 observations that exceed the threshold of ±2.



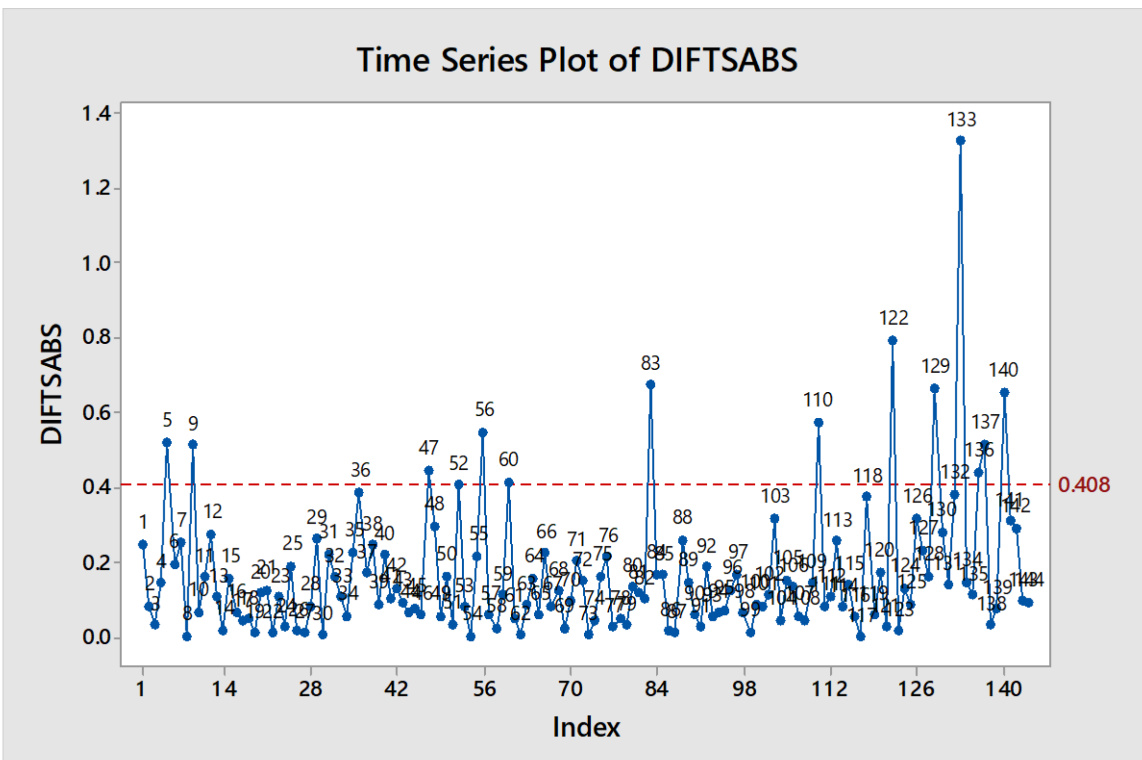
Deleted Residual: they distribution as a t distribution with $n-K-1$ degrees of freedom. Therefore, one can use the distribution to find the threshold. For instance, if we set the confidence level at 95%, then we can find $t_{144-5-1,2.5\%}$=1.977. So we have about 5 observations with a studentized deleted residual whose absolute value exceeds ± 1.977 and can be considered as possible outliers.



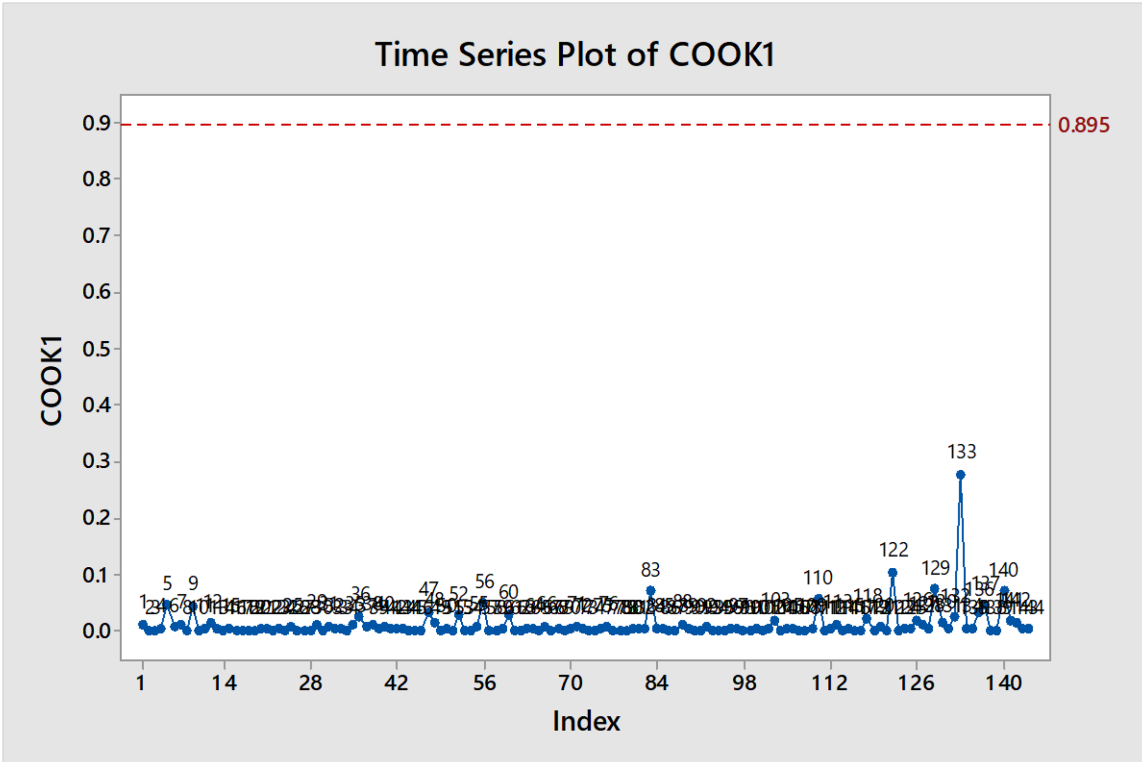
Leverages: Observations with high leverage could potentially drastically alter the regression analysis. Therefore, an observation is considered to be a high leverage point if its leverage exceeds twice of the average of all leverages $2(K+1)/n$. So, we have 13 observations exceeding our threshold of $\frac{2(5+1)}{144} = 0.0833$ and can be considered as possible outliers.



DFFITS: An observation will be considered influential on a single fitted value if $|DFFITS| > 2\sqrt{\frac{K+1}{n}} = 2\sqrt{\frac{6}{144}} =$ .408. Using this approach, we had about 11 observations that were above the threshold of .408 with observations 133 and 122 being the highest above our threshold.



COOK: An observation will be considering to be influential if its Cook's Distance exceeds $F(.50, K+1, n-K-1)=$ $F(.50, 6, 144-6)$=0.895. We found 5 outliers in this approach.



Based on these approaches for checking outliers, we deleted those outliers: 140,133, 56,129,88,110 and,122. We can see that our R-Square has improved from 34.66 %to 35.36%.

Correlations and Multicollinearity:

Multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. When we looked at the variance inflation factor (VIF) for the 9-variable model, we observed that no variables exceeded the thumb rule of 10. Therefore we can conclude there is no multicollinearity problem in our model and we can proceed with 9 variables model. However, High correlations between variables were indicated by the Pearson correlation matrix suggests a possible there are multicollinearity problem in our dataset.

Term	**Coefficient**	**SE Coeff**	**T-value**	**P- value**	**VIF**
Constant	111.26	5.20	21.40	0.00	
LAG	-0.22	0.09	-2.44	0.01	1.06
EMP@EOY	-1.81	0.72	-2.51	0.01	2.76
POS@EOY	3.79	1.01	3.76	0.00	3.67
TYPE	12.57	4.36	2.88	0.00	1.67
EMPLOYED	19.20	5.02	3.83	0.00	1.02

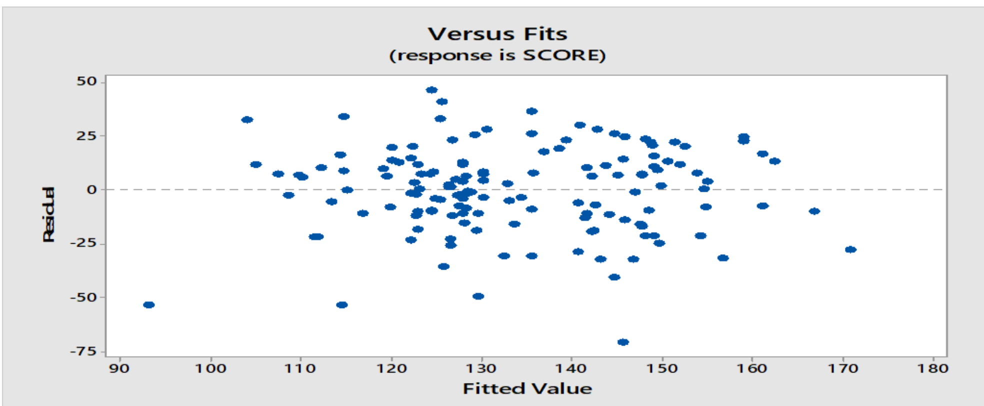
Correlation: SCORE, LAG, EMP@EOY, POS@EOY

	SCORE	LAG	EMP@EOY
LAG	-0.151		
	0.071		
EMP@EOY	0.237	-0.059	
	0.004	0.479	
POS@EOY	0.414	0.048	0.786
	0.000	0.567	0.000

Cell Contents: Pearson correlation
P-Value

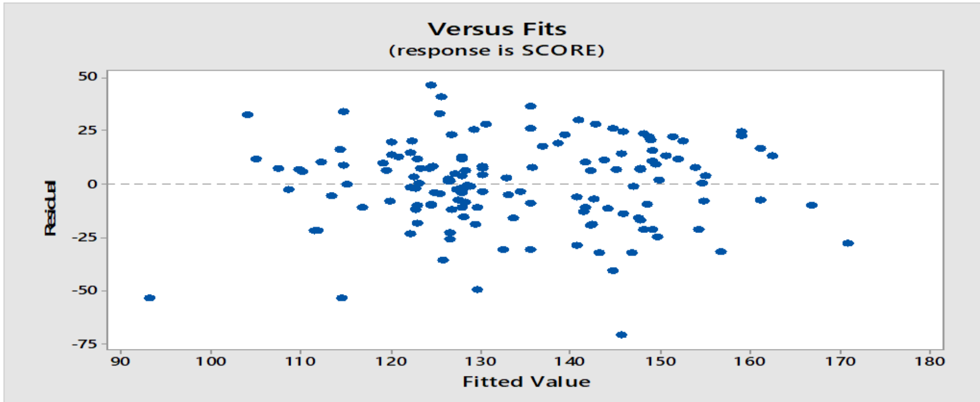
Assumption(before deleting the outlier):

Linear model assumption:

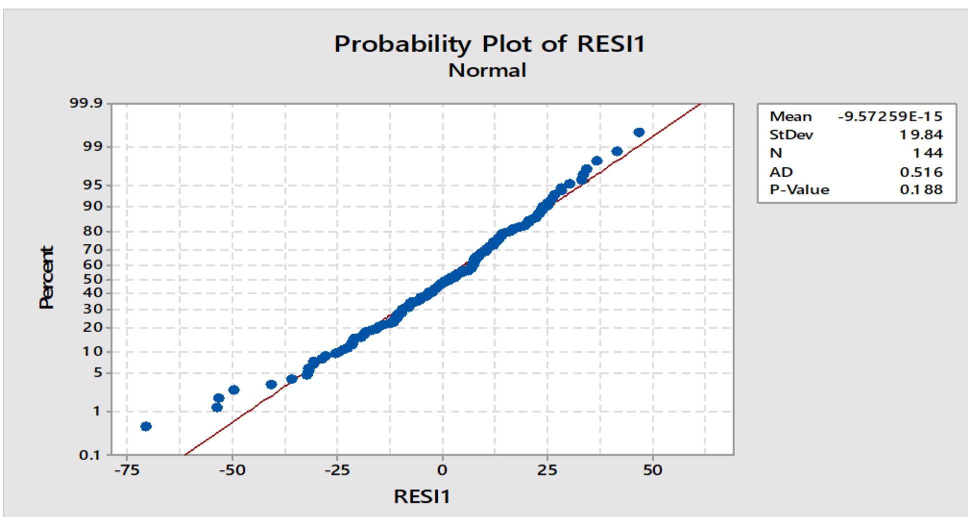


The residuals and the fitted values should be uncorrelated. When we plot two uncorrelated variables on a scatterplot there should not be any "non-random" ("abnormal") pattern observed on the chart. If it does, that should be an indication that the residuals and fitted values are correlated. This is directly linked to possible violation of the model assumption. Our residuals vs filled value indicates no violation of model assumption.

Constant variance assumption:



Our residual versus fitted values also seems hold constant variance assumption.



The residuals appear to fall on a straight line. Therefore, it appears that the normal distribution assumption holds.

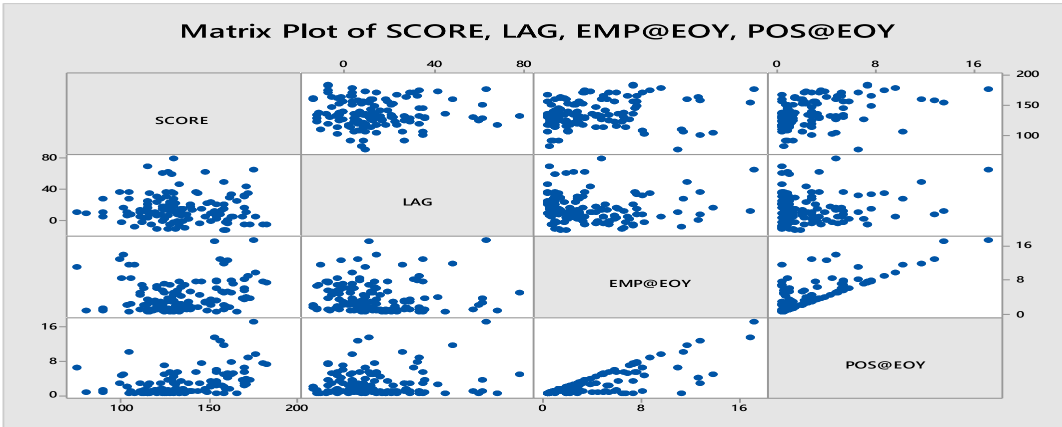
The normality test can be hypothesized as

$H_0: F(x)$ is normally distributed vs.

$H_a: F(x)$ is not normally distributed.

Interpretation: Since P-value › level of significance, therefore Null hypothesis is accepted. We see that it is normally distributed.

Examine the possibility of adding quadratic or interaction terms to the model.



We can see that there are no interaction or quadratic terms that we can include in our model.

The regression model after deleted influencial observations that we already detected

Regression Analysis: SCORE versus LAG, EMP@EOY, POS@EOY, TYPE, Employed

Method

Categorical predictor coding (1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	24413.7	4882.7	14.46	0.000
LAG	1	583.9	583.9	1.73	0.191
EMP@EOY	1	3210.5	3210.5	9.50	0.002
POS@EOY	1	5303.4	5303.4	15.70	0.000
TYPE	1	2736.1	2736.1	8.10	0.005
Employed	1	6108.9	6108.9	18.08	0.000
Error	131	44250.4	337.8		
Total	136	68664.2			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
18.3790	35.56%	33.10%	29.39%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	109.91	4.95	22.21	0.000	

LAG	-0.1222	0.0929	-1.31	0.191	1.10
EMP@EOY	-2.244	0.728	-3.08	0.002	2.75
POS@EOY	4.06	1.02	3.96	0.000	3.71
TYPE					
1	11.76	4.13	2.85	0.005	1.73
Employed					
1	21.08	4.96	4.25	0.000	1.03

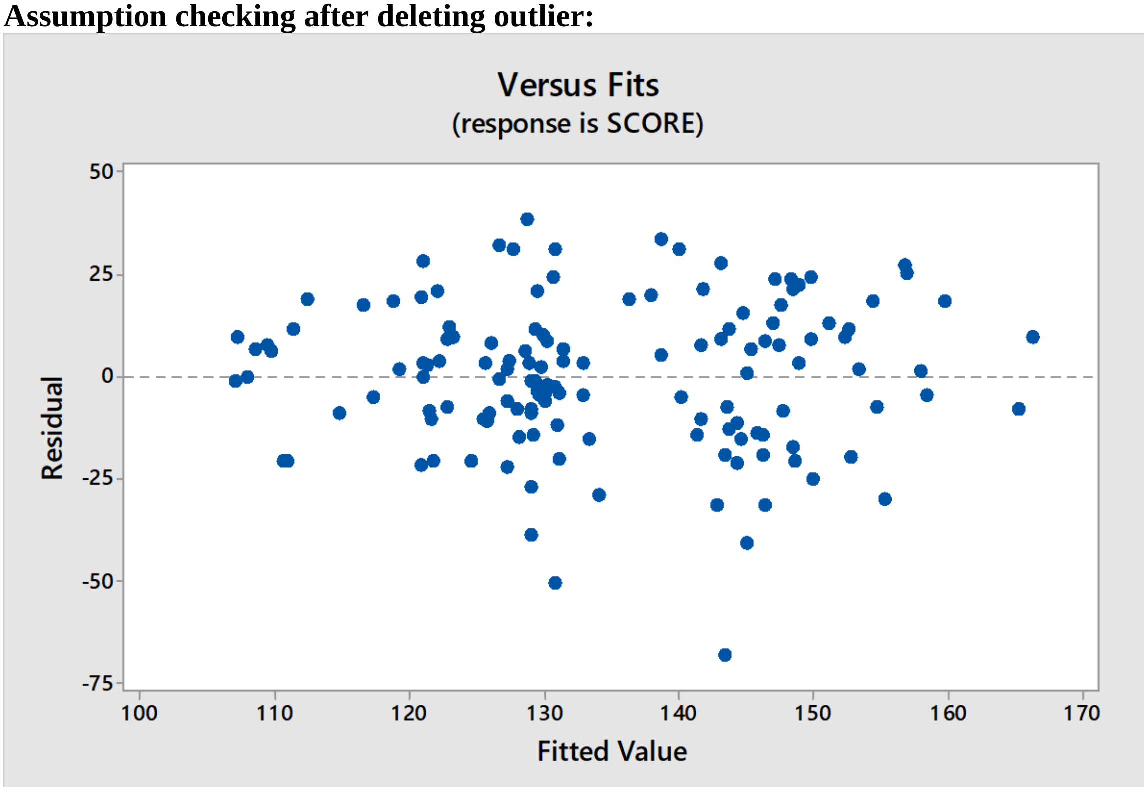
Multicollinearity:

Correlation: SCORE, LAG, EMP@EOY, POS@EOY

	SCORE	LAG	EMP@EOY
LAG	-0.023		
	0.515		
EMP@EOY	0.206	-0.020	
	0.000	0.571	
POS@EOY	0.406	0.106	0.797
	0.000	0.002	0.000

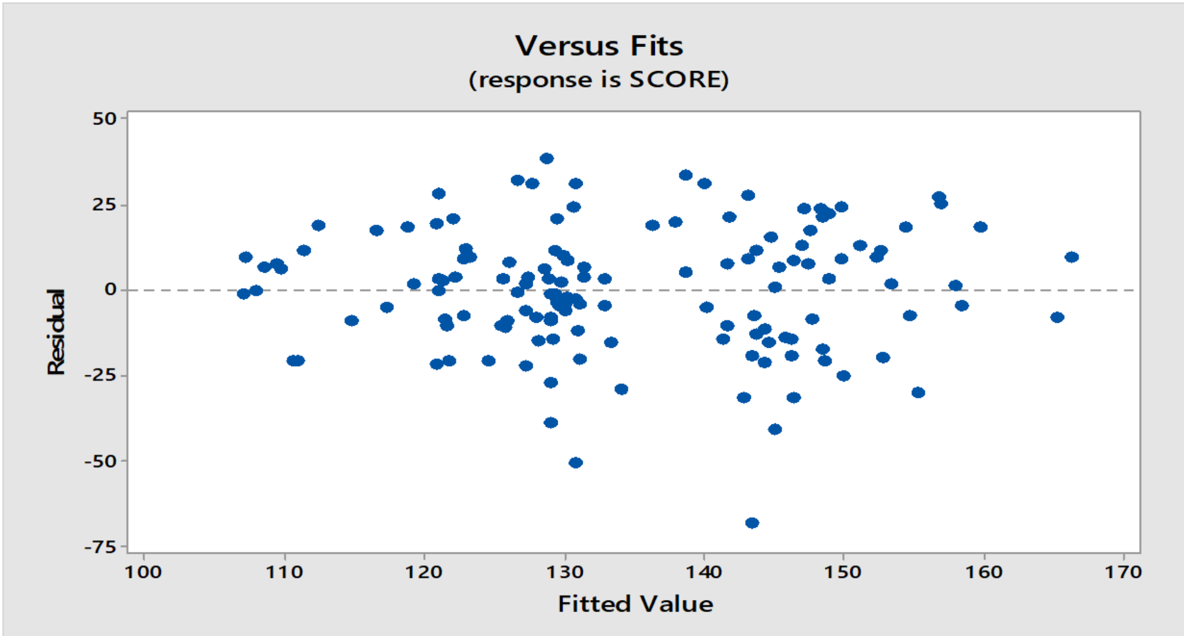
Cell Contents: Pearson correlation
P-Value

Assumption checking after deleting outlier:

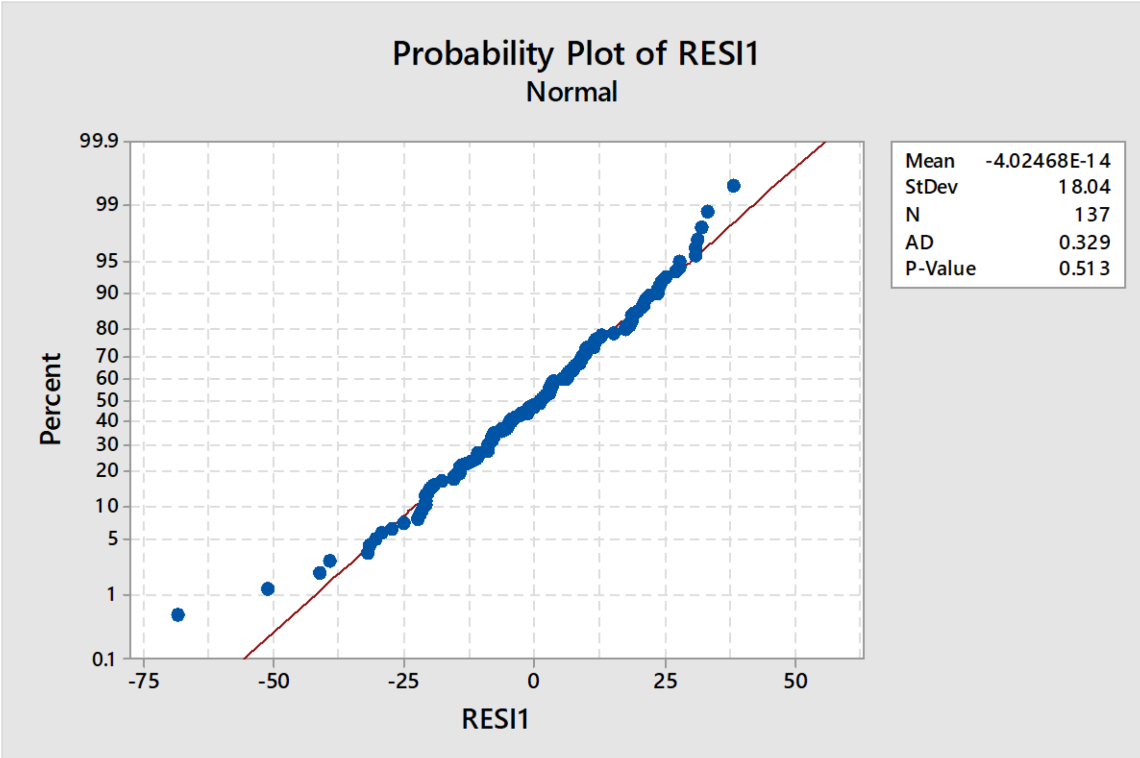


The residuals and the fitted values should be uncorrelated. When we plot two uncorrelated variables on a scatterplot there should not be any “non-random” (“abnormal”) pattern observed on the chart.

If it does, that should be an indication that the residuals and fitted values are correlated. This is directly linked to possible violation of the model assumption. Our residuals vs filled value indicates no violation of model assumption.



Our residual versus fitted values also seems hold constant variance assumption.



The residuals appear to fall on a straight line. Therefore, it appears that the normal distribution assumption holds.

The normality test can be hypothesized as

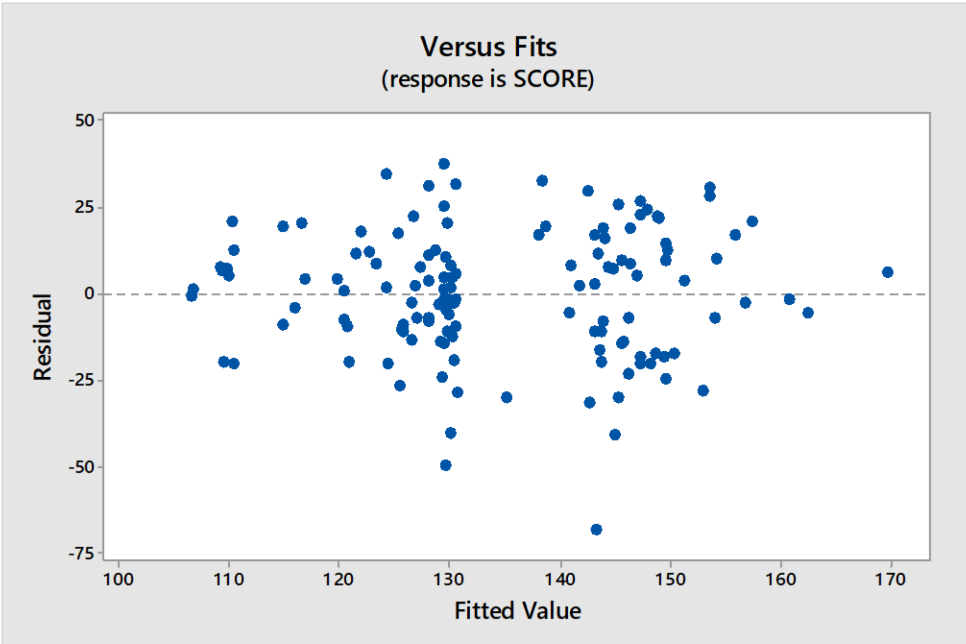
$H_0: F(x)$ is normally distributed vs.

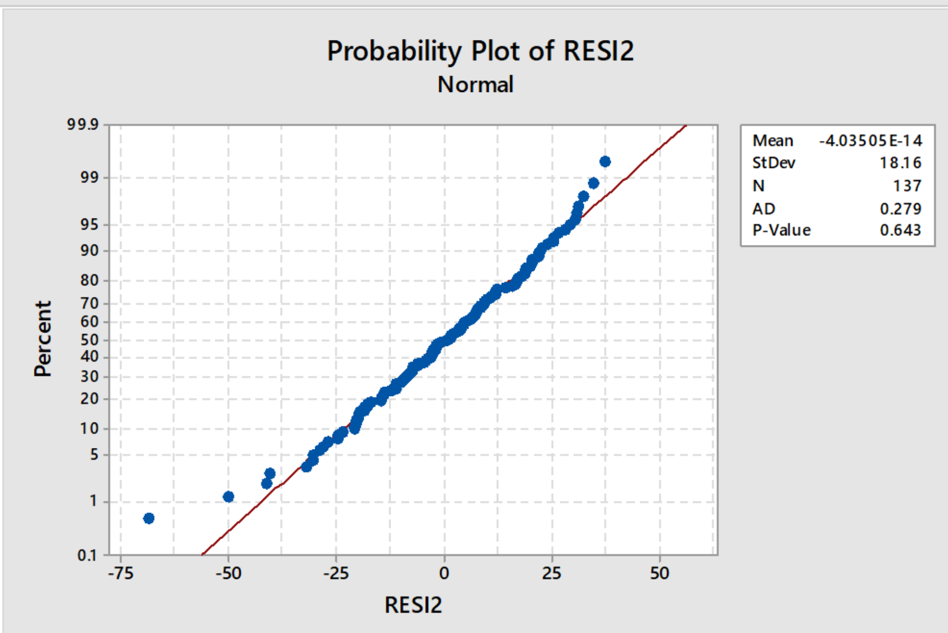
$H_a: F(x)$ is not normally distributed.

Interpretation: Since P-value › level of significance, therefore Null hypothesis is accepted. We see that it is normally distributed.

Our model doesn't have any multicollinearity problem.

Model Assumptions:





Our model seems hold linear model, normality and constant variance assumption.

Final Model:

Regression Analysis: SCORE versus EMP@EOY, POS@EOY, TYPE, Employed

Method

Categorical predictor coding (1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	23830	5957.5	17.54	0.000
EMP@EOY	1	2796	2795.8	8.23	0.005
POS@EOY	1	4731	4730.9	13.93	0.000
TYPE	1	3339	3339.3	9.83	0.002
Employed	1	5788	5787.8	17.04	0.000
Error	132	44834	339.7		
Lack-of-Fit	123	41991	341.4	1.08	0.494
Pure Error	9	2843	315.9		
Total	136	68664			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
18.4297	34.70%	32.73%	29.38%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	108.55	4.85	22.36	0.000	
EMP@EOY	-2.051	0.715	-2.87	0.005	2.63
POS@EOY	3.674	0.984	3.73	0.000	3.41
TYPE					
1	12.76	4.07	3.14	0.002	1.67
Employed					
1	20.41	4.95	4.13	0.000	1.02

The estimated regression equation is

$$\begin{aligned}\widehat{Evaluation\ score} &= 108.14 - 2.051 * EMP@EOY + 3.674 * POS@EOY \\ &+12.76 * Type + 20.41 * Employed\end{aligned}$$

For every years with the company at the end of the year on an average the evaluation score decreases 2.051 while holding current position, type of evaluation and employment status constant. Similarly, for every current position at the end of the year the evaluation score increases on an average 3.674, while holding years with the company at the end of the year, type of evaluation and employment status constant. This rate also applied to both type of evaluation (3 months and 1-year evaluation) and employment status (those who are still employed and already left the company). We can also say that the employees who still stayed in the company and their 1-year evaluation is more than those already left and 3-month evaluation.

Recommendations:

Here are some of the questions raised in the case study. Based on the final model summarized above, we would like to provide evidence to support our conclusion.

a. Evaluators are occasionally late for their evaluations. Rumors at the company suggest that the later the evaluation, the lower the score, as the manager may be attempting to postpone a controversial or hostile discussion with the employee about poor performance.

The belief that the later the evaluation, the lower the scores is not correct. Because in our analysis, we haven't found the lag variable significant. The scatterplot for score and lag variables appears not to be related. Later we also found that the lag variable is not included in the regression model. Consequently, there is no evidence that the later evaluation affects the score.

b. Three-month probationary evaluations are considered unnecessary by the employees, because three months is generally not enough time to adjust to a new position or reach the peak level of efficiency in a new job.

In our analysis the three-month evaluation adversely affects the evaluation scores. The mean score after three-month evaluation is 108.14 and the mean score after annual evaluation is 108.14+20.41=128.55. Therefore, the claim that three-month is not enough time to adjust to the environment is true. Over the time employees are getting used to working environment, and have a higher score.

c. There are rumors that the company is attempting to force early retirement by giving the long-time employees lower scores than others.

This rumor is true because we have found that the longer the employee works at the company, the less they have scores. The regression analysis finds the employment variable significant.