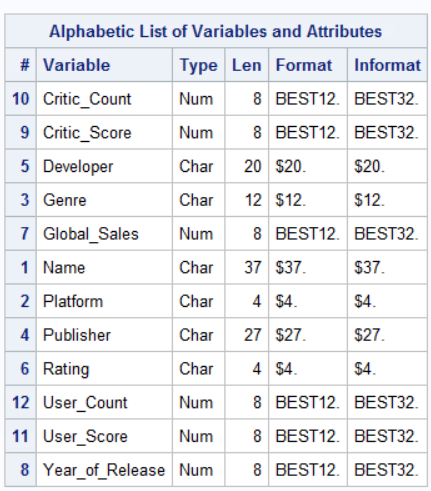
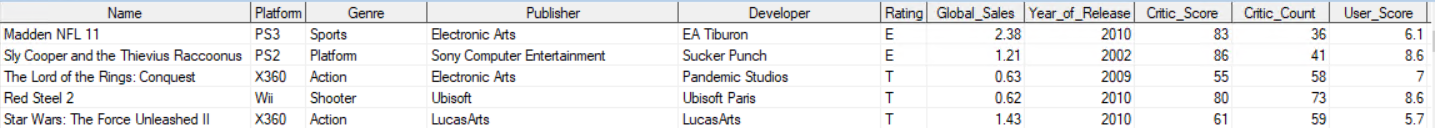
**Predictive with SAS - Homework 3**

**Dataset:**

The data is about the global sales and critics of video games released between 2001 and 2012. This dataset contains 4413 observations with 12 variables.

Of the 12 variables, 6 are numeric and other 6 are character with each having defined length formats.

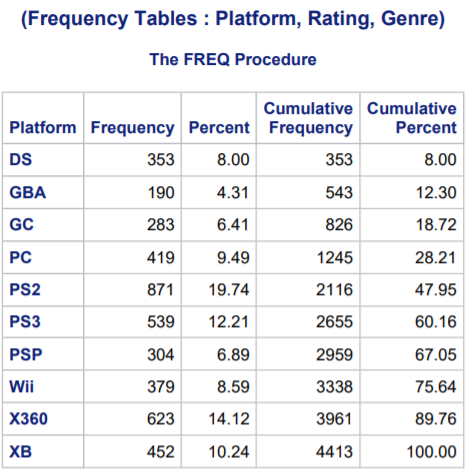




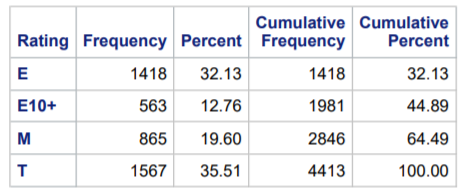
**Q1. Develop a regression model**

1. **First, use proc freq to create a frequency table of 3 variables: platform, genre, and rating**

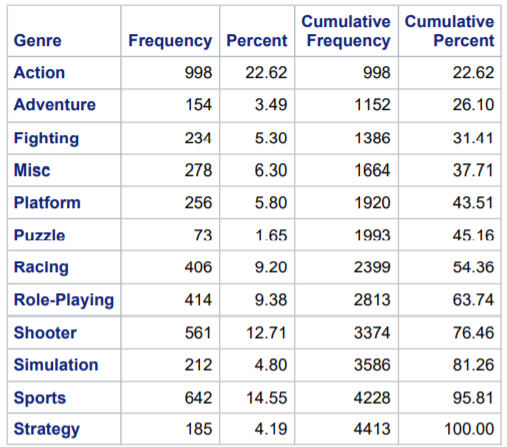
Upon running PROC FREQ with Platform, Rating and Genre, the frequency of each category in the variable, distribution percentage, cumulative percentage and cumulative frequency are displayed.



PS2 has highest frequency and GBA has the lowest frequency in the variable Platform.

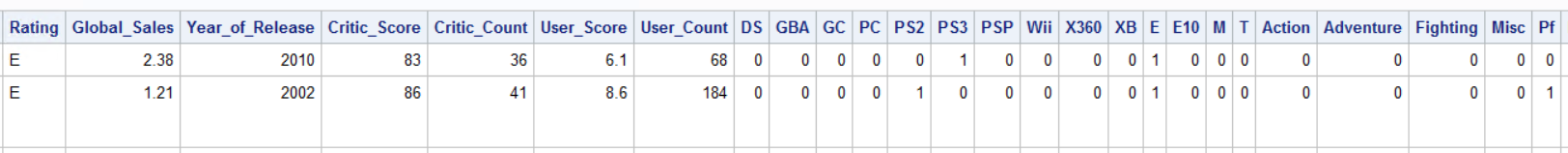


Rating T has the highest frequency and E10+ has the lowest frequency

In Genre, Action has the highest frequency and Puzzle has the lowest frequency.

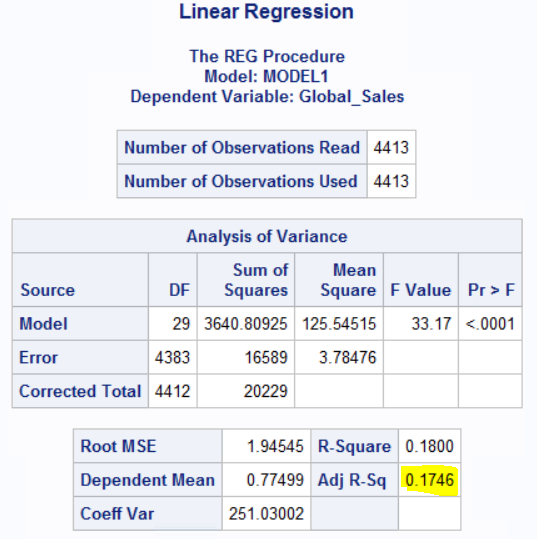
**1.b. In data step, create categorial variables for platform, genre, and rating using if/else statements. Also create a variable for the age of the game relative to year 2013.**

Using IF/ELSE, dummy variables are created for each observation for the variables Platform, Genre and Rating. The last column shows the Age variable that is relative to year 2013, obtained by subtracting Age from 2013.



**1.c. Run a regression with all relevant X variables. Report the adjusted R-squared.**

Considering the following variables, Linear regression is run using the command PROC REG and passing the x and y variables in model.

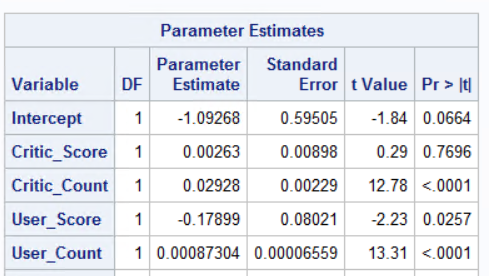


The adjusted R score is a measure of the r score obtained when taking a different set of variables into consideration. It increases when the variables fit the model and decrease when the variables don’t fit the model. **The model has an adjusted R score of 0.1746.**

The PROC REG command also shows the error, p value and t value of each variables in the mode. P value depends on magnitude of association and the precision of the estimate. While the standard error tells about how accurate the mean of the sample is to the population.

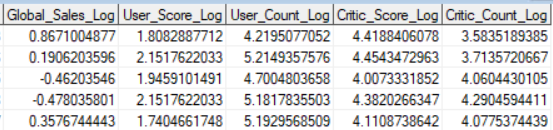
The equation for the regression is

**Global\_sales**= -1.09268+0.00263\*Critic\_score+0.02928\*Critic\_Count-0.17899\*User\_Score+0.00087304\*User\_Count-0.03446\*Age+0.00205\*CU\_Score+0.24619\*Action+0.03951\*Adventure+0.31243\*Fighting +0.61746\*Misc+0.27630\*Pf--0.15651\*Puzzle+0.30161\*Racing-0.06781\*Role+0.20747\*Shooter+0.47968\*Simulation+0.22852\*Sports+0.77537\*DS+0.40535\*GBA+0.19933\*GC-0.32248\*PC+0.61427\*PS2+0.50668\*PS3+0.43502\*PSP+1.34382\*Wii+0.26610\*X360+0.39307\*E-0.09000\*T--0.13206\*M

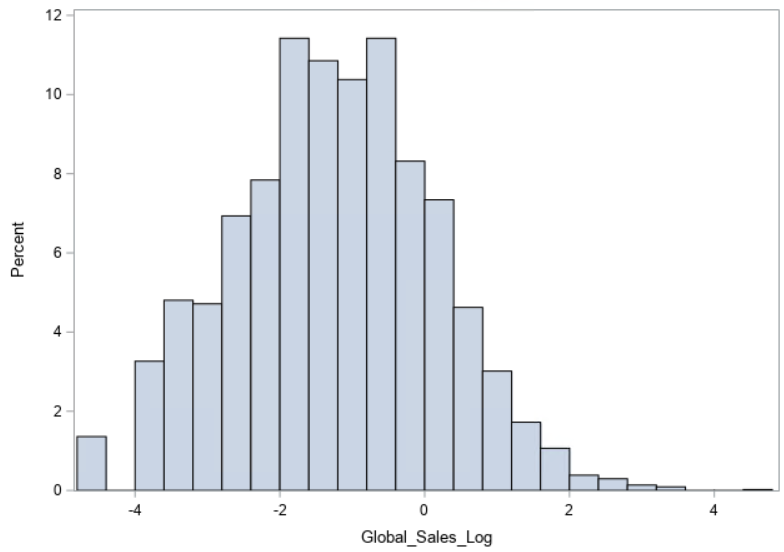


**d. Now, generate natural log of the following variables: global sales, critic\_score, critic\_count, user\_socre, user\_count**

Using the Log() function we can calculate the log of the values. The



The histogram of the distribution of the Global sales Log is plotted.



**e. Run a regression with the log of Y variable and report adjusted R-squared.**

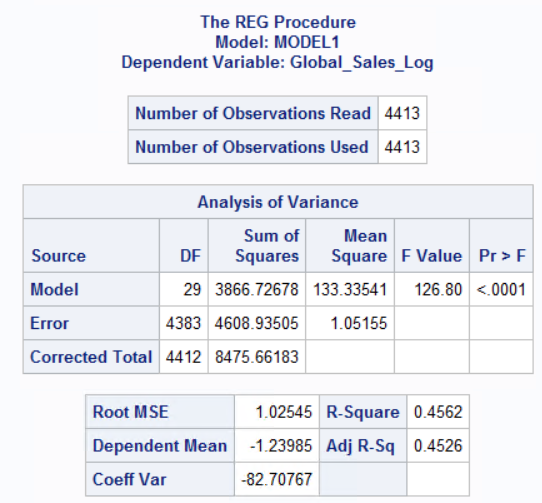
The regression is run with Log(Global sales) as the Y and the significant features as X variables. Since the Y variable is now in logarithmic scale, it affects the ways the model is represented by the other variables. This can be seen by the change in the adjusted R 2 value that is obtained by this model.

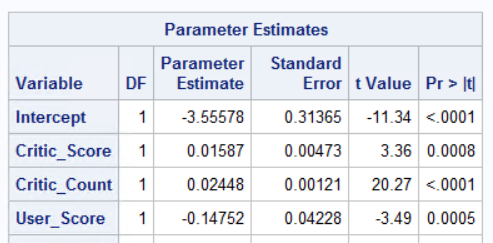
**global\_sales\_log** = -3.55578+0.01587\*Critic\_Score+0.02448\*Critic\_Count -0.14752\*User\_Score+0.00050375\*User\_Count+0.00160\*CU\_Score+0.02158\*Age+

0.56785\*Action+0.11039\* Adventure+ 0.58181\*Fighting+0.74586\* Misc+0.53615\* Pf+0.04344\* Puzzle+0.45970\* Racing+0.27264\* Role+0.43476\* Shooter

+0.83498\*Simulation+0.47272\* Sports+0.43433\* DS+0.46030\* GBA 0.28962\*GC -1.27885\*PC+0.85796\*PS2+0.67909\* PS3+0.36827\* PSP+0.89381\* Wii+0.29100\* X360+0.22863\*E-0.11434\*T--0.20908\* M

The adjusted R squared is now **0.4526**



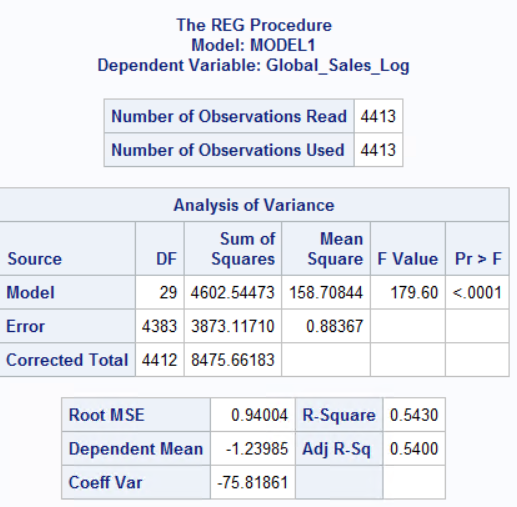


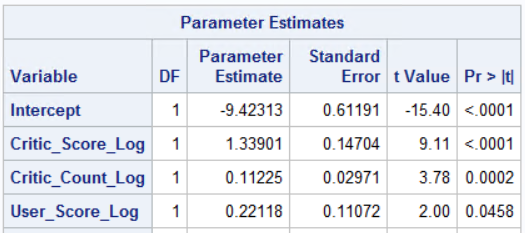
**f. Run a regression with the log of Y variable as well as log of X variables generated in part d). Report adjusted R-squared.**

Now, the dataset considers the log value of the significant terms in the X variables and Y variable is the log of the Global Sales. Since the scale of the model now changes we can see the change in the model representation and the contribution of each variable in this model differs from the last ran model.

**Global\_Sales\_Log** = -9.42313+1.33901\*Critic\_Score\_log+0.11225\* Critic\_Count\_log+0.22118\* User\_Score\_log+0.65492\* User\_Count\_log-0.00136\* CU\_Score-0.00098244\* Age+0.49994\*Action+0.153778\* Adventure+0.50929\* Fighting+0.91150\*Misc+0.36927\*Pf+0.29775\*Puzzle+0.43190\*Racing+0.01774\* Role +0.34012\*Shooter+0.90700\*Simulation+0.61233\* Sports+0.16489\* DS+0.22759\* GBA-0.00214\*GC-2.51798\* PC+0.41300\* PS2+0.08976\* PS3+0.06354\*PSP+0.43728\* Wii-0.15573\* X360+0.15858\*E-0.18455\*T -0.47056\*M;

This model yields an adjusted R square value of **0.5400**





**g. Which model (out of part c, e, and f) offers the highest adjusted R-squared? What would be the economic reasoning on why that model provides the best fit? Also interpret coefficients.**

The model in F has the highest adjusted R score is 0.5400 which is the highest among the three models. This is because of the LOG-LOG model used in this where the X and Y variables are in log scale.

For nonlinear effects, logarithmic transformation is the best method to show the data model. When x is small it has a large effect but while x is very large, the effect is diminished.

The advantage of the log function is that it doesn’t show as unit based but rather as percentage based. The violations discussed in question 2 are well reduced because of the use of this model.

**Model Interpretation:**

**Global\_Sales\_Log** = -9.42313+1.33901\*Critic\_Score\_log+0.11225\* Critic\_Count\_log+0.22118\* User\_Score\_log+0.65492\* User\_Count\_log-0.00136\* CU\_Score-0.00098244\* Age+0.49994\*Action+0.153778\* Adventure+0.50929\* Fighting+0.91150\*Misc+0.36927\*Pf+0.29775\*Puzzle+0.43190\*Racing+0.01774\* Role +0.34012\*Shooter+0.90700\*Simulation+0.61233\* Sports+0.16489\* DS+0.22759\* GBA-0.00214\*GC-2.51798\* PC+0.41300\* PS2+0.08976\* PS3+0.06354\*PSP+0.43728\* Wii-0.15573\* X360+0.15858\*E-0.18455\*T -0.47056\*M;

For 1% increase in Critic\_score there is 1.33% change in global sales.

For 1% increase in Critic\_Count there is 0.11% change in global sales.

For 1% increase in User\_Score there is 0.22% change in global sales.

For 1% increase in User\_Count there is 0.65% change in global sales.

For 1% decrease in CU\_Score there is 0.001% change in global sales.

For 1% decrease in Age there is 0.0009% change in global sales.

For 1% increase in Action there is 0.49% change in global sales.

For 1% increase in Adventure there is 0.15% change in global sales.

For 1% increase in Fighting there is 0.50% change in global sales.

For 1% increase in Misc there is 0.91% change in global sales.

For 1% increase in Pf there is 0.36% change in global sales.

For 1% increase in Puzzle there is 0.297% change in global sales.

For 1% increase in Racing there is 0.43% change in global sales.

For 1% increase in Role there is 0.01% change in global sales.

For 1% increase in Shooter there is 0.34% change in global sales.

For 1% increase in Simulation there is 0.90% change in global sales.

For 1% increase in Sports there is 0.61% change in global sales.

For 1% decrease in DS there is 0.16% change in global sales.

For 1% increase in GBA there is 0.22% change in global sales.

For 1% decrease in GC there is 0.002% change in global sales.

For 1% decrease in PC there is 2.51% change in global sales.

For 1% increase in PS2 there is 0.41% change in global sales.

For 1% increase in PS3 there is 0.08% change in global sales.

For 1% increase in PSP there is 0.06% change in global sales.

For 1% increase in Wii there is 0.437% change in global sales.

For 1% increase in X360 there is 0.155% change in global sales.

For 1% increase in E there is 0.158% change in global sales.

For 1% decrease in T there is 0.18% change in global sales.

For 1% decrease in M there is 0.47% change in global sales.

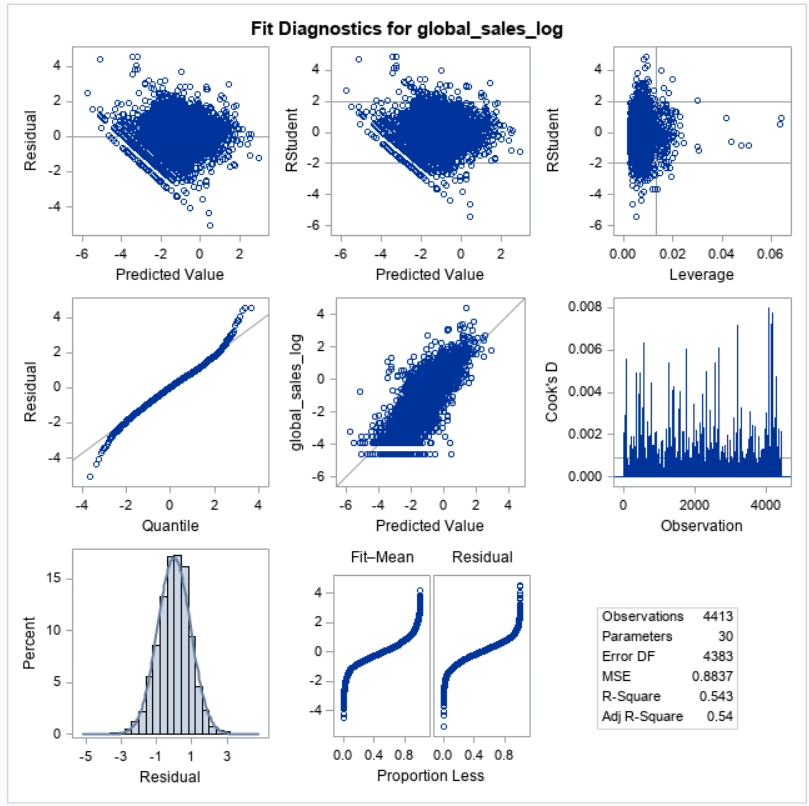
**2. For the model you constructed in Q.1part g), verify whether the various regression assumptions discussed in class are satisfied. If an assumption is violated, discuss how it can be handled, and implement the solution if possible. Discuss whether the solution had a practically significant impact on your model results. If no practical solution is possible, acknowledge it.**

**(i)Outliers and influential points detection**

The outliers and influential points in the dataset drag the regressor to different alignment towards the extreme values. The influential points are not seen at first because of the outliers but once the outliers are removed, the influential points then cause the regressor to have a biased orientation.

1)New dataset is considered with studentized residuals and Cook's d value for each observation.

The Cook’s d plot helps us to identify outliers as well as data points that are extreme and cannot be easily spotted because of the existence of the outliers. This is obtained by using the COOKD and Student=residuals in the reg model.



The observations from the model:

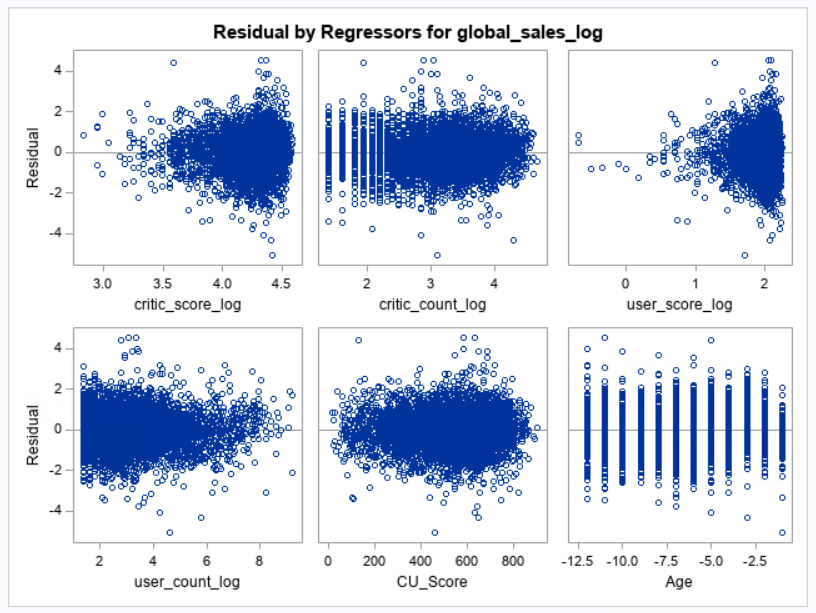
1)We see that in the Cook’s D plot certain data points have value peaking at 0.008 highest and other peaking from 0.004-0.008 these are the outliers and extreme values.

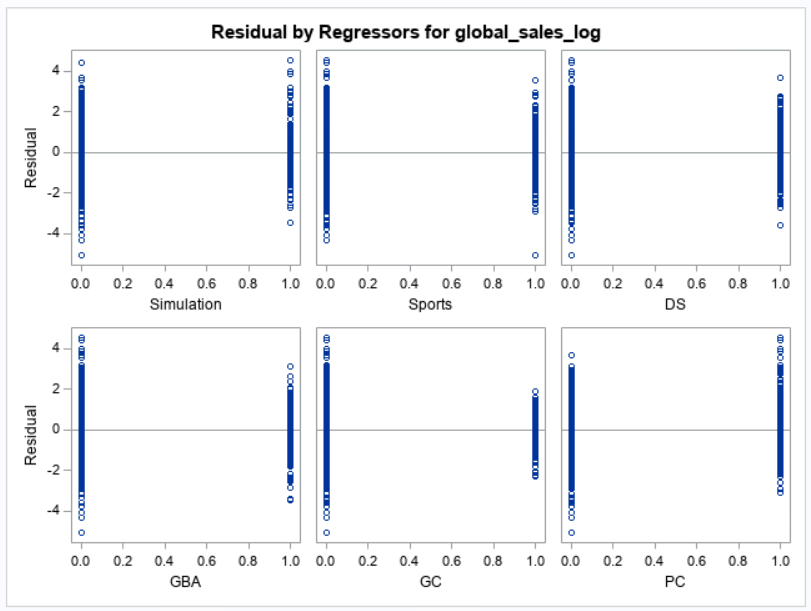
2) The residual and the predicted value show the difference in the predicted value and actual value. The outliers can be seen in the model with one data point extremely at the top.

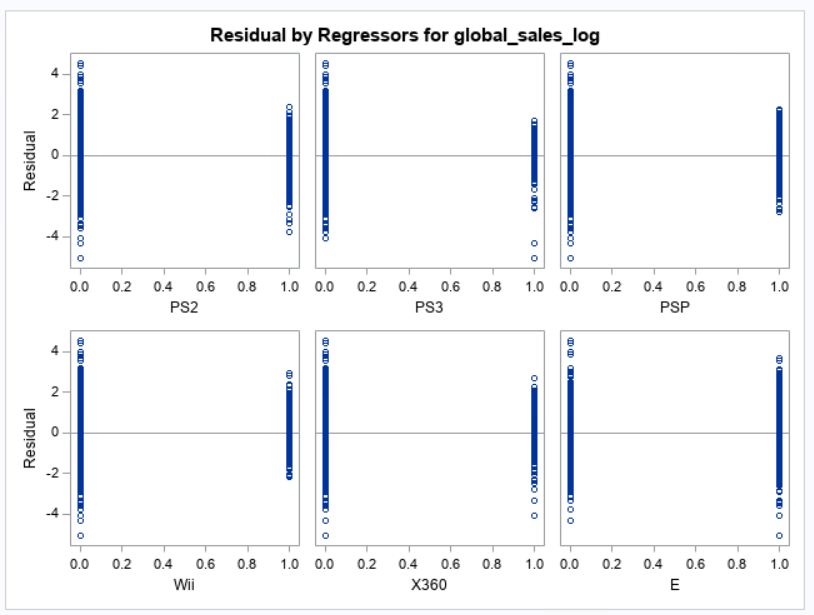
3) The leverage plot shows the data points which are isolated out and have a higher leverage over the dataset. We can see two data points farthest at 0.06 with high leverage

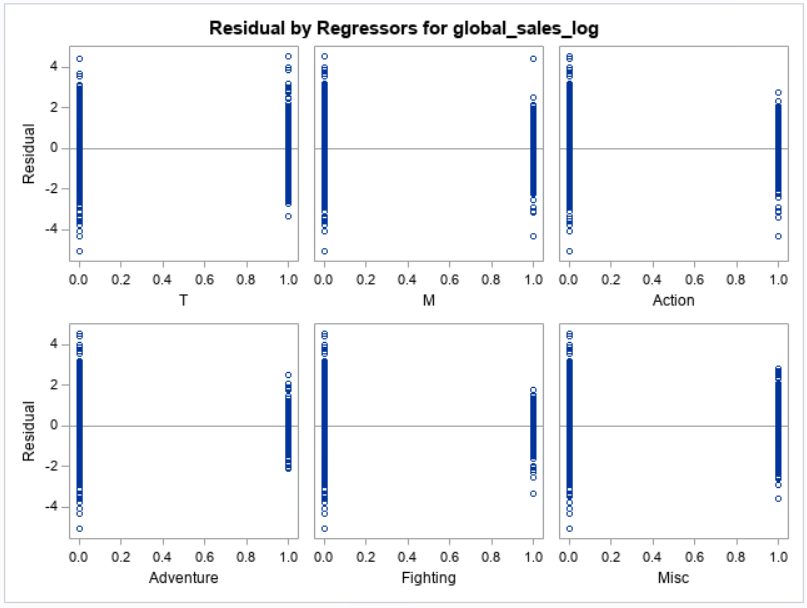
4)Few data points fall outside 2 standard deviation.

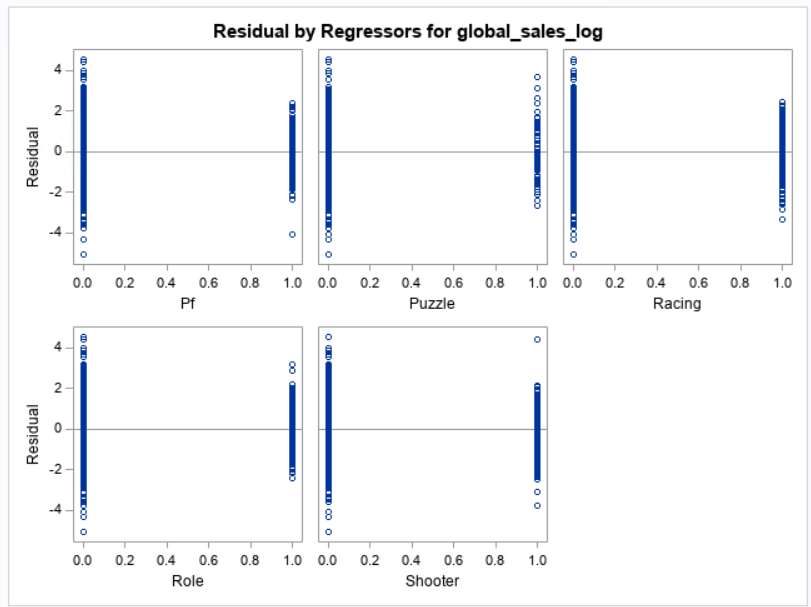
5) The normal distribution curve plot shows the data points edging outside the curve.





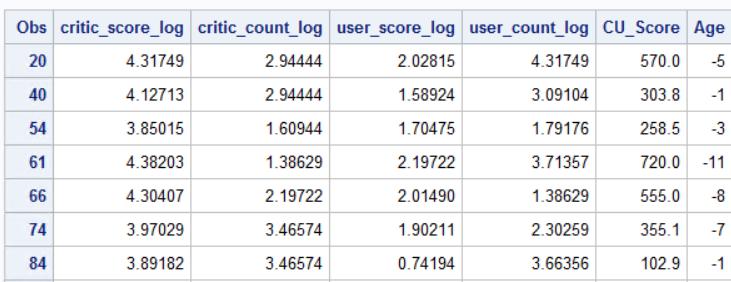






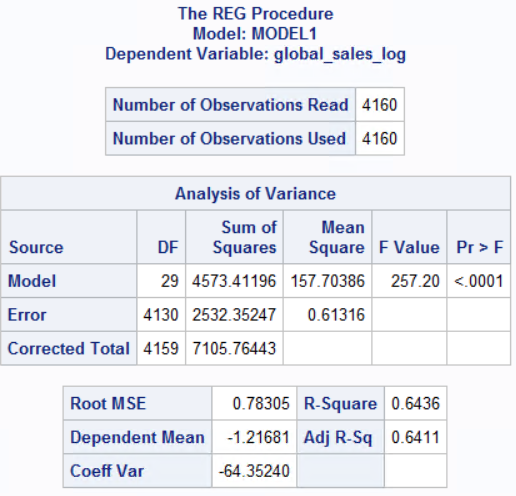
2) Printing influential observations:(Cook's d > 4/n)

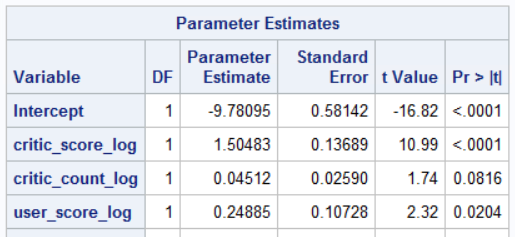
Using rule: Cook’s D value > 4/n we can find the observations that contributed to the model, eliminating the remaining outliers and the extreme values. The dataset of the resultant elimination is shown below.



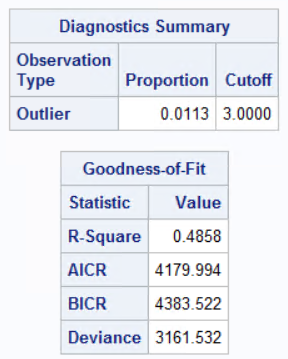
3) Estimating model without influential observations

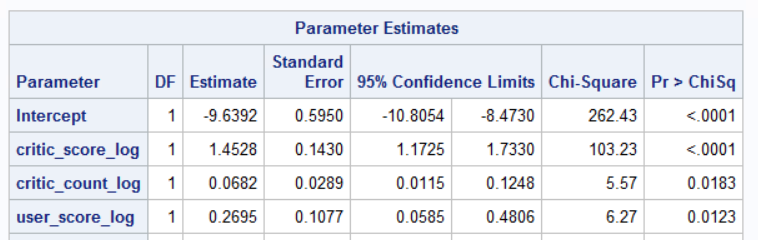
This violation can affect the score of the r2 obtained for the model. Thus, the model is run after eliminating the outliers and influential observations this constricts the models to data points that are near the regressor then this helps to increase the Adjusted R 2 value. The value of this regression run is the 0.6411. This shows that the variables in the model and highly significant to the model and are contribute to the Y variable.





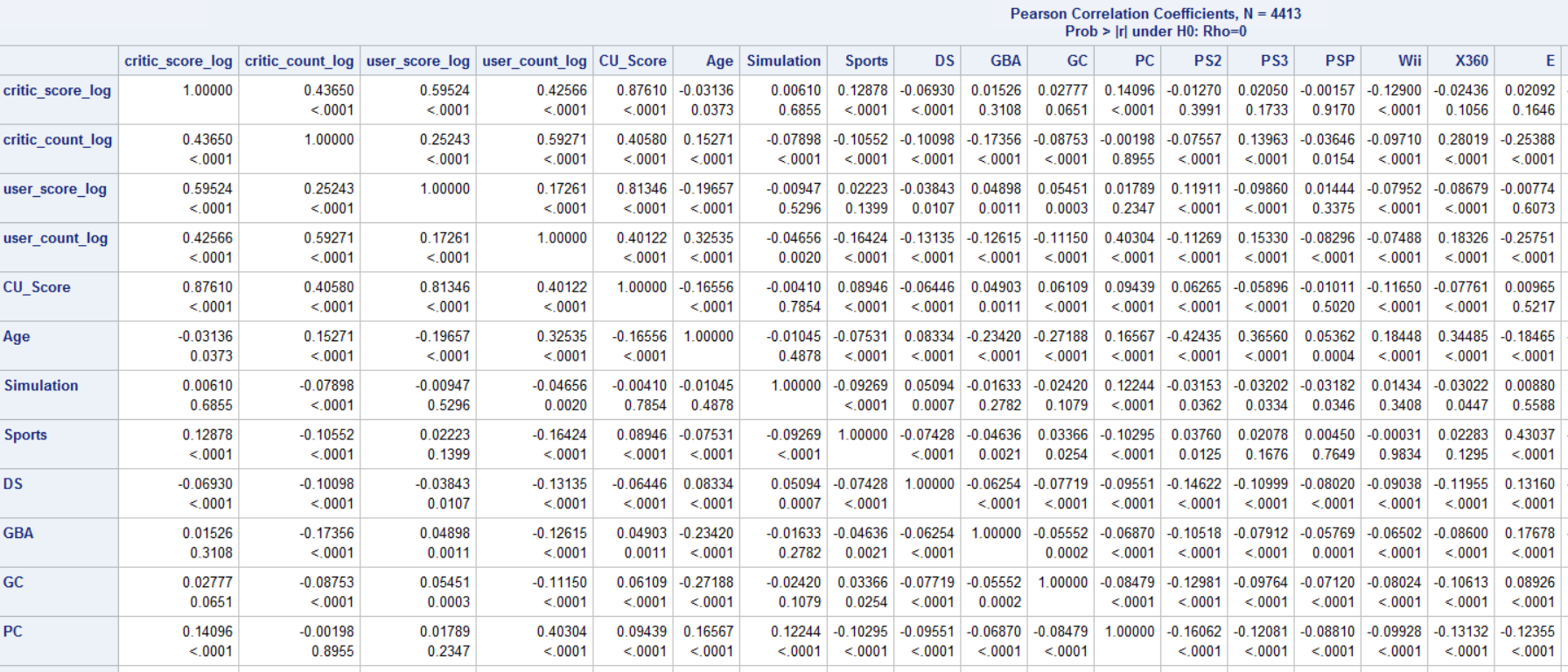
**4) Robust Regression** – The robust regression is a manual method to find the outliers and influential points in the dataset and its run with **proc robustreg** with MM method.





(**ii) Multicollinearity detection**

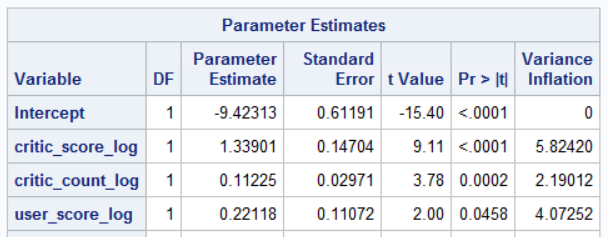
Multicollinearity is checked using a correlation matrix. It cn be seen that mostly all the values are small thus showing not much collinearity.



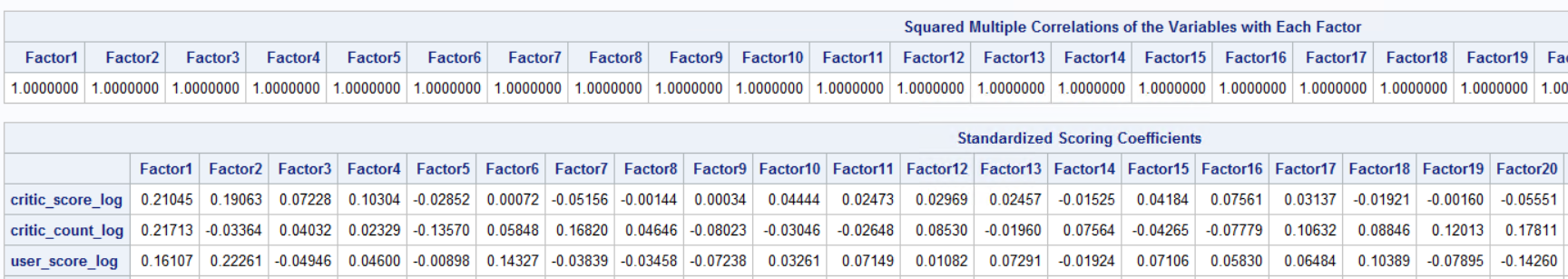
The correlation matrix defines the collinearity of the variables with its significance given below.

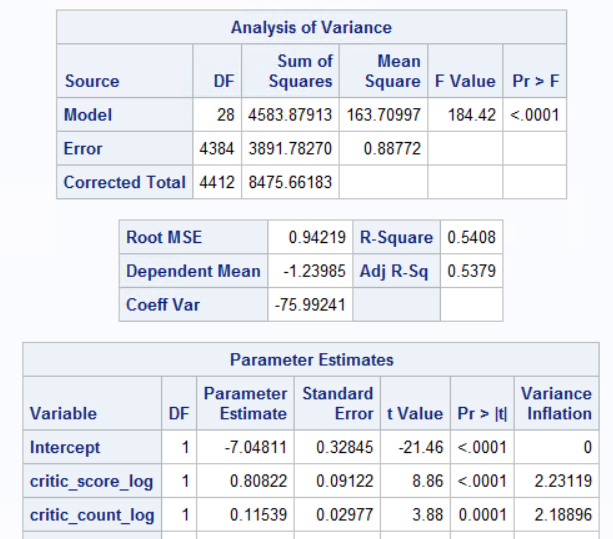
**Variation inflation**

The variation inflation can also be used to test for collinearity. We can see that most of the values of the terms are less than 0. With only CU\_score having a greater than 10 value.



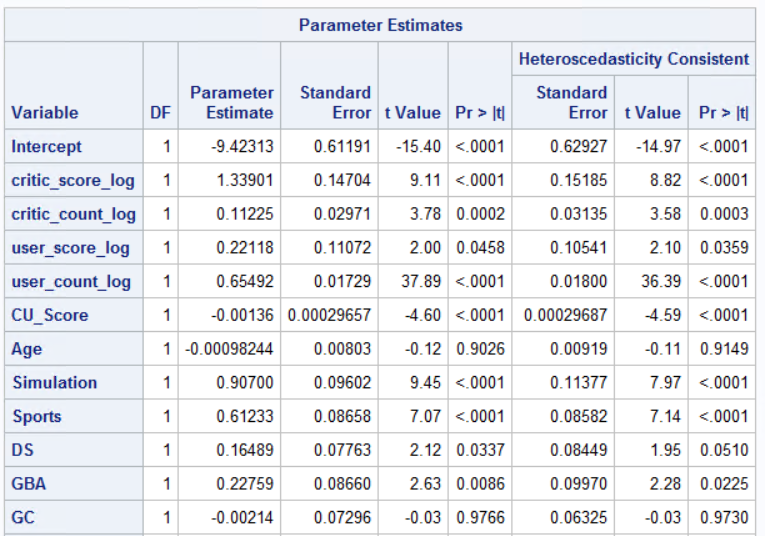
Generating factors



****

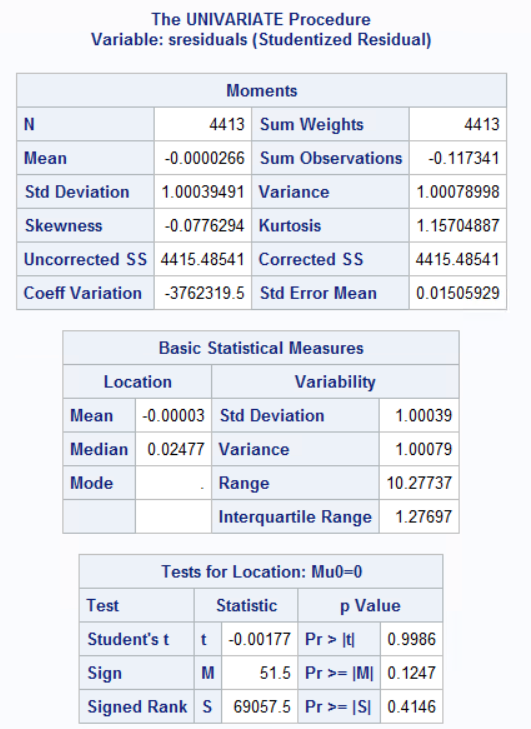
**(iii) Heteroscedasticity detection**

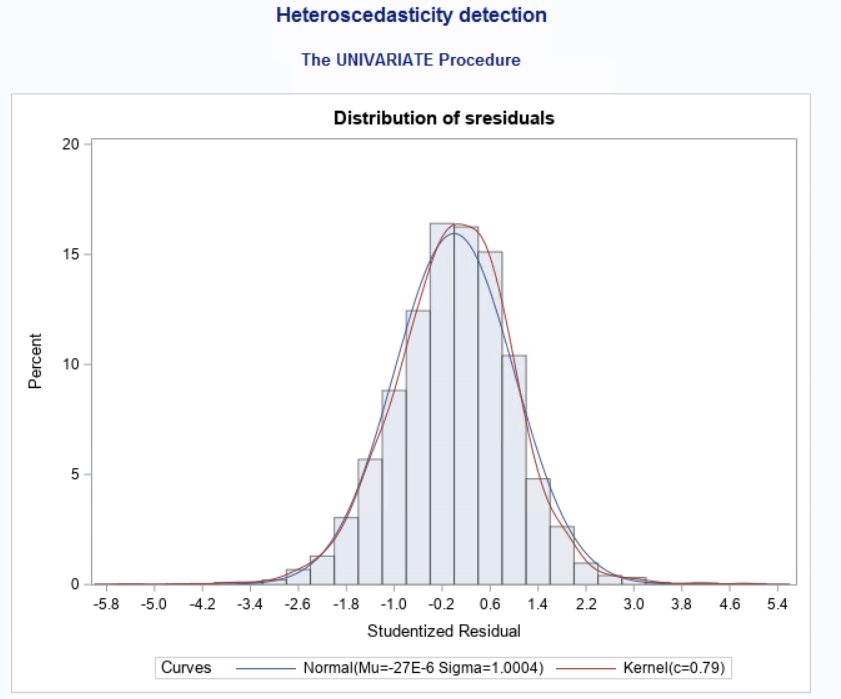
Heteroscedasticity error in the data can be seen using the white’s test. It can be seen in this that the t value generated is much lesser than the actual t values thus the error doesn’t exist. Thus, the spread of the residuals around the regressor doesn’t vary across the plot.



**(iv) Normality of error term**

The error must follow normal distribution else the p term would signify wrong.

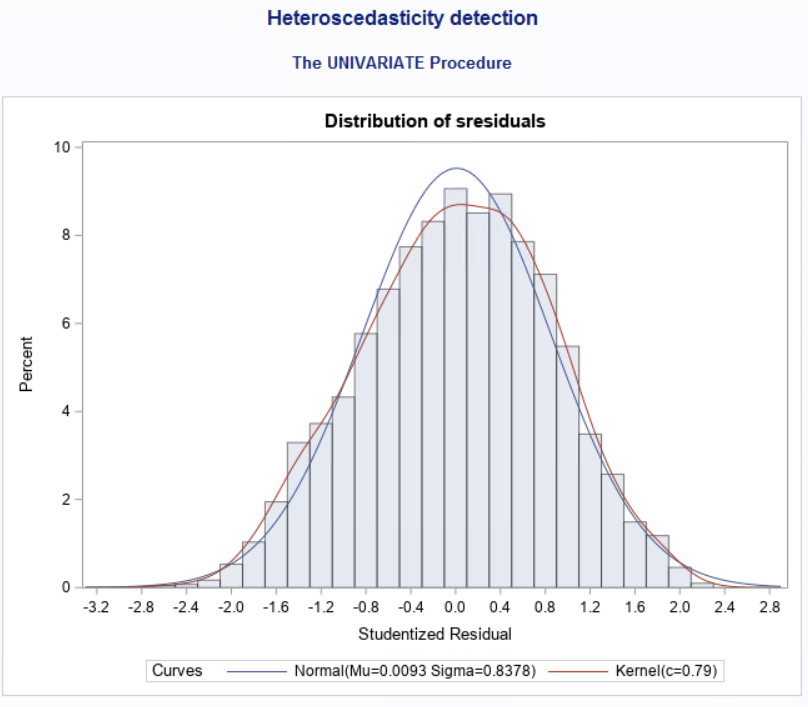


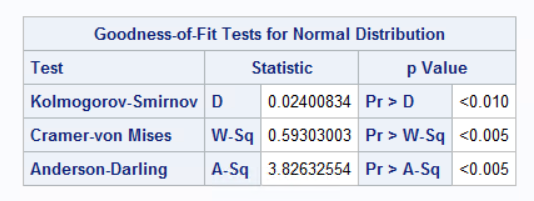


The cook’s d is used to eliminate influencers and then run the test for normality with PROC UNIVARIATE. We can see that both the curves had followed a normal distribution but after eliminating the influencers the scale of the y axis has reduced and the plot as a good spread now with the kernel displayed by the red line showing the curves of the graph while the blue normal distribution smoothens the curves

Thus, violation didn’t exist in the model.



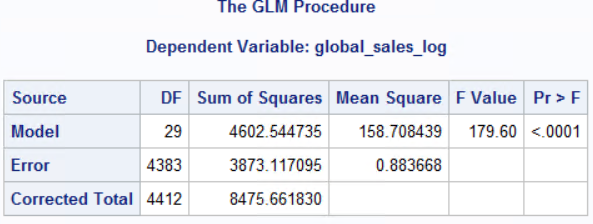




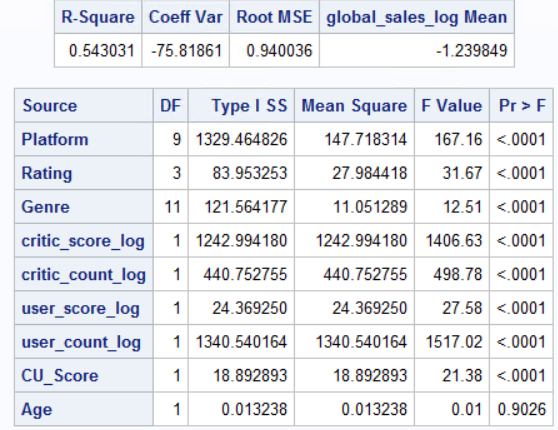
Since the last models has LOG-LOG model it dint have the violations except for outliers .

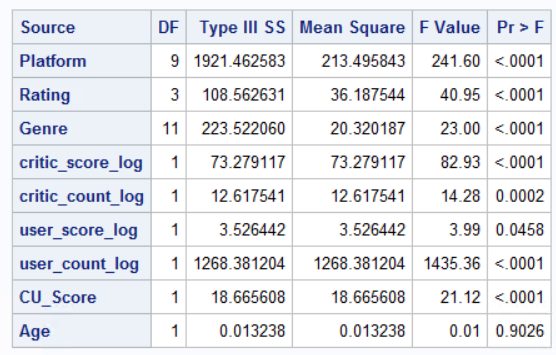
**3. Replicate your final model in Q.1 part g) using proc glm and class commands. Note that proc glm allows you quickly check many basic results. However, it does not provide many diagnostic outputs that proc reg provides.**

PROC GLM command is used to generate this model. The major advantage with GLM is that the variables can be provided by classes to the models and also the interaction terms are generated immediately without creating variables separately for each term.









The below table shows the values when the model is run through PROC GLM. There are 2 tables that describe the sum of squares and other details with 2tpes of tests namely the Type 1 and Type 3 tests. If we want to consider all the independent variables in the model, then we take the Type 3 F-test table. If we don't want to consider all the independent variables, then we take the Type 1 F-test table.

SAS labels the F-tests based on the incremental sum of squares with the heading Type I Sum of Squares. The F-tests corresponding to the t-tests is labelled under the heading of Type III Sum of Squares.