**Homework 3 – Regression Analysis**

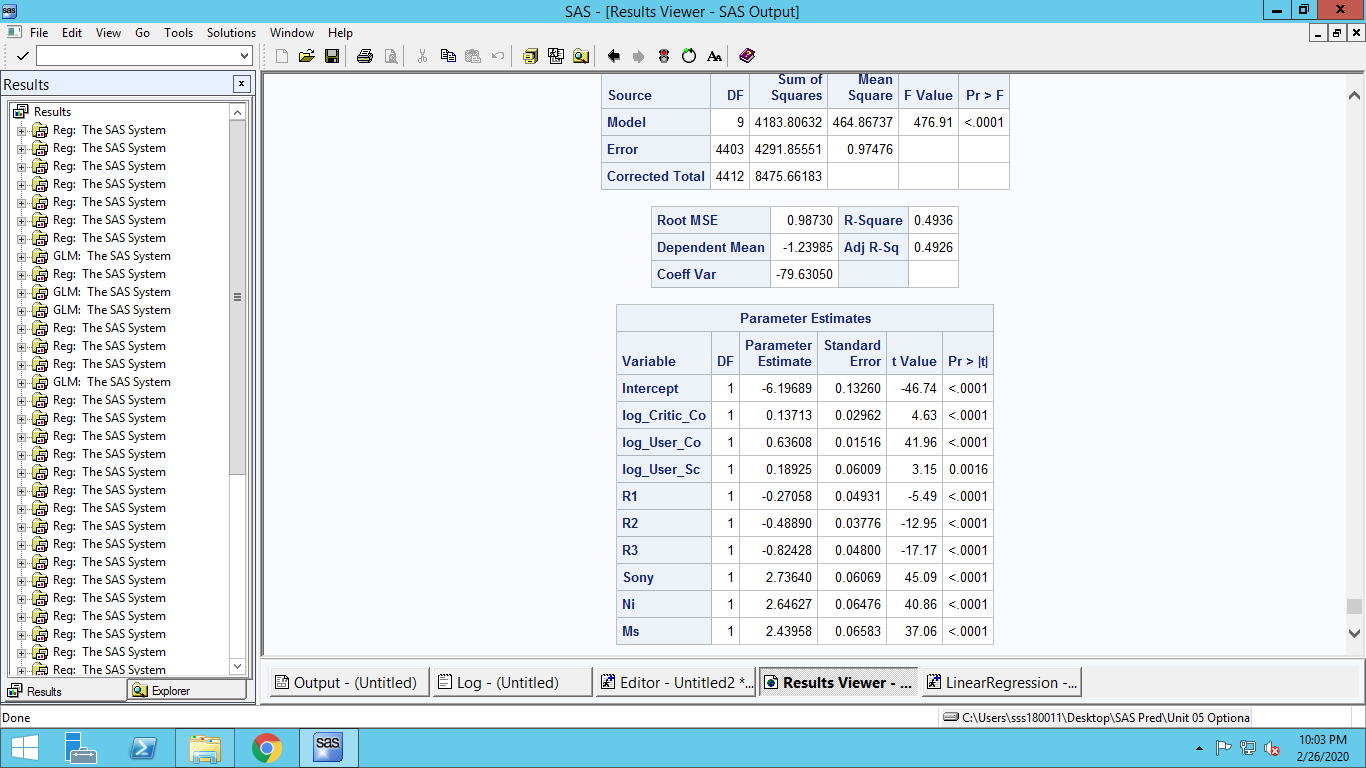
The dataset for this exercise is available in VideoGamesSales\_Main.7bdat. This dataset contains information on the global sales and critic and user review ratings for videogames launched between 2001 and 2012 (from [www.vgchartz.com](http://www.vgchartz.com)). The variables are:

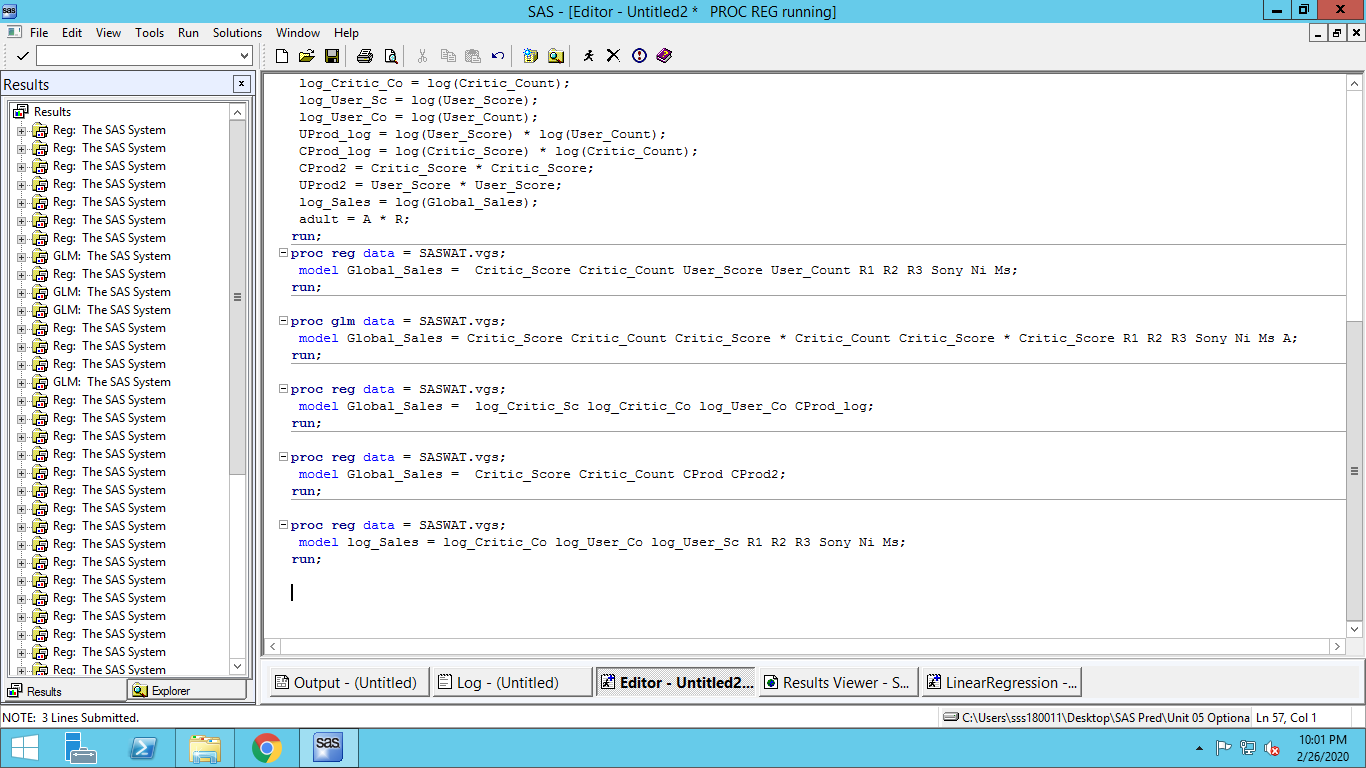
* Name of the game
* Videogame platform on which it was released.

|  |  |
| --- | --- |
| **Platform** |  |
| **DS** | Nintendo DS |
| **GBA** | Nintendo Game Boy Advance |
| **GC** | Nintendo Game Cube |
| **PC** | Personal Computer |
| **PS2** | Sony PlayStation 2 |
| **PS3** | Sony PlayStation 2 |
| **PSP** | Sony PlayStation Portable |
| **Wii** | Nintendo Wii |
| **X360** | Microsoft XBOX 360 |
| **XB** | Microsoft XBOX |

* Videogame Genre (e.g., Action, Sports, Shooter etc.)
* Publisher
* Developer
* Rating: E = Everyone, E10+ = Everyone 10+, T = Teen, M = Mature
* Global Sales (Millions of units)
* Year of release
* Critic Score (0 – 100): Average critic rating
* Critic Count : Number of critic ratings
* User Score (0 – 10): Average user rating
* User Count: Number of user ratings

1. Develop a regression model that links global sales to video game reviews. Explore ways in which the model fit could be improved through suitable changes to the model specification and variables.
   1. Present the final model and results.





**Final Model:** log(Global Sales) = -6.19689 + 0.13713 \* log(Critic Count) + 0.63608 \* log(User Count) + 0.18925 \* log(User\_Score) + -0.27058 \* R1 +

-0.48890 \* R2 + -0.82428\* R3 + 2.73640\* Sony + 2.64627\* Ni + 2.43958\* Ms

**R1: Everyone 10+**

**R2: Teen**

**R3: Mature**

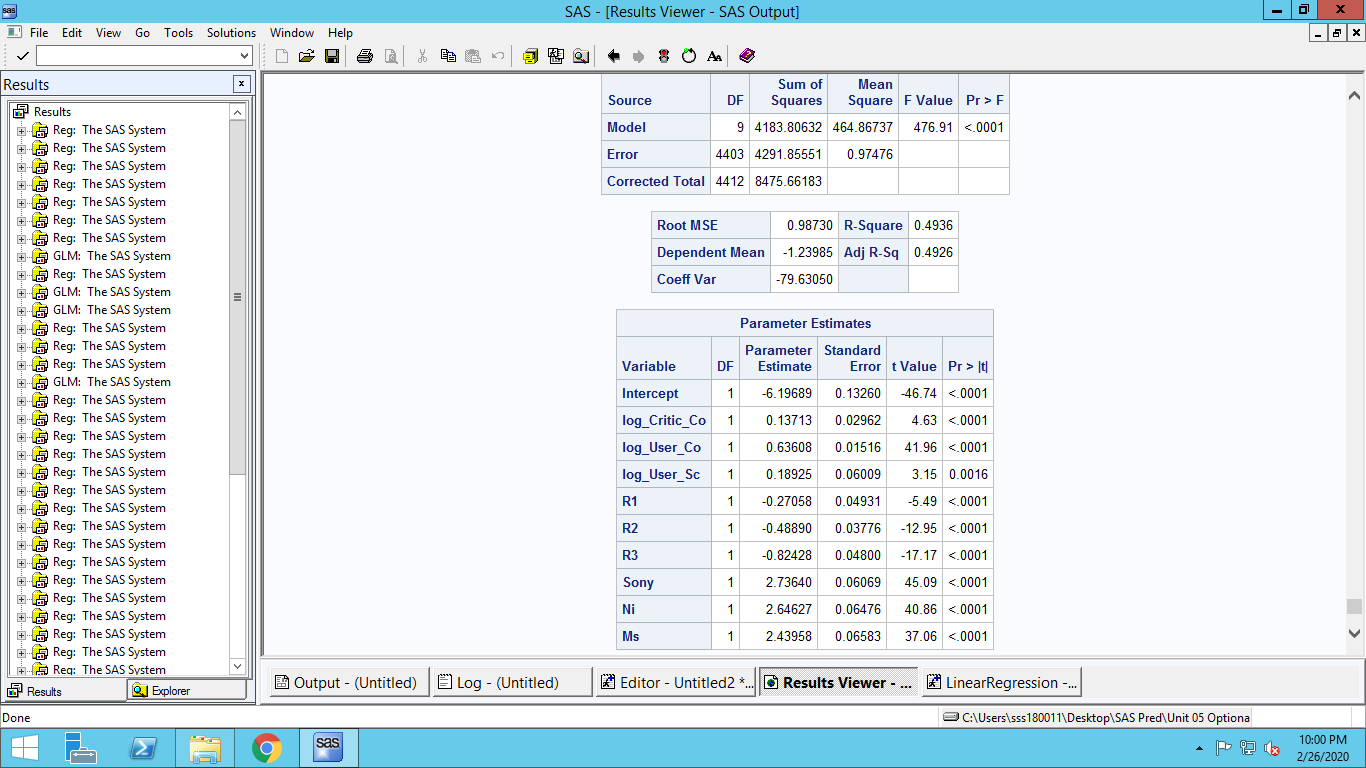
**Base: Everyone**

**Sony: Sony Products**

**Ni: Nintendo Products**

**MS: Microsoft Products**

**Base: Personal Computer**

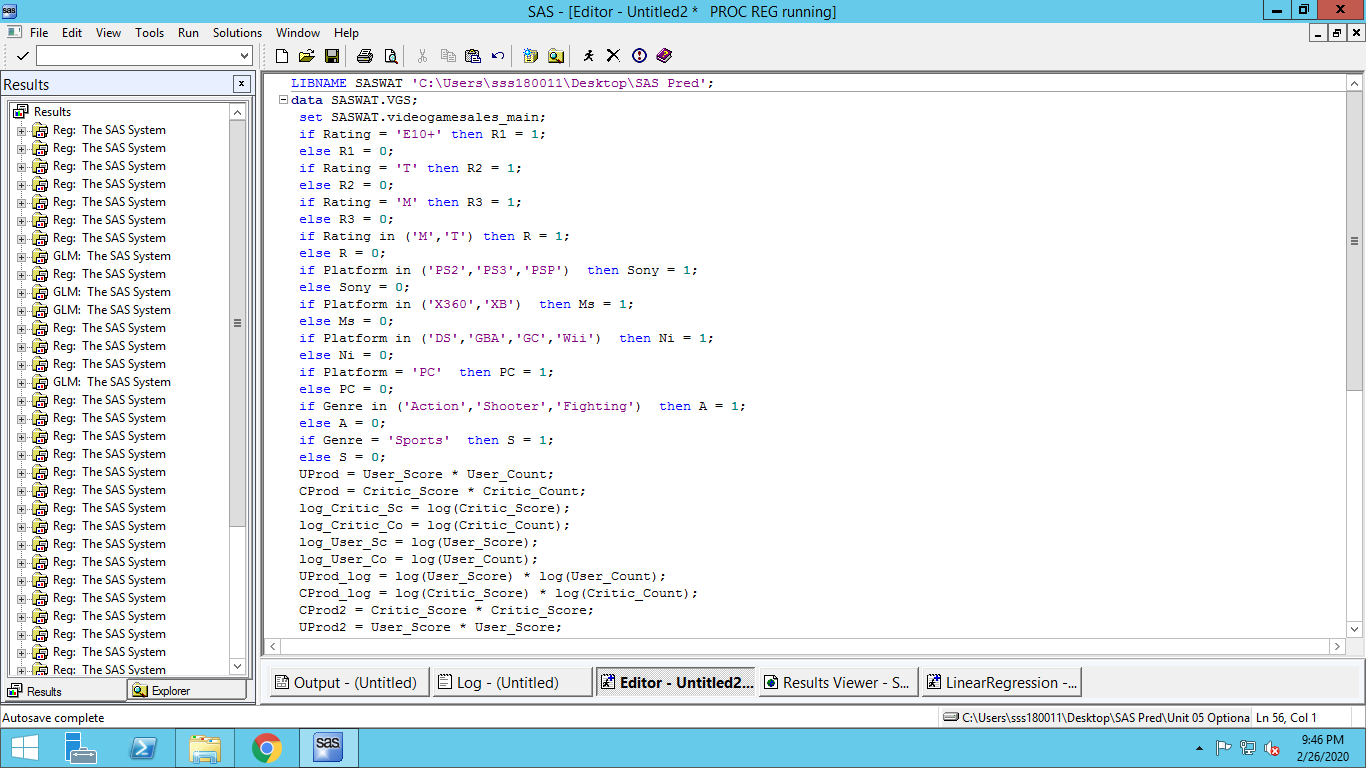


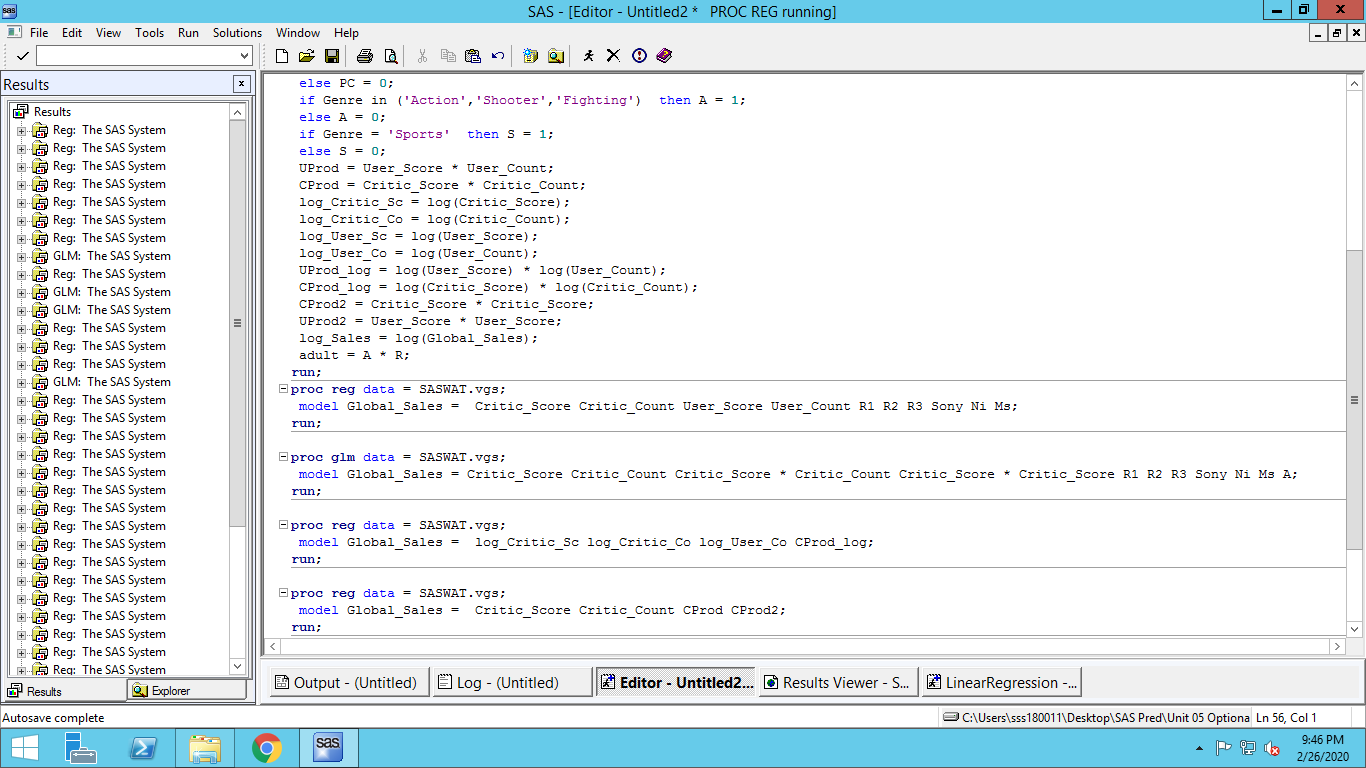
**R\_Square = 0.4936**

**Adjusted R-Square = 0.4926**

* 1. Explain how you developed your model (what was your initial model, what were the key variations you tried and how did you arrive at the final model – and the thought process behind these steps).

**STEP 1:**





**Nominal Explanatory variables were created on Rating column. ‘Everyone’ is base. R1 = E10+.**

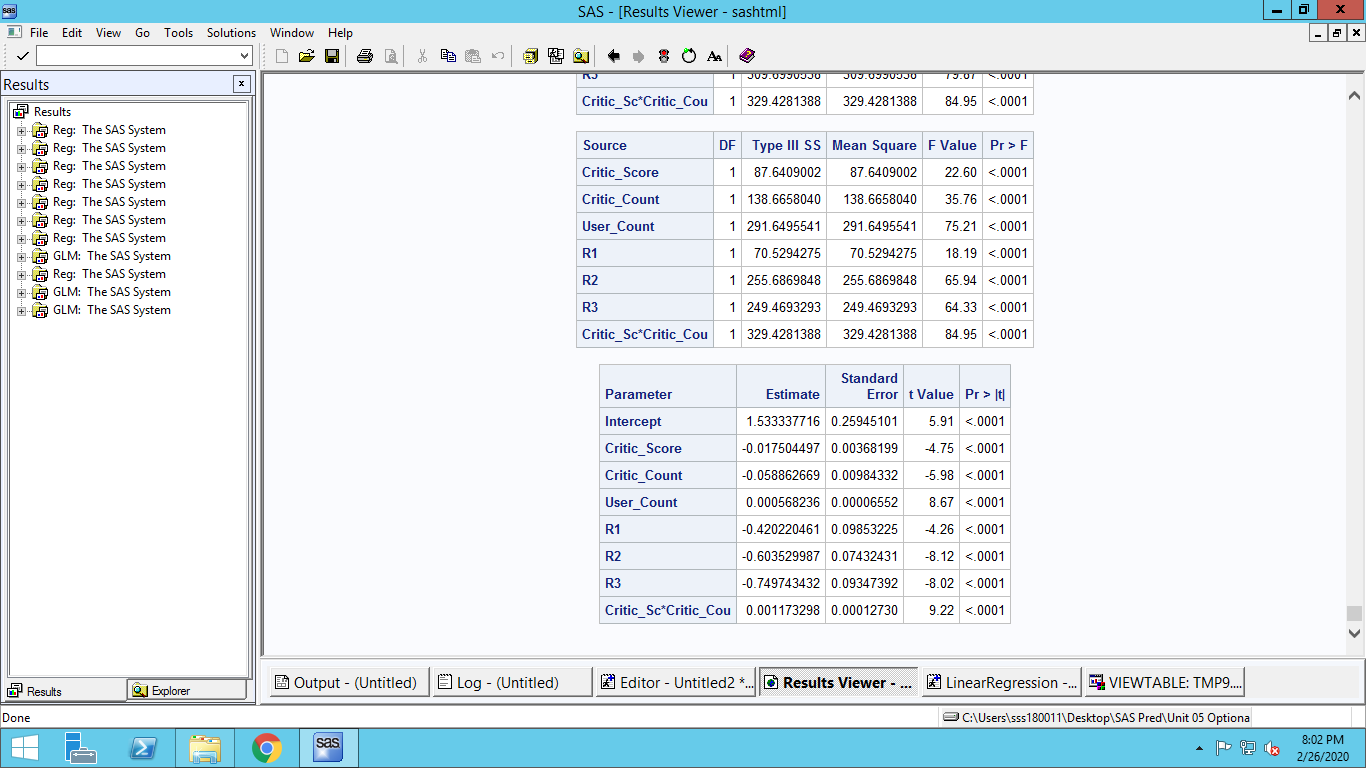
**R2 = Teen. R3 = Mature**

**Segment Response models were created to see interaction effects**

**Interaction between Number of reviews with Review score was explored**

**User\_Score \* User\_Count had to be removed as it is not statistically significant**

**Model R\_Square value still remains low**



**Step 2:**

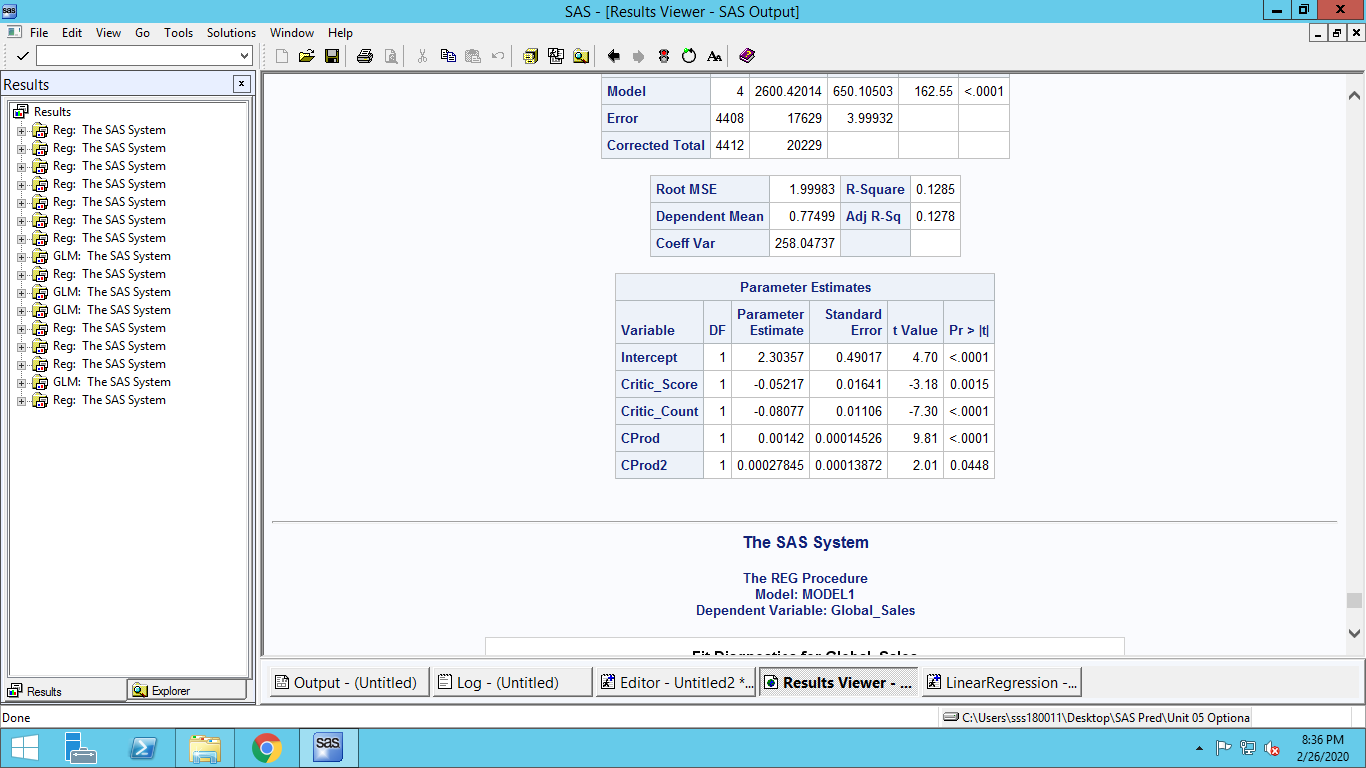
**Additional Nominal Explanatory variables created for Gaming Platforms**

**Assumption: Sales and reviews have non-linear relationship. Both log and polynomial relationship are explored**

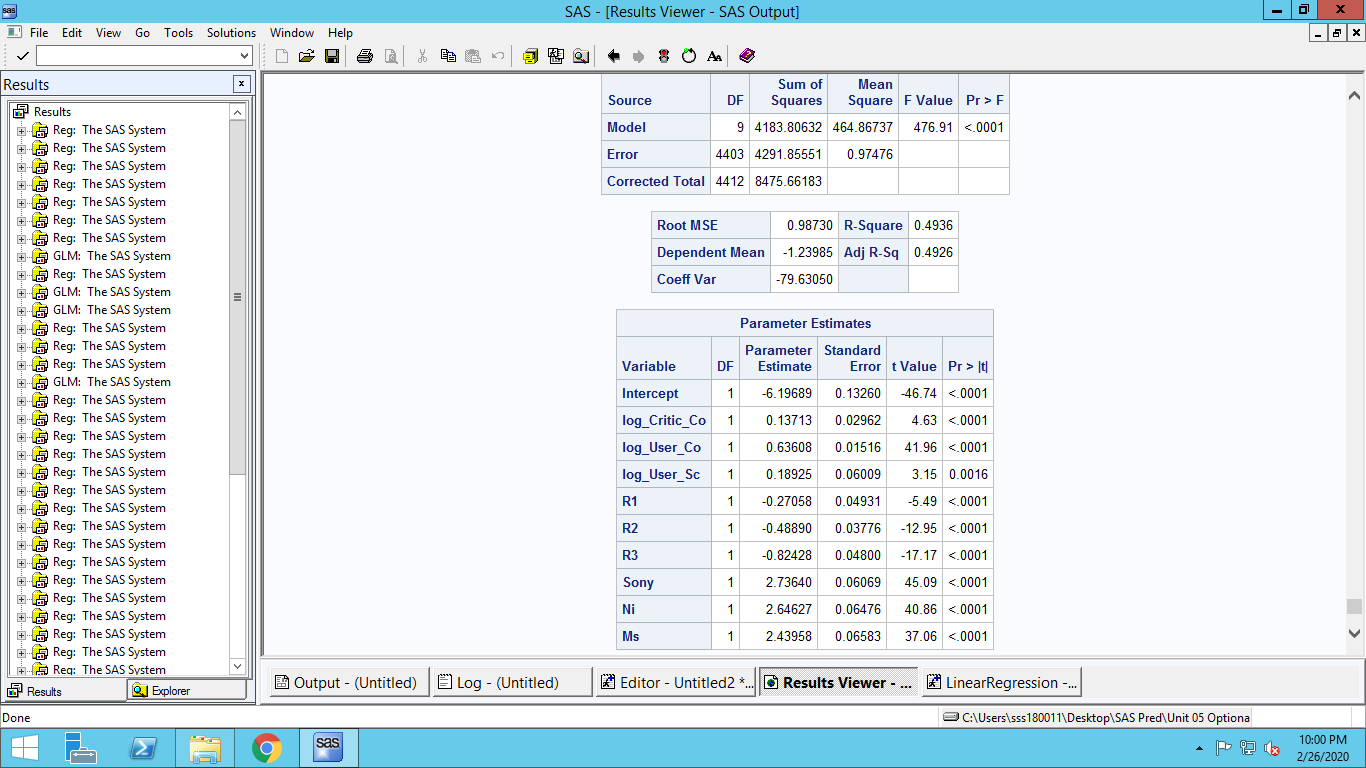


**Using non-linear log relationships did not increase R-square value any further**

**Using non-linear polynomial relationships also did not increase R-Square value any further**



**A log-log relationship finally gives higher R-Squared value**



* 1. Interpret the model results.

**All the model parameters are statistically significant.**

**Games which are rated E (for Everyone) increases sales than other ratings.**

**R1 has negatively affect sales than base. R2 affects even more. R3 most negatively affects. This is expected because with higher ratings eligible population base decreases.**

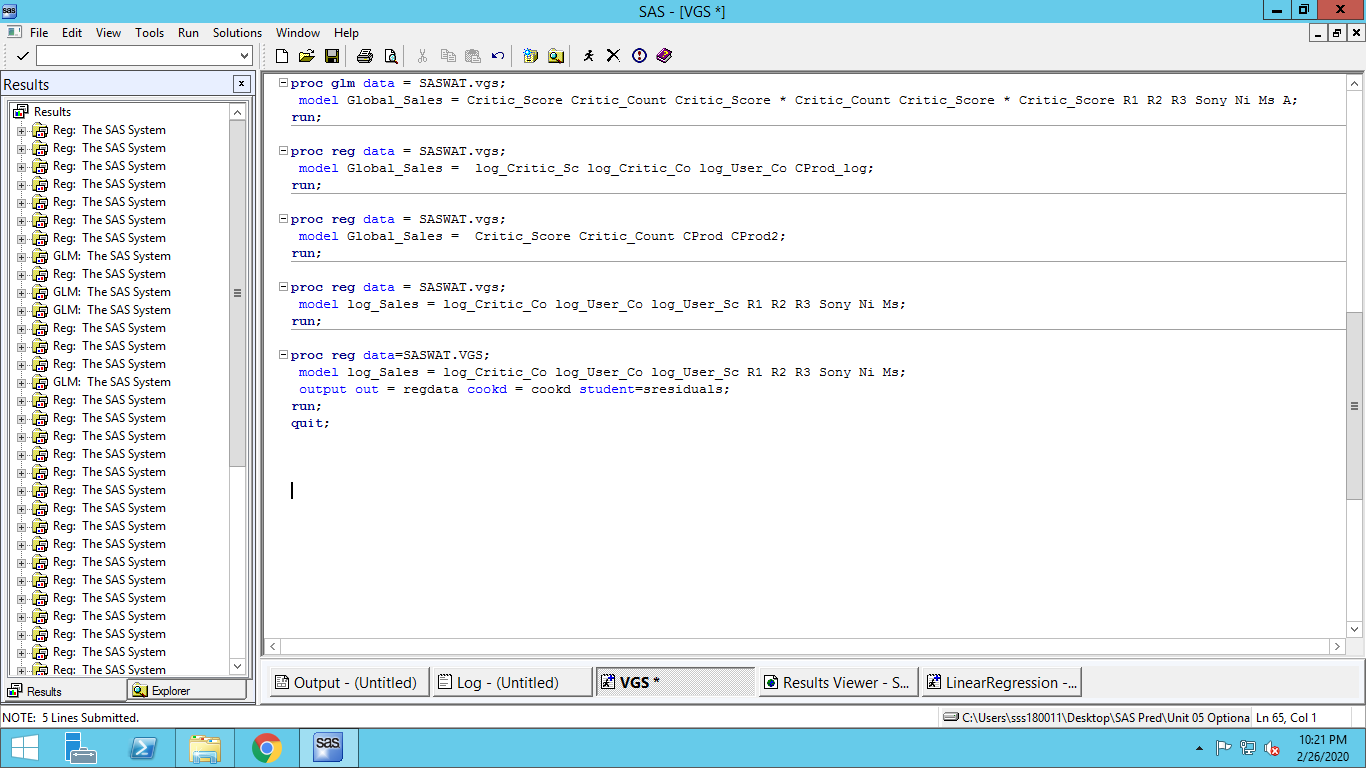
**Sony increases sales higher than other platforms. All platforms with respect to base PC increase sales positively.**

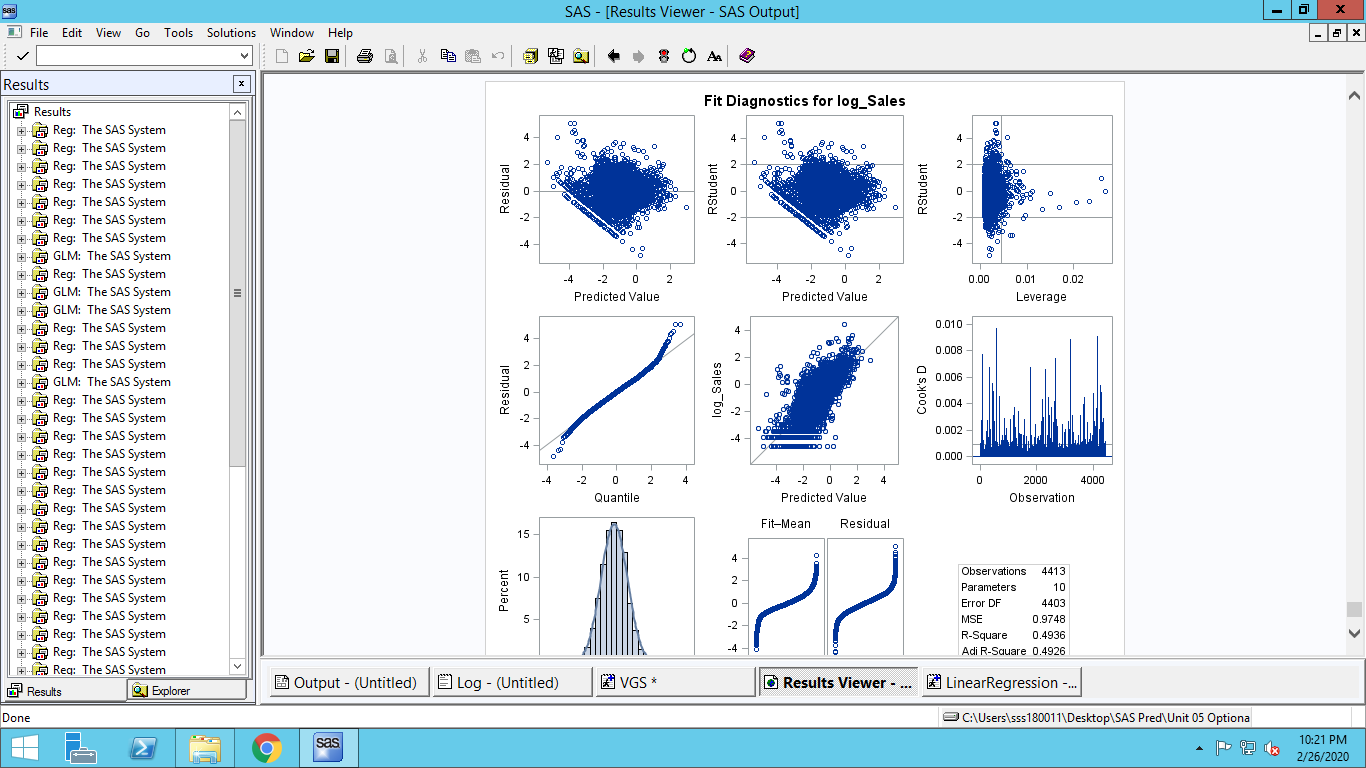
**The relationship between % Change in Global sales is explained by % change in Critic Counts, User Counts and User scores. Hence a log-log relationship is used.**

**Parameter estimates of Critic Counts, User Counts and User scores are all positive. Therefore, they all increase sales with increase in their values but increase stagnate at higher values.**

**Similarly, For global sales effects of independent variables vary for low and higher values. For games with already higher sales increase in independent values may not create the same effect as compared with games with less global sales**

1. For the model you constructed, 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 same. Discuss whether this change had a practically significant impact on your model results.
2. **Outliers and Influential observations**





**From the first plot we can see many residuals far away from mean 0 line. From second plot we can see many Studentized residuals fall outside 2 standard deviation margin line.**

**In leverage plot no observation lies in high leverage and high residual quadrant, but many lie in low leverage and high residual regions**

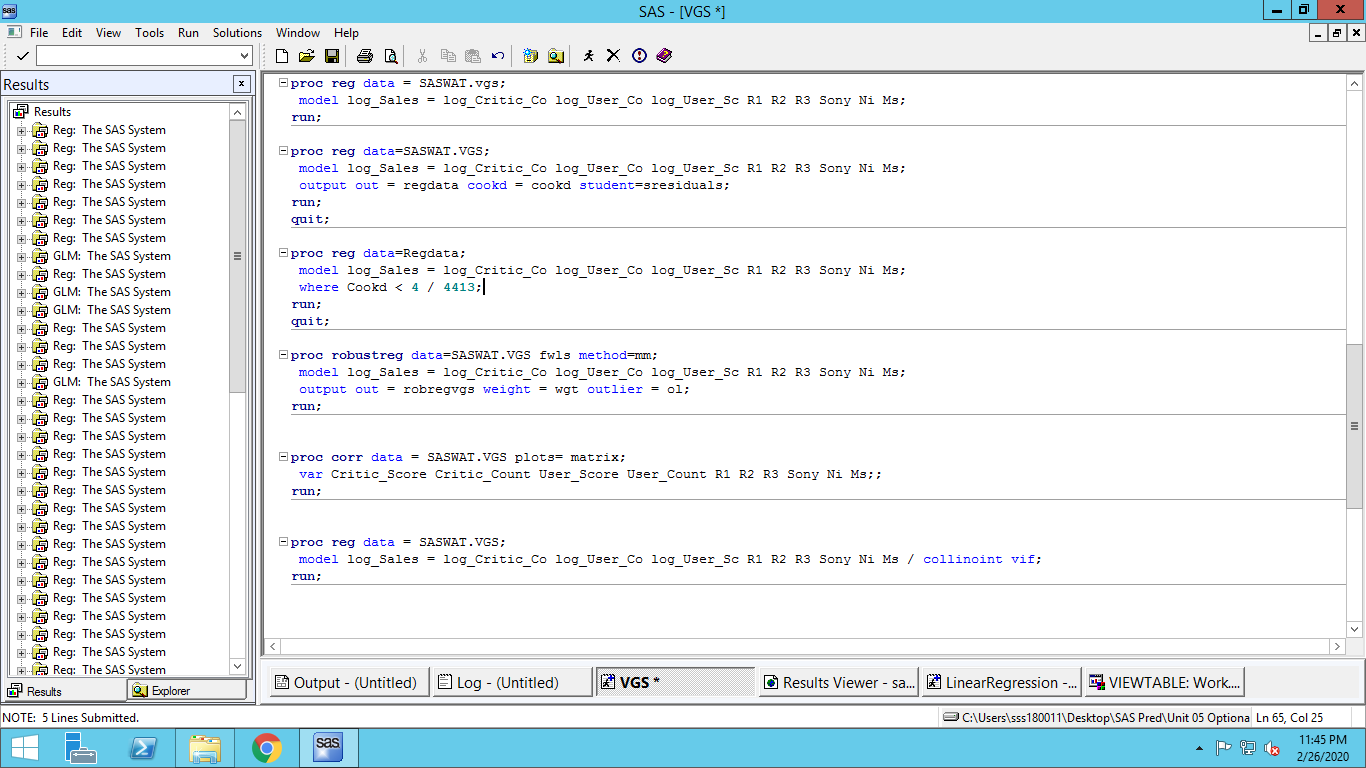
**In Cook’s plot many observations have high CookD value**

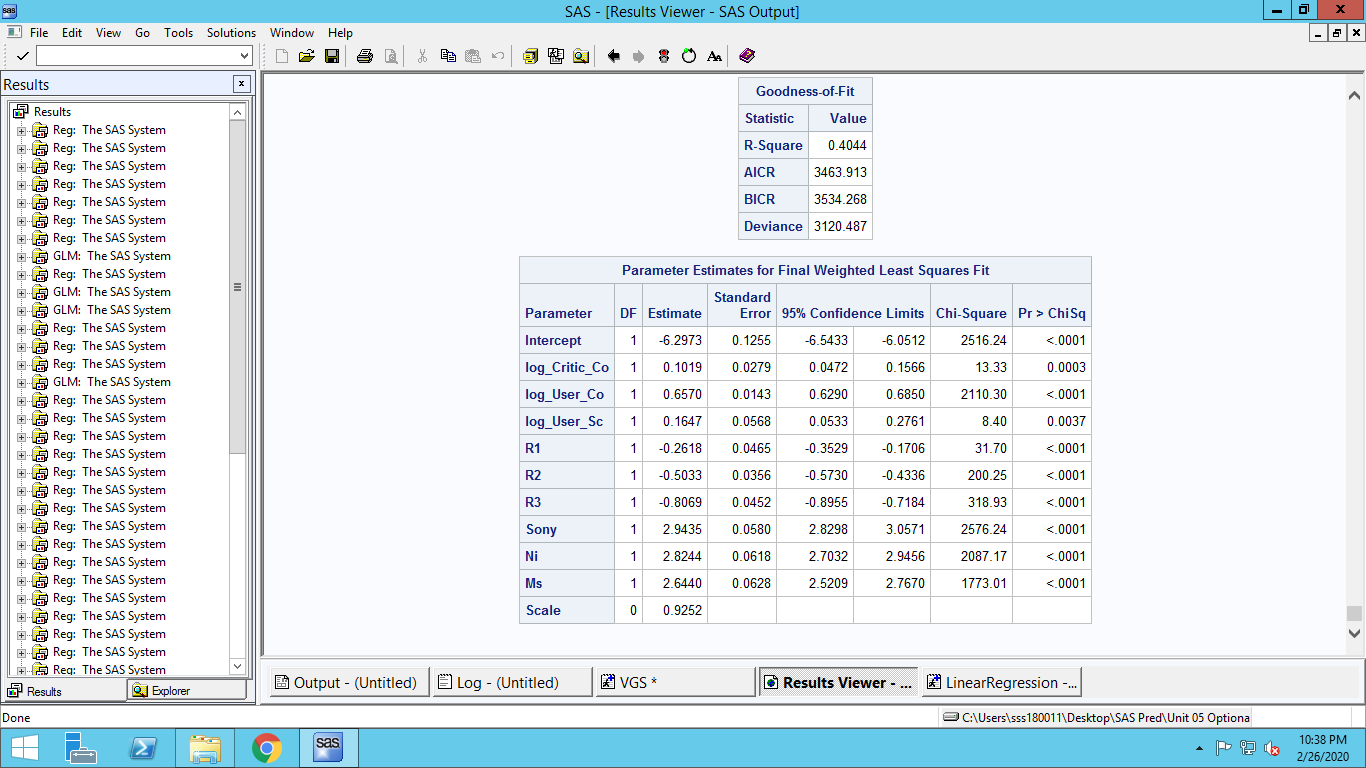
1. **Regression was rerun removing observations having higher Cookd value**



**This didn’t affect R-Square value so much. Neither it affected any Parameter estimates**

1. **Robust Regression**

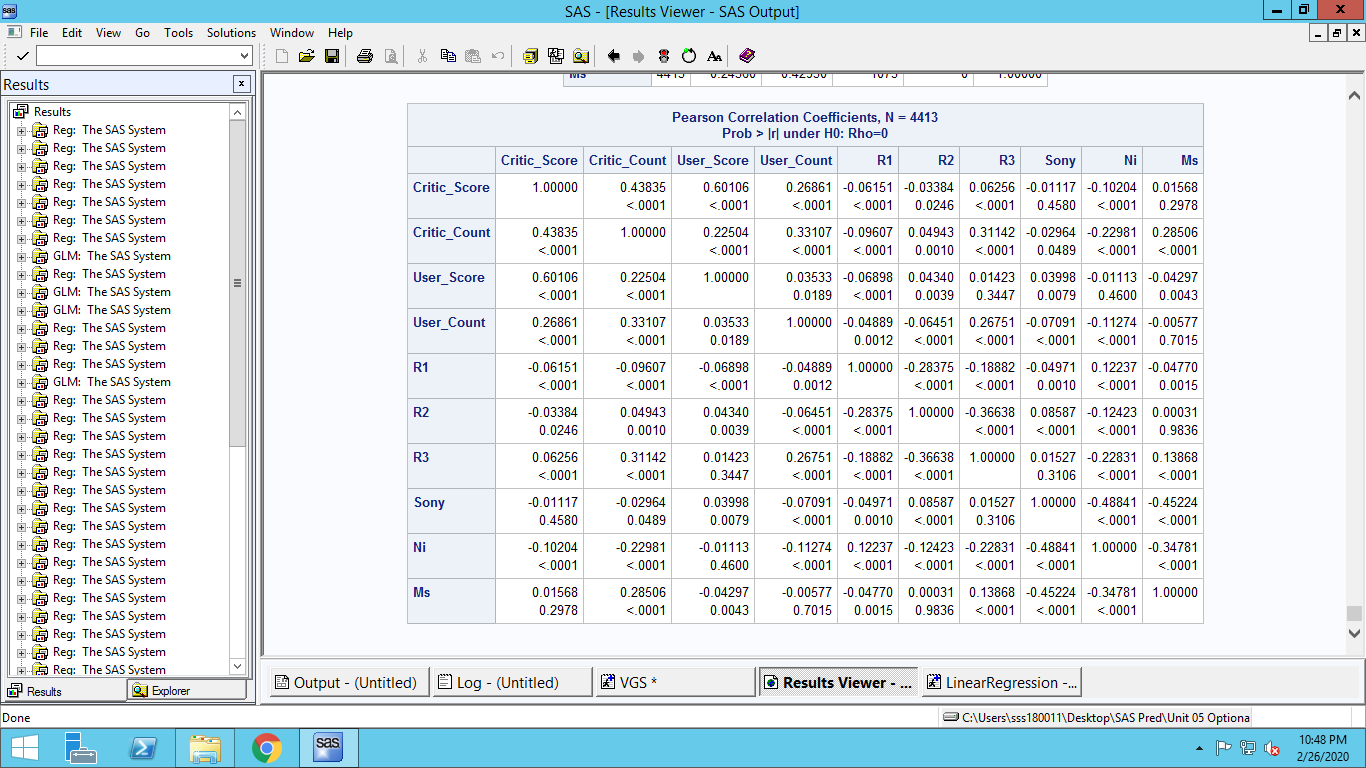




**Parameter estimates sign didn’t change anywhere. Neither did the values got affected by any significant value**

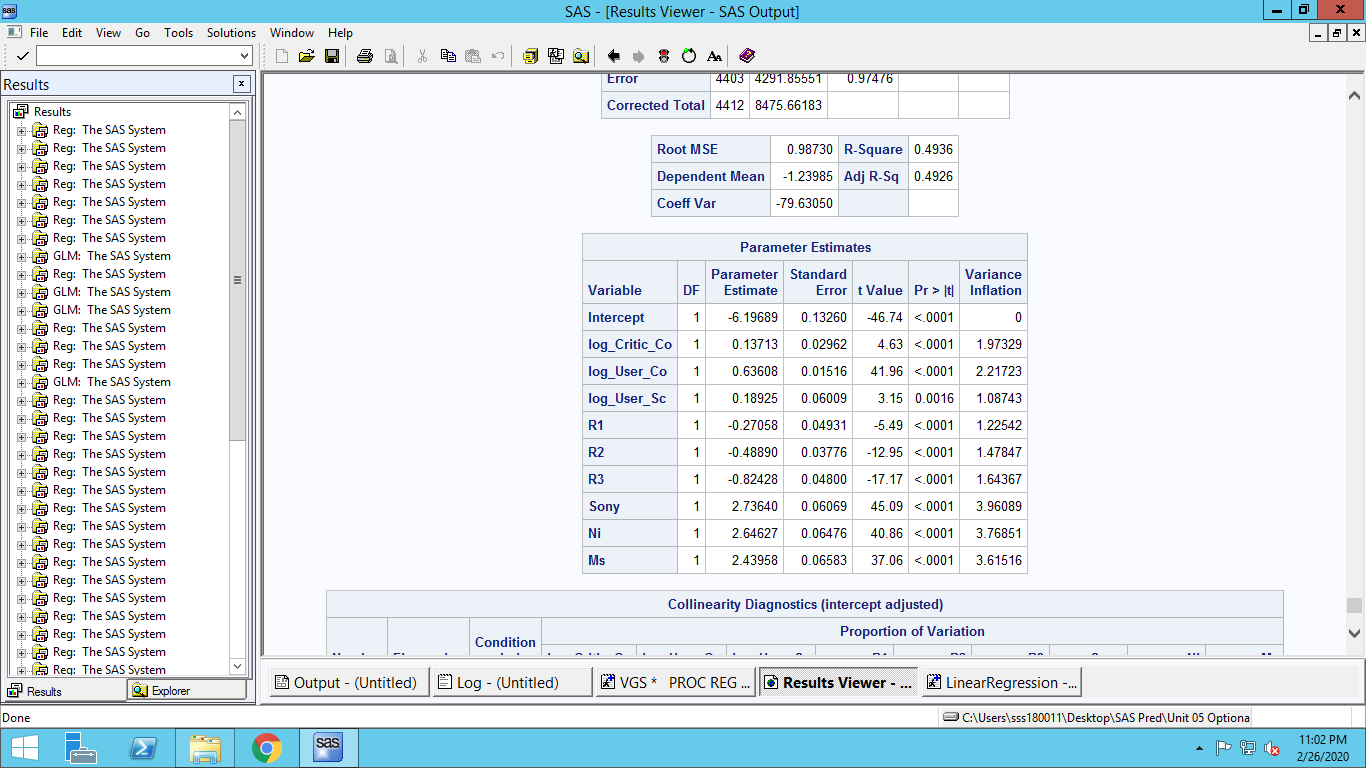
**All chi-Square tests show statistical significance**

1. **Multicollinearity amongst independent variables**
2. **Principal components for collinearity amongst variables**

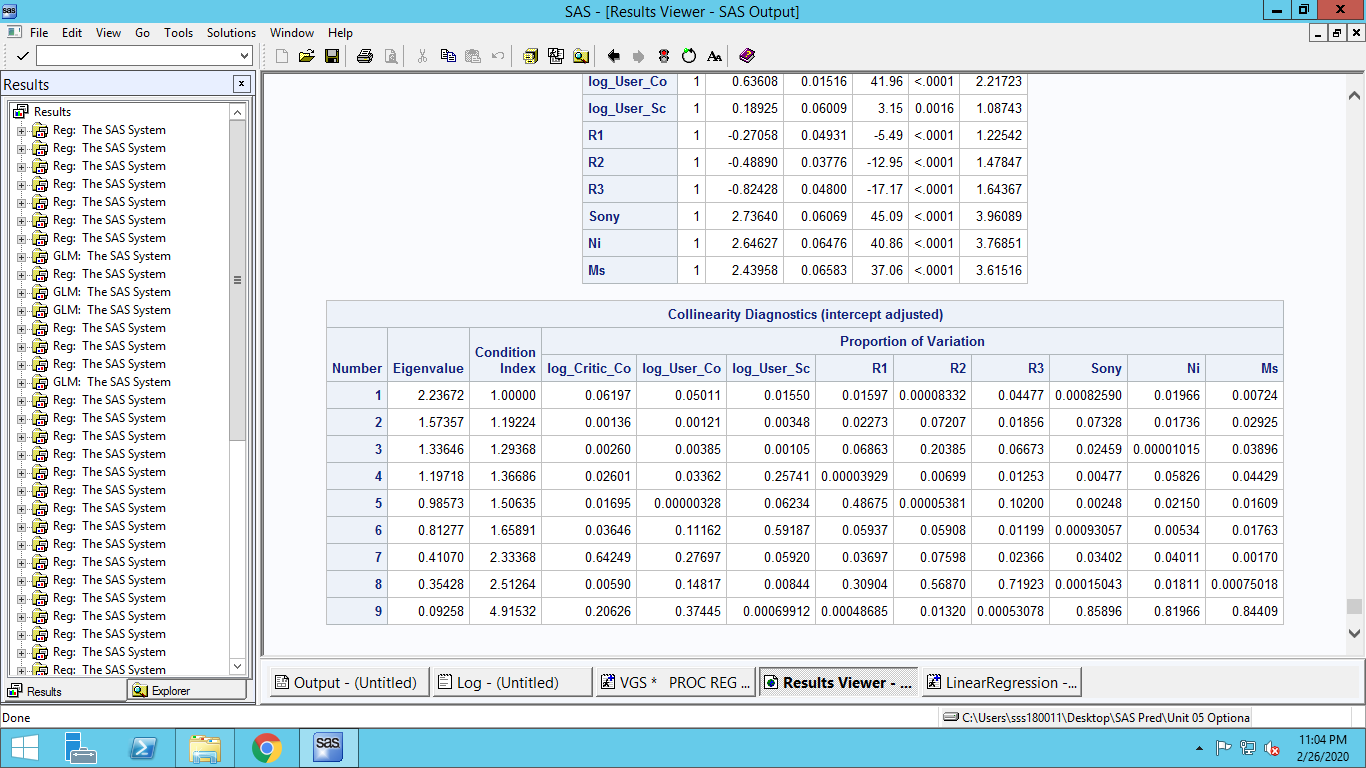


**Correlation noted between Critic score and User Score. Medium correlation noted between Critic count and Score. They are also statistically significant. All other correlation values are negligible.**

1. **Regression with collinearity diagnostics and vif**



**None of the values of variance inflation factors are over 10.Their effects can be ignored**

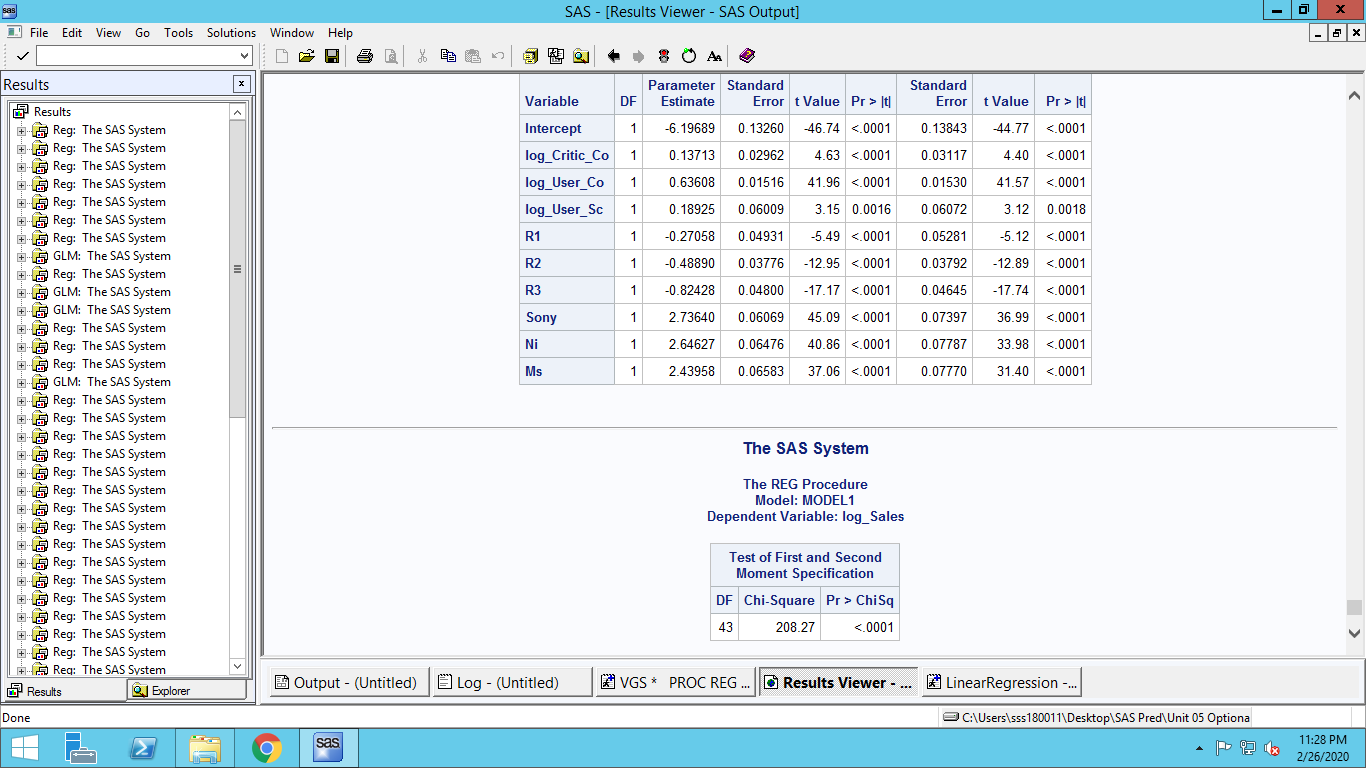


**None of the condition index is high enough to cause any problem.**

**So this model doesn’t have any multicollinearity problem**

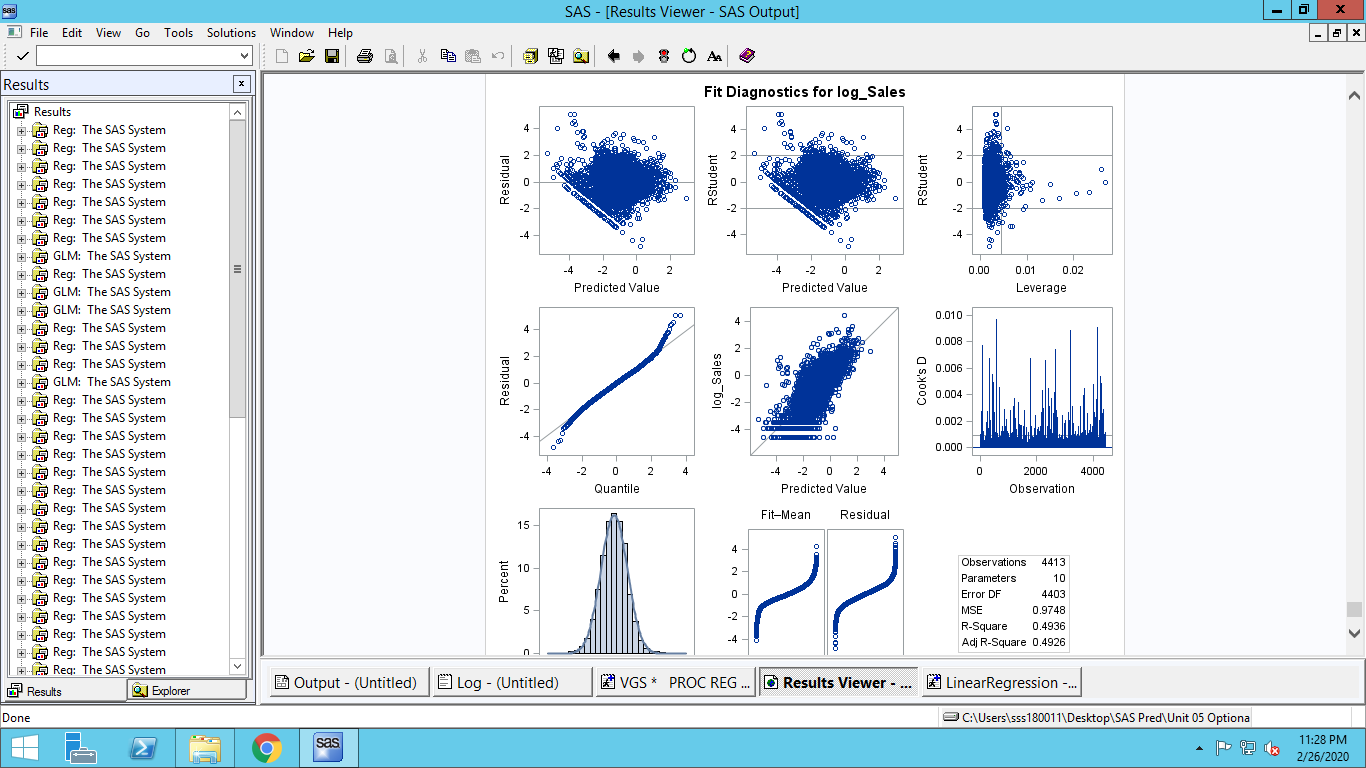
1. **Heteroscedasticity in error term**

**Heteroscedasticity Detection:**



**Parameters are statistically significant.**

**White’s test show that the null hypothesis which states that the regression assumptions are satisfied is rejected. This means there is Heteroscedasticity problem**

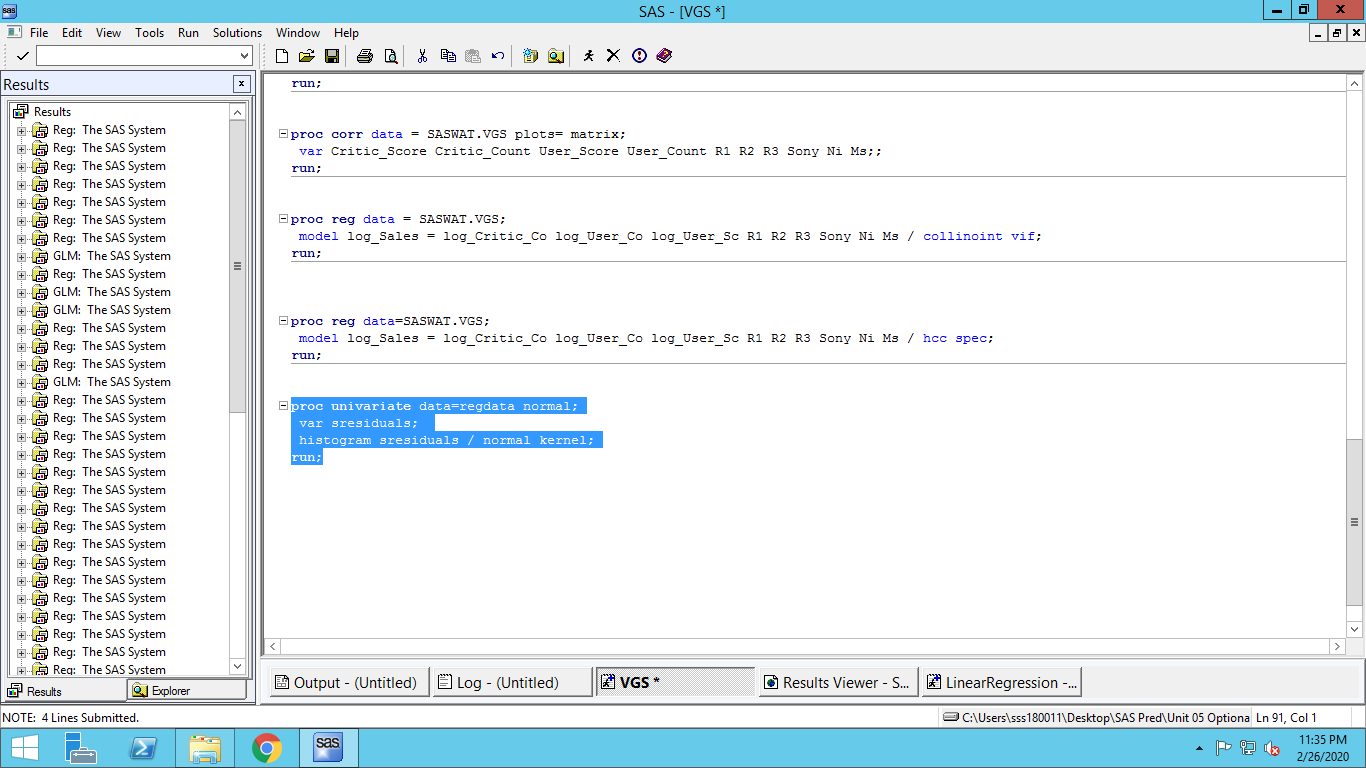


**For intermediate predicted values residuals are high. Studentized plot also shows the same conclusion**

**Standard errors and t-value are different for Heteroscedasticity column.**

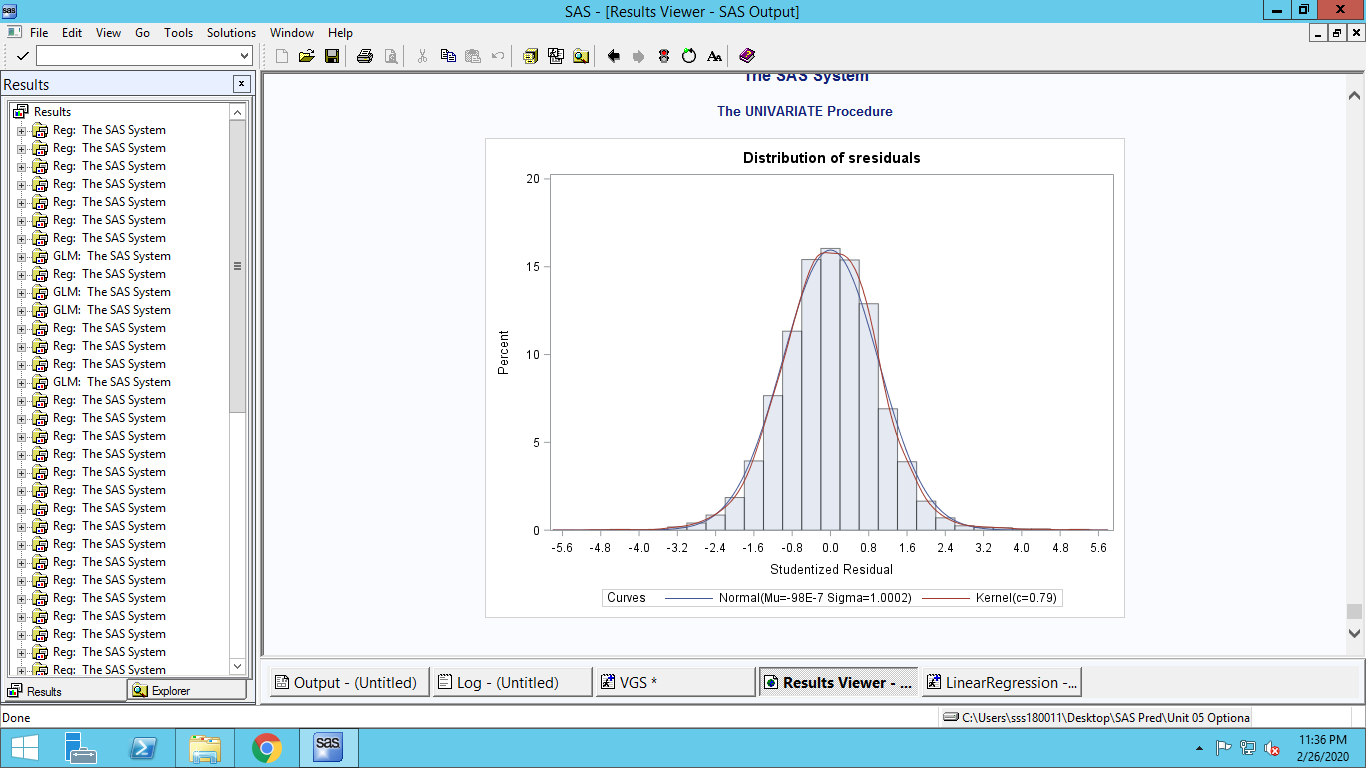
**This problem still persists even though we have used log transformations in our model**

1. **Normality of error term**



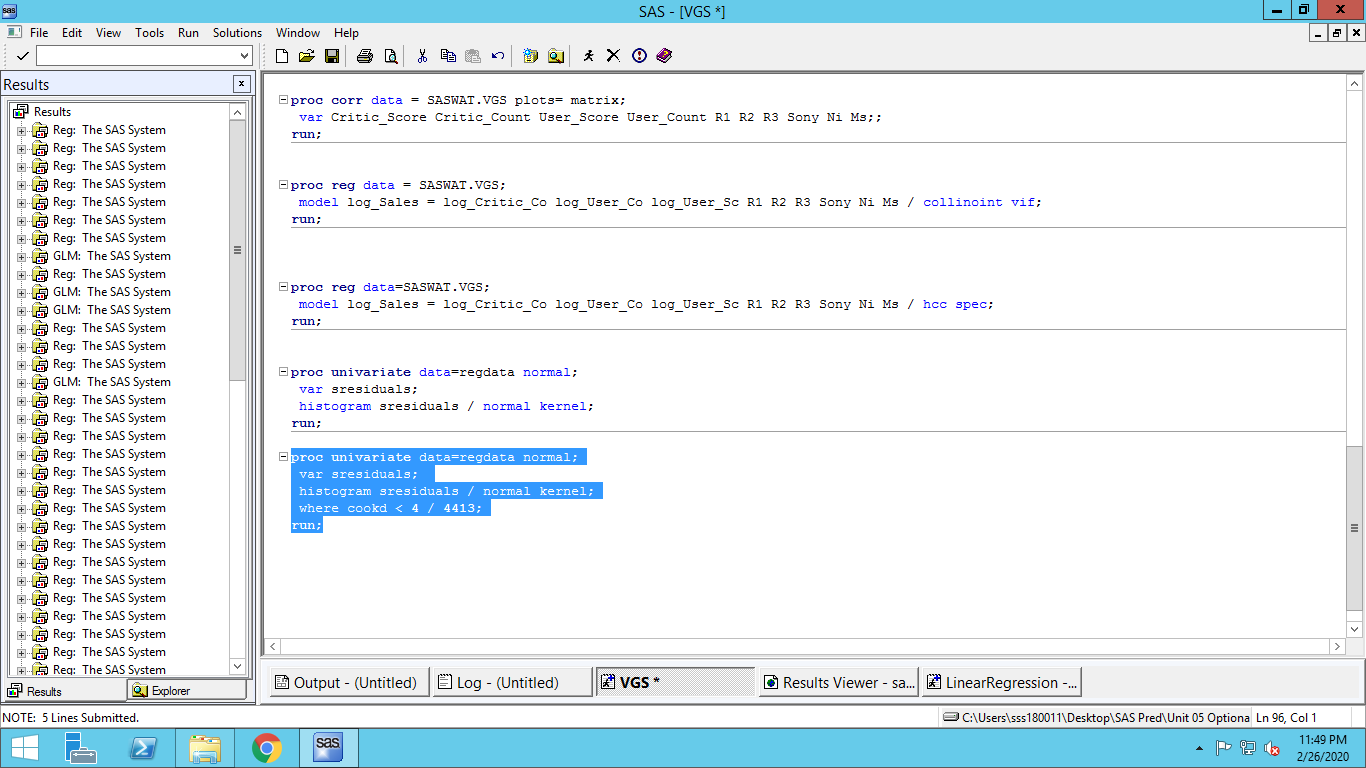


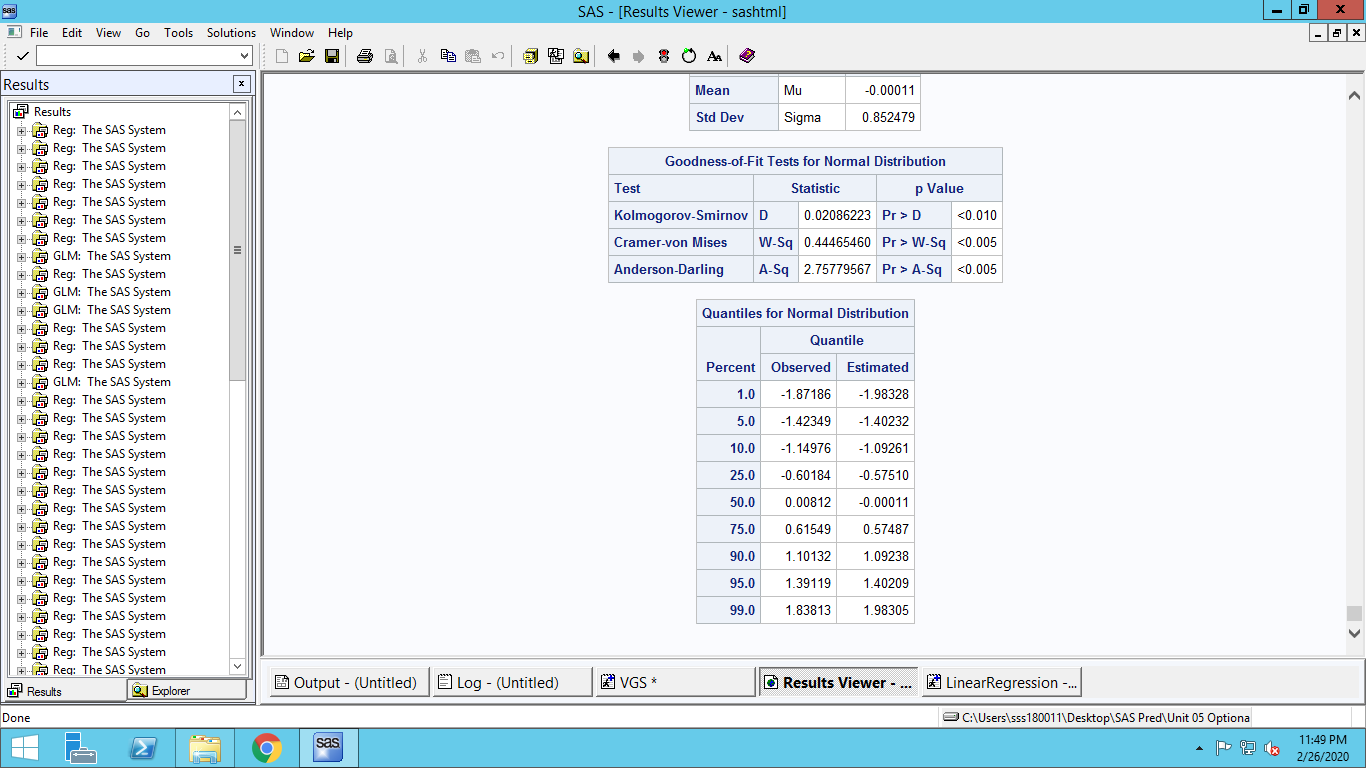
**We reject the null hypothesis that distribution follows normal distribution**

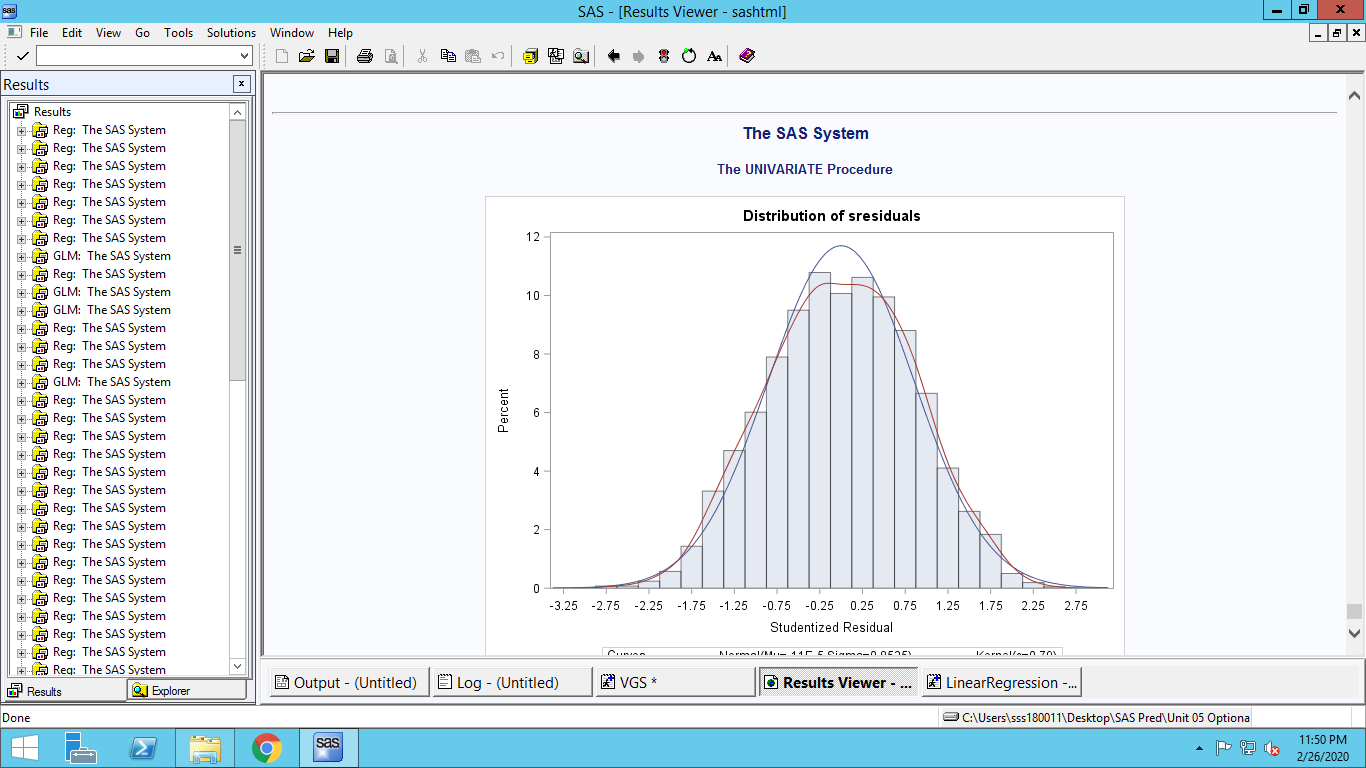


**The graph shows near normality, even though statistical tests refute this**

**Even after removing cookD’s values the error terms fail to achieve normality**







**The graph shows similarity with normal distribution**