BUAN 6337 Case Study

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).

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

Question 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.

a) Present the final model and results.
b) 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).
c) Interpret the model results.

Answer

a)

Our final model used user and critic scores, user and critic counts, and platforms as the independent variables and global sales as the dependent variable. We saw that this model passed the P-value test and most variables were significant. However, it still struggled with the R-square and we believe that was due to regression assumptions being violated. We did observe R-square increasing as we made changes to the model (from .05 to .16). We have further improvised the model fit by addressing the skewness of the dependent variable in Q2 which has increased the Adj R-square to 0.4237

Final Regression Model:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	3268.47303	251.42100	65.21	<.0001
Error	4399	16961	3.85563		
Corrected Total	4412	20229			

Root MSE	1.96358	R-Square	0.1616
Dependent Mean	0.77499	Adj R-Sq	0.1591
Coeff Var	253.36957		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.33057	0.20258	-6.57	<.0001
Critic_Score	1	0.02173	0.00300	7.25	<.0001
User_Score	1	-0.07697	0.02760	-2.79	0.0053
Critic_Count	1	0.02425	0.00212	11.46	<.0001
User_Count	1	0.00082364	0.00006463	12.74	<.0001
DS	1	0.70089	0.14012	5.00	<.0001
GBA	1	0.55522	0.17193	3.23	0.0012
GC	1	0.26251	0.14958	1.76	0.0793
PC	1	-0.63859	0.13909	-4.59	<.0001
PS2	1	0.59655	0.11442	5.21	<.0001
PS3	1	0.27856	0.12790	2.18	0.0295
PSP	1	0.23465	0.14582	1.61	0.1076
Wii	1	1.27848	0.13802	9.26	<.0001
X360	1	0.11625	0.12802	0.91	0.3639

b)

Our first model was a basic regression analysis that considered Critic_Score and User_Score (the independent variables) on Global Sales (the dependent variable). The resulting model passed the P-value test which rejects the null that the variances are 0 and can conclude that the model is significant, but R-square looked to be weak (we want to see a number closer to 1) which implies it does not explain the variation in the data very well. The P-Values of the parameter estimates also look good stating that the variables are statistically significant with each other and do have an impact on global sales. However, we can see that the User Score variable causes a decrease in global sales for every 1-point increase in user score, and we wanted to explore this further by adding more variables as this does not make sense.

Regression Model with Critic Score and User Score as Independent Variables:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	995.22810	497.61405	114.09	<.0001
Error	4410	19234	4.36149		
Corrected Total	4412	20229			

Root MSE	2.08842	R-Square	0.0492
Dependent Mean	0.77499	Adj R-Sq	0.0488
Coeff Var	269.47848		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.19103	0.17992	-6.62	<.0001
Critic_Score	1	0.03928	0.00281	13.99	<.0001
User_Score	1	-0.10666	0.02771	-3.85	0.0001

The second variation included User_Count and Critic_Count as additional independent variables. It seems this strengthened the original outcomes of the prior model, but R-Square was still relatively weak and User_Score was now not showing no relationship based on the P-value and still negative, so we wanted to try a variation of other variables in attempt 3 to see if these significantly contributed to revenue and would smooth out the negative impact of user score. It is interesting to note that this model says User_Score does not have an impact on global sales and User_Count has minor impact on global sales. It is appearing Critic_Score and Critic_Count are influence global sales where as User_Score and User_Count do not or have much less of an impact.

Regression Model Including Critic_Count & User_Count:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2409.65359	602.41340	149.02	<.0001
Error	4408	17820	4.04259		
Corrected Total	4412	20229			

Root MSE	2.01062	R-Square	0.1191
Dependent Mean	0.77499	Adj R-Sq	0.1183
Coeff Var	259.43980		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.75860	0.17484	-4.34	<.0001
Critic_Score	1	0.01506	0.00300	5.02	<.0001
User_Score	1	-0.04013	0.02704	-1.48	0.1379
Critic_Count	1	0.02351	0.00189	12.43	<.0001
User_Count	1	0.00066670	0.00006312	10.56	<.0001

We then compared just Critic_Score and Critic_Count and see both variables pass the P-value test but R-square is still pretty low.

Regression model with just Critic_Score and Critic_Count:

The REG Procedure
Model: MODEL1
Dependent Variable: Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1918.15682	959.07841	230.98	<.0001
Error	4410	18311	4.15221		
Corrected Total	4412	20229			

Root MSE	2.03770	R-Square	0.0948
Dependent Mean	0.77499	Adj R-Sq	0.0944
Coeff Var	262.93370		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.19445	0.15707	-7.60	<.0001
Critic_Score	1	0.01631	0.00244	6.70	<.0001
Critic_Count	1	0.02863	0.00186	15.42	<.0001

For the third variation we thought the platform could have an effect on sales and wanted to compare these different variables. Xbox (XB) is the variable that was left off (similar to Q4 in our lecture). R-square has improved, and we can see all models had higher sales than Xbox besides PC. The P-value for Game Cube, PSP, and Xbox 360 were above the rejection region of .05 which means sales from these platforms are not significantly different than Xbox and all other platform sales are. We also concluded that this would be our final model as we observed outliers and many residuals and wanted to focus on resolving issues in order to build a more accurate model supporting global sales. We anticipate that R-square will improve once changes are made.

Final Regression Model:

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	3268.47303	251.42100	65.21	<.0001
Error	4399	16961	3.85563		
Corrected Total	4412	20229			

Root MSE	1.96358	R-Square	0.1616
Dependent Mean	0.77499	Adj R-Sq	0.1591
Coeff Var	253.36957		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.33057	0.20258	-6.57	<.0001
Critic_Score	1	0.02173	0.00300	7.25	<.0001
User_Score	1	-0.07697	0.02760	-2.79	0.0053
Critic_Count	1	0.02425	0.00212	11.46	<.0001
User_Count	1	0.00082364	0.00006463	12.74	<.0001
DS	1	0.70089	0.14012	5.00	<.0001
GBA	1	0.55522	0.17193	3.23	0.0012
GC	1	0.26251	0.14958	1.76	0.0793
PC	1	-0.63859	0.13909	-4.59	<.0001
PS2	1	0.59655	0.11442	5.21	<.0001
PS3	1	0.27856	0.12790	2.18	0.0295
PSP	1	0.23465	0.14582	1.61	0.1076
Wii	1	1.27848	0.13802	9.26	<.0001
X360	1	0.11625	0.12802	0.91	0.3639

We did perform testing for non-linear effects of independent variables ie., user/critic score and count but did not perceive significant changes in the r-square or to the model overall. We have further improvised the model fit by addressing the skewness of the dependent variable in Q2 which has increased the Adj R-square to 0.4237

C)

Interpreting the results: Our model fell within the rejection region of the P-Value so we could state it as statistically significant. It did have a low R-Square value (.16 on a range of 0 to 1. 1 Being a strong representation of the variances) which we noted as an issue. We tested several variations, and this model had the strongest R-Square value. We also tested for non-linear effects among the key parameters and observed little change in R-Square. The key parameters Critic_Score, User_Score, Critic_Count, and User_Count all fell within the rejection region of the P-Value which meant the variables were statistically significant in changing global sales. It is notable that the Critic_Score and Critic_Count appeared to carry much more weight based on coefficient change than User_Score and User_Count. We further compared platforms which showed that the Wii sold the most and only the PC sold less than Xbox. Most of the platforms were within the rejection region confirming that they were a significantly different sales number than Xbox. Overall, we see the model to be significant, but it did have a high number of residuals which we believe the testing in part 2 will identify the source of issue and fixes will lead to a higher R-Square and a stronger model overall.

Question 2

For the final model you constructed in question 1, verify whether the various regression assumptions discussed in the lecture 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.

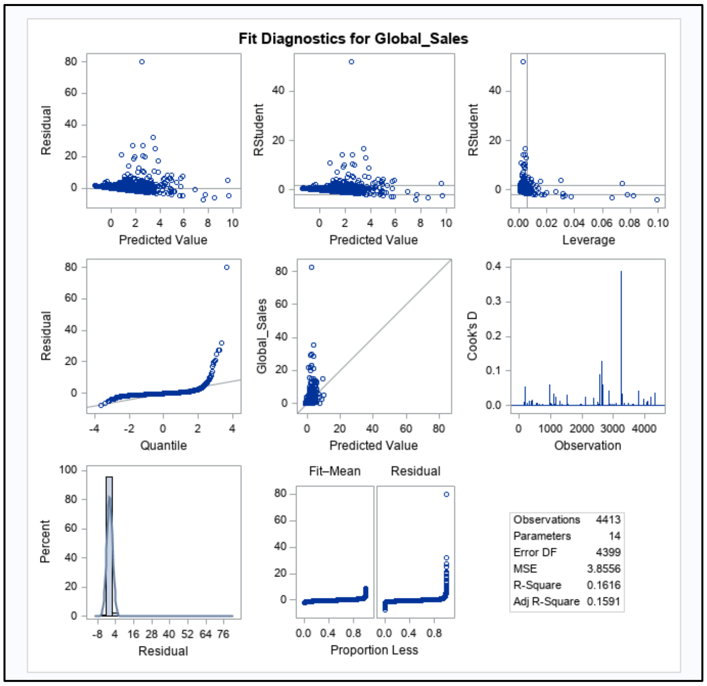
Answer

A) Outliers

We did observe numerous outliers (almost 10%!!!) and after applying the fix, the variable were all considered significant.

It can be seen that most observations fall within the predicted range, but several residuals fall way beyond the predicted range (graphs 1 and 2). Looking at Cooks D plot it can also be observed that many outliers exist. Outliers can either be handled by dropping observations that are beyond the CookD threshold or by running robust regression. We went for robust regression since it doesn't involve any loss of information.

Graphic Interpretation of Model:



After running robust regression, we see that all variables are now significant, but R-Square is still considerably low and went down a few points.

Robust Regression Image:

Diagnostics Summary		
Observation Type	Proportion	Cutoff
Outlier	0.0968	3.0000

Goodness-of-Fit	
Statistic	Value
R-Square	0.1273
AICR	3740.941
BICR	3848.301
Deviance	596.5340

Parameter Estimates for Final Weighted Least Squares Fit							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-0.3070	0.0351	-0.3757	-0.2382	76.58	<.0001
Critic_Score	1	0.0080	0.0005	0.0069	0.0090	230.02	<.0001
User_Score	1	-0.0183	0.0048	-0.0276	-0.0089	14.67	0.0001
Critic_Count	1	0.0061	0.0004	0.0053	0.0068	244.00	<.0001
User_Count	1	0.0002	0.0000	0.0002	0.0002	152.00	<.0001
DS	1	0.0979	0.0244	0.0500	0.1458	16.07	<.0001
GBA	1	0.1573	0.0299	0.0987	0.2159	27.67	<.0001
GC	1	0.0716	0.0255	0.0215	0.1217	7.86	0.0051
PC	1	-0.2691	0.0243	-0.3167	-0.2215	122.64	<.0001
PS2	1	0.1905	0.0198	0.1517	0.2293	92.65	<.0001
PS3	1	0.2385	0.0221	0.1952	0.2818	116.48	<.0001
PSP	1	0.1063	0.0249	0.0575	0.1550	18.26	<.0001
Wii	1	0.2549	0.0242	0.2075	0.3023	111.07	<.0001
X360	1	0.0998	0.0221	0.0564	0.1431	20.38	<.0001
Scale	0	0.3281					

B) Multi-collinearity

In running a correlation matrix we can see primary variables (Critic_Score, Critic_Count, User_Score, and User_Count) are moderately correlated, some have low correlation. This moderate to low correlation hints that we do not have a multicollinearity problem.

Correlation Matrix:

The SAS System

The CORR Procedure

4 Variables:	Critic_Score User_Score Critic_Count User_Count

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Critic_Score	4413	69.72558	14.01001	307699	17.00000	98.00000
User_Score	4413	7.24453	1.41985	31970	0.50000	9.50000
Critic_Count	4413	29.06005	18.38325	128242	4.00000	107.00000
User_Count	4413	136.88760	520.00640	604085	4.00000	9851

Pearson Correlation Coefficients, N = 4413
Prob > |r| under H0: Rho=0

	Critic_Score	User_Score	Critic_Count	User_Count
Critic_Score	1.00000	0.60106 <.0001	0.43835 <.0001	0.26861 <.0001
User_Score	0.60106 <.0001	1.00000	0.22504 <.0001	0.03533 0.0189
Critic_Count	0.43835 <.0001	0.22504 <.0001	1.00000	0.33107 <.0001
User_Count	0.26861 <.0001	0.03533 0.0189	0.33107 <.0001	1.00000

Furthermore, our variance inflation factors are all 1 or slightly higher which is very low (well below the threshold of 10), and we do not need to be worried about multicollinearity affecting our model. Additionally, our condition index is well below 10 which is a beginning indicator of multicollinearity with 100 or more being strong multicollinearity.

Model: MODEL1
Dependent Variable: Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2409.65359	602.41340	149.02	<.0001
Error	4408	17820	4.04259		
Corrected Total	4412	20229			

Root MSE	2.01062	R-Square	0.1191
Dependent Mean	0.77499	Adj R-Sq	0.1183
Coeff Var	259.43980		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.75860	0.17484	-4.34	<.0001	0
Critic_Score	1	0.01506	0.00300	5.02	<.0001	1.92839
User_Score	1	-0.04013	0.02704	-1.48	0.1379	1.60909
Critic_Count	1	0.02351	0.00189	12.43	<.0001	1.31811
User_Count	1	0.00066670	0.00006312	10.56	<.0001	1.17592

Collinearity Diagnostics (intercept adjusted)						
			Proportion of Variation			
Number	Eigenvalue	Condition Index	Critic_Score	User_Score	Critic_Count	User_Count
1	1.99492	1.00000	0.09855	0.07714	0.09525	0.05235
2	1.04575	1.38117	0.02186	0.19814	0.07945	0.41716
3	0.62229	1.79047	0.00930	0.05650	0.72710	0.45899
4	0.33704	2.43289	0.87029	0.66822	0.09820	0.07150

C) Heteroscedasticity

In checking for heteroscedasticity we do observe that the parameter estimates are mostly significant with User Score being the only insignificant parameter so we may have an issue.

The REG Procedure
Model: MODEL1
Dependent Variable: Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2409.65359	602.41340	149.02	<.0001
Error	4408	17820	4.04259		
Corrected Total	4412	20229			

Root MSE	2.01062	R-Square	0.1191
Dependent Mean	0.77499	Adj R-Sq	0.1183
Coeff Var	259.43980		

Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	-0.75860	0.17484	-4.34	<.0001	0.11715	-6.48	<.0001
Critic_Score	1	0.01506	0.00300	5.02	<.0001	0.00194	7.75	<.0001
User_Score	1	-0.04013	0.02704	-1.48	0.1379	0.01715	-2.34	0.0193
Critic_Count	1	0.02351	0.00189	12.43	<.0001	0.00268	8.78	<.0001
User_Count	1	0.00066670	0.00006312	10.56	<.0001	0.00013750	4.85	<.0001

Overall the results of the Whites test show that the regression assumptions are not satisfied by being able to reject the Null Hypothesis. This shows that we do have a Heteroscedasticity problem.

Whites Test:

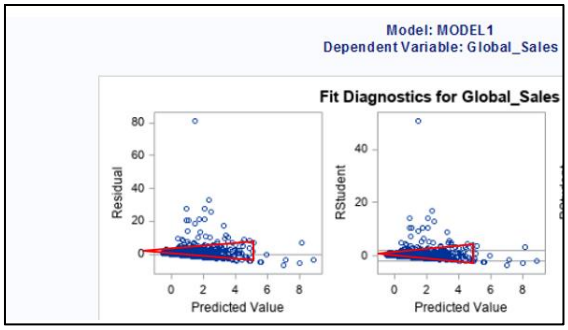
The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: Global_Sales

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
14	64.52	<.0001

In reviewing the first two graphs we can see that it does appear that the beginning residuals are expanding as you move further down the x-axis.

Visual Check for Heteroscedasticity:



Finally, the heteroscedasticity Consistent T-Values in all instances are quite different than the parameter T-Values, another indicator of heteroscedasticity.

Parameter Estimates Chart:

Parameter Estimates						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	-0.75860	0.17484	-4.34	<.0001	0.11715	-6.48	<.0001
Critic_Score	1	0.01506	0.00300	5.02	<.0001	0.00194	7.75	<.0001
User_Score	1	-0.04013	0.02704	-1.48	0.1379	0.01715	-2.34	0.0193
Critic_Count	1	0.02351	0.00189	12.43	<.0001	0.00268	8.78	<.0001
User_Count	1	0.00066670	0.00006312	10.56	<.0001	0.00013750	4.85	<.0001

To fix heteroscedasticity we can do so by transforming the dependent variable Global_Sales using the log transform of y variable process. We then see that the T-Values are much closer though not exact, and the P-values are much better. Additionally, observing the first few residual graphs shows a wider distribution that does not resemble the triangles observed earlier. However, it does still appear to slightly shape from left to right in the opposite direction. the Whites test is still within the rejection region which indicates we are still being affected by heteroscedasticity. It is nice to note that our R-Square is improving with a score of .4254 which is a much healthier model than what we started with.

Results of Log Transform of Y Variable:

The REG Procedure
Model: MODEL1
Dependent Variable: log_Global_Sales

Number of Observations Read	4413
Number of Observations Used	4413

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	3605.81417	277.37032	250.55	<.0001
Error	4399	4869.84766	1.10704		
Corrected Total	4412	8475.66183			

Root MSE	1.05216	R-Square	0.4254
Dependent Mean	-1.23985	Adj R-Sq	0.4237
Coeff Var	-84.86174		

Parameter Estimates						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	-3.92714	0.10855	-36.18	<.0001	0.10440	-37.62	<.0001
Critic_Score	1	0.03074	0.00161	19.14	<.0001	0.00157	19.63	<.0001
User_Score	1	-0.07172	0.01479	-4.85	<.0001	0.01451	-4.94	<.0001
Critic_Count	1	0.02014	0.00113	17.76	<.0001	0.00120	16.85	<.0001
User_Count	1	0.00048176	0.00003463	13.91	<.0001	0.00005903	8.16	<.0001
DS	1	0.55378	0.07508	7.38	<.0001	0.07986	6.93	<.0001
GBA	1	0.55702	0.09213	6.05	<.0001	0.10870	5.12	<.0001
GC	1	0.32624	0.08015	4.07	<.0001	0.07300	4.47	<.0001
PC	1	-1.35801	0.07453	-18.22	<.0001	0.07959	-17.06	<.0001
PS2	1	0.85804	0.06131	13.99	<.0001	0.05740	14.95	<.0001
PS3	1	0.79992	0.06853	11.67	<.0001	0.05517	14.50	<.0001
PSP	1	0.39035	0.07813	5.00	<.0001	0.07722	5.05	<.0001
Wii	1	1.09560	0.07396	14.81	<.0001	0.07462	14.68	<.0001
X360	1	0.45278	0.06860	6.60	<.0001	0.05790	7.82	<.0001

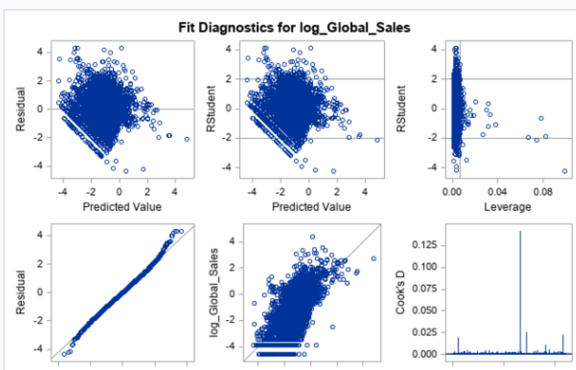
The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: log_Global_Sales

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
60	334.62	<.0001

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: log_Global_Sales



D) Normality of error term

In testing for normality of error term we see that in the Goodness of Fit Test in all of the tests the P-Value are significant. We reject the null hypothesis that the residuals follow the normal distribution and thus we do have a normality of error problem. The histogram also shows that the distribution is effected by outliers.

Reviewing Residuals Information with Proc Univariate:

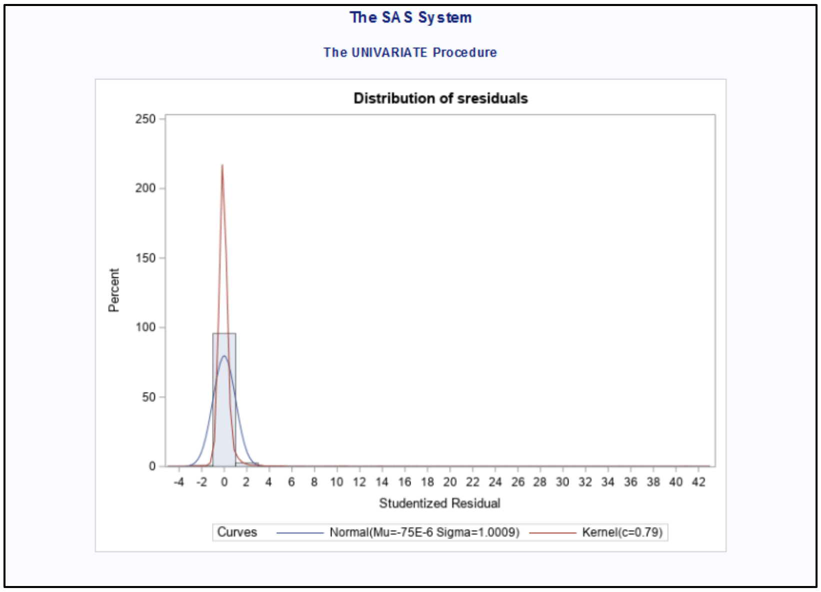
The SAS System

The UNIVARIATE Procedure
Fitted Normal Distribution for sresiduals (Studentized Residual)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	-0.00008
Std Dev	Sigma	1.000875

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.247135	Pr > D	<0.010
Cramer-von Mises	W-Sq	110.452658	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	591.630538	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	-0.95553	-2.32846
5.0	-0.62567	-1.64637
10.0	-0.49741	-1.28275
25.0	-0.30233	-0.67515
50.0	-0.09345	-0.00008
75.0	0.12873	0.67500
90.0	0.40256	1.28260
95.0	0.70093	1.64622
99.0	2.58538	2.32831



We run a Cooks D Test to remove the outliers and see a much more normal distribution. However, we still reject the null hypothesis due to P=Values falling in the rejection region for all 3 tests. This normality of error problem can also be commonly solved by A Log Transformation of the Y Variable which we performed earlier when checking for heteroscedasticity and saw a positive result.

Reviewing Residuals Information with Proc Univariate after Removing High Cooks D Values:

The SAS System

The UNIVARIATE Procedure
Fitted Normal Distribution for sresiduals (Studentized Residual)

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	-0.06399
Std Dev	Sigma	0.392712

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.0722008	Pr > D	<0.010
Cramer-von Mises	W-Sq	7.1472213	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	46.2542243	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	-0.87638	-0.97758
5.0	-0.61545	-0.70995
10.0	-0.49388	-0.56727
25.0	-0.30310	-0.32887
50.0	-0.09678	-0.06399
75.0	0.12170	0.20089
90.0	0.36385	0.43929
95.0	0.59289	0.58196
99.0	1.29701	0.84959

The SAS System

The UNIVARIATE Procedure