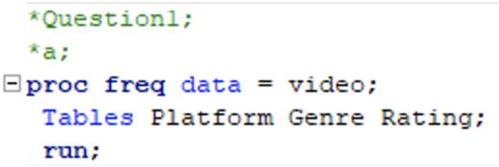
Homework 3

Jeffrey Chang

Question 1. (a)

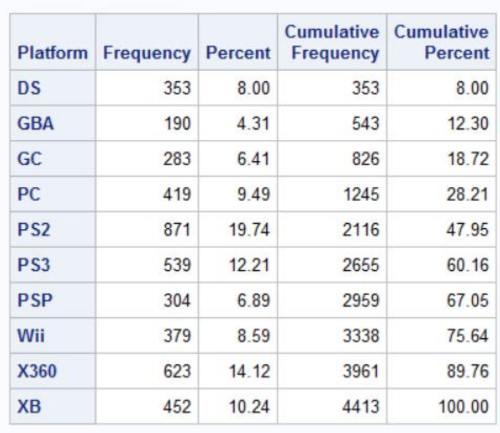
Used proc freq to create a frequency table of 3 variables: platform, genre, and rating.

Code:

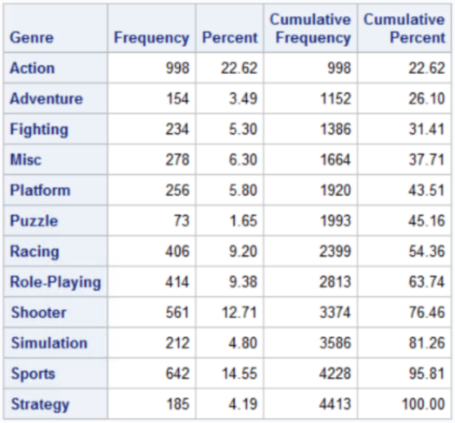


```
*Question1;
*a;
proc freq data = video;
 Tables Platform Genre Rating;
 run;
```

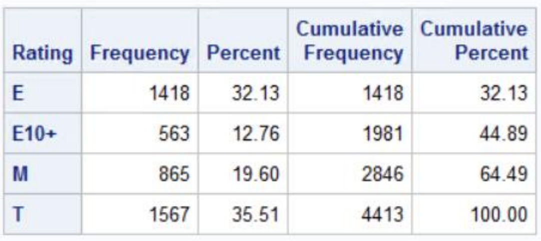
Output:



Platform	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DS	353	8.00	353	8.00
GBA	190	4.31	543	12.30
GC	283	6.41	826	18.72
PC	419	9.49	1245	28.21
PS2	871	19.74	2116	47.95
PS3	539	12.21	2655	60.16
PSP	304	6.89	2959	67.05
Wii	379	8.59	3338	75.64
X360	623	14.12	3961	89.76
XB	452	10.24	4413	100.00



Genre	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Action	998	22.62	998	22.62
Adventure	154	3.49	1152	26.10
Fighting	234	5.30	1386	31.41
Misc	278	6.30	1664	37.71
Platform	256	5.80	1920	43.51
Puzzle	73	1.65	1993	45.16
Racing	406	9.20	2399	54.36
Role-Playing	414	9.38	2813	63.74
Shooter	561	12.71	3374	76.46
Simulation	212	4.80	3586	81.26
Sports	642	14.55	4228	95.81
Strategy	185	4.19	4413	100.00



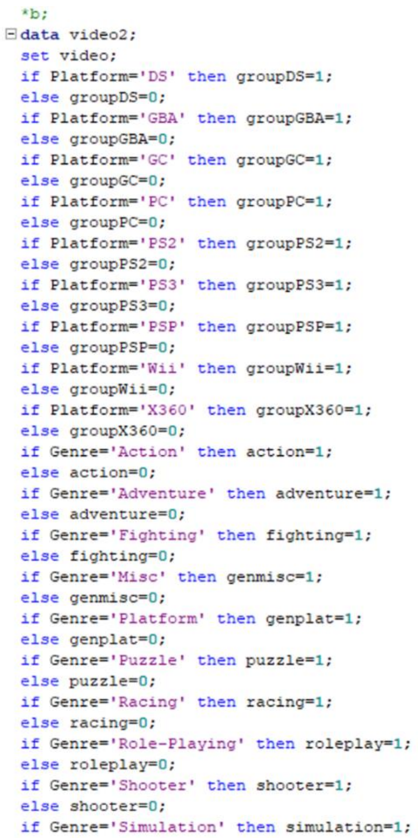
Rating	Frequency	Percent	Cumulative Frequency	Cumulative Percent
E	1418	32.13	1418	32.13
E10+	563	12.76	1981	44.89
M	865	19.60	2846	64.49
T	1567	35.51	4413	100.00

Question 1. (b and c)

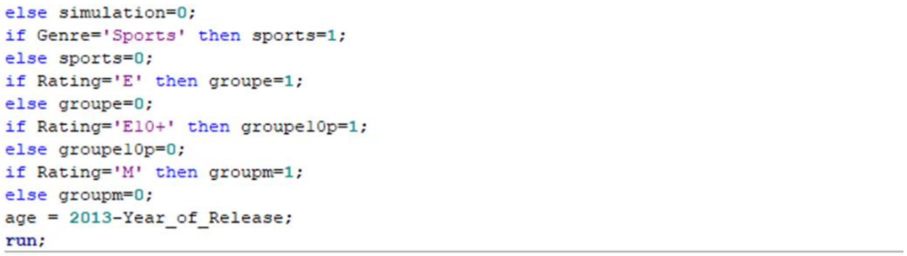
create categorial variables for platform, genre, and rating using if/else statements. Also create a variable for the age of the game relative to year 2013. Run a regression with all relevant X variables.

Adjusted R-squared is 0.1742.

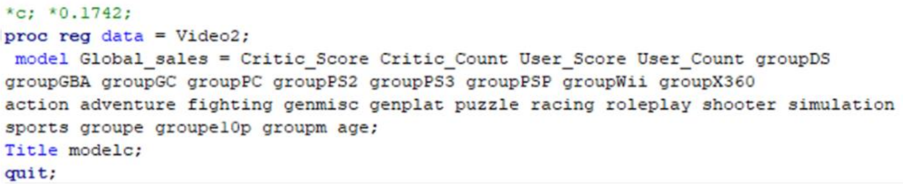
Code:



```
*b;
data video2;
set video;
if Platform='DS' then groupDS=1;
else groupDS=0;
if Platform='GBA' then groupGBA=1;
else groupGBA=0;
if Platform='GC' then groupGC=1;
else groupGC=0;
if Platform='PC' then groupPC=1;
else groupPC=0;
if Platform='PS2' then groupPS2=1;
else groupPS2=0;
if Platform='PS3' then groupPS3=1;
else groupPS3=0;
if Platform='PSP' then groupPSP=1;
else groupPSP=0;
if Platform='Wii' then groupWii=1;
else groupWii=0;
if Platform='X360' then groupX360=1;
else groupX360=0;
if Genre='Action' then action=1;
else action=0;
if Genre='Adventure' then adventure=1;
else adventure=0;
if Genre='Fighting' then fighting=1;
else fighting=0;
if Genre='Misc' then genmisc=1;
else genmisc=0;
if Genre='Platform' then genplat=1;
else genplat=0;
if Genre='Puzzle' then puzzle=1;
else puzzle=0;
if Genre='Racing' then racing=1;
else racing=0;
if Genre='Role-Playing' then roleplay=1;
else roleplay=0;
if Genre='Shooter' then shooter=1;
else shooter=0;
if Genre='Simulation' then simulation=1;
```



```
else simulation=0;
if Genre='Sports' then sports=1;
else sports=0;
if Rating='E' then groupe=1;
else groupe=0;
if Rating='E10+' then groupe10p=1;
else groupe10p=0;
if Rating='M' then groupm=1;
else groupm=0;
age = 2013-Year_of_Release;
run;
```

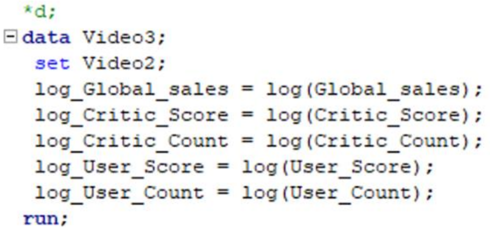


```
*c; *0.1742;
proc reg data = Video2;
 model Global_sales = Critic_Score Critic_Count User_Score User_Count groupDS
groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
sports groupe groupe10p groupm age;
Title modelc;
quit;
```

Question 1. (d)

Generate natural log of the following variables:
global sales, critic_score, critic_count, user_socre, user_count.

Code:

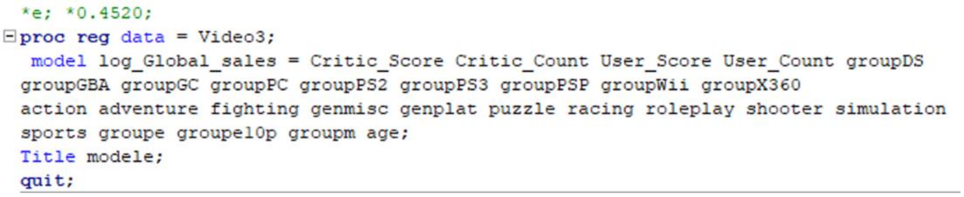


```
*d;
data Video3;
 set Video2;
 log_Global_sales = log(Global_sales);
 log_Critic_Score = log(Critic_Score);
 log_Critic_Count = log(Critic_Count);
 log_User_Score = log(User_Score);
 log_User_Count = log(User_Count);
run;
```

Question 1. (e)

Run a regression with the log of Y variable. Adjusted R-squared is 0.4520.

Code:

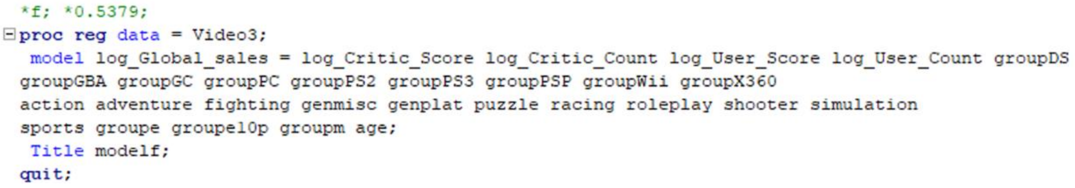


```
 *e; *0.4520;
proc reg data = Video3;
  model log_Global_sales = Critic_Score Critic_Count User_Score User_Count groupDS
 groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
 action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
 sports groupe groupe10p groupm age;
 Title modele;
 quit;
```

Question 1. (f)

Run a regression with the log of Y variable as well as log of X variables generated in part (d). Adjusted R-squared is 0.5379.

Code:



```
 *f; *0.5379;
proc reg data = Video3;
  model log_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count groupDS
 groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
 action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
 sports groupe groupe10p groupm age;
  Title modelf;
 quit;
```

Question 1. (g)

The Regression with the log of Y variable and log of X variable has the highest Adjusted R-squared which is 0.5379.

The first reason is to respond to skewness towards larger values; for example, when the score increase, the increase of the Global Sales is not uniform. The second is to show percent change or elasticity of the Global sales with respect to the critic and user reviews. From the coefficients of the log X values we can see that the Global sales is highly elastic to changes in user and critic reviews score and number of reviews.

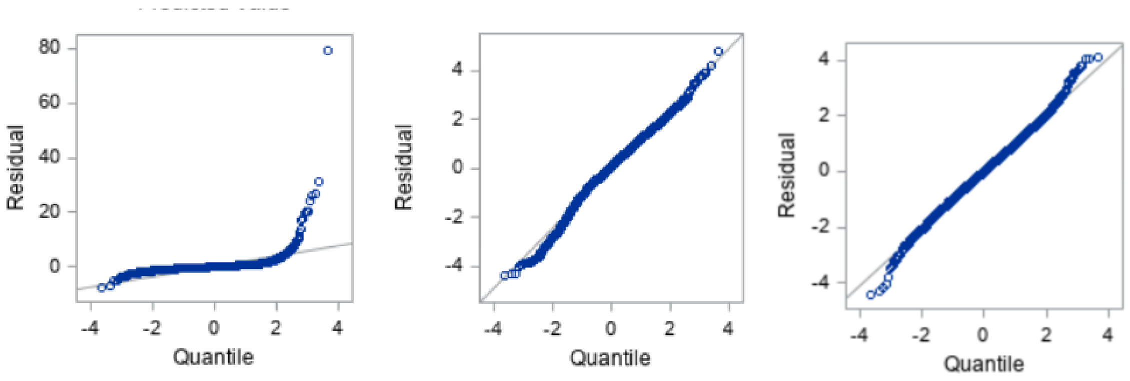
Interpretation of Coefficients from the log-log model:

1. For every 1% increase in critic review score the Global sales will increase by 0.81%
2. For every 1% increase in number of critics reviews the Global sales will increase by 0.115%
3. For every 1% increase in user review score the Global sales will decrease by 0.17%
4. For every 1% increase in number of users reviews the Global sales will increase by 0.63%
5. If the Platform for the game is the Nintendo DS then the Global sales will be 18% more than Microsoft Xbox games.
6. If the Platform for the game is the Nintendo GBA then the Global sales will be 22.6% more than Microsoft Xbox games.
7. If the Platform for the game is the Nintendo Game Cube then the Global sales will be 0.69% more than Microsoft Xbox games.
8. If the Platform for the game is the Personal Computers then the Global sales will be 245.77% less than Microsoft Xbox games.
9. If the Platform for the game is the Sony PS2 then the Global sales will be 43.7% more than Microsoft Xbox games.
10. If the Platform for the game is the Sony PS3 then the Global sales will be 13.26% more than Microsoft Xbox games.

11. If the Platform for the game is the Sony PSP then the Global sales will be 22.6% more than Microsoft Xbox games.
12. If the Platform for the game is the Nintendo Wii then the Global sales will be 45.68% more than Microsoft Xbox games.
13. If the Platform for the game is the Microsoft Xbox360 then the Global sales will be 11.12% less than Microsoft Xbox games.
14. If the Genre of the game is action then the Global sales will be 51.9% more than strategy games.
15. If the Genre of the game is adventure then the Global sales will be 15.31% more than strategy games.
16. If the Genre of the game is fighting then the Global sales will be 51.13% more than strategy games.
17. If the Genre of the game is miscellaneous then the Global sales will be 90.6% more than strategy games.
18. If the Genre of the game is platform then the Global sales will be 38.62% more than strategy games.
19. If the Genre of the game is puzzle then the Global sales will be 29.2% more than strategy games.
20. If the Genre of the game is racing then the Global sales will be 44.27% more than strategy games.
21. If the Genre of the game is roleplay then the Global sales will be 3.78% more than strategy games.
22. If the Genre of the game is shooter then the Global sales will be 35.59% more than strategy games.
23. If the Genre of the game is simulation then the Global sales will be 90.74% more than strategy games.
24. If the Genre of the game is sports then the Global sales will be 60.37% more than strategy games.
25. If the rating of the game is E then the Global sales will be 34.63% more than M rated games.
26. If the rating of the game is E10+ then the Global sales will be 18.45% more than M rated games.
27. If the rating of the game is T then the Global sales will be 27.83% lesser than M rated games.
28. For every one year increase in age the global sales will decrease by 0.41%

The following are the plots of three regression models. And we choose coefficients that minimize squared deviation from predicted values.

1. Regression with all relevant X variables.
2. Regression with the log of Y variable and log of X variables.
3. Regression with the log of Y variable.



Question 2.

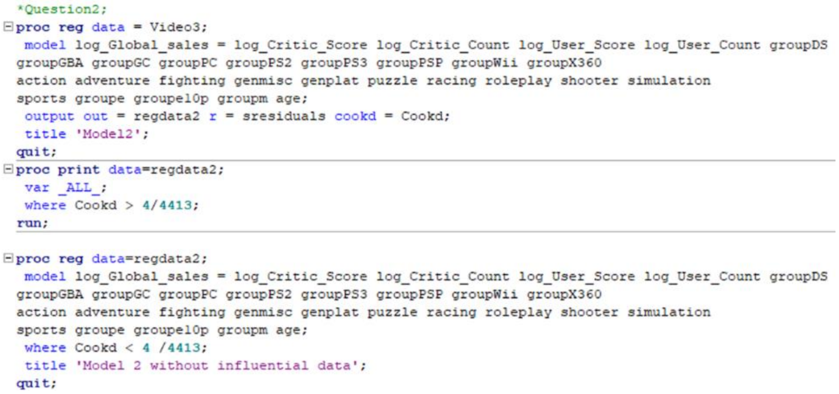
For the model you constructed in Q.1 part (g), verify whether the various regression assumptions discussed in class are satisfied. If an assumption is violated, discuss how it can be handled, and implement the solution if possible. Discuss whether the solution had a practically significant impact on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

Assumption 1 – Outlier

Method 1: COOK's Distance

We use COOK'D to measure of the influence of each observation and test the effect of deleting that observation on regression estimates.

Code:

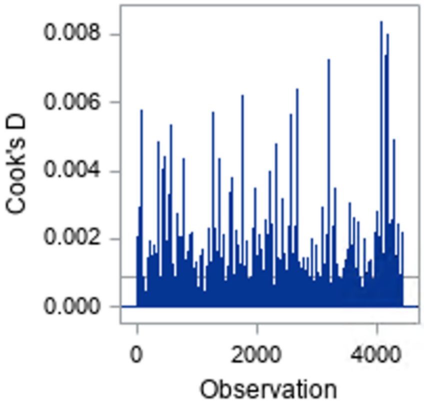


```
 *Question2;
proc reg data = Video3;
 model log_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count groupDS
groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
sports groupe groupe10p groupm age;
 output out = regdata2 r = sresiduals cookd = Cookd;
 title 'Model2';
quit;
proc print data=regdata2;
 var _ALL_;
 where Cookd > 4/4413;
run;

proc reg data=regdata2;
 model log_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count groupDS
groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
sports groupe groupe10p groupm age;
 where Cookd < 4 /4413;
 title 'Model 2 without influential data';
quit;
```

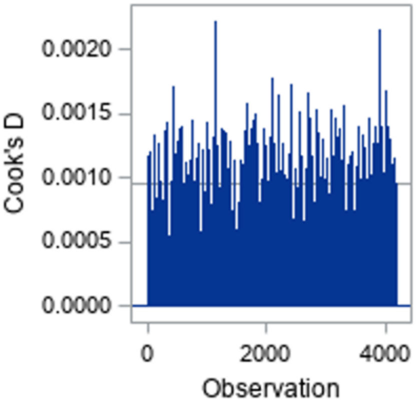
Output:

Before removing outlier



Adjusted R-squared is 0.5379.

After removing outlier



Adjusted R-squared is 0.6384.

Conclusion: based on the cook's distance, we identify the outlier and then use the rule of thumb to remove.

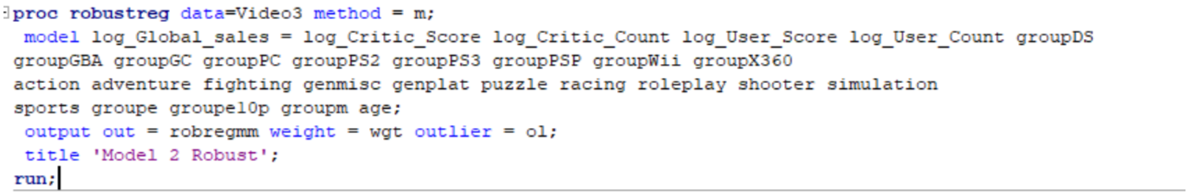
Solution: thus, we solve this problem by just removing the outlier.

Method 2: Robust Regression (MM)

We also use robust regression without manually removing the outlier. Measuring the influence of each observation with different weights.

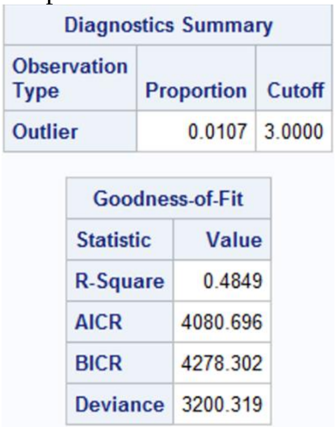
R-squared after implementing the robust regression become 0.4849, which is lower than the original value. In this case, we prefer to use COOK's distance to remove the outliers.

Code:



```
proc robustreg data=Video3 method = m;
 model log_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count groupDS
groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
sports groupe groupe10p groupm age;
 output out = robregmm weight = wgt outlier = ol;
 title 'Model 2 Robust';
run;
```

Output:



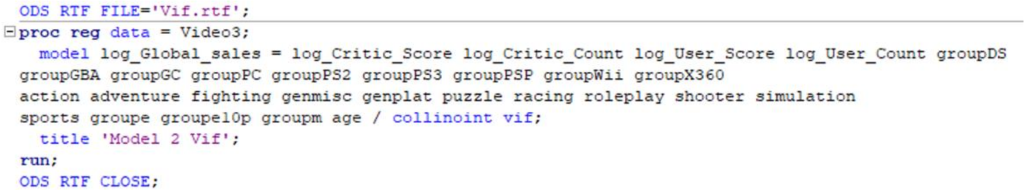
Diagnostics Summary		
Observation Type	Proportion	Cutoff
Outlier	0.0107	3.0000

Goodness-of-Fit	
Statistic	Value
R-Square	0.4849
AICR	4080.696
BICR	4278.302
Deviance	3200.319

Assumption 2 – Multicollinearity

Method: VIF, If VIF>10 for a variable there could be multicollinearity (MC)

Code:



```
ODS RTF FILE='Vif.rtf';
proc reg data = Video3;
   model log_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count groupDS
groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
sports groupe groupe10p groupm age / collinoint vif;
   title 'Model 2 Vif';
run;
ODS RTF CLOSE;
```

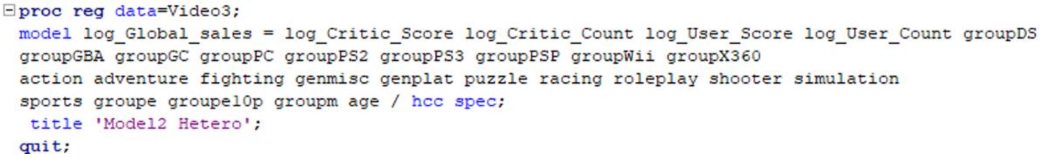
Output:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-7.23259	0.32499	-22.25	<.0001	0
log_Critic_Score	1	0.80822	0.09122	8.86	<.0001	2.23119
log_Critic_Count	1	0.11539	0.02977	3.88	0.0001	2.18896
log_User_Score	1	-0.16769	0.07158	-2.34	0.0192	1.69417
log_User_Count	1	0.62734	0.01625	38.61	<.0001	2.79744
groupDS	1	0.17909	0.07774	2.30	0.0213	2.21122
groupGBA	1	0.22686	0.08680	2.61	0.0090	1.54307
groupGC	1	0.00679	0.07310	0.09	0.9260	1.59421
groupPC	1	-2.45766	0.08360	-29.40	<.0001	2.98572
groupPS2	1	0.43712	0.05629	7.76	<.0001	2.49564
groupPS3	1	0.13260	0.07709	1.72	0.0855	3.16722
groupPSP	1	0.08887	0.07582	1.17	0.2412	1.83315
groupWii	1	0.45682	0.07787	5.87	<.0001	2.36645
groupX360	1	-0.11122	0.07490	-1.48	0.1376	3.38116
action	1	0.51192	0.07922	6.46	<.0001	5.45985
adventure	1	0.15312	0.10540	1.45	0.1464	1.86004
fighting	1	0.51129	0.09713	5.26	<.0001	2.35501
genmisc	1	0.90597	0.09373	9.67	<.0001	2.57805
genplat	1	0.38618	0.09592	4.03	<.0001	2.49919
puzzle	1	0.29197	0.13551	2.15	0.0312	1.48496
racing	1	0.44273	0.08783	5.04	<.0001	3.20332

roleplay	1	0.03784	0.08625	0.44	0.6609	3.14381
shooter	1	0.35592	0.08478	4.20	<.0001	3.96475
simulation	1	0.90742	0.09624	9.43	<.0001	2.10575
sports	1	0.60370	0.08676	6.96	<.0001	4.65137
groupe	1	0.34635	0.04464	7.76	<.0001	2.15993
groupe10p	1	0.18448	0.05036	3.66	0.0003	1.40345
groupm	1	-0.27826	0.04519	-6.16	<.0001	1.59983
age	1	-0.00415	0.00797	-0.52	0.6027	2.94727

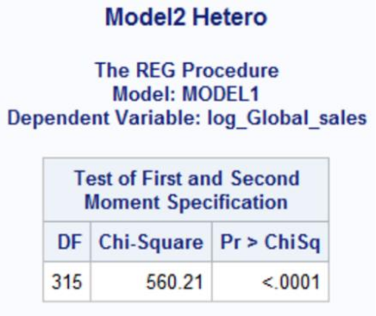
Conclusion: there is no variance inflation factor greater than 10, which means we conclude that there is no collinearity among the independent variables (x) in the model.
Solution: there is no need for a solution.

Assumption 3 – Heteroscedasticity
Use White Test to detect whether has heteroscedasticity problem.
Code:



```
proc reg data=Video3;
model log_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count groupDS
groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
sports groupe groupe10p groupm age / hcc spec;
 title 'Model2 Hetero';
quit;
```

Output:



Model2 Hetero

The REG Procedure
Model: MODEL1
Dependent Variable: log_Global_sales

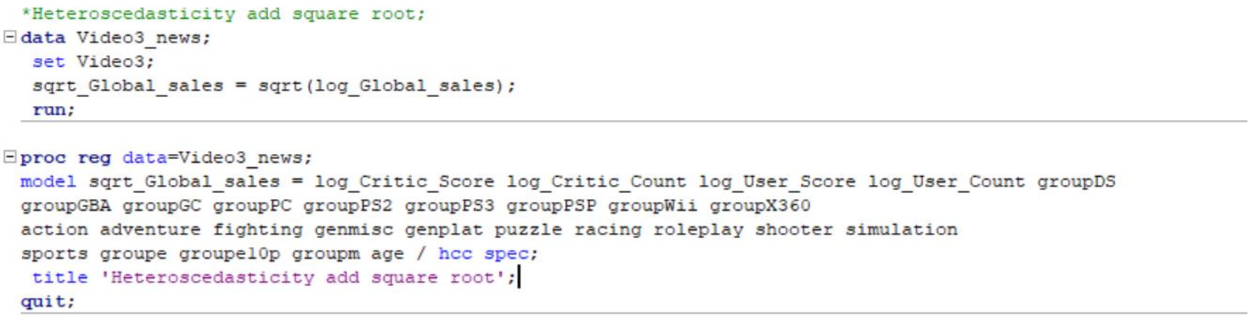
Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
315	560.21	<.0001

Conclusion: According to the white test, the null hypothesis is that there is homoscedasticity; the alternative hypothesis is that there is heteroscedasticity. Based on p < 0.05 in the test of first and second moment specification, we reject H0 and conclude that there is heteroscedasticity in the model.

Method: Square Root Function

We can try to use square root function of log_Y to solve Heteroscedasticity problem. From the results of this model after White Test, we can see that the null hypothesis can now be accepted. But as this new Y brings down our Adjusted R – squared to 0.3736, it is not recommended.

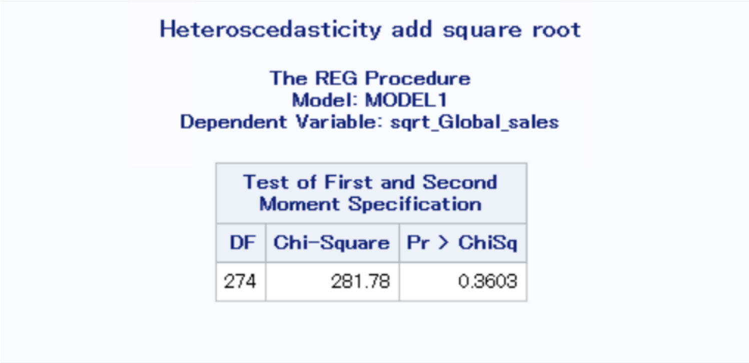
Code:



```
*Heteroscedasticity add square root;
data Video3_news;
 set Video3;
 sqrt_Global_sales = sqrt(log_Global_sales);
 run;

proc reg data=Video3_news;
model sqrt_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count groupDS
groupGBA groupGC groupPC groupPS2 groupPS3 groupPSP groupWii groupX360
action adventure fighting genmisc genplat puzzle racing roleplay shooter simulation
sports groupe groupe10p groupm age / hcc spec;
 title 'Heteroscedasticity add square root';
quit;
```

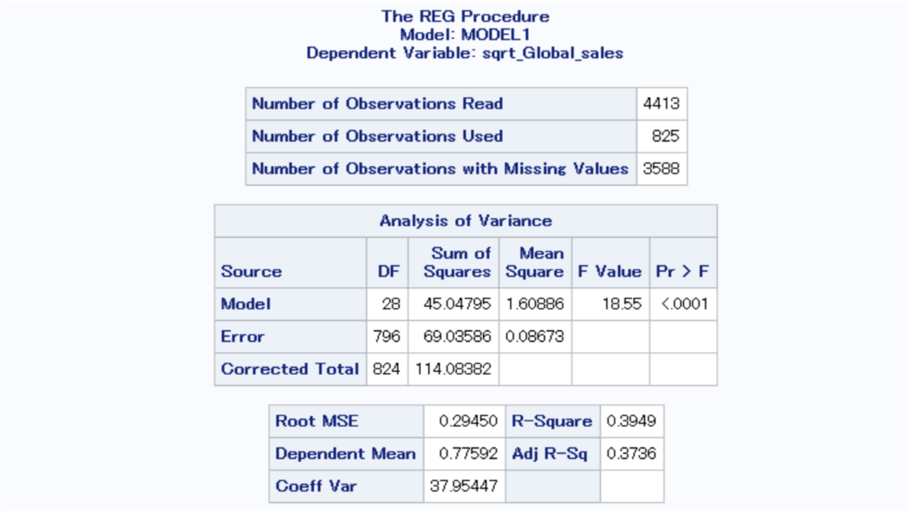
Output:



Heteroscedasticity add square root

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_Global_sales

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
274	281.78	0.3603



The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_Global_sales

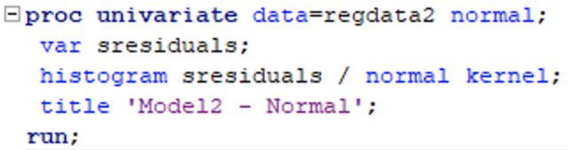
Number of Observations Read	4413
Number of Observations Used	825
Number of Observations with Missing Values	3588

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	28	45.04795	1.60886	18.55	<.0001
Error	796	69.03586	0.08673		
Corrected Total	824	114.08382			

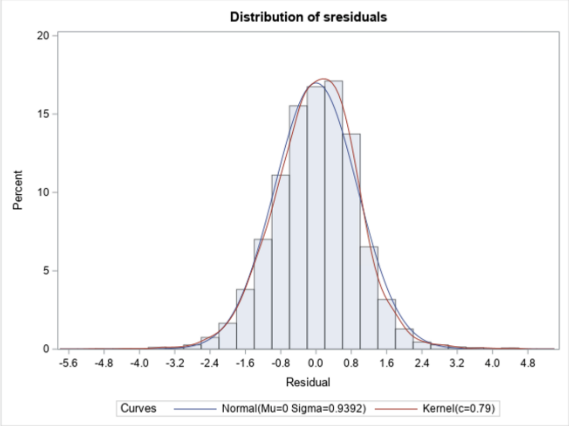
Root MSE	0.29450	R-Square	0.3949
Dependent Mean	0.77592	Adj R-Sq	0.3736
Coeff Var	37.95447		

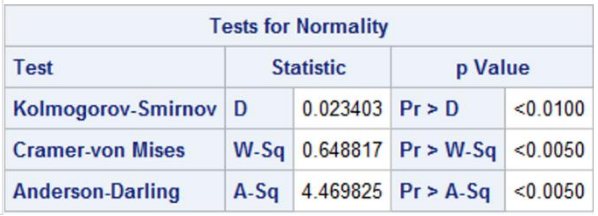
Assumption 4 - Normality of error term
Before removing outlier
Code:



```
proc univariate data=regdata2 normal;
  var sresiduals;
  histogram sresiduals / normal kernel;
  title 'Model2 - Normal';
 run;
```

Output:



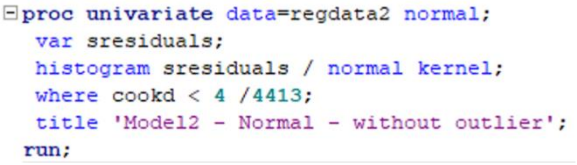


Tests for Normality

Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.023403	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.648817	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.469825	Pr > A-Sq	<0.0050

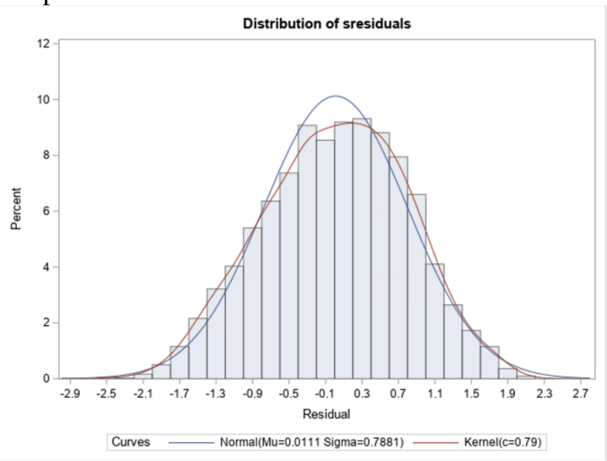
After removing outlier

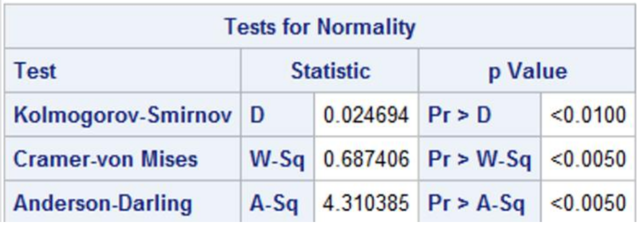
Code:



```
proc univariate data=regdata2 normal;
  var sresiduals;
  histogram sresiduals / normal kernel;
  where cookd < 4 /4413;
  title 'Model2 - Normal - without outlier';
run;
```

Output:





Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.024694	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.687406	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	4.310385	Pr > A-Sq	<0.0050

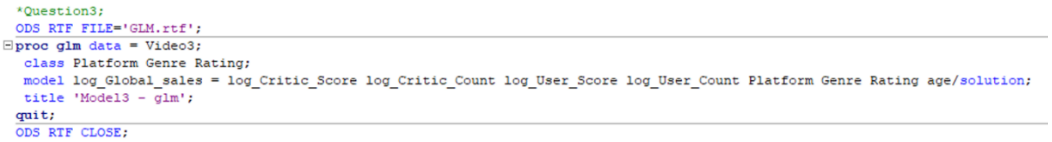
Method: Histogram Plot & Test for Normality (Kolmogorov-Smirnov Test)

Conclusion: It is perfectly fitted with a normal distribution. But the test for normality shows slightly different. After removing the outlier, the residual distribution is still not significantly normal distributed.
Solution: The distribution is closed to a normal distribution with or without removing the outlier. there is no need for a solution.

Question 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.

Code:



```
*Question3;
ODS RTF FILE='GLM.rtf';
proc glm data = Video3;
 class Platform Genre Rating;
 model log_Global_sales = log_Critic_Score log_Critic_Count log_User_Score log_User_Count Platform Genre Rating age/solution;
 title 'Model3 - glm';
quit;
ODS RTF CLOSE;
```

Output:
Like Question 1, we create three tables.

Class Level Information		
Class	**Levels**	**Values**
Platform	10	DS GBA GC PC PS2 PS3 PSP Wii X360 XB
Genre	12	Action Adventure Fighting Misc Platform Puzzle Racing Role-Playing Shooter Simulation Sports Strategy
Rating	4	E E10+ M T

Number of Observations Read	4413
Number of Observations Used	4413

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	28	4583.879127	163.709969	184.42	<.0001
Error	4384	3891.782703	0.887724		
Corrected Total	4412	8475.661830			

R-Square	Coeff Var	Root MSE	log_Global_sales Mean
0.540828	-75.99241	0.942191	-1.239849

Source	DF	Type I SS	Mean Square	F Value	Pr > F
log_Critic_Score	1	888.531881	888.531881	1000.91	<.0001
log_Critic_Count	1	629.954102	629.954102	709.63	<.0001
log_User_Score	1	15.426965	15.426965	17.38	<.0001
log_User_Count	1	279.549399	279.549399	314.91	<.0001
Platform	9	2329.174107	258.797123	291.53	<.0001
Genre	11	334.044556	30.367687	34.21	<.0001
Rating	3	106.957594	35.652531	40.16	<.0001
age	1	0.240523	0.240523	0.27	0.6027

Source	DF	Type III SS	Mean Square	F Value	Pr > F
log_Critic_Score	1	69.689433	69.689433	78.50	<.0001
log_Critic_Count	1	13.339357	13.339357	15.03	0.0001
log_User_Score	1	4.872653	4.872653	5.49	0.0192
log_User_Count	1	1323.222754	1323.222754	1490.58	<.0001
Platform	9	1902.800388	211.422265	238.16	<.0001
Genre	11	216.703996	19.700363	22.19	<.0001
Rating	3	107.197987	35.732662	40.25	<.0001
age	1	0.240523	0.240523	0.27	0.6027

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-7.232592638	B	0.32498743	-22.25	<.0001
log_Critic_Score	0.808217770		0.09121871	8.86	<.0001
log_Critic_Count	0.115389567		0.02976721	3.88	0.0001
log_User_Score	-0.167693016		0.07157663	-2.34	0.0192
log_User_Count	0.627344644		0.01624908	38.61	<.0001
Platform DS	0.179086494	B	0.07774489	2.30	0.0213
Platform GBA	0.226858708	B	0.08679848	2.61	0.0090
Platform GC	0.006790872	B	0.07309884	0.09	0.9260
Platform PC	-2.457661664	B	0.08360259	-29.40	<.0001
Platform PS2	0.437121189	B	0.05629419	7.76	<.0001
Platform PS3	0.132602797	B	0.07708519	1.72	0.0855
Platform PSP	0.088871787	B	0.07582290	1.17	0.2412
Platform Wii	0.456816390	B	0.07786952	5.87	<.0001
Platform X360	-0.111222777	B	0.07489885	-1.48	0.1376
Platform XB	0.000000000	B	.	.	.
Genre Action	0.511918278	B	0.07922014	6.46	<.0001
Genre Adventure	0.153123011	B	0.10540302	1.45	0.1464
Genre Fighting	0.511294136	B	0.09713099	5.26	<.0001
Genre Misc	0.905972573	B	0.09373280	9.67	<.0001
Genre Platform	0.386177559	B	0.09591692	4.03	<.0001
Genre Puzzle	0.291974084	B	0.13550554	2.15	0.0312
Genre Racing	0.442728829	B	0.08782803	5.04	<.0001
Genre RolePlaying	0.037835867	B	0.08624990	0.44	0.6609
Genre Shooter	0.355919713	B	0.08477922	4.20	<.0001
Genre Simulation	0.907422326	B	0.09624215	9.43	<.0001

Genre Sports	0.603700650	B	0.08675622	6.96	<.0001
Genre Strategy	0.000000000	B	.	.	.
Rating E	0.346354873	B	0.04463637	7.76	<.0001
Rating E10+	0.184481997	B	0.05036400	3.66	0.0003
Rating M	-0.278262267	B	0.04519006	-6.16	<.0001
Rating T	0.000000000	B	.	.	.
age	-0.004147019		0.00796704	-0.52	0.6027
Note:			The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.		

Conclusion:

Comparing the results of this model to the model in question 1(f), we can see that the log-log model has been replicated exactly using GLM.