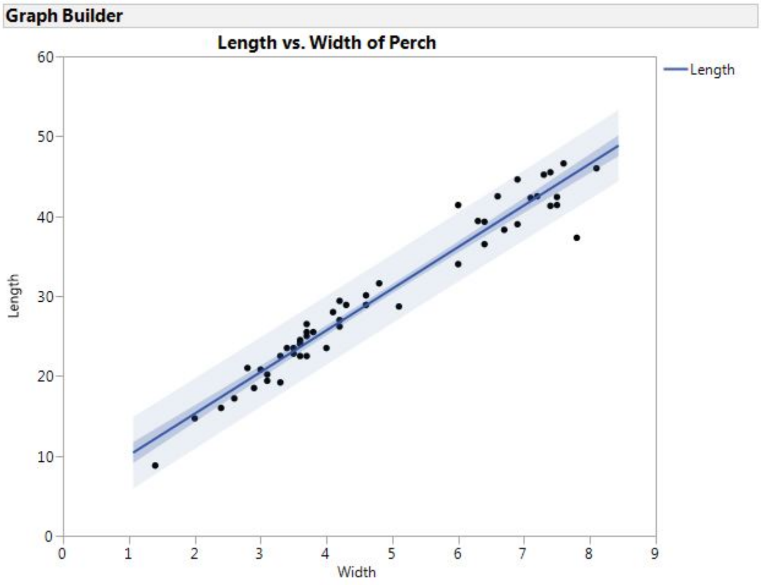
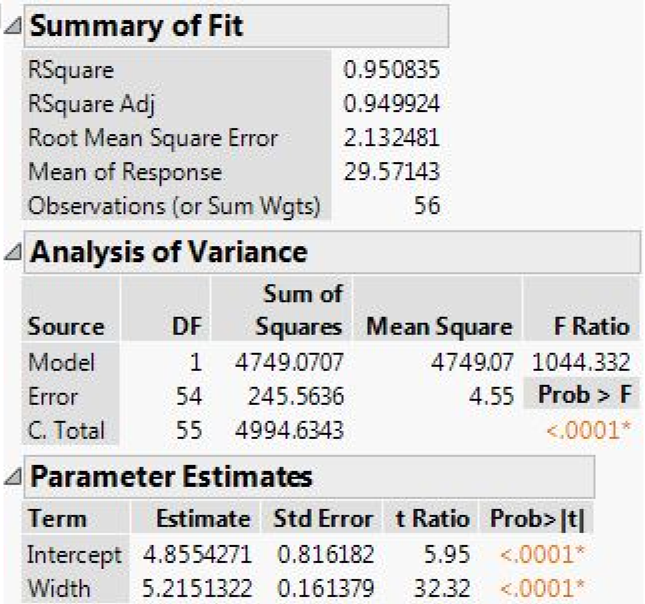
Question 1

Suppose you plan to enter a fishing contest to try to catch the perch (type of fish) that is the longest. You then become curious if the width of a fish is a good predictor of the length of the fish. You find data on a sample of 56 perch that contains the length of the perch (in centimeters) and the width of the perch (in centimeters).

Below is a scatterplot of the data with prediction and confidence bands, as well as the least squares regression analysis.





Summary of Fit

RSquare	0.950835
RSquare Adj	0.949924
Root Mean Square Error	2.132481
Mean of Response	29.57143
Observations (or Sum Wgts)	56

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	4749.0707	4749.07	1044.332
Error	54	245.5636	4.55	Prob > F
C. Total	55	4994.6343		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.8554271	0.816182	5.95	<.0001*
Width	5.2151322	0.161379	32.32	<.0001*

Also make note of the following results:

- The prediction interval for a perch with a width of 6.7 cm is 35.44 to 44.16.
- The confidence interval to estimate the population mean for a perch with a width of 6.7 cm is 39.80 to 40.65.

You may assume that the conditions for conducting inference with simple linear regression were reasonably met.

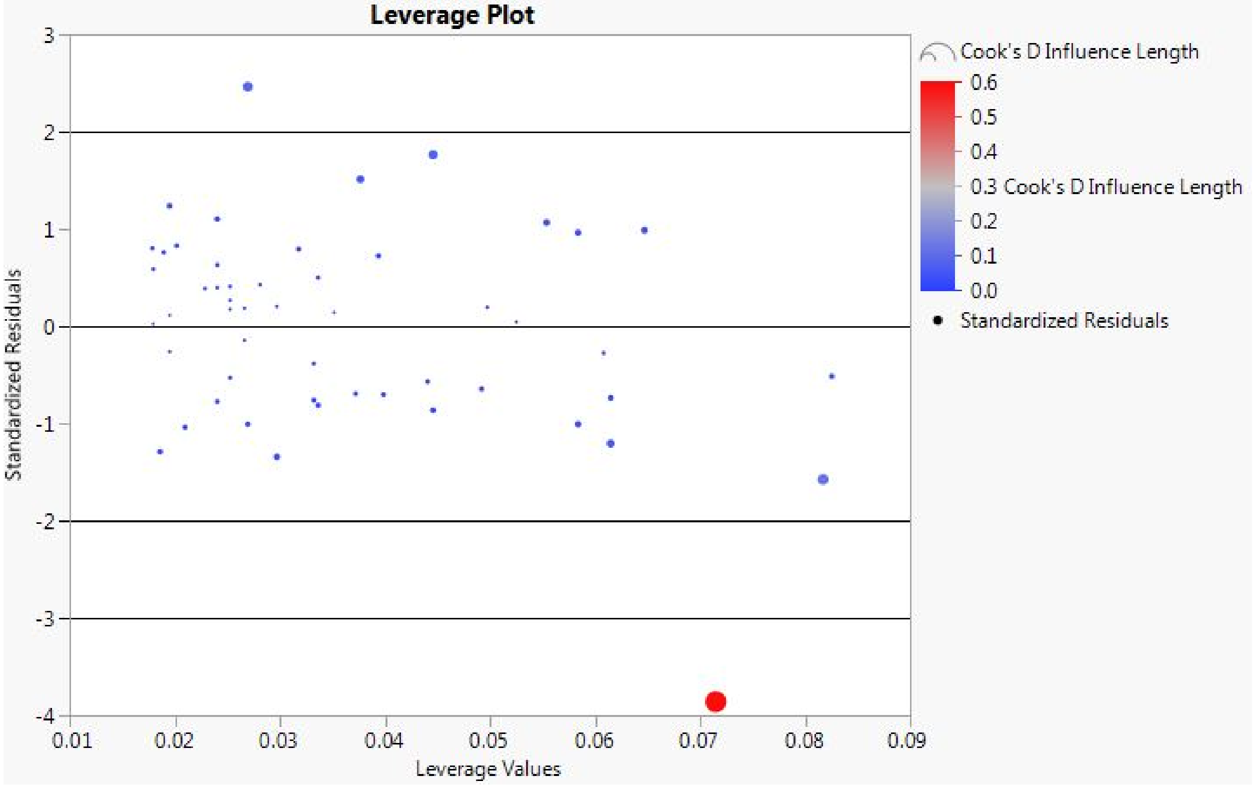
a. Provide an interpretation of the confidence interval for the population mean for a perch with a width of 6.7 cm.

b. Looking at the scatterplot, which bands represent the prediction intervals and which bands represent the confidence intervals? Explain your reasoning.

c. Is there statistical evidence to conclude that perch with higher widths tend to have higher lengths? Report the appropriate analysis to answer this question. Write up your results as a short report to the *Big Fish Magazine* company (1-2 paragraphs). Make sure that it is written in a way that the editors for *Big Fish Magazine* would understand the numbers. Note that you will lose points if your answer is not written in the form of a report. Be sure to include the following in your report:

- What statistical method you used to answer the research question and why;
- the relevant parameter;
- hypotheses (if applicable);
- the results from your statistical method; and
- your interpretation of the results.

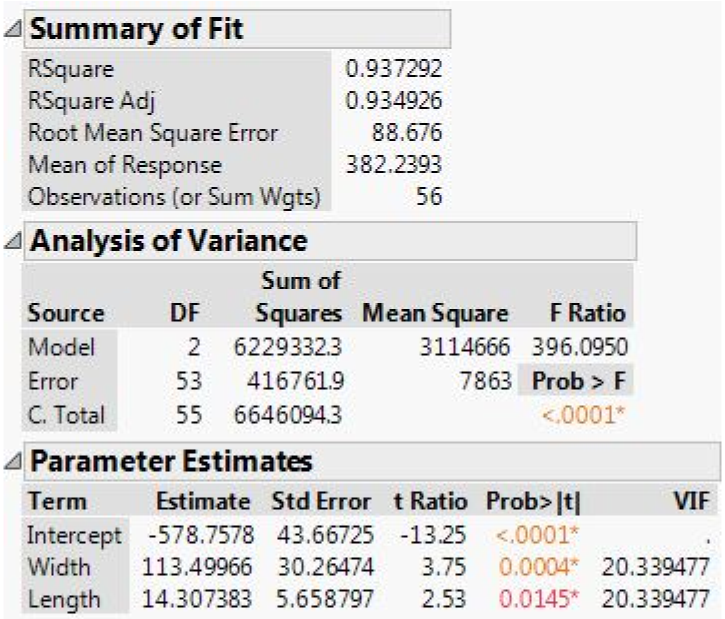
Below is a leverage plot.



d. Do you have any concerns about outliers? Explain.

Question 2

Suppose you would now like to predict the weight (in grams) of perch using length and width from the same dataset. Below is output from the multiple regression analysis.



Summary of Fit

RSquare	0.937292
RSquare Adj	0.934926
Root Mean Square Error	88.676
Mean of Response	382.2393
Observations (or Sum Wgts)	56

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	6229332.3	3114666	396.0950
Error	53	416761.9	7863	Prob > F
C. Total	55	6646094.3		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-578.7578	43.66725	-13.25	<.0001*	.
Width	113.49966	30.26474	3.75	0.0004*	20.339477
Length	14.307383	5.658797	2.53	0.0145*	20.339477

a. What is the multiple least squares regression equation?

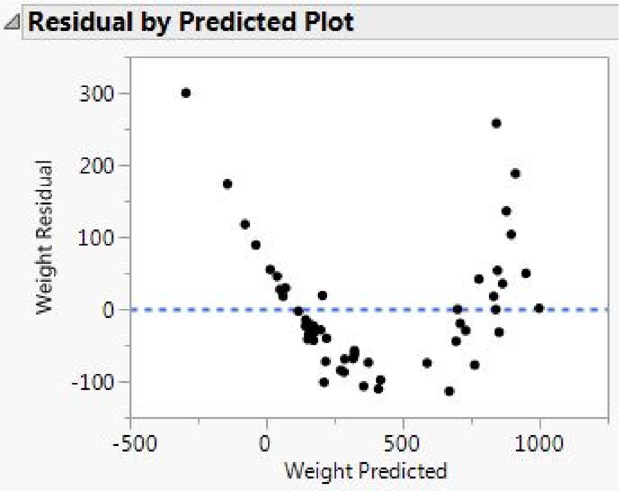
b. Is it appropriate to interpret the value of the Y-intercept in this example? If no, explain why. If yes, interpret the value of the Y-intercept.

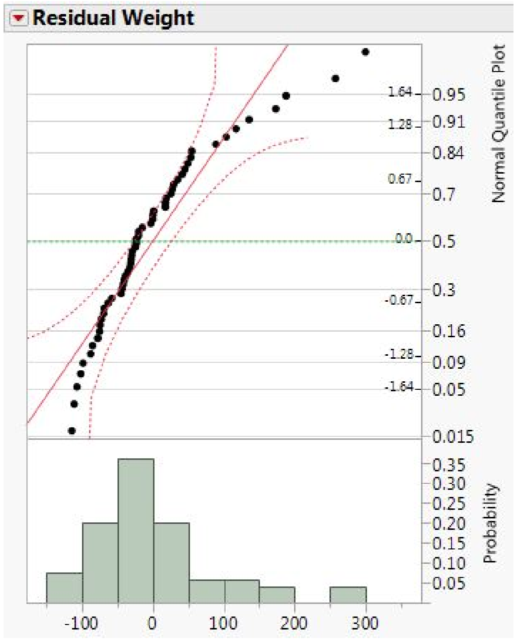
c. Interpret the value of the slope for length.

d. Interpret the value of adjusted R2.

Question 3

Continue with the previous question. Below is output from JMP that can be used to check the necessary assumptions.



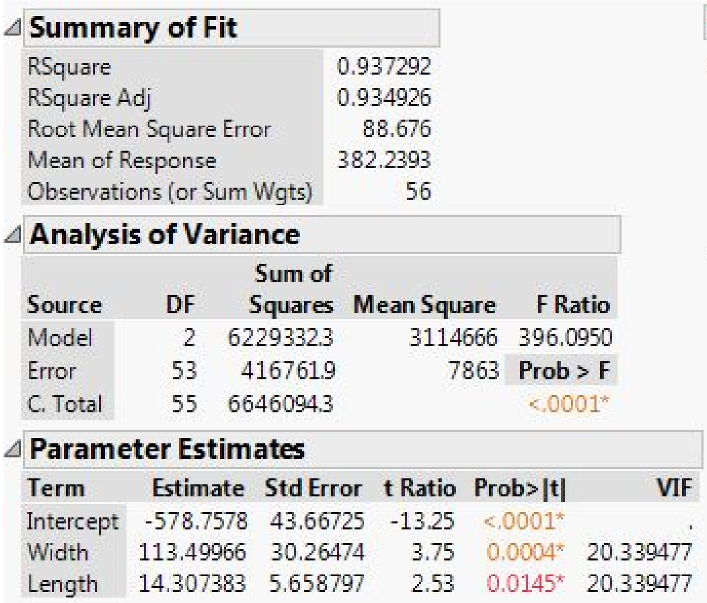


e. Do the multiple regression assumptions appear to be met?

f. Based on your answer to the previous question, was it was appropriate to interpret the Y-intercept, slopes, and adjusted R2? Explain.

Question 4

Continue with the previous question. Below is the regression output and a scatterplot matrix and correlation matrix from JMP.



Summary of Fit

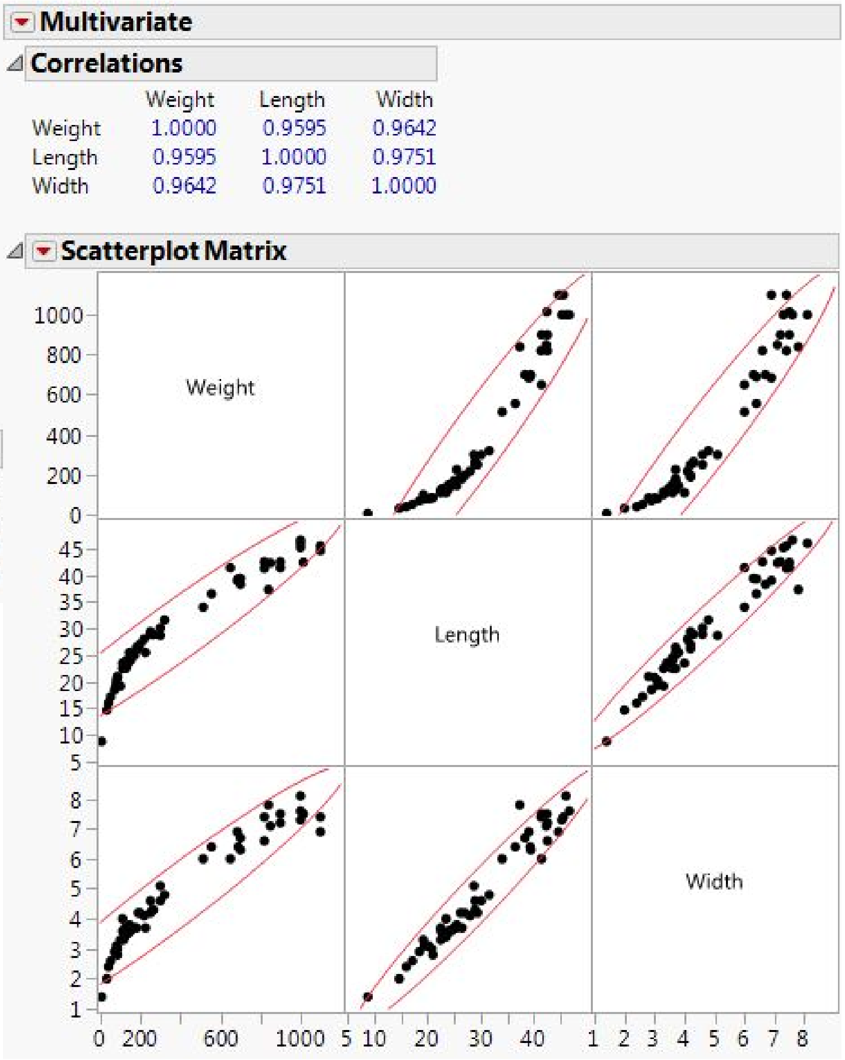
RSquare	0.937292
RSquare Adj	0.934926
Root Mean Square Error	88.676
Mean of Response	382.2393
Observations (or Sum Wgts)	56

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	6229332.3	3114666	396.0950
Error	53	416761.9	7863	Prob > F
C. Total	55	6646094.3		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-578.7578	43.66725	-13.25	<.0001*	.
Width	113.49966	30.26474	3.75	0.0004*	20.339477
Length	14.307383	5.658797	2.53	0.0145*	20.339477



Multivariate

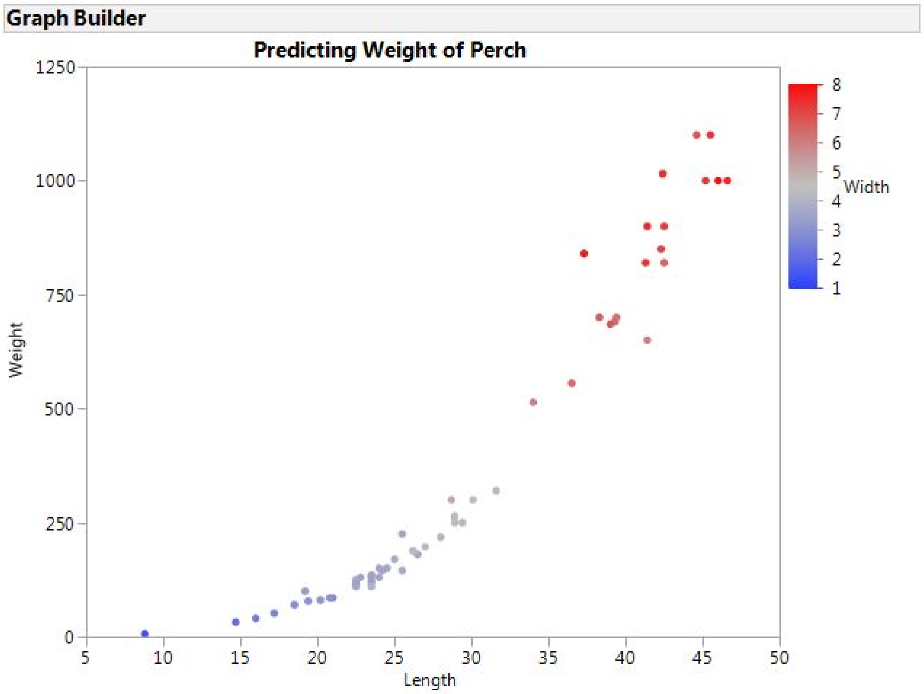
Correlations

	Weight	Length	Width
Weight	1.0000	0.9595	0.9642
Length	0.9595	1.0000	0.9751
Width	0.9642	0.9751	1.0000

Scatterplot Matrix

g. Interpret the correlation and scatterplot matrices.

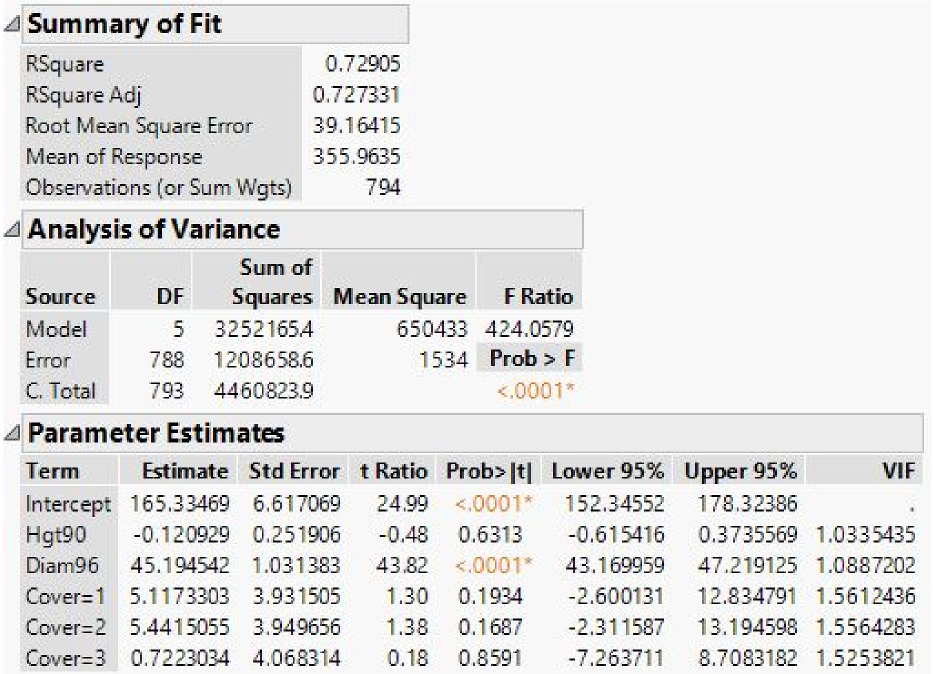
h. Based on previous output, do you believe there would be any issues with multicollinearity?



i. Interpret the plot shown above by comparing the relationships among each pair of variables.

Question 5

Data was collected for 1000 white pine trees. Pine trees were planted as seedlings in 1990 at the Brown Family Environmental Center. The planters did their best to control for confounding variables by planting them the same distance apart and in similar soil. We would like to predict the height of the pine tree in 1997 (cm) using their height at time of planting in 1990 (cm), the diameter of the trunk in 1996 (cm), and amount of thorny cover in 1995. Amount of cover was coded as 0 for no cover, 1 for some cover, 2 for moderate cover, and 3 for lots of cover.



⊿ Summary of Fit

RSquare	0.72905
RSquare Adj	0.727331
Root Mean Square Error	39.16415
Mean of Response	355.9635
Observations (or Sum Wgts)	794

⊿ Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	3252165.4	650433	424.0579
Error	788	1208658.6	1534	Prob > F
C. Total	793	4460823.9		<.0001*

⊿ Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	165.33469	6.617069	24.99	<.0001*	152.34552	178.32386	.
Hgt90	-0.120929	0.251906	-0.48	0.6313	-0.615416	0.3735569	1.0335435
Diam96	45.194542	1.031383	43.82	<.0001*	43.169959	47.219125	1.0887202
Cover=1	5.1173303	3.931505	1.30	0.1934	-2.600131	12.834791	1.5612436
Cover=2	5.4415055	3.949656	1.38	0.1687	-2.311587	13.194598	1.5564283
Cover=3	0.7223034	4.068314	0.18	0.8591	-7.263711	8.7083182	1.5253821

a. Is it appropriate to interpret the value of the Y-intercept in this example? If no, explain why. If yes, interpret the value of the Y-intercept.

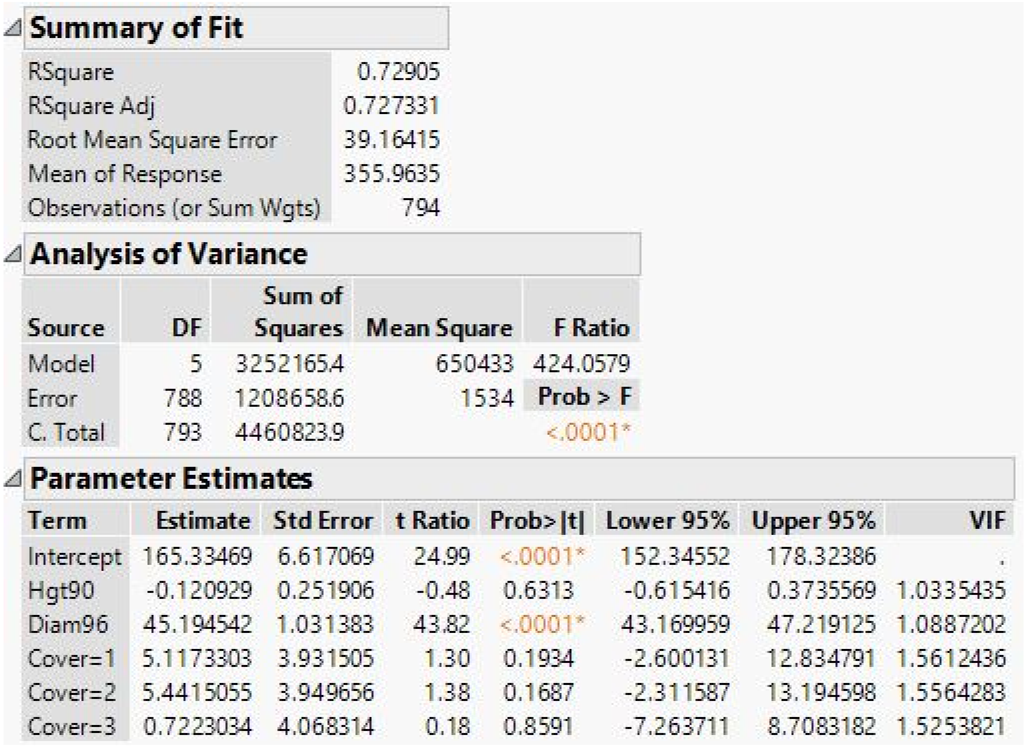
b. Interpret the slopes for the amount of thorny cover in 1995.

c. Interpret the confidence interval for the slope of diameter of the tree in 1996.

d. Predict the height of a tree in 1997 for a tree that was 14cm tall at time of planting in 1990, had a 5cm diameter in 1996, and had a moderate amount of thorny cover in 1995.

Question 6

Continue with the previous question. Output from JMP is provided below.



Summary of Fit

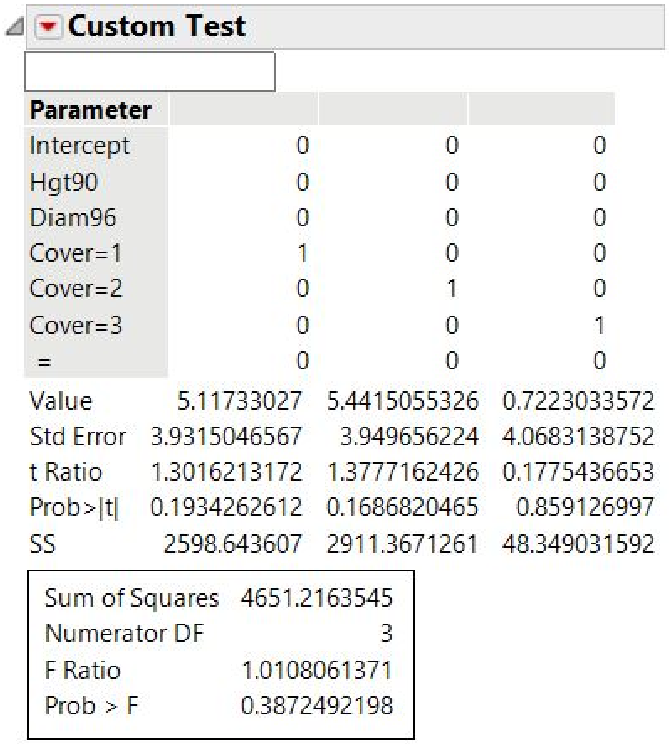
RSquare	0.72905
RSquare Adj	0.727331
Root Mean Square Error	39.16415
Mean of Response	355.9635
Observations (or Sum Wgts)	794

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	3252165.4	650433	424.0579
Error	788	1208658.6	1534	Prob > F
C. Total	793	4460823.9		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	165.33469	6.617069	24.99	<.0001*	152.34552	178.32386	.
Hgt90	-0.120929	0.251906	-0.48	0.6313	-0.615416	0.3735569	1.0335435
Diam96	45.194542	1.031383	43.82	<.0001*	43.169959	47.219125	1.0887202
Cover=1	5.1173303	3.931505	1.30	0.1934	-2.600131	12.834791	1.5612436
Cover=2	5.4415055	3.949656	1.38	0.1687	-2.311587	13.194598	1.5564283
Cover=3	0.7223034	4.068314	0.18	0.8591	-7.263711	8.7083182	1.5253821



Custom Test

Parameter			
Intercept	0	0	0
Hgt90	0	0	0
Diam96	0	0	0
Cover=1	1	0	0
Cover=2	0	1	0
Cover=3	0	0	1
=	0	0	0

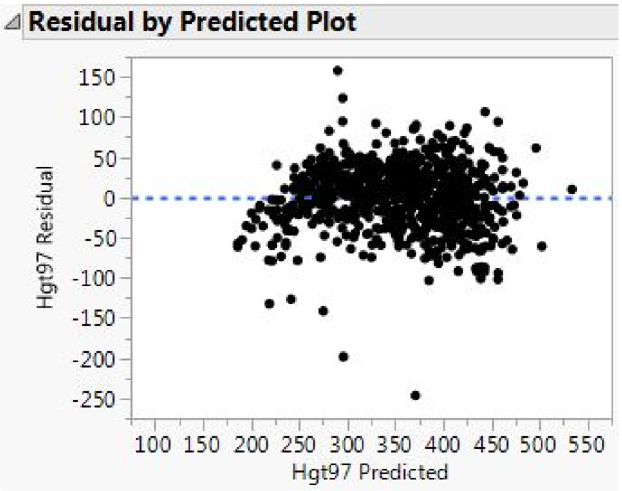
Value	5.11733027	5.4415055326	0.7223033572
Std Error	3.9315046567	3.949656224	4.0683138752
t Ratio	1.3016213172	1.3777162426	0.1775436653
Prob>|t|	0.1934262612	0.1686820465	0.859126997
SS	2598.643607	2911.3671261	48.349031592

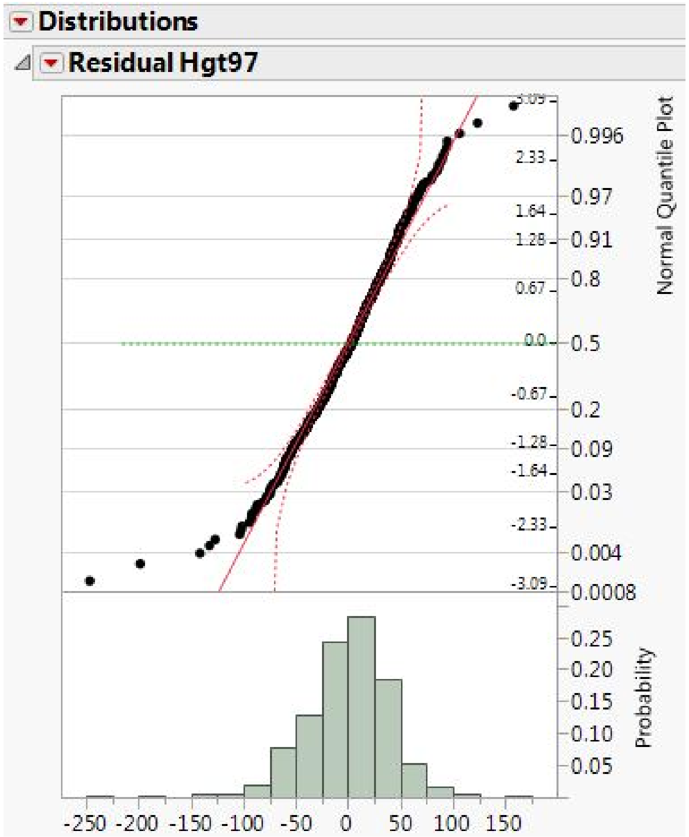
Sum of Squares	4651.2163545
Numerator DF	3
F Ratio	1.0108061371
Prob > F	0.3872492198

e. Is there evidence that the amount of cover is a predictor of the height in 1997 when diameter of the trunk in 1996 and height of the tree when planted in 1990 are included in the model? Explain by referring to an appropriate p-value. You do not have to show all steps of the hypothesis test.

f. Is there evidence that the diameter of the trunk in 1996 is a predictor of the height in 1997 when height of the tree when planted in 1990 and the amount of thorny cover in 1995 are included in the model? Explain by referring to an appropriate p-value. You do not have to show all steps of the hypothesis test.

g. Conduct a hypothesis testing for the overall model. Show all steps of a hypothesis test.





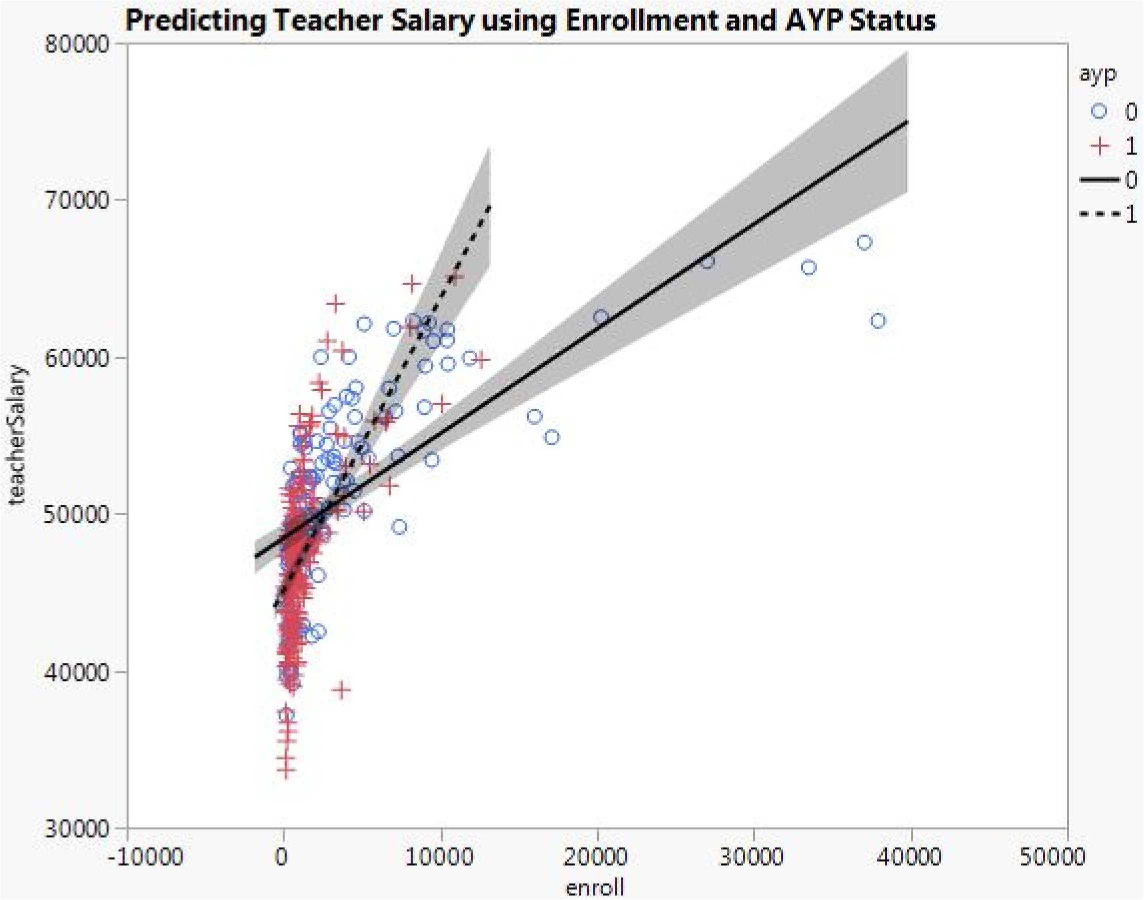
h. Check the assumptions for regression and comment on each. Use the graphs included above.

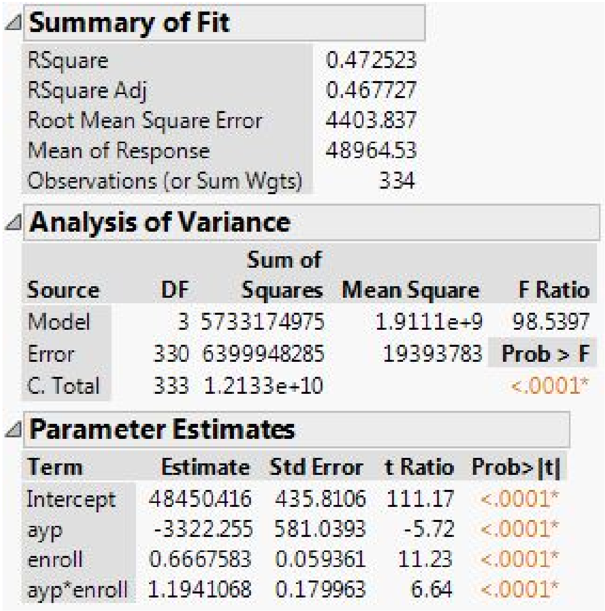
Question 7

"Adequate yearly progress (AYP) is the measure by which schools, districts, and states are held accountable for student performance under Title I of the No Child Left Be- hind Act of 2001 (NCLB) Under NCLB, states are required to show that public school students are making yearly progress toward meeting state academic content standards. The goal is to have all students reaching proficient levels in reading and math by 2014 as measured by performance on state tests."1, 2

Data was collected from 344 school districts in the 2011-2012 school year in Minnesota. For each district, it was recorded if the district met the AYP standards (0=no, 1=yes), the average teacher salary for the district, estimated proportion of school ages children in poverty (0=below 0.10, 1=between 0.10 and 0.15, 2=above 0.15), and the number of students enrolled in the district.

We would like to predict the average teacher salary in the district using whether or not the district met the AYP standards, estimated proportion of school ages children in poverty, and the number of students enrolled in the district.





Summary of Fit

RSquare	0.472523
RSquare Adj	0.467727
Root Mean Square Error	4403.837
Mean of Response	48964.53
Observations (or Sum Wgts)	334

Analysis of Variance

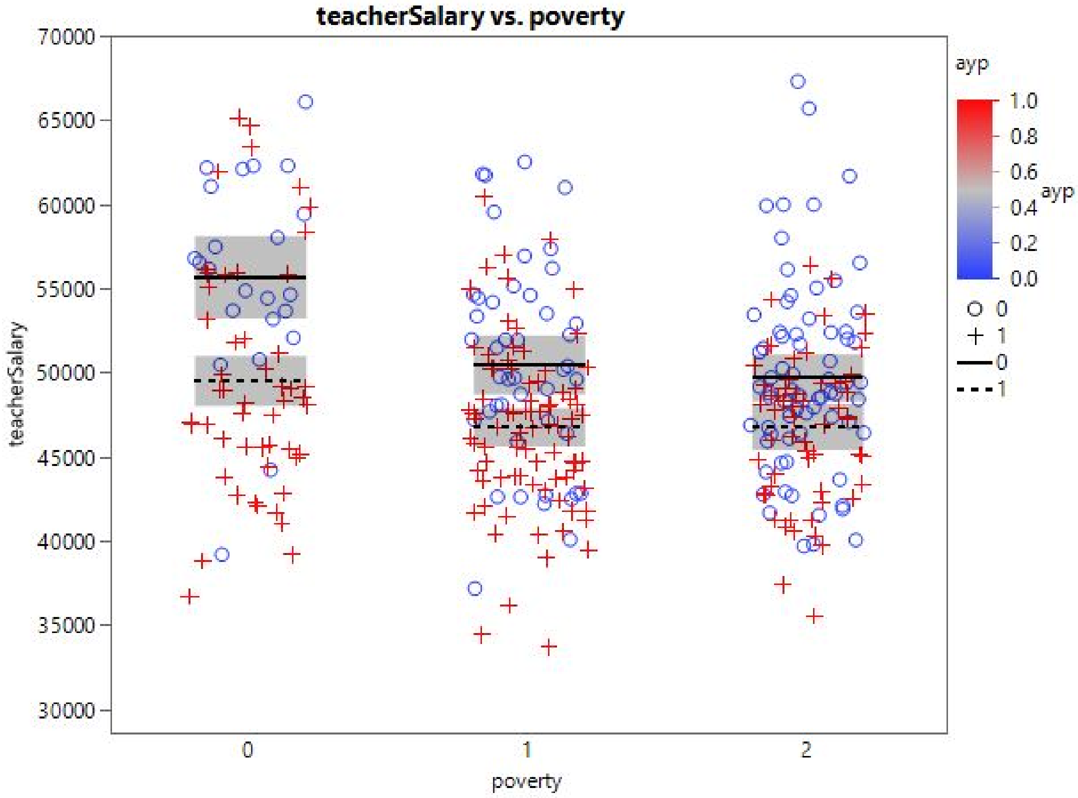
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	5733174975	1.9111e+9	98.5397
Error	330	6399948285	19393783	Prob > F
C. Total	333	1.2133e+10		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	48450.416	435.8106	111.17	<.0001*
ayp	-3322.255	581.0393	-5.72	<.0001*
enroll	0.6667583	0.059361	11.23	<.0001*
ayp*enroll	1.1941068	0.179963	6.64	<.0001*

a. Based on the graph and regression output above, does it appear that there is an interaction between ayp and enrollment? Explain.

b. What does it mean to have an interaction between ayp and enrollment?

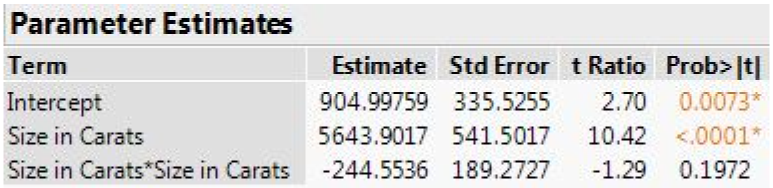


c. Based on the graph above, does it appear that there is an interaction between ayp and poverty? Explain.

Question 8

Suppose you are planning to purchase a diamond and are curious how much the size of a diamond (in carats) affects the price of a diamond. A sample of 351 diamonds was taken and their price (in dollars) was recorded. The sample was not selected randomly.

There is a non-linear relationship between the size of a diamond and price of the diamond so a quadratic model was fit to the data. Below is the regression output.

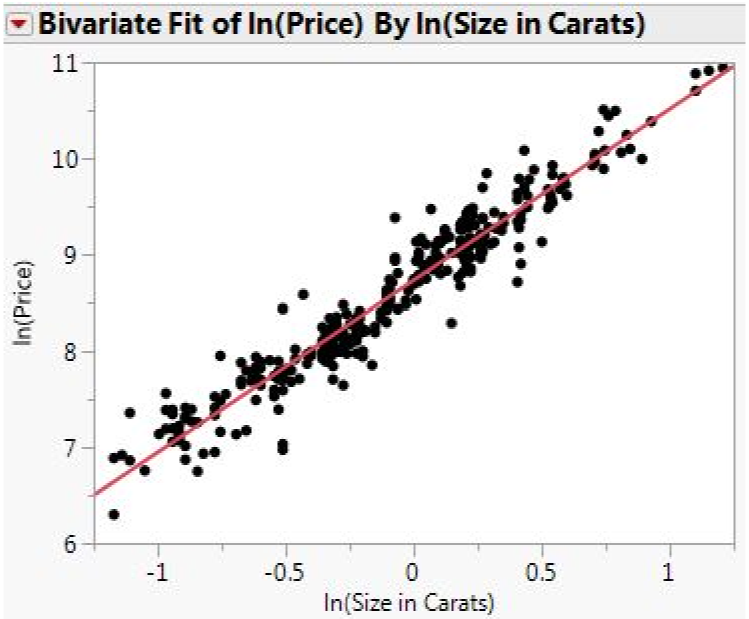


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	904.99759	335.5255	2.70	0.0073*
Size in Carats	5643.9017	541.5017	10.42	<.0001*
Size in Carats*Size in Carats	-244.5536	189.2727	-1.29	0.1972

a. Is there evidence of a quadratic relationship between the size of a diamond (in carats) and price of the diamond? Write out all of the steps of the hypothesis test.

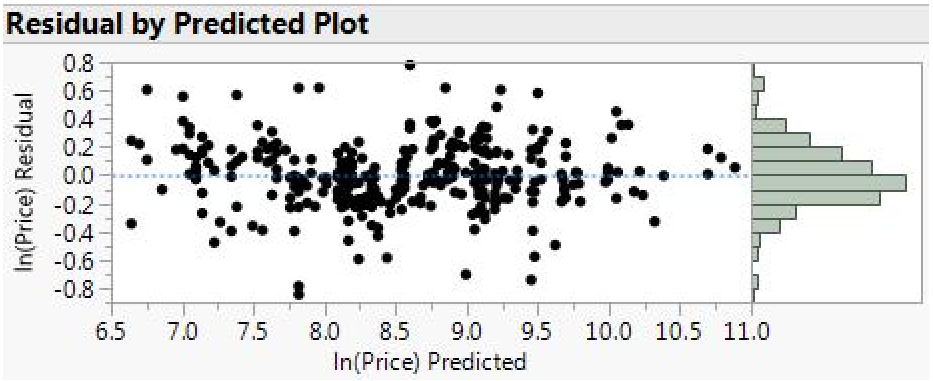
Suppose it was decided that the quadratic model was not a good fit and that a natural log transformation on both variables would be better. Below is a scatterplot of the transformed data.

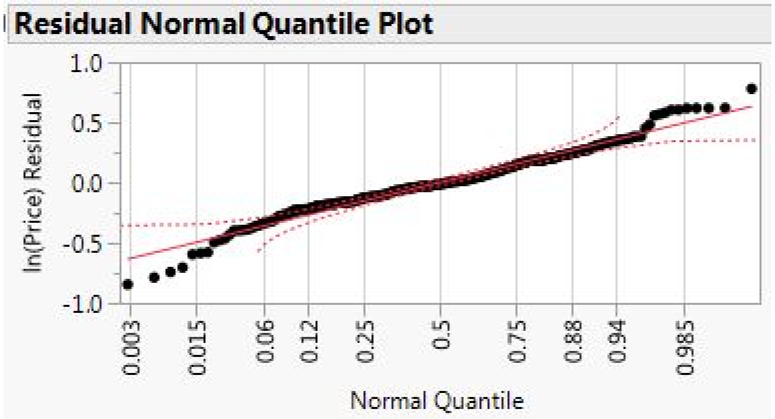


Bivariate Fit of ln(Price) By ln(Size in Carats)

b. Describe the relationship between the natural log of price of diamonds and the natural log of size of diamonds (in carats).

Below is a residual plot and normal quantile plot of the transformed data.

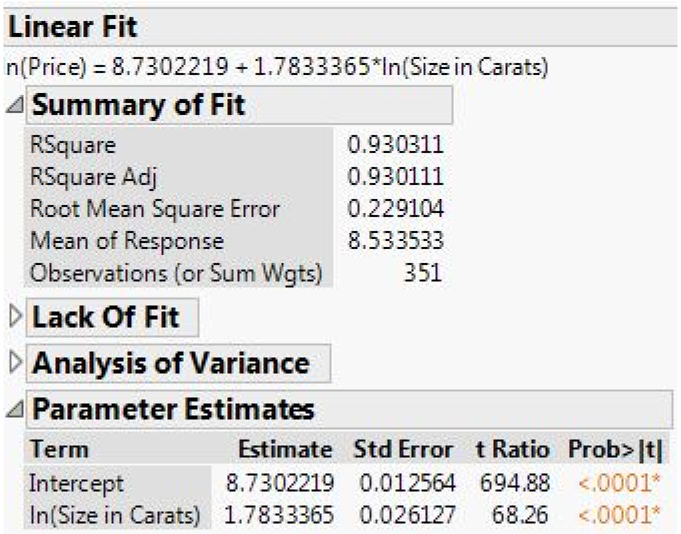




c. Check the assumptions for simple linear regression and comment on each.

Question 9

Continue with the previous question. Below is the simple linear regression output for the transformed data.



```
Linear Fit
n(Price) = 8.7302219 + 1.7833365*ln(Size in Carats)
Summary of Fit
  RSquare                        0.930311
  RSquare Adj                    0.930111
  Root Mean Square Error         0.229104
  Mean of Response               8.533533
  Observations (or Sum Wgts)          351
Lack Of Fit
Analysis of Variance
Parameter Estimates
  Term               Estimate  Std Error  t Ratio  Prob>|t|
  Intercept         8.7302219   0.012564   694.88   <.0001*
  ln(Size in Carats) 1.7833365  0.026127    68.26   <.0001*
```

d. What is the estimated least squares equation in this case?

e. What is the predicted median price for a diamond that is 1.5 carats?

f. Provide a reasonable interpretation of the estimated slope (or function of the estimated slope) in the context of this example.