M349R (Unique 54230)

**Instructor:** Gustavo Cepparo

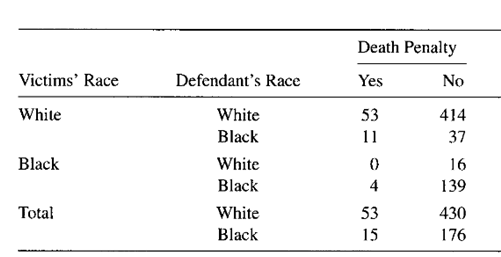
Project 3

Juan Acosta

JA45384

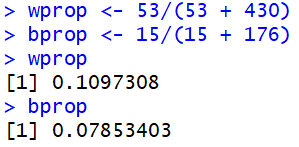
Fall 2018

Problem 1



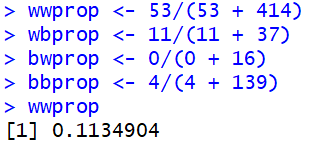
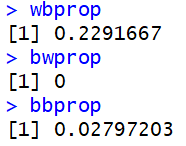


[a] Compute the proportion of Black and the proportion of White defendants that received Death Penalty.



[b] Now proceed to control for Victim’s Race. So calculate the proportion of Black defendants that received death penalty given that the victim was white. Also do the same for a black victim.

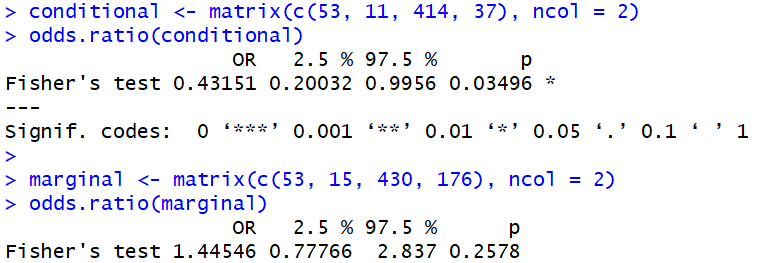
Defining xyprop as the proportion of the race of the victim x and defendant y:

[c] Explain Simpson’s Paradox in the context of this problem.

Even though the proportion of whites (11%) that get the death penalty is greater than that of blacks (7.9%). Controlling for the victims race we can see that among white victims, black defendants received the death penalty more often. Similarly, among black victims, black defendants received the death penalty more often than white ones. This is due to the amount of murders analyzed. Therefore, the marginal association is due to a tendency of whites to kill whites, therefore, the sheer amount of murders of whites on whites dilutes the proportions.

[d] Calculate odds ratios for the partial tables (conditional) and for the marginal table. Also interpret the odd ratios.



Looking at the conditional odds ratio value of 0.43 and the CI, we can see that we can reject the null hypothesis in favor of the alternative. Therefore, we can conclude there is conditional dependence.

Looking at the marginal odds ratio value of 1.44 and the CI we can see that we cannot reject the null hypothesis of the odds ratio (aka OR=1) . Therefore, we conclude that there is marginal independence.

Therefore, given that the conditional and marginals have opposite behaviors we can conclude that this is a case of Simpson’s Paradox.

Problem 2



(7 pts each)

4) The dataset “Cocaine” contains 56 observations on the variables related

to sales of cocaine powder in northeastern California over the

period 1984-1991. The data are a subset of those

used in the study Caulkins, J. P. and R. Padman (1993),

"Quantity Discount and Quality Premia for Illicit Drugs,"

Journal of the American Statistical Association, 88, 748-757.

The variables are

price = price per gram in dollars for a cocaine sale,

quant = number of grams of cocaine in a given sale,

qual = quality of the cocaine expressed as percentage purity,

trend = a time variable with 1984 = 1 up to 1991 = 8.

Consider the regression model

price = beta1 + beta2 quant + beta3 qual + beta4 trend + e

a) What signs (positive or negative) would you expect for the

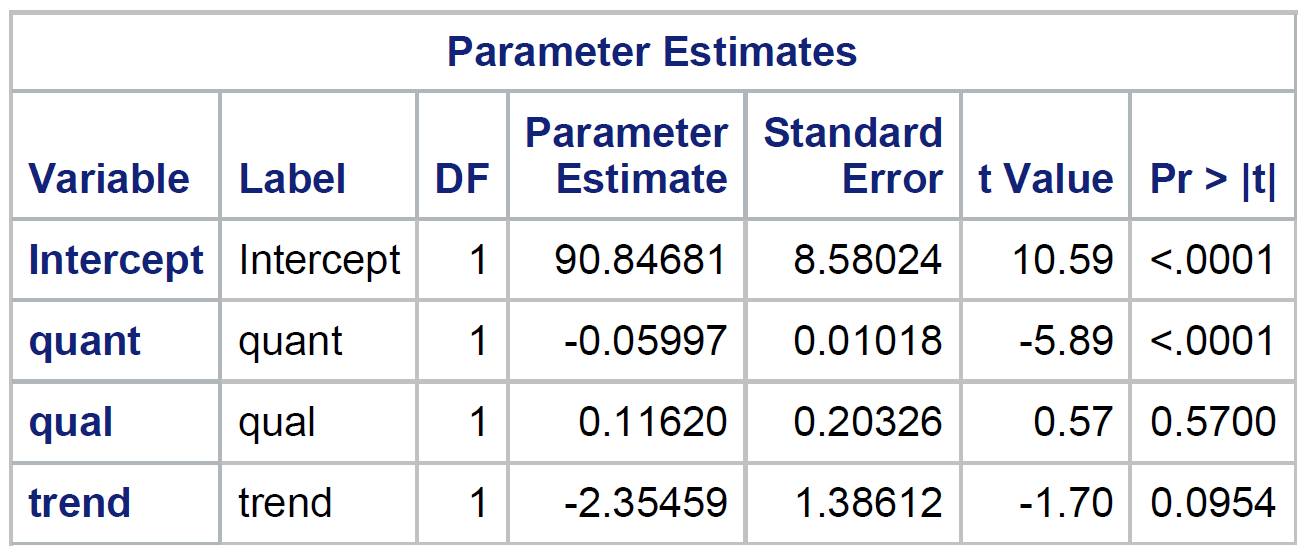
estimates of beta2 , beta3? (3 pts)

beta2 should be negative due to economies of scale. Nonetheless, due to the nature of the business a higher quantity carries higher risk, which could imply that quantity is correlated to e in some way. Further analysis required.

beta3 should be positive due to a higher quality product being more costly to produce.

b) Use SAS to find the equation. Report the results and interpret

the coefficients. Have the signs turned as you expected ? (6 pts)



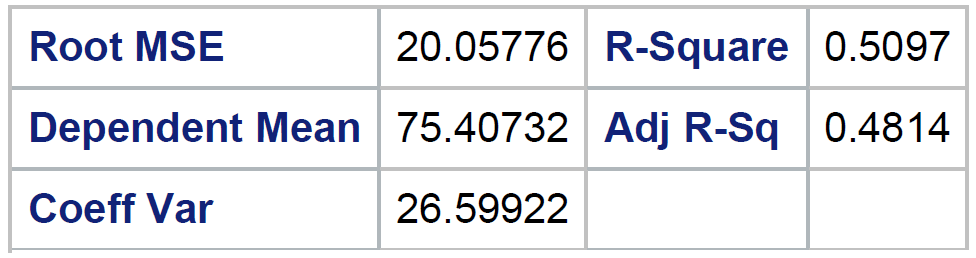
Price = 90.84681 – 0.05997\*quant + 0.11620\*qual – 2.35459\*trend

(8.58024) (0.01018) (0.20326) (1.38612)

Yes the signs are what we expected, but we can see that qual and trend are not statistically significant under our regular thresholds.

c) What proportion of variation in cocaine price is explained by

variation in quantity, quality, and time? (3 pts)



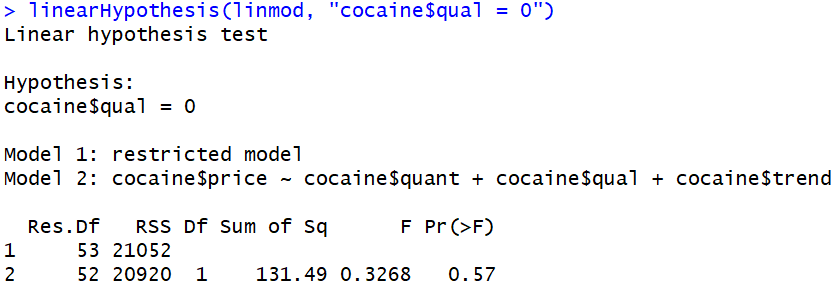
From the SAS output we can see that 50.97% of the variance of cocaine prices in California between the years of 1984 to 1991 can be attributed to quantity, quality and the overall time trend

d) It is claimed that the greater the number of sales, the higher the risk of getting caught; and thus sellers are willing to accept a lower price if they can make sales in large quantities. Set up H0and Ha that would be appropriate to test this hypothesis. Carry out the hypothesis test. (5 pts)

H0: beta2 < 0

HA: beta2 >= 0

e) Test the hypothesis that the quality of cocaine has no influence on price. (5 pts)



Therefore, we cannot reject the null hypothesis that quality does not affect the price of cocaine.

f) What is the average annual change in the cocaine price? Can you suggest why price might be changing in this direction? (5 pts)

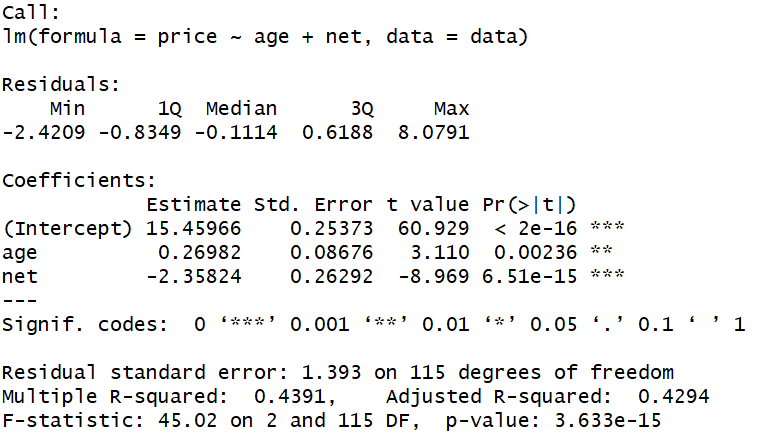
–2.35459 (1.38612) Possibly two reasons:

1: As the DEA narrowed down on the trafficking routes, the drug dealers were willing to lower the price in order to sell the merchandise faster.

2: As the drug cartels consolidated their trafficking routes, bringing cocaine into California was easier, therefore because o economies of scale they sold it cheaper in order to be able to sell more



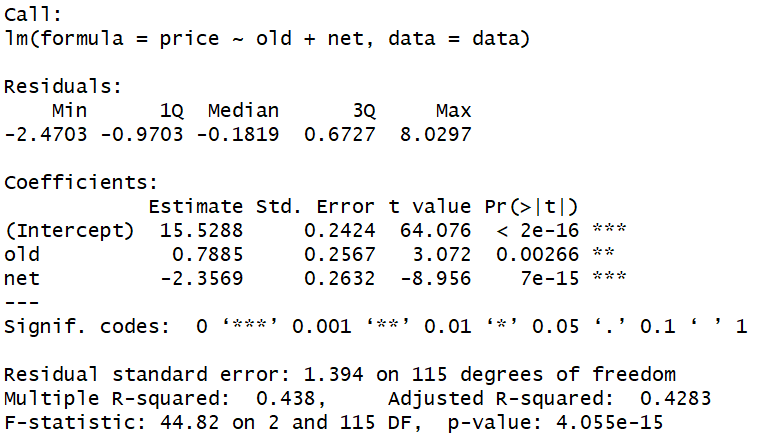
a)



We first note that both age and net are statistically significant, and despite the R2 = 0.4391 we can see a low RMSE = 1.393. Therefore these variables are a good representation of at least a portion of the variability of price.

We can see that approximately on the average age adds $0.26982 worth of value to a CD, while it being on the net **decreases** its price approximately on the average by $2.35824. Which is a considerable amount when taking into account the price range of CD’s is between 11.49 and 21.99 in this sample.

b)



We first note that both old and net are statistically significant, and despite the R2 = 0.438 we can see a low RMSE = 1.394. Therefore these variables are a good representation of at least a portion of the variability of price.

It is interesting to note that old and net being dummy variables can at most affect the price of a CD by 3$ and this is also (almost) the difference in price between the 1Q (12.99) and 3Q (15.88) of the distribution of price in this sample. And remembering that 50% of the data lies within 1Q and 3Q it is even more interesting that the R2 is close to 50%.

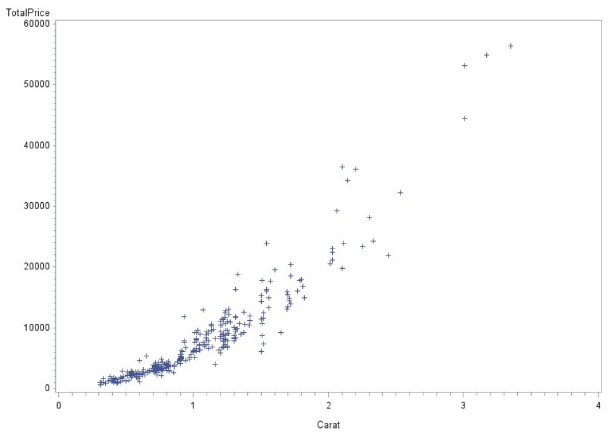
We can see that approximately on the average, a disc being old adds $0.7885 worth of value to a CD, while it being on the net **decreases** its price approximately on the average by $2.35824. Which is a considerable amount when taking into account the price range of CD’s is between 11.49 and 21.99 in this sample.

**Problem 7.** Mark is shopping for a diamond and is interested in learning more about how these

gems are priced. He has heard about the four C’s: carat, color, cut, clarity. Now he wants to see

if there is any relationship between these diamond characteristics and the price. (SAS problem)

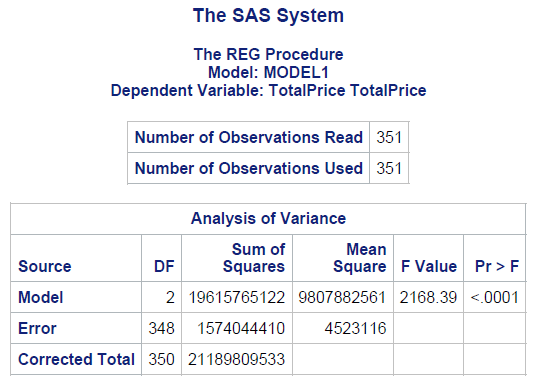
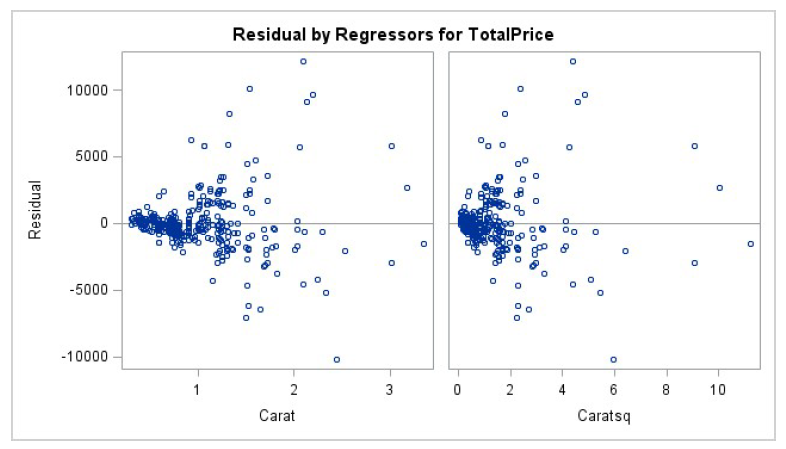
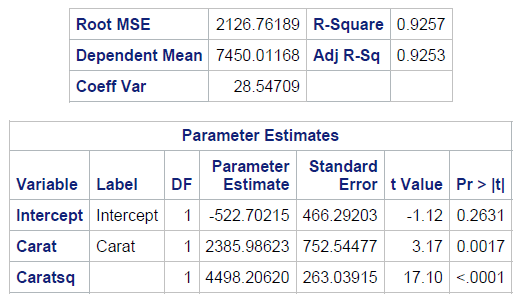
Mark has a dataset from AwesomeGems.com (July, 2005) n = 351.

a. Plot TotalPrice versus Carat for diamonds and describe the graph.

We can see a clear relationship between Price and Carat, yet it does not look linear as it is convex. There is also an abundance of data in the range of (0,2) Carats, and not much afterwards’

b. Run a Quadratic Regression Model with TotalPrice as response and Carat as explanatory.

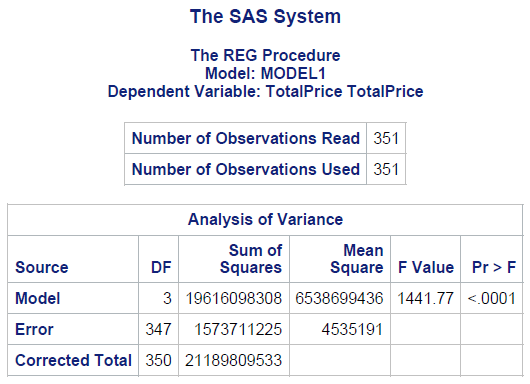
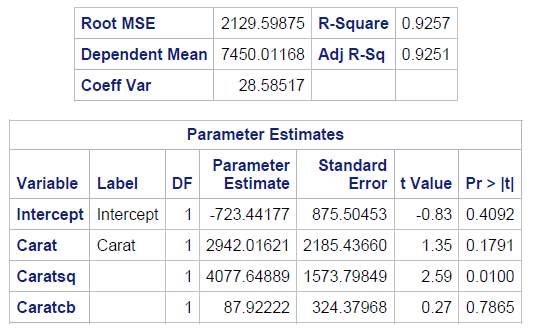
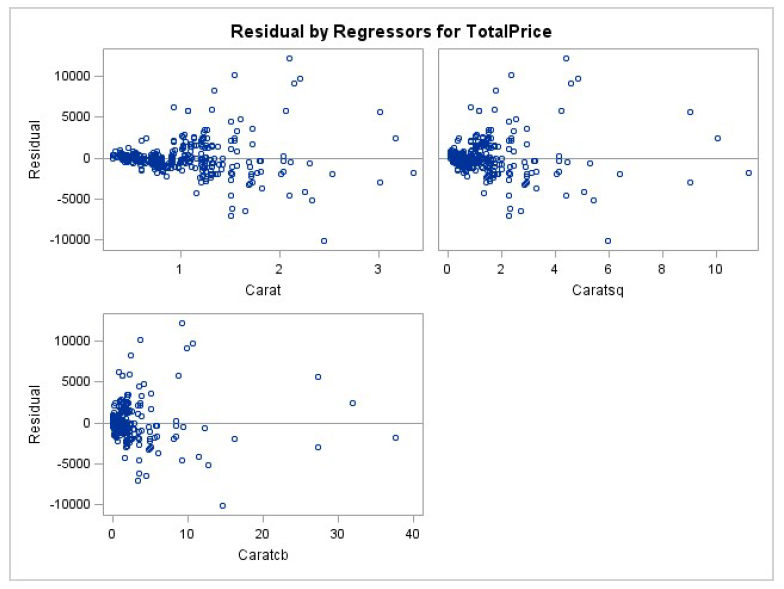
c. Describe the residuals.

In this regression we can see that carat and caratsq are statistically significant, and there is a high R2 yet the RMSE is also very high which is a clear indicator that something is not quite right. And in the residual plots we can see what is wrong, as there is some clear clustering and the residuals of carat clearly funnel out. An indicator of heteroskedasticity.

d. Run a Cubic Regression Model with TotalPrice as response and Carat as explanatory.

e. Describe the residuals.

In this regression we can see that carat and caratcb are not statistically significant, while caratsq barely is. Again, we can see there is a high R2 yet the RMSE is also very high which is a clear indicator that something is not quite right. And in the residual plots we can see what is wrong, as there is some clear clustering in all three plots and the residuals of carat clearly funnel out. An indicator of heteroskedasticity.

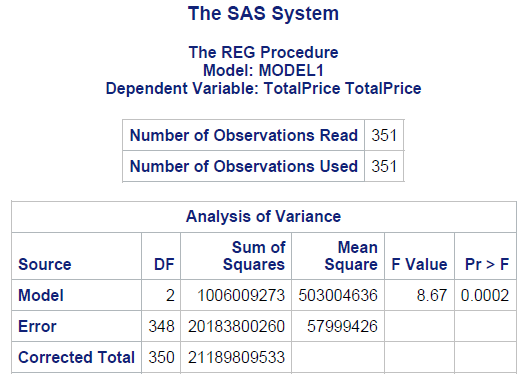
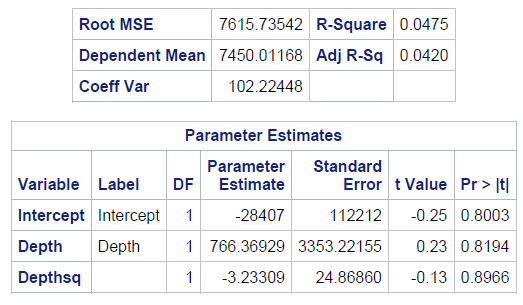
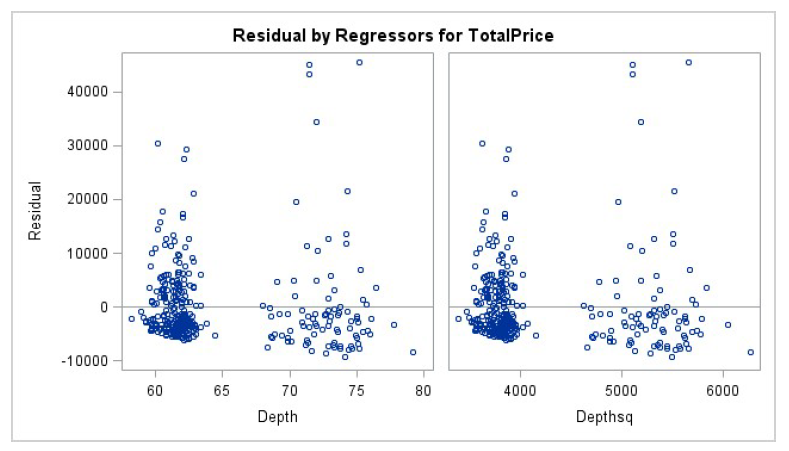
f. Make a decision between these models by looking at RMSE, Adjusted R^2 and residuals

and p-values for individual tests.

Model b is better, as adding the caratcb variable increased the standard errors of every other variable.

g. Run a quadratic model using Depth.

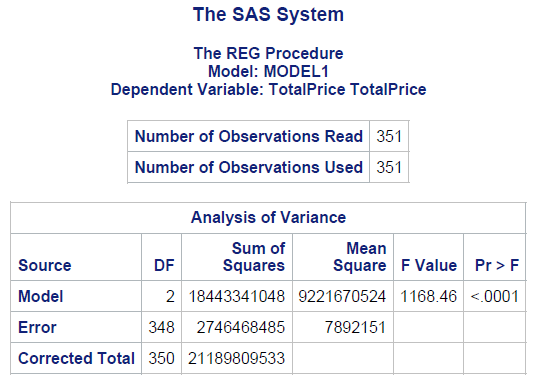
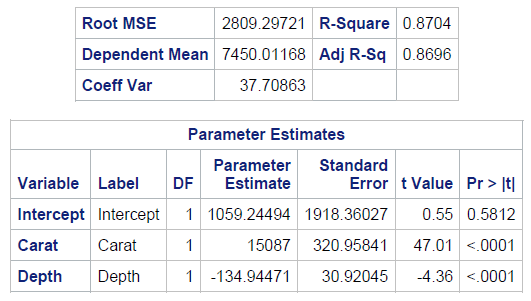
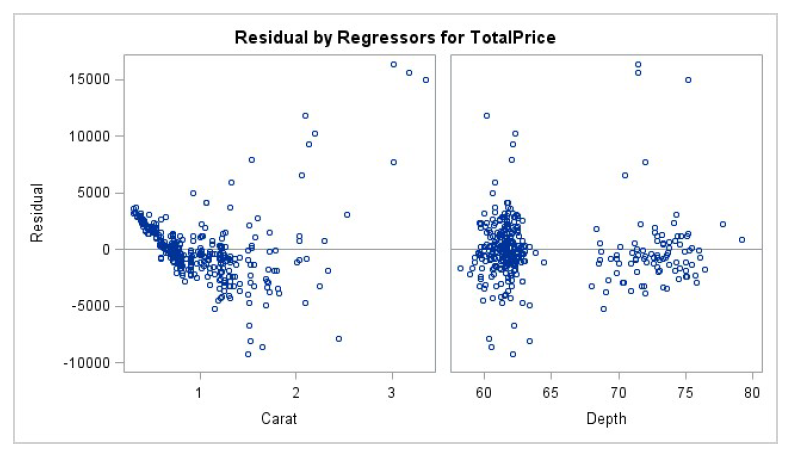
h. Describe Residuals, Interpret RMSE.

In this regression we can see that neither depth or depthsq is statistically significant, and there is a very small R2 and a very high RMSE. Therefore, we can conclude that this is not a very good model. And in the residual plots we can see what is wrong, as there is two very distinct groups with vertical patterns.

i. Run a two predictor model using Carat and Depth.

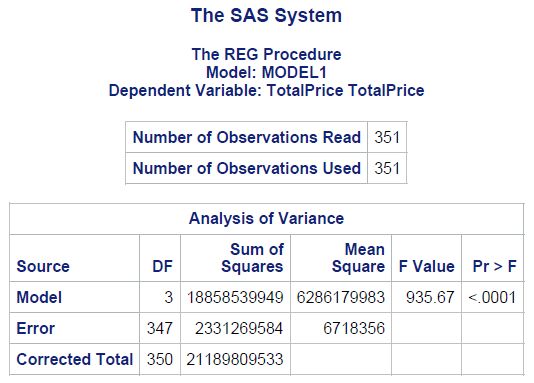
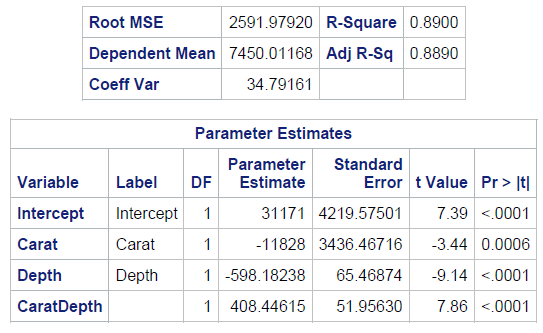
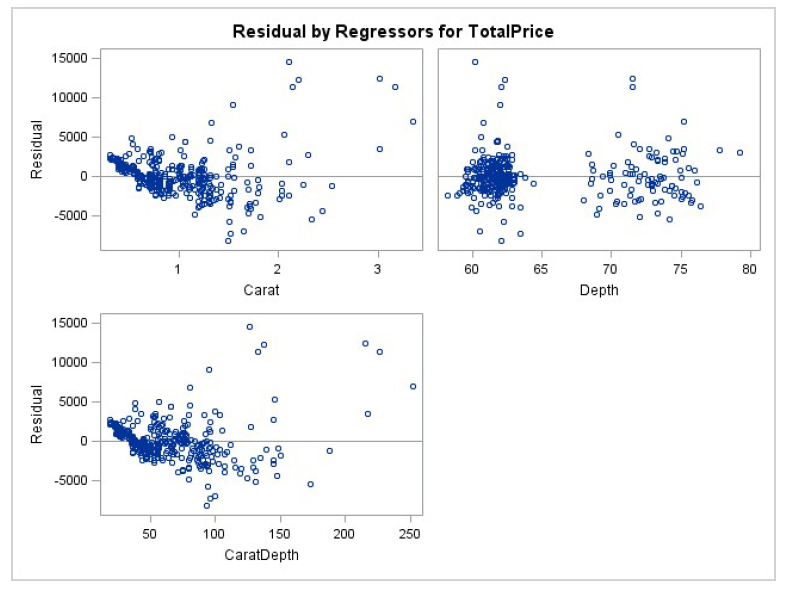
j. Describe Residuals, Interpret RMSE.

In this regression we can see that all regressors are statistically significant, and there is a high R2 yet the RMSE is also very high which is a clear indicator that something is not quite right. The residual plots indicate some of the issues, as there is some clear clustering on the depth residuals plot, and the residuals of carat clearly funnel out

k. Run a three predictor model that adds interaction for Carat and Depth.

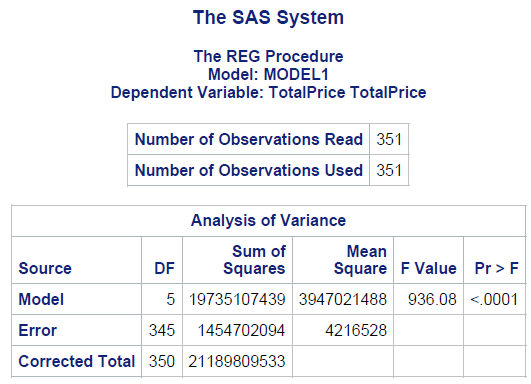
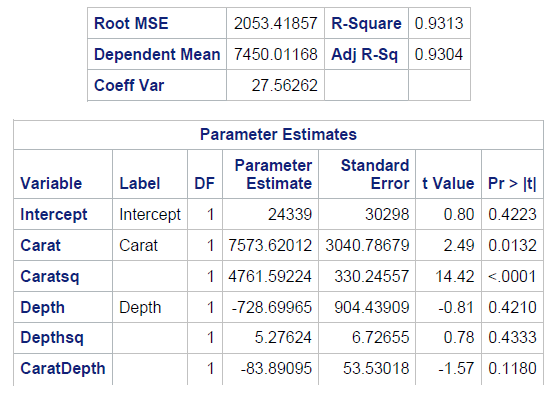
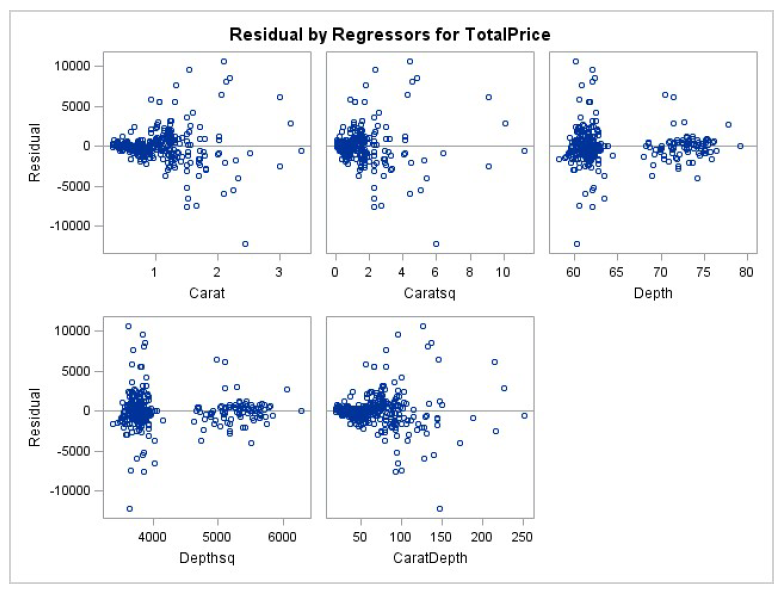
l. Describe Residuals, Interpret RMSE.

In this regression we can see that all regressors are statistically significant, and there is a high R2 yet the RMSE is also very high which is a clear indicator that something is not quite right. The residual plots show the same issues as before, as there is some clear clustering in the residual plot od depth and the residuals of carat clearly funnel out. Moreover, CaratDepth does not present the clustering that Depth does, yet it still funnels out.

m. Run a complete second order model using Carat and Depth. Note: A complete 2nd order

model has two predictors with linear terms, quadratic terms and an interaction term.

In this regression we can see that the carat regressors are statistically significant, yet the depth ones are not. There is a high R2 yet the RMSE is also very high which is a clear indicator that something is not quite right. On the residual plots we see all kinds of clustering and funneling out.

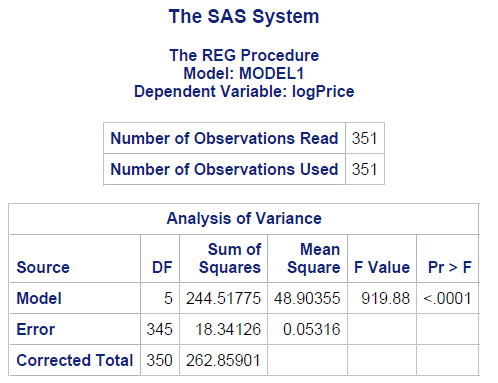
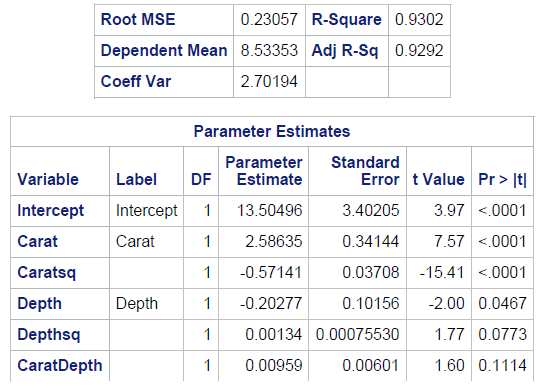
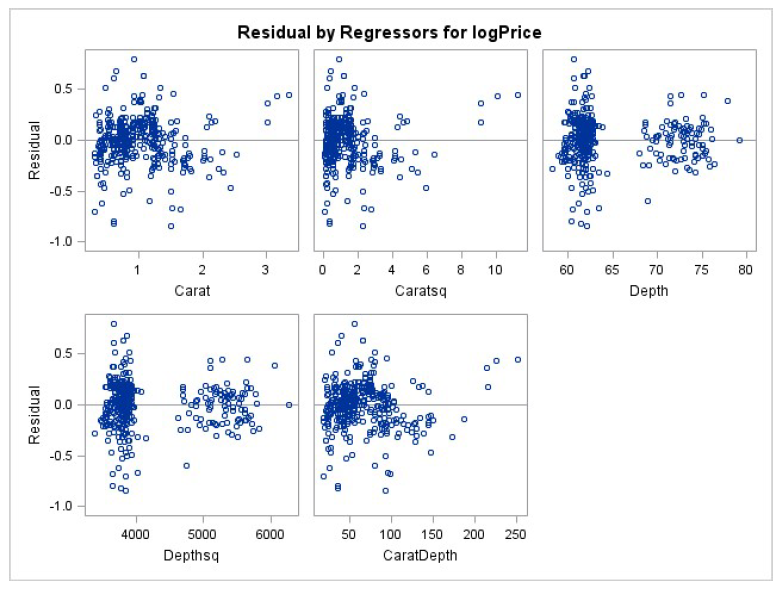
n. From a to m which model would you recommend using for predicting TotalPrice.

M, as it provides the lowest RMSE with highest R2.

o. I’m aware that we have a problem with residuals. The variability of the residuals tends to

increase for the more expensive diamonds. Using the model you picked on part n

transform the response variable to logPrice. Checked the residuals.

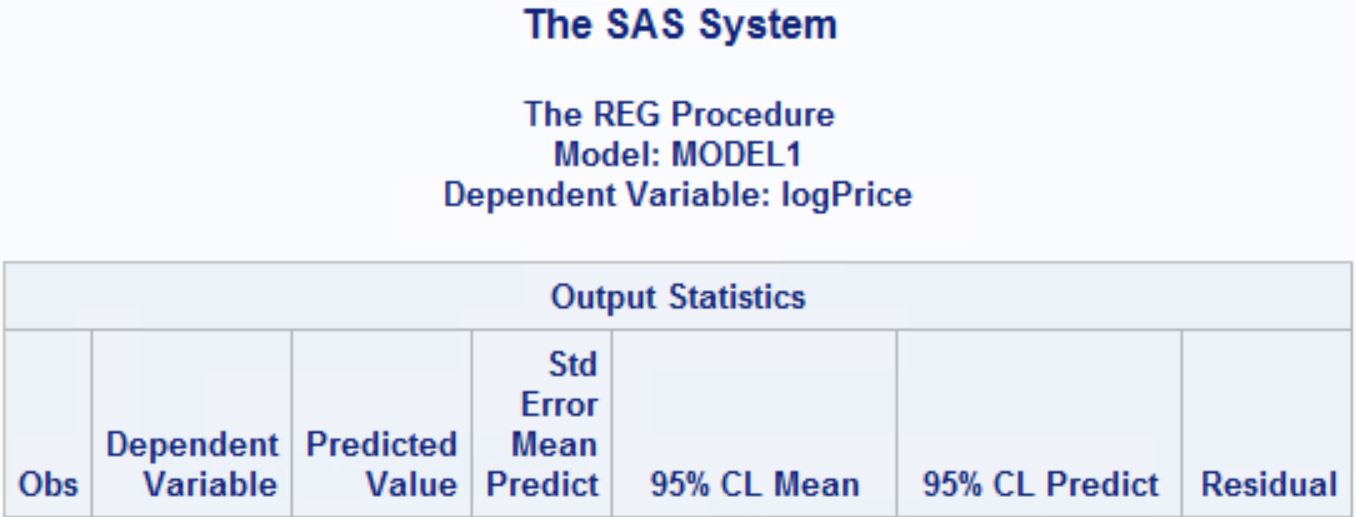
Depth continues to be heavily clustered into two groups, but carat seems to have become more random. Therefore, we can see that the residuals involving carat have less defined patterns.

p. What average total price does the quadratic model predict for a 0.5 carat diamond? (with

logPrice as response, remember to use the inverse of the natural log).

Ybar = 6.13042 + 3.05963(0.5) – 0.52730(0.25) = 7.52841

Exp(Ybar) = 1860.14551962



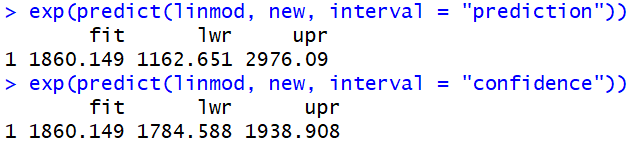


q. Compute 95% confidence intervals and 95% prediction intervals for question p. Write

conclusions.

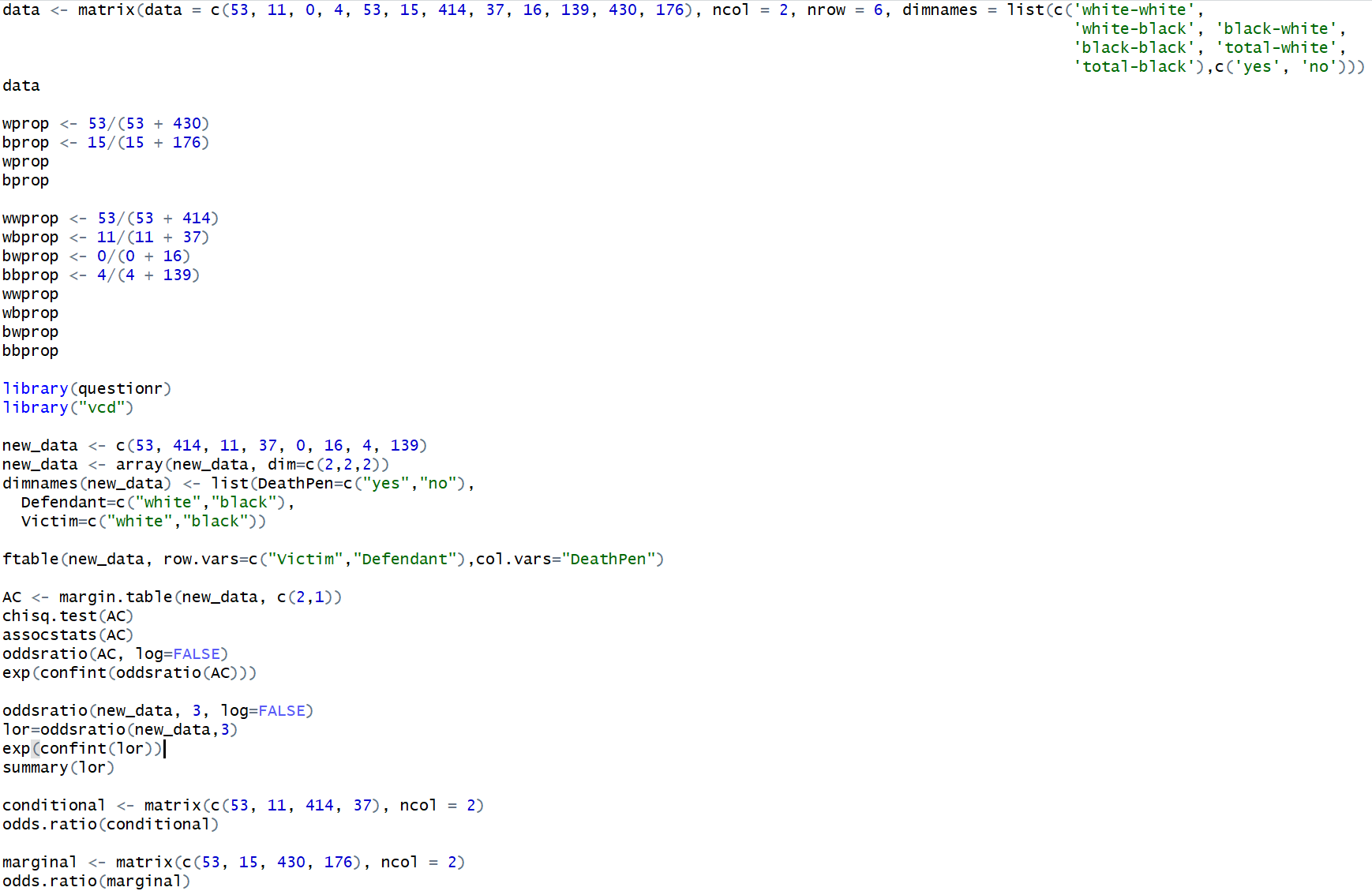
CI = (exp(7.4869),exp(7.5699)) = (1784.51152268, 1938.94637723)

PI = (exp(7.0585), exp(7.9984)) = (1162.69980737, 2976.19226785)

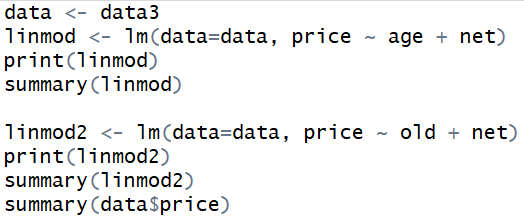


Code

Problem 1



Problem 5



Problem 7

