**11.0 – Terms Based on Nominal or Ordinal Predictors**

In section 9 we considered terms created from nominal or ordinal predictors and we review them below.

**Dummy terms - for dichotomous nominal/ordinal predictors**

Dichotomous predictors have two levels, e.g. a predictor that is “Yes” vs. “No” as in the Fireplace predictor in the Saratoga, NY homes data.

or

**Factor terms – for nominal or ordinal variables with more than 2 levels**

Suppose the predictor is a nominal or ordinal variable with levels (. Then we chose one of the levels as the ***reference group***and create dummy terms for the remaining  
 levels.

Here level is the reference group. In JMP the level that is alpha-numerically last is the reference group, in R the level that is alpha-numerically first is the reference group.

We consider again the coding of the Fuel Type predictor in the Saratoga, NY homes data.

e.g. Suppose

JMP would choose *4 = Oil* as the reference group and create dummy variables for the other two fuel types, i.e.

In R, would be the reference group.

**Example 11.1 – Saratoga NY, Homes**

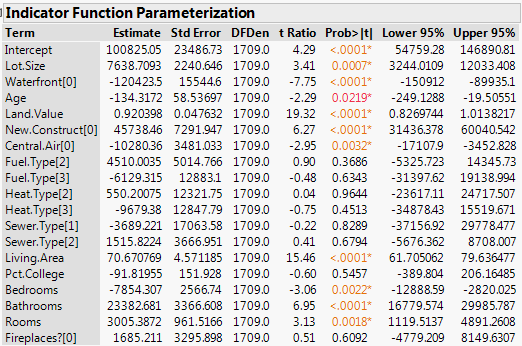
We again consider the prices of homes in Saratoga, NY dataset. The variables we will consider in this example are as follows:

**Variables:**

* Price – price of the home in dollars
* Lot.Size – acres
* Waterfront – is the home located on a waterfront (0 = no, 1 = yes)
* Age – age of the home (yrs.)
* Land.Value – assessed value of the land the home is on
* New.Construct – is the home a new construction (0 = no, 1 = yes)
* Central.Air – does the home have central air conditioning (0 = no, 1 = yes)
* Fuel.Type – type of fuel used to heat the home (2 = Gas, 3 = Electric, 4 = Oil)
* Heat.Type – type of heating system (2 = hot air, 3 = hot water, 4 = electric)
* Sewer.Type – type of sewer system (2 = none, 3 = private, 4 = public)
* Living.Area – living area (ft2)
* Pct.College – percentage of college housing in the home’s neighborhood
* Fireplaces – number of fireplaces in home (0,1,2,3, or 4)
* Bedrooms – number of bedrooms
* Rooms – number of rooms
* Fireplace? – does the home have at least one fireplace (0 = no, 1 = yes)

As we can see there are a number of nominal variables in these data some which that are dichotomous (e.g. ) and some that are factors with more than 2 levels (e.g. ). We will begin by fitting a full model in JMP using all available predictors, with the nominal variables being converted to terms automatically.

Below are the **Indicator Parameterization Estimates** of the regression coefficients:



Terms for dichotomous predictors

Terms for multi-level factors

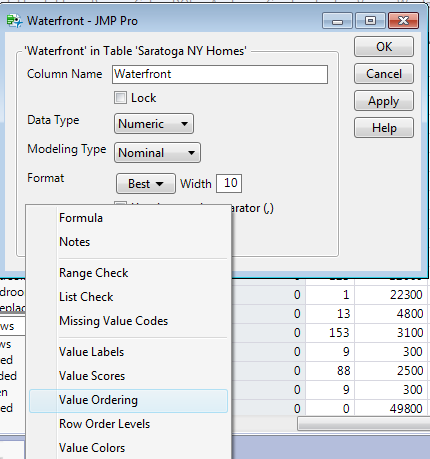
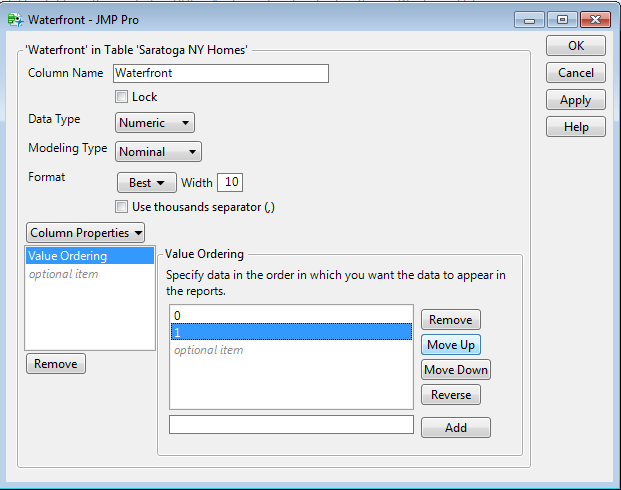
The coding on some of the terms is not ideal, for example consider the term for



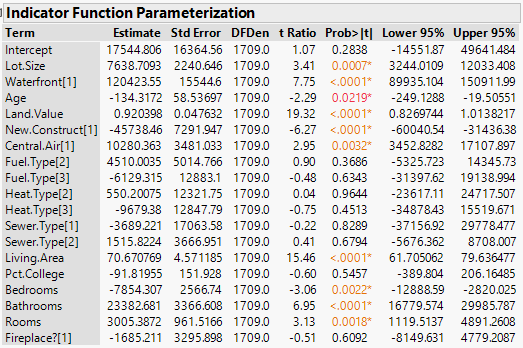
In JMP the level that is considered as the 1-level is placed in square brackets, thus

Waterfront[0]

We would probably prefer that the waterfront term be 1 if the home is on the waterfront and 0 if not. This can be accomplished by reordering levels of the predictor . Similar comments could be made for the central air conditioning, new construction, and the fireplace dummy variables. To reorder the values of a nominal/ordinal variable right-click on the column you wish reorder and select **Column Info**… . Then from the **Column Properties** drop-down menu select **Value Ordering** and **Move Up** the 1-level to the top of the list.

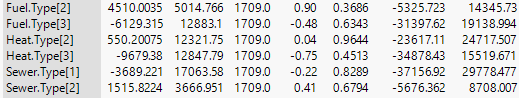
 

After moving the 1-level to the top for all of the dichotomous nominal predictors we obtain the following parameter estimates.



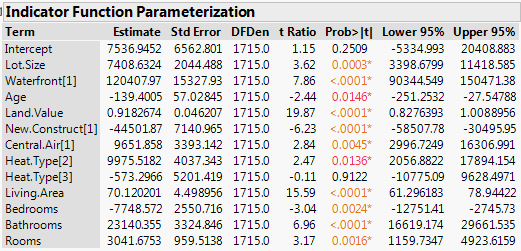
What are the estimated changes in the mean response for   
these terms?

For the factors (nominal variable with more than 2 levels) we can see for the fuel type and heat type the reference group is level-4 for each and level-3 for sewer type.



We can change the reference group by moving level we want to be the reference group to the bottom of the list using the **Column Info…> Column Properties > Value Ordering** option. However before pursuing this we would need to consider that some of the factors included in the current model are not significant thus we might remove some of them from our model.

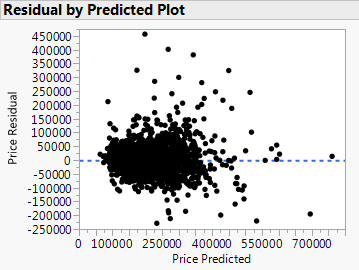
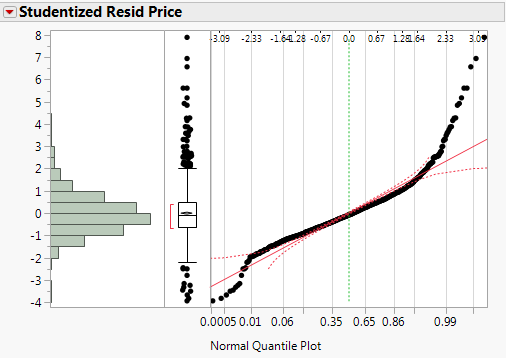
Performing backward elimination for the full model we arrive at the following model:



Interpretation of some of the estimated coefficients:

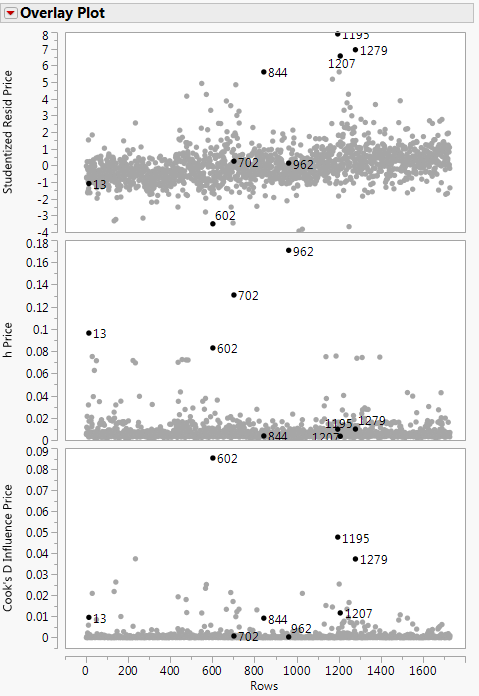
Examining residual and case diagnostics for the “final” model chosen via backward elimination we obtain the following.

**Residual plots** - and normal quantile plot of standardized residuals ()

**Case diagnostics**

**Comments:**



**Example 11.2 – Testing Wool Samples**

In some experiments, data are collected at all combinations of levels of a few factors. For example, a factorial experiment has three factors, and the first factor has two levels, the second has three levels, and the third factor has four levels. In a single replicate of the experiment one observation is taken for each of the combinations of the factors, thus in total there will be observations per replicate. Often times these experiments are done with more than one replicate, e.g. if this experiment was replicated 5 times we would have a total of n = 120 observations. In experiments with a large number of factors fractional replication is used in which some factor combinations are not run. This is called a Fractional Factorial Experimental Design and are commonly used in experimentation in an industrial setting. To learn more about this type of experimentation, take STAT 321 – Industrial Design of Experiments.

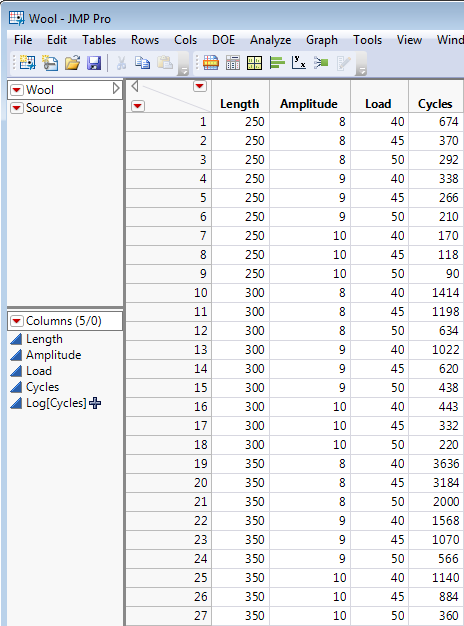
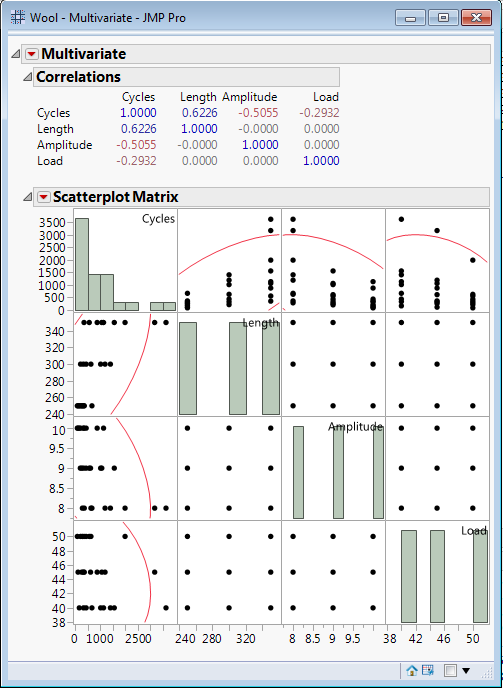
In this example we analyze data collected from a factorial experiment. As a single replicate requires n = 27 experimental runs, this experiment was not replicated.

The variables in this experiment are:

* Cycles – number of cycles until the wool specimen fails
* Length – length of the test specimen (250, 300, 350 mm)
* Amplitude – amplitude of the loading cycle (8, 9, 10 mm)
* Load – load put on the specimen (40, 45, 50 g)

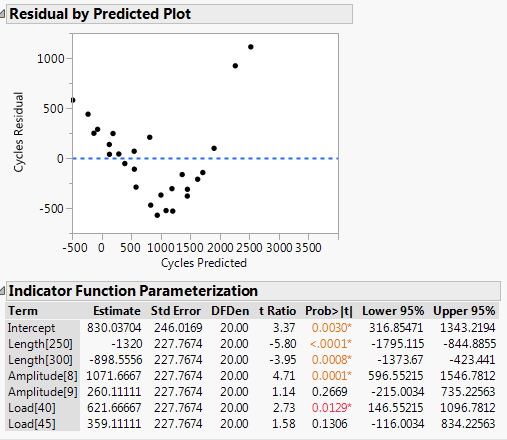
The data table and a scatterplot matrix of these data are shown below.

**Comments:**

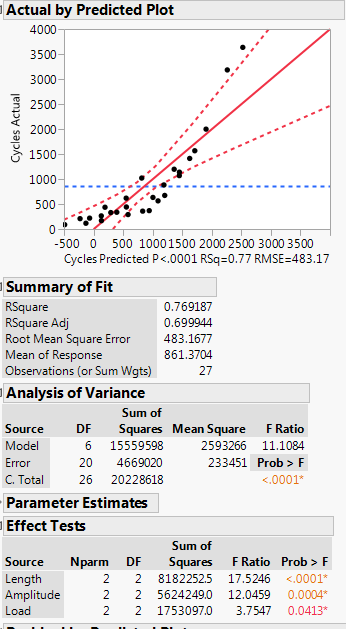
 

Clearly the predictors can be interpreted as numeric/continuous variables, although it is more standard to treat them as factors, i.e. nominal variables with three levels each in this case. Also we could treat them as ordinal variables and think of the setting for these predictors as “low”, “medium”, and “high”. We will compare and contrast the “final” model developed under each of these situations, but will begin by treating them as factors. We first fit a model using all three factors coded automatically by JMP, i.e.

where the terms are denoted



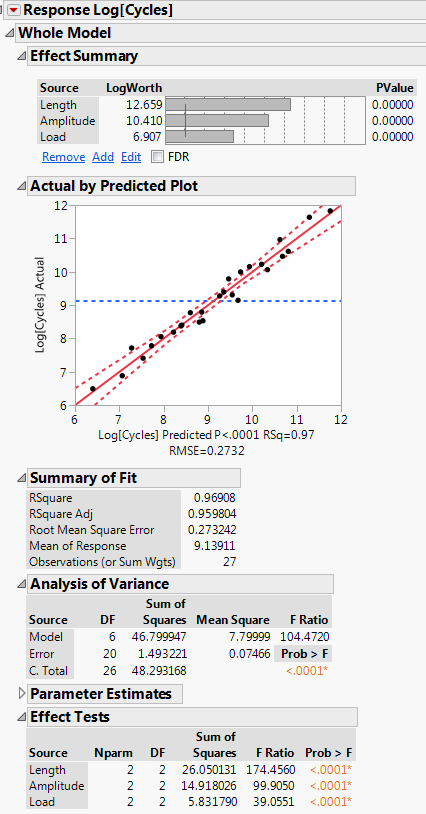
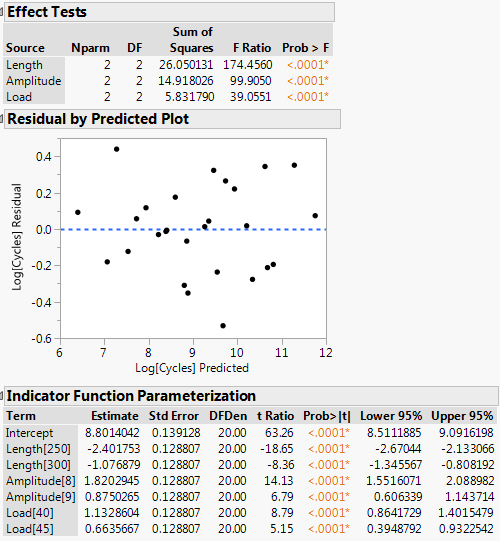
**Comments:**

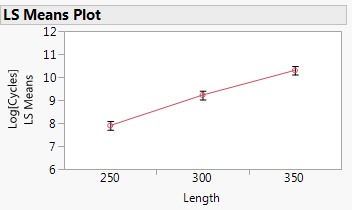
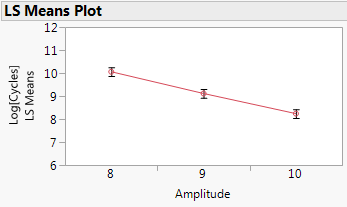


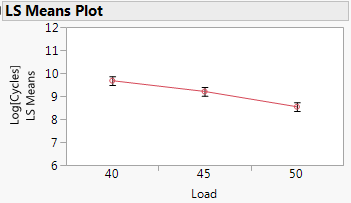
Clearly this model is deficient in terms of both the mean and variance functions.

The typical fix for nonconstant variation is to log transform the response. Also from the scatterplot matrix above we know that the response is skewed right.

Fitting the model using as the response gives the following.



The reference group for each of the factors is the highest setting for each. We can change the reference group to the lowest setting by using the **Value Ordering** option in the **Column Info…** as described above.

The use of the as the response complicates the interpretation of the estimated coefficients, however the LS Means Plots show the effects leading to conclude that longer specimens have longer cycle times, as amplitude and/or load increase the cycle time decreases.

To interpret the estimated coefficients we need to consider the multiplicative effect on the number of cycles until the specimen fails associated with the different settings relative to the reference group.

**Examples**:

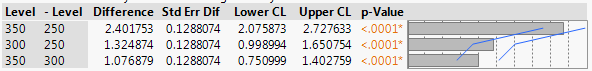
Length[250] vs. Length[350] – the estimated coefficient for Length[250] is , thus the multiplicative decrease in the number of cycles is or the number of cycles for length = 250 mm specimens is 18.95% of the number of cycles for length = 350 mm specimens, or a 81.05% decrease in the number of cycles. Another way to interpret this is that the number of cycles for 350 mm specimens is times larger than the number of cycles for 250 mm specimens.

Amplitude[8] vs. Amplitude[10] –

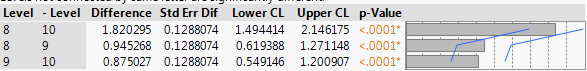
Load[40] vs. Load[50] –

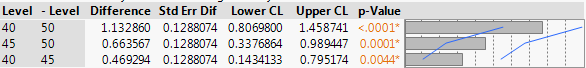
95% CI’s for the Differences in the Median Number of Cycles

Length



Amplitude



Load  


**Example 1: Diamond Prices and the 4 C’s**

**A diamond’s quality (and hence price) is determined by the 4C’s.**

**Color** – colorless to light yellow (see below)

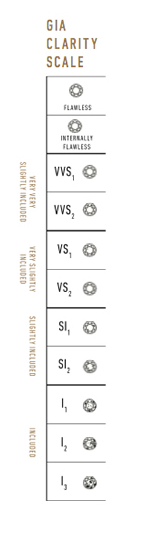
**Clarity** – see GIA scale on the right

**Cut** – grade of cut   
 (good, very good, excellent, ideal)

**Carat** – size/weight of the diamond   
 (1 carat – 200mg or 1/5 of a gram)



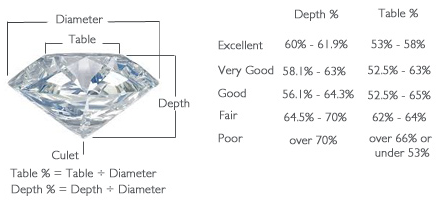


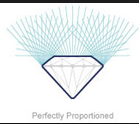
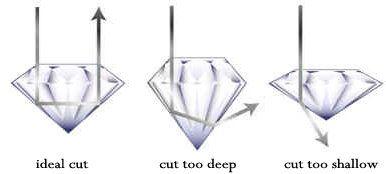


Also the Total Depth (%) and Table (%) are important characteristics

that deal with the shape of the diamond. These measurements are   
shown in the figure below, along with guidelines for quality of the

diamond based on these measurements.





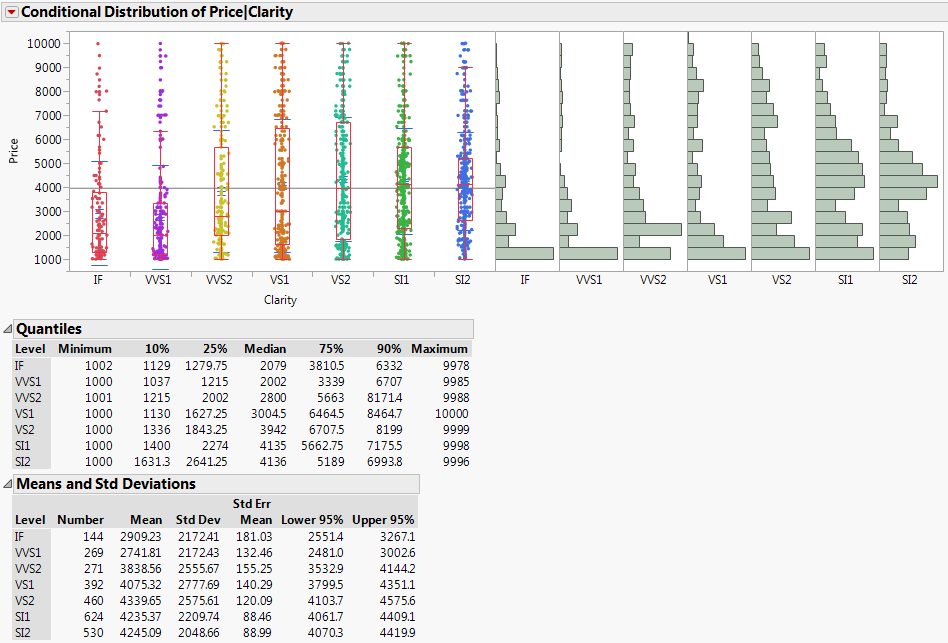
Our goal is to develop an accurate model for predicting or determining the price of a diamond based upon these diamond characteristics. We will first develop a multiple linear regression model for these data.

**Datafile:** **Diamonds.csv**  
These data contain the 4C’s, depth (%), table (%), and price for a sample of n = 2,690 diamonds. There are additional columns containing the natural logarithm of the price (Log Price), the difference between the Table(%) and Depth (%) values as well as their ratio (Table/Depth), i.e. two new terms created by using functions of these predictors. Our goal is to develop a regression model for the diamond price using the available predictors and terms.

**Predictor/Term Descriptions:**

* Carat Size – carat weight of the diamond (e.g. 0.5 = ½ carat)
* Color – D,E,F,G,H,I,J,K (see diagram above)
* Clarity – (ordinal, lowest = *SI2*, highest = *IF*)
* Depth – depth (%)
* Table – table (%)
* Cut – good, very good, excellent, ideal
* *Report – GIA, AGS (ignore this)*
* Price – price ($)
* Log Price – natural logarithm of price (ln($))
* Table-Depth – the difference between table (%) – depth (%)
* Table/Depth – table(%)/depth(%), i.e. the ratio of table to depth
* *Test – used later in the course*

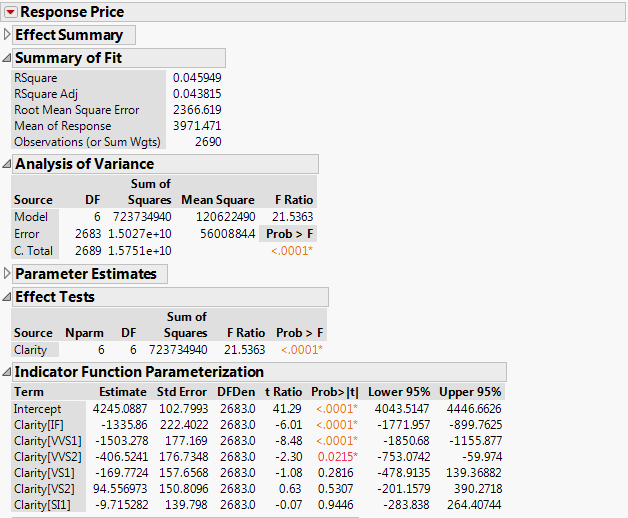
We begin by considering the regression of Price ($) on Clarity . Below is a descriptive analysis of the conditional distribution of Price ($) given Clarity using **Analyze > Fit Y by X**.

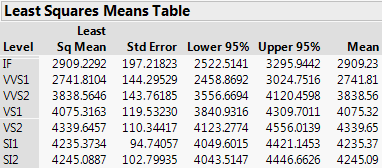


**Comments:**

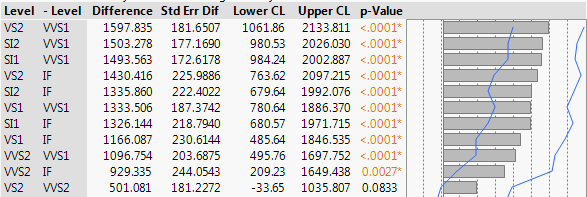
Does anything about these results seem counterintuitive?

Fitting the regression model using dummy terms created from the Clarity using as the reference group using **Analyze > Fit Model** we obtain the following.





CI’s for the differences in mean prices



Estimated Means for the Different Clarity Levels

The estimated mean function is given by:

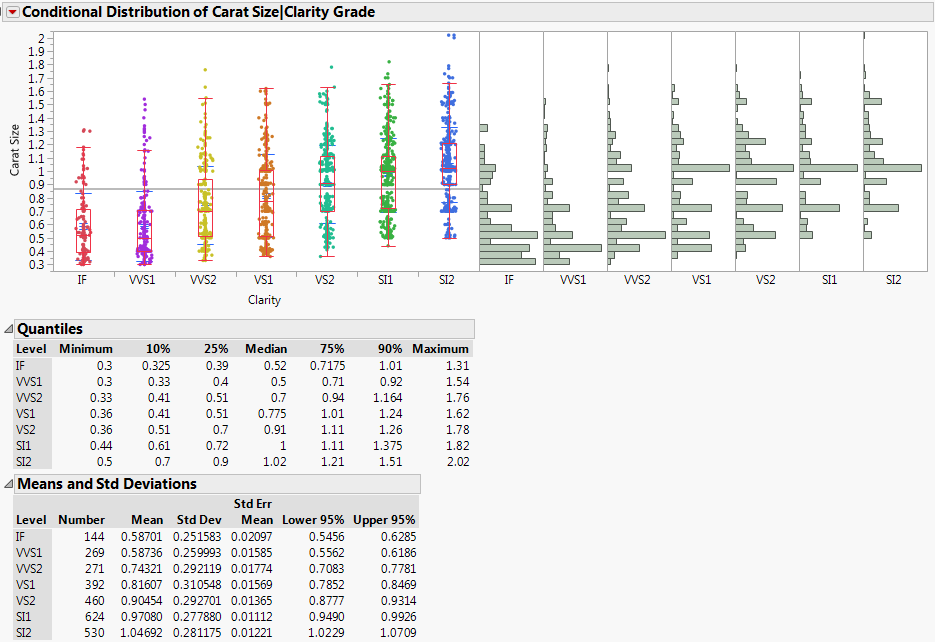
where

The confidence intervals for the differences in the mean prices any two levels are found using Tukey’s HSD, thus they are wider than the CI’s for the individual coefficients.

For example, the 95% CI for the is given by ($679.64 , $1992.08), thus we estimate the mean price of diamonds with clarity (the lowest grade in these data) is between $679.65 and $1992.08 larger than the mean price of diamonds with clarity (the highest grade in these data). Huh? A similar conclusion is reached when comparing clarity diamonds to clarity diamonds which are the second highest clarity grade (95% CI for ).

Why do you think this is happening?

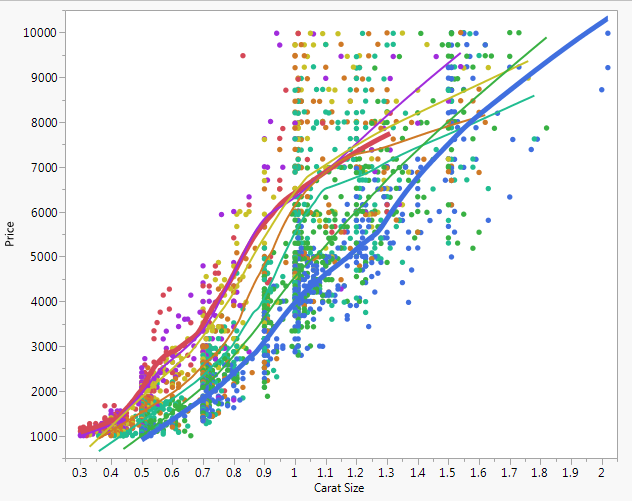
Below is an examination of the conditional distribution of given .

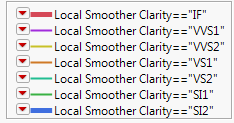


**Comments:**

What are the obvious consequences of the relationship between carat size and clarity?

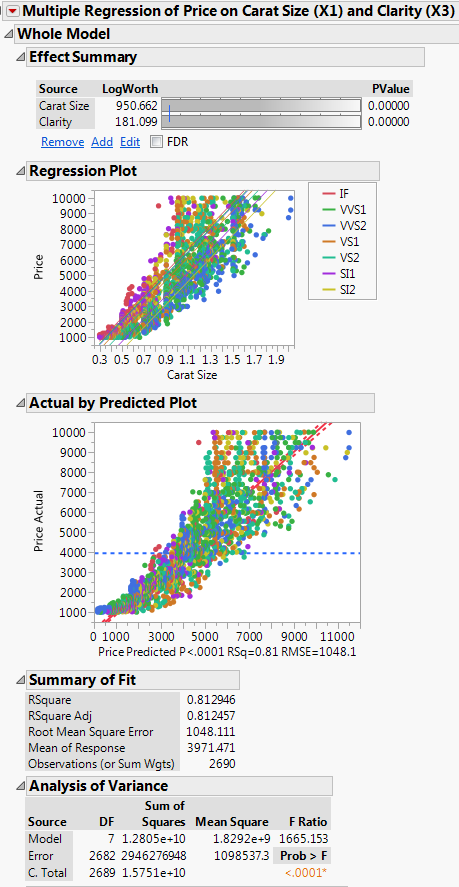
To confirm this, consider a scatterplot of a separate kernel smooth to visualize for each clarity level.

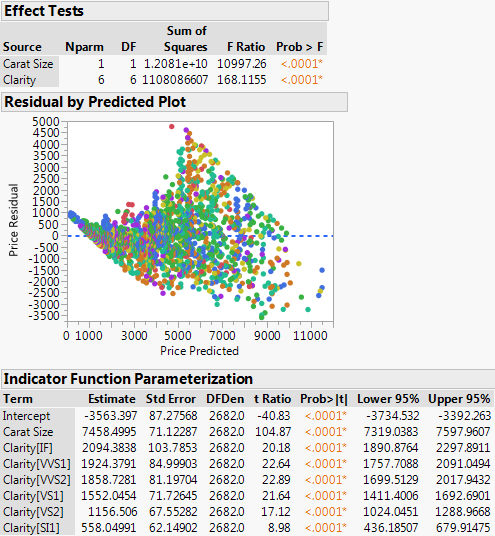


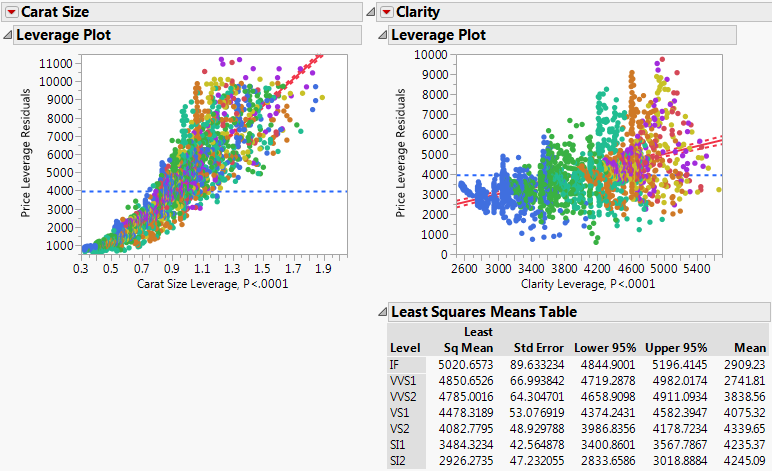


Clearly when comparing diamonds of the **same carat size** the highest clarity diamonds sell for a much higher price than the lowest clarity diamonds . In order the study the effect of clarity on diamond price we definitely need to take carat size into account.

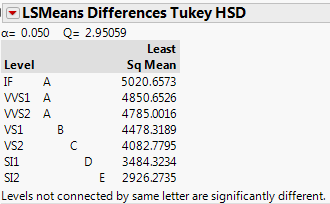
We now consider a multiple linear regression model that incorporates both predictors, and fit using **Analyze > Fit Model**.

Summary of the fitted model:

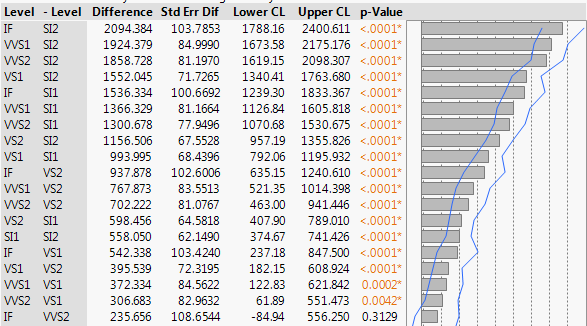




**Residual Plot**



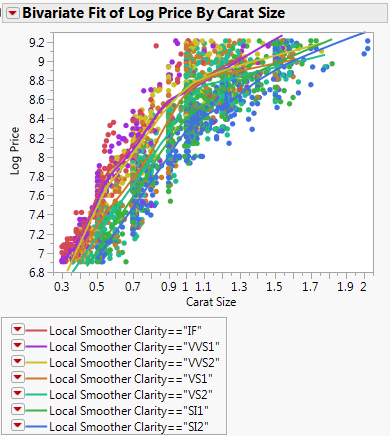
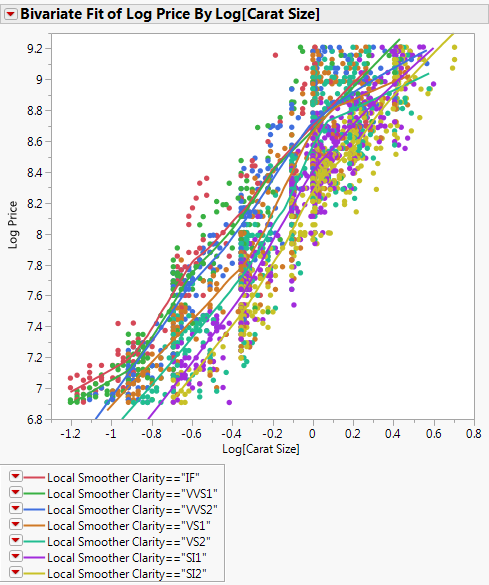
CI’s for mean price differences between clarity levels **holding carat size fixed**

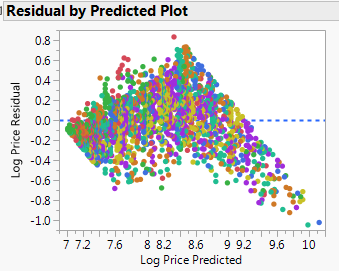


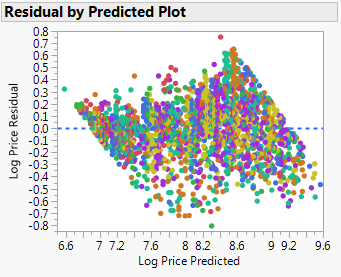
95% CI for

95% CI for

Clearly this model is deficient in terms of the assumed mean and variance functions. How can we address these deficiencies? Probably take the logarithm of the diamond price for starters and possible transform as well.  
  
log-linear – i.e log-log – i.e.





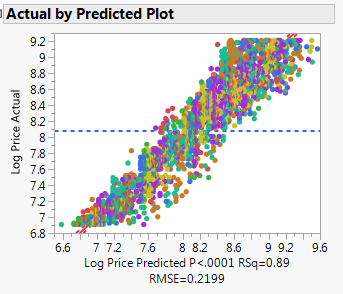
The log-log multiple regression model certainly seems to be a marked improvement, but interpretation of the estimated parameters will be a challenge.

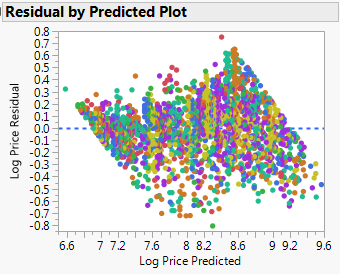
We will consider the following model using log(Price) as the response, the term , and the interaction of this term with factor terms for .

**Model with Clarity dummy terms, log(Carat Size), and their interaction.**

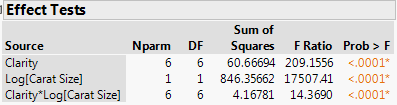
Thus the mean function has a total of parameters (1 for intercept, 6 dummy variables for the 7 level nominal/ordinal predictor clarity, 1 for the continuous term

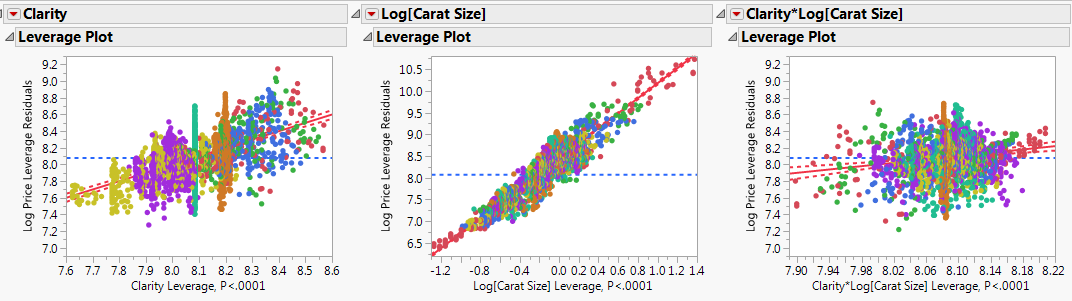
, and 6 more for the interaction of and the dummy variables. This model essentially fits a separate log-log line (unrelated lines) for each clarity level. The overall model summary is shown below. If dropped the interaction terms would have the parallel lines model.

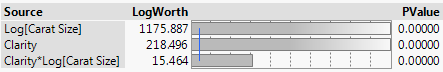




Big F-tests for dropping terms

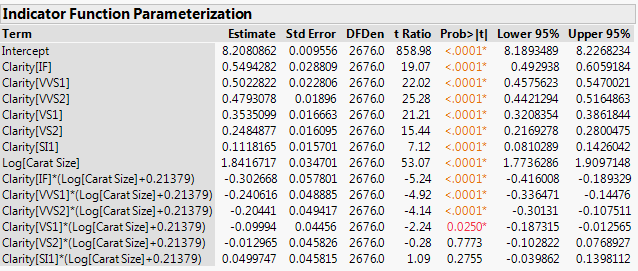


AVP’s for Clarity, log(Carat Size), and the Interaction Terms  




from the ***Big F-test*** for dropping the term(s) in question.

The parameter estimates using indicator parameterization are shown below.

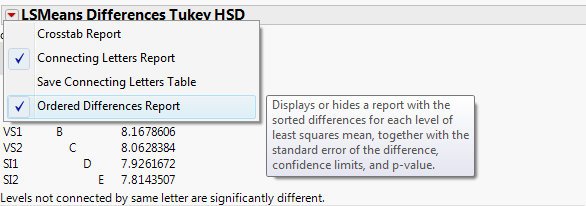


Estimated Mean Function for Clarity = *IF* Diamonds

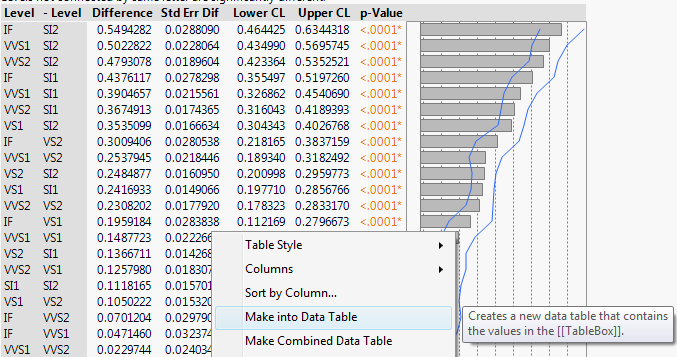
What is expected increase in price associated with a increase in the carat size of a diamond with Clarity = *IF?*  or a 15.4% increase (i.e. a percentage increase).

To compare the diamond prices across clarity taking diamond size into account we can use Tukey’s HSD which is an available option above the added variable plot for any factor term, e.g. clarity.



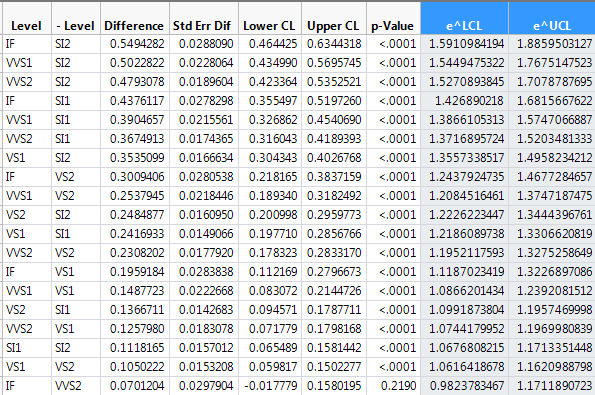


Tukey’s HSD CI’s for comparing “means” across levels of a factor



Right-clicking on table and selecting the **Make into Data Table** option will save them to a data table.

Converting CI’s for the mean differences in the log-price across clarity levels back to the original scale we have:

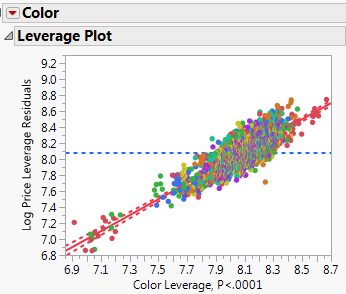
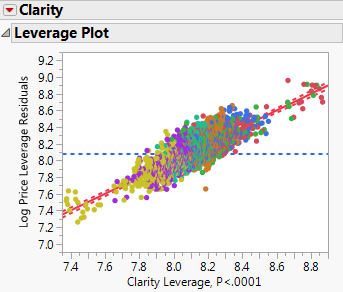
  
Thus comparing the diamonds to diamonds we see that for diamonds with the same carat weight, the IF diamond will typically be 59% to 88.6% more expensive, or 1.59 to 1.89 times more expensive. Similar comparisons can be made for other pairs of clarity levels.

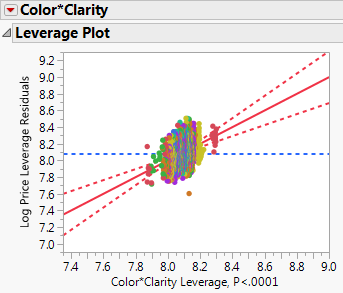
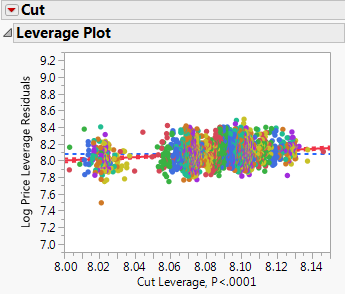
Recall

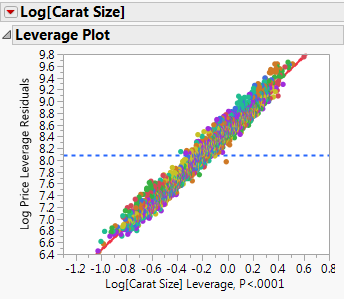
Thus exponentiating the log of the ratio gives the ratio,

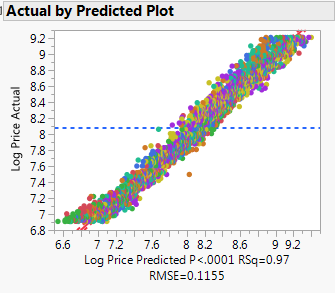
or

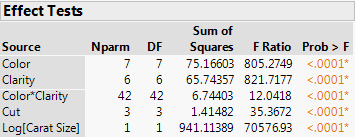
**Model using the 4C’s -** (more details later)

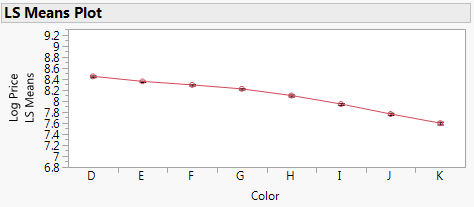
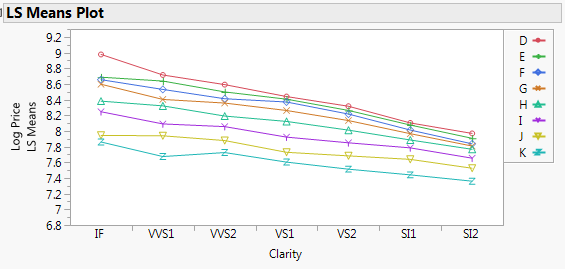
 

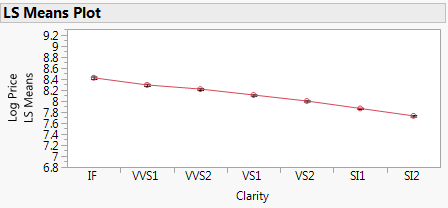
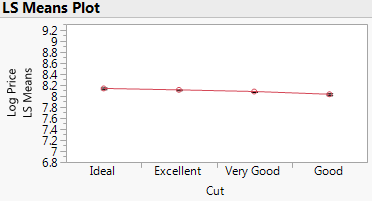
 

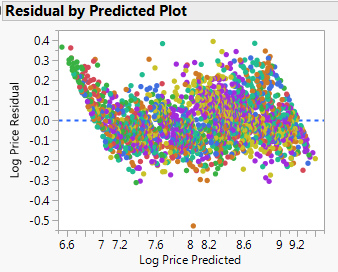
 AVP’s and tests for all terms







This model has 60 parameters, so we will not make any attempt to right it out!   
  
The breakdown of the is given below.  
Intercept 1 y-intercept parameter

Color 7 dummy terms

Clarity 6 dummy terms

Cut 3 dummy terms

Color\*Clarity 42 dummy term products

Log(Carat Size) 1 “slope” parameter

Total 60 parameters