**14.1 – Transformations of the Response**

The form of the multiple regression model is given by

and typically we assume . However, if examination of the residuals provides visual evidence that the assumed mean function is not correct and/or the then it may be the case that a transformation of the response, satisfies the assumed model. For example we have used as the response in several models in previous sections, which is one of the most common response transformations used. We have also seen that normality is a desirable property in regression, thus we may consider whether the distribution of is approximately normal in our choice of response transformation.

We can summarize the reasons for transforming the response as follows.

**Transforming to Linearity -** In cases where we have clear visual evidence or a curvature test result suggests that it is possible that the assumed model holds when the response is transformed , i.e. . As mentioned above we have seen numerous examples where transforming the response to the log scale, , remedied problems with the model using the untransformed response.

**Transforming to Stabilize the Variance –** In cases where we have clear visual evidence (and/or a NCV plot) or nonconstant variance test (score test) results to suggest that it is possible a response transformation will satisfy , i.e. is constant.

**Transforming to Normality –** In cases where the errors do not appear to be normally distributed, as evidenced by a normal quantile plot or histogram of the residuals or standardized residuals , a transformation of the response, , can improve normality of the errors. Such a transformation will also imply that the conditional distribution of .

**14.2 - Variance Stabilizing Transformations**

In section 13 we considered models for the variance, here we will assume

where is a function called the *kernel variance function*. For example if has a Poisson distribution, then the variance function is equal to the mean function, so .When this relationship holds the variance stabilizing transformation is .

Note:This is because the mean and variance for a Poisson random variable are the same. The Poisson random variable typically used to model the number of occurrences per time/space unit, thus the response is typically represents a count.

The table below from Cook & Weisberg (pg. 318) summarizes common variance stabilizing transformations used in OLS regression.

|  |  |
| --- | --- |
|  | **Comments** |
|  | Appropriate when for example when has a  Poisson distribution. The latter form is called the Freeman-Tukey deviate, and it gives better results if the are small or some . |
|  | Though most commonly used to achieve linearity, this is also a variance stabilizing transformation when . It can be appropriate if the errors are a percentage of the response, like , rather than an absolute deviation like units. |
|  | The inverse transformation stabilizes variance when  . It can be appropriate when responses are close to zero, but occasionally large values occur. |
|  | This is usually called the *arcsine square-root* transformation. It stabilizes the variance when is a proportion between zero and one, but it can be used more generally if has a limited range by first transforming to the range (0,1) and then applying the transformation. To scale a variable to [0,1] range use the formula . |

An alternative to variance stabilizing transformations is to use *generalized linear models* which we will cover towards the end of the course. For example, in cases where the response is Poisson distributed counts we can use *Poisson Regression* and when the response is a proportion, i.e. a binomial probability of “success”, rather than use , we can use *Binomial Regression (*which is more commonly referred to as *Logistic Regression).*

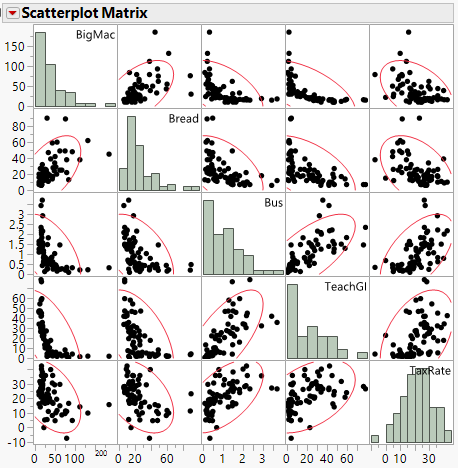
**Example 14.1 – Big Mac Data**

The data come from a study of prices in many world cities conducted in 2003 by the Union Bank of Switzerland.   
  
Source: Union Bank of Switzerland report, Prices and Earnings Around the Globe (2003 version) from <http://www.ubs.com/1/e/ubs_ch/wealth_mgmt_ch/research.html>

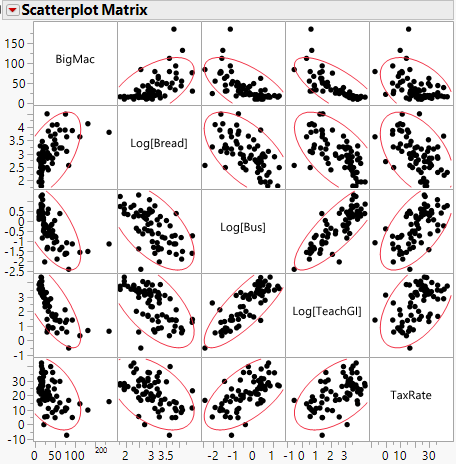


The variables in this dataset:

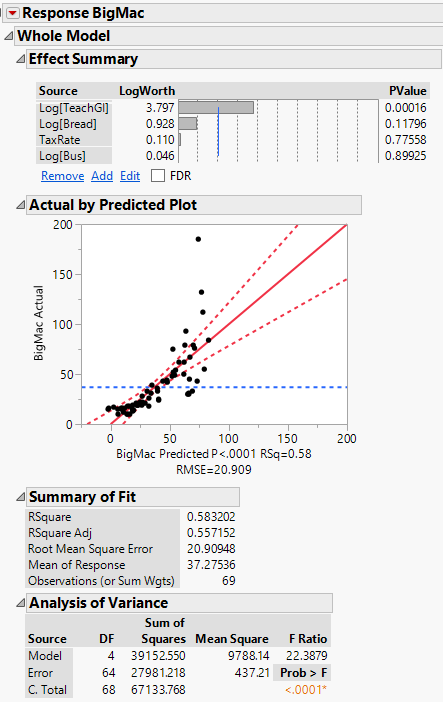
* Big Mac – minutes of labor required to purchase a Big Mac
* Bread – minutes of labor to purchase 1 kg of bread
* Rice – minutes of labor to purchase 1 kg of rice
* FoodIndex – index of the cost of food (Zurich = 100)
* Bus – cost in U.S. dollars for a one-way 10 km ticket
* Apt – normal rent in U.S. dollars for a 3 room apartment
* TeachGI – primary teacher’s gross income, 1000’s of U.S. dollars
* TeachNI – primary teacher’s net income, 1000’s of U.S. dollars
* TaxRate – tax rate paid by a primary teacher
* TeachHours – primary teacher’s hours of work per week.

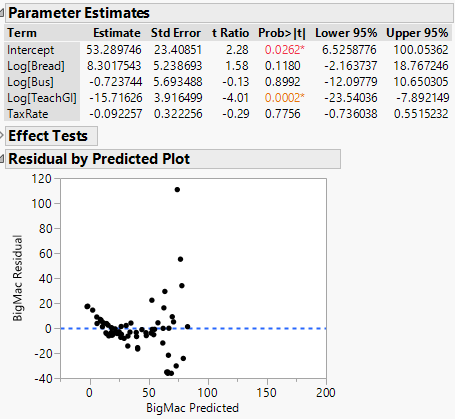
For this analysis we will consider modeling the response using potential predictors: .  


**Comments:**

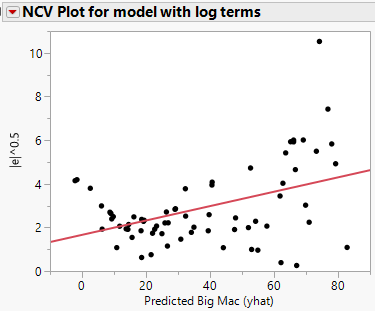
Taking the natural logarithm of the the scatterplot matrix becomes.  


Using terms we fit the model.

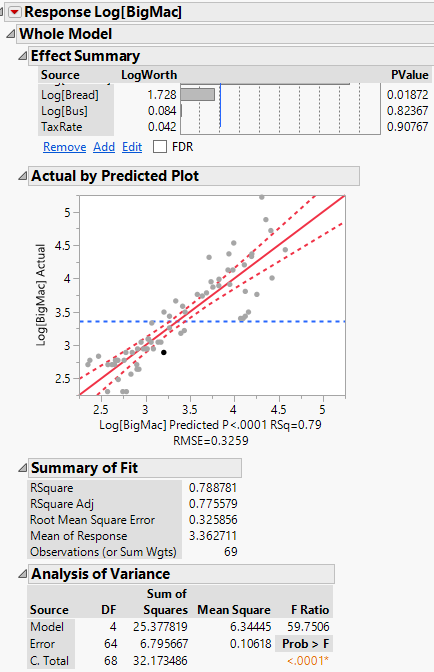
A summary of the model with a residual plot is shown below.  




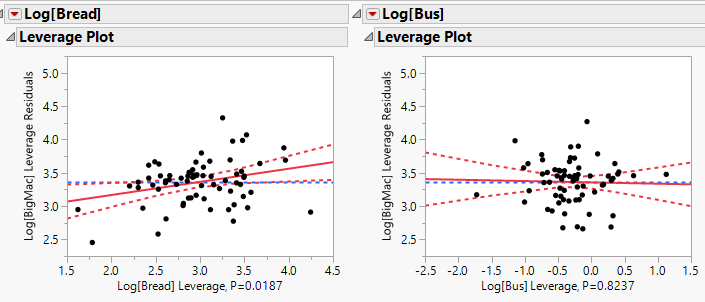
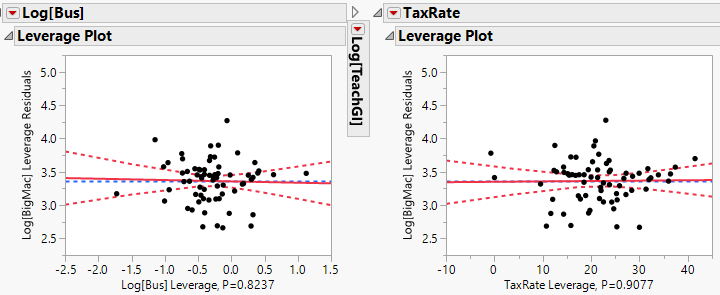
A nonconstant variance (NCV) plot is show below. Clearly there is evidence of increasing variation and curvature in both residual plots.

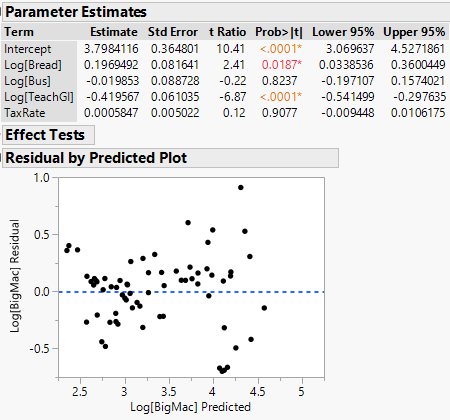


Given there is evidence of curvature and nonconstant variation we can consider transforming the response and we will first consider . A summary of the model using the log of the response gives.

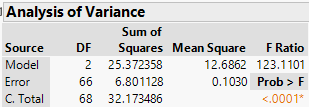
   
It appears that the terms and are not significant.

ANOVA Table for AH Model

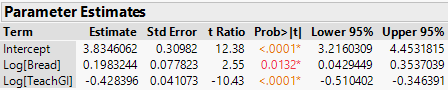


We can use the “Big F-Test” to compare the full model to the model dropping these terms.

ANOVA Table for NH Model  


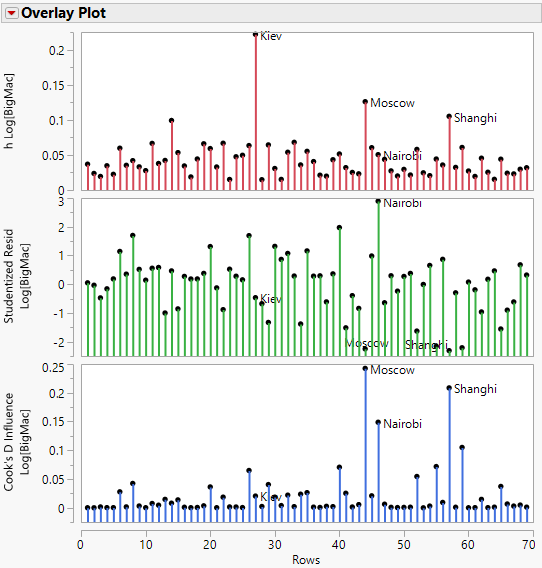
> pf(.0235,2,64,lower.tail=F)

[1] 0.9767824

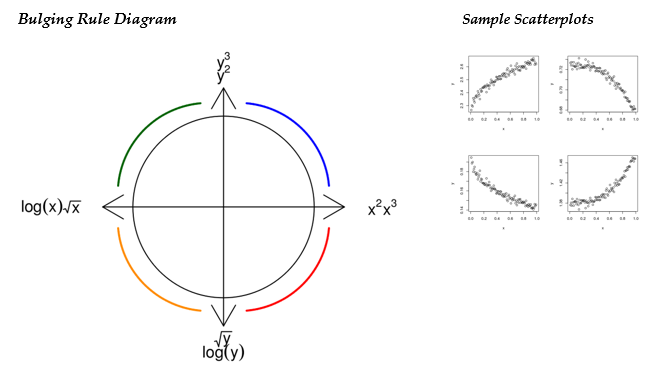
There is no evidence to support the inclusion of the deleted terms (p >> .05).   


The fitted model is:

Interpretation of the coefficients: Examining cases diagnostics:



**14.3 – Transformations to Linearity**

We have previously used the *Bulging Rule* to find suitable transformations of the response and/or the predictor to linearize the relationship in the case of simple linear regression.   


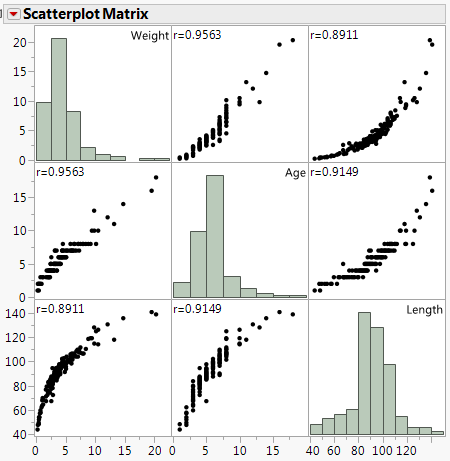
In multiple regression we can use an *inverse fitted value plot* to visualize the transformation to linearize the relationship, i.e.

The inverse fitted value plot is simply a plot of and a smooth added to this plot will give a visualization of . However for this procedure to work, i.e. for the plot of to give a visualization of the optimal transformation , we must have that for each nonconstant term in the model ( the following mean function holds,

This basically says the terms in the model have to be **linearly related** with each other. This will be satisfied if the terms follow a *multivariate normal distribution*.

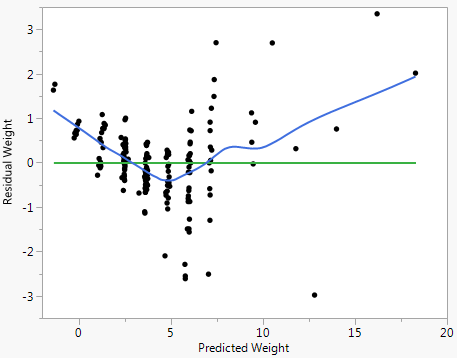
FYI – Some properties of the Multivariate Normal Distribution

1. All variables will have a normal distribution.
2. All pairs of variables have a bivariate normal distribution.
3. Any conditional expectation will be linear. For example if the terms/predictors and the response have a multivariate normal distribution then it is guaranteed that the and the variance function .

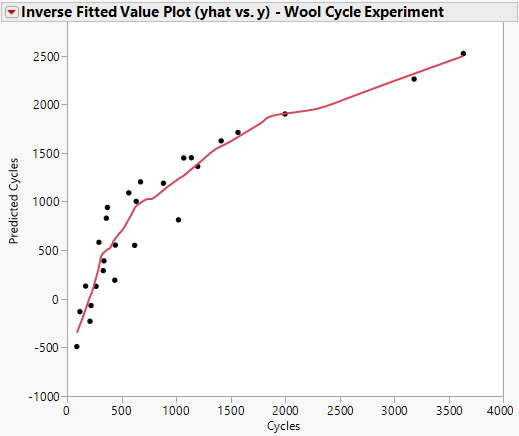
If the terms are simply the predictors themselves then each plot of one predictor/term vs. another in a scatterplot matrix should look linear, in fact each panel should look as if it is a sample from a bivariate normal distribution. If any panel in a scatterplot matrix has a clearly curved mean then the condition above may fail.   
  


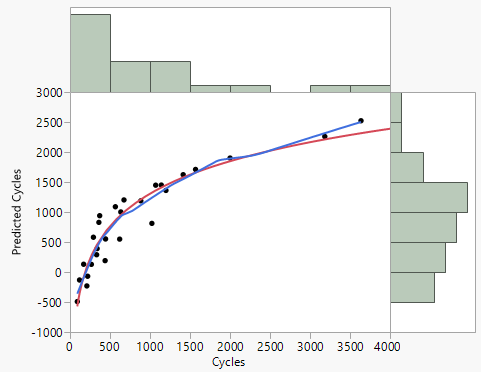
**Example 14.2:** The scatterplot matrix on the left is for the paddlefish data that we have examined previously. The response is the weight of the paddlefish (kg) and the two potential predictors are age (yrs.) and length (cm). Clearly the relationship between and is nonlinear, and the condition above may fail for these data. Thus an inverse fitted value plot obtained from fitting the model may not give a proper visualization of a linearizing transformation for the response.

Also joint distribution of is clearly not MVN, why?

We should not use an inverse fitted value plot to potentially transform the response in the regression of , even though the residuals from fitting this model show clear evidence of curvature.  
  
Plot of for   
  
Thus in order to address curvature we may have to make modifications to the mean function model (i.e. add nonlinear terms based upon the predictors to the model), however as the variance does appear to be constant we may still need to consider a response transformation for this reason.

**Example 14.3 – Wool Experiment**These data were examined previously in Example 11.2 in Section 11 of the notes. We initially fit the model . The residual plot () from fitting this model to these data is shown below and clearly shows curvature indicating the mean function above is not correct.  
To address this curvature we could add polynomial and interaction terms to the model or potentially transform the response. We will use the inverse fitted value plot visualize an appropriate response transformation.

Below is an inverse fitted value plot from fitting this model.   
  
We have seen previously that the model with as the response fit these data well, i.e. the model below was adequate:



Using **Fit Special** with log(x), i.e.log(y) in this case as y is on the horizontal axis, we see that the log transformation of the response is a reasonable approximation to .

??

**Example 14.4 – Volume of Black Cherry Trees**

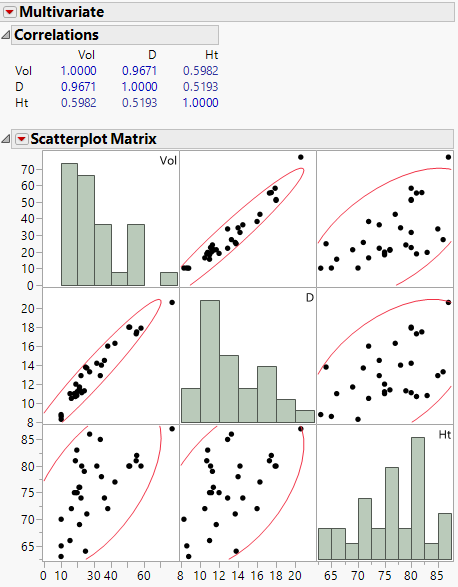
These record the girth in inches, height in feet and volume of timber in cubic feet of each of a sample of 31 felled black cherry trees in Allegheny National Forest, Pennsylvania. Note that girth is the diameter of the tree (in inches) measured at 4 ft. 6 in. above the ground.



*Black Cherry Tree*

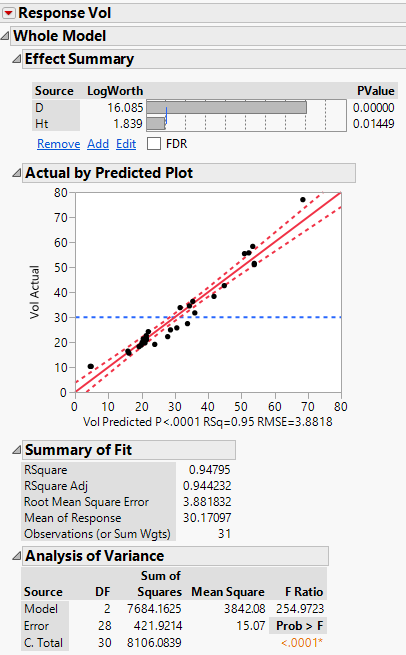
The variables in this dataset:

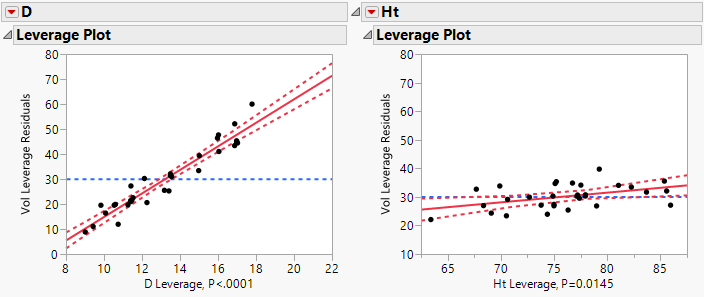
* Vol – volume of the black cherry tree (ft3)
* – girth/diameter of tree (in.)
* – height of tree (ft.)

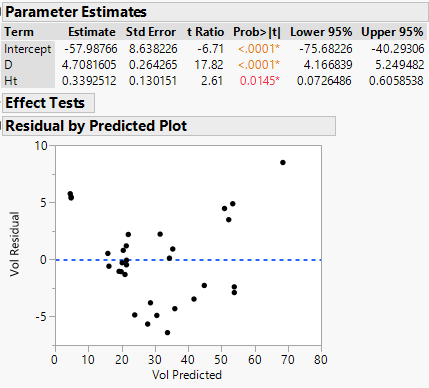
Below is a scatterplot matrix of these data.  


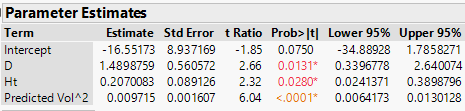
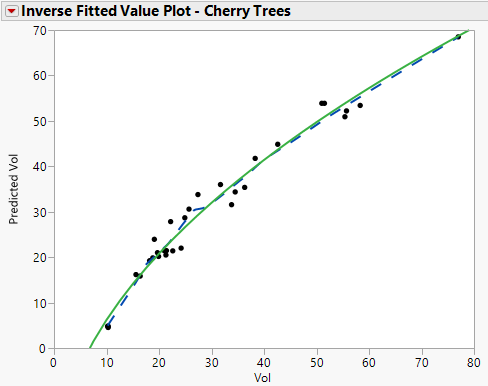
We first fit the model with terms equal to the predictors diameter (D) and height (Ht), and

Geometry of a Tree?

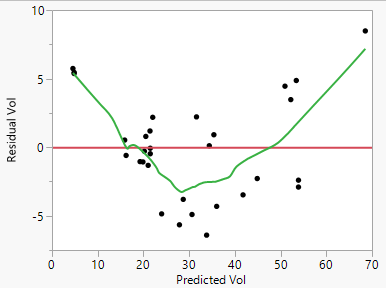




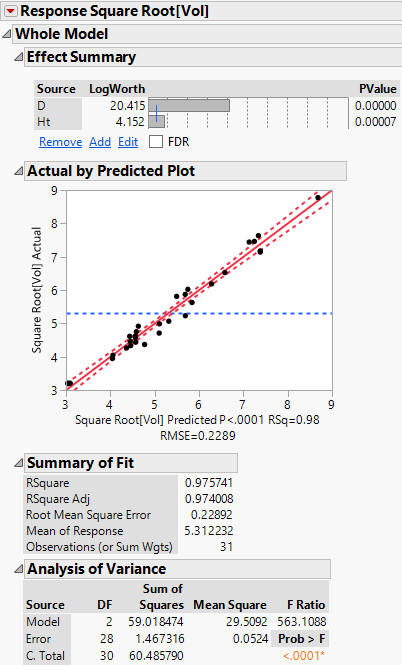


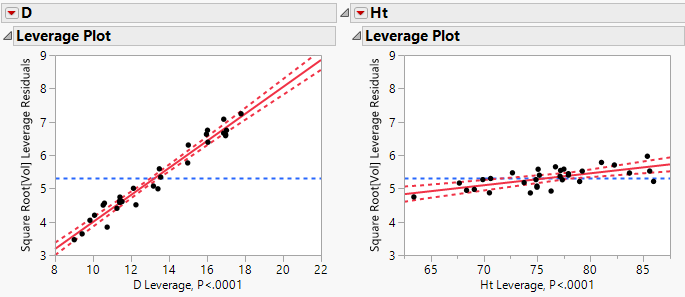
Test for Curvature – Tukey’s Test for Nonadditivity  
  
We have clear evidence of curvature based on the test and the residual plot shown above. We will again use the inverse fitted value plot see if a response transformation is suggested.  


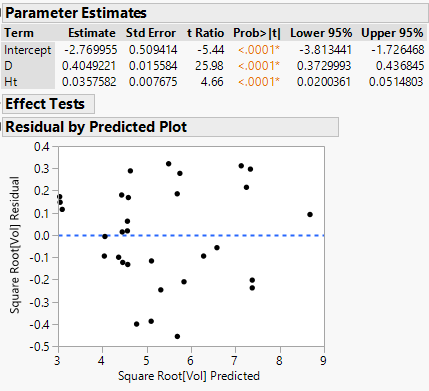
The kernel smoother (--- line) and the Fit Special with (x-axis) agrees with the smooth perfectly (solid line). Thus we will fit the model using the as the response.



Summary of Fitted Model -







Interpretation of the estimated coefficients when the square root response transformation is used is not possible. The fact that both estimated coefficients are positive indicates that as each increases (probably not independently) the square root volume, and hence volume, increases. However this model appears to be able to predict the volume accurately using these two easy obtain measurements.

In regression, the goal is to either (1) understand how the terms in the model are related to the response by interpreting the estimated coefficients or (2) predict the response accurately. We will discuss prediction accuracy when discuss strategies for model selection and development.

**14.4 – Transformations to Normality**

As we saw in Section 4 – Bivariate Normality and Regression normality is desirable distributional property when conducting regression analysis. Tukey’s Ladder of Powers is useful when transforming a numeric variable using the power transformation family to have an approximately normal distribution in the transformed scale.

There is a numerical method for finding the “optimal” transformation to achieve approximate normality called the *Box-Cox Method*. It can be used find the transformation to achieve normality in the univariate and the multivariate case. In situations where the we have evidence of non-normality in the residuals from a fitted model , which will usually be accompanied by curvature and/or nonconstant variation also, we generally start by using the Box-Cox method to find an optimal response transformation.

**Box-Cox Transformation Family & Method**

The Box-Cox method chooses the optimal by maximizing ,

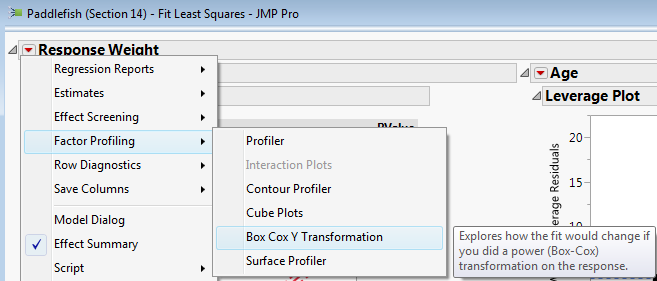
(or minimizing )

Here is the residual sum of squares from fitting the model with as the response. The maximization is usually done by evaluating over a sequence of values from - 2 to 2.

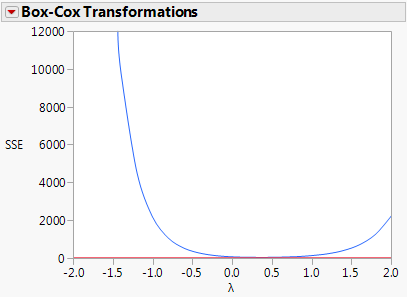
We will consider the use of the Box-Cox transformation procedure for each of the examples above and compare contrast the Box-Cox transformations to those chosen using the methods above.

**Example 14.2 – Weight of Paddlefish**

To obtain the Box-Cox transformation for the response in a multiple regression select **Factor Profiling > Box-Cox Y Transformation** as shown below.



The results for the Box-Cox procedure for transforming are shown below.



JMP plots the criterion to be **minimizmed** rather than maximized, thus we wish to choose based on this plot with the smallest SSE.

Here the minimizing is difficult to discern but it appears to be around suggesting a square root transformation for the weight of the paddlefish.

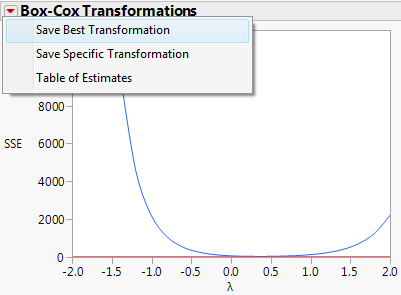
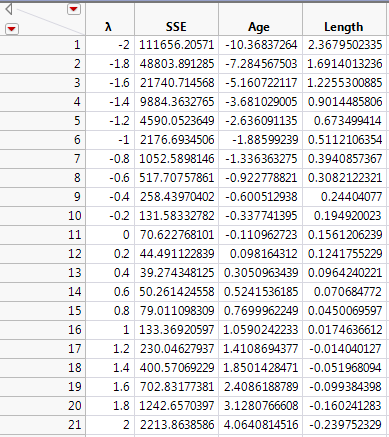
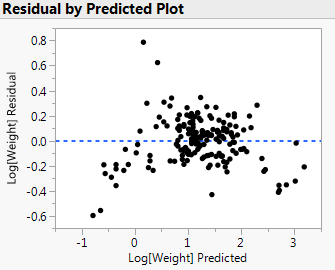
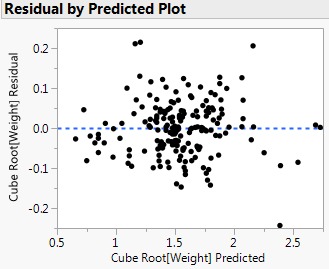
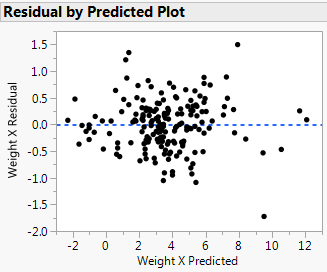
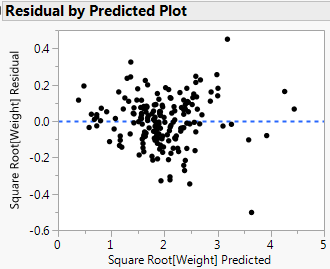
We can have JMP save the “optimal” (**Save Best Transformation**) or save the results for a sequence of the values to a table (**Table of Estimates**) as shown below.  


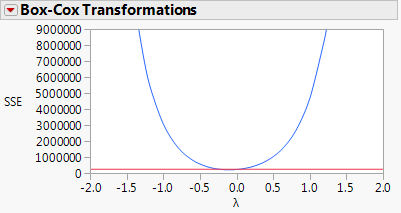
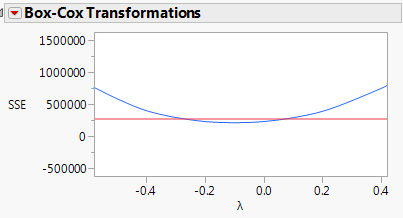
Table of Estimates for   


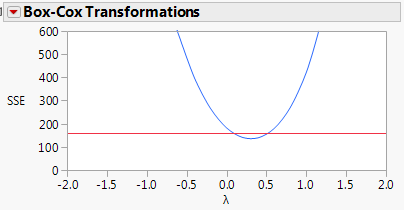
Here the optimal choice for chosen using the Box-Cox procedure is .   
  
However it appears using any power between may be reasonable as well.

Below are the residual plots from using

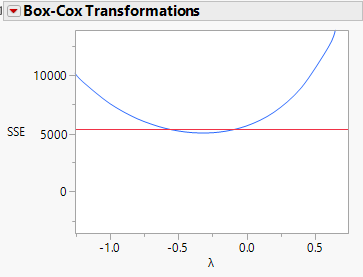
Which transformation should we use?

**Example 14.3 – Wool Data**Original Box-Cox Plot Zoomed in region containing optimal  Clearly , i.e. is nearly optimal and is contained within the CI for .

**Example 14.4 – Black Cherry Trees**

The Box-Cox method suggests is optimal but (cube root) and (square root) are within the CI for

**Example 14.1 – Big Mac Data – cautionary note about the Box-Cox Method**



The optimal transformation chosen by the Box-Cox Method is .

This seems like a strange choice but we need to realize that the Box-Cox Method is very susceptible to outliers and influential points. These data have several observations (i.e. cities) that seem to fit this description.

The Box-Cox Method is not a “silver bullet”, it can be useful when attempting to find a response transformation but the “optimal” transformations suggested by it should not be blindly trusted. I generally will run it and look at the range of values within the CI. If a common transformation like square root or log are contained in the CI for , I will opt for those over the optimal value. Also remember that transformations other than the logarithm can make interpretation of estimated coefficients difficult, if not impossible. Thus is parameter interpretation is of primary interest we should generally avoid response transformations unless they are absolutely necessary.

**14.5 – Summary of Response Transformations**

Response transformations can stabilize variance, address curvature, and improve normality in a regression setting. In this section we have examined some guidelines, plots, and methods for finding these transformations. Here are some additional general guidelines to consider when transforming the response and, though not the focus of this section, the predictors as well.

**The Log Rule –** If the values of a variable range over more than one order of magnitude (i.e. max/min > 10) and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful in addressing model deficiencies.

**The Range Rule –** If the range of a variable is considerably less than one order of magnitude, then any transformation of that variable is unlikely to be helpful.

**Ladder of Powers –** Use Tukey’s Ladder of Powers to transform variables to approximate normality.

**Bulging Rule –** Use the Bulging Rule straighten scatterplots, particularly if there are nonlinear relationships amongst the predictors.

**Inverse Fitted Value Plot –** If the relationships amongst the predictors are approximately linear, draw the inverse fitted value plot (). If this plot shows a clear nonlinear trend then the response should be transformed to match the nonlinear trend. The transformation could be selected by using a smoother. If there is no clear nonlinear trend, transformation of the response is unlikely to be helpful.

**Box-Cox Method –** This method can be used to select a transformation to normality. This method is susceptible to the effects of outliers, so blindly trusting the transformation suggested is not a good idea. Also the method will give a range of potential transformations, so choosing a common transformation within the range suggested is generally preferred.

In the next section we will discuss predictor transformations (i.e. creation of terms) to address model deficiencies.