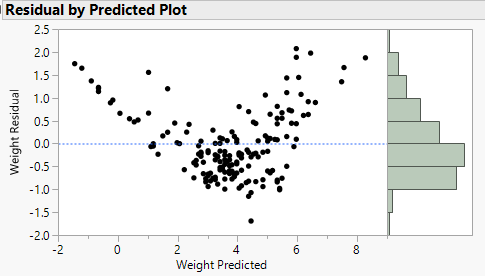
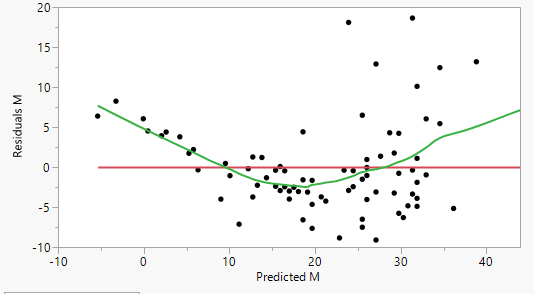
**12.0 – Review of Residual Plots**The form of the multiple regression model is given by

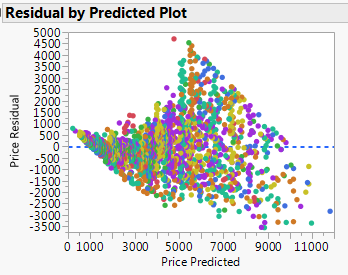
and typically we assume .

Graphically we generally check these assumptions by plotting the residuals . For checking the adequacy of both the mean function and the variance function we can plot the residuals vs. fitted values .

Paddlefish Data: Mussels Data:

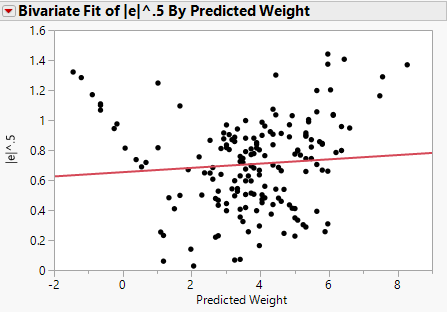
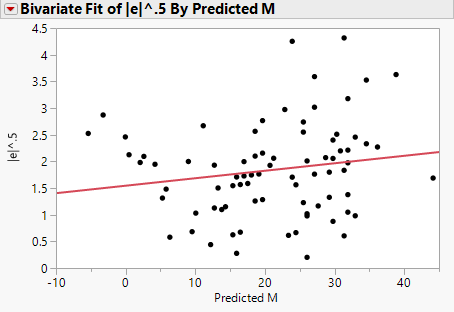
 

Diamond Data:

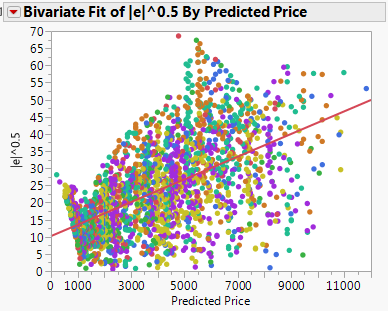
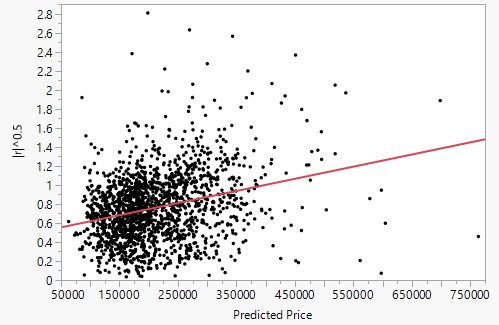


These examples have residual plots that suggest obvious curvature, i.e. the linear mean function is inadequate, and there is also evidence of nonconstant variation as well. The curvature we see in these plots indicates is not constant which implies the assumed mean function does not adequately represent the relationship between the response the set of predictors.

To better examine the variance assumption we can plot . The residuals plots for both models are shown below. The curvature present in both models makes these plots less useful for assessing nonconstant variance. Ideally the mean function is correct when assessing the variance assumption.

Paddlefish Data: Mussels Data:  

The NCV plot using residuals & fitted values from the model above clearly shows increasing variation. As a final example of the NCV plot consider again the Saratoga, NY homes data. Below is a plot of for the model fit using terms from all available predictors, clearly there is evidence of nonconstant variance.

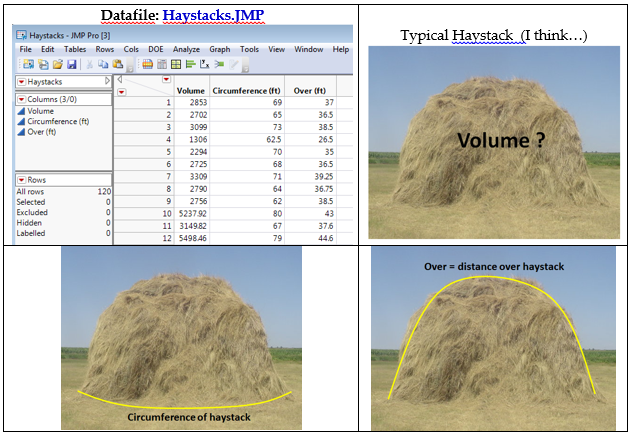
**12.1 – Testing for Curvature**

Suppose we plot the residuals versus a linear combination of the u-terms in our regression model **,**

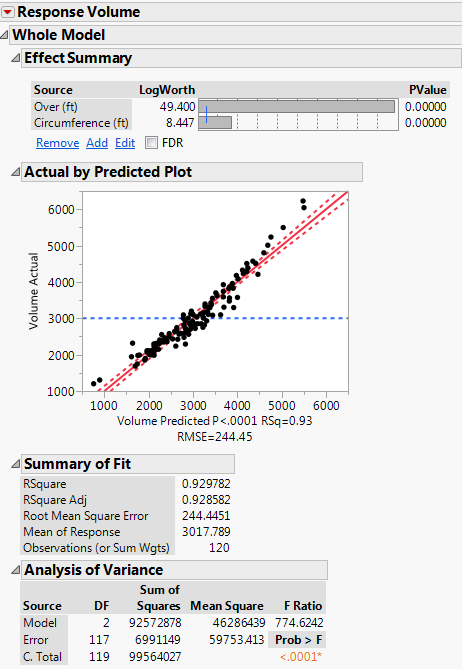
and judge through visual inspection that the residual mean function indicates curvature. The usual residual plot ( vs. ) we have previously examined is a special case of this situation where , i.e. the fitted values. We could also plot the residuals ( vs. individual terms in the model () or any linear combination of them defined by .

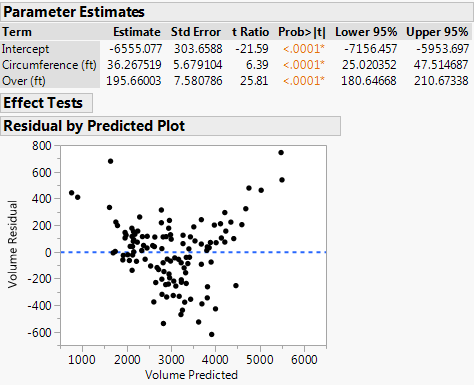
**Example 12.1 – Haystack Volumes Datafile: Haystacks.JMP**

We examined these data earlier in the course, but we know consider the multiple regression of haystack volume on both circumference and the over dimension of the haystacks.

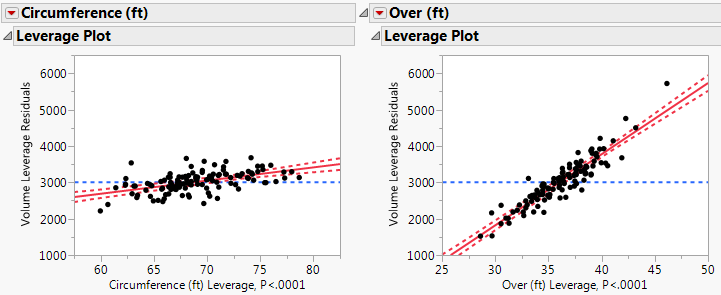


We consider the model with constant variance .



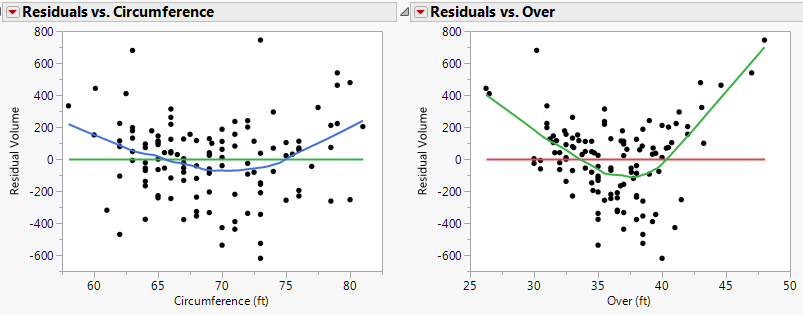


Does this residual plot suggest model violations?



Which predictor/term, Circumference (C) or Over (O), has the strongest adjusted relationship with Volume?

We could also plot the residuals vs. each term/predictor in the model.



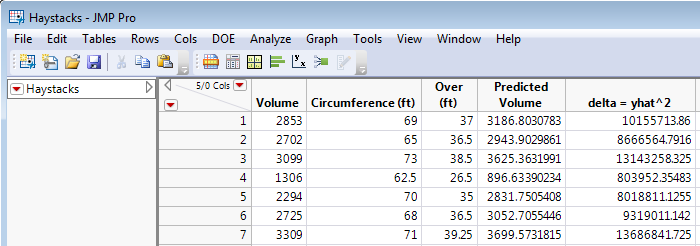
Clearly there is curvature present in the plots of the residuals vs. each of the individual terms/predictors in the model. While it is clear this model is deficient and the curvature needs to be addressed, we will now consider statistical tests that can be used to identify situations where these is significant curvature present.

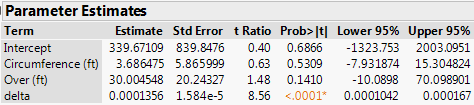
**Test for Curvature**

To test for curvature we fit the model below:

and test whether these is evidence that the parameter . The default choice for is to use the fitted values from the original, which is the model above without the squared term. Using the is called *Tukey’s Test for Nonadditivity*.

Conducting this test for the haystack data requires that we first save the Predicted Values (i.e. from the model above and then use them as a term in the multiple regression model that includes both C and O.

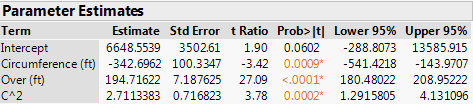


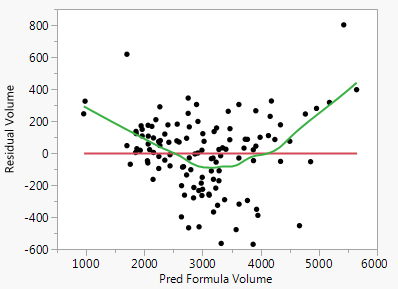
**Tukey’s Test for Nonadditivity for Haystack Data**

Conclusion:

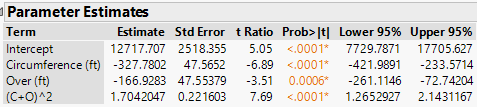
What about other choices of ? Other choices for these data would be:

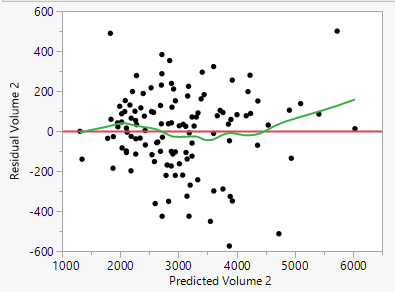
There are certainly infinitely many other choices.

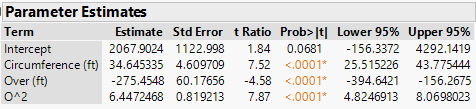
**Adding Squared Circumference**   


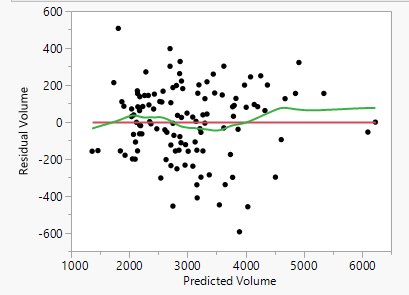
Plot of   


**Adding**



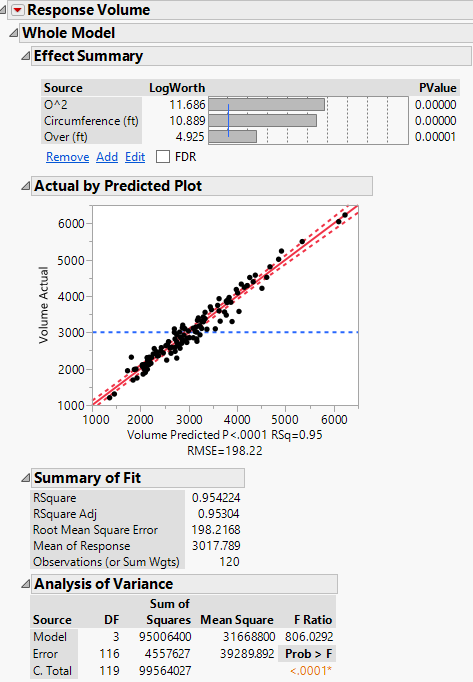
Plot of   


**Adding Squared Over**   


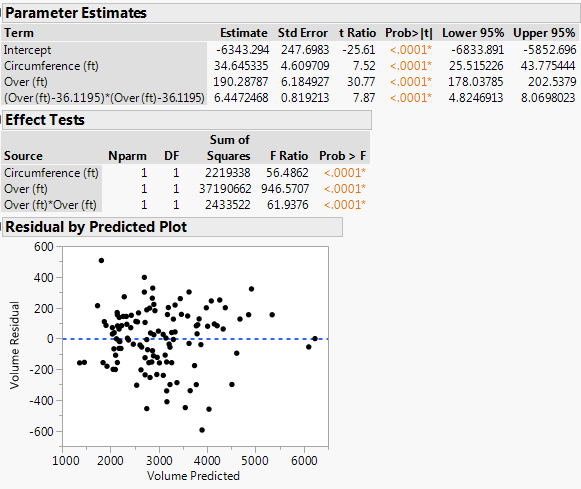
Plot of   


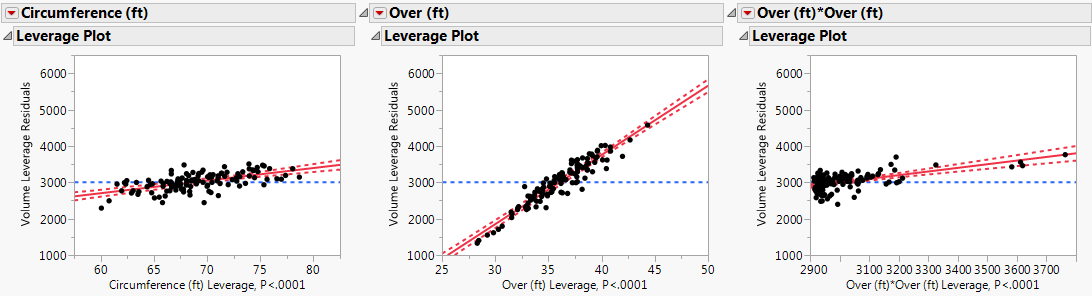
These tests suggest the curvature seen in the original model can be addressed by adding the polynomial term to the base model. Thus we find a “final” adequate model for the volume of the haystack is given by

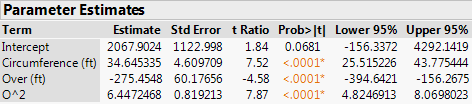
A complete summary of this model is on the following page.



You can avoid the mean centering on the squared term for the Over dimension by squaring Over in the data table and adding it as a term in the model. The parameter estimates using this approach are the same as those for the curvature test above and are presented below. The rest of the regression output is identical except for the scaling in the AVP’s. My recommendation is to allow JMP to do the mean centering. We will see why in the next section.



AVP’s  




**12.2 – Testing for Nonconstant Variance**

The NCV plot is a useful graphical tool for assessing constant the constant variance assumption. To further study the variance function it is useful to have a functional form for the variance assumption. Let consist of terms constructed like the -terms in the mean function. The components of are functions of the predictors . Like the -terms, the -terms can be the predictors themselves, some other function of them, or may be the same as some of the -terms in the mean function, i.e. or .

The model for the variance function is

with the parameter being the variance of when **,** and is a vector of parameters like the ’s in the mean function. The exponential function ( in this formulation of the variance function guarantees the variance is positive for all values of . The constant variance function, which is what we typically assume to be the case for our model, is also a special case of the model above when .

Taking the logarithm of the variance function above gives the following:

The role of the intercept is taken by so the -terms need not contain an intercept term, i.e. a column of ones.

Often times the variance is a function of the mean; generally, as the mean increases, so will the variance. This is an important special case where the -terms are the same the -terms in the mean function, i.e. . This gives , so the model above becomes,

The variance is constant when . To test for nonconstant variance using this model above we need to test,

To test these hypotheses we use the ***score test*** which is outlined below.

**Score Test Procedure**

1. Fit the OLS regression of on using the residuals and fitted values from the model under consideration, .
2. Compute the test statistic,
3. Find *p-value* =   
   In R: > pchisq(test statistic value, df = 1, lower.tail = F)

In JMP: Chi-square Probability Calculator.JMP linked below this section of   
 notes on my website.

1. If *p-value < .05* we reject NH and conclude the depends on .

If we wish to use a more general form for where again the are terms created from the predictors, then the numerator in the score test statistic is found by regressing the squared residuals on and the p-value is found using a chi-squared distribution with , i.e. .

**Example 12.3 – Bank Transactions**

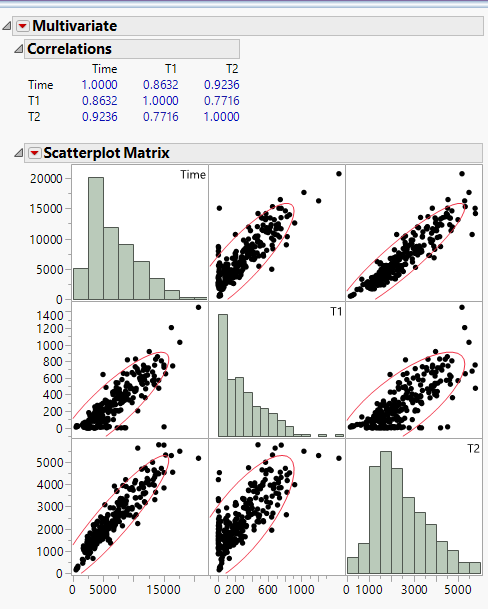
These data come from a study of total transaction time coming from two types of transactions labeled and .

The variables in the data file Transactions.JMP contains the following variables:

* – total time for transactions of types and .
* total number of transactions of type 1.
* total number of transactions of type 2.

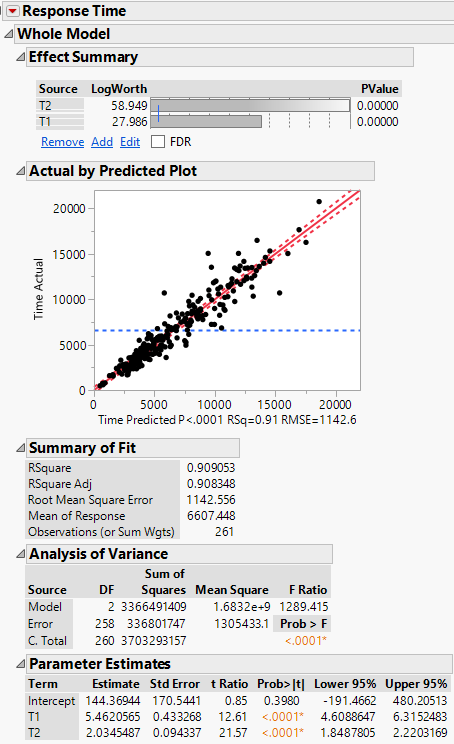
We first consider fitting the model,

Below are a correlation matrix and scatterplot matrix for these data.

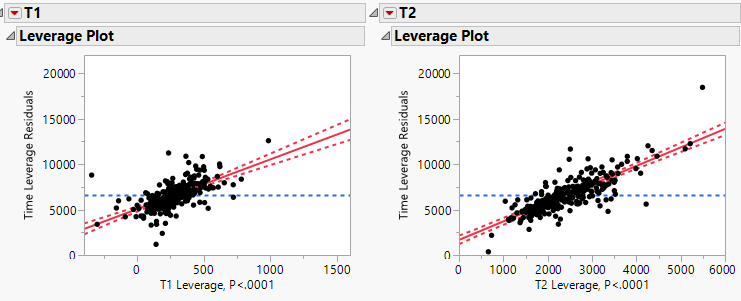


A summary of the fitted model is shown on the next page.

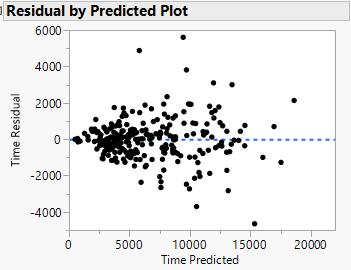
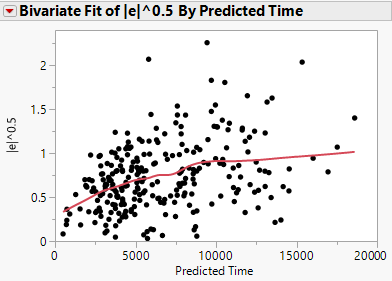
Summary of Model - and



AVP’s for and



Residual Plots - and

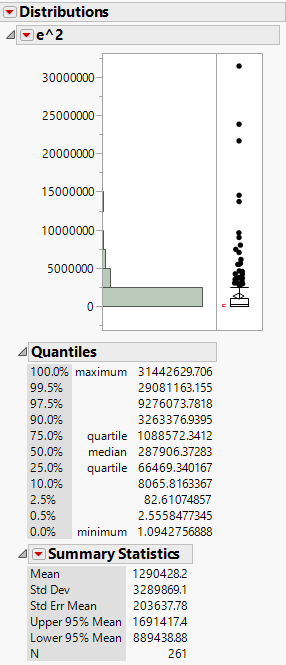
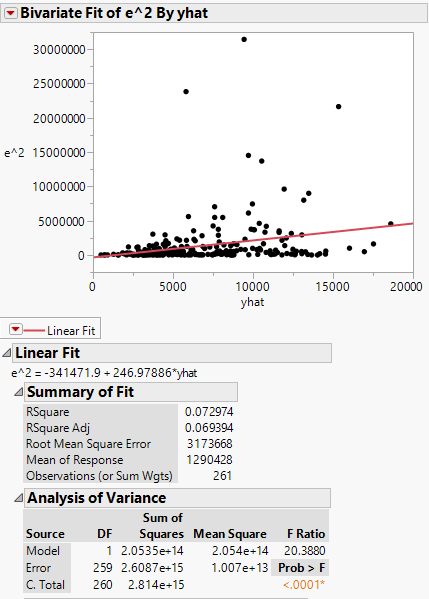
 

There is clear visual evidence of nonconstant variation, though the mean function seems adequate.

Is finding that is not constant important? This depends on the goal of the analysis. If the only goal is estimate the number of minutes on average each type of transaction takes then this model is fine. Our estimates from the analysis above is each Type 1 transaction takes 5.46 minutes and each Type 2 transaction takes 2.03 minutes. However, any computation that depends on the estimated variance such as standard errors of the estimated coefficients, confidence/prediction standard errors, etc. will be wrong. For example the standard error the mean total time spent will be too large for smaller banks and too small for larger ones. Also confidence intervals for the point estimates for the length time for each transaction type will wrong because the standard errors depend on

We will now conduct score tests using different -terms.

1. Assume and we will test vs. .



Score Test Statistic =

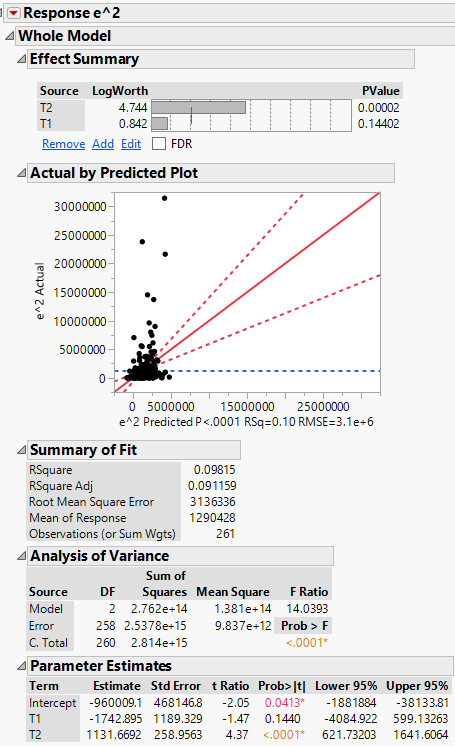
Using R as chi-square calculator

> pchisq(61.66,df=1,lower.tail=F)

[1] 4.081888e-15

Thus p < .0001 therefore we reject NH and conclude the variance function is significantly related to the mean function.

1. Assume



Score Test Statistic =

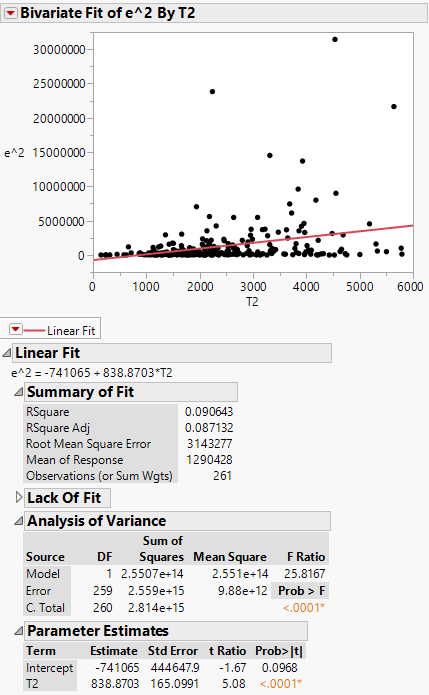
Using R as chi-square calculator

> pchisq(82.93,df=2,lower.tail=F)

[[1] 9.817012e-19

Thus p < .0001 therefore we reject NH and conclude the variance function is significantly related to the terms used in the model for the mean function. We also note that is not significant, thus we might want to test if the nonconstant variation is due primarily to Type 2 () transactions.

1. Assume



Score Test Statistic =

Using R as chi-square calculator

> pchisq(76.58,df=1,lower.tail=F)

[1] 2.114705e-18

Thus p < .0001 therefore we reject NH and conclude the variance function is significantly related to the number of Type 2 transactions, .

In Section 17 of the course notes we will examine Weighted Least Square (WLS) regression as means of modeling the nonconstant variance. The results from these test results can provide some guidance on how to choose the form of the weights.

**Example 12.3 – Green Mussels from Marlborough Sound, New Zealand**

**Datafile: Mussels NZ.JMP**

These data were collected on green horse mussels sampled from the Marlborough Sounds off the coast of New Zealand. These data were collected as part of a larger ecological study of mussels in this region.

Variables in the dataset:

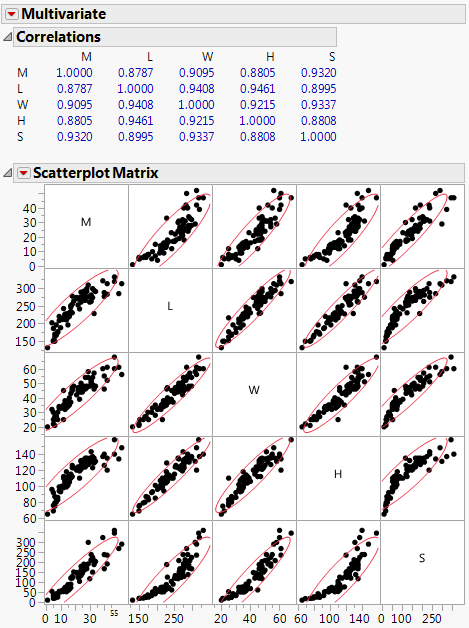
* M = mass of the mussels muscle (g), i.e. the edible part of the mussel
* L = length of the mussel’s shell (mm)



* W = width of the mussel’s shell (mm)
* H = height of the mussel’s shell (mm)
* S = mass of the mussel’s shell (g)

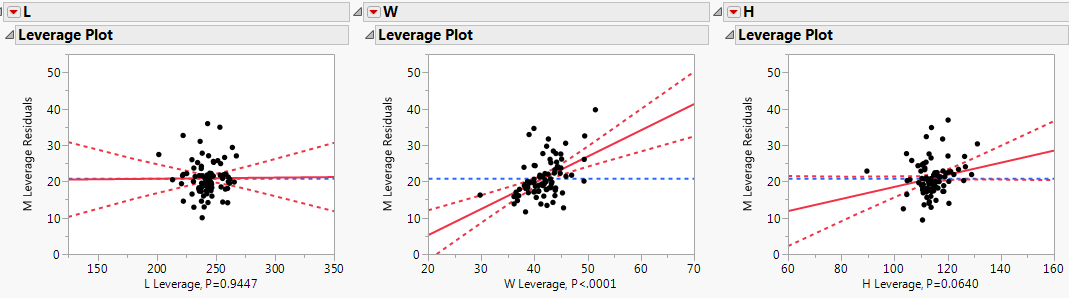
The goal of this regression analysis is to develop a model for the mean muscle mass (M) using the shell-based measurements.

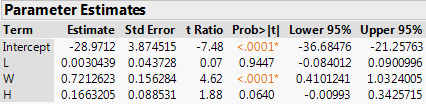
We begin by examining a scatterplot matrix of these data.

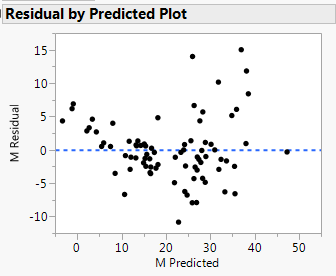


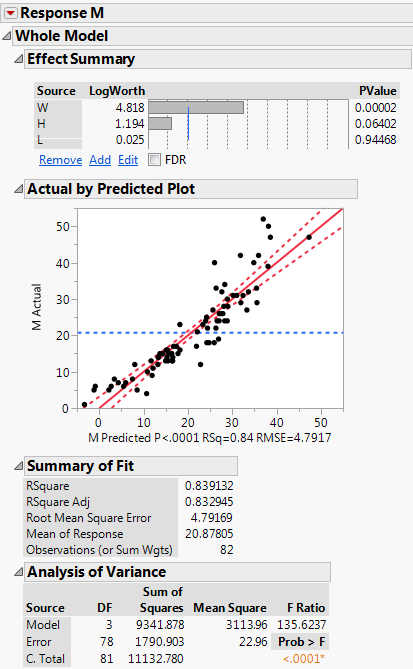
We begin we consider the MLR model using the shell-dimensions only as the terms in the model,

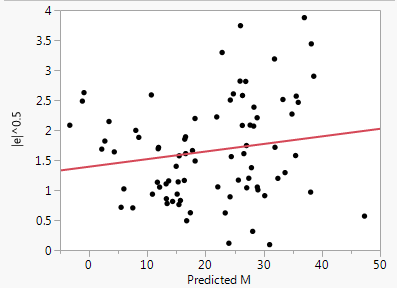
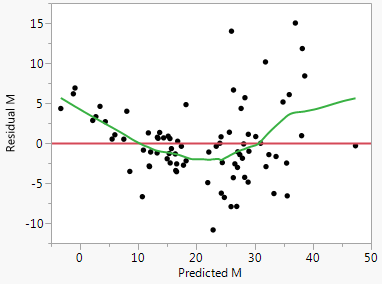
A summary of the fitted model is shown below.



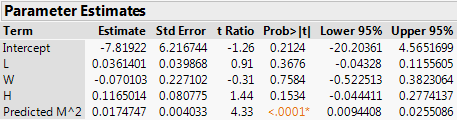




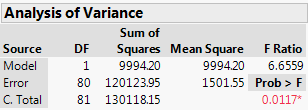
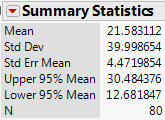


Tukey’s Test for Nonadditivity



Score Test for Nonconstant Variance -

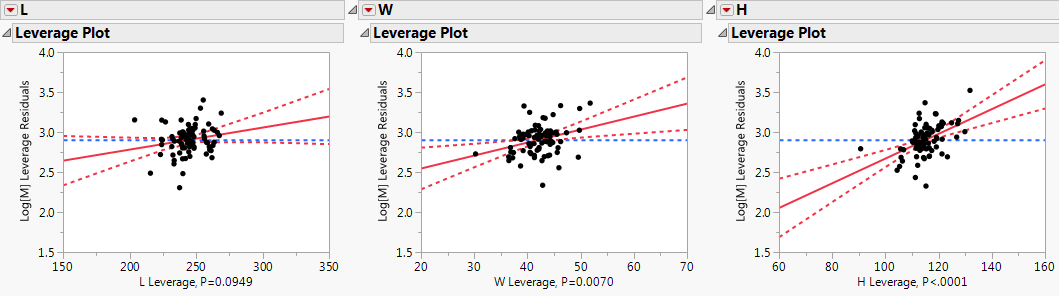
 

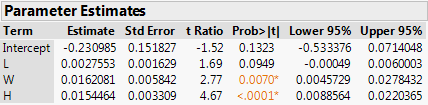
> pchisq(11.55,df=1,lower.tail=F)

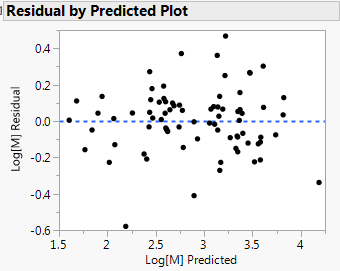
[1] 0.0006774926

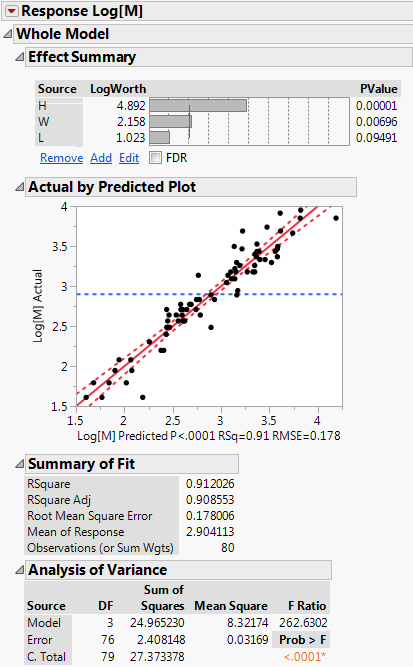
As there is evidence of both curvature and nonconstant variance, from both visual inspection and tests, starting with a response transformation such as is a better approach than adding terms to the model to address the curvature (e.g. ).

Fitting the model with the mussel muscle mass log-transformed gives the following.









Testing for curvature and nonconstant variance using the same test methods above confirms that both deficiencies have addressed by this model (p > 0.05).