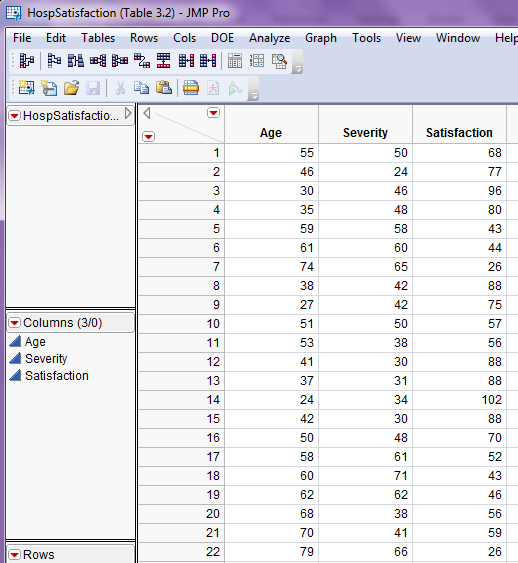
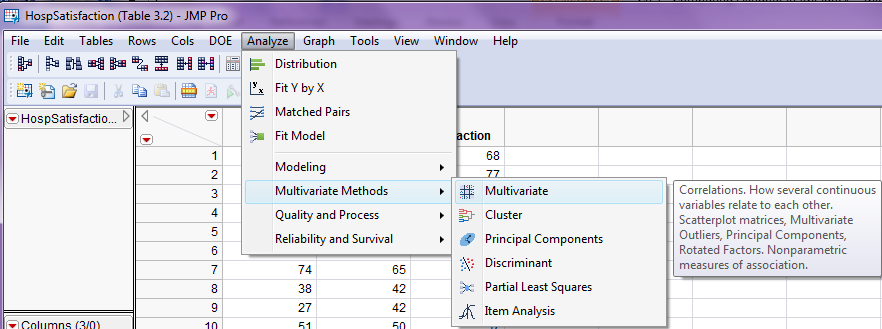
**Ch 3. – Regression Analysis – Computing in JMP**

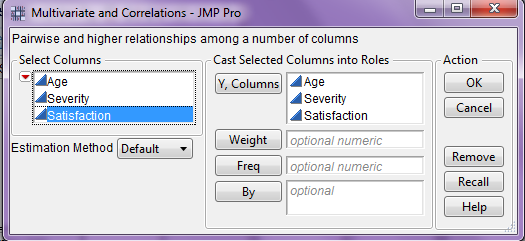
**Example 1: Patient Satisfaction Survey** (pgs. 79 – 83)

**HospSatisfaction (Table 3.2).JMP**

Before fitting any regression model it is good idea to plot the response variable and the potential predictors in a scatterplot matrix. To do this in JMP select **Analyze > Multivariate Methods > Multivariate** as shown below.



The completed dialog box is shown below.

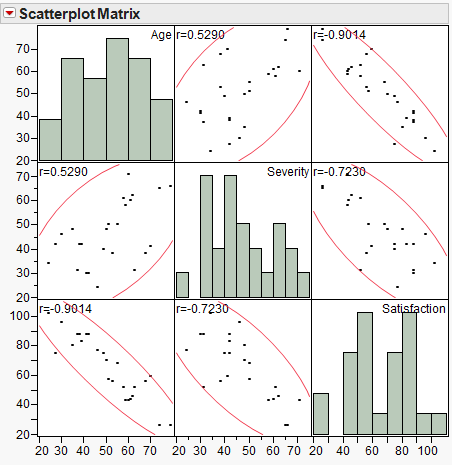


Comments:

The output below gives a correlation matrix for Age, Severity, and Satisfaction. It gives all pairwise correlations between these variables. The diagonals are all 1.000 because the correlation of any variable with itself is 1.



The scatterplot matrix for these variables is shown on the next page. I have added the following options to the display: **Show Correlations** and **Horizontal** from the **Show Histograms** option.



We can see both age and severity of illness are negatively correlated with the response satisfaction, and are positively correlated with each other as would be expected.

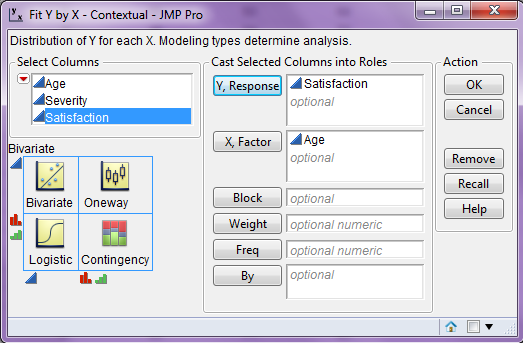
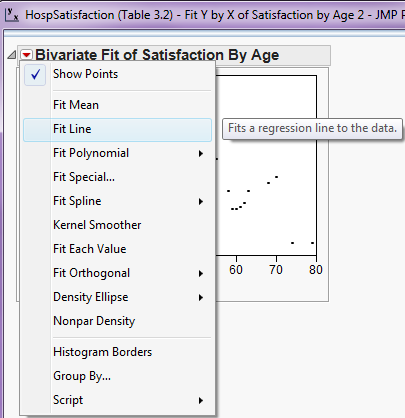
We first examine the simple linear regression model for satisfaction using patient age as the sole predictor, i.e. we will assume the following holds for the mean satisfaction as a function of patient age: .

Thus we will fit the following model to the observed data from the patient survey:

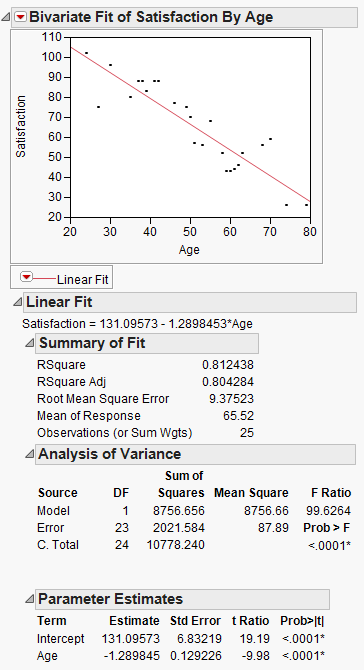
where,

We can do this two ways in JMP, using either **Analyze > Fit Y by X** or **Analyze > Fit Model**.

**1)** **Analyze >** **Fit Y by X** approach

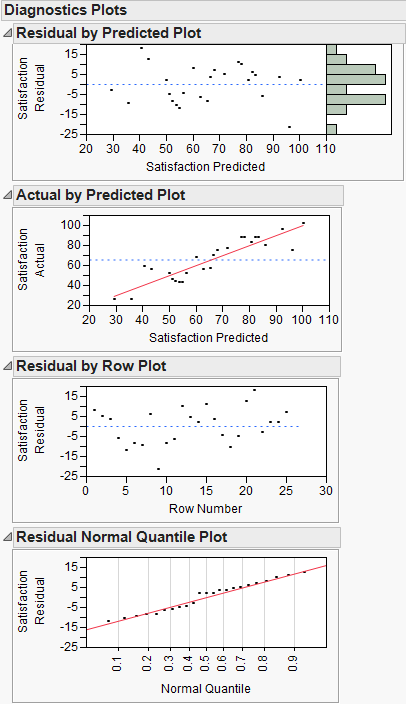
 

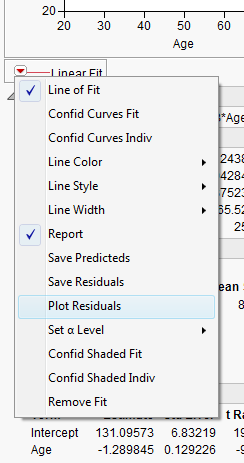
The output from fitting the simple linear regression model to these data is shown below:



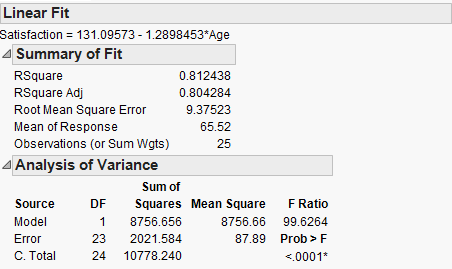
Comments:

**Examining residuals** (

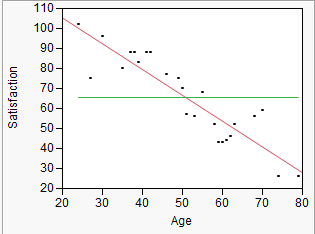




To obtain the plots of the residuals shown to the left, select **Plot Residuals** from the **Linear Fit** pull-down menu.

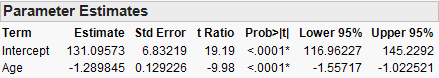


**R-square =** .8124 or 81.24% of the variation in the patient satisfaction responses can be   
 explained by the simple linear regression on patients age.

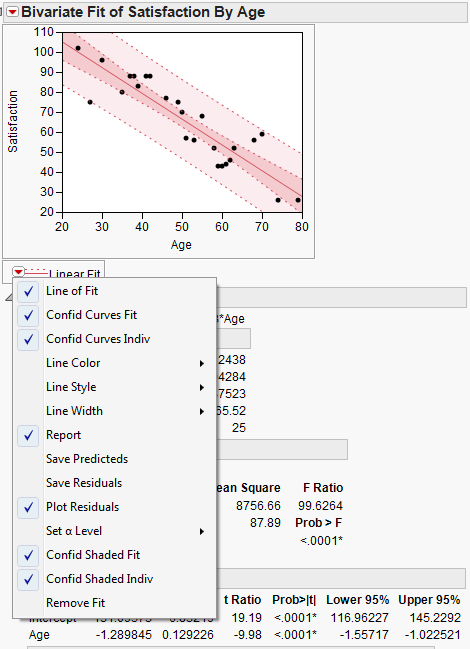


**Mean Square Error (MSE)**

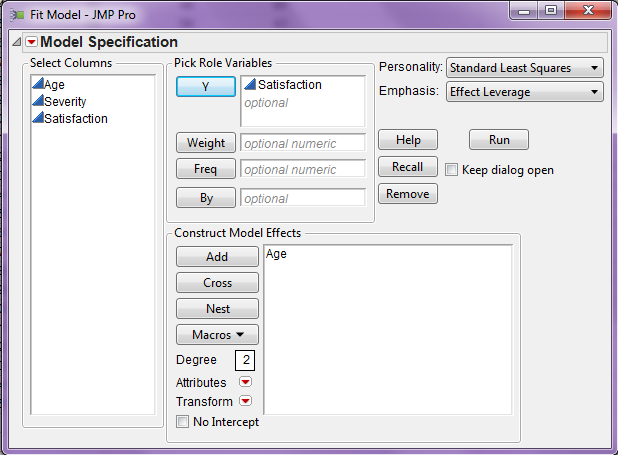
**Testing Parameters and CI’s for Parameters:**



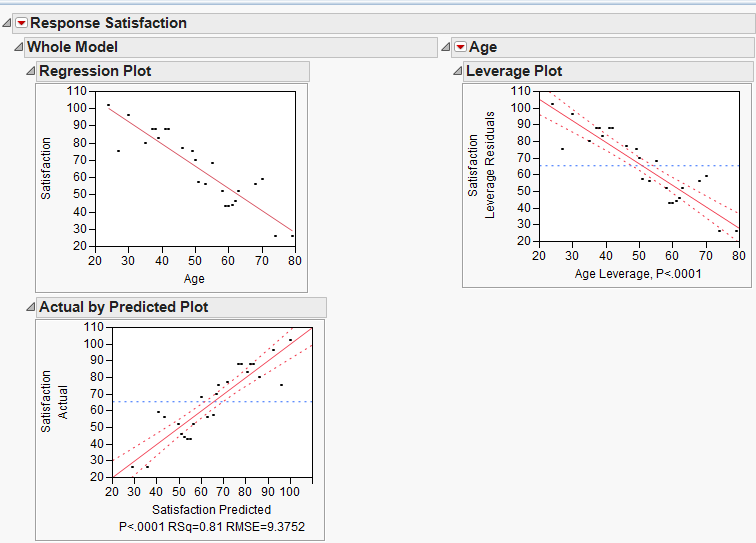
**Confidence Intervals for the Mean of Y and Prediction of a New Y**

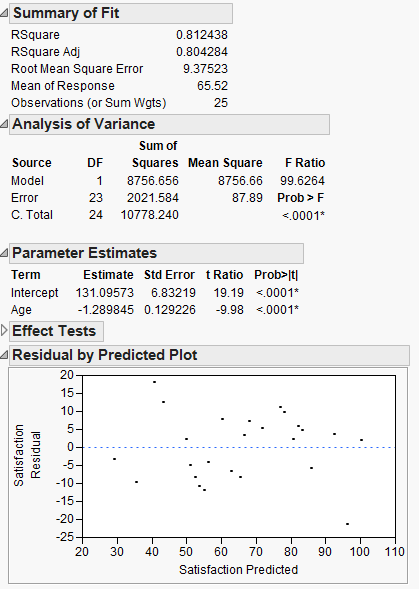


**2) Analyze > Fit Model** approach – more information and allows for multiple   
 regression

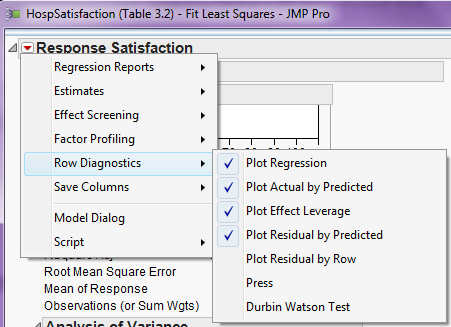


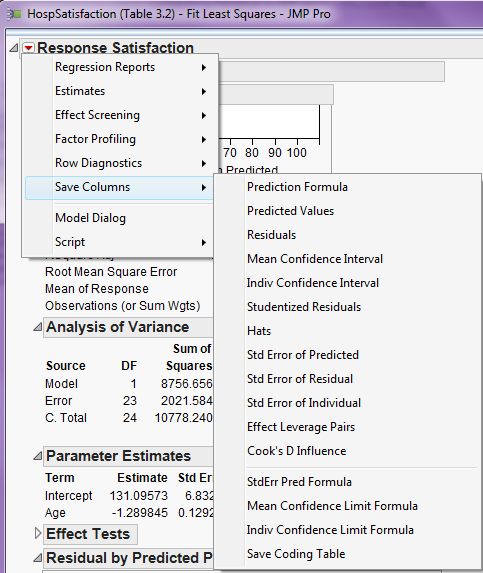
Basic output from **Fit Model** approach is shown below.



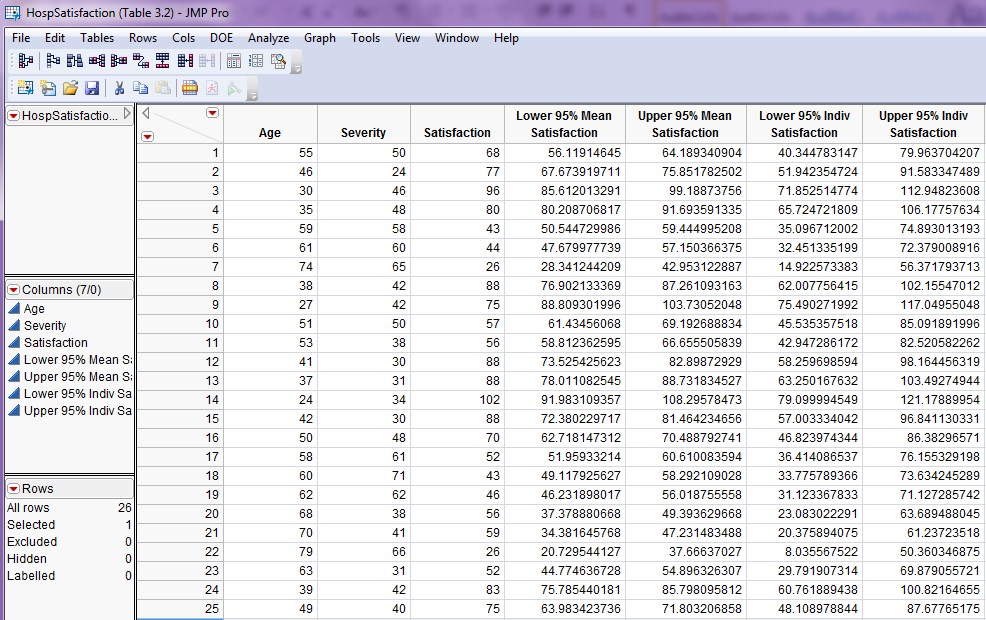


**Additional Plots and Summary Statistics**





**Confidence Intervals and Prediction Intervals**



**For adding a new patient who is 65 years of age gives…**



**Leverage and Influence – Basic Idea**

Basic idea:

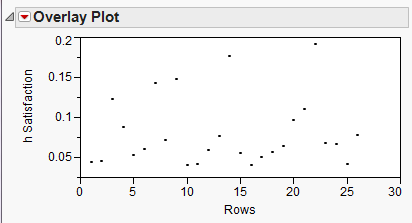
**Guidelines for Leverage and Cook’s Distance**

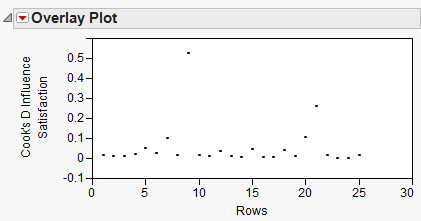
High leverage:

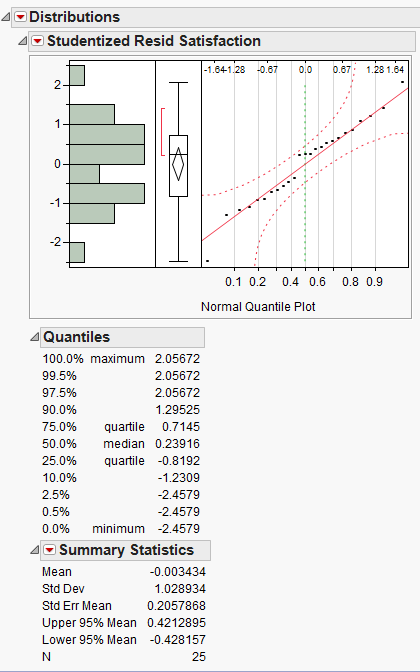
Cook’s Distance:

What makes Cook’s Distance large?

**Plotting Studentized Residuals, Leverage and Cook’s Distance**



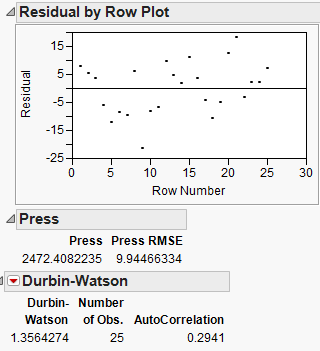




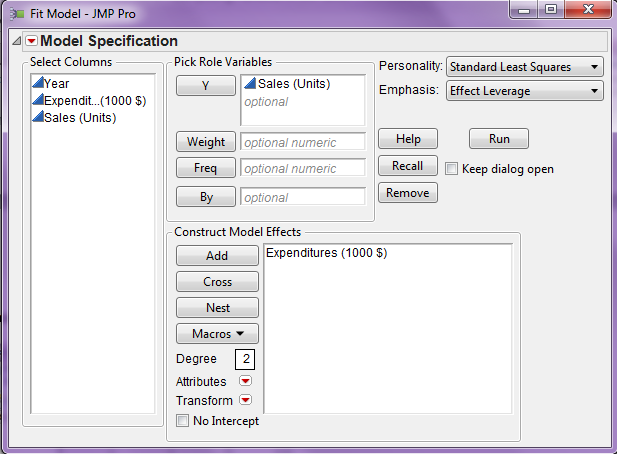
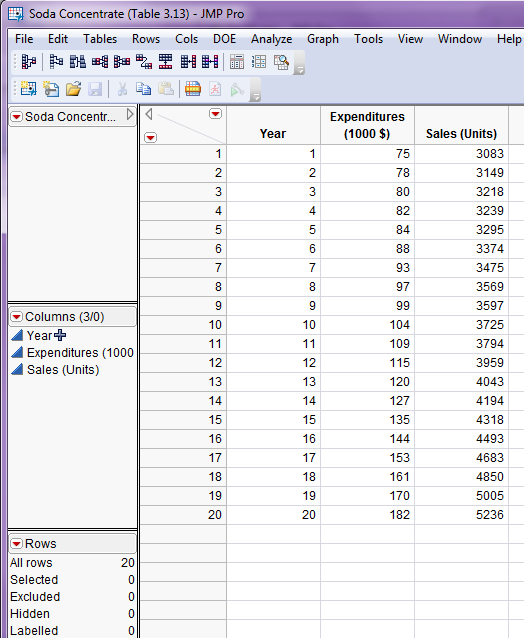
**PRESS and Durbin-Watson Statistics**

Basic idea behinds these statistics:

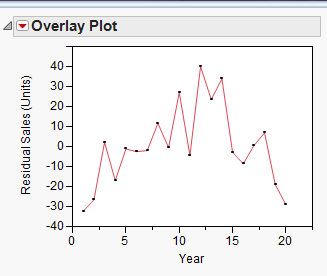
**Output from Patient Satisfaction Example:**

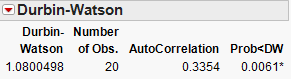


**Example 2: Soda Concentrate Sales (Durbin-Watson statistic/test)**



After fitting the simple linear regression model of Sales on Expenditures and saving the residuals from the fit we can plot the residuals vs. year (time). Autocorrelation is clearly evident from the both the plot and the Durbin-Watson test. Given that the observations are collected over time, the significance of this test suggests addressing the significant autocorrelation through our modeling is important.

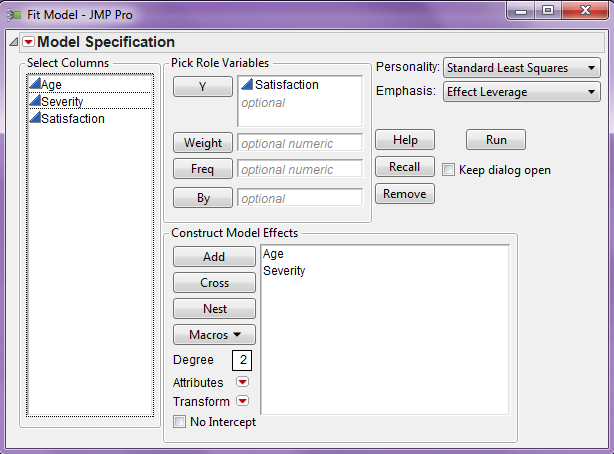




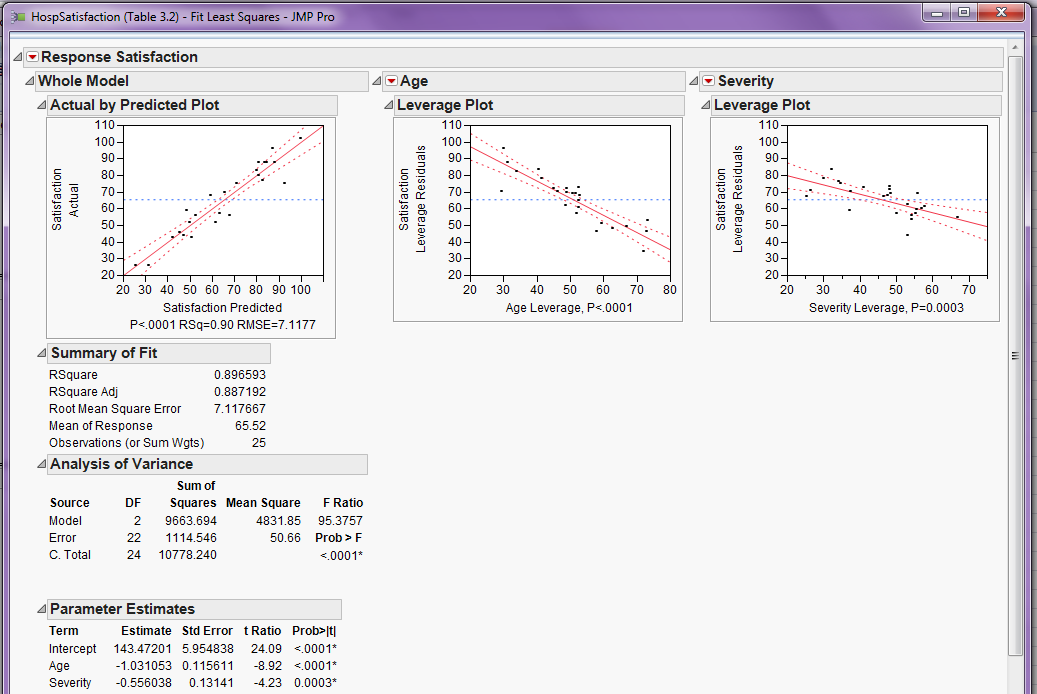
**Fitting a Multiple Regression Model**To fit a multiple regression in JMP we must use **Analyze > Fit Model** and add the terms we wish to use in our model. As an example we consider the for hospital satisfaction using both age of the patient and severity of illness as terms in our model, i.e. we fit the following model to our survey results,

where,

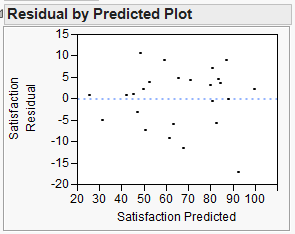
The dialog box for Fit Model is shown below.



The output is shown on the following page.



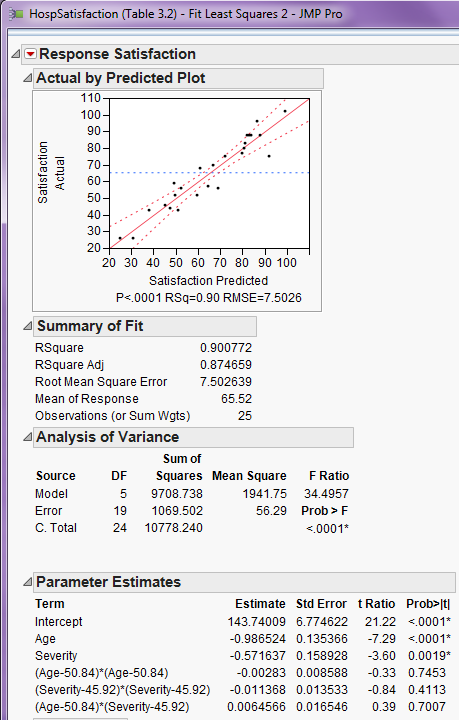
Comments:



We can see that both predictors, Age  
& Severity are significant (p < .001).   
  
The residual plot shows no model   
deficiencies.

The authors consider adding more terms based on the predictors Age and Severity in Example 3.3 on page 91 of the text. The terms they are add are: , giving us the model:

To fit this model in JMP

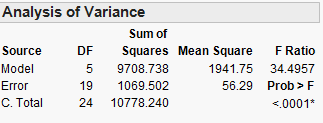
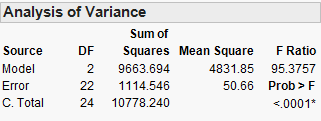


None of the additional terms are significant. We can test dropping these terms one at a time sequentially starting with as it has the largest p-value (p = .7453). Another approach is consider dropping all of these terms simultaneously using the “General F-test” as shown below.

The General F-test for comparing *nested* models.

where,

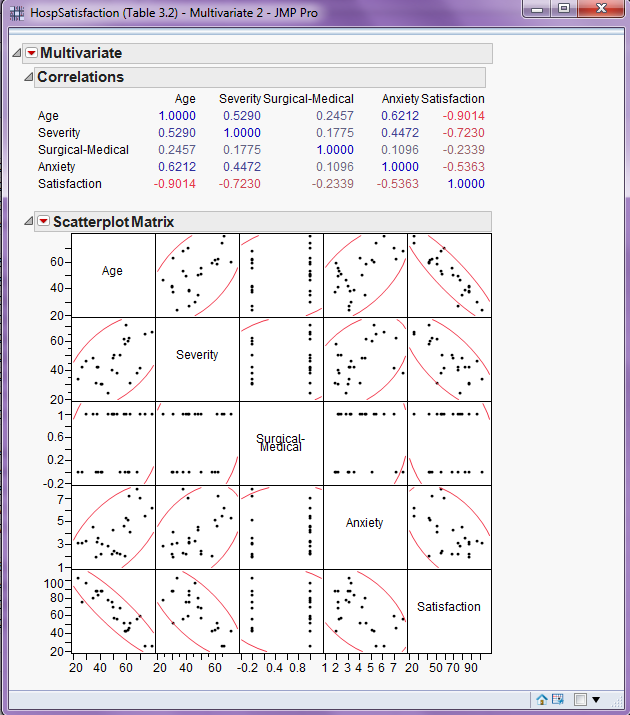
The relevant output for both models is shown below.

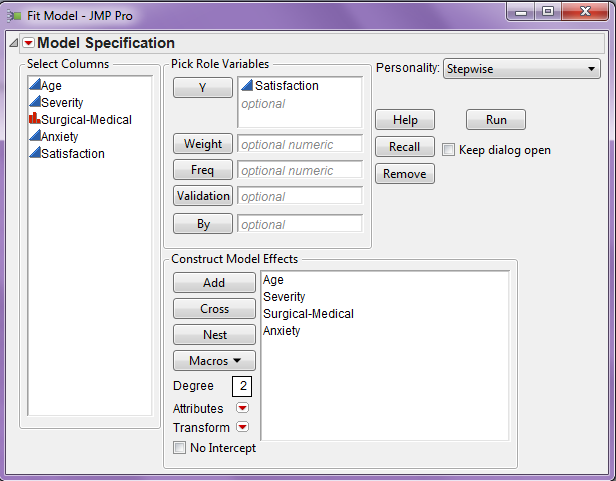
**Variable Selection Methods - Stepwise Regression** (see section 3.6 of text)

*Backward Elimination* – include potential term of interest initially and then remove variables one-at-time until no more terms can be removed.

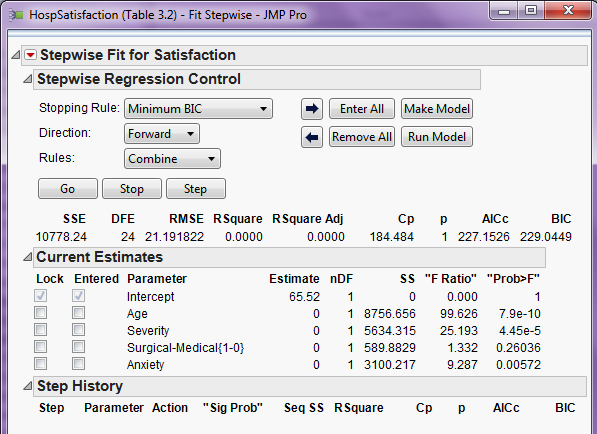
*Forward Selection* – add the best predictor first, then add predictors successively until none of the potential predictors are significant.



To use stepwise selection in JMP change the Personality menu to Stepwise in the Fit Model dialog box, as shown below.



The stepwise control panel is displayed below:



By default the control panel is initialized to carry out forward selection using the BIC model criterion (similar to adjusted-R2).

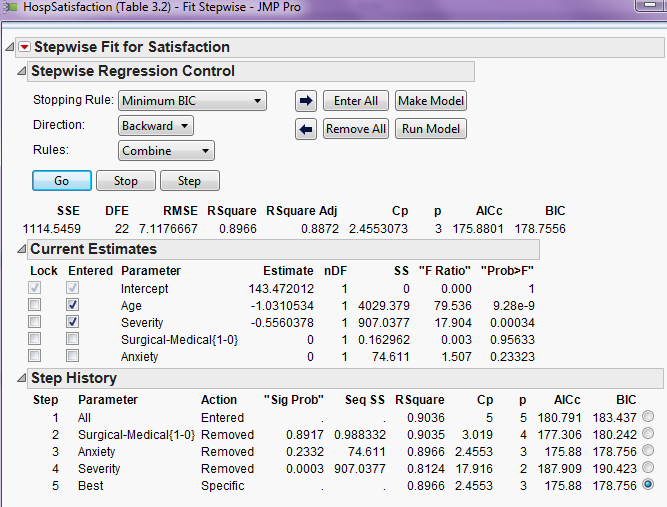
***Backward Elimination***: Click Enter All and change Direction to Backward and click Go.

***Forward Elimination***: Click Remove All and change Direction to Forward and click Go.

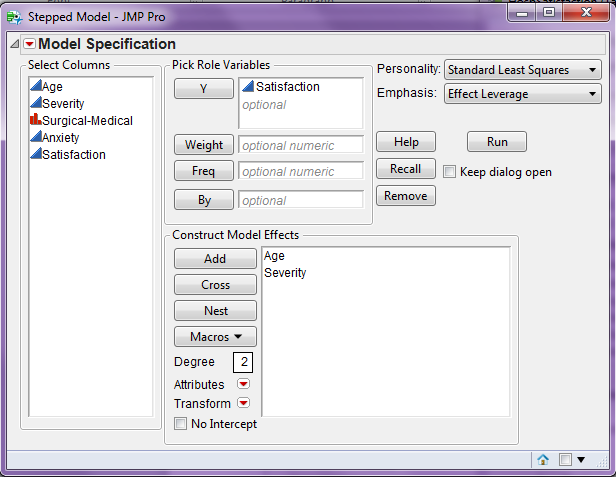
***Mixed***: You can either start no terms (Remove All) or all terms (Enter All) and click Go.

***Stopping Rule***: Is either set to AIC, BIC, or p-value based selection. The three generally will give similar results unless the list of candidate terms is large. For this example the three choices lead to similar models.

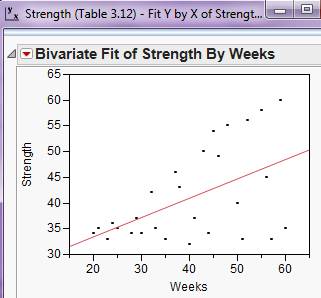
The final model chosen by using Backward Elimination with BIC is shown below.



If you agree with the “Best” model chosen select Make Model to run this model and examine the results which will be the same as if we used Fit Model run this model in the first place. The best model uses only Age and Severity as terms in the model. The other candidates: patient Anxiety and whether they were in the Surgical or Medical unit are not used as they don’t significant improve the model.



**Weighted Least Squares (WLS) – Example 3.9 (**pg . 117**)**In weighted least squares we give more weight to some observations in the data than others. The general concept is that we wish to give more weight to observations where the variation in the response is small and less weight to observations where the variability in the response is large. In the scatterplot below we can see that the variation in strength increases with weeks and strength. That is larger values of strength have more variation than smaller values.

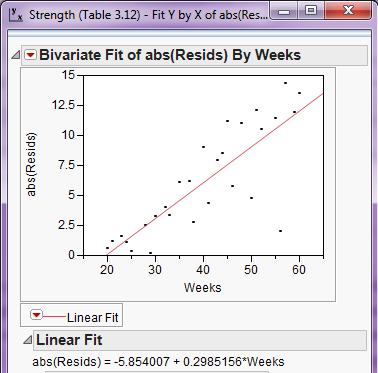


Comments:

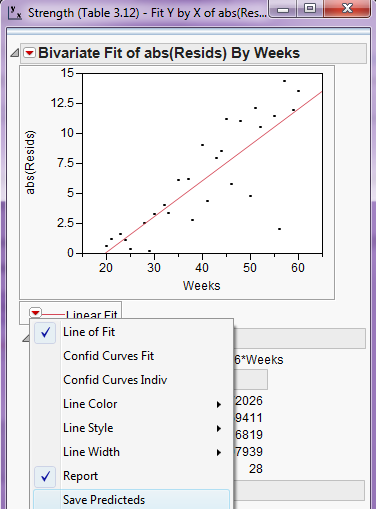
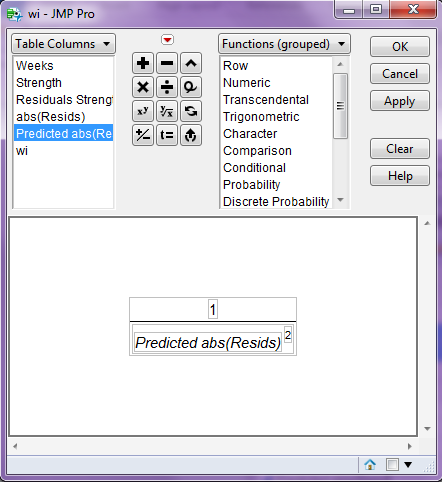
In WLS the parameter estimates are found by minimizing the weight least squares criteria:

The problem with weighted least squares is how do we determine appropriate weights? A procedure for doing this in the case of nonconstant variance as we have in this situation is outlined on pg. 115 of the text. We will implement this procedure in JMP. First we need to save the residuals from an OLS (equal weight for all observations) regression model fit to these data. Then we either square these residuals or take the absolute value of these residuals and fit a regression model using either residual transformation as the response. In general, the absolute value of the residuals from OLS fit will work better.

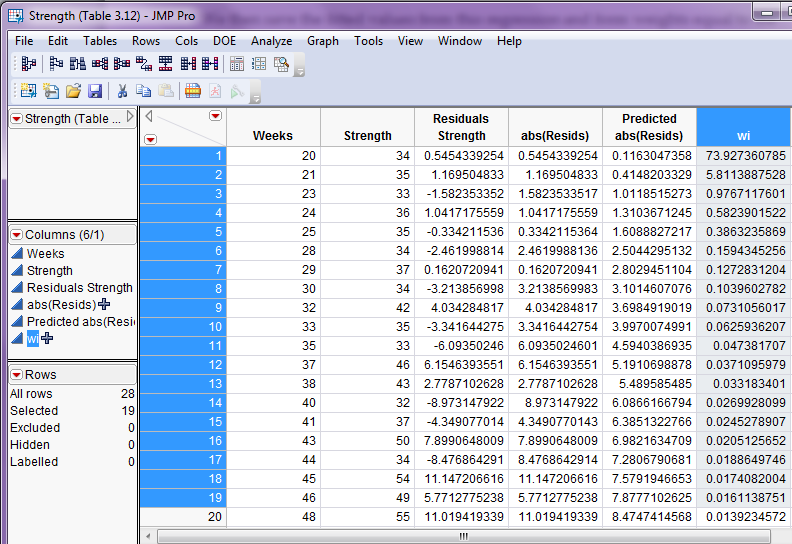
Below is a scatterplot of the absolute residuals vs. week (X) with an OLS simple linear regression fit added.



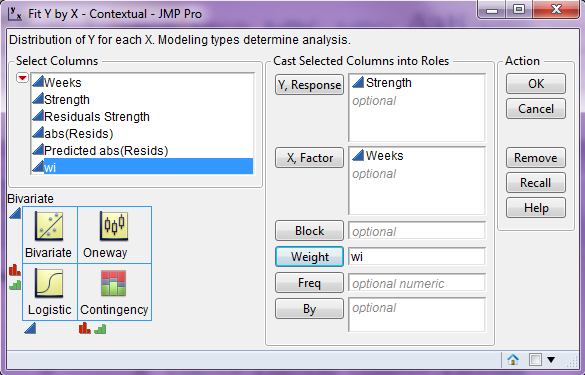
We then save the fitted values from this regression and form weights equal to the reciprocal of the square of the fitted values, i.e.

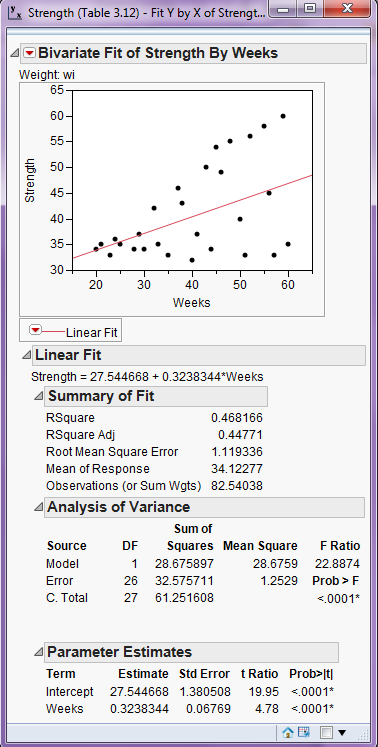
The resulting JMP spreadsheet is shown below.



Finally we fit the weighted least squares (WLS) line using these weights.



The resulting WLS fit is shown on the following page.



Contrast these parameter estimates with those obtained from the OLS fit.  
OLS WLS

If the parameter estimates change markedly after the weighting process described above, we typically repeat the process using the residuals from first WLS fit to construct new weights and repeat until the parameter estimates do not change much from one weighted fit to the next. In this case the parameter estimates do not differ much, we do not need to consider repeating. The process of repeating WLS fits is called ***iteratively reweighted least squares*** (IRLS).

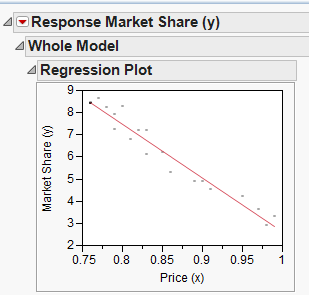
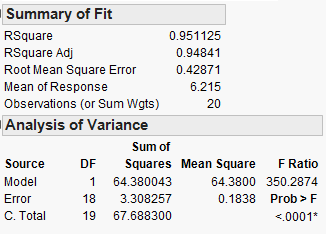
While we will not be regularly fitting WLS models, we will see later in the course many methods for modeling time series involve giving more weight to some observations and less weight to others. In time series modeling we typically give more weight to observations close in time to the value we are trying to predict or forecast. This is precisely idea of Discounted Least Squares covered in section 3.7.3 (pgs. 119-133). We will not cover Discounted Least Squares because neither JMP nor R have built-in methods for performing this type of analysis and the mathematics required to properly discuss it are beyond the prerequisites for this course. However, in Chapter 4 we will examine exponential smoothing which does use weights in the modeling process. In exponential smoothing all observations up to time *t* are used to make a prediction or forecast and we give successively smaller weights to observations as we look further back in time.

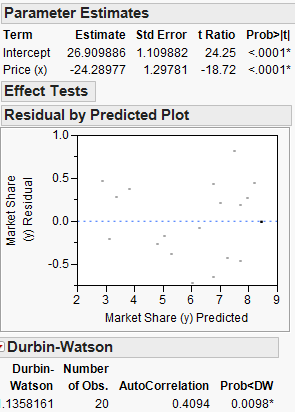
**Detecting and Modeling with Autocorrelation**

* Durbin-Watson Test
* Cochrane-Olcutt Method
* Use of lagged variables in the model

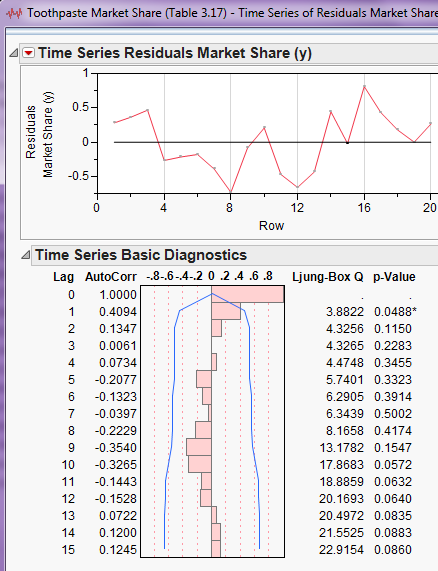
**Example 3.14** (pg. 143-145)

In this example the toothpaste market share is modeled as function of price. The data is collected over time so there is the distinct possibility of autcorrelation. We first use Analyze > Fit Model to fit the simple linear regression of market share (Y) on price (X).



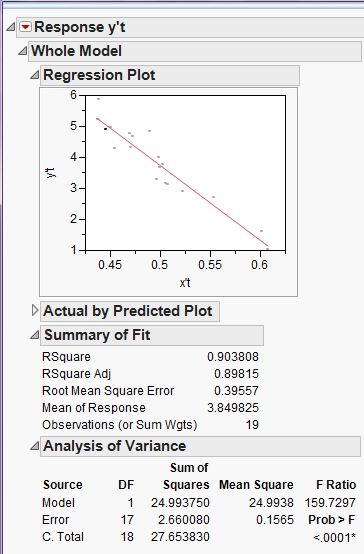
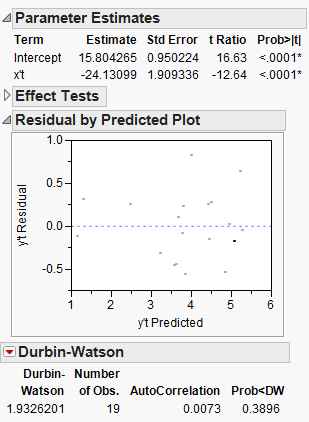
The Durban-Watson statistic *d = .136* is statistically significant (p = .0098). Thus we conclude a significant autocorrelation ( exists in the residuals and we conclude the errors are not independent. This is a violation of the assumptions required for OLS regression. The plot below shows the autocorrelation present in the residuals from the OLS fit.



The Cochrane-Olcutt method for dealing with this autocorrelation transforms both the response series for market share and the predictor series for price using the lag 1 autocorrelation in the residuals. We form the series and using the following formulae:

where , the lag-1 autocorrelation in the residuals from the OLS fit. These transformed version of the response and predictor incorporate information from the previous time period in the modeling process, thus addressing the autocorrelation.

Below is the regression of on fit using **Analyze > Fit Model**.

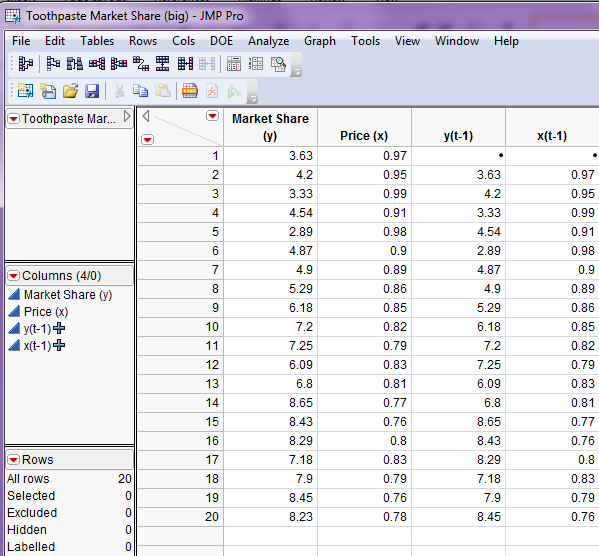
There is no significant autocorrelation in the residuals from the model fit using the Cochrane-Olcutt method.

Another approach to dealing with autocorrelation in a OLS regression is to include lags of the response series and the predictors series in the regression model. Equations 3.119 and 3.120 in the text give two models for doing this. The first incorporates first order lags of the response series and the predictor series as terms in the model and second uses only the first order lag of response series.

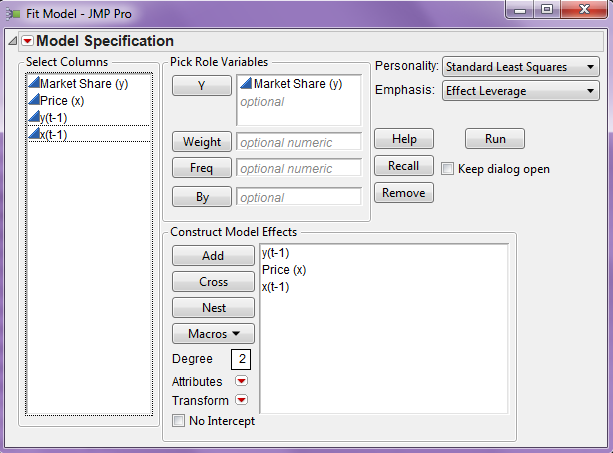
where (3.119)

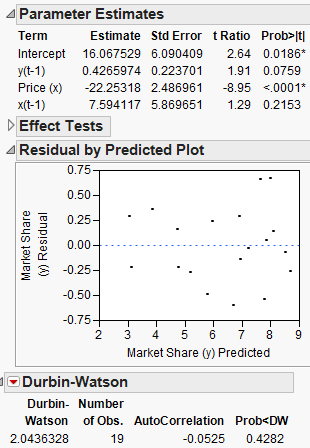
where (3.120)

We will fit both of these models to toothpaste market sales data. First use the JMP calculator to form lag-1 series for both the market share (Y) and price (X) series.

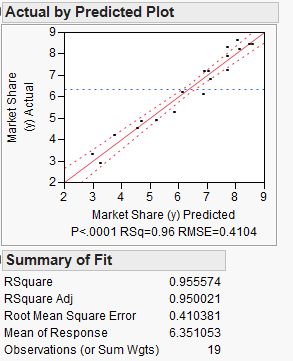
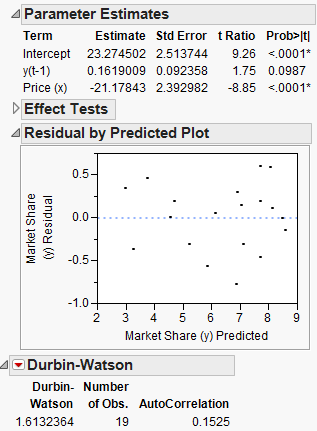


First we use **Fit Model** to fit the first model (3.119)





First notice the Durbin-Watson test does not suggest significant autocorrelation. As the lag term () is not significant we can drop it from the model and fit the model in equation 3.120. This model fits the data well and does not suffer from autocorrelation.

**Final Examples: Beverage Sales**

We return again to the beverage sales example and consider developing a few regression models that can be used predict or forecast future sales. One model will make use of trigonometric functions of time to model the seasonal behavior in this time series (Model 1), the other will use month as a factor term (Model 2). Autocorrelation issues will be explored for both models. The mean functions for these two modeling approaches are outline below.

Model 1:

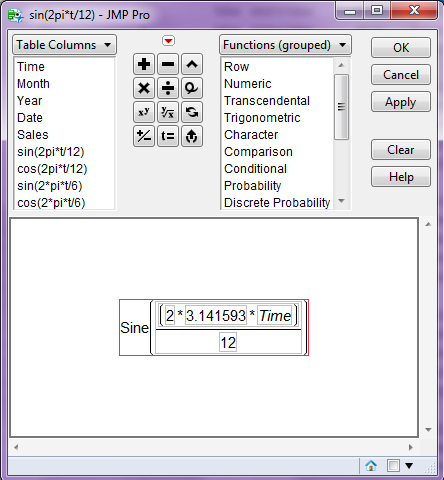
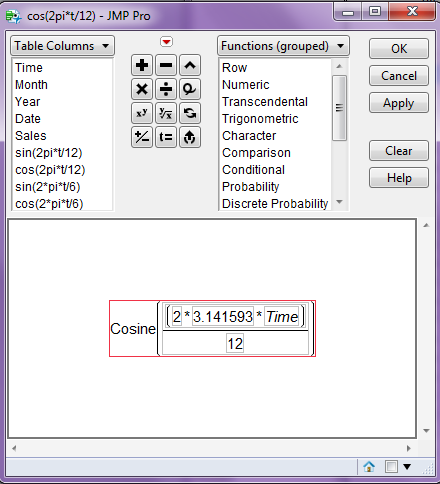
Additional terms may be added to the model for example other trigonometric   
 terms at other periodicities or lag terms based on the series

Model 2:

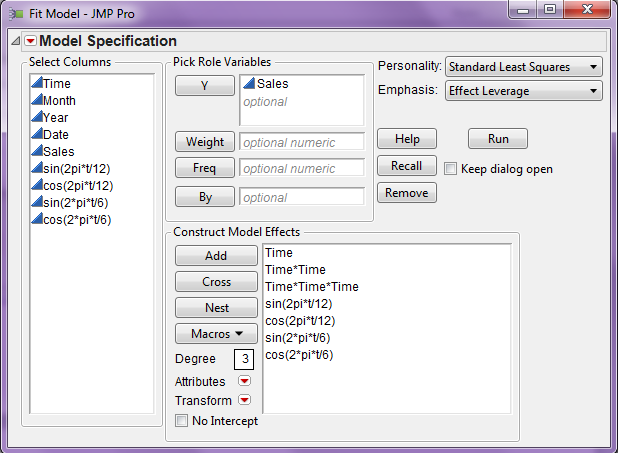
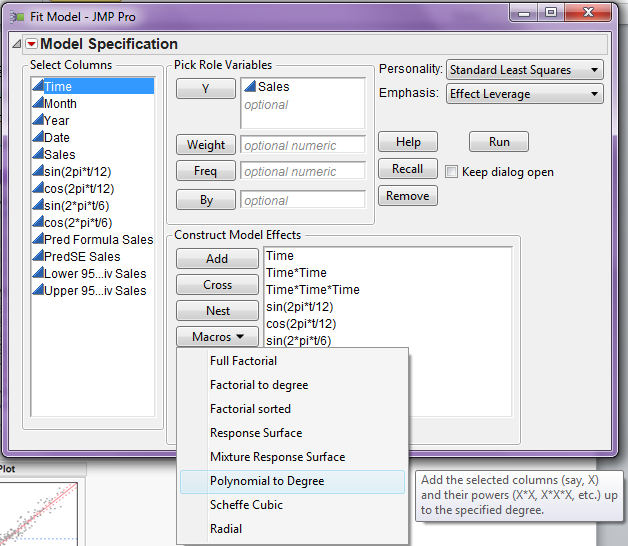
where,

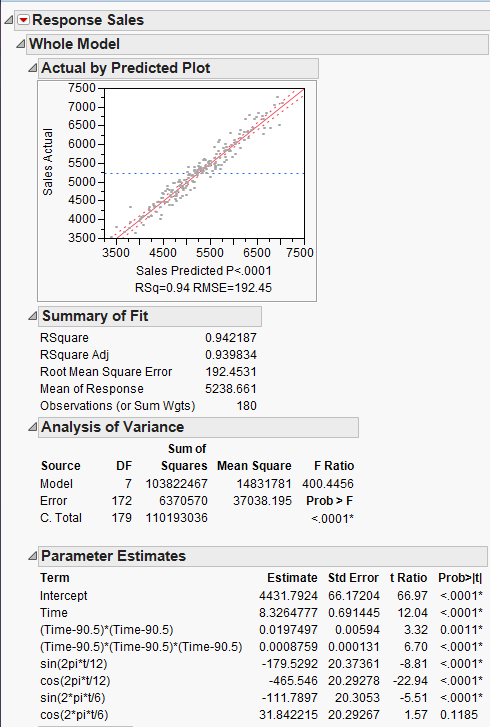
Additional lag terms of the time series may be added to deal with   
 potential autocorrelation issues.

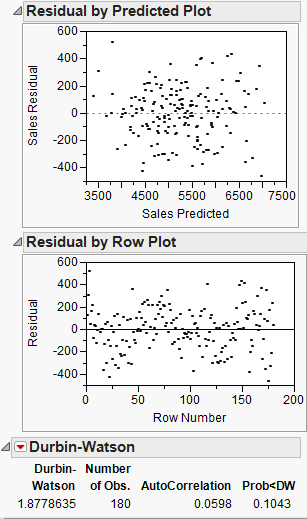
Fitting Model 1: In order to fit model 1 we first need to form the trigonometric terms in the JMP calculator. There is no built-in constant for so we will use 3.141593.

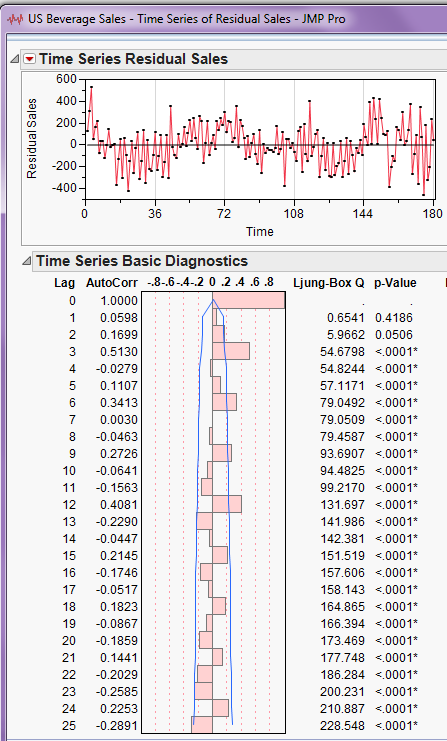
After also creating two 6-month period terms we fit the model using a cubic polynomial to capture the long term trend in the time series and the four trigonometric terms. To fit the cubic polynomial change the box labeled Degree to 3 then highlight Time in the variable list and select Polynomial to Degree from the Macros pull-down menu.

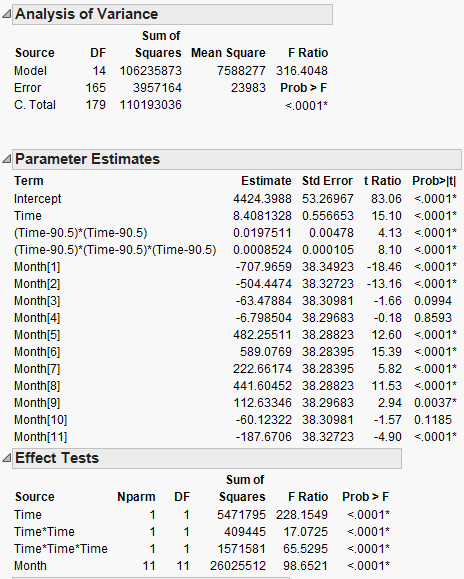
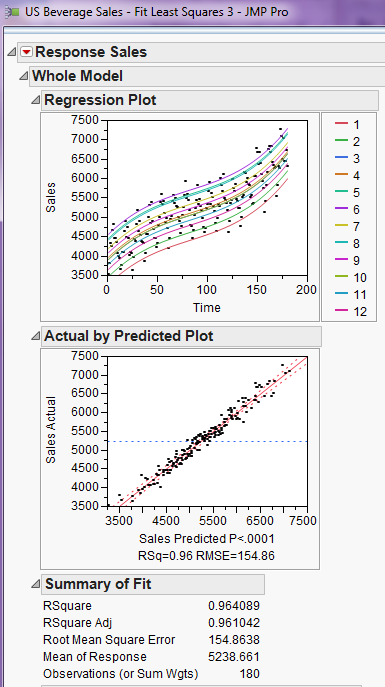




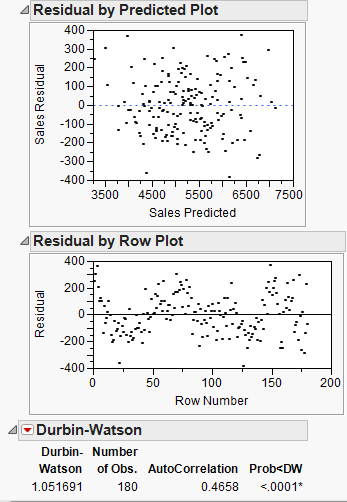
Despite the fact that the Durbin-Watson test statistic is not significant, examination of an ACF for the residuals from this model suggests there is still some autocorrelation structure in the residuals that should be accounted for.



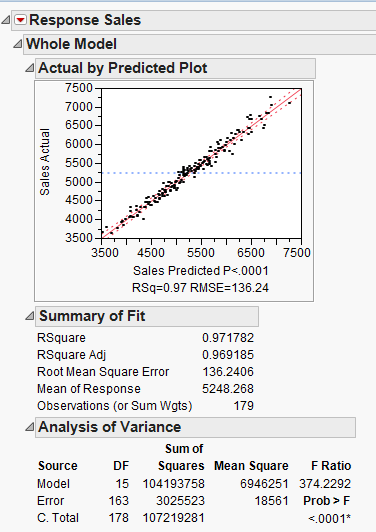
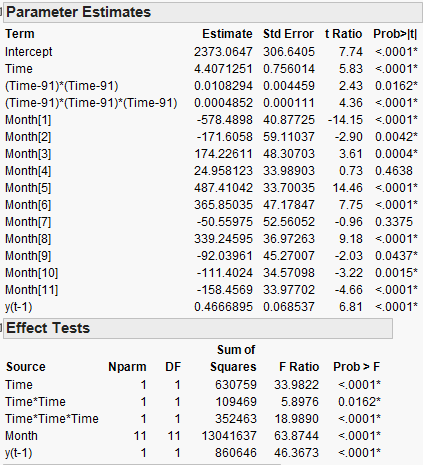
Fitting Model 2: Fitting this model requires first building terms for the cubic polynomial in time and then specifying that Month should be treated as categorical/nominal, i.e. as a factor. JMP will then create dummy variables for the first 11 months and include these in the model.

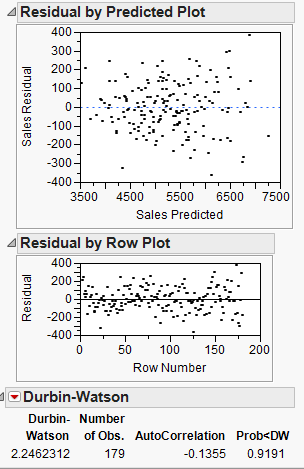


Examining the residuals and testing for autocorrelation we find.



We could add the term to the model in attempts to deal with the autocorrelation.



The inclusion of the lag-1 response term has addressed the autocorrelation in the residuals, which is confirmed by the Durbin-Watson test (p = .9191) and the ACF plot below.

