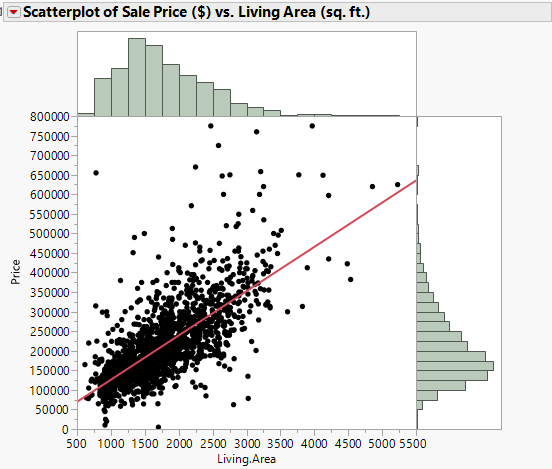
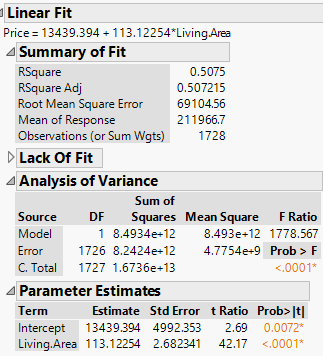
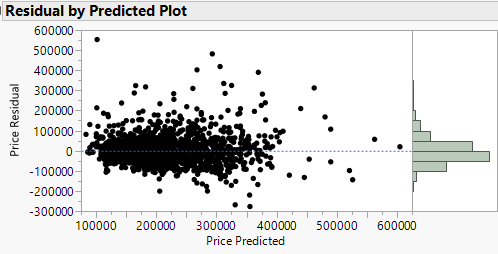
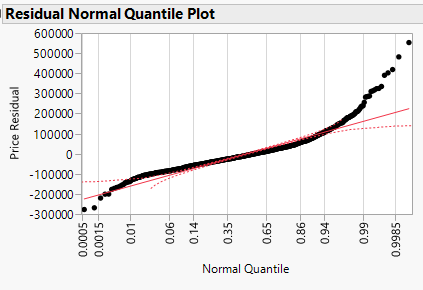
**8.0 – Introduction and Motivating Example**

For the remainder of the course our focus will be on multiple regression models. In multiple regression we study the conditional distribution of a response variable () given a set of potential predictor variables . These variables can have any data type – continuous/discrete, ordinal, or nominal. Before covering the details of multiple regression, we will first consider a motivating example that illustrates some important concepts about developing multiple regression models and how to create ***terms***from predictors.

**Example 8.1 – Sale Price of Homes in Saratoga, NY  
Datafiles: Saratoga NY Homes.JMP, Saratoga NY Homes.txt**

In this example we will examine regression models for the sale price of home using two potential predictors: and square footage of the main living area and whether the home has a fireplace of not (Yes or No). We begin by considering the simple linear regression of sale price (Y) on living area ().

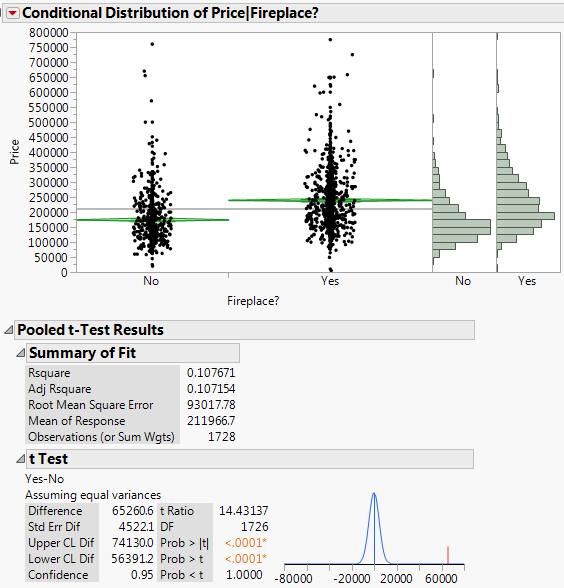
Residual Plots:  
 

While we may not be completely satisfied with the assumptions that,

we will now consider the regression of sale price on whether or not the home has a fireplace ().

As this predictor is nominal (Yes or No) we will essentially be comparing:

vs.

which could be done using an independent samples t-test. Here a pooled t-test is used even though there is evidence the variances in the sale prices of homes with and without fireplaces are not equal.  


**Conclusion from t-test:**

We have extremely strong evidence to conclude that the mean selling price of homes with a fireplace exceeds the mean selling price of homes without one (p < .0001). Furthermore we estimate the mean selling price of homes with a fireplace exceeds the selling price of homes without a fireplace by between $56,391and $74130 with 95% confidence.

**8.1 – Regression with a Dichotomous Nominal Variable**

Can we formulate the two-sample t-test above a simple linear regression? We can if we recode the nominal variable fireplace status (Yes/No) as a numeric variable. This can be done one of two ways.

**Dummy Variable Coding (R default)** **Contrast Coding (JMP default)**

Dummy variable coding is called “Indicator Function Parameterization” in JMP.

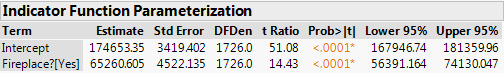
The standard SLR models can then be written as follows for the two parameterizations.

***Model using Dummy Coding***

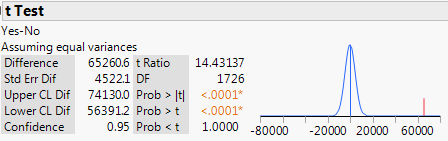
or more explicitly,

Thus represents the shift in the mean selling price for homes with fireplaces.

Below are the parameter estimates obtained fitting this model to our data.



Here we can see that and the associated t-statistic for testing the vs. is given by .   
Compare this result to the t-test conducted above.

  
Here we see the pooled t-test is equivalent in every way to a simple linear regression on a dummy variable indicating group membership! Remember most statistical analyses are just regression in disguise!

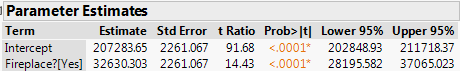
***Model using Contrast Coding***

or more explicitly,

Thus if we consider the difference in the mean selling price between homes with and without fireplaces we have:

Thus an estimate of the difference in the population between means is .

The summary from fitting the model using contrast coding is given in the JMP output below.



Here we see that , thus the estimated difference in the population means is , which is same the results from the pooled t-test and for the dummy coding approach above. Notice that despite the change in the coding the t-statistic and associated p-value are exactly the same as the results above.

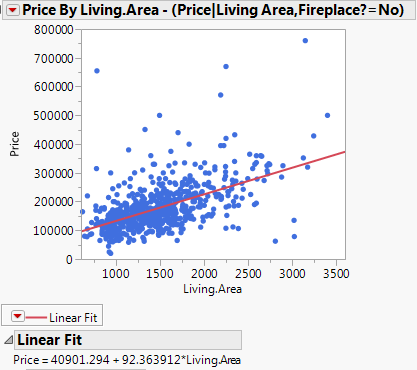
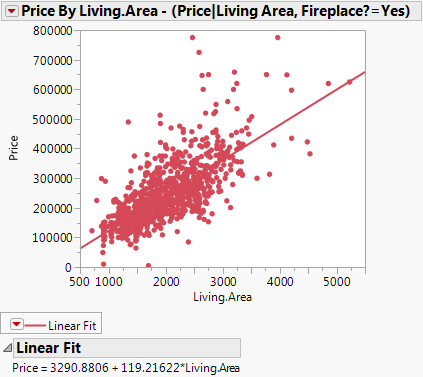
To obtain a confidence interval for the difference in the mean selling price for homes with and without a fireplace we need to adjust our CI by a factor of 2 as well, i.e.

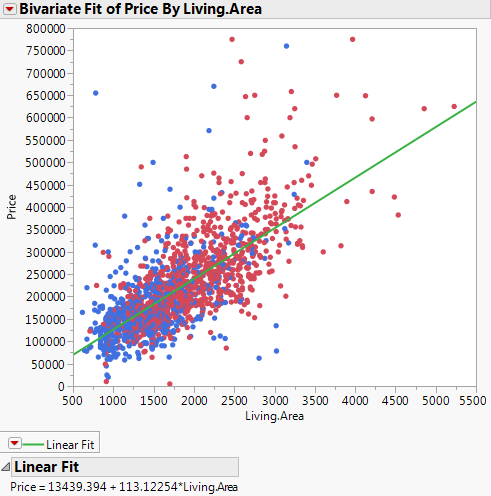
.

This again agrees with the results from the pooled t-test and dummy coding regression above.

Do the results above suggest that if someone will put a fireplace in my home that currently does not have one for $25,000 I should do it and more than double my investment? The results of the t-test above certainly seems to suggest it would be financially sound to do so. However this conclusion is wrong!!

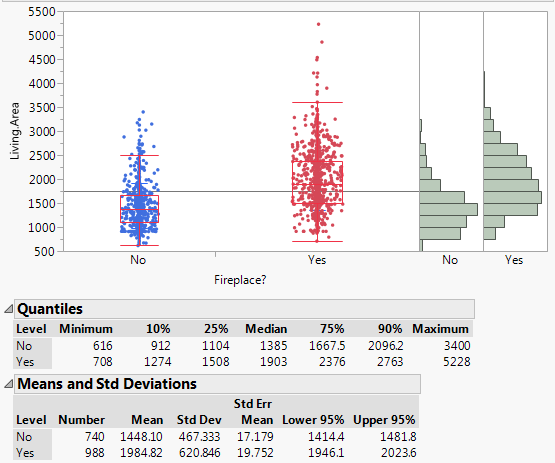
To understand why, consider again the scatterplot of selling price () vs. living area (), but this time we will color code the observations according to whether or not the home has a fireplace.



What would you say is generally true for homes   
with fireplaces in terms of the size of the main living area ()?

We can see this more clearly by examining the conditional distributions of living area given fireplace status. Homes with fireplaces are over 500 larger!



What is the estimated increase in the mean selling price associated with a 500 increase in the size of main living area?

Not quite $65,260, but certainly explains the bulk of this figure!

The punchline of the discussion on the previous page is that when considering what role having a fireplace in your home not has on the mean selling price we need to take other information about the home into account. Intuitively this should make sense, if I have crappy one bedroom house, putting a fireplace in my home will not likely net me an additional $65,000 when I sell it. If you consider all of the other variables we have in this dataset (e.g. # of bathrooms, # of bedrooms, lot size, etc.) we should immediately realize there are many other potential predictors to consider when developing a model for the mean selling price of a home!

**8.2 – Multiple Regression Model – one nominal and one continuous**

We now will consider a multiple regression model for selling price that takes both of these predictors:

into account.

The basic form of a multiple linear regression model for the mean is:

The are ***terms*** that are the functions of the ***predictors*** () and must be numeric. The terms may be the predictors themselves as might be the case with the *Living Area* (). However when we considered the predictor *Fireplace?* , which is a nominal variable (No or Yes), we needed to create a numeric term using either dummy (0/1) or contrast (-1/+1) coding in order to use it in a regression model. If we wanted to use the logarithm of *Living Area* in our model instead, we would create the term . We will discuss **terms** and **predictors** in more detail in the next section.

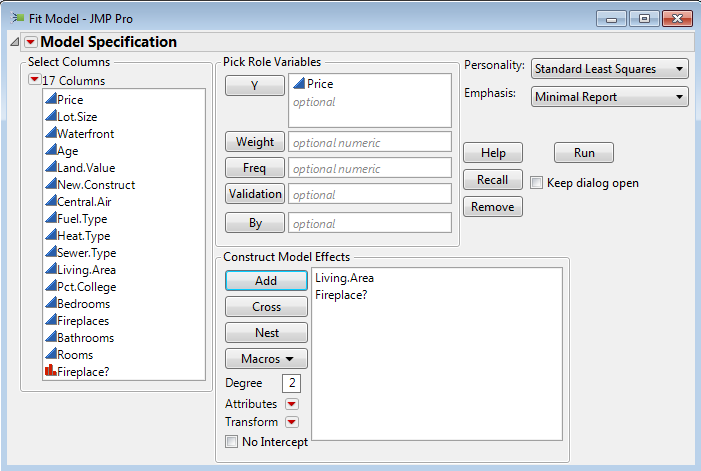
One potential multiple regression model using both predictors would be:

where and

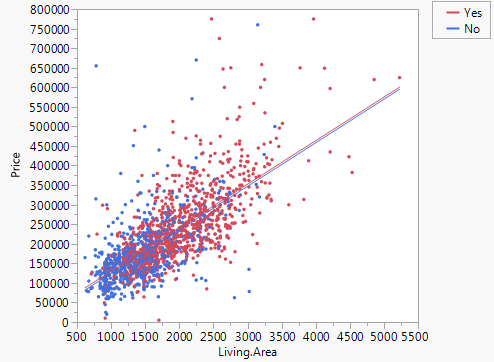
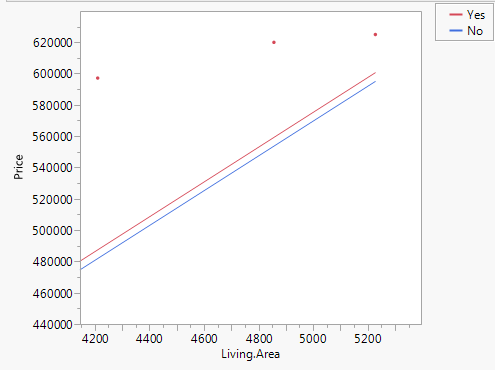
To better understand this model consider the following:

Thus the y-intercept is for homes with a fireplace and for homes without a fireplace and the slope is the same for both home types.

To fit this model in JMP select **Analyze > Fit Model** and place both *Living.Area* and *Fireplace?* in the Construct Model Effects box and the response Price in Y box as shown below.

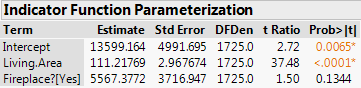


The model summary is shown below:

The gap between these parallel lines is almost imperceptible in this plot. Zooming in shows a small gap.

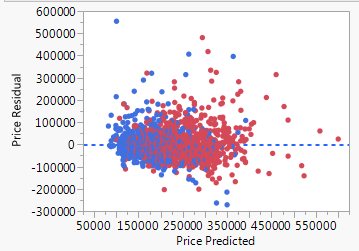
Below are the parameter estimates and significance tests for the three parameters in our model (.

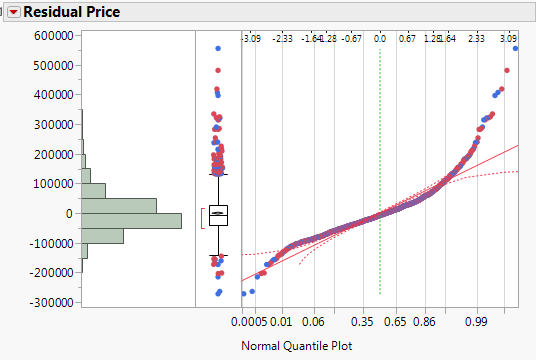


We can see that main living area is highly significant (p < .0001), however now that we have incorporated information about the size of the home in our model, we do NOT have evidence of a difference in the mean selling price associated with having a fireplace or not (p = .1344)!

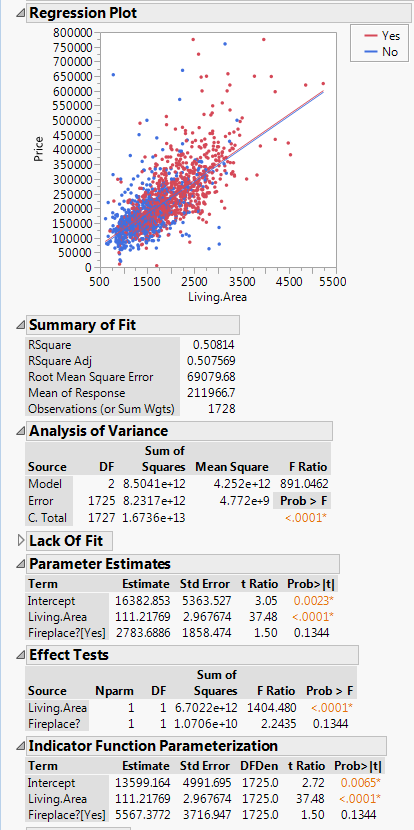
The model on the previous page is called the ***Parallel Lines model*** as the slope () is the same for homes with and without fireplaces.

**Residual Plots**





**Complete Summary for the Parallel Lines Model**



It still may be the case that having a fireplace or not does significantly affect the mean selling price of a home as we do not necessarily need the slope of the regression line for homes with and without fireplaces to be the same.

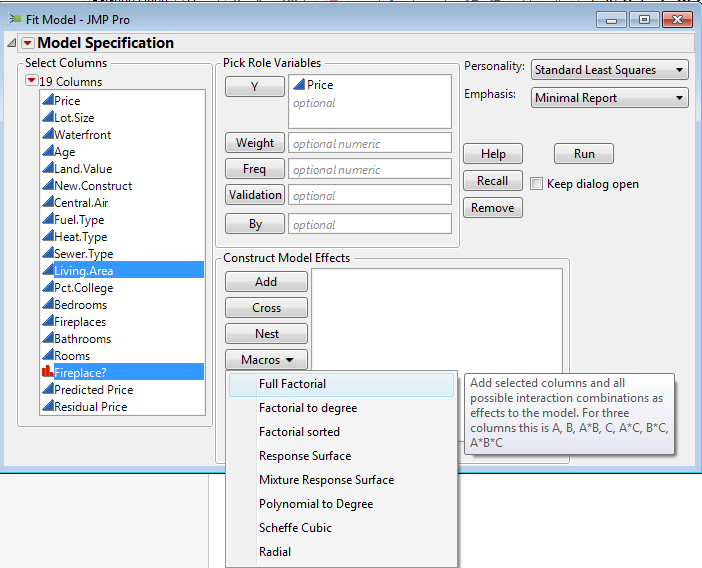
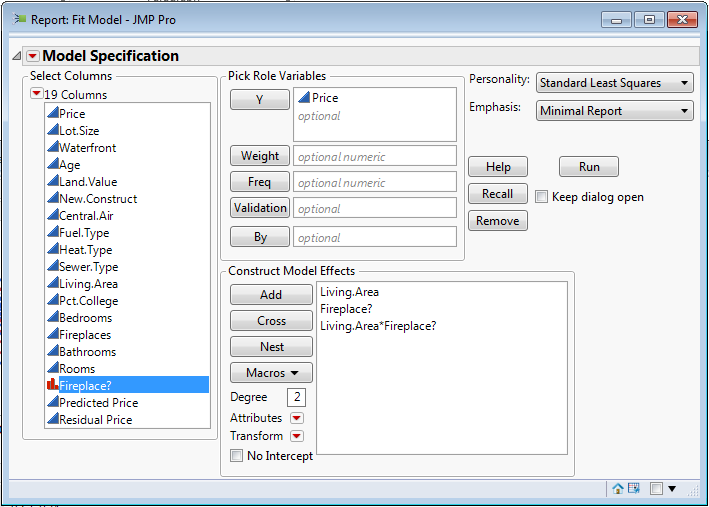
***The Unrelated Lines Model***

We now consider a model that will allow the slopes of the lines to be different.

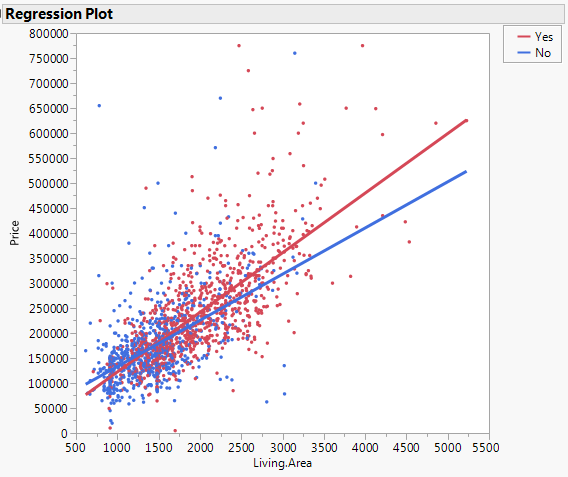
where again and

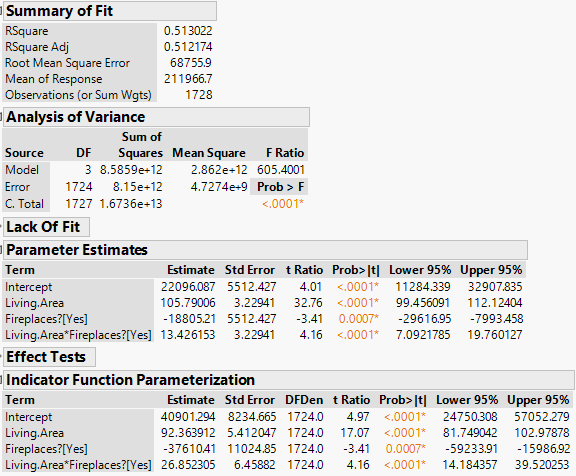
The term is called an ***interaction*** term, in this case it represents the interaction between the size of main living area and whether a home has a fireplace or not. Interactions allow for the effect of one term to be dependent on the value of another. Here that would simply mean that change in the mean price associated with a change in the main living area is not the same for homes with and without fireplaces. For example, it is possible that the per square foot increase in price ($/ is greater for homes with a fireplace than for those without or vice versa.

To better understand this model and how this happens from a parameter standpoint, we again consider the mean functions for homes with and without fireplaces separately.

To fit this model in JMP we first select **Analyze > Fit Model**, then highlight both *Living.Area* & *Fireplace?* simultaneously, and finally click **Macros > Full Factorial** as shown below.  
 

A summary of this model fit to these data is on the following page.





Interaction term

Apparently having a fireplace DOES matter! It just depends on the size of the home as seen by the significance of the interaction term (p < .0001). JMP does something interesting with interaction term that we will discuss on the following page.

**8.3 - Estimates and Predictions for Multiple Regression Models**

The estimated regression model with Dummy Coding (Indicator Parameterization) for selling price is:

Recall that,

and

Thus we can write this model out for homes with and without fireplaces as follows.

**With Fireplace(s)**

**Without Fireplace(s)**

Find the estimated mean selling price for the following homes:

1. 1000 sq. ft. home without a fireplace
2. 1000 sq. ft. home with a fireplace
3. 3000 sq. ft. home without a fireplace
4. 3000 sq. ft. home with a fireplace

**Mean Centered Terms/Predictors (** **)**

Some statistical software packages, JMP included, will ***mean center*** continuous terms involved in an interaction or raised to a power (i.e. polynomial terms) in a regression model. While this has no effect on the qualities of the fit, the estimated regression parameters will differ substantially. The reason this is done, particularly in the case of polynomial terms, is that squaring and cubing large numbers will make them much larger forcing parameter estimates associated with these to be near zero in value to compensate.

Below is the previous estimated regression model with **mean centered** and dummy/indicator function parameterization.

and

Again, it is easier to see how to work with this parameterization if we write out the model for homes with and without fireplaces.

**With Fireplace(s)**

🡨 EXACTLY THE SAME AS ABOVE!!

**Without Fireplace(s)**

🡨 EXACTLY THE SAME AS ABOVE!!

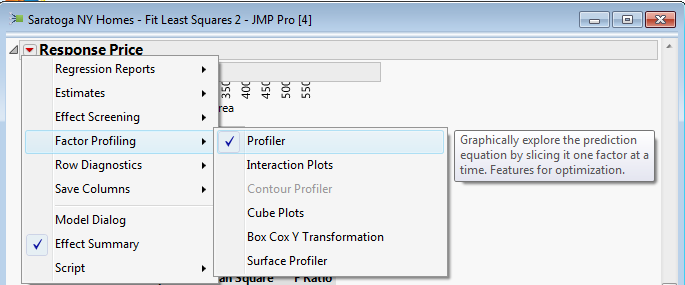
So though the model looks more complex, the end result is exactly the same. Let us consider using the model using mean centering for living area above for the find the estimated mean selling price for the same hypothetical homes used above:

1. 1000 sq. ft. home without a fireplace
2. 1000 sq. ft. home with a fireplace

1. 3000 sq. ft. home without a fireplace
2. 3000 sq. ft. home with a fireplace

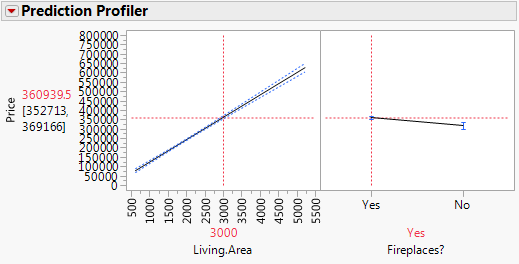
**Using the Prediction Profiler in JMP**

We can explore the fitted values/estimated conditional means, along with confidence intervals, using the Profiler in JMP. To obtain the profiler select **Factor Profiling > Profiler** from the drop down menu for the response at the top of the regression output as shown below.

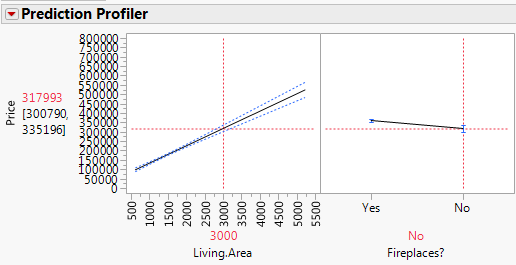


Below are some views of the prediction profiler corresponding to 3000 sq. ft. homes with and without fireplaces.

**Prediction for homes with *Living Area* = 3000 sq. ft. and at least one fireplace**



**Prediction for homes with *Living Area* = 3000 sq. ft. and no fireplace**



**Analysis of Covariance (ANCOVA)**

The previous example could be viewed as an example of an Analysis of Covariance (ANCOVA). In ANCOVA we are interested the potential effect of a nominal variable two or more levels, fireplace (Y/N) in this case, but we examine this relationship adjusting for a continuous predictor or covariate which is a potential confounder. Here the square footage of the main living area is being used as the covariate. Our analysis above allowed us to more accurately estimate the effect of fireplace by adjusting for the size of the home, which clearly was a confounding factor.

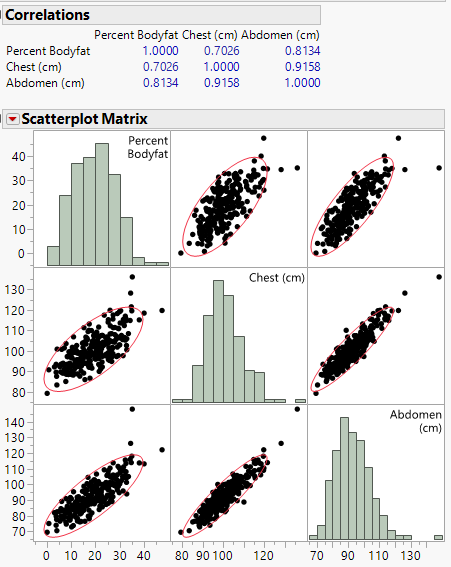
**8.4 – Multiple Regression – with at least two continuous predictors**

We now consider the case of regression with two numeric predictors. Adding a second predictor to simple linear regression model is not simply a combination of the SLR models for each predictor. To illustrate these concepts we again consider the percent body fat example.

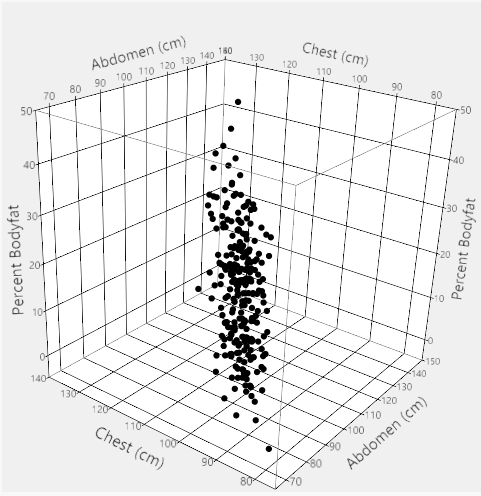
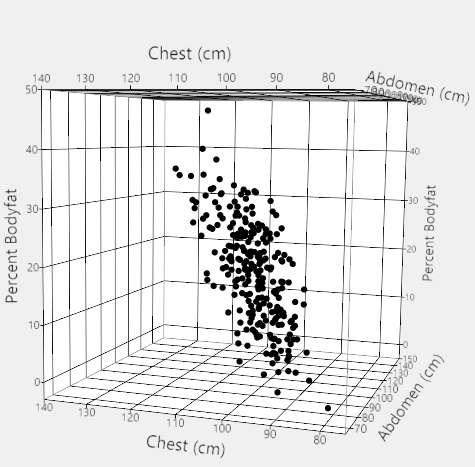
**Example 8.2 –Body Fat (%), Chest and Abdominal Circumference**

In this example we will develop a multiple regression model for percent body fat using both and as potential predictors. A scatterplot matrix of these data is shown below.

**Comments:**



A 3-D scatterplot of these data is shown below.

We will be fitting the multiple regression model

Here the terms and are just the predictors themselves in their original scales.

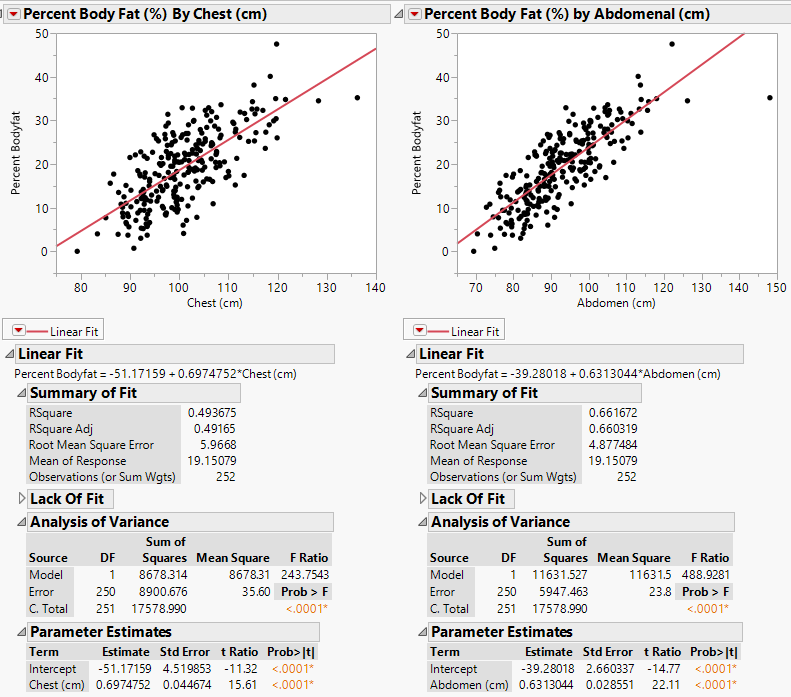
Estimated regression equation is given by:

How do the parameter estimates in the multiple regression compare to those obtained from fitting simple linear regression models to estimate percent body fat using each predictor separately (see models on the following page).

**Abdominal circumference**

**Chest circumference**

Clearly the estimated parameters in the multiple regression model are **NOT** equal to the parameter estimates for these same predictors in the SLR models!!!



As we saw above chest and abdominal circumference are highly correlated with each other. Thus the information contained one is shared in part with the other. When adding a continuous predictor to a model already containing another continuous predictor we need to take this shared information into account.

To illustrate this consider adding chest circumference to a model already containing abdominal circumference consider the following concepts and associated procedures:

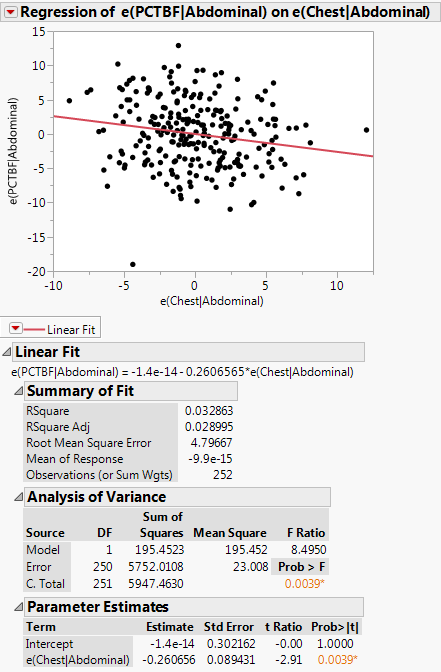
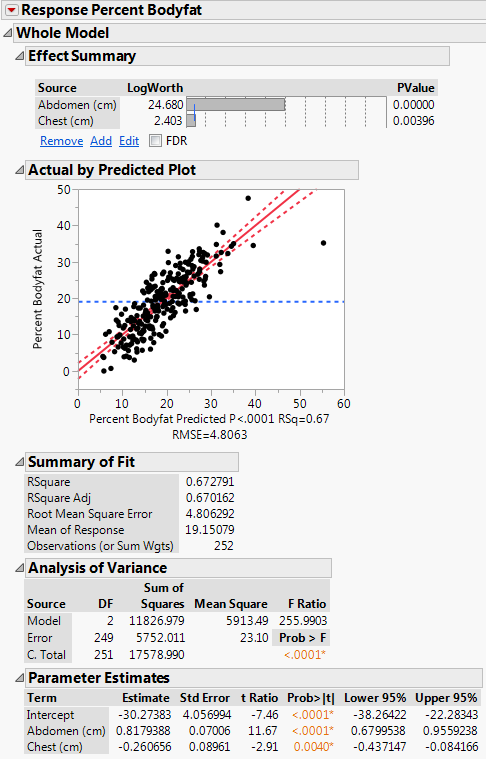
1. Chest circumference can only explain variation left over from the regression of percent body fat on abdominal circumference. That is it can only explain the unexplained (residual) variation from the model fit using only abdominal circumference. This is represented by the residuals from the regression of Percent body fat on the abdominal circumference. We will denote these residuals as .
2. Because chest circumference and abdominal circumference are correlated they have information in common, thus only the part of chest circumference that is NOT explained by abdominal circumference can be used to explain the left over variation discussed in (1.) above. This is represented by the residuals from the regression of chest circumference on abdominal circumference. We will denote these residuals as .
3. To get the estimated parameter in the multiple regression model for chest circumference we need to find the “slope” from the regression of .
4. We must realize that the order of this process is irrelevant, i.e. we could equivalently have considered adding abdominal circumference to a model already containing chest circumference instead. In fact, to get the parameter estimated for abdominal and chest circumferences above, we NEED to do this.

Thus to find the estimated parameter we reverse the roles of chest and abdominal circumference.

We will now implement the process outlined above with chest circumference being added to a model for percent body fat already containing abdominal circumference.

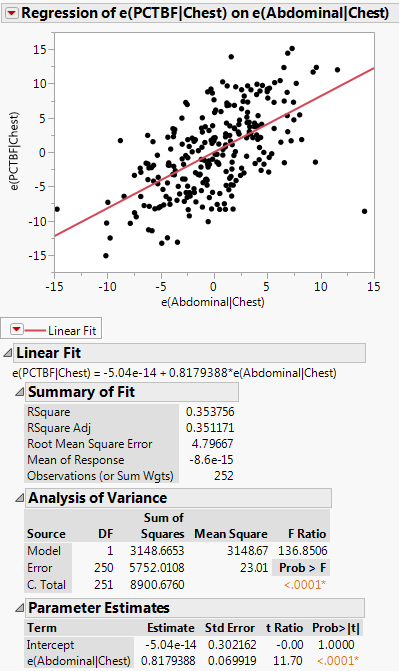
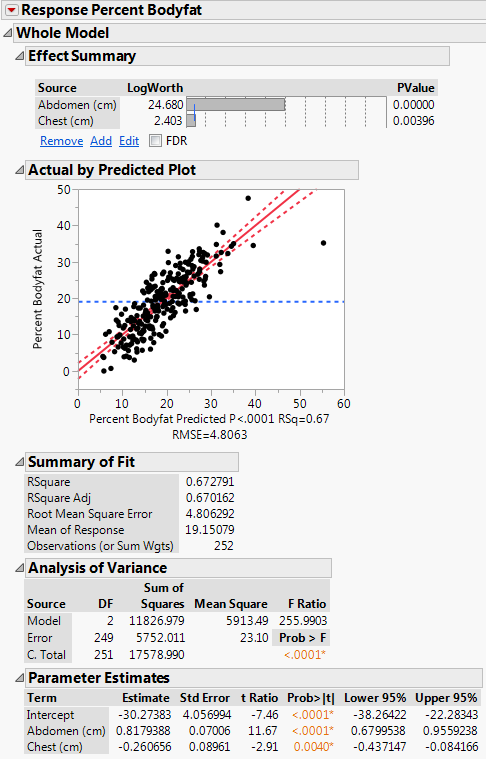
1. First we fit the model and save the residuals () or using the notation above, .
2. Now fit the model and save the residuals which we will denote as .
3. Finally we perform the regression of on . The estimated slope is the coefficient of chest circumference in the multiple regression model.

Result from Step 3 Multiple Regression of % Body Fat on Chest & Abdominal

Reversing the roles of and we obtain the following.

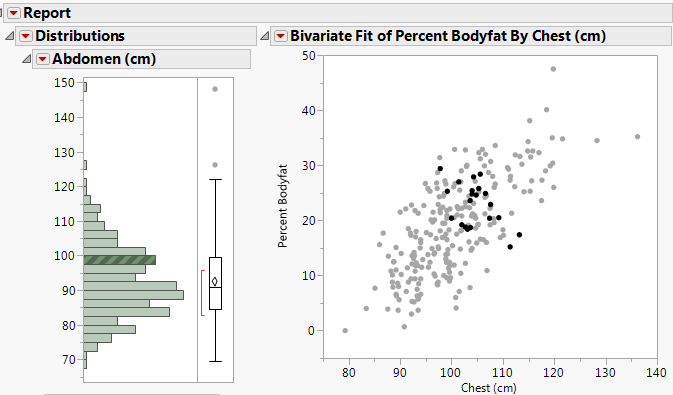
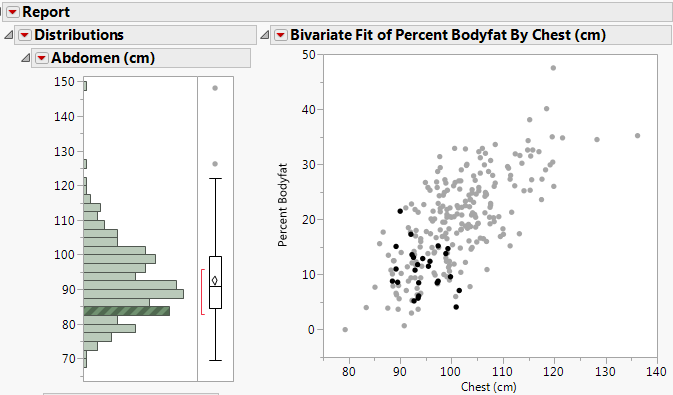
Result from Step 3 Multiple Regression of % Body Fat on Chest & Abdominal

This process leads us to the estimated multiple regression model

Something seems odd here, what is it?

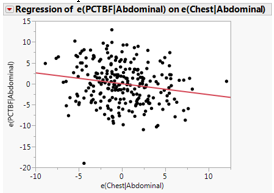
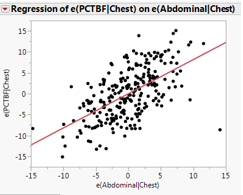
To help understand why this happens consider the graphs below:

When conditioning on abdominal circumference, the relationship between % body fat and chest circumference is a negative association!

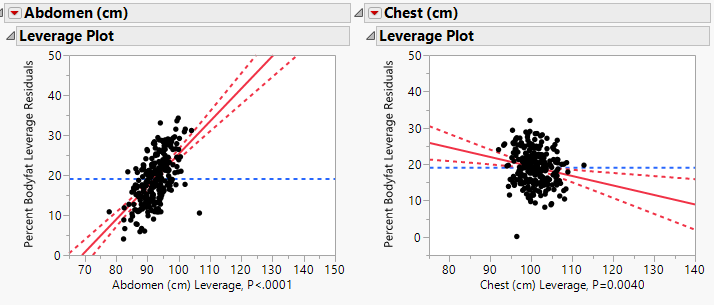
**Added Variable Plots (Effect Leverage Plots)**

The plots constructed above and shown again below are called Added Variable Plots (AVPs) or Effect Leverage Plots (JMP). AVPs show the estimated effect of a terms in a multiple regression model adjusted for the other terms in the model.

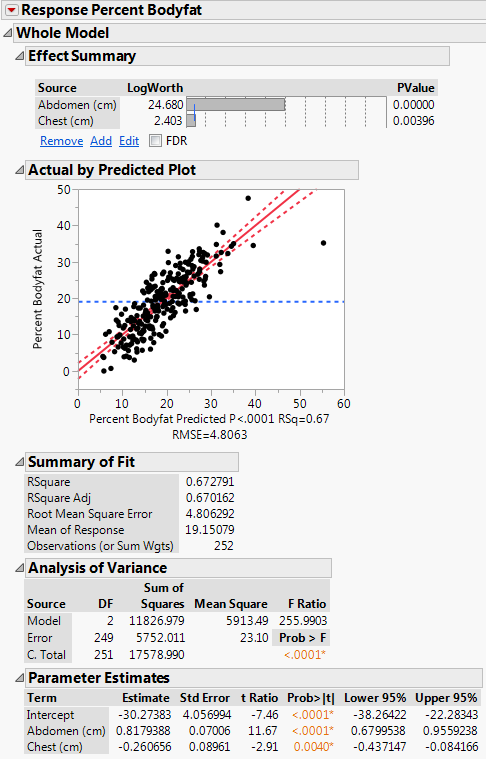
The general form of an AVP for a term in a multiple regression model is:

which is a plot of the residual variation from the regression of the response on the rest of the terms in the model vs. the residual variation from regression of the term in question on the rest of the terms in the model. Put another way it is part of the response not explained by the other terms in the model vs. the part of the term in question not explained by the other terms in the model. Below are the Effect Leverage Plots for both and .



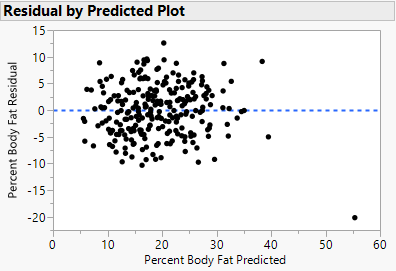
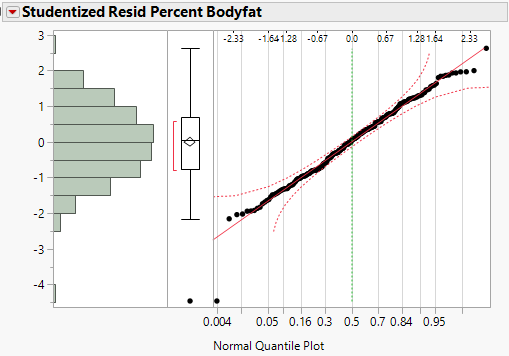
The plots in JMP use different scaling on the horizontal and vertical axes but the scatterplot and trend exhibited are the same as those constructed through the process outlined above. The relative important of each term in the model can be determined by comparing these plots to one another.

**Model Summary:**



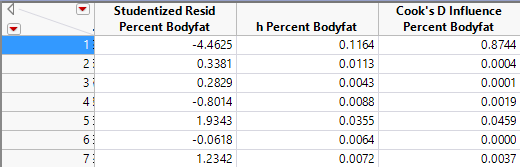
**Discussion:**

**Residual Plots:**

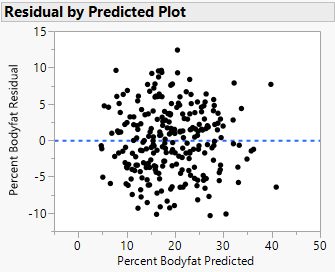
 

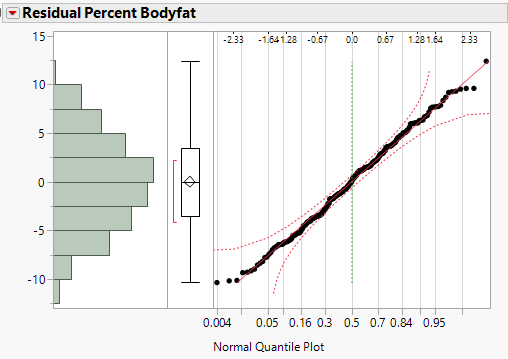
Obs. #1

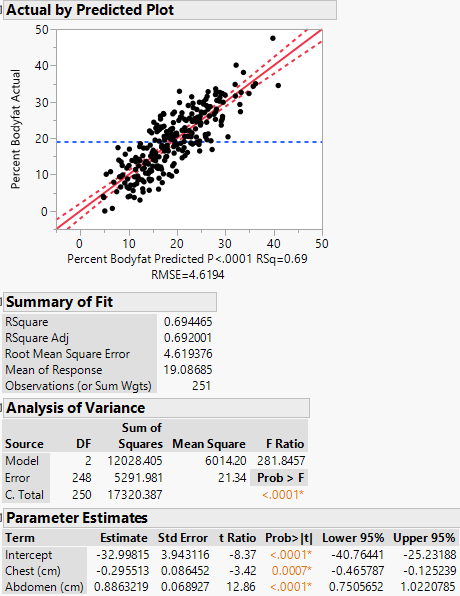
**Leverage, Outliers, and Influence**

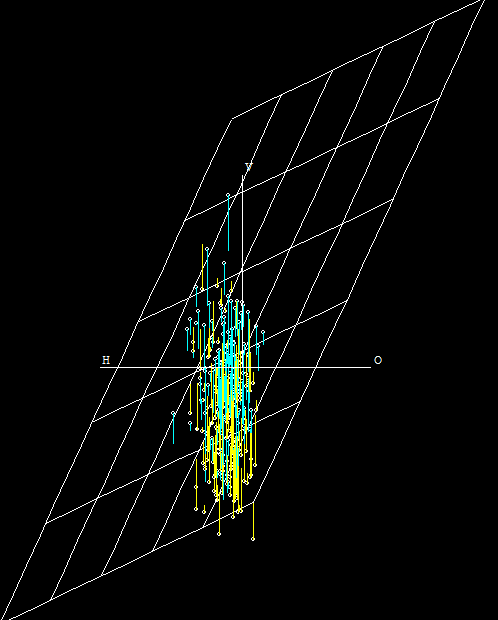
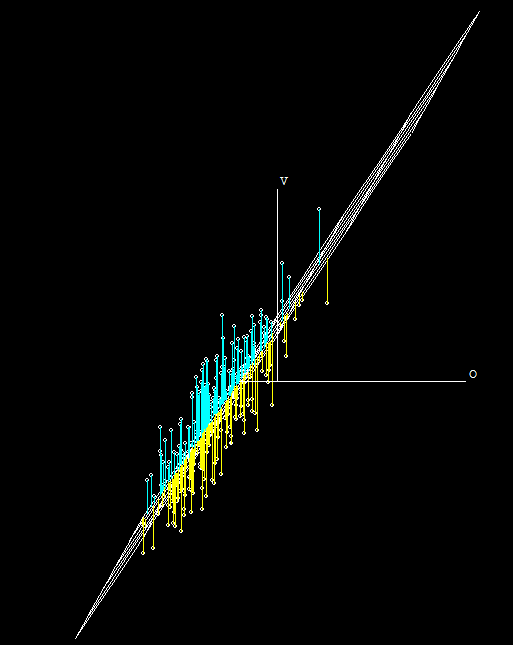
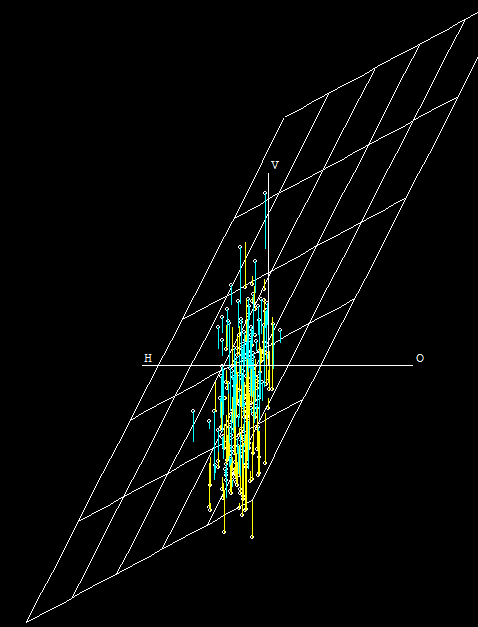


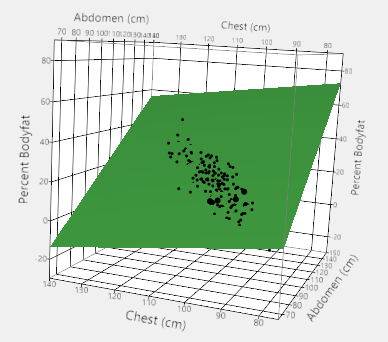
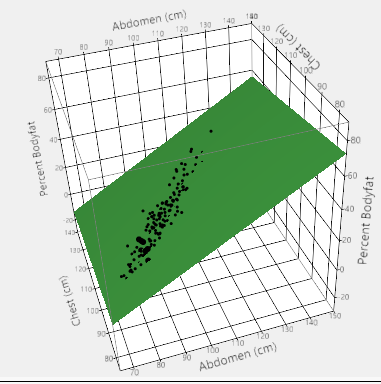
**Model Summary without Obs. #1 –** investigating the effects of an influential observation







In the next section we examine multiple regression models in more detail.