**13.1 – Case Diagnostics: Potential, Influence, and Outliers**

In section 7 we examined case diagnostics for simple linear regression, namely potential/leverage , influence , and outliers . Unusual cases on were fairly easy to identify in a simple linear regression setting by simply examining a scatterplot of the response vs. the predictor. In multiple regression the measures used are the same but identifying these cases by visual inspection (e.g. a scatterplot matrix) is not always easy. We review these measures and consider a multiple regression example below.

**Review of MLR in Matrix Notation:**The form of the multiple regression model is given by

and typically we assume .

In matrix notation,

where,

in our regression model and the are the observed values of the *jth* term. As before, the OLS estimates of the parameters are found using matrices as:

The **residuals** are given by .

**Leverage or Potential –** *Case diagnostic measuring the potential for influence.*  
One can show that the matrix results above imply that the following important result for case diagnostics.

This may seem like a violation of the assumption that , but keep in mind that the residuals ( are actually estimated quantities like the parameter estimates ( , etc. Thus they have estimated variances as do the parameters. The diagonal elements of the hat matrix (for short) is called the ***potential*** or ***leverage*** value for the case. One can show that,

2. The mean of the leverage values is:

where # of terms/parameters in the model.   
  
Note: for the model where the mean leverage value is .

Thus the closer the leverage value for the case is to 1, the closer the variance of the residual associated with it is to zero.

Rule of Thumb: Leverage values () exceeding twice the average () should be considered noteworthy. However, when is large it is unlikely any single observation will exceed this cutoff.

**“Standardized/Studentized” Residuals and Outliers –** *Poorly fit cases*The value of residuals ( are **scale dependent**, for example if the response is in millions are residual may be in the hundreds of thousands or millions, while if the response is a decimal value the residual may be in the hundredths. Thus putting the residuals in standardized scale would allow us to more easily determine if a residual is extreme (i.e. if the case is a potential outlier).

Because and the sum (and hence the mean) of the residuals is ALWAYS zero a standardized residual (think *z-score*) is given by:

The are called the ***standardized residuals*** (though they are referred to as studentized residuals in JMP). If the errors are normally distributed then we would expect approximately 95% the observations to have and 99.7% to have . Any observations with beyond these ranges are potentially outlying. HOWEVER, this is form of the standardized residual is not satisfactory for identifying outlier.

Because outliers will inflate our estimate of the error standard deviation (RMSE) they can potentially mask themselves. Thus a better statistic/residual for identifying outliers is the ***studentized residual*** or ***externally studentized residual***  which is given by,

Another expression for the studentized residual () is:

Because the case is excluded when calculating the estimated error standard deviation ( the outlier can inflate the estimate and subsequently mask itself. One can show that the studentized residuals have a t-distribution with , where # of terms/parameters in the model. The NH and AH being tested by this t-statistic are as follows:

🡨   
  
This is called the *mean shift model*. The mean shift model says that for the case we need a special shift in the mean function. To test vs. for the case we use

We can then find the p-value for testing if the case is an outlier and compare it to the Bonferroni Corrected significance level ().

**Measures of Influence – Cook’s Distance and DFBETAS**A case is considered influential if the results of a regression change appreciably when it’s exclusion from the analysis produces markedly different results.   
  
***Cook’s Distance*** measures influence by measuring changes in the fitted values when the case is deleted. As the fitted values are determined by the parameter estimates we can view these as changes in the parameter estimates when the case is deleted. The following two formulations of Cook’s Distance ( are equivalent:

There is no significance test based upon Cook’s Distance () but there are some general guidelines for identifying cases with a high degree of influence. It is generally useful to investigate cases where and you should always investigate cases where . For large datasets it is unlikely any case with exceeding these guidelines. For that reason some suggest using a size-adjusted cut-off of

A more “intuitive” formulation of the Cook’s Distance for the case is given in terms of the standardized residual and the leverage/potential value (.

This formulation allows to better understand what has to happen in order for a case to have unduly high influence on the results of a regression.

There several other measures of influence that have been proposed and we will not discuss them all, but one that can be important to consider if interpretation of particular estimated coefficient in the model is . The measure of influence looks specifically at the standardized change in the estimated coefficient when the case is deleted,

where the notation indicates the case is deleted. Cutoffs for identifying points with undue influence on are as follows. Extremely influential cases will have , however for large datasets (i.e. large ) it is unlikely any influential case would exceed 1.0 thus we use the size-adjusted cutoff .

**Example 13.1 – Pharmacokinetic Study in Rats**

Response: = % of initial dose in the rat’s liver

Predictors:

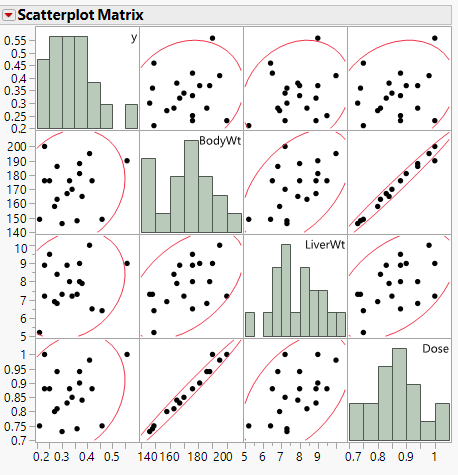
Body weight (g)

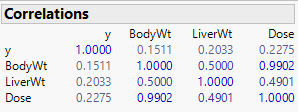
- Liver weight (g)

- Relative Dose, assigned as approximately 40 grams/kilogram of body   
 weight

The goal of the analysis is to see what percent of the drug remains in the rat’s liver. The research hypothesis is that there is no relationship between the response () and body weight, liver weight, and relative dose based on the method of determining the dose.

We begin by examining a scatterplot matrix of these data.

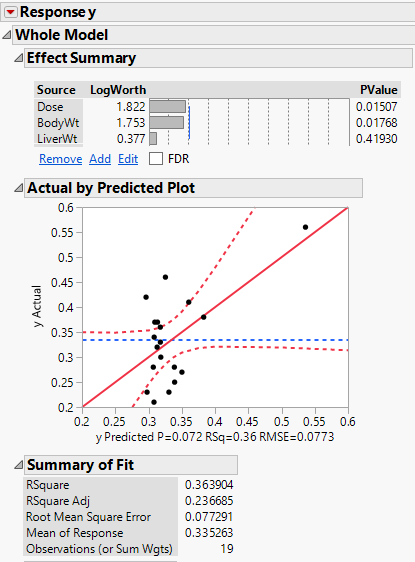


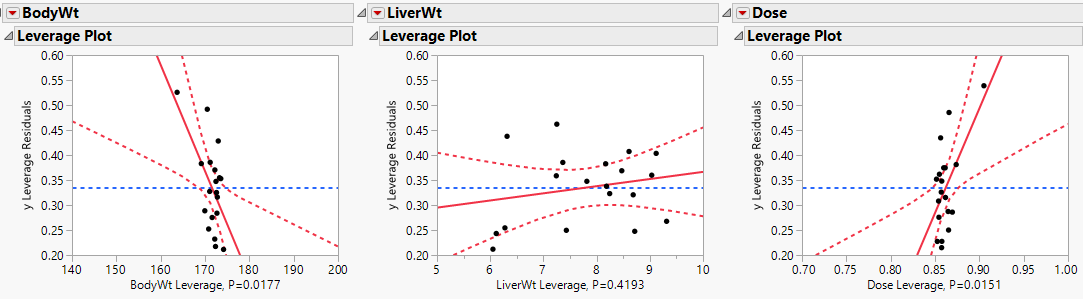


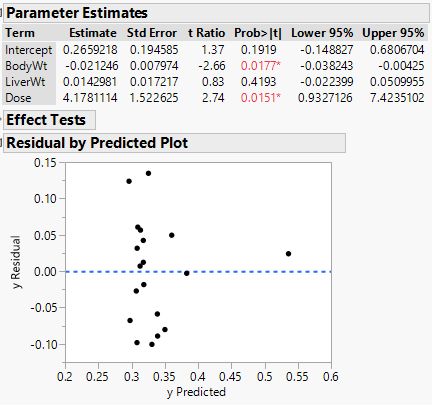
The correlations between the response and each of the predictors are weak. Dose and body weight are highly correlated as expected because the dose assigned is determined by the rat’s body weight.

We begin by fitting the MLR model

which is summarized below.



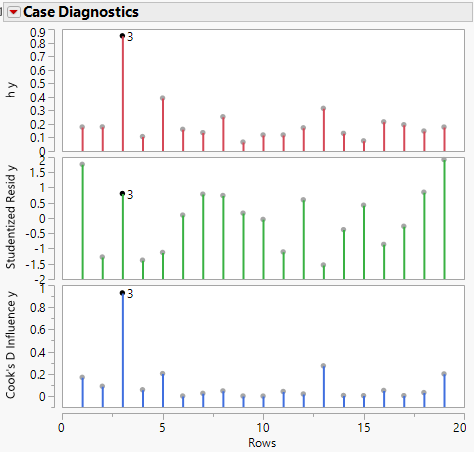




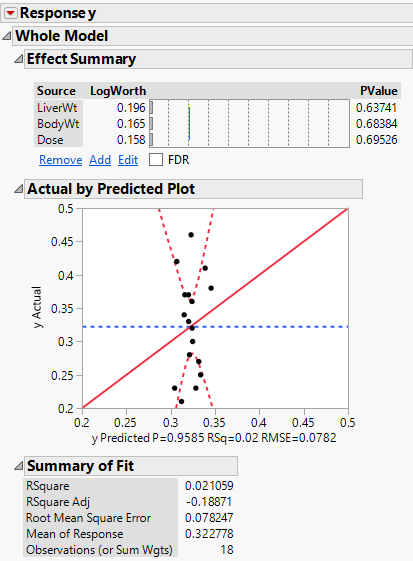
The fact two of the predictors/terms body weight and dose are statistically significant would suggest the research hypothesis is not supported by these data, i.e. it appears that both body weight and dose are related to the response. However, one case () seems to stand out from the rest in both a plot of the actual response values vs. the predicted values and the residual plot ().

We will examine plots the leverage/potential values (), standardized residuals (), and Cook’s Distance ). Again you save these quantities to the data table using **Save Columns** from the main drop-down menu labelled **Response y** for this model.

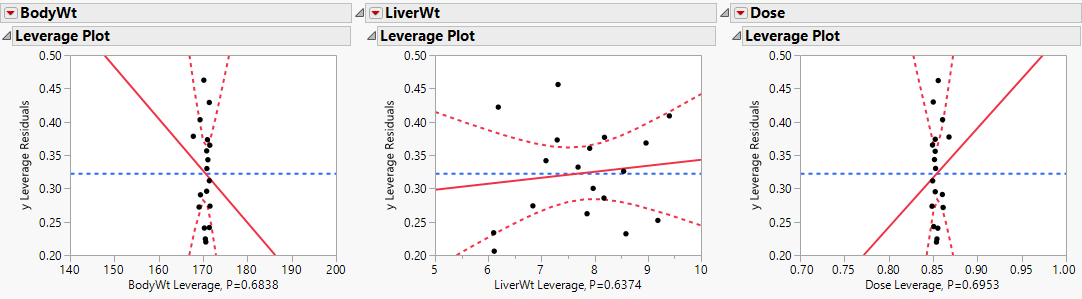
Use **Graph > Overlay Plot** to plot these diagnostics vs. case number with **Overlay Plots > No Overlay** and **Y Options > Needle** selected from the main drop-down menu.

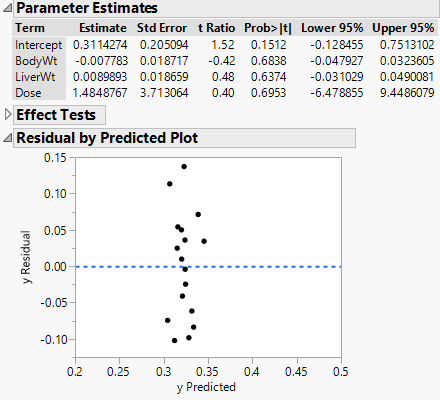


The Cook’s Distance for case #3 is nearly 1.0 and the leverage/potential value is very large as well. This case is clearly highly influential. With such cases it is a good idea to delete the point from the regression and investigate the changes.

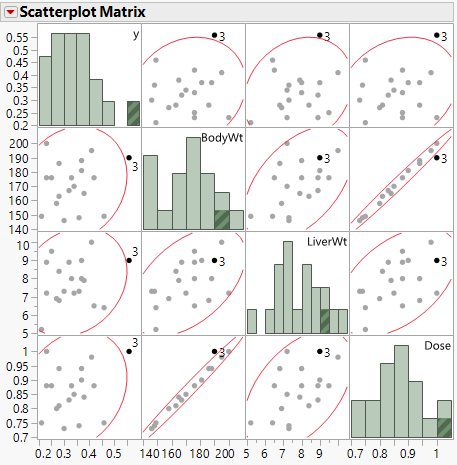


Without case #3 the research hypothesis appears to be confirmed as none of the terms are significant and .





Consider again the scatterplot matrix with the high influence case (#3) highlighted.



**Summary:**

Other heavier rats received the same dose as Rat #3. Was the Body Weight or Dose recorded in error? Rat #3 is not in error, but the hypothesized relationship used for assigning dose is not correct. Rat #3’s values “break up the correlation” between Dose and Body weight. We would need additional data, with dose assigned differently to sort out the models.

**Collinearity (or Multicollinearity)**  
When predictors/terms are highly correlated with one another or if predictors/terms are well explained in a regression sense (think ) by other terms/predictors in the model then standard errors of estimated quantities are inflated. This is called collinearity or multicollinearity.

Suppose we are fitting the model

and consider the term . One can show the standard error of from fitting the simple linear regression model is given by

whereas the standard error of in the multiple regression model above is

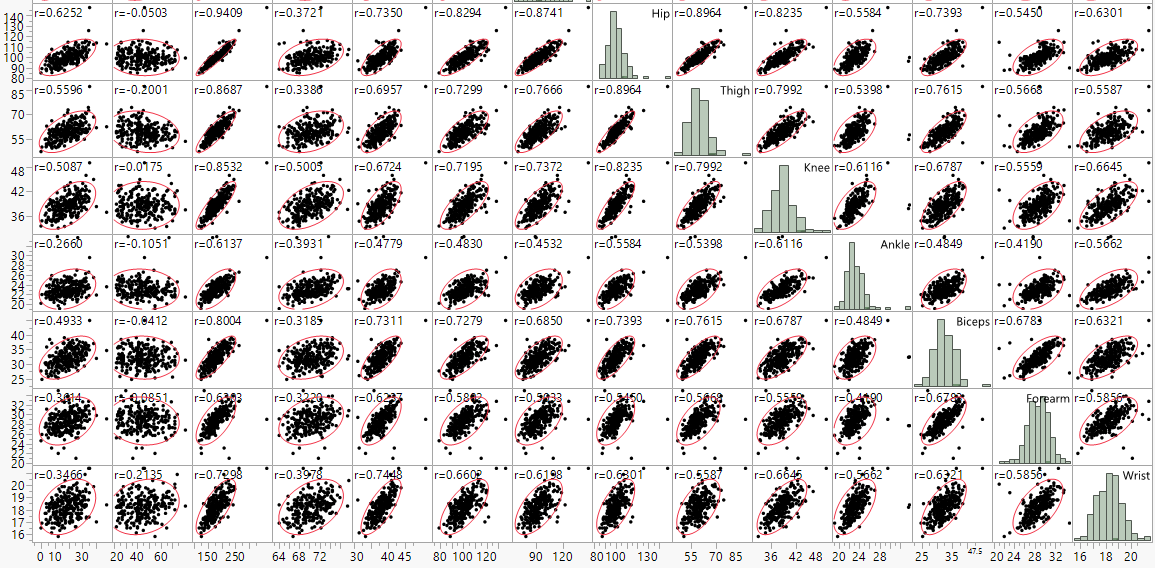
where the R-square from the regression of on the other terms in the model. When there is a 10-fold increase in the standard error. The multiplier of in the multiple regression model is called the Variance Inflation Factor (VIF), i.e.

When you have large VIF inflation factors (VIF > 10) it is generally a good idea to resolve them. Usually this means eliminating some terms from the model in which case the collinearity could be resolved. Another option is to collect more data which could reduce some of the high correlations between the terms. This is generally not an option and will only work in cases where the number of observations is small to begin with.

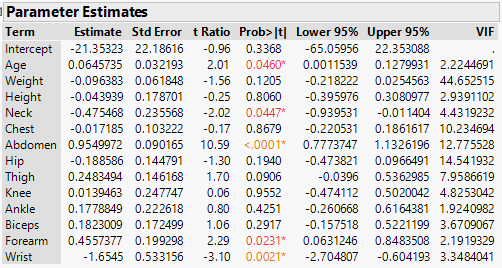
**Example 13.2 – Percent Body Fat Study**

Consider again the body fat study that we have used in numerous previous examples. A scatterplot matrix with pairwise correlations added for these data is shown below.

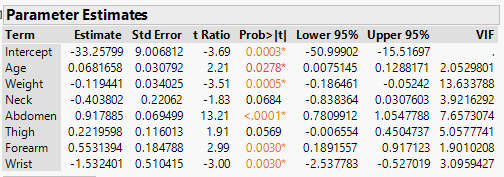




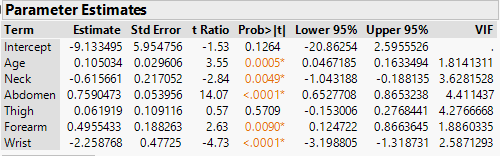
We can see that several of the variables are highly correlated, e.g. chest & abdominal circumferences () and hip & thigh (). Below are the parameter estimates for model using all available predictors as terms. Also added to the parameter estimates table are the CI’s for each parameter and the VIF for term.



Here we see that the VIF > 10 for Weight, Abdomen, Hip, and Chest. After performing backward elimination we arrive at the reduced model containing Age, Weight, Neck, Abdomen, Thigh, Forearm, and Wrist however the VIF for Weight is still above 10.



After dropping Weight from the model, the VIFs are all less than 10.

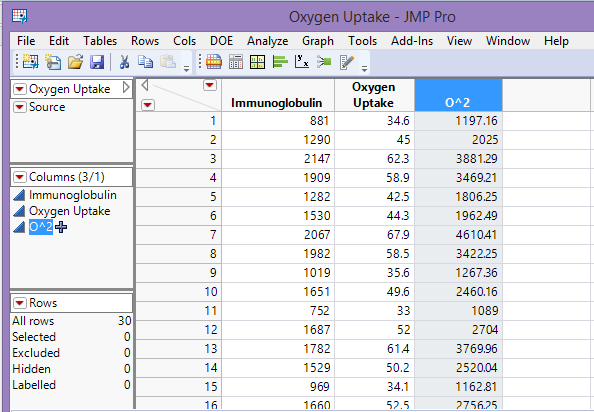
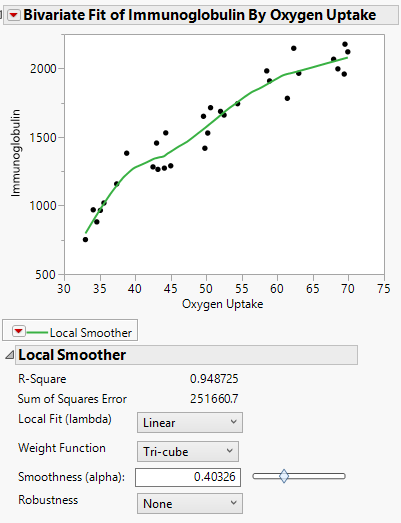


Another situation where collinearity can be an issue is when we added polynomial terms to a regression model as we will see in the example below.

**Example 13.3 – Immunoglobulin and Oxygen Uptake**

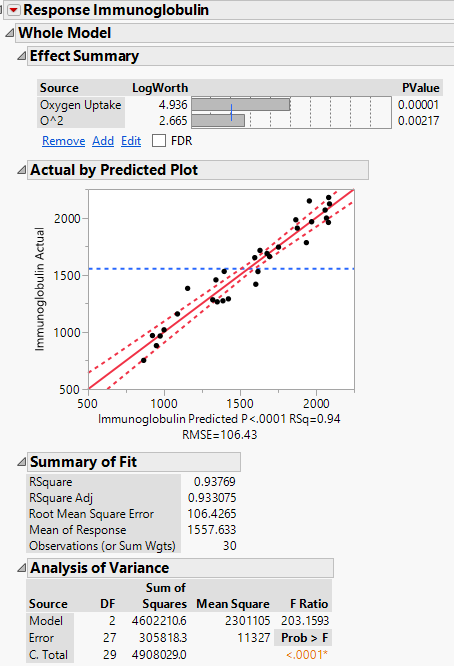
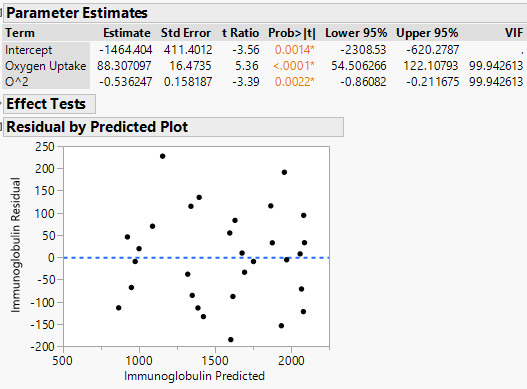
What is the impact of exercise on the human immune system? In order to answer this very global and general research question, one has to first quantify what "exercise" means and what "immunity" means. Of course, there are several ways of doing so. For example, we might quantify one's level of exercise by measuring his or her "maximal oxygen uptake." And, we might quantify the quality of one's immune system by measuring the amount of "immunoglobin in his or her blood." In doing so, the general research question is translated into the much more specific research question: "How is the amount of immunoglobin in blood related to maximal oxygen uptake ?"

Below is a scatterplot of immunoglobulin in blood vs. maximal oxygen uptake .

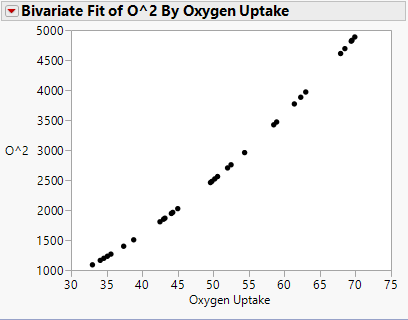


To address the curvature present in the scatterplot we might consider adding a square term to a simple linear regression of Immunoglobulin on Oxygen Uptake.

We will fit the quadratic model

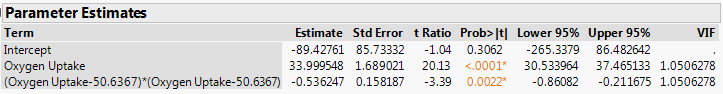
 

This appears to be a good model with a high and no evidence of assumption violations. However, the VIF for the estimated coefficients are both nearly 100!

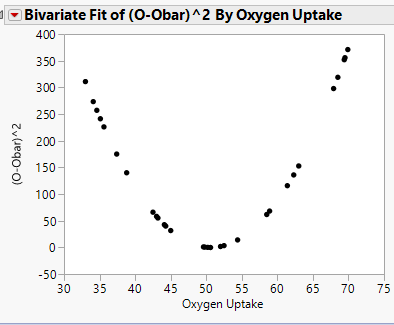
To see why the VIF’s for both terms are so large consider the scatterplot of the terms vs. .  


We can clearly see that even though we squared oxygen uptake there is a very strong correlation between the terms ().

If we fit the polynomial model with degree = 2 (i.e. quadratic) in JMP the results differ in that the squared term is mean centered as shown in parameter estimates below. Note the sample mean of oxygen uptake = 50.6367 units.



Below is a plot of mean centered square term used in the model above vs. oxygen uptake (). We can see that mean centering takes away the collinearity exhibited when mean centering is not used. This is one of the reasons that JMP mean centers polynomial terms.



**Summary:**Perfect or near perfect collinearity or multicollinearity will result in a situation where does not exist. This will also happen whenever the number of terms is greater than the number of observations (i.e. . In either situation, alternatives to OLS regression have been proposed to deal with these situations. We will discuss these methods in Section 17 of the notes.

Identifying collinearity when it exists can be an important part of the model development process, however model selection methods can alleviate VIF issues by eliminating redundant terms from the model even when we don’t initially notice the problems with variance inflation in our preliminary model.