MINI EXERCISE QUESTIONS

Open the course data set **Bodyfat.jmp** in JMP. (Note that a version of this file is available in the JMP Sample Data Library, but you will use the course file for this practice.) Recall that the data set consists of **%Fat** and other physical measurements from 252 men.  
  
In this exercise, you conduct a regression analysis for **%Fat** as a function of the other variables. You'll explore the variables, fit a full model, explore multicollinearity, conduct a residual analysis, check for influential values, and develop a reduced model for **%Fat**.  
  
Notice that the purpose of this exercise is to provide you with an opportunity to practice applying what you learned throughout this lesson. Show the solutions after each question to check your work. In a real-life modeling situation, you might make different decisions regarding the best course of action at each step.

1. Create histograms for all of the variables. Are there any missing values? Are there any unusual observations that should be hidden or excluded from the analysis?
2. Create a scatterplot for all of the variables.
   1. Is the response, **%Fat**, highly correlated with any of the predictors (with a correlation coefficient of |+/- 0.7| or higher)?
   2. Are any of the predictors highly correlated with one another? (There are many pairs of variables, so use the scatterplot matrix to generally assess correlations.)
   3. Are there any unusual patterns or observations in the scatterplot matrix?
3. Fit a full model, with **%Fat** as the response and all the other variables as predictors. Which two variables are the most significant in this model?
4. Use VIFs to investigate multicollinearity in this model. To requests VIFs, right-click the **Parameter Estimates** table and select **Columns**and then**VIF**.
   1. Overall, do the VIFs indicate that there is a problem with multicollinearity?
   2. For this exercise, remove **BMI** from the model. (Notice that a simplistic approach is taken for the purposes of this exercise.) What happens to the VIFs for the other variables after **BMI** is removed?
5. Look at the residual plots. Are there any unusual patterns or observations?
6. Save **Cook's D Influence** values to the data table and use the **Distribution** platform to graph the values. Are there any influential values that you should be concerned with?
7. The initial "Full" model has all the predictors except **BMI**.  
     
   Hint: For these questions, you might need to select the **Summary of Fit** and **Analysis of Variance** tables from the top red triangle under **Regression Reports**.
   1. Is this model significant?
   2. What are the RSquare Adjusted and RMSE for this full model?
8. Use the Effect Summary table to slowly remove nonsignificant terms from the model, and use a cutoff *p*-value of 0.05. Stop when the largest *p*-value that remains is <0.1.
   1. How many terms are in the reduced model?
   2. What are the RSquare Adjusted and RMSE for this reduced model?
   3. Compare the fit statistics (RSquare Adjusted and RMSE) between the full model and the reduced model. Which model performs better?
   4. Assume that you use one of these two models for predictive purposes. You take several physical measurements and use the model to predict a man's **%Fat**. From a practical perspective, which model would you rather use, the larger full model or the reduced model? Why?