Statistics 740/840, Design of Experiments I, Fall 2019

# Home Work Assignment #6, Due 11/25/2019

This assignment pertains to **Definitive Screening Designs**; you should have read the notes on DSDs and watched the assigned videos on DSDs before attempting the assignment.

**Questions** **2 and 3** of this assignment are based on the dataset **FermentationProcess.JMP**. The data is based a definitive screening design that was used to optimize the fermentation step of a bio-manufacturing process. The actual factor settings have been replaced with the generic settings and the response is simulated based on the results of the original experiment. The fermentation step is used to grow bacteria and induce the bacteria to produce a therapeutic molecule. The molecule is extracted from the bacteria cells and the yield measured in mg/L. The Column Info window for each factor contains explanatory notes. The experiment consists of 5 experimental factors: **pH** of the fermentation solution; **%DO**, the target dissolved oxygen once molecule induction is initiated; **Induction Temperature**, the solution temperature at induction; **Induction OD600**, the microbial mass as measured by optical density, when induction is initiated; and **Feed Rate**, during the induction phase the rate of glucose feed to the reactor. The response is the **Yield** of the molecule in mg/L.

1. (10 pts) Define the following terms.  
     
   a) Supersaturated design.

**A design that utilized sometimes fewer then half as many runs as factors. It makes it an attractive design to not assume some factors won’t affect the resultant without any data on the choice.**  
  
b) Full Quadratic Model.

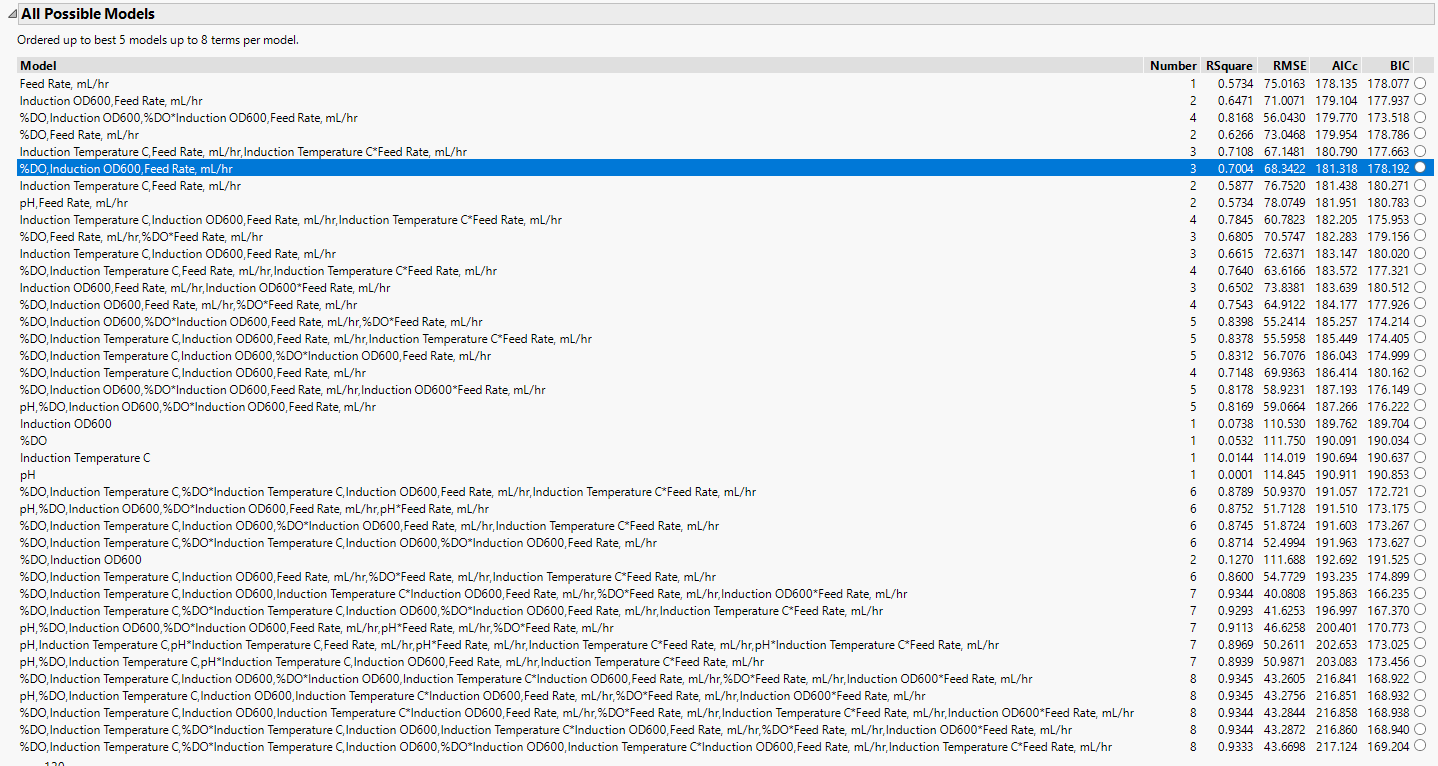
**The basis for optimization and Prediction. Used as a second order approximation to the behavior of a physical system.**  
  
c) Partial Alias.

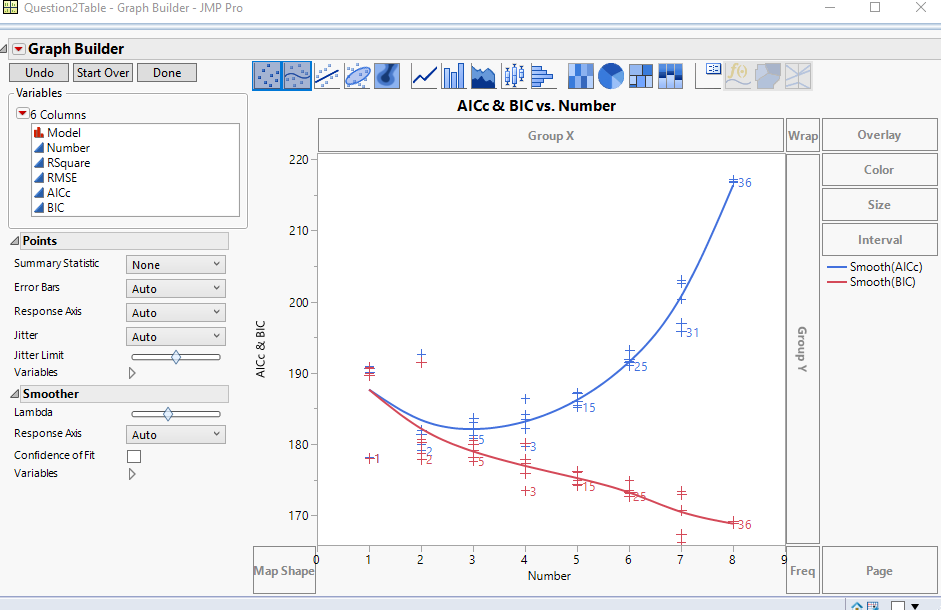
**When fundamental affects on the resultant are neither orthogonal nor fully aliased/confounded. So, what this means is that the affects are not 0 or 1, but somewhere in the middle.**   
  
d) I-optimality

**Integrated Mean Square Error optimality. Used for when prediction is of paramount concern. Great because a I-optimal RSM design can use far fewer runs than a traditional RSM design. It provides the most precise model predictions over the region studied.**  
e) D-optimality.

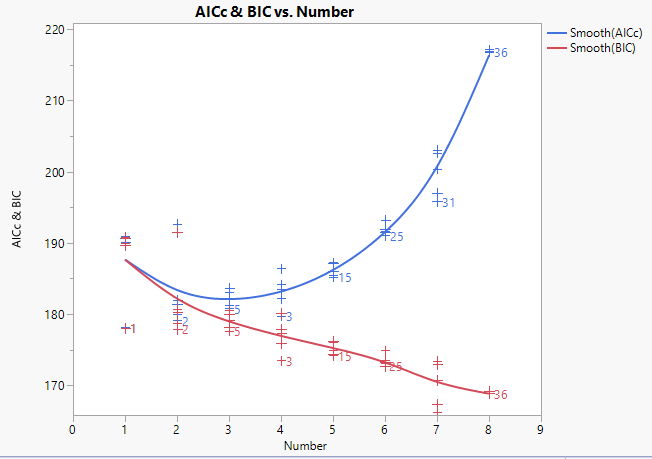
**Perform bad for prediction, but has great D-efficiency. Find a best design in the sense that the precision in the estimates of the model coefficients is maximized relative to a specified model and total number of runs. Find most precise estimates of experimental effects**

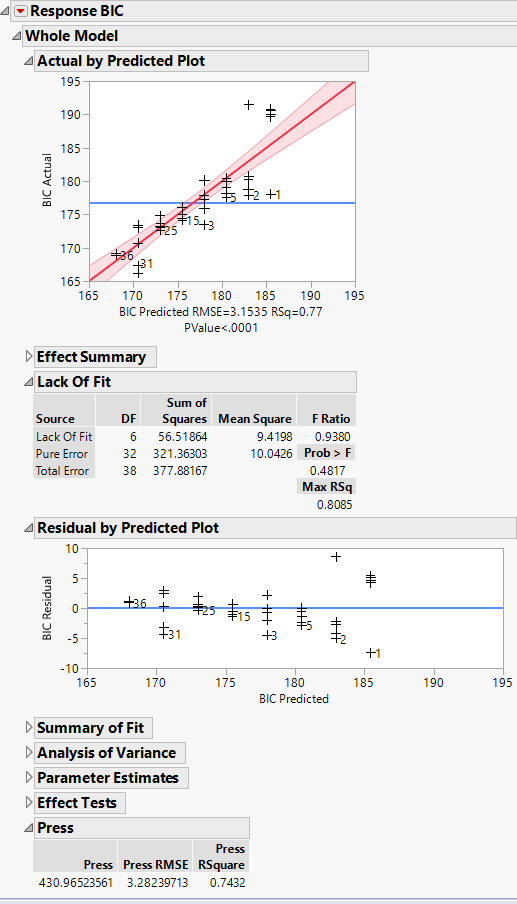
**Instructions for Question 2:** Go to the **Fit Model** platform and define a full quadratic model as you have been shown during lectures and in the screening design notes. Next, select the **Stepwise** fitting personality. Once in the Stepwise I want you to use **All Possible Models** regression to find a set of three possible best models for further analysis. Use the following heuristic to find the set of candidate models – we have used this approach repeatedly during lectures and in the Screening Design notes.  
  
a. Select the **All Possible Models** options in the **Stepwise** platform, set the largest possible model to 8, select the **Heredity Restriction** option, and set the number of models to display to 5.  
  
b. Sort the All Possible Models output in ascending order by AICc.  
  
c. Make the All Possible Models output into a data table.  
  
d. Use **Graph Builder** to create an overlay plot of AICc and BIC vs Number (X axis). Also, as shown in class move BIC to the right Y axis to make the plot more interpretable.

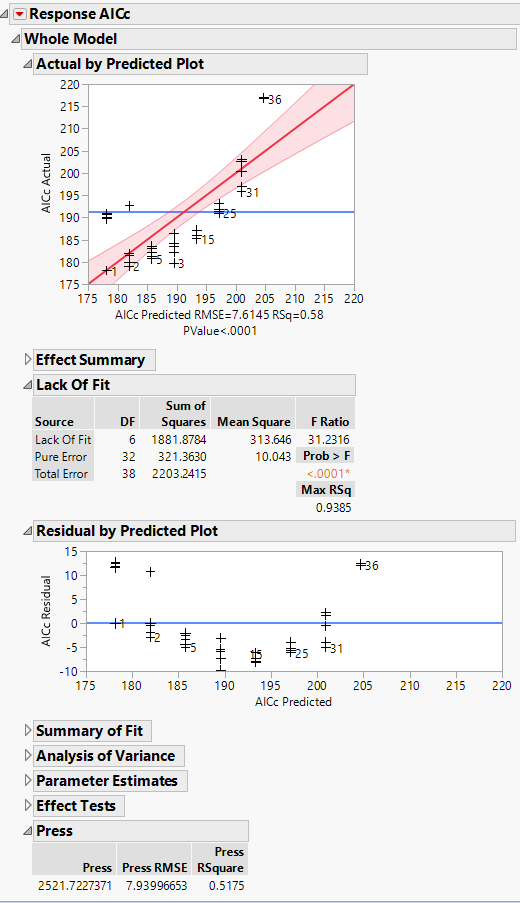


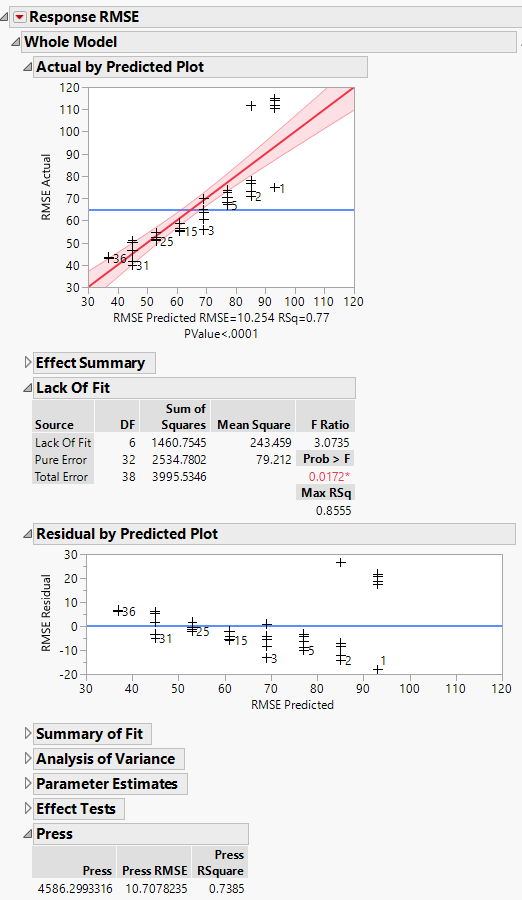


1. (20pts) Using the **Graph Builder** overlay plot select three models to examine further. You should use a combination of AICc, BIC, and RMSE to narrow the selection to three models. Fit each of the three models in the Stepwise platform to form a **Fit Group**. Compare these three models based on **Press**, **Actual by Predicted, Residual** plots, and **lack of fit**. Finally, save the prediction formulas for the three models to the data table. In your answer, you must show the Fit Model reports for each of the three models. Clearly explain why you chose these final models. Note, since this is a simulation there exists an actual best model, which you should be able to find using the steps you have been given.

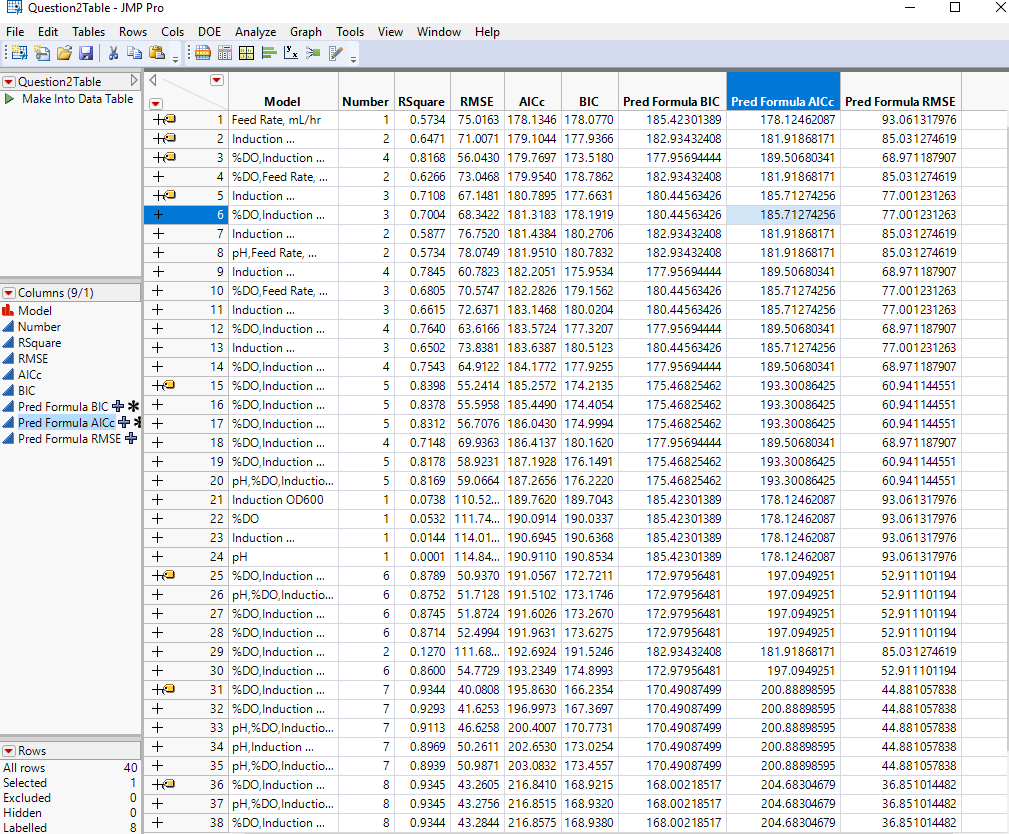




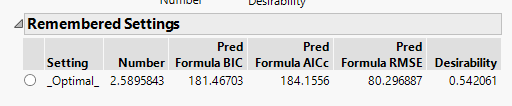




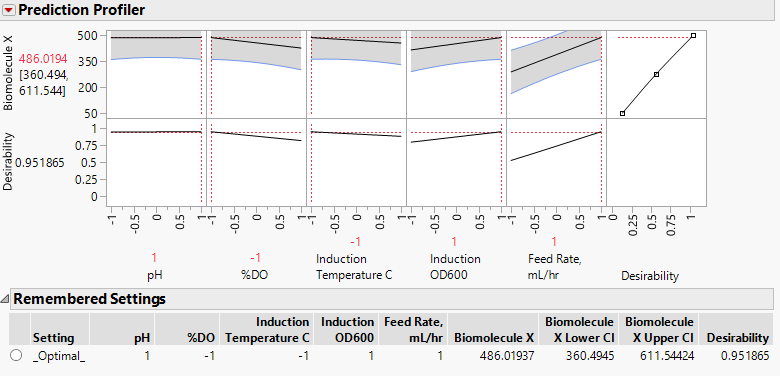
The actual by predicted plots show a strong linear relation with points scattered randomly along it. With that, there are no systematic large deviation among any of the three models. From the lack of fit table , a p-value would show us that there is a significant lack of fit. Because each of the models have a larger p-value (>.10), so there seems to be no lack of fit. The residual by predicted plot show also that the points sit close together along the mean line with the RMSE residual values plotted with the predicted from the software. The Press statistics are all fairly large for each of the models, which isn’t ideal as a low press stat means that there are small prediction errors.



Setting were adjusted and a desirability to maximize the output was given to output the desired values of the predicted models, showing that the AICc overall is the best model to use as it outputs the highest value.



**Instructions for Question 3**: Recall, in the videos on DSDs I showed how to use the Model Comparison platform in JMP Pro (**Analyze 🡪 Modeling 🡪 Model Comparison**). Navigate to the Platform; in the launch window select the three prediction formula columns of the data table as the **Y, (Predictors)**. In the Report window, click on the red arrow to open the main menu and from the menu select **Model Averaging**; this creates a new prediction column in the data table, which is just an average of your three prediction formula columns. At this point you are now ready for question four of this assignment.

1. (10 pts) The goal of the experiment is to maximize yield of the biomolecule. Open the Profiler platform (**Graph 🡪 Profiler**) and in the launch window enter your **Model Average** column as the Predictor, next be certain to select the **Expanded Intermediate Formulas** option just below the **Select Columns** window. Once in the Profiler report window, use the **Desirability Functions** to find the settings of the factors that result in the highest average yield. What are the settings of the factors resulting in the highest predicted yield? What is the average predicted yield at those settings? Be certain to include a screenshot of your Prediction Profiler results.  
     
     
   

The optimal settings to maximize Biomolecule X yielded 486 mg/L, and includes the pH, Induction and Feedrate at the high setting, while the %DO, Induction Temperature are set to their low settings.