Final Data Analysis Assignment
Due 5/3/2010 by 5pm in 244 box in 211 Old Chem.

You may ask questions of the TA or Prof; otherwise all work should be your own.

A chemical engineering experiment was run to study heat transfer in a shallow fluidized bed. Data were collected on the following four candidate predictors: \( X_1 \), fluidizing gas flow rate in pounds per hour; \( X_2 \), supernatant gas flow rate in pounds per hour; \( X_3 \), supernatant gas inlet nozzle opening in millimeters; \( X_4 \), supernatant gas inlet temperature, \(^\circ\)F. The measured responses are \( Y_1 \), heat transfer coefficient, and \( Y_2 \) thermal efficiency. Twenty observations were gathered with the data at http://www.stat.duke/courses/Spring10/sta244/exams/chem.dat

Using various methods discussed in class develop model(s) for predicting the thermal efficiency, \( Y_2 \) using the predictor variables \( X_j \), \( j = 1, \ldots, 4 \). (You may need to investigate transformations of the response and consider quadratic and interaction terms, and check that model assumptions are appropriate – don’t forget about EDA at all stages of model building/fitting/checking). Are there outliers/influential points? Which approach(es) do you recommend? Which variables are important? Do the different methods tend to agree? How do changes in the variables affect thermal efficiency (give appropriate estimates and measures of uncertainty with interpretations).

Suppose that for a new run of the experiment, we are only given that \( x_1 = 116.9 \) and \( x_2 = 172.1 \). Construct a predictive distribution for the thermal efficiency using the available data and estimate the probability that the thermal efficiency exceeds 50.

Please turn in a typed solution that addresses the questions above (some suggested steps are given below). Provide an “executive summary” of the results in a manner interpretable by a chemical engineer interested in the study conclusions that 1) addresses which variables are important and how they impact thermal efficiency and 2) provides the estimated probability for the new run. You should have at most 2 pages of text and tables summarizing the analysis, with up to an additional 4 pages of figures. You should include the R-code in an appendix (suggestion for the energetic student: explore Sweave for preparing your solution) Please be rigorous in providing full justification for each of your answers, including all relevant statistical details, calculations and results. Approaches that are dead-ends do not need to be discussed.

Suggested things to consider in your analysis/write-up.

1. Conduct an exploratory data analysis. What does this suggest about relationships among the variables? Need for transformations? Interactions? (recall lattice plots or coplots)

2. Fit the full model using main effects, interactions and quadratic terms in all of the predictors and carry out a residual analysis. Does this, added variable plots or the
EDA above suggest that you should transform any variables? Are there any outliers or influential points? Are there any indications of problems due to multicollinearity?

3. Use a classical model selection method (forward, backward, or stepwise, best Cp) to see if you could identify a simpler model. (see help(step)) for R implementations. What is the final model? Briefly interpret results.

4. Use the BAS package (or the BMA package) to explore the posterior distributions over all models (use the Zellner-Siow prior, method="ZS-null" for BAS or the g-prior with g=n). What does this suggest about the most important models or most important variables? Is there substantial model uncertainty?

5. Implement at least one of the shrinkage methods (ridge, lasso, Bayesian lasso or horseshoe). Summarize your findings and compare to the frequentist model selection and Bayesian Model averaging.

6. For each of the approaches above (full model, stepwise model selection, BMA, and the shrinkage methods if you implemented that) carry out a cross-validation exercise, where you fit using all data but the $i$th case to obtain an out-of-sample prediction of the $i$th case, $\hat{Y}_{(i)}$. Do this for each of case, $i = 1, \ldots, n$ and compute CV residual $Y_i - \hat{Y}_i$. Compute the cross-validation MSE$^1$

$$\text{CV-MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_{(i)})^2$$

Also create side-by-side boxplots of the CV residuals for each method, and a scatterplot matrix of residuals under each method. The predict.bma function in BAS is similar to the predict function for linear models and will give predictions under BMA and all other models; the variable “best” in the output can be used to identify the predictions for the highest probability model.

How do the methods compare?

7. Should you construct a model for the missing $X_3$ and $X_4$ or use a model without them? If you drop $X_3$ and $X_4$ will your predictions be biased? If you decide to model the missing data, make sure that you account for uncertainty in their values (recall missing data from Sta290)

---

$^1$This is leave-one-out cross validation, which often selects more complex models, a better approach is to do leave $k$-out CV.