# FINAL EXAM

## DIRECTIONS AND POLICIES:

- Final is due on Wednesday, December 10, 3:00 pm. As a submission mechanism drop the exam off with Gina Martin in the Statistics main office in Eckhart 108 anytime during regular business hours before the due date and time.

- Exam is to be done individually. Evidence of joint work may result in forfeiting some or all of the points in the exam, as we already did at times for the homework.

- When asked to use a method whose class is defined but is otherwise unspecified use only methods covered in class.

- When asked to carry out a calculation in R attach a brief summary of R output to your answer in addition to other items specified by the question, but otherwise do not attach R code, unless asked to do so.

- Please show your work and justify your answers by R output or explanations, as appropriate. In particular explain what method/routing you have used. **However there is no need for the answers to consist of more than a few sentences.**

- If the answer to one of the points exceeds 1/3 of a page, please provide a 3-4 line summary of your answer, and mark it clearly ("Summary answer to point ZZZ")

1. Consider the airfoil data set whose data table is provided in chalk. The column names are: "frequency","angle","chordLength","velocity","thickness","pressure". The data set is also in the UCI machine learning repo if you need more information about it.

a) (5 points) Fit a linear model of pressure on all other variables. Diagnose it to decide whether you see problems *other than nonlinearity* with it. In particular decide whether the output and the variables appear to be in in need of transformation. Run the transformed model and decide whether things have improved.
b) (5 points) With your best model at (a) decide whether it appears that you observe nonlinearity in the fit. If so, decide how to fix this nonlinearity in the context of the classical linear model and decide whether the changes are significant.
c) (6 points) Discuss pluses and minuses of the models at part a) and b), decide which is the best one in your opinion and how you would explain or deal with its imperfections.
d) (8 points) To fix the issues at a) and b) attempt now to fit a tree model. Read the manual of "rpart" and decide what seems to be a good choice for the complexity parameter and why. With this complexity parameter fit the best model, and diagnose its residuals.
e) (6 points) Discuss tree models versus linear model in general and for this example (e.g discuss their pluses and minuses and whether you see evidence of that here).

Problem 2. Consider the CCH.csv data set that is in chalk. It denotes the power output from a combined cycle heating unit as a function of the ambient conditions. The file "ReadmeCCH.txt" in chalk gives additional details about the problem and meaning of the variables. The file is originally from the UCI machine learning repository if you feel the need to get even more detail.

a) (8 points) Fit a linear model of PE on all other variables (temperature, humidity, etc). Do the proper diagnostics that indicate whether it seems that
   i. the normality assumption is holding
   ii. are there any outliers
   iii. (iii) do you see any nonlinearity?

b) (16 points) Focus now on deciding whether there are some outliers under the normality assumption. Design a sharp compute-intensive test to identify the worst q outliers by means of
   i. the Cook Distances and
   ii. (ii) the externally studentized residuals.

   Implement it for q <=10. Here by sharp I mean that under the proper assumptions the probability that the q-th worst outlier exceeds the test-value that you will compute is exactly the probability level used to compute the said test value. Find some graphical way to illustrate how the test value depends on the p-value. Decide now at p-value of 0.01, 0.05 and 0.1 whether the point with the q-th most extreme value for the statistics above is in fact an outlier, for 1 <=q <=10.

c) (6 points) Looking at the data, would you conclude that the departure from the standard assumptions can be best represented by declaring a number of points as outliers, or can you think of a more satisfactory answers? Can you imagine an outside-the-problem hypothesis that explains this departure? What information would you like to have to decide the issue?

3. Consider the data set "NIR" in the R package 'chemometrics'. We aim to model the glucose content as a function of the NIR absorbance values in the various spectrum ranges provided in the data set.

a) (3 points) Which of the variable selection or shrinking procedures learned in class seem suitable for this situation of n<p? Why?

In the following questions, consider a setup where we permute the observations by using the seed "1968", and then we use the first 126 values for training and the last 40 values for testing.

b) (5 points) Fit a principal component regression model, the latter using both a scree plot and a K-fold crossvalidation approach to select the number of components.
c) (5 points) Fit now a partial least squares model, with a number of components chosen by a K-fold crossvalidation approach, with K the same as at point b.
d) (6 points) For PCR and PLS redo the calculations using now a leave-one-out crossvalidation approach. Are the estimates of the error better or worse than the K-fold CV? What would you have expected?
e) (7 points) Fit now a LASSO model with the complexity parameter tuned by crossvalidation. If you find somewhere a routine that does that – which you do not implement yourselves – describe its working principle. Which of the 3: LASSO, PCR and PLS do you prefer?
f) (4 points) Discuss the theoretical advantages and disadvantages of LOO and K-fold CV relative to each other. Focus on ability to compute, accuracy, and determinism, justify your argument.