Homework 4 STAT 8210 Spring 2015

Problem 1

Use the data in table B11 (the quality of Pinot Noir wine data) to build regression models for this problem. The response variable is *quality*.

a) Begin with all predictors except *region*. Use forward, backward and stepwise regression to find the "best" model. For each method, indicate what variables are selected to be in each model and what your criteria for entry and removal were. **6 pts.**

b) Use all possible subsets approach and choose the two models with best values of Mallows Cp statistic.

- What variables are included in each of the two best models chosen using the Cp criteria? **4 pts**
- Do the models meet Mallows criteria? Hockings? **2 pts.**
- Fit both of these models and give the equation for each. Comment on what you see as far as significance of the variables in each. **6 pts.**
- Is there much difference in their prediction ability (PRESS)? **2 pts**

c) Incorporate indicator variables for *region* and find the model selected as best using the Cp criteria? Fit this model. Compare it with the models found in part b). (Adjusted R-squared value, Cp, Press, etc.) Is there an indication that including the region information substantially improves the regression? **8 pts.**

Problem 2

Do problem 13.5 a-d in your text.

a**) 2 points b) 2 pts c) 4 pts d)4 pts**

Problem 3 No Collaboration on this problem. Treat as a take-home exam question.

A logistic regression model is fit to predict whether or not a bank note is counterfeit. For 200 notes, the value of the response **y** is 1 if the note is a counterfeit and 0 if it is not a counterfeit. Although there were several possible predictors in the original data set, we focus on only two: **top**, which is the width of the note's top margin in mm; and **right** which is the note's right edge width in mm. **The output contains results on two logistic model fits.**

$$\text{(M1): log-odds} = \beta_0 + \beta_1 * \text{right} + \beta_2 * \text{top}$$

$$\text{(M2): log-odds} = \beta_0 + \beta_1 * \text{right} + \beta_2 * \text{top} + \beta_3 * \text{right} * \text{top}$$

a) Use the output to write down the fitted model for the log odds that a bank note is counterfeit if t**op** and **right** are the only predictors and the interaction is not included. **1 pt.**

b) Is there any indication of a lack of fit (inadequate model) for M1? Briefly explain. **2 pts.**

Regardless of your answer to part b, assume for the remainder of question 6 that the model M1 was OK.

c) Interpret in the context of this problem the meaning of β_1 in M1. **2 pts.**

d) Which model should be preferred: M1 or M2? Justify your choice as convincingly as you can. For full points, you will need to use the most straightforward justification. (Note: the choice can be convincingly justified without writing a lot.) **1 pt.**

e) Using your preferred model, give an estimate of the probability that a bank note is counterfeit if the right edge width is 130 mm and the top margin is 10 mm. **2 pts.**

f) Suppose a bank note has estimated **odds of 0.5** (under Model 1) of being counterfeit (based on its values for right and top). If the right edge measurement had gone up by 1 mm but the top measurement is unchanged, what would the new estimated odds (under model 1) of the note being counterfeit now be? **1 pt**

g) Suppose that we are using **Model 2**, and the estimated odds of a particular note being counterfeit are 0.5 (based on its values for right and top). We **can't tell** what the new estimated odds would be if the right edge measurement had gone up by 1 mm (and the top measurement were unchanged) without also knowing an **additional piece of information** (this is a piece of information we didn't need for part f.) What is the information we need, and why do we need it for model 2 but not for model 1? **1 pt**