Stat 203

Practice Take Home Final Exam.

[**Note to 2019 students:** Instructions for this year's final will likely change slightly as solutions will now be uploaded.]

Instructions: There are 4 problems of equal points value, please submit each problem in hard copy on separate sheet(s) and put your name on each sheet. The datasets mentioned below are in the directory `Final` on the course website.

In writing solutions to the data problems, **please follow the conventions** in the `Sample Solution` posted in the `Final` directory: include in your answer only the pieces of output (and plots) needed for your answer, and collect the `R` code you used (without output) at the end of the answer. Include only the plots and output that are necessary for your analysis. Poorly organized presentation and/or errors in additional plots/analyses may be penalized. For each question, explain clearly (i) your objectives, (ii) any hypotheses that you are testing, (iii) the statistical procedure, and (iv) the statistical support for your conclusions.

Honor Code: Please respect the honor code in completing this exam. You can use books and computers, but not other people or their solutions.

1. [25 pts.] Suppose that $N = r + 2s$ observations are to be taken as follows: r observations with mean λ, along with s observations with mean $\lambda - \mu$ and then s further observations with mean $\mu - \lambda$. The N observations are assumed to have errors which are independent with mean 0 and variance σ^2.

(a) Write down the standard linear model that describes these observations. What are the least squares estimates $\hat{\lambda}$ and $\hat{\mu}$, and the estimate of σ^2?

(b) Evaluate the entries in the covariance matrix of the least squares estimates, and also evaluate $\mathrm{Var}(\hat{\lambda} - \hat{\mu})$.

(c) Suppose that the total number of observations N is fixed. How would you choose r and s so that the larger of $\mathrm{Var}(\hat{\lambda})$ and $\mathrm{Var}(\hat{\mu} - \hat{\lambda})$ is as small as possible?

2. [25 pts.] The dataset `temps.txt` records Fahrenheit normal body temperatures and heart rate, in beats per minute, of 65 males, coded by 1, and 65 females, coded by 2.

(a) for males and females separately, make scatter plots of heart rate versus body temperature and show the linear least squares regression lines. Comment on possible presence or absence of relationships.

(b) Use an F test to assess if the two regression lines are identical.

(c) Use an appropriate permutation test to test the same hypothesis.

(d) test whether the slopes are the same but the intercepts are unequal. [No need to do a permutation test here.]

[P.T.O.]

3. [25 pts.] The file brca.txt contains 668 cases of potentially cancerous tumors, of which 235 are in fact malignant. Variable definitions are in brca.info. The traditional method for determining if a tumor is malignant is by surgery. The aim of this study was to learn if a new method called fine needle aspiration (which draws only a small tissue sample) could be effective in determining tumor status and prognosis.

(a) Fit a logistic regression with Class as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Based on material discussed in the course, can this information be used to determine if this model fits the data? Explain.

(b) Starting from the model in (a), use AIC as the criterion to determine the "best" subset of variables (use the step function).

(c) Use the reduced model to predict the outcome for a new patient with predictor variables 1, 1, 3, 2, 1, 1, 4, 1, 1 (in the same order as in the dataset). Give a confidence interval for the prediction.

(d) Suppose that a cancer is classified as benign if $p > 0.5$ and malignant if $p < 0.5$. [Here p is the estimated probability of being benign]. Compute the number of classification errors of both types that will be made if this method is applied to the current data with the reduced model.

(e) Suppose we change the cutoff to 0.9, so that $p < 0.9$ is classified as malignant. Compute the number of errors in this case. Comment on the issues in determining the cutoff.

(f) Split the data into two parts: assign every third observation to a test set, and the remaining two thirds of the data to a training set. Use the training set to determine the model, and use the test set to select the cutoff (e.g. to minimize the sum of errors of both types in the test set over a grid of possible values for the cutoff).

4. [25 pts.] The dataset in dairy.txt was extracted from a Canadian record book. Random samples of 10 mature (five-year-old and older) and 10 two-year-old cows were taken from each of 5 breeds. Information about the variables is in dairy.info. The average butterfat percentages of these cows were recorded, yielding a total of 100 butterfat percentages, broken down into 5 breeds and into 2 age classes (see dairy.info for details of coding of the variables).

Analyze and discuss your results.

[While this is deliberately formulated in an open-ended way, it is not necessary to go into more detail than is found in the parts of the lecture notes that you deem to be relevant].