STA244 **4/23/2003**

Take Home Final Exam

This is an open book/note exam. All work must be your own; you may neither give or accept help from others. Please cite appropriate results from the text or lecture where necessary and clearly define notation that you adapt – remember I am not a mind reader! Your solution for Part I may be neatly handwritten (but, if I cannot read it, I cannot grade it). You have a total of 48 hours from the time you start to complete the exam, but the latest the exam may be turned in for credit is by noon on 4/29/2003. If you have any questions please email me.

Part I

Consider the linear model with response $Y = (y_1, \dots y_n)^T$ and $n \times p$ design matrix X (which may include a column of ones for the intercept), with

$$Y \sim N(X\beta, \sigma^2 D)$$

where D is a known $n \times n$ diagonal matrix with elements $d_i > 0$, or equivalently,

$$Y_i \overset{ind}{\sim} N(x_i^T\beta, \sigma^2 d_i)$$

with x_i^T equal to the ith row of X.

1. Show that the maximum likelihood estimate of β is $(X^T D^{-1} X)^{-1} X^T D^{-1} Y$. (you may assume that X is full rank).

2. Show that the MLE above is equal to the ordinary least squares estimator in the model

$$z_i = u_i^T \beta + e_i$$

   where $z_i = y_i/\sqrt{d_i}$ and $u_i = d_i^{-1/2} x_i$, and $E(e_i) = 0$ and $\text{Var}(e_i) = \sigma^2$.

3. Weighted least squares finds the value of β that minimizes

$$\sum w_i (y_i - x_i^T \beta)^2$$

   What choice of w_i makes weighted least squares and the maximum likelihood estimate above equivalent?

4. Find the MLE of σ^2. Find its expected value. Is it unbiased? If not, construct an unbiased estimate of σ^2.

In R and S-Plus weighted least squares can be carried out using `lm(Y ~ X, weights = w)` *where* `w` *is the vector of weights. Programs such as* `leaps` *and boxcox also utilize weights. (If you instead choose to implement weighted least squares by the transformation approach in part (2), be careful of how you treat the intercept!) Use these results in Part II.*

Part II

Largemouth bass were studied in 53 different Florida lakes to examine the factors that influence the level of mercury contamination. Water samples were collected from the surface of the middle of each lake in August 1990 and then again in March 1991. The pH level, the amount of chlorophyll, calcium, and alkalinity were measured in each sample. The average of the August and March values were used in the analysis. Next, a sample of fish was taken from each lake with sample sizes ranging from 4 to 44 fish. The age of each fish and mercury concentration in the muscle tissue was measured. (Note: Since fish absorb mercury over time, older fish will tend to have higher concentrations). Thus, to make a fair comparison of the fish in different lakes, the investigators used a regression estimate of the expected mercury concentration in a three year old fish as the standardized value for each lake. Finally, in 10 of the 53 lakes, the age of the individual fish could not be determined and the average mercury concentration of the sampled fish was used instead of the standardized value. The smallest level of mercury concentration that the measuring instrument can detect is 40 parts per billion. Any level below that was set to 40 parts per billion.

The URL is `http://www.isds.duke.edu/courses/Spring04/sta244/final/bass.dat`. Variable names in the file are

- ID: ID number
- Lake: Name of the lake
- Alkalinity: Alkalinity (mg/L as Calcium Carbonate)
- pH: pH
- Calcium: Calcium (mg/l)
- Chlorophyll: Chlorophyll (mg/l)
- Avg.Mercury: Average mercury concentration (parts per million) in the muscle tissue of the fish sampled from that lake
- No.samples: How many fish were sampled from the lake
- min: Minimum mercury concentration amongst the sampled fish
- max: Maximum mercury concentration amongst the sampled fish
- Standard.mercury : Regression estimate of the mercury concentration in a 3 year old fish from the lake (or = Avg Mercury when age data was not available)
- age.ind: Indicator of the availability of age data on fish sampled

Your job is to develop a model for predicting mercury concentrations in three year old fish, addressing issues discussed below. The write up of your analysis should be typed with a maximum of 5 pages including any figures and tables; additional material may be included in an appendix, i.e. calculations or derivations used to support conclusions made in the write up. Your report should be in narrative form rather than addressing point-by-point the questions below. Please provide a brief introduction to the problem, your solution, conclusions and recommendations. You do not need to describe all intermediate models or unsuccessful directions, just make sure that you provide enough details so that someone could replicate your results. As always, make sure that all notation is clearly defined.

1. Give a non-technical explanation of why weighted regression is more appropriate for this problem, explaining what the weights mean.

2. Is the regression surface the same for the lakes with missing age data? Explain what method you use, giving any test statistics, distributions and conclusions. Should you include the lakes with missing age data in your analysis?

3. What are the most important variables (if any) for predicting mercury concentration? Do you need to transform the data?

4. Are there any outliers or influential observations?

5. (For the appendix) Using a non-informative prior distribution, $p(\beta, \phi) \propto 1/\phi$, where ϕ is $1/\sigma^2$ or the inverse of the error variance and β are the regression coefficients in your linear model, find the joint predictive distribution of standardized mercury concentrations for the 53 lakes (make sure your results are in the original units!). Describe how you could simulate values from this joint distribution using `rnorm` and `rchisq`, etc ... in R or using another language. Explain how you could calculate (approximately using simulated data) expected posterior rankings of the lakes. Implement this approach in your favorite language, with the code in an appendix). *In R refer to the functions* `sort, sort.list, rank, order`. In the body of your write-up, explain why this approach for ranking will not necessarily lead to the same order as with the observed standardized mercury value in the data set.

6. Florida has set a standard of 1/2 part per million as the unsafe level of mercury concentration in edible foods. Based on this sample, which lakes would be considered unsafe? Which is the most relevant for addressing the question of safety: thresholds based on predictions of expected mercury concentration, confidence intervals for expected mercury concentration, prediction intervals for mercury concentration, or some other measure? Explain.

7. Provide a table with (1) ranking of lakes, (2) any associated statistics used in developing the ranking, (3) the criterion by which you judged the lakes to be unsafe, and (4) the observed standardized mercury value (and of course lake names). Please make sure that you have explained the approaches used to develop the table somewhere in the write up (either the body or appendix).

8. Are there any lakes where the measured value is considered safe, but under your analysis would declare them unsafe? Can the converse occur (i.e. measured value is unsafe, but your model would classify the lake as safe)? Explain how this can occur. What are your recommendations to the state in these cases?

9. In your summary, briefly address any limitations of your analysis and recommendations for how to improve the study.