STA244 **4/23/2003**

Take Home Final Exam

Due 5/1/2003 by 5pm

This is an open note/open book test. All work must be your own.

Study of the growth of plants can be a crucial element in understanding how they compete for resources. For example, soybean varieties with large early leaves may be desirable in no-till farming, because larger leaves can shade the ground and inhibit weed growth (competition). In a study comparing soy bean varieties on leaf growth, it was necessary to develop an inexpensive method for predicting leaf area from measurements that can be taken without removing leaves from the plant. The data consist of $n = 145$ leaves of the soybean variety Heifeng-25. The goal is to predict *Area* which is measured using a destructive method in the laboratory, using non-destructive measures of leaves that are easy to obtain. The variables in the data file consist of Year, DAP (days after planting), Length (length of terminal leaflet in cm), Width (width of terminal leaflet in cm) and Area (actual area of terminal leaflet in cm^2). The URL is `http://www.isds.duke.edu/courses/Spring03/sta244/final/final.asc`. Use this data to answer the following questions. Please provide neat solutions and justify your work where needed!

Write a summary of your analyses (5 pages max) and discuss which model you would prefer and why and what it implies about leaf growth. Try to address the following points in your analysis.

1. Construct scatter plots of all variables. Briefly describe the information in the plots and what they suggest about modeling the data.

2. A model suggested by approximating the area of leaflets by a rectangle suggests the following mean in the log scale:

$$\mu_i = E(\log Area_i | Width_i, Length_i, \eta) = \eta_0 + \eta_1 \log Length_i + \eta_2 \log Width_i \qquad (1)$$

   Fit this model to the data. Provide a brief interpretation of the results of the F-test in the summary analysis of variance table and individual t-tests. Do the assumptions for regression/testing seem satisfactory? Are there any outliers/influential cases?

3. Conduct an F-test for the hypothesis that the area of a leaf can be well approximated by a rectangle, $H_o : \mu_i = \log(Length)_i + log(Width)_i$ versus the model given in (1) and carry it out. Briefly summarize your conclusions. *Hint: reparameterize the model so that you have a test in the form $\eta = 0$ versus $\eta \neq 0$.*

4. Consider an alternative mean function, without transforming Area:

$$\theta_i = E(Area | Width, \beta) = \beta_0 + \beta_1 Width_i + \beta_2 Width_i^2 \qquad (2)$$

   Fit this model and interpret. Give a short justification of this model based on geometric arguments. Examine diagnostic/residual plots.

5. Based on informal plots/diagnostics, which model seems preferable (1) log-normal where log(Area)is normally distributed or (2) normal where Area is normally distributed?

6. Using Zellner's g-prior with g = n, $p(\beta|\phi, g) \sim N(0, g(X'X)^{-1}/phi)$ and $p(\phi) \propto 1/\phi$, where var(Area) = $1/\phi$ and X refers to the covariates in model (2), find and compute the marginal distribution of Area under the model in (2). Call this the Normal model marginal.

7. Using Zellner's g-prior with g = n, $p(\eta|\tau, g) \sim N(0, g(X'X)^{-1}/tau)$ and $p(\tau) \propto 1/\tau$, where $1/\tau$ is the variance of log(Area) and X refers to the covariates in model (1), find the marginal distribution of **Area**. Evaluate this for the leaf area data and call this the Log-Normal model marginal. *Hint: find the marginal under the normal assumption for log(Area) and then perform a change of variables*

8. Using equal prior probabilities on the models, find the posterior probabilities of the Normal model and the Log-Normal model given **Area**. Which would you select? Is this result consistent with any exploratory analyses regarding assumptions/significance that you have conducted?

9. Construct a prediction interval (classical or Bayesian) for **Area** for a leaf with Length = 7 cm and Width= 4 cm. Give a careful interpretation of this interval.

10. Neither of the above models include Year or DAP. Should either or both of these (or some transformations) of these be included in the model? Would this change any of your conclusions about the log-normal versus normal error distribution. Are there any other models that you would propose? Investigate and justify.