Statistics 135 – Midterm
Due: Wednesday, November 16, 2005

As with the assignments this term, please submit your answers typeset in some word processing package. And please include your **R** code for all your calculations.

The examination is open book, you may refer to the lecture notes, standard references (Venables and Ripley, Krause and Olson, etc), and so on. However you should not discuss this exam with anybody. Direct all questions to me (`irwin@stat.harvard.edu` or 617-495-5617 or drop by my office).

1. (35 points) In studying the breeding habits of the common puffin, 38 sites at Great Island in Newfoundland were studied. One measure the researchers were interested in was the nesting frequency (the number of burrows per $9m^2$) (`nesting`). Four variables thought to be useful predictors of nesting frequency are the percentage of grass cover (`grass`), the mean soil depth in cm (`soil`), the angle of slope in degrees (`angle`), and the distance from the cliff edge in metres (`distance`). One possible model for this dataset,

$$\text{nesting}_i = \beta_0 + \beta_1 \text{grass}_i + \beta_2 \text{soil}_i + \beta_3 \text{angle}_i + \beta_4 \text{distance}_i + \epsilon_i$$

   (a) (5 points) The data from this study is available on the datasets page in the file `puffin.txt`. Read in the data, calculate the standard summary statistics (mean, standard deviation, and 5 figure summary) for each of the variables and create a scatter plot matrix of the data. Does it appear that any of the potential predictor variables are associated with nesting frequency?

   (b) (5 points) Run the linear regression model for the above model and give the standard summaries (parameter estimates and standard errors, ANOVA table, etc). What evidence is there that some of the variables are useful in describing nesting frequency?

   (c) (5 points) Create the residual plot of the residuals against the fitted values. In addition, in a single figure, plot the residuals against each of the predictor variables. Do any of these figures suggest a problem with the regression model?

   (d) (5 points) Suppose that a new site was found where `grass` = 95, `soil` = 25, `angle` = 5, and `distance` = 60. Predict the number of nests for this site based on the original 38 sites. Any comments about this prediction?

   (e) (5 points) Calculate the F test for comparing the model with all four predictors in the model with the model having only `soil` and `distance` in the model. What does this F test imply about the predictors `nesting`?

   (f) (5 points) Now fit the model

$$\text{nesting}_i = \beta_0 + \beta_1 \text{soil}_i + \beta_2 \text{distance}_i + \beta_3 \text{soil}_i \times \text{distance}_i + \epsilon_i$$

   Is there any evidence of an interaction between distance and soil on the nesting frequency?

(g) (5 points) Now fit the model

$$\text{nesting}_i = \beta_0 + \beta_1 \text{soil}_i + \beta_2 \text{distance}_i + \beta_3 \text{soil}_i^2 + \beta_4 \text{distance}_i^2 + \epsilon_i$$

Is there any evidence of a nonlinearity in the relationship of distance or soil on the nesting frequency?

2. (30 points) The Ryan-Joiner test considers the following hypotheses

$$H_0 : \{x_1, x_2, \ldots, x_n\} \text{ is a random sample from a normal population}$$
$$H_A : \{x_1, x_2, \ldots, x_n\} \text{ is not a random sample from a normal population}$$

The test statistic is r, the correlation coefficient of the coordinate pairs,

$$\left(\Phi^{-1}\left(\frac{i - 0.375}{n + 0.25} \right), x_{(i)} \right), \qquad i = 1, \ldots, n$$

of a normal Q-Q plot, where $\Phi^{-1}(\cdot)$ denotes the inverse CDF of a standard normal RV and $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ are the ordered data values. Under H_0, r should be close to 1.

(a) (5 points) For the dataset `douglasfir.txt`, available on the datasets page for the course, calculate in **R**, the value of the statistic r described above.

(b) (5 points) Create the Q-Q plot (aka Normal Scores plot) for this dataset.

(c) (5 points) The function `qqline` will add a straight line to Q-Q plots created by either `qqnorm` or `qqplot`. This line is a description of the main trend in the plot. By examining the code for `qqline`, what line is added to the `qqnorm` plot (i.e. when `datax = FALSE`).

(d) (10 points) The distribution of r under H_0 is complicated, thus getting the p-value for this test statistic requires some work. One way of determining it is by simulation. One scheme for approximating the p-value for this statistic is the following

i. For $j = 1, \ldots, m$ generate $\{x_1^{(j)}, \ldots, x_n^{(j)}\} \overset{iid}{\sim} N(0, 1)$. (i.e. generate m datasets under H_0.)

ii. For each dataset j, calculate r_j, the value of the Ryan-Joiner test statistics.

iii. Approximate the p-value by

$$\hat{p} = \frac{1}{m} \sum_{j=1}^{m} I(r_j \leq r)$$

where r is the value of the Ryan-Joiner statistic for the dataset in question.

Create a function in **R** to implement this procedure. The inputs to the function should be a vector containing the data and m, which should have a default value of 1000. The function should return 3 values, $\hat{p}$, r, and m. Evaluate your function with the Douglas fir dataset for $m = 1000$ and $m = 10000$.

(e) (5 points) Based on the above analysis, is there any evidence to suggest that the Douglas fir data is not normally distributed?