Statistics 135 – Final Exam
Due: Friday, January 13, 2006

As with the assignments this term, please submit your answers typeset in some word processing package. And please include your **R** or **SAS** 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).

Please turn your completed exam to my mailbox on the 7th floor of the Science Center by 5:00 p.m. on Friday,

1. (??? points) In this question, we will analyze data from a study involving 51 women (Brumback and Rice, 1998). The dataset, `pdg.txt`, contains a subset of the data for 50 women in the study; the file `missing.txt` contains data for the remaining woman. It is thought that the pregnancy outcome of either **conception** or **non-conception**, is related to the daily levels of a hormone known as pregnanediol-3-glucuronide (PDG). In this study, the PDG level was measured on the natural log scale from 8 days pre-ovulation (day $= -8$) to 15 days post-ovulation (day $= 15$), using a urinary hormone assay. The columns of both datasets are:

   - `Group`: outcome of the pregnancy (0 = conception, 1 = non-conception)
   - `Subject`: subject number
   - `Day`: day in a subject's cycle, ranging from -8 to 15 (8 days pre-ovulation to 15 days post-ovulation)
   - `PDG`: level of the PDG hormone

   Write **SAS** programs to carry out the following tasks

   (a) (5 points) Read in the data, creating a dataset `PDG`. In this data set create a new variable `GroupLabel` which has value `conception` or `nonconception` depending on the variable `Group`. Then create a new dataset `AfterDay5` which contains the subset of the data corresponding only to days 5 to 15 post-ovulation. As part of this new dataset, add the variable $\log(PDG)$ stored in `logPDG` and get the dataset in the form necessary to do part (b).

   (b) (10 points) Using the dataset `AfterDay5`, fit a regression model to **each subject in each group**, that predicts $\log(PDG)$ level using the day, post-ovulation. Store the results of these regression models in a new dataset `RegAfterDay5`.

   Hint: in the `PROC REG` command, you will need to use the `BY` command, along with the `OUTEST` option (see the **SAS** help for more details).

(c) (10 points) Create a new dataset from `RegAfterDay5`, keeping only the group, subject, intercept, and slope variables (This is an important step, otherwise you cannot carry out statistical inference). Then using high-resolution graphics (not line-printer graphics) and numerical summaries, describe the distributions of the slope and intercept parameter estimates obtained for each subject in each group.

(d) (5 points) Test whether or not the slope and intercept parameters are the same for the conception and non-conception groups. Check the assumptions of the statistical tests you carry out.

(e) (5 points) The one subject in the file `missing.txt` has a missing group value. Download the data for this subject, and predict whether of not this woman conceived using the $\log(PDG)$ values. Explain in detail how you made the prediction.

Note: While more formal procedures can be used to derive a classification rule for forecasting whether a conception occurred or not, a close approximation to these analyzes can be found using the simple exploratory data analysis techniques in this example. While the more advanced procedure will lead to a satisfactory answer to this question, these simple techniques are all that are required to answer this question.

2. (25 points) The dataset `ethanol.txt` contains data from an experiment where ethanol fuel was burned in a single-cylinder engine. For various settings of the engine compression (`C`) and equivalence ratio (`E`), the emissions of nitrogen oxides were recorded.

   For this question, please do all analysis in **SAS**.

   (a) (5 points) Read in the data file and create a dataset `ethanol`. Create high resolution scatter plots plotting `NOx` against `C` and `E`. Do these plots suggest that either `C` or `E` will be useful predictors of `NOx`?

   (b) (5 points) Fit a linear regression model predicting `NOx` with linear and quadratic effects of `C` and `E` and the `C*E` interaction. Is there any evidence of an interaction between `C` and `E` on the response of `NOx`? Is there any evidence of a non-linearity of the effect of `C` or `E` on `NOx`?

   (c) (5 points) Fit a generalized additive model predicting `NOx` by a linear effect in `C` and a smoothing spline in `E` where the amount of smoothing is selected by generalized cross validation.

   (d) (5 points) Fit a generalized additive model predicting `NOx` by a smoothing spline in `C` and `E` where the amount of smoothing is fixed by `C` having 2 degrees of freedom and `E` having 5 degrees of freedom.

   (e) (5 points) Based on all these analyzes, is there any evidence that the effect of `C` on `NOx` is non-linear? What about `E`?

3. (20 points) In the situations where standard distributional results may not hold, the bootstrap can be used to get properties of an estimator, such as the bias, or confidence intervals for a parameter. This question will examine one approach to calculating confidence intervals for the median of a distribution using **R**.

   For a data vector $x = (x_1, x_2, \ldots, x_n)$ denote the sample median by $M(x)$.

   The bootstrap involves generating B bootstrap samples based on the observed data vector x. For $b = 1, \ldots, B$, the bth bootstrap sample

   $$x^{*b} = (x_1^{*b}, x_2^{*b}, \ldots, x_n^{*b})$$

   is generated by sampling, **with replacement**, n items from the original data $x = (x_1, x_2, \ldots, x_n)$. So if $n = 5$, two possible bootstrap samples are

   $$x^{*1} = (x_2, x_4, x_1, x_2, x_5)$$
   $$x^{*2} = (x_3, x_2, x_2, x_4, x_2)$$

   Note that repeated observations are to be expected.

   Then for each bootstrap sample, calculate the sample median

   $$m_b = M(x^{*b})$$

   To calculate the percentile interval (one bootstrap approach to confidence intervals), is to determine the interval

   $$(m_l, m_u) = (m_{(\alpha)}, m_{(1-\alpha)})$$

   where $m_{(\alpha)}$ is the αth quantile of $m_1, \ldots, m_B$. So if $\alpha = 0.05$ and $B = 2000$, you need the 100th and 900th ordered values of the m_b's. If B is large enough, (m_l, m_u) is an approximate $100(1 - 2\alpha)\%$ confidence interval for the population median.

   (a) (15 points) Write a function `medCI` in **R** to implement this procedure. The input arguments to this function are to be the data vector `x`, the number of bootstrap samples `B`, and the confidence level of the interval `conflevel`. Set default levels for `B` at 1000 and `conflevel` at 0.95. The output of the function is the confidence interval, returned as a vector of length 2. In writing this function, check to make sure that the input values of `conflevel` and `B` are valid (0 < `conflevel` < 1 and `B` is a positive integer).

   Hint: the bootstrap samples can be generated using the function `sample`. See the help page or other documentation sources for further assistance.

   (b) (5 points) The dataset `law82.txt` contains average LSAT and GPA scores for incoming students for 82 law schools. Read in the data and for both variables calculate 5 confidence intervals using your function for confidence levels 0.9 and 0.95 and B = 1000 and 2000 (for a total of 20 intervals for each variable). Does changing the confidence level and the number of bootstrap samples act like you would expect it to?

   While it shouldn't be needed to answer the question, for further information about the bootstrap, see the file `bootstrap.pdf` available on the course web site. This contains lecture notes from Statistics 221 on the bootstrap. It also includes other references about the bootstrap.