STA 2201S 2014 Assignment 3. *due Friday, March 21 at the beginning of class*

When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. DO NOT include in this summary printouts of computer code with the relevant selections highlighted. All computer code used to obtain the results summarized in the response should be provided as an appendix. In this appendix you may highlight the relevant results.

1. Data on the Winter Olympic winning times for various events are available at Fact-Monster, and doubtless other places. For most events this web site stops at 2002. Pick an event of interest to you, for which there is data from 1924, and fit a linear and a nonlinear model to the winning times, as a function of year, from 1924 to 2002. Use this model to predict the winning times in 2006, 2010 and 2014, along with prediction intervals, and compare these predictions to the actual winning times in these three years. You can use the nonlinear model I presented in class (Feb 7), or one of your own devising.

2. SM: Exercise 10.7.3. By writing $\Sigma\{y_j - \hat{g}(x_j)\}^2 = (y - \hat{g})^{\mathrm{T}}(y - \hat{g})$ and recalling that $y = g + \epsilon$ and $\hat{g} = Sy$, where S is a smoothing matrix, show that

$$E\left[\sum_{j=1}^{n}\{y_j - \hat{g}(x_j)\}^2\right] = \sigma^2(n - 2\nu_1 + \nu_2) + g^{\mathrm{T}}(I - S)^{\mathrm{T}}(I - S)g.$$

   Hence explain the use of $s^2(h)$ as an estimator of σ^2. Under what circumstances it is unbiased?

3. The article "An estimate of the science-wise false discovery rate and application to the top medical literature" by Jager & Leek (*Biostatistics*, 2014), is posted on the course web page and available via the link in (c). In this paper they attempted to estimate the rate of false discoveries in papers published in leading medical journals.

   (a) Construct a 2×2 table with "Null hypothesis true/false" as the two column headings, and "Discovery/No Discovery" as the two row headings. Give a definition (algebraic) of the false discovery rate as a function of the entries in your table.

   (b) What model did Jager & Leek use for the distribution of p-values?

   (c) Their conclusion was that the rate of false discoveries among published results was 14% with an estimated standard error of 1%. How was the standard error estimated?

   (d) There were several discussants of this paper, and all the discussions can be found at `biostatistics.oxfordjournals.org/content/15/1.toc`. Choose one discussion and summarize in a paragraph the main point of the discussant. Comment briefly on this point, and on the reply by Jager & Leek.