Bio 754: Final Project

May 6, 2011

See the class site for due date, guidance on how to answer, and notice of any corrections or clarifications.

Problem statement:

Your instructor lives in downtown Annapolis, Maryland, and will be attending the Joint Statistical Meeting (JSM) in Miami, Florida from July 30-Aug 4th 2011. He will travel from his home to Miami and back for the conference. He doesn't mind leaving Annapolis a few days early or leaving Miami a few days late, but once he starts traveling he wants to get there fast. He also doesn't like it when his flight is cancelled or delayed. Your final project is to choose which flight, on which carrier, at what time, Jeff should fly to Miami and back to minimize headaches and travel time.

Jeff will pick his JSM flight based on the most convincing analysis (with some leeway depending on price, availability, and his wife's preferences).

The data

The **flights** data on the class website contains information on all commercial flights between the Annapolis area and the Miami area in 2007 and 2008. The Annapolis area airports are Reagan National (DCA), Baltimore-Washington International (BWI), and Dulles International Airport (IAD). The Miami area airports are Miami International Airport (MIA) and Fort Lauderdale-Hollywood International Airport (FLL).

This data is abstracted from a *much* larger (and ridiculously interesting) data set on commercial flights available from `http://stat-computing.org/dataexpo/2009/`. You will be able to find the variable descriptions, etc. on that website. You may use the rest of the flight data if you like, but you don't have to.[1] You may also collect additional data from outside sources if you would like to augment the flights data, although this is not required.

[1] If you decide to use the whole data set you should (1) probably run your analysis on the cluster, (2) talk to Marvin about appropriate safety precautions when dealing with large data sets (in particular the use of mem_free, h_vmem, and h_fsize when running jobs on the cluster). You might also want to look into the `bigmemory`, `biglm`, `filehash`, and/or `ff` packages in `R`.

Any additional data sources must be documented carefully.

Choice of statistical methods

You may use any of the techniques we learned in class or any other statistical methods you know/have picked up/look up to analyze the data. You must justify your choice of statistical method. You should describe your all statistical methods precisely and interpret the parameters in simple language, with appropriate characterization of uncertainty. Appropriate references should be included.

Format

Your final project should have the following sections: Title, Abstract, Introduction, Description of Statistical Methods, Results and Conclusions. R code for performing the analyses in your paper may be included as an appendix. Remember, your goal for this project is to convince me to take the flights you have suggested. There is an optional Latex template available from the class website. You may use different names for the sections or add sections to your paper if it makes the exposition clearer.

Length

The final paper must be less than 2,500 words (5 pages of text not including references). In addition, you may have up to four display items (tables and figures). Don't forget to include figure captions and axis labels!!

Good luck!!!