Final Exam Takehome Portion / Last HW Assignment
Due by 11:59pm Tues Dec 4, 2012
On Blackboard

- This assignment is about the same length ad difficulty as a weekly homework assignment, but it will count as half of your final exam grade.

- Unlike other homework assignments, *DO NOT* consult others as you do this work. Acceptable resources are:

  - The Gelman & Hill text.
  - R, and anything you can install in R as a library package.
  - Any incidental computing aid such as a calculator, excel spreadsheet, etc. Please cite any such resources in your final submission.
  - Class notes or anything else I have handed out in class, or posted on either of these websites:
    - `http://www.stat.cmu.edu/~brian/463/`
    - `http://www.cmu.edu/blackboard/`
  - Talking with or emailing the instructor (BJ) or TA's (Xiaolin or Spencer).
  - *Static* resources on the www or in the library, For example:
    - **These are OK**: books, webpages and pdf's; *Cite these as references in your final submission.*
    - **These are NOT OK**: email (except with BJ or the TA's), chats, IM, social networking sites, etc.; *Don't use them.*
  - Anything not in the above list? Check with BJ *first*.

  Submitting your exam work constitutes your personal guarantee that you worked on your own, following strictly the resource guidelines above. Violation of your guarantee will result in a grade of 0 (zero) on this exam.

- Please assemble your work for this exam into a single document (pdf preferred, doc or docx acceptable), and submit on Blackboard by Midnight Tue Dec 4. Late submissions will not be accepted[1].

  - Use the filename "463-final-junker-b.pdf" or "663-final-junker-b.pdf" (with your last name and first initial, not mine!).
  - Organize your work so that it is easy to read and easy to find the answers to each part of each question below. Answers that are not easy to find will not be graded!

- This takehome will be worth half of your final exam grade, which is equivalent to 10% of your course grade.

[1] If there is some good reason you can't turn it in on time, please talk to me about *well in advance*.

Background

Lillard & Panis (2000) discuss data on the decisions of 501 mothers to deliver babies in a hospital vs. at home or elsewhere. The mothers have varying numbers of children, ranging from 1 to 10, and make separate decisions about where to deliver each child. There are a total of 1060 births in the data set. The available variables are

```
'data.frame':   1060 obs. of  6 variables:
 $ hospital: int  0 0 1 0...    1 = hospital birth, 0 = birth elsewhere
 $ loginc  : num  4.33 5.62... logarithm of family income (log dollars)
 $ distance: num  1.7 7.9...    distance (miles) from nearest hospital
 $ dropout : int  0 0 0 0 0... 1 = mother did not complete hs, 0 = completed hs
 $ college : int  1 0 0 0 0... 1 = mother attended college, 0 = did not
 $ mom     : int  1 2 2 2 2... unique identifier for each mother
```

Note that family income varies from the birth of one child to the next, hence family income is recorded for each child, rather than once only for each mother. Note also that if both `dropout` and `college` are zero, then the mother completed high school but did not go on to college. The mother's `group` identifier appears once for each of her children; thus, the number of children per mother could be obtained as `n.kids <- table(mom)`.

The data are available in the file `hosp.txt` under a link for the takehome portion of the final, at `http://www.stat.cmu.edu/~brian/463`.

Exercises

1. A simple model (perhaps good, perhaps bad) would be to allow the probability of birth to depend on family income at the time of birth, with a random intercept for each mother. As a multilevel model, this is

   Level 1:

   $$\text{logit}(P[y_i = 1]) \;=\; \alpha_{0j[i]} + \beta_1 x_i, \quad i = 1, \ldots, 1060$$

   Level 2:

   $$\alpha_{0j} \;=\; \beta_0 + \eta_j, \quad \eta_j \sim N(0, \tau^2), \quad j = 1, \ldots, 501$$

   where $y_i = 1$ if the i^{th} child was born in a hospital, $j[i]$ is the mom on the i^{th} child, and x_i is the log(income) for the family at the time of that child's birth. Equivalently in `lmer()`'s modeling language

   ```
   lmer.inc <- lmer(hospital ~ loginc + (1|mom), data=hosp, family=binomial)
   ```

   (a) Write this as a hierarchical Bayes model, adding prior distributions wherever needed.

   (b) Write this as s variance components model.

2. In the file `2012-final-takehome-rcode.r` in the takehome final area on `http://www.stat.cmu.edu/~brian`, there is R code to fit two different WinBUGS models to the hosp.txt data:

   - model.00 and rube.00
   - model.01 and rube.01

   (a) Fit each model and inspect the output and diagnostic graphs. Is there anything remarkable to point out, or is the MCMC algorithm working well all parameters in each model? Write a few sentences about your findings, illustrated with appropriate graphs or numerical output.

   (b) Write model.01 as a hierarchical Bayes model. Find a `glm()` or `lmer()` model that is more or less equivalent to model.01, and fit it. Submit the model you found, the output of `summary()` or `display()` on the fitted model, and a sentence or two comparing parameters estimates between your models and rube.01.

   (c) Write model.00 as a hierarchical Bayes model. Find a `glm()` or `lmer()` model that is more or less equivalent to model.00, and fit it. Submit the model you found, the output of `summary()` or `display()` on the fitted model, and a sentence or two comparing parameters estimates between your models and rube.00.

   (d) Using only the fitted model objects rube.00 and rube.01, and any numerical or graphical output you can derive from them, which model fits the data better? Why? [Advice: Don't kill yourself doing lots of different things to answer this question!]

3. Assume the children are listed in birth order for each mother, in the data set. We can obtain the birth order as follows:

```
b.ord <- unlist(lapply(split(mom,mom),function(x){1:length(x)}))
```

   Do mothers tend to be more likely to have hospital births with each successive child?

   (a) Answer this question using `glm()` models that relate the probability of hospital birth to the `b.ord` variable. Feel free to add other covariates from `hosp.txt` if they make for a better model.

   (b) Does your model fit better if you allow a random intercept or slope, or both? Try to assess this, using `lmer()` and related tools (not `WinBUGS`).

   (c) Does the answer to the question change if you use an `lmer()` model from part (b) rather than a `glm()` model from part (a)? Explain.

   Vocabulary note: Models in which the outcome varies with time, within each subject, are called growth curve models. In part (b) we are trying to fit a growth curve model, for the propensity to chose hospital birth.