36-463/663: Multilevel and Hierarchical Models
Project, Part I
Due: FRIDAY NOVEMBER 18.

Please submit as a single pdf on Blackboard.

Instructions & General Information

- DO NOT WORK WITH OTHERS. Please do this project entirely on your own. Acceptable resources are:
  - The Gelman & Hill text, and the Lynch 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/463-663/`
    - `http://www.cmu.edu/blackboard/`
  - Talking with or emailing the instructor (BJ) or TA (Nicholas Kim).
  - *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), chats, IM, social networking sites, etc.; *Don't use them.*[1]
  - Anything not in the above list? Check with BJ *first*.
  - Anything you used? *Cite it* **whether it is on the OK list or the not-OK list**.

  Submitting your work for this part of the project 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 part of the project.
- Please assemble your work for this part of the project into a single pdf document and submit on Blackboard. Late submissions will not be accepted[2].
  - Use the filename "project-part-one-junker-brian.pdf" (with your last name and first name, 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. Clearly label all sections / subparts / tables / graphs / etc. **Answers that are not easy to find or read will receive no credit!**
- This part of the project will be worth 10% of your final grade.

[1] Except for the class Q&A Discussion Board on Blackboard, please don't use discussion boards either. You may post questions on the Q&A board, but let me or Nick answer them.

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

Problem

The Setting

Derek Lomas, a former grad student in the Human-Computer Interaction Institute at Carnegie Mellon, is a learning scientist and educational game designer in the Design Lab at the University of California, San Diego. As part of his graduate work, Dr. Lomas developed and refined a set of visual estimation games collectively called *Battleship Numberline*. You can (and should) play part of the game to see what it is like at `http://playpowerlabs.org/bsnl/v21/brainpop/BSNL.html`, and you can learn about the design of Battleship Numberline and some of Dr. Lomas' research questions from the short research report Lomas et al. (2013); a copy of the article is provided in the same subdirectory as this problem statement.

The basic "game play" is that a number between 0 and 1 is shown to the player, indicating the location of a submarine hidden in a body of water on the screen. The player clicks on a number line overlaid on the body of water to indicate the location of the submarine. A missile drops to the location indicated by the player, and if the missile hits the submarine squarely, a correct answer is recorded.

Battleship Numberline was an extraordinarily successful game that was played on computers around the world. The data that we have to work with comes from one of Lomas' studies, involving 414 players answering estimation questions using decimal numbers between 0 and 1. Each player has a unique id, and we also know the IP number (numerical web address, such as 108.129.0.203) of the computer the player was playing on.

In this part of the project we will be interested in building hierarchical generalized linear models, actually hierarchical logistic regression models, to determine what factors influence the probability of correct responses.

Since answers to questions from the same player are likely to be more similar to each other than answers from different players, players are a natural choice for a random effect. It's also possible that geographic location, as coded by part or all of the IP number, might be a second random effect.

References

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* NY: Cambridge Univ Press.

Lomas, D., Patel, K., Forlizzi, J.L., and Koedinger, K.R. (2013). Optimizing Challenge in an Educational Game Using Large-Scale Design Experiments. Paper presented at CHI 2013, Paris, France. Obtained online at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.480.2493&rep=rep1&type=pdf

Lynch, Scott M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.* New York: Springer.

The Data

The data are contained in the text file `dec-data.csv` which can be found in the same directory as these instructions. There are 8,257 lines in the data set; each line represents one player's response to one question. There are 31 variables in this data set. Some of the main variables to consider are listed below; other variables are described in the accompanying file `dec-data.r`; some of these other variables may also be useful for our analyses.

Some of the main variables are:

Variable	Definition & Comments
SID	Student (player) ID number; there are 414 unique SID's.
ip	IP number (numerical web address) for the computer player was using. IP numbers are organized hierarchically. For example, 108.129.0.203 denotes a specific computer; 108.129.0 denotes a group of computers in an organization or physical location, 108.129 denotes a group of organizations or physical locations, etc. See also the variable ip2 and ip3 in the data set.
isGuidesEnabled	Are there hashmarks printed on the numberline that player is using to guess location of sub?
experimentName	An "experiment" is a specific subset of questions that was asked of a specific subset of players. experimentName is the descriptive name of the experiment for this player $\times$ question interaction.
levelName	Groups together experimental conditions (as identified by "experimentName")
currentLevelNo	Numerical game level announced to player for this player $\times$ question interaction.
currentQuestion	The "correct value" of the location of the submarine that player is trying to locate on the number line. Since all the questions have the same format (find the hidden submarine on the number line), this "correct value" uniquely identifies each question in the game. There are 20 unique questions.
answerDec	The decimal value of the player's guessed locaton of the submarine.
curReactionTime	The time it took player to produce a response for this question. See also the variable reacTime in the data set.
hitType	Factor with 5 levels: "Miss!!", "Near Miss!!", "Partial Hit!!", "Perfect Hit!!", "Time Out!!". The value "Perfect Hit!!" indicates that `answerDec` was close enough to `currentQuestion` to count as a "correct answer"; all other levels indicate "wrong" or "partial credit".
resp	resp=1 if hitType="Perfect Hit!!"; else resp=0.

The Goal

Analyze this data to determine how difficult each question is, and whether there are other factors besides the individual question, that affect the probability of a correct response. Account appropriately for any important grouping effects, and any other variables of interest.

This part of the project has has five sections/problems, and is worth a total of 100 points.

Please adhere to the page limits suggested for each part of the project.
Also, please attach two appendices to your hw paper:

- **Appendix I** (mandatory) a list of all materials you referred to, formatted in the same style as the list of references on pp. 575 ff. of Gelman & Hill. *Please adhere closely to the style of references in G&H. Failure to do so may result in misunderstanding about your work and loss of points on this takehome.*
- **Appendix II** (optional) any material (R code, computer output, figures, tables, etc.) that you think is important for your homework but would not fit in the page limits for each part of the project above. *The more stuff that is in here, the less likely that I will consider it when grading your project, so please choose wisely what (if anything) to put here.*

1. *Exploratory Data Analysis.* [20 pts]

   Perform an exploratory data analysis of the data, keeping these questions in mind:

   - How many questions does each player try?
   - How are the "experiments" related to questions or to players?
   - How are players related to ip numbers or partial ip numbers?
   - What is the distribution of proportion-correct scores for players?
   - What fraction of players get each question right?
   - What is the distribution of reaction times across players? Across questions?

   Turn in appropriate tables & graphs, with appropriate descriptions for each.

   Page limit for section 1: 2 pages.

2. *Logistic regression for question difficulty.* [30 pts]

   (a) Fit a logistic regression model giving the probability that a question will be answered correctly, using currentQuestion *as a factor*, and omitting the intercept. Summarize the fit, and comment on why taking the intercept out might be a useful idea.

   (b) Plot the coefficients against the fraction of participants who got the corresponding question right. Try both the raw fraction of particpants, and the logit of the fraction. Provide the better plot, explain why it is better, and interpret the plot.

   (c) Try replacing, or augmenting, currentQuestion with other variables in the data set. Provide (i) a short paragraph describing what method or mixture of methods you used to find the best among all of these models; and (ii) a paragraph interpreting that model that would be useful to Dr. Lomas.

   Page limits for section 2: 1 page for each of the 3 subparts.

3. *Logistic regression for player proficiency.* [20 pts]

   (a) Fit a logistic regression model giving the probability that a player will provide a correct answer, using SID *as a factor*, and omitting the intercept. Summarize the fit. *Note: glm is estimating over 400 parameters, and it may take a few minutes for it to fit the model.*

   (b) Plot the coefficients against the proportion correct for each player. Try both the raw proportion correct, and the logit of the proportion correct. Provide the better plot, explain why it is better, and interpret the plot.

   Page limits for section 3: 1 page for each of the 2 subparts.

4. *Mixed Effects Models* [40 pts]

   Dr. Lomas is primarily interested in the questions—what makes them difficult or easy, for example—and it is quite likely that questions answered by the same person will be dependent on each other

through that player's ability. For these reasons, we will focus on questions as fixed effects and players as random effects.

(a) Use `glmer()` to fit a mixed effects logistic regression predicting the probability of a correct response, using currentQuestion as a fixed effect *factor*, omitting the intercept, and with a random intercept grouped by SID (that is, all these should be implemented in your model).

Note: this model will take a while to fit, as did the model in question 3. In addition you will probably see some error message indicating that the algorithm did not converge. This is a sign of lack of fit, and suggests that we should be careful how we try to use the estimated parametrs.

(b) Plot the fixed effects from this model against the fraction of players who got the corresponding question correct, or against the logit of that fraction, if it makes a better plot. Provide a brief interpretation of the fixed effects based on this plot.

(c) Plot the random effects from this model against the proportion correct for each player, or against the logit of the proportion correct, if it makes a better plot. Provide a brief interpretation of the random effects, based on this plot.

(d) Try replacing or augmenting currentQuestion with other variables to try to obtain a better model (again, many of these models will take a few minutes to fit). Describe briefly (a paragraph or less) the method or methods you used to find the best among all of these models; and show the best model you found. *Note: some of the models you try will take a long time to fit. There may also be some convergence issues, as in part (a). Try to find a model or models that fit well, do not have convergence issues, and make some sense given the nature of the problem. You may (or may not) find your fits from sections 2 and 3 useful to guide your work here.*

(e) Now take the best model you have so far, and see whether adding a second random intercept corresponding to ip, ip2, or ip3 helps. Do you still need the random effect corresponding to SID? Describe briefly (a paragraph or less) the method or methods you used to find the best among all of these models; and show the best model you found.

Note: The mixed effects logistic model you fitted in part (a) is a variation of the "Rasch model" in psychometrics (a blend of statistics and psychology that focuses in part on modeling standardized testing data).

Page limits for section 4: 1 page for each of the 5 subparts.

5. *Summary.* [10 pts]

Review your work. Can you find any interesting relationships between question difficulty and other variables in the data set (whether you found them with modeling or with EDA, or both)?

Now write a careful text summary of your work. Briefly describe the data set, the methods you used, and your conclusions. Be sure to mention any relationships you found between item difficulty and other variables in the data set. Your language should be precise, but intelligible to someone like Dr. Lomas, who knows some statistics but is not an expert.

Page limits for section 5: 2 pages.