General Linear Model I

Midterm Exam

Due December 6, 2016 (11:59 P.M.)

1. (3 pt) Two items A and B are weighed on a balance, first separately and then together, to yield observations y_1, y_2, and y_3. Say, suppose the true weights of A and B are α_A and α_B, we have

$$\begin{aligned} y_1 &= \alpha_A + \varepsilon_1 \\ y_2 &= \alpha_B + \varepsilon_2 \\ y_3 &= \alpha_A + \alpha_B + \varepsilon_3 \end{aligned}$$

   (a) If $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $i = 1, 2, 3$, find the reasonable estimates of α_A and α_B. Show your work.

   (b) If $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ for $i = 1, 2$, and $\varepsilon_3 \sim N(0, k^2\sigma_\varepsilon^2)$, where constant $k > 1$, find the reasonable estimates of α_A and α_B. Show your work.

   (c) Let $y_1 = 41$, $y_2 = 53$, $y_3 = 97$, $k = 1.25$. Choose a suitable function in `R`, and find the estimates of α_A and α_B in (a) and (b). Include your `R` codes, and highlight the key `R` function you use.

2. (5 pt) Consider the data from Wikipedia article *List of world records in masters athletics* at https://en.wikipedia.org/wiki/List_of_world_records_in_masters_athletics. We use man's marathon records at https://en.wikipedia.org/wiki/List_of_world_records_in_masters_athletics#Marathon, and woman's marathon records at https://en.wikipedia.org/wiki/List_of_world_records_in_masters_athletics#Marathon_Women.

   First, summarize these records by using a table with columns `Time (second)`, `Age (year)`, and `Gender (Female or Male)`.

   For marathon records, if there are more than one record of a certain age group, click "progression" and get to the wikipedia page about *marathon world record progression*, and choose the fastest record in this list of progression. For example, In Masters W40 group, there are two records by Mariya Konovalova (2:22:27) and Irina Mikitenko (2:24:54). The you need to get to the wikipedia page *Masters W40 marathon world record progression* at https://en.wikipedia.org/wiki/Masters_W40_marathon_world_record_progression, and get the record for Masters W40 group as 2:24:53.6.

   For age, use the age in the column of age group. For instance, Fauja Singh ran a marathon in 5:40:01 at age 92 years old in age group M90. Take the age as 90 years.

   (a) Consider time as the response variable (y) and age as the explanatory variable (x). For female students, use woman's record; for male students, use man's record. Fit the models

$$y = \beta_{10} + \beta_{11}x$$

   and

$$y = \beta_{20} + \beta_{21}x + \beta_{22}x^2.$$

   Include your `R` codes and report your estimates. Does the extra quadratic term appear necessary?

(b) Denote the estimates in part (a) of the intercept of model $y = \beta_{10} + \beta_{20}x$ as b_{0F} in woman's record model, and as b_{0M} in man's record model.
Include gender as an additional explanatory variable (v), and $v = 1$ corresponds to woman's record, and $v = 0$ corresponds to man's record. Consider the model

$$y = \beta_{30} + \beta_{31}x + \beta_{32}v.$$

Include your R codes and report your estimates. How does gender appear to affect the records?

(c) For female students, compare $\hat{\beta}_{30} + \hat{\beta}_{32}$ and b_{0F}. For male students, compare $\hat{\beta}_{30}$ and b_{0M}. Explain the difference.

(d) For female students, use woman's record; for male students, use man's record. Using the data fit a Gamma generalized linear model. Interpret your findings and compare with part (a). Include your R codes, and write down the link function you choose, and the equation of your fitted model.

(e) Show that the density of *inverse Gaussian distribution* lies in the exponential family, and write the distribution in the canonical form of a generalized linear model. Then repeat part (d) using an inverse Gaussian generalized linear model.

3. (3 pt) Read the research article especially the methods section, at `http://www.sciencedirect.com/science/article/pii/S0195666316302100`. The authors used R package `lme4` to analyze the raw data.

Stephanie M. Manasse, Hallie M. Espel, Leah M. Schumacher, Stephanie G. Kerrigan, Fengqing Zhang, Evan M. Forman, Adrienne S. Juarascio, Does impulsivity predict outcome in treatment for binge eating disorder? A multimodal investigation, *Appetite*, Volume 105, 1 October 2016, Pages 172-179.

This paper is also available on Blackboard.

(a) Manasse et al. did not release their data. Based on their paper, find the structure of their data set. Say, find all variables and the relations between these variables. You may use plots to show the relations.

(b) Generate a small R-readable data set, the numbers in your data set are arbitrary, but the structure of your data set is similar with Manasse et al.'s data set. Say, it looks like you randomly pick several rows from Manasse et al.'s data set. Clearly define all variables in your data set, and this data set can be directly analyzed by functions in R package `lme4`.

(c) Write down the commands by using functions in R package `lme4`, which are able to process your small data set. Report the results based on your small data set.

Instructions

- There are 3 questions, each question is between 3-5 points. A perfect score is 10 points.
- Show all work. You will receive partial credit for partially completed problems.
- You may use any references, any texts and any online media.
- Discussion is allowed and encouraged. But it is not allowed to directly copy solutions from other students. If you have a group discussion (on-line or face-to-face), please mention it in your solution (including the names of participants in your discussion group). Mentioning the general discussion will not influence your score.

End of the Midterm of Stat 706 (Instructor: Jiangtao Gou)