Hunter College: Statistics 706 Fall 2017

General Linear Model I

Midterm Exam

Due December 10, 2017 (11:59 P.M.)

1. (3 pt) There are a series of papers by de Souza et al. on the generalized linear models in astronomy. These articles can be found either from the links provided below, or from the Blackboard where three pdf files were uploaded. In this question, we focus on the second paper. You may read other papers if you are interested in.

   - R.S. de Souza, E. Cameron, M. Killedar, J. Hilbe, R. Vilalta, U. Maio, V. Biffi, B. Ciardi, J.D. Riggs (2015). The overlooked potential of generalized linear models in astronomy, I: binomial regression, *Astronomy and Computing*, **12**, 21–32. `https://doi.org/10.1016/j.ascom.2015.04.002`.
   - J. Elliott, R.S. de Souza, A. Krone-Martins, E. Cameron, E.E.O. Ishida, J. Hilbe (2015). The overlooked potential of generalized linear models in astronomy, II: gamma regression and photometric redshifts. *Astronomy and Computing*, **10**, 61–72. `https://doi.org/10.1016/j.ascom.2015.01.002`.
   - R. S. de Souza, J. M. Hilbe, B. Buelens, J. D. Riggs, E. Cameron, E. E. O. Ishida, A. L. Chies-Santos, M. Killedar (2015). The overlooked potential of generalized linear models in astronomy, III: Bayesian negative binomial regression and globular cluster populations. *Monthly Notices of the Royal Astronomical Society*, 453, 1928-1940. `https://doi.org/10.1093/mnras/stv1825`

   (a) Read the second paper by Elliott et al. on gamma regression and photometric redshifts. Write a brief summary of section 2 *overview of regression methods*, page 62-64.

   (b) Appendix A (page 68-69) provides instructions to perform the photometric redshift estimation using the `R` package. Run these `R` codes line by line, and explain the purpose and output of each command line. Elliott et al. also provide `python` codes in Appendix B. If you prefer `python`, you can run and explain the `python` codes. Note, you only need to choose either `R` or `python`.

2. (6 pt) Consider the data from *All Time World Rankings*. We use man's 100 meter dash records at `http://www.mastersathletics.net/fileadmin/html/Rankings/All_Time/100metresmen.htm`, and woman's 100 meter dash records at `http://www.mastersathletics.net/fileadmin/html/Rankings/All_Time/100metreswomen.htm`.

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

For the time record in each age and gender group, you should use the fastest times without wind assistance. Based on *Rule 260.14(c) of IAAF Competition Rules 2016-2017*, if a tail wind exceeds 2 meters per second the result cannot be registered as a record on any level. So you should use the fastest times among the wind speed less than or equal to +2 m/s.

For age, use the lower bound of each age group. For instance, the age for age group M35-39 is 35, the age for age group W90-94 is 90.

(a) Summarize the record of each age and gender group and form an `R`-readable table. For example, the first several rows of the table may be

Gender	Age	Time
M	35	9.97
M	40	10.29
	......	
W	35	10.74
W	40	10.99
	......	

(b) 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?

(c) 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?

(d) 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.

(e) 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 (b). Include your `R` codes, and write down the link function you choose, and the equation of your fitted model.

(f) 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 (e) using an inverse Gaussian generalized linear model.

3. (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 $y_1 = 41$, $y_2 = 53$, $y_3 = 97$, $k = 1.2$. 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. Compare the estimates of α_A and α_B in (a) and (b) and explain the differences.

Instructions

- There are 3 questions, each question is between 3-6 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 between classmates in Stat 706 is 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)