# *Exercises 10 · GLMs*

**Due Monday, April 25, 2016**

*(1) Breast-cancer screening (take-home exam question in 2013)*

The data in "brca.csv" (from the course website) consist of 987 screening mammograms administered at the Group Health Cooperative in Washington state in 2002. Five radiologists, each of whom frequently read mammograms, were selected at random from those in the cooperative. For each radiologist, roughly 200 of the mammograms they had read were selected at random. Each row corresponds to a single woman's mammogram; the radiologist who read it is identified by a three-number code (1-999).

For each patient, two outcomes are recorded. The first outcome is an indicator of whether the subject was recalled by the radiologist for further diagnostic screening after the mammogram (1=Recalled for further diagnostic screening, 0=Not recalled). The second outcome is an indicator of whether there was an actual diagnosis of breast cancer within 12 months following the screening mammogram (1=Yes, 0=No). In addition, several risk factors identified in previous studies are provided; referent values for a "typical female" are indicated by asterisks:

*age:* 40-49*, 50-59, 60-69, 70 and older
*family history of breast cancer:* 0=No*, 1=Yes
*history of breast biopsy/surgery:* 0=No*, 1=Yes
*breast cancer symptoms:* 0=No*, 1=Yes
*menopause/hormone-therapy status:* Pre-menopausal, Post-menopausal & no HT, Post-menopausal & HT*, Post-menopausal & unknown HT
*previous mammogram:* 0=No*, 1=Yes
*breast density classification:* 1=Almost entirely fatty, 2=Scattered fibroglandular tissue*, 3=Heterogenously dense, 4=Extremely dense

All entries are numeric. For risk factors with just two levels, the referent level is represented by zero and the alternative by one. For risk factors having more than two levels, the referent level is specified and columns are presented only for incidence of the non-referent levels. A separate "brcanames.txt" file is provided specifying the column names.

(A) Given a set of risk-factor levels and a particular radiologist, there is a conceptual $2 \times 2$ table of recall outcome by cancer outcome. To learn about the probabilities associated with any such table,

construct two models: one for the probability of post-screening recall given risk-factor levels and radiologist; and another for the probability of cancer given recall outcome, risk-factor levels, and radiologist. Fit your chosen models to the given data. Explain your choice of model. Then interpret and comment upon the results, paying careful attention to issues of uncertainty.

(B) For a "typical" patient (no history of breast biopsy or surgery or family history of breast cancer, age between 40 to 49, post-menopausal and using hormone replacement therapy, has density breast classification 2, and has no reported symptoms), estimate the chance of the joint event of no recall and no cancer. Assume an "average" radiologist, and explain how you operationalized this assumption. Recall that $P(X, Y) = P(X)P(Y \mid X)$.

(C) For a typical patient (as above), estimate the chance of a false positive—that is, the chance that the patient is recalled, yet does not develop cancer within 12 months following the screening mammogram. Again, assume an "average" radiologist.

(D) In light of your analysis, are there any risk factors the radiologists should place higher or lower weight upon in deciding whether to recall a patient for further screening after an initial mammogram? How did you come to this conclusion?

*(2): Count outcomes*

Please read the following paper: Long, J. S. (1990). The Origins of Sex Differences in Science. Social Forces, 68(4), 1297–1316. You can find this easily through the UT Library website. Next, download the data in "biochem.csv" from the course website. This is Long's data set on the number of articles written by people in a sample of 915 biochemistry graduate students in biochemistry Ph. D. programs in the U.S. Each row corresponds to a student. The variables are:

- articles: articles produced during last 3 years of Ph.D.
- sex: male or female student
- married: whether the student is single or married
- kidsUnder5: number of children aged 5 or younger
- prestige: a measure of the national reputation of the department
- mentorArticles: how many articles were produced by the Ph.D. mentor during last 3 years of student's Ph.D.

Analyze the data with the Long's article in mind. Describe your approach. Do you reach qualitatively similar conclusions as he did?