Exercises 9 · More on GLMs

To turn in (Due Monday, April 13, 2015)

(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 2×2 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 the joint probability $P(X, Y)$ is the same as $P(X)P(Y \mid X)$.

(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) Journal versus journal

It's good to get practice reading and digesting scientific papers that use unfamiliar statistical methods or terms. Here are two!

The files "bisph-bmj.pdf" and "bisph-jama.pdf" contain two articles, one from the *British Medical Journal*, one from the *Journal of the American Medical Association*. Both appeared in the summer of 2010. Both address the question of whether a class of osteoporosis drugs called oral bisphosphonates[1] increase the risk of esophageal cancer. Both describe observational studies involving the same medical-records database. Yet the study in JAMA concluded:

[1] http://en.wikipedia.org/wiki/Bisphosphonate

> Among patients in the UK General Practice Research Database, the use of oral bisphosphonates was not significantly associated with incident esophageal or gastric cancer.

On the other hand, the study reported in BMJ concluded:

> The incidence of oesophageal cancer was increased in people with one or more previous prescriptions for oral bisphosphonates compared with those with no such prescriptions.

As you can see, the imaginary newspaper headlines one would write about these two studies differ sharply!

Write a two-page report (three in a pinch) comparing the methodologies and conclusions of the two papers. Based upon the statistical evidence, do you think their conclusions are really as different as the imaginary newspaper headlines would suggest? If so, does one article seem more trustworthy than the other, or is there little to choose between them?

In addition to your report, include a supplement that provides explicit definitions, in your own words, of the following terms. Some are from class, some may be new.

- relative risk
- Cox proportional hazards model
- Kaplan–Meier curve
- censoring
- hazard ratio
- case–control study
- cohort study

You may write out any equations by hand, if you wish. Please also define any other unfamiliar statistical terms that you encounter in the papers.