Homework 3: Categorical Data

Nicholas G Reich, for Biostats 743 at UMass-Amherst

Your assignment should be submitted in two separate files by 8am on Thursday October 19th. The first, should be an RMarkdown (.Rmd) or another format that dynamically compiles your write up and runs the code inside it. The second should be the PDF file that was reproducibly compiled using the first file. All figures should be generated by the code, none should be loaded directly. The homework files should be submitted using your shared Google Drive folder with the instructor.

Question 1

Assume the following data generation model:

$$Y_i \sim Poisson(\mu_i)$$

where

$$log\mu_i = \alpha + \beta \cdot x_i$$

Fix $\alpha = log(15)$, $\beta = log(2)$, and draw x_i independently from the distribution $Unif(0, 20)$ for $i = 1, ..., N$.

(a) Simulate $N = 500$ observations from this model. Plot the data.
(b) Fit a Poisson log-linear GLM to the data.
(c) Write a function that can calculate the likelihood of your data given (α, β). Across a fine grid of possible values for α and β, compute the likelihood for each point. Plot the resulting likelihood surface, showing the MLE, contour lines representing the 80% and 95% likelihood based credible regions, and the estimated 80% and 95% confidence ellipses based on the estimated asymptotic covariance of $(\hat{\alpha}, \hat{\beta})$.
(d) Repeat a-c for a new sample of $N = 20$.
(e) Describe your results, with particular attention paid to (i) comparisons between the confidence/credible regions for each of the two sample sizes, and (ii) any differences you observe across the two different samples of data. Note: be sure to use `set.seed()` to ensure your results and interpretations are reproducible and consistent when you re-knit the file.
(f) Re-run the code for (d) 10 times (with different samples of y each time, hold x fixed across the iterations) and qualitatively assess the sensitivity of your results in (d) and interpretations in (e). Do your results change substantially with each new sample? Summarize your results graphically.

Question 2

Background

This question builds on unpublished research from the Women's Health and Aging Study (WHAS), a population-based prospective cohort of women age 65 and older designed to sample the one-third most-disabled women in a twelve zip code area of Baltimore, Maryland (Guralnik J, 1995). This study was conducted in the 1990s and 2000s. Broadly, WHAS was designed to identify and answer questions about risk factors for older women becoming frail. The definition of frailty used here was based on a method developed and validated by Fried and colleagues in the Cardiovascular Health Study (CHS) prospective cohort (Fried, J Gerontol Med Sci 2001). WHAS women were evaluated at baseline on five clinically measured criteria: (1) shrinking, defined as either body mass index less than 18.5 kg/m2 or greater than 10% loss of body weight since age 60; (2) weakness, defined as the lowest quintile of grip strength of the dominant hand; (3) poor endurance and energy, defined as self-report of being either more tired or weaker than usual in the past 30 days; (4) slowness, defined as the lowest quintile in time to complete a 4 m walk; and, (5) low activity level, defined by the lowest quintile of self-report of weekly activity determined by a subset of questions from the

Minnesota Leisure Questionnaire. Participants were defined as frail if they met three or more of these five criteria, pre-frail if they met one or two criteria and non-frail if they failed to meet any of the criteria. To create a binary variable for frailty, `bin_frail`, women were defined as frail if they met three or more of the five criteria and not frail otherwise.

Baseline age, self-report race, education, and cognitive impairment have been shown to have associations with frailty and mortality (Fried 2001; Hirsch, 2006; Buchman, 2008; Szanton, 2010). Education is provided as a continuous variable measuring years of formal education completed. The Mini–Mental State Examination (MMSE) is a standard measure of cognitive impairment, based on a 30-point questionnaire. MMSE scores were assessed categorically, with seven categories from 24 to 30 (Folstein, 1975).

Recent work has shown an association between low levels of nutrition and decreasing physical function in elderly adults (Bartali, JAMA 2008) and that low levels of some micronutrients such as caretenoids, alpha-tocopherol or 25-hydroxyvitamin D were associated with frailty in older adults according to the clinical definition developed by Fried et al. (Semba et al, J Gerontol A Biol Sci Med Sci, 2006).

Modeling tasks

Your goal is to use data from this observational cohort study of community-dwelling older women to develop an objective and evidence-based micronutrient summary aimed at identifying older adults at risk of frailty.

The dataset `whas.csv` in the class Google Drive folder (with associated codebook) contains observations from 682 women. Some values (but not all) that were missing in the original dataset have been imputed to create this dataset.

Based on this data, can micronutrient and demographic data be used to predict incident frailty in older women? Design a modeling experiment to answer this question. **Before** running any models, look at the data available to you and write down a one-paragraph analysis plan that describes what models you will fit (you should fit and compare at least five GLMs). In the design phase, you may use descriptive and exploratory plots of the data, but **do not** assess the bivariate relationship between the outcome and any of your covariates (i.e. don't "unblind" yourself to the results before beginning the analysis phase). Include in your analysis plan what kind of model selection (if any) you will perform, and whether you will consider transformations of your covariates and/or using non-linear relationships to model the relationship between any covariate(s) and the outcome. Based on previous literature, a good reference model would be one that includes basic demographic data as well as the nutrient data of carotenoids and selenium. In general, when evaluating a set of models for predictive accuracy, it is good practice to include at least one *a priori* very simple model that can serve as a reference model to other more complex models. Use a formal validation exercise (e.g. k-fold cross-validation or a held-out test sample) to evaluate your models. Evaluate your models' predictive performance using an ROC curve. Include in your write up for this problem your analysis plan, summary output of key models that you fit (including up to 3 tables and figures as appropriate), and a one-paragraph summary of the results. Use mathematical notation to write down the model equations for at least one of the fitted models.

Extra credit

Because the WHAS study was a cohort study, population-based sampling weights are provided (the `swts` variable) so that the weighted sample more accurately reflects the demographic make-up of the population. Re-run one of the prediction models above, incorporating the survey weights. Describe which package/software you used to incorporate the survey weights. Compare the weighted results to your results above, including coefficient estimates, standard errors, and area under the ROC curve. Are they different? Explain why the results are similar or different to the ones obtained above. Reference specific observations in the dataset to show the differences.

Question 3

The previous question asked you to optimize your model choice for the best predictive model. This question will ask you to use the same data, but with an inferential question in mind. The `frail` variable is an ordinal categorical variable with three levels indicating a participant's level of frailty: "not frail", "pre-frail", and "frail".

Based on this data, do low-levels of serum plasma carotenoids (α-carotene, β-carotene, β-cryptoxanthin, zeaxanthin and lycopene), and demographic variables impact the development of the frailty phenotype in older women? Use an appropriate GLM to answer this question. Describe your model selection process and show your final model in an equation. Present your results in 1 paragraph and no more than 2 tables or figures. Summarize any relevant model diagnostics to show how well your model fits the data.

Question 4

Fit your chosen model in Question 3 using Stan, a probabilistic programming language for implementing Bayesian analysis. For ease of implementation, you can use the `rstanarm` package in R. Read this vignette before starting. Use the `shinystan` package in R to inspect your model fit. Find 2-3 plots and/or tables that summarize the results clearly, and include these in your write up in a dynamically loadable way (i.e. don't save a posterior plot as a .jpg file and load it as a static file into your report). Discuss similarities and differences with the model fit in Question 3.

Question 5

This question uses the `frisk_with_noise.dat` dataset in the `CDA` Google Drive folder. The data resulted from an investigation by the NY State Attorney General's Office into the "stop and frisk" policies of the NYPD. These data are an anonymized version of that data, collected over a 15-month period in 1998 and 1999. For this analysis, we are interested in looking at the impact of ethnicity and precinct on the rate of police stops.

(a) Based on this data, does ethnicity play a role in rate of police stopping? **Before** running any models, look at the data available to you and write down a one-paragraph analysis plan that describes what models you will fit. For this section, do not consider any models with overdispersion. Include in your write-up: the analysis plan, summary output of key models that you fit, and a one-paragraph summary of the results. Use mathematical notation to write down the model equations for at least one of the fitted models.

(b) For one of the chosen models from part (a), investigate whether a correction for overdispersion is needed. If it is, fit a new model with the correction in place. Show the results, describe how and why the results changed (if they did) and and re-interpret the results in light of this model.

(c) Considering the analysis plan that you wrote in (a) if you had it to do again, is there anything you would change about the analysis plan if you had to do it over again? Why?