STA4030: Categorical Data Analysis

Assignment 3

Due Date and Time: November 30, 2020 (Monday), 10:00PM

INSTRUCTION:

- Please scan your answers in **one single .pdf file** and submit via Blackboard System.
- **Late submissions** will receive a mark of zero.
- Students may discuss set problems with others, but your final submissions must be your own work.
- All these questions should be answered using a pen, paper, calculator (good practice for your midterm and final).
- You may use any software you like, e.g., R, Python, Excel, etc., to find the percentiles regarding relative distributions (for example, to find p-values).
- Show and write down your solutions in detail and clearly.

Problem Set 3:

1. (Exercise 1.10 of Agresti (2015)) GLMs normally use a hierarchical structure by which the presence of a higher-order term implies also including the lower-order terms. Explain why this is sensible, by showing that,

   (a). a model that includes an x^2 explanatory variable but not x makes a strong assumption about where the maximum or minimum of $E[Y]$ occurs.

   (b). a model that includes x_1x_2 but not x_1 makes a strong assumption about the effect of x_1 when $x_2 = 0$.

2. Show that the gamma distribution is a member of the exponential dispersion family and identify the natural parameter. The pdf for the gamma distribution can be written as,

$$f(y; k, \mu) = \frac{(k/\mu)^k}{\Gamma(k)} \exp\{-ky/\mu\} y^{k-1}, \ y > 0. \tag{1}$$

   Derive that $E(Y) = \mu$ and $Var(Y) = \mu^2/k$.

Note: Another form of the pdf of the gamma distribution is,

$$f(y;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} \exp\{-y/\beta\}, \ y > 0. \tag{2}$$

Compare Eq. (1) and Eq. (2), we have the following relationship between the parameters,

$$\alpha = k,$$

and

$$\beta = \frac{\mu}{k}.$$

Then $E(Y) = \alpha\beta$ and $Var(Y) = \alpha\beta^2$.

3. A study reports n_i independent binary observations $\{y_{i,1}, \ldots, y_{i,n_i}\}$ at level $X = x_i, i = 1, \ldots, N$ with $\sum_{i=1}^{N} n_i = n$. Consider the model,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i,$$

where $\pi_i = P(Y = 1 \mid X = x_i)$.

   (a). Show that the kernel of the likelihood function is the same as treating the data as n Bernoulli observations or as N binomial observations.

   (b). For the saturated model, explain why the likelihood function is different for these two data forms. Hence, the deviance reported by software depends on the form of data entry.

   (c). Explain why the difference between deviances for two unsaturated models does not depend on the form of data entry.

4. In the first 9 decades of the 20th century in baseball's National league, the percentage of times the starting pitcher pitched a complete game were: 72.7 (1900-1909), 63.4, 50.0, 44.3, 41.6, 32.8, 27.2, 22.5, 13.3 (1980-1989).

   (a). Treating the number of games as the same in each decade, the linear probability model has ML fit

   $$\hat{\pi} = 0.7578 - 0.0694x,$$

   where X = decade with $x = 1, 2, \ldots, 9$. Try to interpret the fitted probabilities.

   (b). Substituting $x = 12$, predict the percentage of complete games for 2010-2019. Interpret your results.

   (c). The logistic regression ML fit is

   $$\hat{\pi} = \frac{\exp\{1.148 - 0.315x\}}{1 + \exp\{1.148 - 0.315x\}}.$$

   Obtain $\hat{\pi}$ for $x = 12$. Which link function do you prefer?

5. For a study using the logistic regression model to determine characteristics associated with remission in cancer patient. Table 1 shows the most important explanatory variable, a labeling index (LI). This index measures proliferative activity of cells after a patient receives an injection of tritiated thymidine, representing the percentage of cells that are "labelled". The response Y measured whether the patient achieved remission (1 = yes). Software reports for a logistic regression model using LI to predict the probability of remission. Table 1 contains the output.

		Criterion	Intercept Only	Intercept and Covariate
		$-2\log L$	34.372	26.073
Parameter	Estimate	S.E.	Chi-Square	pr > ChiSq
Intercept	-3.7771	1.3786	7.5064	0.0061
LI	0.1449	0.0593	5.9594	0.0146
Odds Ratio	Estimates			
		Effect	Point Estimate	95% CI
		LI	1.156	$(1.029, 1.298)$

Table 1: Computer Output for Cancer data

(a). Show how software obtained $\hat{\pi} = 0.068$ when LI $= 8$.

(b). Show that $\hat{\pi} = 0.5$ when LI $= 26.06694$.

(c). Show that the rate of change in π is 0.009 when LI $= 8$ and 0.036 when LI $= 26.06694$.

(d). The lower quartile and upper quartile for LI are 14 and 28. Show that $\hat{\pi}$ increases by 0.42, from 0.15 to 0.57, between those values.

(e). For a unit change in LI, show that the estimated odds of remission would be multiplied by 1.156.

(f). Explain how to obtain the confidence interval reported for the odds ratio. Try to interpret your results.

(g). Conduct a likelihood ratio test for the effect ($\beta = 0$), showing how to construct the test statistic using the $-2\log L$ values reported.

6. (Open Question. For this question, you may attach your R codes screenshots.)
For the horseshoe crab data,

(a). Try to download the crabs dataset which can be found in the glm2 R package. Check the following URL,

$$https://cran.r-project.org/web/packages/glm2/index.html$$

(b). Try different values of the arguments in the R command glm2(). Try at least two models, i.e., at leat two settings of the arguments. [Hint: The negative binomial modeling treats colour as nominal-scale; or quantitatively assign scores to the colour variable.]

(c.) Try model comparison regarding your models proposed in item (b). Interpret your results.

(d). Try in R about the likelihood-ratio test regarding the null hypothesis,

$$H_0: \quad \text{No colour effect.}$$

Interpret your results.

THE END