THE CHINESE UNIVERSITY OF HONG KONG
Department of Statistics

STAT4006: Categorical Data Analysis
Problem Sheet 3

The deadline for this Problem Sheet is 5.30pm on Monday 30th November. Please submit your solutions via the link provided on the course Blackboard page - if you must submit your solutions in hard copy, please contact me at jawright@sta.cuhk.edu.hk in advance. **No late submissions will be accepted. A late submission will receive a mark of zero.** Students may discuss set problems with others, but their final submissions must be their own work.

Please answer the following problems. Questions should be answered using a pen, paper, calculator (good practice for your midterm and final). That said, you may use any software you like to find percentiles (i.e. for finding p-values). Show your working.

1. **(Adapted from 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 $\mathbb{E}[Y]$ occurs.

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

2. **(Adapted from Exercise of Agresti (2015))** Show that the gamma distribution is a member of the exponential dispersion family and identify the natural parameter. The p.d.f. for the gamma distribution can be written as

$$f(y;k,\mu) = \frac{(k/\mu)^k}{\Gamma(k)} e^{-ky/\mu} y^{k-1}, y \geq 0,$$

   for which $\mathbb{E}[Y] = \mu$, $Var(Y) = \mu^2/k$.

3. **(Adapted from Exercise 7.32 of Agresti (2015))** For the horseshoe crab data, the negative binomial modeling shown in the R output below treats colour as nominal-scale and then in a quantitative manner, with the category numbers as scores. Interpret the result of the likelihood-ratio test comparing the two models. For the simpler model, interpret the colour effect and interpret results of the likelihood-ratio test of the null hypothesis of no colour effect.

```
> fit.nb.color <- glm.nb(y ~ factor(color)) # Using Crabs.dat file
> summary(fit.nb.color)
                 Estimate    Std. Error  z value     Pr(>|z|)
(Intercept)       1.4069     0.3526      3.990       6.61e-05f
factor(color)2   -0.2146     0.3750     -0.572       0.567
factor(color)3   -0.6061     0.4036     -1.502       0.133
factor(color)4   -0.6913     0.4508     -1.533       0.125
---
> fit.nb.color2 <- glm.nb(y ~ color) # using color scores (1,2,3,4)
> summary(fit.nb.color2)
              Estimate    Std. Error  z value    Pr(>|z|)
(Intercept)   1.7045      0.3095      5.507      3.66e-08
color        -0.2689      0.1225     -2.194       0.0282
---
> anova(fit.nb.color2, fit.nb.color)
Likelihood ratio test of Negative Binomial Models
Response: y
Model  theta    Res.df   2 x log-lik. Test      df  LR stat. Pr(Chi)
```

```
1       0.7986   171      -762.6794
2       0.8019   169      -762.2960    1 vs. 2   2   0.3834   0.8256
---
> 1 - pchisq(767.409-762.679, df=172-171) # LR test vs. null model
[1] 0.0296
```

4. In the first nine decades of the twentieth 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 ($x = 1, 2, \ldots, 9$). Interpret -0.0694.

   (b) Substituting $x = 13$, predict the percentage of complete games for 2020-2029. Interpret.

   (c) The logistic regression ML fit is $\hat{\pi} = exp(1.148 - 0.315x)/[1 + exp(1.148 - 0.315x)]$. Obtain $\hat{\pi}$ for $x = 13$. Which link function do you prefer?

5. For a study using the logistic regression model to determine characteristics associated with remission in cancer patient, the following table 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 $\hat{\pi}$ 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 multiply by 1.156.

   (f) Explain how to obtain the confidence interval reported for the odds ratio. Interpret.

   (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. **(Adapted from Exercise 5.16 of Agresti (2015))** A study has n_i independent binary observations $\{y_{i1}, \ldots, y_{in_i}\}$ at level $X = x_i$, $i = 1, \ldots, N$, with $\sum_i n_i = n$. Consider the model $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, where $\pi_i = P(Y_{ij} = 1)$.

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

THE END