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.

GLMs

Question 1

Write your own R code for obtaining the MLEs and SEs for a Poisson GLM using two covariates. Conduct a small simulation study designed to stress-test your code and show that it can work in a variety of situations. Compare the results to the `glm()` function in R. Summarize the results and your experience in 1-2 paragraphs. [Acknowledgments to Ciprian Crainiceanu for this question.]

Question 2

CDA exercise 4.11

Table 4 is based on a study with Brithish doctors. (a) For each age, find the sample coronary death rates per 1000 person-years for nonsmokers and smokers. To compare them, take their ratio and describe its dependence on age.

(b) Fit a main-effects model for the log rates using age and smoking as factors. In discussing lack of fit, show that this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over age.

(c) From part (a), explain why it is sensible to add a quantitative interaction of age and smoking. For this model, show that the log ratio of coronary death rates changes linearly with age. Assign scores to age, fit the model, and interpret.

Question 3

CDA exercise 4.12

Table 1: 4.9 Data for Question 6 on Coronary Death Rates

	Person-years		Coronary Deaths	
Age	Nonsmokers	Smokers	Nonsmokers	Smokers
35-44	18,793	52,407	2	32
45-54	10,673	43,248	12	104
55-64	5710	28,612	28	206
65-74	2585	12,663	28	186
75-84	1462	5317	31	102

Table 2: Q7

	Histology								
Disease Stage	I			II			III		
Follow-upTimeInterval(months)	1	2	3	1	2	3	1	2	3
0-2	9	12	42	5	4	28	1	1	19
	(157	134	212	77	71	130	21	22	101)
2-4	2	7	26	2	3	19	1	1	11
	(139	110	136	68	63	72	17	18	63)
4-6	9	5	12	3	5	10	1	3	7
	(126	96	90	63	58	42	14	14	43)
6-8	10	10	10	2	4	5	1	1	6
	(102	86	64	55	42	21	12	10	32)
8-10	1	4	5	2	2	0	0	0	3
	(88	66	47	50	35	14	10	8	21)
10-12	3	3	4	2	1	3	1	0	3
	(82	59	39	45	32	13	8	8	14)
12+	1	4	1	2	4	2	0	2	3
	(76	51	29	42	28	7	6	6	10)

Table 4.10 describes survival for 539 males diagnosed with lung cancer. The prognostic factors are histology (H) and state (S) of disease. The assumption of a constant rate over time is often not sensible, and this study divided the time scale (T) into two-month intervals and let the rate vary by the time interval. Let μ_{ijk} denote the expected number of deaths and t_{ijk} the total time at risk for histology i and state of disease j, in follow-up time interval k. Analyses suggested a lack of interaction between T and either prognostic factor (i.e., such proportional hazards models have the same effects of H and S for each time interval).

(a) The main effects model

$$log(\mu_{ijk}/t_{ijk}) = \alpha + \beta_i^H + \beta_j^S + \beta_k^T$$

has deviance 43.9. Explain why $df = 52$. Does the model seems to fit adequately?

(b) For this model, interpret the estimated effects of S,

$$\hat{\beta_2^S} - \hat{\beta_1^S} = 0.470(SE = 0.174)$$

$$\hat{\beta_3^S} - \hat{\beta_1^S} = 1.324(SE = 0.152)$$

(C) The model that adds an $S \times H$ interaction term has deviance 41.5 with $df = 48$. Test whether a significantly improved fit results by allowing this interaction.

Question 4

CDA exercise 4.16

For binary data, define a GLM using the log link. Show that effects refer to the relative risk. Why do you think this link is not often used? Hint: What happens if the linear predictor takes a positive value?

Question 5

CDA exercise 4.21

A binomial GLM $\pi_i = \phi(\sum_j \beta_j x_{ij})$ with arbitrary inverse link function ϕ assumes that $n_i Y_i$ has a $bin(n_i, \pi_i)$ distribution. Find w_i, in (4.29) and hence $cov(\hat{\beta})$. For logistic regression, show that $w_i = n_i \pi_i (1 - \pi_i)$.

Question 6

CDA exercise 4.27

For known k, show that the negative binomial distribution (4.13) has exponential family form (4.1) with natural parameter $log[\mu/(\mu + k)]$.