Homework Exercise Set 3: STA 7249

1. (11.23) Consider the model $\mu_i = \beta$, $i = 1, \ldots, n$, assuming $v(\mu_i) = \mu_i$. Suppose actually $\mathrm{Var}(Y_i) = \mu_i^2$. Using the univariate version of GEE, show $u(\beta) = \sum_i (y_i - \beta)/\beta$ and $\hat{\beta} = \bar{y}$. Show V in $\mathbf{V} = \left[\sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\right]^{-1}$ equals β/n, the actual asymptotic variance $\mathbf{V}\left[\sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)' [v(\mu_i)]^{-1} \mathrm{Var}(Y_i) [v(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\right]\mathbf{V}$ simplifies to β^2/n, and its consistent estimate is $\sum_i (y_i - \bar{y})^2/n^2$.

2. (11.25) Consider the model $\mu_i = \beta$, $i = 1, \ldots, n$, for independent Poisson observations. For $\hat{\beta} = \bar{y}$, show the model-based asymptotic variance estimate is $\bar{y}/n$, whereas the robust estimate of the asymptotic variance is $\sum_i (y_i - \bar{y})^2/n^2$. Which would you expect to be better (a) if the Poisson model holds, (b) if there is severe overdispersion?

3. (11.27)
   (a) For a univariate response, how is quasi-likelihood (QL) inference different from ML inference? When are they equivalent?
   (b) Explain the sense in which GEE methodology is a multivariate version of QL.
   (c) Summarize advantages and disadvantages of the QL approach.
   (d) Describe conditions under which GEE parameter estimators are consistent and conditions under which they are not. For conditions in which they are consistent, explain why.

4. (11.4) Refer to the depression data set analyzed in class. Analyze the data using the scores (1, 2, 4) for the week number, using ML or GEE with various working correlation structures. Interpret estimates and compare substantive results to those in the class example which used scores (0, 1, 2).

5. (11.7) Table 1 is from a Kansas State Univ. survey of 262 pig farmers. For the question "What are your primary sources of veterinary information?", the categories were (A) Professional Consultant, (B) Veterinarian, (C) State or Local Extension Service, (D) Magazines, and (E) Feed Companies and Reps. Farmers sampled were asked to select all relevant categories. The $2^5 \times 2 \times 4$ table shows the (yes, no) counts for each of these five sources cross-classified with the farmers' education (whether they had at least some college education) and size of farm (number of pigs marketed annually, in thousands).
   (a) Explain why it is not proper to analyze the data by fitting a multinomial model to the counts in the $2 \times 4 \times 5$ contingency table cross-classifying education by size of farm by the source of veterinary information, treating source as the response variable. (This table contains 453 positive responses of sources from the 262 farmers.)

Table 1:

			A = yes								A = no							
			B = yes				B = no				B = yes				B = no			
			C = yes		C = no		C = yes		C = no		C = yes		C = no		C = yes		C = no	
			Response on D															
Educ	Pigs	E	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
No	<1	Y	1	0	0	0	0	0	0	0	2	1	1	2	1	1	5	3
		N	0	0	0	0	0	0	0	1	1	0	0	5	4	7	7	0
	1-2	Y	2	0	0	0	0	0	0	0	4	0	0	4	1	0	0	4
		N	0	0	0	0	0	0	0	0	0	0	0	5	0	3	4	0
	2-5	Y	3	0	0	0	0	0	0	0	3	0	0	1	2	0	1	1
		N	1	0	0	0	0	0	0	3	0	0	0	2	0	1	4	0
	>5	Y	2	0	0	0	0	0	0	0	1	0	1	0	0	1	0	2
		N	1	0	0	2	1	0	1	6	0	1	1	1	0	0	6	0
Some	<1	Y	3	0	0	0	0	0	0	0	4	0	1	1	0	0	2	11
		N	0	0	0	0	0	0	0	0	4	0	1	2	4	6	14	0
	1-2	Y	0	0	0	0	0	0	0	0	2	0	0	1	0	0	1	6
		N	0	0	0	0	1	0	0	1	2	1	0	4	2	7	14	0
	2-5	Y	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	3
		N	1	0	0	0	0	0	0	0	0	0	0	5	0	4	4	0
	>5	Y	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	2
		N	1	1	0	0	0	1	0	10	0	0	0	4	1	2	4	0

Source: Prof. Tom Loughin, Kansas State Univ.

(b) For a farmer with education i and size of farm s, let $\pi_j(is)$ denote the probability of responding 'yes' on the jth source. Table 2 shows output for using GEE with exchangeable working correlation to estimate parameters in the model lacking an education effect,

$$\text{logit}[\pi_j(is)] = \alpha_j + \beta_j s, \quad s = 1, 2, 3, 4.$$

Explain how to interpret the working correlation matrix. Explain why the results suggest a strong positive size of farm effect for source A and perhaps a weak negative size effect of similar magnitude for C, D, and E.

(c) Constraining $\beta_3 = \beta_4 = \beta_5$, the ML estimate of the common slope is -0.184 ($SE = 0.063$). Explain why it is advantageous to fit the marginal model simultaneously for all sources rather than separately to each. (Agresti and Liu 1999 discussed analyses for data of this form.)

Table 2:

Working Correlation Matrix

	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.0997	0.0997	0.0997	0.0997
Row2	0.0997	1.0000	0.0997	0.0997	0.0997
Row3	0.0997	0.0997	1.0000	0.0997	0.0997
Row4	0.0997	0.0997	0.0997	1.0000	0.0997
Row5	0.0997	0.0997	0.0997	0.0997	1.0000

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Std Error	Z	Pr > |Z|
source	1	-4.4994	0.6457	-6.97	<.0001
source	2	-0.8279	0.2809	-2.95	0.0032
source	3	-0.1526	0.2744	-0.56	0.5780
source	4	0.4875	0.2698	1.81	0.0708
source	5	-0.0808	0.2738	-0.30	0.7680
size*source	1	1.0812	0.1979	5.46	<.0001
size*source	2	0.0792	0.1105	0.72	0.4738
size*source	3	-0.1894	0.1121	-1.69	0.0912
size*source	4	-0.2206	0.1081	-2.04	0.0412
size*source	5	-0.2387	0.1126	-2.12	0.0341