Categorical Data Analysis Exam II (2019)

1. (25 points) For each of the following statements, answer true (T) or false (F):

   (a) ( ) Cochran-Mental-Haenszel statistic (M^2) is a model-based test statistic for testing conditional independence.

   (b) ( ) When Y is binary, testing the Goodness-of-fit of loglinear model (XY, XZ, YZ) is equivalent to testing that there is no interaction between X and Z in their effects on Y in a logit model for Y.

   (c) ( ) Let Y be a response variable and X and Z be independent variables. The logit models corresponding to loglinear model (XY, YZ) and (XY, XZ, YZ) are not same.

   (d) ( ) A difference between logit and loglinear models is that the logit model is a generalized linear model assuming a binomial random component whereas the loglinear model is a generalized linear model assuming a Poisson random component. Here, when both are fitted to a contingency table having 50 cells, the logit model treats the cell counts as 25 binomial observations whereas the loglinear model treats the cell counts as 50 Poisson observations.

   (e) ( ) The cumulative logit model assumes that the response variable Y is ordinal; it should not be used with nominal variables. By contrast, the baseline-category logit model treats Y as nominal. It can be used with ordinal Y, but it then ignores the ordering information.

   (f) ( ) The cumulative logit model for J response categories corresponds to a logistic regression model holding for each of the $J-1$ cumulative probabilities, such that the curves for each cumulative probability have exactly the same shape (i.e., the same β parameter); that is, they increase or decrease at the same rate, so one can use $\hat{\beta}$ to describe effects that apply to all $J-1$ of the cumulative probabilities.

   (g) ( ) If X and Y are binary, and Z has K categories, so the data can be summarized in $2 \times 2 \times K$ contingency table, one can test conditional independence of X and Y, controlling for Z, using a Wald test or a likelihood-ratio test of $H_0 : \beta = 0$ in the model

   $$\text{logit}[P(Y = 1)] = \alpha + \beta x + \beta_1 z_1 + \cdots + \beta_{K-1} z_{K-1},$$

   where $z_i = 1$ for observations in category i of Z and $z_i = 0$ otherwise.

   (h) ( ) For a sample of retired subjects in Florida, a contingency table is used to relate X =cholesterol (8 ordered level) to Y =whether the subject has symptoms of heart disease (yes= 1, no= 0). For the linear logit model logit$[P(Y =$

1)] = $\alpha+\beta x$ fitted to the 8 binomials in the 8×2 contingency table by assigning scores to the 8 cholesterol levels, the deviance statistic equals 6.0. Thus, this model provides a poor fit to the data.

(i) ( ) In the example just mentioned, at the lowest cholesterol level, the observed number of heart disease cases equals 31. The standardized residual equal 1.35. This means that the model predicted 29.65 cases (i.e., 1.35=31-29.65).

(j) ( ) Matched pairs are the categorical responses to compare for two samples that have a natural pairing between each subject in one sample and a subject in the other sample. The samples are statistically independent.

2. (25 points) Let Y =political ideology (on an ordinal scale from 1 =very liberal to 5 =very conservative), x_1 =gender (1 =female, 0 =male), x_2 =political party (1 =Democrat, 0 =Republican).

(a) A main effects model with a cumulative logit link gives the output shown. Explain why the output reports four intercepts.

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept1		1	-2.5322	0.1489	-2.8242	-2.2403
Intercept2		1	-1.5388	0.1297	-1.7931	-1.2845
Intercept3		1	0.1745	0.1162	-0.0533	0.4023
Intercept4		1	1.0086	0.1232	0.7672	1.2499
gender	female	1	0.1169	0.1273	-0.1327	0.3664
gender	male	0	0.0000	0.0000	0.0000	0.0000
party	democ	1	0.9636	0.1297	0.7095	1.2178
party	repub	0	0.0000	0.0000	0.0000	0.0000

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
gender	1	0.84	0.3586
party	1	56.85	<.0001

(b) Explain how to describe gender effect on political ideology with an odds ratio.

(c) Give the hypotheses to which the LR statistic for gender refers, and explain how to interpret the result of the test.

(d) When we add an interaction term to the model, we get the output shown. Explain how to find the estimated odds ratio for the gender effect on political ideology for Republicans.

Parameter			DF	Estimate	Standard Error
Intercept1			1	-2.6743	0.1655
Intercept2			1	-1.6772	0.1476
Intercept3			1	0.0424	0.1338
Intercept4			1	0.8790	0.1389
gender	female		1	0.3661	0.1784
gender	male		0	0.0000	0.0000
party	democ		1	1.2653	0.1995
party	repub		0	0.0000	0.0000
gender*party	female	democ	1	-0.5091	0.2550
gender*party	female	repub	0	0.0000	0.0000
gender*party	male	democ	0	0.0000	0.0000
gender*party	male	repub	0	0.0000	0.0000

(e) Using the interaction model, show how to find the estimated probability that a female Republican is in the first category (very liberal).

3. (25 points) Consider the loglinear model for a two-way contingency table. Let μ_{ij} be expected frequencies in an $I \times J$ contingency table.

   (a) Using the following equation for expected frequencies, prove that X and Y are independent.

   $$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y.$$

   (b) To allow for association between X and Y, this model is extended to

   $$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}.$$

   For a 2×2 contingency table, express the log odds ratio in terms of expected frequencies, and use it to show that the odds ratio for this model equals $\exp(\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY})$.

4. (25 points) Polio was largely eradicated as a national health threat by the Salk vaccine, introduced by Jonas Salk and investigated in a nationwide clinical trial in the 1950's. The following $2 \times 2 \times 2$ table classifies 174 polio cases by age of subject (child, adult), paralytic status (yes, no), and vaccination status (yes, no). The data was collected in a cross-sectional study conducted in Des Moines, Iowa, in 1961.

		Salk Vaccine	
Paralysis	Age	Yes	No
Yes	Child(< 20)	32	47
	Adult(≥ 21)	3	7
No	Child(< 20)	45	17
	Adult(≥ 21)	13	10

The preceding three-way table was initially analyzed with loglinear models using PROC GENMOD in SAS. The following table features all possible loglinear models which could be fit to the data, their goodness-of-fit statistics and the associated p-values. (The variables age, vaccination status, and paralytic status are respectively denoted by A, V, and P.)

Model	G^2	p-value
(AVP)	–	–
(AV, AP, VP)	0.08	0.7810
(AV, AP)	16.79	0.0002
(AP, VP)	2.36	0.3073
(AV, VP)	9.18	0.0101
(AV, P)	24.01	0.0000
(AP, V)	17.19	0.0006
(VP, A)	9.58	0.0224
(A, V, P)	24.42	0.0001

(a) Suppose we try to find an appropriate loglinear model for the data. Using the preceding table, explain how the steps of the algorithm would proceed, starting with the saturated model and ending with a model which provides an acceptable fit.

(b) Use the BIC to determine which of the models (AV, AP, VP) or (AP, VP) is preferred. ($BIC = -2 \times$ maximized loglikelihood $+ p \times \log(N)$, where $p =$the number of parameters and $N =$ the number of observations)

(c) Assume that P is regarded as a response variable and A and V are regarded as explanatory variables. Which of the 9 different loglinear models would you select to characterize the data? (Choose one.) For this selected model, list the form of the equivalent logit model.

Henceforth, assume that the fitted model (AP, VP) holds.

(d) Let $\theta_{11(k)}$ denote the conditional (partial) odds ratio for A and V given that P is fixed at level k ($k = 1, 2$). Using the parametric form of the model (AP, VP), prove that $\theta_{11(k)} = 1$ for $k = 1, 2$. (Hint: $\theta_{11(k)} = 1$ if and only if $\log \theta_{11(k)} = 0$.)