STA 4504/5503
Sample questions for exam 2

1. True-False questions.

   (a) _____ For General Social Survey data on Y = political ideology (categories liberal, moderate, conservative), X_1 = gender (1 = female, 0 = male), and X_2 = political party (1 = Democrat, 0 = Republican), the ML fit of the cumulative logit model is $\text{logit}[\hat{P}(Y \leq j)] = \hat{\alpha}_j + .12x_1 + .96x_2$. Hence, for each gender, according to this model fit the estimated odds that a Democrat's response is liberal rather than moderate or conservative, and the estimated odds that a Democrat's response is liberal or moderate rather than conservative, is $e^{.96} = 2.6$ times the corresponding estimated odds for a Republican's response. This odds ratio estimate indicates that in this sample Democrats tended to be more liberal than Republicans.

   (b) _____ Subjects suffering from mental depression are measured after 1 week of treatment, 2 weeks of treatment, and 4 weeks of treatment in terms of a (normal, abnormal) response outcome. Covariates are severity of condition at original diagnosis (1 = severe, 0 = mild) and treatment used (1 = new, 0 = standard). Since each subject contributes three observations to the analysis, we can use the GEE (generalized estimating equations) method to fit the model. To use this method, we must choose a "working" correlation matrix for the form of the dependence among the three responses, but the method is robust in the sense that it still gives appropriate estimates and standard errors for large n even if the actual correlation structure is somewhat different from the one we assumed.

   (c) _____ 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. Hence, 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.

   (d) _____ 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.

   (e) _____ 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.

(f) _____ If X and Y are binary, and Z has K categories, so the data can be summarized in a $2 \times 2 \times K$ 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.

(g) _____ For a sample of retired subjects in Florida, a contingency table is used to relate X = cholesterol (8 ordered levels) to Y = whether the subject has symptoms of heart disease (yes = 1, no = 0). For the linear logit model $\text{logit}[P(Y=1)] = \alpha + \beta x$ 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.

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

2. Multiple choice question. Circle the letter(s) for the correct response(s). More than one response may be correct.

Let π denote the probability that a randomly selected respondent supports current laws legalizing abortion, predicted using gender of respondent ($G = 0$, male; $G = 1$, female), religious affiliation ($R_1 = 1$, Protestant, 0 otherwise; $R_2 = 1$, Catholic, 0 otherwise; $R_1 = R_2 = 0$, Jewish), and political party affiliation ($P_1 = 1$, Democrat, 0 otherwise; $P_2 = 1$, Republican, 0 otherwise, $P_1 = P_2 = 0$, Independent). The logit model with main effects has prediction equation

$$\text{logit}(\hat{\pi}) = .11 + .16G - .57R_1 - .66R_2 + .47P_1 - 1.67P_2$$

For this prediction equation,

a. Females are estimated to be more likely than males to support legalized abortion, controlling for religious affiliation and political party affiliation.

b. Controlling for gender and religious affiliation, the estimated odds that a Democrat supports legalized abortion equal $e^{.47-(-1.67)}$ times the estimated odds that a Republican supports legalized abortion.

c. The estimated probability that a male Jewish Independent supports legalized abortion equals $e^{.11}/(1+e^{.11})$.

d. The estimated probability of supporting legalized abortion is highest for female Jewish Independents.

3. 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).

4. You decide to use GEE methods to handle dependent observations because of repeated measurment or clustering of some type.

   a. Explain what is meant by an exchangeable "working correlation matrix."

   b. If you ignore the dependence, will there be bias in your (i) parameter estimates, (ii) standard error estimates?

5. Consider the loglinear model of independence for a two-way contingency table. This has equation for expected frequencies $\{\mu_{ij}\}$ in an $I \times J$ contingency table,

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

   Motivate this model, by showing how the definition of statistical independence of two categorical variables implies that a loglinear model of this form holds.

6. (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 $\exp(\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY})$. (Hence the two-factor interaction parameters provide information about the XY association.)

7. Consider the baseline-category logit model, for a multinomial response variable having J categories,

$$\log[P(Y = j)/P(Y = J)] = \alpha_j + \beta_j x, \quad j = 1, ..., J - 1.$$

Show how to use this model to generate a related logit model for $\log[P(Y = a)/P(Y = b)]$ using an arbitrary pair a and b of the response categories.

8. For the effect of a particular explanatory variable on an ordinal response variable, explain why the cumulative logit model has the same parameter for each logit, rather than a different parameter for each logit as is the case for the baseline-category logit model. Explain why the P-value for the effect is usually smaller with the cumulative logit model than with the baseline-category logit model.

Solutions

1. a, b, c, d, e, f, are True and g, h are False

2. a, b, c are correct.

3. a. Model refers to four cumulative probabilities, and they differ for any fixed value of explanatory variables. b. For females, estimated odds of response in liberal direction rather than conservative direction (for any of the four cutpoints) are 1.12 times estimated odds for males. c. $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$. If null were true, probability would equal 0.36 of obtaining LR statistic at least as large as observed (0.84). There is not much evidence of a gender effect. d. $e^{0.366} = 1.44$. e. $e^{-2.674+0.366}/[1 + e^{-2.674+0.366}] = 0.09$

4. a. Guess that each pair of observations on the response in a cluster has the same correlation. b. (i) no, (ii) yes.

5., 6., 7. See class notes.

8. For the cumulative logit model, each logit refers to the same thing – namely, the odds of falling below rather than above some point on an ordinal scale. For the baseline-category logit model, each logit deals with a different pair of outcome categories, so there is no reason to expect effects to be constant. The P value is usually smaller for the cumulative logit model because the effect is focused on fewer parameters, so df for the chi-squared test is smaller. Concentrating an effect on a smaller df value means the test statistic tends to be farther out in the right tail, hence smaller.

Formulas

$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \qquad \pi = \frac{\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\cdots+\beta_k x_k)}$

Baseline-category logit model: $\log[P(Y=j)/P(Y=J)] = \alpha_j + \beta_j x$

$$P(Y=j) = \frac{e^{\alpha_j+\beta_j x}}{1+e^{\alpha_1+\beta_1 x}+...+e^{\alpha_{J-1}+\beta_{J-1} x}},\ j=1,2,...,J-1.$$

Cumulative logit model: $\text{logit}\ [P(Y \le j)] = \alpha_j + \beta x$

$$P(Y \le j) = \exp(\alpha_j+\beta x)/[1+\exp(\alpha_j+\beta x)],\ j=1,2,...,J-1.$$

$z = (n_{12}-n_{21})/\sqrt{n_{12}+n_{21}}$ (McNemar)

SE for diff of matched proportions: $\frac{\sqrt{(n_{12}+n_{21})-(n_{12}-n_{21})^2/n}}{n}$

Kappa : $\kappa = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1-\sum_i \pi_{i+}\pi_{+i}}$

Independence loglinear model : $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$

(XY, XZ, YZ) : $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

(XZ, YZ) : $\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$