

1. 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

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

a. Explain what is meant by a “working correlation matrix.”

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

(ii) standard error estimates?

3. 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.

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

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

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.

5. 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. b

6. For the following questions, answer true (T) or false (F).

\_\_\_\_\_ 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.

\_\_\_\_\_ 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.

### Formulas for exam 3:

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

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

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

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

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

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

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

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

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

For comparing proportions with  $n$  matched pairs and counts  $b$  and  $c$  for numbers of different outcomes for the two observations, difference of sample proportions has estimated standard error

$$\frac{\sqrt{(b+c) - (b-c)^2/n}}{n}$$