Comprehensive Exam (2021)
Introduction to Categorical Data Analysis

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

   (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 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) ( ) 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.

   (c) ( ) 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.

   (d) ( ) 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

   (e) ( ) In the cumulative logit model with one independent variable x, for $j < k$, the curve for $P(Y \leq k)$ is the curve for $P(Y \leq j)$ translated by $(\alpha_k - \alpha_j)/\beta$ unit in the x direction.

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

   (g) ( ) 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.

   (h) ( ) 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.

   (i) ( ) Zero-inflated Poisson models can be used for both zero inflation and zero deflation.

(j) (   ) 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.

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.

```
                                       Standard   Wald 95% Confidence
Parameter             DF    Estimate      Error          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
                                     Chi-
            Source         DF      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.

```
                                            Standard
Parameter                  DF     Estimate     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) The logit model may be used for categorical data when there is a binary response variable. Consider a three-way table with variables X, Y and Z of levels I, 2 and K, respectively. The logit is defined as the log of the odds of success for someone having an i and j classification:

$$\log\left(\pi_{i1k}/\pi_{i2k}\right) = \log\left(\mu_{i1k}/\mu_{i2k}\right).$$

Consider the loglinear model (XY, YZ), where Y is the response variable. Show that under this model, the logit is of the form $\alpha + \beta_i + \gamma_k$ where α is a constant, β_i depends on $X = i$ alone and γ_k depends on $Z = k$ alone.

4. (25 points) Prove the following result:
For three-way table for three random variables (X, Y, Z), XY marginal and conditional odds ratios are identical if either (1) Z and X are conditionally independent or (2) Z and Y are conditionally independent.