Bios 765: Homework # 3, Fall 2020
due October 7, 2020

1. (**Applied 10 pts.**) The table below displays a 4-way cross-classification of data related to complaints of symptoms of a respiratory disease, byssinosis, which occurs among textile mill workers. Since the data arise from a cross-sectional study, loglinear model methods are a natural choice.

   a. Identify a good-fitting hierarchical loglinear model. Include in the model the three-way interaction of workplace, years employment, and smoking, regardless of its statistical significance. Write down the model using the ANOVA-like notation (effect coding) used in class notes 8, clearly defining all model terms. Report parameter estimates and their standard errors.

   b. Summarize goodness of fit of the model in **a** and interpret any interactions in the model involving the variable COMPLAINTS in terms of conditional odds ratios.

   c. With COMPLAINTS as the response variable, and using effect coding, fit the logistic regression model that corresponds to the loglinear model in **a**. Write down the logistic regression model, clearly defining all terms. Report parameter estimates and their standard errors. Provide interpretations for the results of this model fit (In other words, estimate the odds ratio effects of the explanatory variables and interpret them).

   d. Express, mathematically, relationships between the parameters in the log-linear model in **a.** and the logistic model in **c.**

   e. Referring to the results in **a.** and **c.**, demonstrate numerically the relationships established in **d.**

Table. Frequency table of Byssinosis Complaints

WORKPLACE	YEARS		COMPLAINTS	
CONDITIONS	EMPLOYMENT	SMOKING	yes	no
Dusty	<10	yes	30	203
Dusty	<10	no	7	119
Dusty	>=10	yes	57	161
Dusty	>=10	no	11	81
Not Dusty	<10	yes	14	1340
Not Dusty	<10	no	12	1004
Not Dusty	>=10	yes	24	1360
Not Dusty	>=10	no	10	986

2. (**Applied 10 pts.**) Provide a further analysis of the dumping syndrome severity data for duodenal ulcer patients (from Table 4 of the "Chapter 10" manuscript and considered in class Notes 10). Fit the equal adjacent odds ratio model with operation treated as a nominal variable. Specifically do the following:

   a. Write down the regression model equation, defining all terms using β-notation. Provide parameter estimates and their standard errors. In your solution, include the computer code you used to fit your model.

   b. For all 6 pairs of operations, provide the model predicted odds ratios for comparison of adjacent severity levels. Provide interpretations.

   c. For the 3 pairs of operations corresponding to a difference of 25% gastric tissue removed (i.e., V+D versus V+A; V+A versus V+H; V+H versus GR), provide 95% confidence intervals for the model predicted odds ratios for comparison of adjacent severity levels.

   d. Provide the predicted proportions with dumping syndrome severity for each level of severity, and for each operation.

   e. Provide a goodness-of-test for the linear by linear association model considered in class notes 10, slides 61-67, relative to the equal adjacent odds ratio model with operation treated as a nominal variable.

3. (**Theory, 10 pts.**) Derive the score statistic for logistic regression. Use the following notation and framework. Let X_A denote design matrix for the primary model. Let $X_E = [X_A, W]$ denote the expanded model. Assume both models have full rank. Let $\hat{\boldsymbol{\beta}}_A$ denote the m.l.e. for X_A. Let $\tilde{\boldsymbol{\beta}}_E = (\tilde{\boldsymbol{\beta}}'_A, \tilde{\boldsymbol{\beta}}'_W)'$ denote the m.l.e. for X_E. Let $\bar{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}'_A, \mathbf{0}')'$ denote restriction of $\tilde{\boldsymbol{\beta}}_E$ to have $\hat{\boldsymbol{\beta}}_W = \mathbf{0}$ relative to hypothesis $\boldsymbol{\beta}_W = \mathbf{0}$. Let

$$\boldsymbol{U}(\boldsymbol{\beta}_E) = \frac{\partial \log L}{\partial \boldsymbol{\beta}_E} = X'_E\big[\boldsymbol{y} - D_{\boldsymbol{n}}\boldsymbol{\pi}(\boldsymbol{\beta}_E)\big]$$

$$I(\boldsymbol{\beta}_E) = -\frac{\partial^2 \log L}{\partial \boldsymbol{\beta}_E \partial \boldsymbol{\beta}'_E} = X'_E D_{\boldsymbol{v}} X_E$$

$$D_{\boldsymbol{v}} = \text{Diag}(v_1, \ldots, v_s) \text{ where } v_i = n_i \pi_i(\boldsymbol{\beta}_E)(1 - \pi_i(\boldsymbol{\beta}_E))$$

$$\pi_i(\boldsymbol{\beta}_E) = \frac{\exp(x'_{iA}\boldsymbol{\beta}_A + w'_i\boldsymbol{\beta}_W)}{1 + \exp(x'_{iA}\boldsymbol{\beta}_A + w'_i\boldsymbol{\beta}_W)}$$

Note that $\pi_i(\bar{\boldsymbol{\beta}}) = \dfrac{\exp(x'_{iA}\hat{\boldsymbol{\beta}}_A)}{1 + \exp(x'_{iA}\hat{\boldsymbol{\beta}}_A)} = \pi_i(\hat{\boldsymbol{\beta}}_A)$.

Show that $Q_s = [\boldsymbol{U}(\bar{\boldsymbol{\beta}})]'[I(\bar{\boldsymbol{\beta}})]^{-1}[\boldsymbol{U}(\bar{\boldsymbol{\beta}})]$ equals the expression for the score statistic in the course notes 9 (slide 16) on logistic regression.

4. (**Theory, 10 pts.**) Derive background statistical theory for the multi-category logistic regression model that was summarized in notes 10, slides 22-27. Be sure to include derivations for the likelihood, score equations, Fisher information and asymptotic variance of $\hat{\boldsymbol{\beta}}$, and the asymptotic variance of the predicted probabilities $\hat{\boldsymbol{\pi}} = \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$. (Hint: Study the background theory for the single multinomial and the background theory for logistic regression presented in class.)