BIOS 765 Homework 3

Leo Li
730031954

Problem 1(a).

Our goal is to identify a good-fitting hierarchical loglinear model. Since the three-way interaction of workplace, years employment, and smoking should be included, it is natural to include all the main effects and two-way interactions associated with workplace, years employment, and smoking. In addition, all the three-way interactions have p-values larger than 0.05, so we choose to exclude all the three-way interaction other than the one associated with workplace, years employment, and smoking.

Therefore, in our model, we choose to include all the main effect, all the two-way interactions, as well as the three-way interactions of workplace, years employment, and smoking. The ANOVA-like representation of the model can be written as,

$$\begin{aligned}\log(\mu_{ijkl}) = \lambda_0 + \lambda^W_{1(i)} + \lambda^Y_{2(j)} + \lambda^S_{3(k)} + \lambda^C_{4(l)} + \lambda^{WY}_{12(ij)} + \lambda^{WS}_{13(ik)} + \lambda^{WC}_{14(il)} + \lambda^{YS}_{23(jk)} \\ + \lambda^{YC}_{24(jl)} + \lambda^{SC}_{34(kl)} + \lambda^{WYS}_{123(ijk)},\end{aligned}$$

where μ_{ijkl} is the expected counts in the corresponding cells resulted from the 4-way cross-classifications, and the notations for the covariates are defined as the following:

- The superscript W represents the workplace conditions, in which $i = 1$ if dusty; $i = 2$ if not dusty;
- The superscript Y represents the years employment, in which $j = 1$ if < 10; $j = 2$ if ≥ 10;
- The superscript S represents the smoking, in which $k = 1$ if yes; $k = 2$ if no;
- The superscript C represents the complaints, in which $l = 1$ if yes; $l = 2$ if no.

The ANOVA-like "sum to zero" constraints are specified as,

$$\sum_{i=1}^{2} \lambda^W_{1(i)} = \sum_{j=1}^{2} \lambda^Y_{2(j)} = \sum_{k=1}^{2} \lambda^S_{3(k)} = \sum_{l=1}^{2} \lambda^C_{4(l)} = 0,$$

$$\sum_{i=1}^{2} \lambda^{WY}_{12(ij)} = \sum_{j=1}^{2} \lambda^{WY}_{12(ij)} = 0,$$

$$\sum_{i=1}^{2} \lambda^{WS}_{13(ik)} = \sum_{k=1}^{2} \lambda^{WS}_{13(ik)} = 0,$$

$$\sum_{i=1}^{2} \lambda^{WC}_{14(il)} = \sum_{l=1}^{2} \lambda^{WC}_{14(il)} = 0,$$

$$\sum_{j=1}^{2} \lambda^{YS}_{23(jk)} = \sum_{k=1}^{2} \lambda^{YS}_{23(jk)} = 0,$$

$$\sum_{j=1}^{2} \lambda^{YC}_{24(jl)} = \sum_{l=1}^{2} \lambda^{YC}_{24(jl)} = 0,$$

$$\sum_{k=1}^{2} \lambda^{SC}_{34(kl)} = \sum_{l=1}^{2} \lambda^{SC}_{34(kl)} = 0,$$

$$\sum_{i=1}^{2} \lambda^{WYS}_{123(ijk)} = \sum_{j=1}^{2} \lambda^{WYS}_{123(ijk)} = \sum_{k=1}^{2} \lambda^{WYS}_{123(ijk)} = 0.$$

Table 1 presents the parameter estimates and their standard errors resulted from fitting the model above.

Table 1: Analysis of Parameter Estimates

Effect	Level	Estimate	SE
Intercept		4.391	0.049
W	1	-0.417	0.043
Y	1	-0.085	0.042
WY	11	0.069	0.023
S	1	0.391	0.047
WS	11	0.088	0.023
C	1	-1.581	0.049
WC	11	0.667	0.042
YS	11	-0.030	0.022
YC	11	-0.156	0.043
SC	11	0.153	0.048
WYS	111	-0.021	0.022

Problem 1(b).

According to our data resulted from the 4-way cross-classification, we realize that the sample size is sufficiently large, so that it is appropriate to use χ^2 approximation to the deviance. From model in part (a), we have a deviance of 8.103, which asymptotically follow

a χ^2_4 distribution, and the resulting p-value is 0.088, suggesting that the model in part (a) fits well.

Now, we would like to interpret the interaction terms in the model involving the variable COMPLAINTS in terms of conditional odds ratio.

- **Interpretation of $\lambda^{WC}_{14(il)}$:** $\lambda^{WC}_{14(11)}$ is estimated to be 0.667 (with 95% confidence intervals (0.584, 0.750)), so that $\widehat{OR}_{WC} = \exp(4\hat{\lambda}^{WC}_{14(11)}) = 14.42$ (with confidence intervals (10.34, 20.10)). This result can be interpreted as that, conditional on years employment and smoking, a person who work in dusty workplaces is 14.42 times more likely to complain about the symptom of byssinosis than a person who do not work in dusty workplaces. The corresponding confidence interval is (10.34, 20.10).

- **Interpretation of $\lambda^{YC}_{24(jl)}$:** $\lambda^{YC}_{24(11)}$ is estimated to be -0.156 (with 95% confidence intervals (-0.240, -0.073)), so that $\widehat{OR}_{YC} = \exp(4\hat{\lambda}^{YC}_{24(11)}) = 0.54$ (with confidence intervals (0.38, 0.75)). This result can be interpreted as that, conditional on workplace conditions and smoking, a person whose years of employment less than 10 years is 0.54 times as likely to complain about the symptom of byssinosis as a person whose years of employment is at least 10 years. The corresponding confidence interval is (0.38, 0.75).

- **Interpretation of $\lambda^{SC}_{34(kl)}$:** $\lambda^{SC}_{34(11)}$ is estimated to be 0.153 (with 95% confidence intervals (0.059, 0.246)), so that $\widehat{OR}_{SC} = \exp(4\hat{\lambda}^{SC}_{34(11)}) = 1.84$ (with confidence intervals (1.27, 2.67)). This result can be interpreted as that, conditional on workplace conditions and years employment, a person who smoke is 1.84 times more likely to complain about the symptom of byssinosis than a person whose do not smoke. The corresponding confidence interval is (1.27, 2.67).

Problem 1(c).

Now, we use COMPLAINTS as the response variable, and use effect coding, we would like to fit a logistic regression model that corresponds to the loglinear model in part (a). Now, assuming the i^{th} population is formed by the cross-classification of workplace conditions, years employment, and smoking. Let π_i represent the probability of complaints in the i^{th} population. Then, the logistic regression model can be written as the following:

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 W_i + \beta_2 Y_i + \beta_3 S_i,$$

where W_i represents the workplace condition for the i^{th} population (1 if dusty; -1 if not dusty), Y_i represents the years employment for the i^{th} population (1 if < 10; -1 if ≥ 10), and S_i represents the smoking for the i^{th} population (1 if yes; -1 if no). Table 2 presents the parameter estimates and their standard errors.

Now, we would like to provide interpretations for the results of this model fit. We have that,

Table 2: Analysis of Parameter Estimates

Effect	Estimate	SE
Intercept	-3.162	0.098
W	1.334	0.085
Y	-0.313	0.085
S	0.305	0.095

- **Odds ratio effects of the workplace conditions:** in this logistic regression model, β_1 is estimated to be 1.334 (with 95% confidence intervals (1.168, 1.500)), so that $\widehat{OR}_{WC} = \exp(2\hat{\beta}_1) = 14.42$ (with confidence intervals (10.34, 20.10)). This result can be interpreted as that, conditional on years employment and smoking, a person who work in dusty workplaces is 14.42 times more likely to complain about the symptom of byssinosis than a person who do not work in dusty workplaces. The corresponding confidence interval is (10.34, 20.10).

- **Odds ratio effects of the years employment:** β_2 is estimated to be -0.313 (with 95% confidence intervals (-0.479, -0.146)), so that $\widehat{OR}_{YC} = \exp(2\hat{\beta}_2) = 0.054$ (with confidence intervals (0.38, 0.75)). This result can be interpreted as that, conditional on workplace conditions and smoking, a person whose years of employment less than 10 years is 0.54 times as likely to complain about the symptom of byssinosis as a person whose years of employment is at least 10 years. The corresponding confidence interval is (0.38, 0.75).

- **Odds ratio effects of the smoking:** β_3 is estimated to be 0.305 (with 95% confidence intervals (0.119, 0.492)), so that $\widehat{OR}_{SC} = \exp(2\hat{\beta}_3) = 1.84$ (with confidence intervals (1.27, 2.67)). This result can be interpreted as that, conditional on workplace conditions and years employment, a person who smoke is 1.84 times more likely to complain about the symptom of byssinosis than a person whose do not smoke. The corresponding confidence interval is (1.27, 2.67).

Problem 1(d).

We would like to express the mathematical relationship between the parameters in the loglinear model in part (a) and the logistic model in part (c). We have that,

$$\beta_0 = 2\lambda^C_{4(1)},$$

$$\beta_1 = 2\lambda^{WC}_{14(11)},$$

$$\beta_2 = 2\lambda^{YC}_{24(11)},$$

$$\beta_3 = 2\lambda^{SC}_{34(11)}.$$

Problem 1(e).

Now we refer to Table 1 and Table 2 in part (a) and (c), so that we can demonstrate numerically the relationship established in part (d):

$$2\hat{\lambda}^{C}_{4(1)} = -1.581 \times 2 = -3.162 = \hat{\beta}_0,$$

$$2\hat{\lambda}^{WC}_{14(11)} = 0.667 \times 2 = 1.334 = \hat{\beta}_1,$$

$$2\hat{\lambda}^{YC}_{24(11)} = -0.156 \times 2 = -0.312 \approx -0.313 = \hat{\beta}_2,$$

$$2\hat{\lambda}^{SC}_{34(11)} = 0.153 \times 2 = 0.306 \approx 0.305 = \hat{\beta}_3.$$

Note that there might be some discrepancies due to rounding errors, but the result is sufficient to verify that the relationships established in part (d) is valid.

Problem 2(a).

We would like to fit the equal adjacent odds ratio model with operation treated as a nominal variable. The regression model equation can be written as the following:

$$\log(\frac{\pi_{i1}}{\pi_{i2}}) = \beta_{21} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

$$\log(\frac{\pi_{i2}}{\pi_{i3}}) = \beta_{31} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

where π_1, π_2, and π_3 are the probabilities of none, slight, and moderate severity for i^{th} population, respectively. and x_{i1}, x_{i2}, and x_{i3} are the indicators for operations "V+D", "V+A", and "V+H", respectively. That is, the reference operation is "GR". β's are the corresponding parameters of interest. Table 3 presents the parameter estimates and their standard errors, and the computer code used to fit the model is included in the Appendix.

Table 3: Analysis of Parameter Estimates

Parameter	Estimate	SE
β_{21}	0.384	0.161
β_{31}	0.784	0.196
β_1	0.457	0.208
β_2	0.377	0.200
β_3	0.139	0.189

Problem 2(b).

We would like to provide the model predicted odds ratios for comparison of adjacent severity levels for all 6 pairs of operations:

- **"V+D" verus "GR":** for duodenal ulcer data, the estimate of the common odds ratio for adjacent categories (none vs. slight or slight vs. moderate) is estimated as $\exp(0.457) = 1.579$ corresponding to comparing operations "V+D" to "GR".
- **"V+A" verus "GR":** for duodenal ulcer data, the estimate of the common odds ratio for adjacent categories (none vs. slight or slight vs. moderate) is estimated as $\exp(0.377) = 1.457$ corresponding to comparing operations "V+A" to "GR".
- **"V+H" verus "GR":** for duodenal ulcer data, the estimate of the common odds ratio for adjacent categories (none vs. slight or slight vs. moderate) is estimated as $\exp(0.139) = 1.149$ corresponding to comparing operations "V+H" to "GR".
- **"V+D" verus "V+A":** for duodenal ulcer data, the estimate of the common odds ratio for adjacent categories (none vs. slight or slight vs. moderate) is estimated as $\exp(0.457 - 0.377) = 1.084$ corresponding to comparing operations "V+D" to "V+A".
- **"V+D" verus "V+H":** for duodenal ulcer data, the estimate of the common odds ratio for adjacent categories (none vs. slight or slight vs. moderate) is estimated as $\exp(0.457 - 0.139) = 1.374$ corresponding to comparing operations "V+D" to "V+H".
- **"V+A" verus "V+H":** for duodenal ulcer data, the estimate of the common odds ratio for adjacent categories (none vs. slight or slight vs. moderate) is estimated as $\exp(0.377 - 0.139) = 1.268$ corresponding to comparing operations "V+A" to "V+H".

Problem 2(c).

In Table 4, we provide 95% confidence intervals for the model predicted odds ratio for comparison of adjacent severity levels for the 3 pairs of operations corresponding to a difference of 25% gastric tissue removed.

Table 4: Odds Ratio Estimates and 95% Confidence Intervals

Odds Ratio	Estimate	95% CI	
V+D vs V+A	1.084	0.705	1.665
V+A vs V+H	1.268	0.854	1.882
V+H vs GR	1.149	0.794	1.663

Problem 2(d).

In Table 5, we provide the predicted proportions with dumping syndrome severity for each level of severity, and for each operation.

Table 5: Predicted Proportions with Dumping Syndrome Severity for each Level of Severity, and for each Operation

Operation	Severity	Predicted proportions
V+D	none	0.643
V+D	slight	0.277
V+D	moderate	0.080
V+A	none	0.620
V+A	slight	0.290
V+A	moderate	0.091
V+H	none	0.547
V+H	slight	0.324
V+H	moderate	0.129
GR	none	0.502
GR	slight	0.342
GR	moderate	0.156

Problem 2(e).

We would like to provide a goodness-of-fit test for the linear by linear association model relative to the equal adjacent odds ratio model with operation treated as a nominal variable.

We use the likelihood ratio test to assess the adequacy of the linear by linear association model, and the test statistic is calculated to be:

$$LRT = 4.5898 - 4.4034 = 0.19,$$

which asymptotically follows χ^2_2 distribution, leading to a p-value of 0.91, suggesting that the linear by linear association model can provide sufficient goodness-of-fit.

Problem 3.

In this question, we would like to derive the score statistic for logistic regression. Let us consider each term of $Q_s = [U(\bar{\beta})]'[I(\bar{\beta})]^{-1}[U(\bar{\beta})]$, and we have that,

$$U(\bar{\beta}) = X_E'[y - D_n\pi(\bar{\beta})] = \begin{bmatrix} X_A' \\ W' \end{bmatrix}(y - \hat{\mu}) = \begin{bmatrix} X_A'(y - \hat{\mu}) \\ W'(y - \hat{\mu}) \end{bmatrix} = \begin{bmatrix} 0 \\ W'(y - \hat{\mu}) \end{bmatrix}.$$

The last equation holds because the top part of $U(\bar{\beta})$ is simply the score function of the primary model. In addition,

$$I(\bar{\beta}) = X_E' D_{\hat{\nu}} X_E = \begin{bmatrix} X_A' \\ W' \end{bmatrix} D_{\hat{\nu}} \begin{bmatrix} X_A & W \end{bmatrix} = \begin{bmatrix} X_A' D_{\hat{\nu}} X_A & X_A' D_{\hat{\nu}} W \\ W' D_{\hat{\nu}} X_A & W' D_{\hat{\nu}} W \end{bmatrix}.$$

Then, let $[I(\bar{\beta})]^{-1} = \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix}$, we will have that,

$$Q_s = [U(\bar{\beta})]'[I(\bar{\beta})]^{-1}[U(\bar{\beta})] = \begin{bmatrix} 0 & (y - \hat{\mu})'W \end{bmatrix} \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix} \begin{bmatrix} 0 \\ W'(y - \hat{\mu}) \end{bmatrix} = (y - \hat{\mu})' W I^{22} W'(y - \hat{\mu}),$$

in which I^{22} is the lower right component of $[I(\bar{\beta})]^{-1}$, which can be derived by the formula of the inverse of the block matrix. Then we have that,

$$\begin{aligned} I^{22} &= (W'D_{\hat{\nu}}W - W'D_{\hat{\nu}}X_A(X_A'D_{\hat{\nu}}X_A)^{-1}X_A'D_{\hat{\nu}}W)^{-1} \\ &= (W'(D_{\hat{\nu}} - D_{\hat{\nu}}X_A(X_A'D_{\hat{\nu}}X_A)^{-1}X_A'D_{\hat{\nu}})W)^{-1}. \end{aligned}$$

Plugging into the expression for Q_s, we have that,

$$\begin{aligned} Q_s &= (y-\hat{\mu})'W(W'(D_{\hat{\nu}} - D_{\hat{\nu}}X_A(X_A'D_{\hat{\nu}}X_A)^{-1}X_A'D_{\hat{\nu}})W)^{-1}W'(y-\hat{\mu}) \\ &= (y-\hat{\mu})'WV_G^{-1}W'(y-\hat{\mu}), \end{aligned}$$

as desired.

Problem 4.

In this question, we would like to derive the background statistical theory for the multi-category logistic regression model. We start by writing out the likelihood function as,

$$L = \prod_{i=1}^{s} n_i! \prod_{j=1}^{r} \frac{[\pi_{ij}(X_i;\beta)]^{y_{ij}}}{y_{ij}!}.$$

Then, the log likelihood function can be derived as,

$$\begin{aligned} l &= \sum_{i=1}^{s}[\log(n_i!) + \sum_{j=1}^{r}(y_{ij}\log[\pi_{ij}(X_i;\beta)] - \log(y_{ij}!))] \\ &= \sum_{i=1}^{s}[\log(n_i!) - \sum_{j=1}^{r}\log(y_{ij}!) + \sum_{j=1}^{r} y_{ij}\log[\pi_{ij}(X_i;\beta)]] \\ &= \sum_{i=1}^{s}[\log(n_i!) - 1_r'\log(y_i!) + \sum_{j=1}^{r} y_{ij}\log(\frac{\exp(x_{ij}'\beta)}{1_r'\exp(X_i\beta)})] \\ &= \sum_{i=1}^{s}[\log(n_i!) - 1_r'\log(y_i!) + \sum_{j=1}^{r} y_{ij}x_{ij}'\beta - \sum_{j=1}^{r} y_{ij}\log(1_r'\exp(X_i\beta))] \\ &= \sum_{i=1}^{s}[\log(n_i!) - 1_r'\log(y_i!) + y_i'X_i\beta - n_i\log(1_r'\exp(X_i\beta))]. \end{aligned}$$

Then the score functions are derived as,

$$\frac{\partial}{\partial\beta}l = \sum_{i=1}^{s}[X_i'y_i - \frac{n_iX_i'\exp(X_i\beta)}{1_r'\exp(X_i\beta)}] = \sum_{i=1}^{s}[X_i'y_i - n_iX_i'\pi_i]$$

The score equations can be obtained by setting the score functions above to 0, which lead to,

$$\sum_{i=1}^{s} n_iX_i'\hat{\pi}_i = \sum_{i=1}^{s} X_i'y_i.$$

The Fisher information can be derived as,

$$\begin{aligned}
-\frac{\partial^2}{\partial\beta^2}l &= \sum_{i=1}^{s}[\frac{n_i X_i' \exp(X_i'\beta)X_i}{1_r' \exp(X_i\beta)} - \frac{n_i X_i' \exp(X_i'\beta)\exp(X_i'\beta)'X_i}{(1_r' \exp(X_i\beta))^2}] \\
&= \sum_{i=1}^{s}[n_i X_i' D_{\pi_i} X_i - n_i X_i' \pi_i \pi_i' X_i] \\
&= \sum_{i=1}^{s} n_i X_i' (D_{\pi_i} - \pi_i\pi_i') X_i.
\end{aligned}$$

The asymptotic covariance matrix of $\hat{\beta}$ is the inverse of the Fisher information, so that,

$$Var(\hat{\beta}) = \{\sum_{i=1}^{s} n_i X_i' (D_{\pi_i} - \pi_i\pi_i') X_i\}^{-1}.$$

Finally, we would like to derive the asymptotic variance of the predicted probabilities $\hat{\pi}$. Note that the predicted probabilities can be expressed as $\hat{\pi}' = (\hat{\pi}_1', \ldots, \hat{\pi}_s')$, where the $\hat{\pi}_i$'s are the predicted probabilities from s independent populations. For each $\hat{\pi}_i$, we have that,

$$\frac{\partial}{\partial\beta}\pi_i = \frac{X_i' \exp(X_i'\beta)}{1_r' \exp(X_i'\beta)} - \frac{X_i' \exp(X_i'\beta)\exp(X_i'\beta)'}{[1_r' \exp(X_i'\beta)]^2} = X_i' D_{\pi_i} - X_i'\pi_i\pi_i' = X_i'(D_{\pi_i} - \pi_i\pi_i').$$

Therefore, $\frac{\partial}{\partial\beta}\pi = [X_1'(D_{\pi_1} - \pi_1\pi_1'), \ldots, X_s'(D_{\pi_s} - \pi_s\pi_s')]'$. Then by the delta method based on linear Taylor series approximation to $\hat{\pi}$, in a function of $\hat{\beta}$, the asymptotic covariance of $\hat{\pi}$ can be expressed as,

$$V_{\hat{\pi}}(\pi) = H(\pi) Var(\hat{\beta}) H'(\pi),$$

where $H(\pi) = [X_1'(D_{\pi_1} - \pi_1\pi_1'), \ldots, X_s'(D_{\pi_s} - \pi_s\pi_s')]'$, as desired.

Appendix: SAS Code

```
*********************************************
           Question 1
*********************************************;

data q1;
input W Y S C count;
* cond: 1:dusty, -1:not dusty;
* empl: 1:<10, -1:>=10;
* smok: 1:yes, -1:no;
* comp: 1:yes, -1:no;
cards;
 1  1  1  1 30
 1  1  1 -1 203
 1  1 -1  1 7
 1  1 -1 -1 119
 1 -1  1  1 57
 1 -1  1 -1 161
 1 -1 -1  1 11
 1 -1 -1 -1 81
-1  1  1  1 14
-1  1  1 -1 1340
-1  1 -1  1 12
-1  1 -1 -1 1004
-1 -1  1  1 24
-1 -1  1 -1 1360
-1 -1 -1  1 10
-1 -1 -1 -1 986
;
run;

title '1 (a): full model';
proc genmod data=q1 order=data;
  model count = W|Y|S W|Y|C W|C|S C|Y|S /d=p link=log;
run;

title '1 (a): selected model';
proc genmod data=q1 order=data;
  model count = W|Y W|S W|C Y|S Y|C S|C W*Y*S/d=p link=log;
run;

data q1c;
input W Y S yes total;
* cond: 1:dusty, -1:not dusty;
* empl: 1:<10, -1:>=10;
* smok: 1:yes, -1:no;
* comp: 1:yes, -1:no;
cards;
 1  1  1  30 233
 1  1 -1   7 126
 1 -1  1  57 218
 1 -1 -1  11 92
-1  1  1  14 1354
-1  1 -1  12 1016
```

```
-1 -1  1  24 1384
-1 -1 -1  10 996
;
run;

proc genmod data=q1c;
  model yes/total = w y s / d=b link=logit;
run;

************************************************
           Question 2
************************************************;

data q2;
   input Oper Sev nrow p;
   count=round(nrow*p,1);
cards;
0 1 96  .635
0 2 96  .292
0 3 96  .073
1 1 104 .654
1 2 104 .221
1 3 104 .125
2 1 110 .527
2 2 110 .364
2 3 110 .109
3 1 107 .495
3 2 107 .355
3 3 107 .150
;
run;

title '2 (a): equal adjacent odds ratio model';
proc logistic data=q2 order=data;
  class oper / param=ref;
  weight count;
  model sev = oper/link = alogit aggregate scale=none;
  oddsratio oper;
  output out=out pred=pred;
run;

title '2 (d) predicted probabilities';
proc print data=out;
run;

title '2 (e): linear by linear association model';
proc logistic data=q2 order=data;
  weight count;
  model sev = oper/link = alogit aggregate scale=none;
run;
```