Exercise 4.7 [13 points]

治file race view reagan others using vote.raw
(14 observations read)
glm total=reagan+view
list race view reagan total
race view reagan total
  1.  1  1  1 13
  2.  1  2  2 13
  3.  1  3  3 15
  4.  1  4  4 25
  5.  1  5  5 25
  6.  1  6  6 141
  7.  1  7  7 26
  8.  2  1  1  6
  9.  2  2  2 16
 10.  2  3  3 25
 11.  2  4  4 32
 12.  2  5  5  8
 13.  2  6  6  8
 14.  2  7  7  4

. glm reagan race view, f(b total)
Residual df = 11 No. of obs = 14
Pearson E2 = 11.51111 Deviance = 12.47503
Dispersion = 1.133663

Bilinear (Betaotal) distribution, logit link

reagan | Coef. Std. Err. z P>|z| [95% Conf. Interval]
---------+--------------------------------------------------------------------
 1 4 . 2704
 1 3 . 2629
 1 2 . 2508
 11. 2 4 13 0.973 0.028
 10. 2 3 2 2 0.741 1.484
 9 . 2508 0.603 -0.808
 8. 2 4 1 32 1.521 -0.433
 7. 1 4 155 301 145.941 1.045
 6. 2 3 2 25 0.741 1.484
 5. 1 3 44 115 42.356 0.390
 4. 2 1 2 13 0.603 -0.808
 3. 1 2 2 13 2.308 -0.949
 10. 1 5 92 153 92.707 -0.117
 11. 1 6 100 141 100.852 -0.159
 12. 2 6 2 9 1.107 0.906
 13. 2 7 7 0 4 0.417 -0.683
 14. 1 7 18 26 17.583 0.175

b. [6 points]

Treating vote as the response, there does seem to be a trend in the nominal main effects at the seven levels of political views. This models fits well and has a Pearson $\chi^2 \approx 4.18$ & the likelihood ratio $\chi^2 \approx 4.96$ in comparison with the saturated model with df=6. When one’s political view moves from extremely liberal(1) to extremely conservative(7), the odds likelihood of voting for Reagan rather than Carter or other; though extreme conservatives (7) had slightly lower odds of voting for Reagan than did those who were at (6). From the significant race main effect, we know the above is more so when the respondent is white, as opposed to non-white. Non-whites are $e^{-2.887} = .056$ times less likely to vote for Reagan than white voters, controlling for their political views. (We have only two levels of race here, so with or without dummy coding race, we should have the same estimate for the race main effect.) Examining the observed and estimated counts and the Pearson residuals, we know this model does not have extraordinarily bad fit.

b. [7 points]

Since there seems to be a trend in the ordinal main effects at the seven levels of political views, we can fit a logit model that uses the ordinal nature of political views. This models also fits well and has a Pearson $\chi^2 \approx 11.51$ & the likelihood ratio $\chi^2 \approx 12.47$ in comparison with the saturated model with df=11. $\beta_{\text{view}} = .491$ tells us that on a 7-point scale of political views, as one moves 1 point from liberal to conservative, s/he is $e^{.491} = 1.63$ times as likely to vote for Reagan as for Carter or others, controlling for the race main effect. Being a non-white, one is $e^{-2.937} = .053$ times as likely to vote for Reagan. With such a small odds ratio, it should be noted that 1) the observed and estimated probabilities of non-whites voting for Reagan never exceeds 25%; 2) the whites actually compose the majority of this data set. Therefore, even though this parsimonious model using the ordinal nature of political views can well explain the data and may be considered as a better model than the one in (a), we should be cautious about generalizing our conclusion onto the general population. The race or ethnicity does play an important
role in voters’ behavior. (In fact, if you perform analyses breaking down the race, you will not find a significant effect of political views using its ordinal nature on non-white respondents.)

**Exercise 5.12 [3 points]**

<table>
<thead>
<tr>
<th>X is independent of Y</th>
<th>A father’s gender is independent of his unborn 1st child’s gender.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y is independent of Z</td>
<td>the unborn 1st child’s gender is independent of the mother’s gender.</td>
</tr>
<tr>
<td>X is independent of Z?</td>
<td>the father’s gender is independent of the mother’s??? No! One has to be male and one has to be female!</td>
</tr>
<tr>
<td>⇒ Y is jointly independent of X and Z</td>
<td>the unborn 1st child’s gender is jointly independent of the father’s and the mother’s.</td>
</tr>
</tbody>
</table>

**Exercise 5.13 [4 points]**

<table>
<thead>
<tr>
<th>Religion (X)</th>
<th>Sexual Attitude (Y)</th>
<th>% Opposing abortion (Z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Religious</td>
<td>Conservative</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>Permissive</td>
<td>31%</td>
</tr>
<tr>
<td>Non-religious</td>
<td>Conservative</td>
<td>29%</td>
</tr>
<tr>
<td></td>
<td>Permissive</td>
<td>15%</td>
</tr>
</tbody>
</table>

According to the above hypothetical table, we can see that opposition to the legal availability of abortion is stronger among the religious (25+31=56%) than the non-religious (29+15=44%); and stronger among those with conservative sexual attitudes (25+29=54%) than those with more permissive attitudes (31+15=46%). However, it is not true that the religious (25%) are more likely than the non-religious (29%) to have conservative sexual attitudes. In other words, conditional (in)dependence (when holding Z constant at a level when discuss X & Y) does not imply marginal (in)dependence (disregarding Z when discuss X & Y).

**Exercise 5.14 [7 points]**

<table>
<thead>
<tr>
<th>Model</th>
<th>$\Delta G^2$</th>
<th>Best model</th>
</tr>
</thead>
<tbody>
<tr>
<td>fit(age smoking test)</td>
<td>Goodness-of-fit $\chi^2(0) = 0.000$</td>
<td></td>
</tr>
<tr>
<td>fit(age smoking, smoking test, test age)</td>
<td>Goodness-of-fit $\chi^2(1) = 20.656$ *</td>
<td></td>
</tr>
<tr>
<td>fit(age smoking, smoking test)</td>
<td>Goodness-of-fit $\chi^2(2) = 48.568$</td>
<td></td>
</tr>
<tr>
<td>fit(age smoking, age test)</td>
<td>Goodness-of-fit $\chi^2(2) = 47.613$</td>
<td></td>
</tr>
<tr>
<td>fit(age test, smoking test)</td>
<td>Goodness-of-fit $\chi^2(2) = 32.449$ *</td>
<td></td>
</tr>
<tr>
<td>fit(age, smoking test)</td>
<td>Goodness-of-fit $\chi^2(3) = 65.785$</td>
<td></td>
</tr>
<tr>
<td>fit(smoking, age test)</td>
<td>Goodness-of-fit $\chi^2(3) = 64.830$ *</td>
<td></td>
</tr>
<tr>
<td>fit(test, age smoking)</td>
<td>Goodness-of-fit $\chi^2(3) = 80.951$</td>
<td></td>
</tr>
<tr>
<td>fit(age, smoking, test)</td>
<td>Goodness-of-fit $\chi^2(4) = 98.166$</td>
<td></td>
</tr>
</tbody>
</table>

By fitting various models, we found that the saturated model is, strictly speaking, the best fitting model (of course!). The next best model in line is the no-3-way-interaction model with a high $\Delta G^2=20.656$ with df=1 when compared with the saturated model. Other models obviously do not fit well. Therefore, we proceed to examine how the no-3-way-interaction model fits differently from the saturated model.

The no-3-way-interaction model is actually the logit model treating the breathing test as the response. Since this model does not fit the data well, it tells us that age and smoking habit as the two main effects in the logit model do not explain the data well. The two insignificant coefficients of $AC_{22}$ and $BC_{22}$ in the saturated model also tell us the same. By examining the Pearson residuals, we see that this model fails to explain both counts of young and old non-smokers having abnormal breath test results. Perhaps some other variables must be considered in studying these
Caucasians in certain industrial plants in Houston, e.g. occupation: plant blue-collar workers vs. office personnel; or some other types of association between these three variables should be considered.

Nevertheless, the 3-way-interaction term cannot be dropped out of the saturated model in order to fit the data. That is, each pair of variables may be conditionally dependent, and an odds ratio for any pair may vary across levels of the third variable. (You would reach the same conclusion if you look at conditional odds ratios and marginal odds ratios by pairs of these three variables.)

Exercise 6.3 [7 points]

Model: \[ \Delta G^2 \] & Best model
---
fit(use eject, injure) & Goodness-of-fit chi2(0) = 0.00 & * 
fit(use eject, injure, injury use) & Goodness-of-fit chi2(1) = 3.00 & * 
fit(use eject, use injure) & Goodness-of-fit chi2(2) = 1681.00 & 
fit(use eject, injure, injury) & Goodness-of-fit chi2(2) = 1145.00 & 
fit(use injection, injure) & Goodness-of-fit chi2(2) = 7134.00 & 
fit(use injure, injury) & Goodness-of-fit chi2(3) = 9022.00 & 
fit(use injure, injure) & Goodness-of-fit chi2(3) = 9557.00 & 
fit(use, injure, injure) & Goodness-of-fit chi2(4) = 3568.00 & 
fit(use, injure, injure) & Goodness-of-fit chi2(4) = 11445.00 & 

---

\[ \text{loglin count age smoking test, fit(age smoking, age test, smoking test)} \] resid

<table>
<thead>
<tr>
<th>Variable age</th>
<th>A</th>
<th>Variable smoking</th>
<th>B</th>
<th>Variable test</th>
<th>C</th>
</tr>
</thead>
</table>

Marginals fit: age smoking, age test, smoking test

Note: Regression-like constraints are assumed. The first level of each variable (and all interactions with it) will be dropped from estimation.

---

**Poisson regression**

- **Number of obs = 8**
- **Log Likelihood = -36.078**
- **Pseudo R2 = 0.9645**
- **Prob > chi2 = 0.0000**
- **Goodness-of-fit chi2(1) = 20.656**
- **Model chi2(6) = 1960.091**

---

**Exercise 6.3 [7 points]**

Model: \[ \Delta G^2 \] & Best model
---
fit(use eject, injure) & Goodness-of-fit chi2(0) = 0.00 & * 
fit(use eject, injure, injury use) & Goodness-of-fit chi2(1) = 3.00 & * 
fit(use eject, use injure) & Goodness-of-fit chi2(2) = 1681.00 & 
fit(use eject, injure, injury) & Goodness-of-fit chi2(2) = 1145.00 & 
fit(use injection, injure) & Goodness-of-fit chi2(2) = 7134.00 & 
fit(use injure, injury) & Goodness-of-fit chi2(3) = 9022.00 & 
fit(use injure, injure) & Goodness-of-fit chi2(3) = 9557.00 & 
fit(use, injure, injure) & Goodness-of-fit chi2(4) = 3568.00 & 
fit(use, injure, injure) & Goodness-of-fit chi2(4) = 11445.00 & 

---
By fitting various loglinear models, we found that the no-3-way-interaction model is the best fitting model with \( \Delta G^2 = 3 \) and \( df = 1 \). Other models obviously do not fit well. Examining standardized residuals provides additional evidence that this model fits well. According to the coefficients, by not wearing seat belts, one is less likely not to be ejected (by a factor of \( e^{AB} = e^{2.40} \approx 11.02 \)); by not being ejected, one is less likely to be killed (by a factor of \( e^{BC} = e^{2.80} \approx 16.44 \)); by not wearing seat belts, one is more likely to be killed (by a factor of \( e^{AC} = e^{1.72} \approx 5.58 \)). This model states that all pairs of variables are conditionally dependent. The conditional odds ratios between any two variables are identical at each level of the third variable.

This no-3-way-interaction model is actually the logit model treating the whether killed as the response. Since this model does fit the data well, it tells us that wearing seat belts and being ejected are the two main effects in the logit model in whether being killed. In addition, the loglinear model tells us how wearing seat belts and being ejected are related to each other; whereas in the corresponding logit model no such interaction term is included.

Exercise 6.13 [6 points]

```
. input type dead expose
  type dead expose
  1. 1 10 170.4
  2. 2 18 147.3
  3. end
.
s.n:poisson dead i.type, e(expose)
i.type Itype_1-2 (naturally coded; Itype_1 omitted)
Iteration 0: Log Likelihood = -4.4662323
Iteration 1: Log Likelihood = -4.4473228
Iteration 2: Log Likelihood = -4.4473152
Poisson regression, normalized by expose Number of obs = 2
Goodness-of-fit chi2(0) = 0.000 Model chi2(1) = 3.632
Prob > chi2 = 0. Prob > chi2 = 0.0567
Log Likelihood = -4.447 Pseudo R2 = 0.2899

  +------------------------------------------------------------------
  | Coef. Std. Err. z P>|z| [95% Conf. Interval]
  +------------------------------------------------------------------
  Itype_2 | .7334638 .3944053 1.860 0.063 -.0395565 1.506484
  _cons | -2.835563 .3162278 -8.967 0.000 -3.455358 -2.215768
  +------------------------------------------------------------------
Iteration 0: Log Likelihood = -6.3975105
Iteration 1: Log Likelihood = -6.2634544
Iteration 2: Log Likelihood = -6.2631721
Iteration 3: Log Likelihood = -6.263176
Poisson regression, normalized by expose Number of obs = 2
Goodness-of-fit chi2(1) = 3.632 Model chi2(0) = -0.000
Prob > chi2 = 0. Prob > chi2 = .
Log Likelihood = -6.263 Pseudo R2 = -0.0000

  +------------------------------------------------------------------
  | Coef. Std. Err. z P>|z| [95% Conf. Interval]
  +------------------------------------------------------------------
  _cons | -2.428903 .1889822 -12.853 0.000 -2.799301 -2.058505
  +------------------------------------------------------------------
.
.predict lnrhat
.gen rhat=exp(lnrhat)
.gen mhat=rhat*expose
.list type dead expose lnrhat rhat mhat

  type dead expose lnrhat rhat mhat
  1. 1 10 170.4 -2.428903 .0881335 15.01794
  2. 2 18 147.3 -2.428903 .0881335 12.98206

By fitting the constant model, we found that this equal rate model is not significantly different from the saturated model, but not comfortably (\( \Delta G^2 = 3.632, p = .0567 \)). We further examined the observed and estimated counts, and found that the predicted counts of death actually head toward different direction than the observed counts. We have reasons to believe that actually this equal rate model does not fit data well, i.e. treatment A and B are related to different rates of death. However, we also found the insignificant coefficient of treatment effect (\( z = 1.86, p = .063 \)) in the saturated model. More data are needed to confirm whether treatment A and B are related to different rates of death. Data can be gathered by extending the duration of exposure (longer follow-ups) or retracting more patients receiving same treatments.