36-720 Homework 3 Solutions

Problem 1 (Agresti 7.1)

Let Y = belief in life after death (1=Yes, 2=Undecided, 3=No), x_1 = gender (1=females, 0=males) and x_2 = race (1=whites, 0=blacks). We fit the model

$$\log(\pi_j/\pi_3) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \quad j = 1, 2.$$

```
> mydata <- data.frame(R=c(rep("W",2),rep("B",2)), G=rep(c("F","M"),2))
> mydata <- cbind(mydata, Y=c(371,250,64,25), U=c(49,45,9,5), N=c(74,71,15,13))
> mydata$Rdummy <- ifelse(mydata$R=="W", 1, 0)
> mydata$Gdummy <- ifelse(mydata$G=="F", 1, 0)
> summary(vglm(cbind(Y,U,N)~Rdummy+Gdummy, family=multinomial, data=mydata))
...
Coefficients:
                Value Std. Error  t value
(Intercept):1  0.88305    0.24264  3.63931
(Intercept):2 -0.75801    0.36136 -2.09768
Rdummy:1       0.34177    0.23704  1.44186
Rdummy:2       0.27098    0.35413  0.76519
Gdummy:1       0.41855    0.17125  2.44402
Gdummy:2       0.10506    0.24651  0.42621
```

Note that these are the same estimates as those in the textbook.

(a) The estimated prediction equation for $\log(\pi_1/\pi_2)$ is

$$\begin{aligned}
\log\frac{\hat{\pi}_1}{\hat{\pi}_2} &= \log\frac{\hat{\pi}_1}{\hat{\pi}_3} - \log\frac{\hat{\pi}_2}{\hat{\pi}_3} \\
&= (0.883 + 0.419x_1 + 0.342x_2) - (-0.758 + 0.105x_1 + 0.271x_2) \\
&= 1.641 + 0.341x_1 + 0.071x_2
\end{aligned}$$

(b) The parameter β_1^G is the log of the conditional odds ratio for gender and Yes/No response. To see this, note that the conditional odds ratio is

$$\frac{\pi_1(x_1=1,x_2)/\pi_3(x_1=1,x_2)}{\pi_1(x_1=0,x_2)/\pi_3(x_1=0,x_2)} = \frac{\exp\{\alpha_1 + \beta_1^G + \beta_1^R x_2\}}{\exp\{\alpha_1 + \beta_1^R x_2\}} = \exp\{\beta_1^G\}.$$

A 95% Wald C.I. for β_1^G is $\hat{\beta}_1^G \pm 1.96 \times \widehat{SE}(\hat{\beta}_1^G) = 0.419 \pm 1.96 \times 0.171 = (0.084, 0.754)$, giving a 95% confidence interval for the odds of (1.087, 2.126). Because this confidence interval doesn't contain 1, there is evidence of a conditional effect of gender, controlling for race.

(c)

$$\begin{aligned}\hat{\pi}_1(x_1 = 1, x_2 = 1) &= \frac{\exp\{0.883 + 0.419 + 0.342\}}{1 + \exp\{0.883 + 0.419 + 0.342\} + \exp\{-0.758 + 0.105 + 0.271\}} \\ &= 5.176/(1 + 5.176 + 0.682) \\ &= 0.76\end{aligned}$$

(d) For black males, $x_1 = 0$ and $x_2 = 0$, so

$$\begin{aligned}\log(\hat{\pi}_1/\hat{\pi}_2) = \hat{\alpha_1} = 0.883 > 0 &\Rightarrow \hat{\pi}_1 > \hat{\pi}_3 \\ \log(\hat{\pi}_2/\hat{\pi}_2) = \hat{\alpha_1} = -0.758 < 0 &\Rightarrow \hat{\pi}_2 < \hat{\pi}_3.\end{aligned}$$

For black females, $x_1 = 1$ and $x_2 = 0$, so

$$\begin{aligned}\log(\hat{\pi}_1/\hat{\pi}_2) = \hat{\alpha_1} + \hat{\beta_1^G} = 1.302 > 0 &\Rightarrow \hat{\pi}_1 > \hat{\pi}_3 \\ \log(\hat{\pi}_2/\hat{\pi}_2) = \hat{\alpha_1} + \hat{\beta_1^G} = -0.653 < 0 &\Rightarrow \hat{\pi}_2 < \hat{\pi}_3.\end{aligned}$$

(e) Note that

$$\hat{\pi}_3(x_1, x_2) = \frac{1}{1 + \exp\{\hat{\alpha}_1 + \hat{\beta}_1^G x_1 + \hat{\beta}_1^R x_2\} + \exp\{\hat{\alpha}_2 + \hat{\beta}_2^G x_1 + \hat{\beta}_2^R x_2\}}$$

All of the estimates for the β parameters are positive, which means that $\hat{\pi}_3$ is maximized for $x_1 = 0$ and $x_2 = 0$, ie., for black males.

(f) For each model, there are 4 gender-race combinations and 3 parameters, so $df = 2(4) - 2(3) = 2$. The likelihood ratio statistic for testing the hypothesis that opinion is independent of gender, given race, is $G^2 = 8 - 0.9 = 7.1$, which has p-value 0.029 when compared to a χ^2 distribution with 2 degrees of freedom. Therefore we reject this hypothesis.

Problem 2 (Agresti 7.7) Here is the R code I used to fit the cumulative logit model, using the variable coding given in Table 7.18.

```
> mydata <- expand.grid(seatbelt=c("no", "yes"), location=c("urban", "rural"),
+                       gender=c("female", "male"))
> mydata <- cbind(mydata,
+                 lev1=c(7287,11587,3246,6134,10381,10969,6123,6693),
+                 lev2=c(175,126,73,94,136,83,141,74),
+                 lev3=c(720,577,710,564,566,259,710,353),
+                 lev4=c(91,48,159,82,96,37,188,74),
```

```
+                 lev5=c(10,8,31,17,14,1,45,12))
> mod <- vglm(cbind(lev1,lev2,lev3,lev4,lev5)~I(gender=="female")+
+             I(location=="rural")*I(seatbelt=="no"),
+             family=cumulative(parallel=T), data=mydata)
> summary(mod)
...
Coefficients:
                                                      Value Std. Error  t value
(Intercept):1                                       3.30742   0.035102  94.2233
(Intercept):2                                       3.48185   0.035562  97.9103
(Intercept):3                                       5.34938   0.046950 113.9385
(Intercept):4                                       7.25633   0.091428  79.3664
I(gender == "female")TRUE                          -0.54625   0.027211 -20.0748
I(location == "rural")TRUE                         -0.69885   0.042388 -16.4867
I(seatbelt == "no")TRUE                            -0.76016   0.039382 -19.3021
I(location == "rural")TRUE:I(seatbelt == "no")TRUE -0.12442   0.054765  -2.2718
```

(a) There are four intercepts because 4 logit models are needed to fit the 5 categories of responses. The baseline category is males in urban areas wearing seat belts, so $\alpha_j = \text{logit}[P(Y \leq j|x)], \quad j = 1, 2, 3, 4$. The estimated cumulative probabilities are $\frac{e^{3.3074}}{1+e^{3.3074}} = 0.965$, $\frac{e^{3.4818}}{1+e^{3.4818}} = 0.970$, $\frac{e^{5.3494}}{1+e^{5.3494}} = 0.995$, $\frac{e^{7.2563}}{1+e^{7.2563}} = 0.9993$, and 1. Te estimated distribution of responses is then (0.965, 0.005, 0.025, 0.0043, 0.0007). For example, when a male who is wearing his seatbelt has an accident in an urban area, there is a 96.5% chance that he will be in response category 1 (he will not be injured).

(b) A 95% Wald confidence interval for the log of the cumulative odds ratio for gender is $\hat{\beta}^G \pm 1.96 \times \widehat{SE}(\hat{\beta}^G) = -0.5463 \pm 1.96 \times 0.0272 = (-0.5996, -0.4930)$, so a 95% confidence interval for the cumulative odds ratio is $(0.549, 0.611)$. That is, the odds of being in a category less than or equal to j (where lower categories correspond to better outcomes) are between 0.549 and 0.611 times as much for females as they are for males. The confidence interval for the odds ratio doesn't contain 1, so the gender effect is significant.

(c) In rural areas, the cumulative odds ratio between the response and seat belt use, given gender, is $\exp(-0.7602 - 0.1244) = 0.413$. In urban areas, it is $\exp(-0.7602) = 0.468$. It seems that the effects of not wearing a seatbelt are worse in rural areas than in urban areas.

Finally, we compare the proportional odds model to the more general cumulative logit model. The likelihood ratio statistic is $154.4522 - 18.72489 = 135.7273$, which has a p-value of very nearly zero when compared to a χ^2 distribution with 12 degrees of freedom. This is evidence against the simpler (proportional odds) model.

Problem 3 (Agresti 7.11)

(a) Consider the baseline category logit models with additive factor effects of S and A. First note

$$\begin{aligned} df &= (\#\text{ Logits}) \times (\#\text{ Smoking levels}) \times (\#\text{ Age levels}) - \\ &\quad (\#\text{ Logits}) \times (\#\text{ Intercepts} + \#\text{ Age parameters} + \#\text{Smoking parameters}) \\ &= 2 \times 3 \times 2 - 2 \times (1 + 1 + 2) \\ &= 4. \end{aligned}$$

This model treats all categories as nominal: the coefficients correspond to dummy variables and the logit models are all with reference to an arbitrary baseline category.

(b) Now consider the model

$$\log \frac{P(B = k+1|S = i, A = j)}{P(B = k|S = i, A = j)} = \alpha_k + \beta_1 s_i + \beta_2 a_j + \beta_3 s_i a_j,$$

where B = breathing (1=Normal, 2=Borderline, 3=Abnormal), A = age (0=$<$ 40, 1=40^+), and S = smoking (-1=Never, 0=Former, 1=Current). Fitting the adjacent category model in R, we get the same parameter estimates as those in the book.

```
> mydata <- expand.grid(smoking=c("never", "former", "current"), age=c("<40", "40-59"))
> mydata <- cbind(mydata,
+                 normal=c(577,192,682,164,145,245),
+                 borderline=c(27,20,46,4,15,47),
+                 abnormal=c(7,3,11,0,7,27))
> mydata$smoking.num <- ifelse(mydata$smoking=="never", -1,
+                              ifelse(mydata$smoking=="former", 0, 1))
> mydata$age.dummy <- ifelse(mydata$age=="<40", 0, 1)
> mod <- vglm(cbind(normal, borderline, abnormal)~smoking.num*age.dummy,
+             family=acat(parallel=T), data=mydata)
> coef(mod)
       (Intercept):1        (Intercept):2          smoking.num
          -2.7554739           -1.5322144            0.1152401
           age.dummy smoking.num:age.dummy
           0.3113259            0.6631256
```

The linear part of the model has intercept depending on age and slope depending on smoking status and age. When age is less than 40, the fitted model is

$$\log \frac{P(B = k+1|S = i, A = 0)}{P(B = k|S = i, A = 0)} = \alpha_k + \beta_1 s_i,$$

and when age is greater than or equal to 40, the fitted model is

$$\log \frac{P(B = k+1|S = i, A = j)}{P(B = k|S = i, A = j)} = (\alpha_k + \beta_2) + (\beta_1 + \beta_3) s_i.$$

The interaction term is positive, indicating that the effect of smoking is greater for older workers.

(c) For age 40-59, the estimated odds ratio of abnormal rather than borderline breathing for current smokers compared to former smokers is

$$\begin{aligned}\frac{P(B=3|S=1,A=1)/P(B=2|S=1,A=1)}{P(B=3|S=0,A=1)/P(B=2|S=0,A=1)} &= \frac{\exp\{(\alpha_2+\beta_2)+(\beta_1+\beta_3)1\}}{\exp\{(\alpha_2+\beta_2)+(\beta_1+\beta_3)0\}} \\ &= \exp\{\beta_1+\beta_3\} = \exp\{0.778\} = 2.178.\end{aligned}$$

Compared to never smokers, the estimated odds ratio is

$$\begin{aligned}\frac{P(B=3|S=1,A=1)/P(B=2|S=1,A=1)}{P(B=3|S=-1,A=1)/P(B=2|S=-1,A=1)} &= \frac{\exp\{(\alpha_2+\beta_2)+(\beta_1+\beta_3)1\}}{\exp\{(\alpha_2+\beta_2)+(\beta_1+\beta_3)(-1)\}} \\ &= \exp\{2(\beta_1+\beta_3)\} = \exp\{2(0.778)\} = 4.74.\end{aligned}$$

Note that, because the β coefficients do not vary with k, the odds ratios above ($\exp\{\beta_1+\beta_3\}$ and $\exp\{2(\beta_1+\beta_3)\}$) are also equal to the corresponding odds ratios for borderline rather than normal breathing. Then we may write a given odds ratio for abnormal rather than normal breathing as the product of the two adjacent odds ratios and see that

$$\begin{aligned}\frac{P(B=3|x_1)/P(B=1|x_1)}{P(B=3|x_2)/P(B=1|x_2)} &= \frac{P(B=3|x_1)/P(B=2|x_1)}{P(B=3|x_2)/P(B=2|x_2)} \times \frac{P(B=2|x_1)/P(B=1|x_1)}{P(B=2|x_2)/P(B=1|x_2)} \\ &= \left[\frac{P(B=3|x_1)/P(B=2|x_1)}{P(B=3|x_2)/P(B=2|x_2)}\right]^2.\end{aligned}$$

Problem 4

(a) For testing H_0 : the common odds ratio in the 2×2 tables conditional on race of the victim is equal to 1, the Cochran Mantel Haenszel test for Table 2.6 gives a test statistic of 5.7959 (df=1, p-value=0.0161), and for table 2.13 it gives a test statistic of 1.2097 (df=1, p-value=0.271). So in the first case we reject H_0, but in the second case we do not.

(b) The odds ratios in each 2×2 subtable and the marginal odds ratios for each model are given below.

Model	$OR_{\text{White victims}}$	$OR_{\text{Black victims}}$	$OR_{\text{Collapsed}}$
(D,P,V)	1	1	1
(DV, PV)	1	1	1.654
(DV, PV, DP)	0.644	0.644	1.181
(DVP)	0.680	0	1.181

The model (D,P,V) specifies that all three variables are mutually independent. This implies conditional independence of D and P given V. Therefore all three odds ratios are 1. The model (DV, PV) specifies conditional independence of D and P given V, but D and P may be marginally dependent. This is why the conditional odds ratios are 1 but the odds ratio for the table ignoring victim's race is not. The model (DV, PV, DP) is the "homogenous association" model. It specifies that the relationship between any two variables does not change based on the third variable (ie., no three-way interaction). This is reflected in that the odds ratios are the same for each category of victim. The model (DVP), on the other hand, allows the odds ratios to vary by level of the third variable, as they do here.

Finally, we see that the marginal odds ratios for models (DV, PV, DP) and (DVP) are both equal to the odds ratio given by the raw counts. For both models, consider that the maximum likelihood estimates are given by setting the sufficient statistics equal to their expected values. For the saturated model this means $\hat{\mu}_{ijk} = n_{ijk} \, \forall i, j, k \Rightarrow \hat{\mu}_{ij+} = n_{ij+} \, \forall i, j$, and for the model (DV, PV, DP) this directly gives $\hat{\mu}_{ij+} = n_{ij+} \, \forall i, j$. In both cases, the fitted values in the collapsed table are the same as the raw counts in the collapsed table, meaning that the odds ratios will be equal as well.

(c) First we test whether the two subtables have a common odds ratio. This is the same as testing whether the model (DV, PV, DP) holds assuming that (DVP) holds, corresponding to line 4 in the table. With a p-value of 0.403, we fail to reject the null hypothesis that the odds ratio is the same (ie., that the interaction term is zero). Now we test whether this common odds ratio is equal to 1. This corresponds to line 3 in the table. With a p-value of 0.277, we fail to reject the null hypothesis that the shared odds ratio is 1. This agrees with part (a).

```
> anova(D.P.V, DV.PV, DV.PV.DP, DPV, test="Chisq")
Analysis of Deviance Table

Model 1: n2 ~ D + P + V
Model 2: n2 ~ D * V + P * V
Model 3: n2 ~ D * V + P * V + D * P
Model 4: n2 ~ D * P * V
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4    137.929
2         2      1.882  2  136.047 2.869e-30
3         1      0.701  1    1.181     0.277
4         0  4.123e-10  1    0.701     0.403
```

(d) Based on the output below, we conclude that there is a common odds ratio that is not equal to 1. This also agrees with part (a).

```
> anova(D.P.V, DV.PV, DV.PV.DP, DPV, test="Chisq")
Analysis of Deviance Table

Model 1: n1 ~ D + P + V
```

```
Model 2: n1 ~ D * V + P * V
Model 3: n1 ~ D * V + P * V + D * P
Model 4: n1 ~ D * P * V
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         4     402.84
2         2       5.39  2   397.44 4.974e-87
3         1       0.38  1     5.01      0.03
4         0  4.123e-10  1     0.38      0.54
```

(e) The output from `lors()` and `xtabs()` agrees with the facts given in the homework:

```
> lors(xtabs(n1~D+P))$lor
[1] 0.3689405
> lors(xtabs(n1~D+P+V))$lor
lors(xtabs(n1~D+P+V))$lor
[1] -0.8425514       -Inf
> lors(xtabs(n1~D+V))$lor
[1] 4.465384
> temp <- xtabs(n1~D+P+V)
> sum(temp[,,1])/sum(temp[,,2])
[1] 3.238994
```

Note also that there is a positive association between death penalty and victim's race, and the direction of the association doesn't change dependeing on the race of the defendant. Both white and black defendants are more likely to get the death penalty for killing a white person.

```
> lors(xtabs(n1~P+V))$lor
[1] 1.704546
> lors(xtabs(n1~P+V+D))$lor
[1]      Inf 2.335157
```

This helps explain how the direction of the association reverses from the conditional to the marginal tables above. Due to the strong association between race of defendant and race of victim, white defendants tend to have white victims, making it appear marginally that they are also more likely to get the dealth penalty.