Exercises from: Loglinear models for contingency tables

Chapter 7 Agresti (2007)

Saúl Sotomayor Leytón

February 2012

Setup

```
> library(vcd)
> library(MASS)
> options(contrasts = c("contr.treatment", "contr.poly"), useFancyQuotes = FALSE)
```

Problem 1

a) Remember that the goodness of fit test has the null hypothesis that all the parameters that are not included in the model of interest equal zero. Thus in the present exercise, the values indicate that the model fits well (P-values equal 0 for both the deviance test and the Pearson's chi-squared test), in other words the belief in after life is not independent on gender.

b)$\hat{\lambda}_j^Y$ represents the column (believe in after life) effect, and the difference between columns $\hat{\lambda}_1^Y - \hat{\lambda}_2^Y$ represents the log odds for believing in after-life. This value equals $\exp(1.4165) = 4.122$, thus, given gender, the odds of believing in after-life is 4.12 those of not believing.

Problem 2

According to Agresti (2007, p. 207) if a loglinear model for a two-way table contains an interaction term $\hat{\lambda}_{ij}^{XY}$ (*i.e.* the model is saturated), the interaction term represents the log odds ratio, thus the estimated odds for a female to believe in after-life is $\exp(0.1368) = 1.15$ 15% times higher than males.

Problem 3

a) First we construct the data and then fit the model.

```
> tb.ej3 <- data.frame(expand.grid(Busing = c(1, 2), President = c(1, 2), Home = c(1,
+     2)), Count = c(41, 72, 2, 4, 65, 175, 9, 55))
> fit.ej3 <- glm(Count ~ .^2, data = tb.ej3, family = poisson)
```

Once the model is fitted we look at the deviance value which is 0.48. This value has a approximate chi-squared distribution with degrees of freedom equal to, the number of cells (8) minus the number of parameters in the model (7). Thus the p-value for the null hypothesis is 0.49. The model fits well.

b) In order to calculate the odds ratio, we can append the fitted values (*i.e.* those from the model (BD,BP,DP)) to the data frame, and then calculate the odds ratio with the function with the same name (from the vcd package).

```
> tb.ej3$Fitted <- fitted(fit.ej3)
```

```
> # For the PB pair
> oddsratio(xtabs(Fitted ~ President + Busing + Home, data = tb.ej3))
```

```
log odds ratios for President and Busing by Home

        1         2
0.7210904 0.7210904
```

The value corresponds to the log odds, so to obtain the odds we exponentiate it $\exp(0.7210904) = 2.06$. Thus, conditioning on voting a black person for president, the odds of busing white and black students is twice the odds for not busing them. In a similar way we calculate the odds for the different pairs: PH = 4.72 and BH = 1.60. They can be interpreted as follows, conditioning on voting a black person for president, the odds of having brought a black friend for dinner is 4 times the odds of not having, and, conditioning on being in favor of busing together black and white kids the odds of having brought a black friend for dinner is 60% higher than not having.

c) This question ask to perform a test to see if BP association is significant. In order to do this, we need to fit a model that does not have this interaction, and then compare it to the one fitted in the previous question (*i.e.* fit.ej3). In the following command, which fits the reduced model, note that we just "updated" the previous fit removing the interaction that we want to test

```
> fit2.ej3 <- update(fit.ej3, . ~ . - Busing:President)
> anova(fit2.ej3, fit.ej3, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Count ~ Busing + President + Home + Busing:Home + President:Home
Model 2: Count ~ (Busing + President + Home)^2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2     5.1149
2         1     0.4794  1   4.6355  0.03132 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of 0.03132 provides evidence for the null hypothesis that all the parameters not included in the simpler model equal zero (*i.e.* $H_0 : BP = 0$); since it's below 0.05 we can reject that.
We can apply the same approach to test if more complex models fit better than (BP,BH,HP), for instance one that contains the triple interaction (BPH)

```
> fit.full.ej3 <- glm(Count ~ .^3, family = poisson, data = tb.ej3)
> anova(fit.ej3, fit.full.ej3, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Count ~ (Busing + President + Home)^2
Model 2: Count ~ (Busing + President + Home + Fitted)^3
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1    0.47939
2         0    0.00000  1  0.47939   0.4887
```

Note that in this case, the null hypothesis hold, in other words, the model with all the double interactions fits better than the one with a triple interaction.

d) To construct a 95% confidence interval we use the normal approximation: CI $= exp(Estimate \pm 1.96 * (StandardError))$

```
> # Store the estimate value in a temporary variable
> temp <- summary(fit.ej3)$coef["Busing:President", 1]
> # Store its standard error in another variable
> temp2 <- summary(fit.ej3)$coef["Busing:President", 2]
> # Calculate the CI. Note it's rounded
> round(exp(temp + 1.96 * c(-1, 1) * temp2), 2)
```

```
[1] 1.03 4.12
```

This is interpreted as: Given that one person votes a black person for president, we can be 95% sure that the odds for that person to be in favor of busing white and black kids together is at least 3% higher than being against busing, and at most, 4 times higher. This interval is very wide, though it doesn't include the value 1, which would mean equality in the odds.

Problem 4

a) First we need to construct the data frame and order the factors so the last variable is the one set to zero, then we need to fit the model (GH,GI,HI), one that includes all the two way interactions

```
> tb.ej4 <- data.frame(expand.grid(I = c(1, 2), G = c(1, 2), H = c(1, 2)), C = c(76,
+     6, 114, 11, 160, 25, 181, 48))
> # Correct factor ordering
> tb.ej4$I <- factor(tb.ej4$I, levels = c(2, 1))
> tb.ej4$G <- factor(tb.ej4$G, levels = c(2, 1))
> tb.ej4$H <- factor(tb.ej4$H, levels = c(2, 1))
> # Model fit
> fit.ej4 <- glm(C ~ .^2, family = poisson, data = tb.ej4)
```

Then we can test the goodness of fit with the value of the deviance (0.300721299332787) and a chi-squared approximation p-value = 0.58. The model seems to fit well.

b) As was done before, first we need to append the fitted values to the data frame and with those values we calculate the odds ratio with the oddsratio from the vcd package. Later to calculate the 95% confidence interval we follow the commands described in Thompson (2007), with some modifications, namely we store the estimate and its standard error in two temporal variables.

```
> tb.ej4$Fit <- fitted(fit.ej4)
> # Odds ratio
> log.odds <- oddsratio(xtabs(Fit ~ G + I + H, data = tb.ej4))
> exp(log.odds$coefficients)
```

```
       2        1
1.589807 1.589807
```

```
> # 95% CI
> temp <- summary(fit.ej4)$coef["I1:G1", 1]
> temp2 <- summary(fit.ej4)$coef["I1:G1", 2]
> exp(temp + 1.96 * c(-1, 1) * temp2)
```

```
[1] 0.9920389 2.5477685
```

Note that even though the odds for being in favor of information opinion is 58% higher for males than females, the confidence interval is so wide that includes the value 1, which means equality in the odds.

c) As previous, to test the Independence of G and I we fit a model without that interaction and compare to a model that includes that.

```
> fit.reduc.ej4 <- update(fit.ej4, . ~ . - I:G)
> anova(fit.reduc.ej4, fit.ej4, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ I + G + H + I:H + G:H
Model 2: C ~ (I + G + H)^2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2     4.1267
2         1     0.3007  1    3.826  0.05046 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the p-value, although slightly, is higher than the 0.05 cut point. This, together with the wide confidence interval, makes me think that the GI term shouldn't be included in the model. However, if we take the AIC value as a reference we note that the simpler model, despite having fewer terms, has a higher AIC value (61.51) compared to the model that includes the interaction (59.68).

Problem 5

a) Although the book provides a table for the output, I have fitted the model in R. Interesting, the values for the lower order terms are not the same as those reported in Agresti (2007, p. 235), though the values for the higher order terms are the same, as well as the deviance value.

The commands used to obtain this result were:

```
> tb.ej5 <- data.frame(expand.grid(D = c(1, 0), V = c(1, 0), P = c(1, 0)), C = c(53,
+     11, 0, 4, 414, 37, 16, 139))
> fit.ej5 <- glm(C ~ .^2, data = tb.ej5, family = poisson)
> summary(fit.ej5)
```

```
Call:
glm(formula = C ~ .^2, family = poisson, data = tb.ej5)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.93578    0.08471  58.265  < 2e-16 ***
D           -2.17465    0.26377  -8.245  < 2e-16 ***
V           -1.32980    0.18479  -7.196 6.19e-13 ***
P           -3.59610    0.50691  -7.094 1.30e-12 ***
D:V          4.59497    0.31353  14.656  < 2e-16 ***
D:P         -0.86780    0.36707  -2.364   0.0181 *
V:P          2.40444    0.60061   4.003 6.25e-05 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1225.07955  on 7  degrees of freedom
Residual deviance:    0.37984  on 1  degrees of freedom
AIC: 52.42

Number of Fisher Scoring iterations: 3
```

Perhaps there is some problem with the coding of the factor or their ordering.
More important, the conditional odds ratio is the same as that mentioned in the book, namely (0.42). This could be interpreted that, conditioning on the victim's race, the odds for a white person to receive the death penalty is 41% that for a black victim.

b) The model also yields the same marginal odds ratio (*i.e.* 1.45). This reverses the situation, with whites having an odds ratio of receiving the death penalty 45% higher than blacks. As stated in the book, this is an example of the Simpson's paradox, where "a relation present in different groups (conditional association) reverses when the groups are combined (marginal association)"[1].

c) The goodness of fit, based on the deviance value is calculated as usual p-value = 1. The model fits well.

d) Based on what is show in table 7.12 (Agresti 2007, p. 222), the corresponding logistic model would be $logitP = \alpha + \beta_i^D + \beta_j^V$. To fit this model we need to convert the data frame to include a column of "yes" counts as well as "totals" (see Dalgaard 2008, section 13.2).

```
> tb3.ej5 <- cbind(tb.ej5[1:4, 1:4], tb.ej5[5:8, 4])
> colnames(tb3.ej5) <- c("D", "V", "p", "yes", "no")
> tb3.ej5 <- tb3.ej5[, -3]
> tb3.ej5$total <- tb3.ej5$yes + tb3.ej5$no
```

But before we fit the model we need to create a variable of the proportions of yes responses, which will be used in the model fit

```
> temp <- tb3.ej5$yes/tb3.ej5$total
> fit.logistic.ej5 <- glm(temp ~ D + V, family = binomial, weights = total,
+     data = tb3.ej5)
```

Note that the residual deviance is the same as the loglinear model. More over, as stated in Agresti (2007, section 7.3.2), the calculated values for the odds ratios will be the same, for example for the logistic model, conditional on the victim's race, the odds that a white defendant to receive a death penalty is exp(-0.8678)=0.42, the same as what was reported in a). So the logistic model is $Logit(P = 1) = -3.5961 - 0.8678D + 2.4044V$ Where D=1 and V=1 for white defendants and victims, respectively.

[1] See https://secure.wikimedia.org/wikipedia/en/wiki/Simpson%27s_paradox

Problem 6

First the construction of the data frame

```
> tb.ej6 <- data.frame(expand.grid(J.P = c(1, 0), T.F = c(1, 0), S.N = c(1,
+     0), E.I = c(1, 0)), C = c(77, 42, 106, 79, 23, 18, 31, 80, 140, 52, 138,
+     106, 13, 35, 31, 79))
```

Then we fit the two models, the one for mutual independence and the one for homogeneous association

```
> fit.ind.ej6 <- glm(C ~ J.P + T.F + S.N + E.I, data = tb.ej6, family = poisson)
> fit.ej6 <- glm(C ~ .^2, data = tb.ej6, family = poisson)
```

a) Now, we can see how the two models fit the data. First for the mutual independence model.

```
> summary(fit.ind.ej6)
```

```
Call:
glm(formula = C ~ J.P + T.F + S.N + E.I, family = poisson, data = tb.ej6)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.92723    0.07501  52.354  < 2e-16 ***
J.P          0.12971    0.06185   2.097    0.036 *
T.F         -0.48551    0.06355  -7.640 2.17e-14 ***
S.N          0.87008    0.06765  12.861  < 2e-16 ***
E.I         -0.26439    0.06226  -4.246 2.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 399.94  on 15  degrees of freedom
Residual deviance: 135.87  on 11  degrees of freedom
AIC: 238.7

Number of Fisher Scoring iterations: 4
```

This model doesn't fit well, p-value = 0.

b) For the model of homogeneous association, we have:

```
> summary(fit.ej6)
```

```
Call:
glm(formula = C ~ .^2, family = poisson, data = tb.ej6)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  4.37035    0.09913  44.087  < 2e-16 ***
J.P         -0.95183    0.14661  -6.492 8.45e-11 ***
T.F         -1.00681    0.14898  -6.758 1.40e-11 ***
S.N          0.29141    0.12138   2.401  0.01636 *
E.I          0.01142    0.12516   0.091  0.92732
J.P:T.F      0.55936    0.13512   4.140 3.48e-05 ***
J.P:S.N      1.22153    0.14547   8.397  < 2e-16 ***
J.P:E.I      0.01766    0.13160   0.134  0.89326
T.F:S.N      0.40920    0.15243   2.684  0.00727 **
T.F:E.I     -0.19449    0.13121  -1.482  0.13826
S.N:E.I     -0.30212    0.14233  -2.123  0.03378 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 399.944  on 15  degrees of freedom
Residual deviance:  10.162  on  5  degrees of freedom
AIC: 125

Number of Fisher Scoring iterations: 4
```

Note that the estimate for the pair J.P:S.N is the greatest in magnitude; which reflects a stronger association.

c) For the homogeneous model, note, in the above summary, that the Wald p-value for the interactions J.P:E.I and T.F:E.I are not significant. So it may be appropriate to remove those interactions from the model. We test this with the anova function.

```
> fit2.ej6 <- update(fit.ej6, . ~ . - T.F:E.I - J.P:E.I)
> anova(fit2.ej6, fit.ej6, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ J.P + T.F + S.N + E.I + J.P:T.F + J.P:S.N + T.F:S.N + S.N:E.I
Model 2: C ~ (J.P + T.F + S.N + E.I)^2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         7     12.369
2         5     10.162  2    2.207   0.3317
```

The p-value of 0.3317 indicates that it's safe to remove those terms from the model.

Problem 7

First, note that the results from table 7.22 use a different coding for the factors, T.F and S.N. This affects the sign of the parameter estimates, but does not change the results from the model comparison. Thus for questions a and b we can work with the data and results from the previous exercise. Nevertheless, following is the code for constructing a table and calculating the results as they appear in Agresti (2007).

```
> tb.ej7 <- tb.ej6
> ## tb.ej7 <- within(tb.ej7, {
> ##     rm(Fit1)
> ## })
> tb.ej7$E.I <- factor(tb.ej7$E.I, levels = c(0, 1))
> tb.ej7$S.N <- factor(tb.ej7$S.N, levels = c(1, 0))  #Reverse coding
> tb.ej7$T.F <- factor(tb.ej7$T.F, levels = c(1, 0))  #Reverse coding
> tb.ej7$J.P <- factor(tb.ej7$J.P, levels = c(0, 1))
```

a) With the data with the correct coding we first fit the model with all the pairwise interactions, then the one without T.F:E.I and J.P:E.I.

```
> fit.full.ej7 <- glm(C ~ .^2, family = poisson, data = tb.ej7)
> fit2.ej7 <- update(fit.full.ej7, . ~ . - T.F:E.I - J.P:E.I)
```

Now we compare them with a chi-square test.

```
> anova(fit2.ej7, fit.full.ej7, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ J.P + T.F + S.N + E.I + J.P:T.F + J.P:S.N + T.F:S.N + S.N:E.I
Model 2: C ~ (J.P + T.F + S.N + E.I)^2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         7     12.369
2         5     10.162  2    2.207   0.3317
```

Just like the previous exercise, the value of 0.3317 indicates that the model fit well. Remember that this value is used for the null hypothesis that all the parameters that are not in the simpler model equal zero.

b) To calculate the 95% confidence interval:

```
> temp <- summary(fit2.ej7)$coef["J.P1:S.NO", 1]  #Estimate
> temp2 <- summary(fit2.ej7)$coef["J.P1:S.NO", 2]  #Standard error
> exp(temp + 1.96 * c(-1, 1) * temp2)
```

```
[1] 0.2220917 0.3922867
```

First, the estimate S.N:J.P (exp(-1.22021)=0.2951671) represents the conditional odds ratio for a person who was classified as "intuitive" of being "judging", compared to a person who was classified as "sensing". Now, the 95% CI of, (0.22,0.39), indicate that we can be 95% sure that the true conditional odds ratio, for a person that was first classified as "intuitive", of being classified as "judging" is at least 22% and at most 39% the odds for a person that was first classified "sensing". Note that the interval is narrow.

c) To change the coding of the S.N factor we use the relevel function, then we fit, again, the full and the reduced model[2]. At the end we would see that the S.N parameter and all those that interact with it

[2]Note that if we want to re-use the previous code we need to fit both models, since the reduced model was calculated with the update function.

change sign. This, of course, changes the value for the conditional odds ratio, and more importantly their interpretation. The estimate becomes exp(1.220214)= 3.387912 and for its confidence interval to, (2.55,4.50). This can be interpreted as, the conditional odds ratio for a person, classified in category 1 as "sensing", to be classified in a second category as "judging", is at least 2.5 times, and at most 4.5 times, the odds for a person classified in category 1 es "intuitive". Note that the relation with the other odds ratio and confidence intervals is through the *inverse* function.

Problem 8

a) The models for the mutual independence, conditional association and all the three-way interactions are, respectively:

$$
\begin{aligned}
log\mu_{ijkl} &= \lambda + \lambda_i^{J.P} + \lambda_j^{T.F} + \lambda_k^{S.N} + \lambda_l^{E.I} \\
log\mu_{ijkl} &= \lambda + \lambda_i^{J.P} + \lambda_j^{T.F} + \lambda_k^{S.N} + \lambda_l^{E.I} + \lambda_{ij}^{J.P:T.F} + \lambda_{ik}^{J.P:S.N} + \lambda_{il}^{J.P:E.I} + \lambda_{jK}^{T.F:S.N} + \lambda_{jl}^{T.F:E.I} + \lambda_{Kl}^{S.N:E.I} \\
log\mu_{ijkl} &= \lambda + \lambda_i^{J.P} + \lambda_j^{T.F} + \lambda_k^{S.N} + \lambda_l^{E.I} + \lambda_{ij}^{J.P:T.F} + \lambda_{ik}^{J.P:S.N} + \lambda_{il}^{J.P:E.I} + \lambda_{jK}^{T.F:S.N} + \lambda_{jl}^{T.F:E.I} + \lambda_{Kl}^{S.N:E.I} \\
&\quad + \lambda_{ijk}^{J.P:T.F:S.N} + \lambda_{ijl}^{J.P:T.F:E.I} + \lambda_{ikl}^{J.P:S.N:E.I} + \lambda_{jkl}^{T.F:S.N:E.I}
\end{aligned}
$$

b) To compare the models based on the AIC (akaike information criteria) we need to fit all the models; now, since we have done that for second model, we need to fit the mutual independence model and the one with all the three-way interactions:

```
> fit.ind.ej7 <- glm(C ~ J.P + T.F + S.N + E.I, family = poisson, data = tb.ej7)
> fit.sat.ej7 <- glm(C ~ .^3, family = poisson, data = tb.ej7)
```

Then we can extract the AIC value with the command, model$aic, and see that the values are, respectively: 238.7, 125 and 129.93. Based on these values the homogeneous association model seems to be the best one.

Problem 9

First we fit the data with the following commands. Note that for the factor "Department" we need to specify that it's a factor. After that we can fit the homogeneous association model

```
> tb.ej9 <- data.frame(expand.grid(D = 1:6, G = c(1, 0), A = c(1, 0)), C = c(512,
+     353, 120, 138, 53, 22, 89, 17, 202, 131, 94, 24, 313, 207, 205, 279, 138,
+     351, 19, 8, 391, 244, 299, 317))
> tb.ej9$D <- factor(tb.ej9$D, levels = 6:1)
> fit.ej9 <- glm(C ~ .^2, data = tb.ej9, family = poisson)
```

a) To fit the conditional and marginal odds ratio we need first to calculate the fitted values.

```
> fit.val.ej9 <- fitted(fit.ej9)
> # Conditional odds ratio
> oddsratio(xtabs(fit.val.ej9 ~ A + G + D, data = tb.ej9), log = FALSE)
```

```
 odds ratios for A and G by D

       6        5        4        3        2        1
0.904955 0.904955 0.904955 0.904955 0.904955 0.904955
```

```
> # Marginal odds ratio
> oddsratio(xtabs(fit.val.ej9 ~ A + G, , data = tb.ej9), log = FALSE)
```

```
 odds ratios for A and G

[1] 1.84108
```

According to Agresti (2007, p.367) the reason for such differences is that men apply in greater number to departments 1 and 2 with relatively high admissions rate, while women apply in greater number to departments 3 to 6 with low admission rate.

b) This model has a Deviance (G^2) value of 20.2 on 5 degrees of freedom. Through a chi-square approximation this has a p-value of 0, the fit is poor. However, by looking to the residuals, it is clear that the lack of fit it is only in department 1:

```
> tb.ej9$Res <- resid(fit.ej9, type = "pearson")/sqrt(1 - lm.influence(fit.ej9)$hat)
```

```
   D G A   C        Res
1  1 1 1 512 -4.0272821
2  2 1 1 353 -0.2797222
3  3 1 1 120  1.8808316
4  4 1 1 138  0.1412619
5  5 1 1  53  1.6334924
6  6 1 1  22 -0.3026439
7  1 0 1  89  4.0272889
8  2 0 1  17  0.2797222
9  3 0 1 202 -1.8808313
10 4 0 1 131 -0.1412619
11 5 0 1  94 -1.6334922
12 6 0 1  24  0.3026438
13 1 1 0 313  4.0272826
14 2 1 0 207  0.2797222
15 3 1 0 205 -1.8808313
16 4 1 0 279 -0.1412619
17 5 1 0 138 -1.6334922
18 6 1 0 351  0.3026438
19 1 0 0  19 -4.0272901
20 2 0 0   8 -0.2797222
21 3 0 0 391  1.8808314
22 4 0 0 244  0.1412619
23 5 0 0 299  1.6334922
24 6 0 0 317 -0.3026439
```

c) To delete department 1 we can use the subset function.

```
> tb2.ej9 <- subset(tb.ej9, D != 1)
> fit2.ej9 <- glm(C ~ .^2, family = poisson, data = tb2.ej9)
```

This model has a Deviance (G^2) value of 0 on 0 degrees of freedom. Through a chi-square approximation this has a p-value of 1, the fit seems well.

d) To fit an equivalent logistic model we need to transform the data so we can have a column of proportions of "yes" responses, or equivalent two columns, one of "yes" responses and other of "no" responses or "totals".

```
> tb3.ej9 <- data.frame(expand.grid(D = 2:6, G = c(1, 0)), Y = tb2.ej9$C[tb2.ej9$A ==
+     1], T = tb2.ej9$C[tb2.ej9$A == 1] + tb2.ej9$C[tb2.ej9$A == 0])
> tb3.ej9$D <- factor(tb3.ej9$D, levels = 6:2)
> fit3.ej9 <- glm(I(Y/T) ~ D + G, weights = T, data = tb3.ej9, family = binomial)
```

This yields the equation $logitA(P = 1) = -2.692 + 1.592D5 + 2.011D4 + 2.065D3 + 3.205D2 + 0.0307G$. So, conditioning on department, the odds, for a male, of being admitted is $\exp(0.0307) = 1.03$ those for females. This isn't a significant difference so we can conclude that admission is independent of gender[3].

Problem 10

First we construct the tables needed for the first two questions.

```
> tb.ej10 <- data.frame(expand.grid(E = c(1, 0), S = c(1, 0), I = c(0, 1)),
+     C = c(1105, 411111, 4624, 157342, 14, 483, 497, 1008))
> tb2.ej10 <- data.frame(expand.grid(E = c(1, 0), S = c(1, 0)), N = c(1105,
+     411111, 4624, 157342), Y = c(14, 483, 497, 1008))
> tb2.ej10 <- within(tb2.ej10, {
+     T <- Y + N
+     rm(N)
+ })
```

a) For the first question we first fit a model with all the pairwise interactions, *i.e.* a homogeneous association model. From that model we can use the stepAIC function to perform a stepwise removal of terms based on the value of AIC.

```
> fit.ej10 <- glm(C ~ .^2, family = poisson, data = tb.ej10)
> stepAIC(fit.ej10, scope = list(lower = C ~ 1), direction = "backward",
+         trace = TRUE)
```

```
Start:  AIC=93.85
C ~ (E + S + I)^2

       Df Deviance    AIC
<none>          2.9   93.9
- S:I   1   1144.6 1233.6
- E:I   1   1680.4 1769.4
- E:S   1   7134.0 7223.0

Call:  glm(formula = C ~ (E + S + I)^2, family = poisson, data = tb.ej10)
```

[3]Seeing also the summary output one can notice that the Wald p-value isn't significant for the gender factor

```
Coefficients:
(Intercept)            E            S            I          E:S          E:I
    11.9661      -3.5256       0.9605      -5.0436      -2.3996       2.7978
        S:I
    -1.7173

Degrees of Freedom: 7 Total (i.e. Null);  1 Residual
Null Deviance:            1625000
Residual Deviance: 2.854         AIC: 93.85
```

Note that removing any of the higher order terms significantly increases the deviance value, thus the best model is the homogeneous association model. As explained in Agresti (2007, p. 209) this type of model implies that the odds ratio between any two variables are the same at each level of the third variable.

b) The equivalent logistic model is fitted with:

```
> fit2.ej10 <- glm(I(Y/T) ~ E + S, weights = T, data = tb2.ej10, family = binomial)
```

This yields the equation, $logitI = -5.043 + 2.798E - 1.717S$. Based on this model we can say that the odds of having fatal injuries increases when the victim is ejected (E) from the vehicle (conditional odds ratio, 16.4 times compared to a non-ejected victim) and decreases when the victim wares seat belt (conditional odds ratio, 18% those who don't use seat belt).

c) The dissimilarity index is calculated as follows:

```
> sum(abs(tb.ej10$C - fitted(fit.ej10)))/(2 * sum(tb.ej10$C))
```

```
[1] 4.767967e-05
```

Remember from page 219 that this value represents the proportion of sample cases that must be removed to yield a perfect fit. Agresti (2007) uses this value to compare simpler models, but in this case I don't think it's informative, since, even removing all the pair-wise interactions yield a dissimilarity index of:

```
> #Mutual independence model
> fit.red3.ej10 <- update(fit.ej10, . ~ . - E:I - S:I - S:E)
> sum(abs(tb.ej10$C - fitted(fit.red3.ej10)))/(2 * sum(tb.ej10$C))
```

```
[1] 0.01402832
```

According to this value we need to move only 1% of the cases to have a perfect fit under a model of mutual independence. However, at least on theory, there is a clear relationship among the use of seat-belt and the magnitude of the injury. Moreover consider the implications of saying that there isn't a relationship between those two factors!.

Problem 11

The reasoning that Agresti (2007, p.367) uses is the following: Since there isn't a three way interaction he analyzes all the pairwise interactions that include the "injury" (I) factor; these are, GI =0.58, IL=2.13 and IS=0.44, and they can be interpreted as: For the first case, the conditional odds ratio for a male to get injured is 58% those for a female; for the second case, the conditional odds ratio of injuries in rural locations is 2 times for urban locations; and for the third case, the conditional odds ratio of getting injured if wearing seat belt is 44% the odds ratio for those who don't wear seat belt. Thus combining the information we agree with the information in that "the most likely case for injury is accidents for females not wearing seat belts in rural locations.

Problem 12

First of all, I assume that by "sensible" Agresti (2007) means that a particular model is easy to interpret.

Now, regarding the model fitting, first we need to construct the data. To this end, I constructed a data frame that could be used to fit a loglinear model (*i.e.* with a column of counts) and from that the tables needed for the two logistic models were constructed. Note that the compound function, as.data.frame(xtabs()), works pretty well for collapsing the data over a particular factor.

```
> tb.ej12 <- data.frame(expand.grid(S = c(0, 1), L = c(0, 1), G = c(0, 1), I = c(0,
+     1)), C = c(7287, 11587, 3246, 6134, 10381, 10969, 6123, 6693, 996, 759,
+     973, 757, 812, 380, 1084, 513))
> tb2.ej12 <- data.frame(expand.grid(S = c(0, 1), L = c(0, 1), G = c(0, 1)),
+     N = c(7287, 11587, 3246, 6134, 10381, 10969, 6123, 6693), Y = c(996, 759,
+         973, 757, 812, 380, 1084, 513))
> tb2.ej12 <- within(tb2.ej12, {
+     T <- N + Y
+     rm(N)
+ })
> tb3.ej12 <- as.data.frame(xtabs(C ~ S + L + G, data = tb.ej12))
> tb3.ej12 <- data.frame(expand.grid(S = c(0, 1), L = c(0, 1), G = c(0, 1)),
+     N = tb3.ej12[1:4, 4], Y = tb3.ej12[5:8, 4])
> tb3.ej12 <- within(tb3.ej12, {
+     T <- N + Y
+     rm(N)
+ })
```

On a second thought, "sensible" could mean appropriate. In this case only the second model is appropriate. The first one isn't because it involves collapsing the data over the injury factor, which isn't correct, since it isn't independent of, either, the location or the gender factor.

For that reason we only fit the second model:

```
> fit.ej12 <- glm(I(Y/T) ~ L + G, weights = T, family = binomial, data = tb3.ej12)
> fit2.ej12 <- glm(I(Y/T) ~ L + G + S, weights = T, family = binomial,
+                  data = tb2.ej12)
```

The equation for this model is: $logitI = -1.974 + 0.758L - 0.545G - 0.817S$. According to this model, the conditional odds ratio of getting injured increases in rural locations, decreases for males and

for those who wear seat belts.

Problem 13

First we construct the data. Note that because all the factors have more than two levels, we need to specify their correct ordering with the function factor.

```
> tb.ej13 <- data.frame(expand.grid(H = c(1, 2, 3), E = c(1, 2, 3), C = c(1,
+     2, 3), L = c(1, 2, 3)), N = c(62, 11, 2, 11, 1, 1, 3, 1, 1, 90, 22, 2,
+     21, 6, 2, 2, 2, 0, 74, 19, 1, 20, 6, 4, 9, 4, 1, 17, 7, 3, 3, 4, 0, 0,
+     0, 0, 42, 18, 0, 13, 9, 1, 1, 1, 0, 31, 14, 3, 8, 5, 3, 2, 2, 2, 5, 0,
+     1, 0, 0, 1, 0, 0, 0, 3, 1, 1, 2, 0, 1, 0, 0, 0, 11, 3, 1, 3, 2, 1, 1,
+     0, 3))
> tb.ej13$L <- factor(tb.ej13$L, levels = 3:1)
> tb.ej13$C <- factor(tb.ej13$C, levels = 3:1)
> tb.ej13$E <- factor(tb.ej13$E, levels = 3:1)
> tb.ej13$H <- factor(tb.ej13$H, levels = 3:1)
```

Now we can fit the homogeneous association model:

```
> fit.ej13 <- glm(N ~ .^2, data = tb.ej13, family = poisson)
```

a) This model has a Deviance value of 31.67 on 48 that has a p-value of 0.97; the model seems to fit well.

b)To calculate the conditional odds ratio we need to take into account that the factors of interest (E and H) have, each, more than two levels. For this reason we can not apply directly the oddsratio function over the xtabs one. We need to store the result of the latter function and then apply the former function with the proper indexing (see Dalgaard 2008, p. 21)

```
> temp <- xtabs(fitted(fit.ej13) ~ E + H + C + L, data = tb.ej13)
> oddsratio(temp[-2, -2, , ], log = FALSE)
```

```
 odds ratios for E and H by C, L

   L
C          3        2        1
  3 8.520614 8.520614 8.520614
  2 8.520614 8.520614 8.520614
  1 8.520614 8.520614 8.520614
```

Note that in the indexing we removed the level two of both factors of interest (which are in the same order as they were in the xtabs function). Worth remember is that the oddsratio function works on 2x2 tables, thus if, for instance, factor E had had 4 levels, and we are interested in the extreme ones, then the indexing would have been: `oddsratio(temp[c(-2,-3),-2, ,])`.

The relationship of $\lambda_{11}^{EH} + \lambda_{33}^{EH} - \lambda_{13}^{EH} - \lambda_{31}^{EH}$ can be explained by extending and relating what is mentioned in Agresti (2007, pages 205 and 207), namely, that the log odds ratio is related to the λ_{ij}^{XY} parameters of interest; all the other parameters, for example λ_{ij}^{AB} cancel out. Also note that because we

have set the third level to zero, all the parameters involving the third level equal zero, so the initial relationship simplifies to λ_{11}^{EH}.

Finally to construct the 95% confidence interval we store the estimate and its standard error in temporary variables, to ease the calculation.

```
> temp <- summary(fit.ej13)$coef["H1:E1", 1]
> temp2 <- summary(fit.ej13)$coef["H1:E1", 2]
> round(exp(temp + 1.96 * c(-1, 1) * temp2), 2)
```

```
[1]  2.86 25.37
```

c) As mentioned in the previous question, because the third category was set to zero, all the "too much", "too little" relationships simplify to λ_{11}^{XY}, where XY represent the pair of interest. All the pairs are summarized in the following table.

Pair	Log odds	Odds ratio	Std. Error	z value	Pr(>|z|)
H1:E1	2.1425	8.52	0.5566	3.849	0.000119
H1:C1	-0.1865	0.83	0.4547	-0.410	0.681730
H1:L1	1.8741	6.51	0.5079	3.690	0.000225
E1:C1	1.2000	3.32	0.5177	2.318	0.020448
E1:L1	-0.1328	0.88	0.6378	-0.208	0.835001
C1:L1	0.8735	2.4	0.4604	1.897	0.057811

Based on the Wald p-value it seems reasonable to drop pairs H:C and E:L. In fact these two pairs are dropped with the function stepAIC. Code shown below.

```
> stepAIC(fit.ej13, scope = list(lower = N ~ 1), direction = "backward")
```

```
Start:  AIC=304.5
N ~ (H + E + C + L)^2

       Df Deviance    AIC
- E:L   4   34.103 298.94
- H:C   4   36.966 301.80
<none>      31.669 304.50
- C:L   4   43.330 308.16
- E:C   4   47.357 312.19
- H:E   4   53.121 317.95
- H:L   4   57.216 322.05

Step:  AIC=298.94
N ~ H + E + C + L + H:E + H:C + H:L + E:C + C:L

       Df Deviance    AIC
```

```
- H:C   4   39.411 296.24
<none>      34.103 298.94
- C:L   4   46.276 303.11
- E:C   4   50.303 307.14
- H:E   4   56.292 313.13
- H:L   4   60.387 317.22

Step:  AIC=296.24
N ~ H + E + C + L + H:E + H:L + E:C + C:L

       Df Deviance    AIC
<none>      39.411 296.25
- C:L   4   53.232 302.07
- E:C   4   56.787 305.62
- H:E   4   62.776 311.61
- H:L   4   67.342 316.18

Call:  glm(formula = N ~ H + E + C + L + H:E + H:L + E:C + C:L, family = poisson,
    data = tb.ej13)

Coefficients:
(Intercept)           H2           H1           E2           E1           C2
    0.58191     -1.32407     -0.44957      0.22499      0.10376     -2.50014
         C1           L2           L1        H2:E2        H1:E2        H2:E1
   -2.21749     -0.17697     -0.02426      0.43704      0.75723      1.48833
      H1:E1        H2:L2        H1:L2        H2:L1        H1:L1        E2:C2
    2.16690      1.97215      1.20102      1.99649      1.95638      1.42753
      E1:C2        E2:C1        E1:C1        C2:L2        C1:L2        C2:L1
    1.48891      0.64429      1.15974      1.30742      0.51665      1.16445
      C1:L1
    0.82806

Degrees of Freedom: 80 Total (i.e. Null);  56 Residual
Null Deviance:            1370
Residual Deviance: 39.41         AIC: 296.2
```

Problem 14

First we construct the data frames. Note that the second table is constructed based on the first one. Also note that the P factor was explicitly stated as such.

```
> tb.ej14 <- data.frame(expand.grid(P = c(1, 2, 3), S = c(0, 1), R = c(0, 1),
+     B = c(0, 1)), C = c(99, 73, 51, 8, 20, 6, 73, 87, 51, 24, 50, 33, 15,
+     20, 19, 4, 13, 12, 25, 37, 36, 22, 60, 88))
> tb2.ej14 <- data.frame(expand.grid(P = c(1, 2, 3), R = c(0, 1), B = c(0, 1)),
+     Y = subset(tb.ej14, S == 1)$C, T = subset(tb.ej14, S == 1)$C + subset(tb.ej14,
+         S == 0)$C)
> tb.ej14$P <- factor(tb.ej14$P, levels = 3:1)
> tb2.ej14$P <- factor(tb2.ej14$P, levels = 3:1)
```

With the data in hand, the model selection was done. First a model with all triple interactions was fitted, and from that, a backward selection was performed with the function stepAIC from the MASS library.

```
> fit.full.ej14 <- glm(C ~ .^3, family = poisson, data = tb.ej14)
> fit.aic.ej14 <- stepAIC(fit.full.ej14, scope = list(lower = C ~ 1), trace = FALSE,
+     direction = "backward")
```

The model selected is the one with the triple interaction P:S:R and all the lower terms. However, by looking at the last step of the elimination process (table below) we can see that eliminating all the three-way interactions increases the deviance value only slightly

	Df	Deviance	AIC
<none>		2.508	160.95
P:S:R	2	6.963	161.40
R:B	1	16.203	172.64
P:B	2	28.427	182.87
S:B	1	59.541	215.98

b) The triple interaction P:S:R implies that the odds ratios among any two of those factors would vary across the levels of the third.
Now, regarding the calculation of the odds ratios, first, note that those involving the P factor can not be computed directly because it has three levels; instead we need to store the results of xtabs in temporary variables (The first three lines in the code below). With those variables we can calculate the odds radio for any pair we want, just remember that since we have set the third level of the P factor to zero, the odds ratio involving that level will be also zero. Also important is that the number used in the index doesn't represent the level of the factor but the line of the table formed with this factor. If we wouldn't have reversed the order (with the command (`tb.ej14$P <- factor(tb.ej14$P,levels=3:1)`) both things would mean the same, however since we did reverse the order, when we use, for example the number `-3` it means that we are dropping the first level of the factor.

```
> temp <- xtabs(fitted(fit.aic.ej14) ~ P + S + R + B, data = tb.ej14)
> temp1 <- xtabs(fitted(fit.aic.ej14) ~ P + R + S + B, data = tb.ej14)
> temp2 <- xtabs(fitted(fit.aic.ej14) ~ P + B + S + R, data = tb.ej14)
> oddsratio(temp[-3, , , ], log = FALSE)  #Odds ratio 2-3
```

```
 odds ratios for P and S by R, B

   B
R           0         1
  0 1.6179113 1.6179113
  1 0.7574738 0.7574738
```

```
> oddsratio(temp[-2, , , ], log = FALSE)  #Odds ratio 1-3
```

```
 odds ratios for P and S by R, B

   B
R           0         1
  0 0.5135238 0.5135238
  1 0.4352461 0.4352461
```

Pair	Odds Ratio	Pair	Odds Ratio
P2:S (R=0)	1.6179113	P2:R (S=0)	1.1581004
P2:S (R=1)	0.7574738	P2:R (S=1)	0.5421995
P1:S (R=0)	0.5135238	P1:R (S=0)	0.7698739
P1:S (R=1)	0.4352461	P1:R (S=1)	0.6525201
S:R (P=2)	2.1469	P2:B	0.537588
S:R (P=1)	3.886629	P1:B	0.3940744
S:B	3.147363	R:B	1.814253

According to these values, the strongest relationships are among the S and R or B factors. For the S:R case it means that the conditional odds for a person, who doesn't attend regularly to religious service, to be in favor of premarital sex is more than three times of those for someone who goes regularly to religious service; this, provided that the person's political tendency is liberal, if it is moderate the odds ratio comparing the previous two groups is 2 times of those who go regularly to religious service. For the S:B case, the conditional odds ratio, for people who go regularly to church, to be in favor of make available to teenagers bird control, is three times of those who go regularly to church.

To know exactly the meaning of the odds ratio, it's strongly advisable to check how the xtabs function classifies the values.

c) Because of the way the table was set up, the logistic model calculates the log odds ratio of being against premarital sex, in function of the other three predictors

```
> fit2.ej14 <- glm(I(Y/T) ~ P + R + B, weights = T, data = tb2.ej14,
+                  family = binomial)
```

As it is seen in the following table, almost all the parameters have significant Wald p-values, except for P2.

	Log odds ratio	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.57	0.21	0.19	7.96e-16	
P2	-0.08	0.92	0.17	-0.45	0.655
P1	-0.81	0.44	0.20	-3.96	7.59e-05
R	1.15	3.16	0.17	6.76	1.37e-11
B	1.15	3.16	0.15	7.50	6.29e-14

More over the model's deviance is 5.877 on 4 degrees of freedom (p-value = 0.21). The fit seems decent.

Because the model doesn't include an interaction term for P and R the odds ratios will be different. However the direction of the relationship is similar, for instance, the conditional odds ratio of being against premarital sex increase for those who attend regularly to religious service as well for those who

are against the availability for birth control to teenagers. Note that both values are equal and that for the B factor in the logistic regression model is also equal to the BS parameter for the loglinear model.

d) Since there are all the pairwise interactions, beside the three-way one between P:R:S none of the partial tables would be equal to the marginal ones.

Problem 15

First we set up the data frame. Next we fit a saturated model (*i.e.* one that has a four-way interaction) and with it we calculate the "best" model based on the AIC:

```
> tb.ej15 <- data.frame(expand.grid(C = c(1, 0), A = c(1, 0), R = c("W", "O"),
+     G = c("F", "M"), M = c(1, 0)), Count = c(405, 13, 1, 1, 23, 2, 0, 0, 453,
+     28, 1, 1, 30, 1, 1, 0, 268, 218, 17, 117, 23, 19, 1, 12, 228, 201, 17,
+     133, 19, 18, 8, 17))
> tb2.ej15 <- data.frame(expand.grid(C = c(1, 0), A = c(1, 0), R = c("W", "O"),
+     G = c("F", "M")), Y = c(405, 13, 1, 1, 23, 2, 0, 0, 453, 28, 1, 1, 30,
+     1, 1, 0), N = c(268, 218, 17, 117, 23, 19, 1, 12, 228, 201, 17, 133, 19,
+     18, 8, 17))
> tb2.ej15 <- within(tb2.ej15, {
+     T = Y + N
+     rm(N)
+ })
> fit.sat.ej15 <- glm(I(Y/T) ~ .^4, weights = T, data = tb2.ej15, family = binomial)
> fit.aic.ej15 <- stepAIC(fit.sat.ej15, scope = list(lower = I(Y/T) ~ 1),
+                         direction = "backward")
```

```
Start:  AIC=78
I(Y/T) ~ (C + A + R + G)^4

          Df   Deviance    AIC
- C:A:R:G  1 4.2247e-10 75.996
<none>       2.3088e-10 77.996

Step:  AIC=76
I(Y/T) ~ C + A + R + G + C:A + C:R + C:G + A:R + A:G + R:G +
    C:A:R + C:A:G + C:R:G + A:R:G

        Df Deviance    AIC
- C:A:G  1  0.11529 74.111
- A:R:G  1  0.17395 74.170
- C:A:R  1  0.21824 74.214
- C:R:G  1  1.54948 75.545
<none>      0.00000 75.996

Step:  AIC=74.11
I(Y/T) ~ C + A + R + G + C:A + C:R + C:G + A:R + A:G + R:G +
    C:A:R + C:R:G + A:R:G
```

```
        Df Deviance    AIC
- A:R:G  1  0.35897 72.355
- C:A:R  1  0.41915 72.415
- C:R:G  1  1.62806 73.624
<none>      0.11529 74.111

Step:  AIC=72.35
I(Y/T) ~ C + A + R + G + C:A + C:R + C:G + A:R + A:G + R:G +
    C:A:R + C:R:G

        Df Deviance    AIC
- A:G    1  0.61322 70.609
- C:A:R  1  1.03631 71.032
- C:R:G  1  1.89550 71.891
<none>      0.35897 72.355

Step:  AIC=70.61
I(Y/T) ~ C + A + R + G + C:A + C:R + C:G + A:R + R:G + C:A:R +
    C:R:G

        Df Deviance    AIC
- C:A:R  1  1.15629 69.152
- C:R:G  1  2.09891 70.095
<none>      0.61322 70.609

Step:  AIC=69.15
I(Y/T) ~ C + A + R + G + C:A + C:R + C:G + A:R + R:G + C:R:G

        Df Deviance    AIC
- A:R    1   1.2553 67.251
- C:A    1   1.5620 67.558
- C:R:G  1   2.7422 68.738
<none>       1.1563 69.152

Step:  AIC=67.25
I(Y/T) ~ C + A + R + G + C:A + C:R + C:G + R:G + C:R:G

        Df Deviance    AIC
- C:A    1   1.6080 65.604
- C:R:G  1   2.8474 66.843
<none>       1.2553 67.251

Step:  AIC=65.6
I(Y/T) ~ C + A + R + G + C:R + C:G + R:G + C:R:G

        Df Deviance     AIC
- C:R:G  1    3.177  65.173
<none>        1.608  65.604
- A      1   92.673 154.668
```

```
Step:  AIC=65.17
I(Y/T) ~ C + A + R + G + C:R + C:G + R:G

       Df Deviance     AIC
- C:R   1    3.186  63.181
- R:G   1    3.191  63.187
- C:G   1    4.673  64.668
<none>      3.177  65.173
- A     1   93.913 153.908

Step:  AIC=63.18
I(Y/T) ~ C + A + R + G + C:G + R:G

       Df Deviance     AIC
- R:G   1    3.201  61.197
- C:G   1    4.681  62.677
<none>      3.186  63.181
- A     1   93.940 151.935

Step:  AIC=61.2
I(Y/T) ~ C + A + R + G + C:G

       Df Deviance     AIC
- C:G   1    4.694  60.689
<none>      3.201  61.197
- R     1    5.369  61.364
- A     1   94.265 150.260

Step:  AIC=60.69
I(Y/T) ~ C + A + R + G

       Df Deviance    AIC
<none>       4.69  60.69
- R     1     6.89  60.88
- G     1    15.06  69.05
- A     1    95.48 149.47
- C     1   502.23 556.22
```

a) This approach yields a main effects model, namely one that does not have an interaction term:

$$logitM = -5.4640 + 2.8592C + 2.9873A - 0.2989R^O + 0.3297G^M \tag{1}$$

About this model two things should be noted, first that during the backward selection process a warning message appeared, (`glm.fit: fitted probabilities numerically 0 or 1 occurred`); second, that based on the Wald p-values for the parameters in the final model, it seems sensible to drop the race effect. In fact, fitting a model without the race term and comparing this with the model above (with a chi-squared test) yields a value for the null hypothesis, that all the parameters not included in the reduced model are zero, equal to 0.1384. Thus the final equation is:

$$logitM = -5.5162 + 2.8591C + 3.0202A + 0.3279G^M \tag{2}$$

The code for to obtain this result was:

```
> fit.aic2.ej15 <- update(fit.aic.ej15, . ~ . - R)
> anova(fit.aic2.ej15, fit.aic.ej15, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: I(Y/T) ~ C + A + G
Model 2: I(Y/T) ~ C + A + R + G
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        12     6.8890
2        11     4.6939  1   2.1952   0.1384
```

b) This model is equivalent to the following loglinear model:

$$\begin{aligned} log\mu_{ijkl} &= \lambda + \lambda_i^C + \lambda_j^A + \lambda_k^G + \lambda_l^M \\ &\quad + \lambda_{il}^{CM} + \lambda_{jl}^{AM} + \lambda_{kl}^{GM} + \lambda_{ij}^{CA} + \lambda_{ik}^{CG} + \lambda_{jk}^{AG} \end{aligned}$$

However it should be noted that ending in this model by fitting "all" possible loglinear models isn't easy; first because there are many interactions among the R factor and the others, thus when we see the Wald p-values some seem significant while others don't. This complicates any further analysis after the one performed by the stepAIC function (see the summary for model `fit2.aic.ej15`). Even if we don't take into account the race factor, the stepAIC returns an even more simpler model, one that drops the interaction C:G.
At the end, I think, this underlines how, when applicable, logistic models are simpler to fit and to interpret (see Agresti 2007, section 7.3.2).

Problem 16

a) The equivalence is explained in Agresti (2007, section 7.3.1) and for this particular exercise it's explained with pencil in the notes. There it's said that the equivalence is reflected in the parameters of both models, which should be equal. Thus we can compare those parameters an if they are equal the models are equivalent. For the loglinear model, it was already fitted in the notes (see `fit3.accident`), while for the logistic model we fit it with:

```
> tb.ej16 <- data.frame(expand.grid(Gender = c(0, 1), Location = c(0, 1), Seat = c(0,
+     1), Injury = c(0, 1)), Count = c(7287, 11587, 3246, 6134, 10381, 10969,
+     6123, 6693, 996, 759, 973, 757, 812, 380, 1084, 513))
> tb2.ej16 <- cbind(subset(tb.ej16, Injury == 1), T = subset(tb.ej16, Injury ==
+     1)$Count + subset(tb.ej16, Injury == 0)$Count)
> tb2.ej16 <- within(tb2.ej16, {
+     rm(Injury)
+     Y <- Count
+     rm(Count)
+ })
> fit.ej16 <- glm(I(Y/T) ~ Seat + Location + Gender, weights = Y, family = binomial,
+     data = tb2.ej16)
```

The coefficients of interest are showed in the following table.

Parameter	Odds ratio	Estimate	Est. error	P value
Logistic model				
Seat	-0.83	-0.83176	0.09137	$<$ 2e-16
Locationrural	0.77	0.76560	0.08410	$<$ 2e-16
Gendermale	-0.53	-0.53327	0.08272	1.15e-10
Loglinear model				
Seat:Injury	-0.82	-0.81710	0.02765	$<$ 2e-16
Locationrural:Injury	0.76	0.75806	0.02697	$<$ 2e-16
Gendermale:Injury	-0.54	-0.54483	0.02727	$<$ 2e-16

Note that even though they are not exactly the same they are very similar.

b) As explained in the note, because we have set the last level to zero, for each of the factor, the odds ratios are basically the parameter values themselves. Now for the odds ratio for the seat effect as it is shown in the table above they are also very similar: -0.83 and -0.82 for the logistic model and the loglinear model, respectively

Problem 17

Agresti (2007) mentions that loglinear models are appropriate when working with multiple response variables and it is of interest to establish a relationship among them. If there is only one explanatory variable, logistic models are easier to fit and interpret.

Problem 18

See notes

Problem 19

See notes

Problem 20

See notes

Problem 21

The model fitted in exercise 13 corresponds to the homogeneous association model, *i.e.* one with all the pairwise associations. Thus we need to fit only the model with all the three way interactions, and then compare both with a chi-square test.

```
> fit2.ej13 <- glm(N ~ .^3, data = tb.ej13, family = poisson)
> anova(fit.ej13, fit2.ej13, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: N ~ (H + E + C + L)^2
Model 2: N ~ (H + E + C + L)^3
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        48     31.669
2        16      8.524 32   23.146   0.8737
```

The differences of deviances equals 23.146 on 32 degrees of freedom, which has a p-value of 0.8737. The simpler model fits well.

For the second part, the function stepAIC can be used

```
> fit.aic.ej13 <- stepAIC(fit.ej13, scope = list(lower = N ~ 1),
+                         direction = "backward")
```

```
Start:  AIC=304.5
N ~ (H + E + C + L)^2

       Df Deviance    AIC
- E:L   4   34.103 298.94
- H:C   4   36.966 301.80
<none>      31.669 304.50
- C:L   4   43.330 308.16
- E:C   4   47.357 312.19
- H:E   4   53.121 317.95
- H:L   4   57.216 322.05

Step:  AIC=298.94
N ~ H + E + C + L + H:E + H:C + H:L + E:C + C:L

       Df Deviance    AIC
- H:C   4   39.411 296.24
<none>      34.103 298.94
- C:L   4   46.276 303.11
- E:C   4   50.303 307.14
- H:E   4   56.292 313.13
- H:L   4   60.387 317.22

Step:  AIC=296.24
N ~ H + E + C + L + H:E + H:L + E:C + C:L

       Df Deviance    AIC
<none>      39.411 296.25
- C:L   4   53.232 302.07
- E:C   4   56.787 305.62
- H:E   4   62.776 311.61
- H:L   4   67.342 316.18
```

This results in the model , the same as Agresti. As it's shown in the independence graph (see notes), this model implies that E and L are conditionally independent on the values of CH, and similarly C and H are conditionally independent on EL.

b) *See notes*

Problem 22

See notes

Problem 23

See notes

Problem 24

First we construct the data. The second command adds the score variables which will be used later, while the fourth and fifth command specifies that R and B are factors, as well as their ordering.

```
> tb.ej24 <- data.frame(expand.grid(R = 1:9, B = 1:4), C = c(49, 31, 46, 34,
+     21, 26, 8, 32, 4, 49, 27, 55, 37, 22, 36, 16, 65, 17, 19, 11, 25, 19,
+     14, 16, 15, 57, 16, 9, 11, 8, 7, 16, 16, 11, 61, 20))
> tb.ej24 <- cbind(tb.ej24, u1 = 1:9, v1 = rep(1:4, each = 9))
> tb.ej24$R <- factor(tb.ej24$R, levels = 9:1)
> tb.ej24$B <- factor(tb.ej24$B, levels = 4:1)
```

a) Next we fit the model of independence, and create another table with the observed counts together the fitted values and its residuals. As described in Agresti (2007, p. 229) greater deviations occur at the corner of the table, however they are not the greatest. These occur one row before the lower corners.

```
> fit0.ej24 <- glm(C ~ R + B, family = poisson, data = tb.ej24)
> tb2.ej24 <- tb.ej24[, 1:3]
> tb2.ej24$Fit0 <- round(fitted(fit0.ej24), 2)
> tb2.ej24$Res0 <- round(resid(fit0.ej24, type = "pearson")/
+                        sqrt(1 - lm.influence(fit0.ej24)$hat), 2)
> tb2.ej24
```

```
   R B  C  Fit0  Res0
1  1 1 49 34.15  3.20
2  2 1 31 21.68  2.45
3  3 1 46 36.32  2.03
4  4 1 34 26.29  1.86
5  5 1 21 19.79  0.33
6  6 1 26 25.48  0.13
7  7 1  8 13.55 -1.82
8  8 1 32 58.28 -4.60
9  9 1  4 15.45 -3.52
10 1 2 49 44.09  0.99
11 2 2 27 27.99 -0.24
12 3 2 55 46.89  1.59
13 4 2 37 33.94  0.69
14 5 2 22 25.54 -0.91
15 6 2 36 32.89  0.71
16 7 2 16 17.49 -0.46
17 8 2 65 75.23 -1.67
18 9 2 17 19.94 -0.84
19 1 3 19 26.13 -1.68
20 2 3 11 16.59 -1.61
21 3 3 25 27.78 -0.64
```

```
22 4 3 19 20.11 -0.29
23 5 3 14 15.14 -0.34
24 6 3 16 19.49 -0.94
25 7 3 15 10.37  1.66
26 8 3 57 44.58  2.38
27 9 3 16 11.82  1.41
28 1 4  9 21.63 -3.21
29 2 4 11 13.74 -0.85
30 3 4  8 23.01 -3.72
31 4 4  7 16.66 -2.75
32 5 4 16 12.53  1.12
33 6 4 16 16.14 -0.04
34 7 4 11  8.59  0.93
35 8 4 61 36.92  4.97
36 9 4 20  9.79  3.70
```

b) The next step is to fit the *linear by linear*.

```
> fit1.ej24 <- glm(C ~ R + B + u1:v1, family = poisson, data = tb.ej24)
```

This model has a β parameter of 0.1215 with a standard error of 0.0134. This means that there is a positive linear association between religious attendance and opposition for birth control. People who go frequently to church are more likely to be against birth control availability to teenagers.

c) The model fitted in b) has a deviance value of 19.901 on 23 degrees of freedom (p-value of 0.65), the model fits well. We can also compare both models with a chi-square test to see if the β parameter is significant (*i.e.* different from zero).

```
> anova(fit0.ej24, fit1.ej24, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: C ~ R + B
Model 2: C ~ R + B + u1:v1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        24    112.536
2        23     19.901  1   92.634 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The extremely low p-value indicates that the linear association is significant.

d) To fit the linear by linear model with a different column code we do the following:

```
> tb.ej24$v2 <- rep(c(1, 2, 4, 5), each = 9)
> fit2.ej24 <- glm(C ~ R + B + u1:v2, family = poisson, data = tb.ej24)
```

The new coding changes the strength of the linear association from 0.1215 to 0.0836.

e) As Agresti (2007) notes, the value of β not only shows the direction and strength of the linear association, but also help us to calculate the odds ratio for any 2x2 sub-table. When we use an equally spaced coding and we want to calculate the odds ratio for contiguous rows and columns, the odds ratio simplifies to $\exp(\beta)$. For the model in d), however, the columns aren't equally spaced so the odds ratio becomes $exp(\beta(3-2)(4-2)) = exp(\beta x 2)$.

Problem 25

a) Controlling for a categorical variable the linear by linear model generalizes to:

$$log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \tag{3}$$

If X and Y are conditionally independent then the value of β becomes zero and the above equation simplifies to the model (XZ,YZ).

b) Comparing with a chi-squared test models with and without the linear association we would get a test of conditional independence based on 1 degree of freedom.

c) As explained for the previous exercise, if we calculate local odds ratio, the equation: $exp(\beta(u_{c+1} - u_c)(v_{d+1} - v_d))$ becomes $exp(\beta(1)(1)) = \exp(\beta)$.

d) A parameter of β_k would imply a triple interaction between XYZ. In other words, the linear association will vary across the levels of Z.

Problem 26

See notes

Problem 27

See notes or the book

References

Agresti, Alan (2007). *An Introduction to Categorical Data Analysis*. 2nd ed. Hoboken NJ: Wiley-Interscience. ISBN: 978-0-471-22618-5.

Dalgaard, Peter (2008). *Introductory Statistics with R*. 2nd ed. Springer, p. 380.

Thompson, Laura A. (2007). *S-plus (and R) Manual to Accompany Agresti's "Categorical Data Analysis" (2002)*.