Department of Statistics
COURSE STATS 330
Model answer for Assignment 5, 2004

In a survey of American college students, the following variables were measured:

alcohol: Has the student tried alcoholic drinks? (Yes/No)
cigarettes: Has the student tried smoking (tobacco) cigarettes? (Yes/No)
marijuana: Has the student tried marijuana (cannabis)? (Yes/No)
race: Student's race (White/Other)
sex: Student's gender (F/M)

The data are shown overleaf. Note that unlike previous assignments, you must type these data in yourself, it is not available on the Web page. See Tutorial 9 for ways of doing this.

1. *Create a suitable data frame for the data. Check for errors. Note that the data overleaf are correct, so any errors will be due to your typing or programming.*

The following creates a data frame containing all the variables. Note the use of `expand.grid` to automatically create the factors. The baselines are fixed by the order in which the factor levels are given, so that "Yes" is the baseline for the factors alcohol, cigarettes and marijuana, "Male" is the baseline for sex, and "White" is the baseline for race.

```
ass5.df<-data.frame(expand.grid(cigarettes=c("Yes","No"),
alcohol=c("Yes","No"),marijuana=c("Yes","No"), sex=c("Male","Female"),
race=c("White","Other")),
count=c(405,13,1,1,268,218,17,117,453,28,1,1,228,201,17,133,23,,2,0,0,2
3,19,1,12,30,1,1,0,19,12,8,17))
```

2. *Collapse the table over the factors Race and Sex. Using the variables alcohol, cigarettes, and marijuana only, find a suitable conditional independence model for these data. Use your fitted model to decide if trying alcohol and trying cigarettes are related. If they are, do you think the degree of the relationship depends on trying marijuana?*

To collapse the table, we have to (i) extract the counts from the data frame, (ii) add them over the correct margins, (ii) make a new data frame containg the collapsed counts and the retained factors. This is accomplished by the code overleaf. Note the use of **`as.vector`** to coerce the array produced by **`tapply`** into a vector (as demanded by the function **`data.frame`**)

```
count1<-as.vector(tapply(ass5.df$count,
list(ass5.df$alcohol, ass5.df$cigarettes, ass5.df$marijuana),sum))

q2.df<-data.frame(expand.grid(cigarettes=c("Yes","No"),
alcohol=c("Yes","No"), marijuana=c("Yes","No")), count=count1)

> q2.df
  cigarettes alcohol marijuana count
1        Yes     Yes       Yes   911
2         No     Yes       Yes     3
3        Yes      No       Yes    44
4         No      No       Yes     2
5        Yes     Yes        No   538
6         No     Yes        No    43
7        Yes      No        No   450
8         No      No        No   279
```

Now we fit a saturated model and look for a suitable submodel:

```
> q2.glm<-glm(count~ cigarettes*alcohol*marijuana, data=q2.df,
family=poisson)
> anova(q2.glm, test="Chisq")

Analysis of Deviance Table
Model: poisson, link: log
Response: count
Terms added sequentially (first to last)


                             Df Deviance Resid. Df  Resid. Dev   P(>|Chi|)
NULL                                             7     2845.86
cigarettes                    1  1275.26         6     1570.60  2.689e-279
alcohol                       1   232.36         5     1338.24   1.821e-52
marijuana                     1    54.18         4     1284.06   1.829e-13
cigarettes:alcohol            1   445.71         3      838.35   6.200e-99
cigarettes:marijuana          1   347.97         2      490.38   1.173e-77
alcohol:marijuana             1   490.03         1        0.36  1.407e-108
cigarettes:alcohol:marijuana  1     0.36         0  -1.961e-13        0.55
>
```

It looks like the homogeneous association model A*C + A*M + C*M fits well. Its deviance is 0.36 on 1 df with a p-value of 0.55.

We interpret this that each of pair of the factors alcohol, cigarettes and marijuana are associated, but the degree of association does not depend on the level of the third factor. Thus, for example, alcohol and cigarettes as associated , but the degree to which they are associated is the same for the marijuana group and the non-marijuana group.

3. *Now fit a saturated model using all the variables. Are there any relationships between trying alcohol, cigarettes, and marijuana on the one hand, and gender and race on the other? Fit suitable additional models to explore these issues.*

We first see if the model in which alcohol, cigarettes and marijuana are independent of race and gender fits the data well, using the uncollapsed data:

```
model1<-glm(count~ cigarettes*alcohol*marijuana + sex*race, data=ass5.df,
family=poisson)
summary(model1)
```

```
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        6.03398    0.04023 149.976  < 2e-16 ***
cigarettesNo                      -3.03035    0.15435 -19.633  < 2e-16 ***
alcoholNo                         -5.71593    0.57830  -9.884  < 2e-16 ***
marijuanaNo                       -0.52668    0.05437  -9.686  < 2e-16 ***
sexFemale                          0.02093    0.04363   0.480  0.63134
raceother                         -2.56495    0.11602 -22.107  < 2e-16 ***
cigarettesNo:alcoholNo             2.62489    0.92583   2.835  0.00458 **
cigarettesNo:marijuanaNo           2.85174    0.16705  17.071  < 2e-16 ***
alcoholNo:marijuanaNo              3.18927    0.59962   5.319 1.04e-07 ***
sexFemale:raceother                0.07438    0.16052   0.463  0.64311
cigarettesNo:alcoholNo:marijuanaNo -0.57627    0.94238  -0.612  0.54087
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4836.793  on 31  degrees of freedom
Residual deviance:   41.478  on 21  degrees of freedom
AIC: 199.32

Number of Fisher Scoring iterations: 5

> 1-pchisq(41.478, 21)
[1] 0.004890871
```

It looks like this model doesn't fit well, as the p-value is too small. Fitting a saturated model yields

```
model1<-glm(count~ cigarettes*alcohol*marijuana*sex*race, data=ass5.df, family=poisson)
anova(model1, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: count
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			31	4836.8	
cigarettes	**1**	**232.4**	**30**	**4604.4**	**1.821e-52**
alcohol	**1**	**1275.3**	**29**	**3329.2**	**2.689e-279**
marijuana	**1**	**54.2**	**28**	**3275.0**	**1.829e-13**
sex	1	0.4	27	3274.6	0.5
race	**1**	**1948.8**	**26**	**1325.8**	**0.0**
cigarettes:alcohol	**1**	**445.7**	**25**	**880.0**	**6.200e-99**
cigarettes:marijuana	**1**	**746.0**	**24**	**134.1**	**3.042e-164**
alcohol:marijuana	**1**	**92.0**	**23**	**42.1**	**8.487e-22**
cigarettes:sex	1	1.125e-03	22	42.0	1.0
alcohol:sex	1	2.7	21	39.4	0.1
marijuana:sex	**1**	**9.9**	**20**	**29.5**	**1.691e-03**
cigarettes:race	1	0.9	19	28.6	0.3

```
alcohol:race                          1        9.6        18        19.1  1.979e-03
marijuana:race                        1        2.9        17        16.2        0.1
sex:race                              1        0.2        16        16.0        0.6
cigarettes:alcohol:marijuana          1        0.4        15        15.6        0.6
cigarettes:alcohol:sex                1        0.8        14        14.8        0.4
cigarettes:marijuana:sex              1        1.7        13        13.1        0.2
alcohol:marijuana:sex                 1        0.1        12        13.0        0.7
cigarettes:alcohol:race               1        2.8        11        10.2        0.1
cigarettes:marijuana:race             1        0.1        10        10.1        0.8
alcohol:marijuana:race                1        0.1         9        10.0        0.7
cigarettes:sex:race                   1        1.1         8         8.9        0.3
alcohol:sex:race                      1        5.2         7         3.7  2.224e-02
marijuana:sex:race                    1  1.287e-04         6         3.7        1.0
cigarettes:alcohol:marijuana:sex      1        0.3         5         3.4        0.6
cigarettes:alcohol:marijuana:race     1        0.8         4         2.6        0.4
cigarettes:alcohol:sex:race           1        1.5         3         1.1        0.2
cigarettes:marijuana:sex:race         1        0.9         2         0.2        0.3
alcohol:marijuana:sex:race            1        0.2         1   2.736e-10        0.7
cigarettes:alcohol:marijuana:sex:race 1        0.0         0   4.551e-10        1.0
Warning message:
fitted rates numerically 0 occurred in: method(x = x[, varseq <= i, drop = FALSE], y =
object$y, weights = object$prior.weights,
```

This suggests the model A*C + A*M + C*M + A*S*R + S*M. However, the fit is not very reliable, due to fitting difficulties caused by the zero counts. We refit the indicated model:

```
model2<-glm(count~ cigarettes*alcohol*marijuana -
cigarettes:alcohol:marijuana +sex*marijuana + alcohol*race*sex,
data=ass5.df, family=poisson)

anova(model2, test="Chisq")

Analysis of Deviance Table
Model: poisson, link: log
Response: count
Terms added sequentially (first to last)


                     Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                    31     4836.8
cigarettes            1    232.4        30     4604.4  1.821e-52
alcohol               1   1275.3        29     3329.2 2.689e-279
marijuana             1     54.2        28     3275.0  1.829e-13
sex                   1      0.4        27     3274.6        0.5
race                  1   1948.8        26     1325.8        0.0
cigarettes:alcohol    1    445.7        25      880.0  6.200e-99
cigarettes:marijuana  1    746.0        24      134.1 3.042e-164
alcohol:marijuana     1     92.0        23       42.1  8.487e-22
marijuana:sex         1      5.9        22       36.1  1.485e-02
alcohol:race          1     10.1        21       26.1  1.518e-03
alcohol:sex           1      6.0        20       20.1  1.447e-02
sex:race              1      0.1        19       20.0        0.7
alcohol:sex:race      1      2.8        18       17.1        0.1
```

This suggests the model A*C + A*M + C*M + A*R + S*M. We fit it and find that the residual deviance 26.060 on 21 degrees of freedom, corresponding to a p-value of 0.204.

This seems a reasonable model. Because of the interactions between sex and marijuana, and race and alcohol, there seem to be associations between these variables. However, there are no associations between cigarettes and race and sex (no interactions between these variables).

Data for Assignment 5.

						Race			
						White		Other	
						Alcohol		Alcohol	
						Yes	No	Yes	No
Sex	Male	Marijuana	Yes	Cigarettes	Yes	405	1	23	0
					No	13	1	2	0
			No	Cigarettes	Yes	268	17	23	1
					No	218	117	19	12
	Female	Marijuana	Yes	Cigarettes	Yes	453	1	30	1
					No	28	1	1	0
			No	Cigarettes	Yes	228	17	19	8
					No	201	133	12	17

Data for Assignment 4

crab	colour	spine	width	weight	satelite	crab	colour	spine	width	weight	satelite
1	2	3	28.3	3050	1	88	4	1	25.5	2750	0
2	3	3	22.5	1550	0	89	4	3	23.5	1900	0
3	1	1	26.0	2300	1	90	2	2	24.0	1700	0
4	3	3	24.8	2100	0	91	2	1	29.7	3850	1
5	3	3	26.0	2600	1	92	2	1	26.8	2550	0
6	2	3	23.8	2100	0	93	4	3	26.7	2450	0
7	1	1	26.5	2350	0	94	2	1	28.7	3200	0
8	3	2	24.7	1900	0	95	3	3	23.1	1550	0
9	2	1	23.7	1950	0	96	2	1	29.0	2800	1
10	3	3	25.6	2150	0	97	3	3	25.5	2250	0
11	3	3	24.3	2150	0	98	3	3	26.5	1967	1
12	2	3	25.8	2650	0	99	3	3	24.5	2200	1
13	2	3	28.2	3050	1	100	3	3	28.5	3000	1
14	4	2	21.0	1850	0	101	2	3	28.2	2867	1
15	2	1	26.0	2300	1	102	2	3	24.5	1600	1
16	1	1	27.1	2950	1	103	2	3	27.5	2550	1
17	2	3	25.2	2000	1	104	2	2	24.7	2550	1
18	2	3	29.0	3000	1	105	2	1	25.2	2000	1
19	4	3	24.7	2200	0	106	3	3	27.3	2900	1
20	2	3	27.4	2700	1	107	2	3	26.3	2400	1
21	2	2	23.2	1950	1	108	2	3	29.0	3100	1
22	1	2	25.0	2300	1	109	2	3	25.3	1900	1
23	2	1	22.5	1600	1	110	2	3	26.5	2300	1
24	3	3	26.7	2600	1	111	2	3	27.8	3250	1
25	4	3	25.8	2000	1	112	2	3	27.0	2500	1
26	4	3	26.2	1300	0	113	3	3	25.7	2100	0
27	2	3	28.7	3150	1	114	2	3	25.0	2100	1
28	2	1	26.8	2700	1	115	2	3	31.9	3325	1
29	4	3	27.5	2600	0	116	4	3	23.7	1800	0
30	2	3	24.9	2100	0	117	4	3	29.3	3225	1
31	1	1	29.3	3200	1	118	3	3	22.0	1400	0
32	1	3	25.8	2600	0	119	2	3	25.0	2400	1
33	2	2	25.7	2000	0	120	3	3	27.0	2500	1
34	2	1	25.7	2000	1	121	3	3	23.8	1800	1
35	2	1	26.7	2700	1	122	1	1	30.2	3275	1
36	4	3	23.7	1850	0	123	3	3	26.2	2225	0
37	2	3	26.8	2650	0	124	2	3	24.2	1650	1
38	2	3	27.5	3150	1	125	2	3	27.4	2900	1
39	4	3	23.4	1900	0	126	2	2	25.4	2300	0
40	2	3	27.9	2800	1	127	3	3	28.4	3200	1
41	3	3	27.5	3100	1	128	4	3	22.5	1475	1
42	1	1	26.1	2800	1	129	2	3	26.2	2025	1
43	1	1	27.7	2500	1	130	2	1	24.9	2300	1
44	2	1	30.0	3300	1	131	1	2	24.5	1950	1
45	3	1	28.5	3250	1	132	2	3	25.1	1800	0
46	3	3	28.9	2800	1	133	2	1	28.0	2900	1
47	2	3	28.2	2600	1	134	4	3	25.8	2250	1

48	2	3	25.0	2100	1	135	2	3	27.9	3050	1
49	2	3	28.5	3000	1	136	2	3	24.9	2200	0
50	2	1	30.3	3600	1	137	2	1	28.4	3100	1
51	4	3	24.7	2100	1	138	3	3	27.2	2400	1
52	2	3	27.7	2900	1	139	2	2	25.0	2250	1
53	1	1	27.4	2700	1	140	2	3	27.5	2625	1
54	2	3	22.9	1600	1	141	2	1	33.5	5200	1
55	2	1	25.7	2000	1	142	2	3	30.5	3325	1
56	2	3	28.3	3000	1	143	3	3	29.0	2925	1
57	2	3	27.2	2700	1	144	2	1	24.3	2000	0
58	3	3	26.2	2300	1	145	2	3	25.8	2400	0
59	2	1	27.8	2750	0	146	4	3	25.0	2100	1
60	4	3	25.5	2250	0	147	2	1	31.7	3725	1
61	3	3	27.1	2550	0	148	2	3	29.5	3025	1
62	3	3	24.5	2050	1	149	3	3	24.0	1900	1
63	3	1	27.0	2450	1	150	2	3	30.0	3000	1
64	2	3	26.0	2150	1	151	2	3	27.6	2850	1
65	2	3	28.0	2800	1	152	2	3	26.2	2300	0
66	2	3	30.0	3050	1	153	2	1	23.1	2000	0
67	2	3	29.0	3200	1	154	2	1	22.9	1600	0
68	2	3	26.2	2400	0	155	4	3	24.5	1900	0
69	2	1	26.5	1300	0	156	2	3	24.7	1950	1
70	2	3	26.2	2400	1	157	2	3	28.3	3200	0
71	3	3	25.6	2800	1	158	2	3	23.9	1850	1
72	3	3	23.0	1650	1	159	3	3	23.8	1800	0
73	3	3	23.0	1800	0	160	3	2	29.8	3500	1
74	2	3	25.4	2250	1	161	2	3	26.5	2350	1
75	3	3	24.2	1900	0	162	2	3	26.0	2275	1
76	2	2	22.9	1600	0	163	2	3	28.2	3050	1
77	3	2	26.0	2200	1	164	4	3	25.7	2150	0
78	2	3	25.4	2250	1	165	2	3	26.5	2750	1
79	3	3	25.7	1200	0	166	2	3	25.8	2200	0
80	2	3	25.1	2100	1	167	3	3	24.1	1800	0
81	3	2	24.5	2250	0	168	3	3	26.2	2175	1
82	4	3	27.5	2900	0	169	3	3	26.1	2750	1
83	3	3	23.1	1650	0	170	3	3	29.0	3275	1
84	3	1	25.9	2550	1	171	1	1	28.0	2625	0
85	2	3	25.8	2300	0	172	4	3	27.0	2625	0
86	4	3	27.0	2250	1	173	2	2	24.5	2000	0
87	2	3	28.5	3050	0						