Exercise 4.7 [13 points]

```
. infile race view reagan others using vote.raw
(14 observations read)
. gen total=reagan+others
. list race view reagan total

         race       view     reagan      total
  1.        1          1          1         13
  2.        1          2         13         70
  3.        1          3         44        115
  4.        1          4        155        301
  5.        1          5         92        153
  6.        1          6        100        141
  7.        1          7         18         26
  8.        2          1          0          6
  9.        2          2          0         16
 10.        2          3          2         25
 11.        2          4          1         32
 12.        2          5          0          8
 13.        2          6          2          9
 14.        2          7          0          4

. xi: glm reagan race i.view, f(b total)
i.view                Iview_1-7    (naturally coded; Iview_1 omitted)

Residual df  =         6                              No. of obs =        14
Pearson X2   =  4.182984                             Deviance   =  4.960863
Dispersion   =   .697164                             Dispersion =  .8268105

Binomial (N=total) distribution, logit link
------------------------------------------------------------------------------
  reagan |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    race |  -2.886714   .4706828     -6.133   0.000      -3.809235   -1.964192
 Iview_2 |   1.017621    1.08285      0.940   0.347      -1.104725    3.139967
 Iview_3 |   2.077408   1.055501      1.968   0.049       .008665    4.146152
 Iview_4 |   2.564028   1.044851      2.454   0.014       .5161579   4.611898
 Iview_5 |   2.908747   1.051427      2.766   0.006       .8479873   4.969506
 Iview_6 |   3.436971   1.054705      3.259   0.001       1.369788   5.504154
 Iview_7 |   3.251093    1.11525      2.915   0.004       1.065242   5.436944
   _cons |   .3722484   1.145968      0.325   0.745      -1.873808   2.618305
------------------------------------------------------------------------------

. glmpred est, mu
. glmpred pres, pearson
. format est pres %5.3f
. list race view reagan total mu pres

         race       view     reagan      total     est     pres
  1.        2          1          0          6   0.027  -0.165
  2.        1          1          1         13   0.973   0.028
  3.        2          2          0         16   0.197  -0.447
  4.        1          2         13         70   12.803   0.061
  5.        1          3         44        115   45.131  -0.216
  6.        2          3          2         25   0.869   1.235
  7.        1          4        155        301  154.229   0.089
  8.        2          4          1         32   1.771  -0.596
  9.        2          5          0          8   0.611  -0.813
 10.        1          5         92        153   91.389   0.101
 11.        1          6        100        141  100.893  -0.167
 12.        2          6          2          9   1.107   0.906
 13.        2          7          0          4   0.417  -0.683
 14.        1          7         18         26   17.583   0.175


. glm reagan race view, f(b total)

Residual df  =        11                              No. of obs =        14
Pearson X2   =  11.51111                             Deviance   =  12.4703
Dispersion   =  1.046464                             Dispersion =  1.133663

Binomial (N=total) distribution, logit link
------------------------------------------------------------------------------
  reagan |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    race |  -2.936908   .4722074     -6.220   0.000      -3.862418   -2.011399
    view |   .4908373   .0592682      8.282   0.000       .3746739    .6070008
   _cons |   .9129588   .5409302      1.688   0.091       -.147245    1.973163
------------------------------------------------------------------------------

. glmpred est2, mu
. glmpred pres2, pearson
. format est2 pres2 %5.3f
. list race view reagan total est2 pres2

         race       view     reagan      total    est2    pres2
  1.        2          1          0          6   0.068  -0.262
  2.        1          1          1         13   2.308  -0.949
  3.        2          2          0         16   0.294  -0.547
  4.        1          2         13         70  18.250  -1.429
  5.        1          3         44        115  42.036   0.380
  6.        2          3          2         25   0.741   1.484
  7.        1          4        155        301 145.941   1.045
  8.        2          4          1         32   1.521  -0.433
  9.        2          5          0          8   0.603  -0.808
 10.        1          5         92        153  92.707  -0.117
 11.        1          6        100        141 100.852  -0.159
 12.        2          6          2          9   1.058   0.975
 13.        2          7          0          4   0.715  -0.933
 14.        1          7         18         26  20.906  -1.436
```

a. [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), s/he is more likely to vote 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 nominal 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. $\hat{\beta}_{view} = .491$ 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]

X is independent of Y	A father's gender is independent of his unborn 1st child's gender.
Y is independent of Z	the unborn 1st child's gender is independent of the mother's gender.
X is independent of Z?	the father's gender is independent of the mother's??? No! One has to be male and one has to be female!
$\Rightarrow$ Y is jointly independent of X and Z	the unborn 1st child's gender is jointly independent of the father's and the mother's.

Exercise 5.13 [4 points]

Religion (X)	Sexual Attitude (Y)	% Opposing abortion (Z)
Religious	Conservative	25%
	Permissive	31%
Non-religious	Conservative	29%
	Permissive	15%

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=54%); 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]

Model	ΔG^2	Best model
fit(age smoking test)	Goodness-of-fit chi2(0) = 0.000	
fit(age smoking, smoking test, test age)	Goodness-of-fit chi2(1) =20.656	*
fit(age smoking, smoking test)	Goodness-of-fit chi2(2) = 48.568	
fit(age smoking, age test)	Goodness-of-fit chi2(2) = 47.613	
fit(age test, smoking test)	Goodness-of-fit chi2(2) = 32.449	*
fit(age, smoking test)	Goodness-of-fit chi2(3) = 65.785	
fit(smoking, age test)	Goodness-of-fit chi2(3) = 64.830	*
fit(test, age smoking)	Goodness-of-fit chi2(3) = 80.951	
fit(age, smoking, test)	Goodness-of-fit chi2(4) = 98.166	

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 ΔG^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.)

```
. loglin count age smoking test, fit(age smoking, age test, smoking test) resid
Variable age = A
Variable smoking = B
Variable test = C
Margins 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.

Iteration 0: Log Likelihood = -37.97168
Iteration 1: Log Likelihood = -36.104492
Iteration 2: Log Likelihood = -36.078125

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

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
          A2 |  -1.344605   .0882685    -15.233   0.000      -1.517608   -1.171602
        AB22 |   .3799098   .1114839      3.408   0.001       .1614054    .5984143
        AC22 |   .8945767   .1661834      5.383   0.000       .5688631     1.22029
          B2 |   .1326946   .0559051      2.374   0.018       .0231226    .2422666
        BC22 |   .9394522     .19231      4.885   0.000       .5625316    1.316373
          C2 |  -3.232031   .1784376    -18.113   0.000      -3.581762   -2.882299
       _cons |   6.376379   .0410193    155.448   0.000       6.295983    6.456776
------------------------------------------------------------------------------

count age smoking  test   cellhat    resid   stdres
577    1    1       1    587.795  -10.795   -0.445
 34    1    1       2     23.205   10.795    2.241
682    1    2       1    671.204   10.796    0.417
 57    1    2       2     67.795  -10.795   -1.311
164    2    1       1    153.205   10.795    0.872
  4    2    1       2     14.795  -10.795   -2.807
245    2    2       1    255.796  -10.796   -0.675
 74    2    2       2     63.205   10.795    1.358
```

```
. loglin count age smoking test, fit (age smoking test) resid
Variable age = A
Variable smoking = B
Variable test = C
Margins fit: age smoking test
Note: Regression-like constraints are assumed.  The first level of each
variable (and all interactions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -25.788086
Iteration 1: Log Likelihood = -25.75

Poisson regression                              Number of obs    =       8
Goodness-of-fit chi2(0)     =     0.000         Model chi2(7)    =1980.747
Prob > chi2                 =         .         Prob > chi2      =  0.0000
Log Likelihood              =   -25.750         Pseudo R2        =  0.9747

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
          A2 |  -1.257976    .088491    -14.216   0.000      -1.431415   -1.084537
        AB22 |   .2342048   .1156658      2.025   0.043       .007504    .4609055
        AC22 |  -.8820901   .5359449     -1.646   0.100      -1.932523    .1683425
      ABC222 |   2.166875   .5690713      3.808   0.000       1.051516    3.282235
          B2 |   .1671872    .056563      2.956   0.003       .0563257    .2780487
        BC22 |   .3495035   .2239531      1.561   0.119      -.0894365    .7884436
          C2 |  -2.831482   .1764791    -16.044   0.000      -3.177375   -2.485589
       _cons |   6.357842   .0416305    152.721   0.000       6.276248    6.439437
------------------------------------------------------------------------------

count  age  smoking  test   cellhat    resid   stdres
577    1    1    1   577.000   -0.000   -0.000
 34    1    1    2    34.000    0.000    0.000
682    1    2    1   682.000    0.000    0.000
 57    1    2    2    57.000    0.000    0.000
164    2    1    1   164.000    0.000    0.000
  4    2    1    2     4.000    0.000    0.000
245    2    2    1   245.000   -0.000   -0.000
 74    2    2    2    74.000   -0.000   -0.000
```

Exercise 6.3 [7 points]

Model	ΔG^2	Best model
fit(use eject injury)	Goodness-of-fit chi2(0) = 0.00	
fit(use eject, eject injury, injury use)	Goodness-of-fit chi2(1) =3.00	*
fit(use eject, use injury)	Goodness-of-fit chi2(2) = 1681.00	
fit(use eject, eject injury)	Goodness-of-fit chi2(2) = 1145.00	*
fit(use injury, eject injury)	Goodness-of-fit chi2(2) = 7134.00	
fit(use, eject injury)	Goodness-of-fit chi2(3) = 9022.00	
fit(eject, use injury)	Goodness-of-fit chi2(3) = 9557.00	
fit(injury, use eject)	Goodness-of-fit chi2(3) = 3568.00	*
fit(use, eject, injury)	Goodness-of-fit chi2(4) = 11445.00	

```
. loglin count use eject injury, fit(use eject, use injury, eject injury) resid

Variable use = A
Variable eject = B
Variable injury = C
Margins fit: use eject, use injury, eject injury
Note: Regression-like constraints are assumed.  The first level of each
variable (and all interactions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -40

Poisson regression                              Number of obs    =       8
Goodness-of-fit chi2(1)     =     3.000         Model chi2(6)    =1624863.0
Prob > chi2                 =    0.0833         Prob > chi2      =  0.0000
Log Likelihood              =   -40.000         Pseudo R2        =  1.0000

------------------------------------------------------------------------------
       count |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
          A2 |   1.439133   .0332137     43.329   0.000       1.374036    1.504231
        AB22 |  -2.399635   .0333403    -71.974   0.000      -2.464981    -2.33429
        AC22 |   1.717321   .0540152     31.793   0.000       1.611453    1.823189
          B2 |   5.925269   .0299613    197.764   0.000       5.866546    5.983992
        BC22 |  -2.797794   .0552557    -50.634   0.000      -2.906094   -2.689495
          C2 |  -3.963146   .0694418    -57.071   0.000      -4.099249   -3.827042
       _cons |   7.001366   .0299221    233.987   0.000        6.94272    7.060012
------------------------------------------------------------------------------

 count  use  eject injury   cellhat  resid   stdres
  1105    1    1    1    1098.132    6.868    0.207
    14    1    1    2      20.868   -6.868   -1.503
411111    1    2    1  411118.062   -7.062   -0.011
   483    1    2    2     476.132    6.868    0.315
  4624    2    1    1    4630.867   -6.867   -0.101
   497    2    1    2     490.132    6.868    0.310
157342    2    2    1  157335.094    6.906    0.017
  1008    2    2    2    1014.868   -6.868   -0.216
```

By fitting various loglinear models, we found that the no-3-way-interaction model is the best fitting model with ΔG^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 $e^{AB_{22}} = e^{2.40} \approx 11.02$); by not being ejected, one is less likely to be killed (by a factor of $e^{BC_{22}} = e^{2.80} \approx 16.44$); by not wearing seat belts, one is more likely to be killed (by a factor of $e^{AC_{22}} = e^{1.72} \approx 5.58$). 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 correponding 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

. xi: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                 =         .            Prob > chi2      =  0.0567
Log Likelihood              =    -4.447            Pseudo R2        =  0.2899

------------------------------------------------------------------------------
    dead |      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                 =    0.0567            Prob > chi2      =       .
Log Likelihood              =    -6.263            Pseudo R2        = -0.0000

------------------------------------------------------------------------------
    dead |      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.