This assignment covers first two sub-lessons of Lesson 6. The assignment should be typed, with properly labeled computer output. You are encouraged to work together to a reasonable degree, but the write up should be your own. Please make sure you show your work! You should submit your homework on ANGEL in the HW6 Drop Box. If there is code available from the lecture notes or previous solutions, feel free to use that code and make adjustment as you see fit.

1. (50 pts)Montana Economic Outlook (same data from HW4) This 1992 Montana poll asked a random sample of Montana residents whether their personal financial status was the worse, the same, or better than a year ago, and whether they thought the state economic outlook was better over the next year. This file contains these items and accompanying demographics about the respondents for every other person included in the poll to reduce the size. Both SAS and R help files are available on ANGEL.

   Here is the coding for the data: AGE = 1 under 35, 2 35-54, 3 55 and over
   SEX = 0 male, 1 female
   INC = yearly income: 1 under \$20K, 2 20-35\$K, 3 over \$35K
   POL = 1 Democrat, 2 Independent, 3 Republican
   AREA = 1 Western, 2 Northeastern, 3 Southeastern Montana
   FIN = Financial status 1 worse, 2 same, 3 better than a year ago
   STAT = State economic outlook 0 better, 1 not better than a year ago

   (a) In Homework 5 you answered questions regarding the association between income (INC) and party affiliation (POL) through the chi-square test of independence. State the independence log-linear model in a log-linear notation and with these two variables, e.g., λ_i^{INC}. Fit this model for income and party affiliation? Comment on the fit of this model. What is the estimated count in the (1,1) cell based on this model? Interpret the parameter estimates of this model. What do you learn from this model about associations between income and party affiliation.

   Solution: The independence model is:

   $$log(\mu_{ij}) = \lambda + \lambda_i^{INC} + \lambda_j^{POL}. \quad (1)$$

   From output $G^2 = 6.704$ with df=4. The p-value is 0.15, indicating that the independent model fits the data well.

   The estimated counts in cell (1,1) is:

   $$\mu_{11} = exp(\lambda + \lambda_1^{INC} + \lambda_1^{POL}) = exp(3.145 - 0.2709 + 0.0403) = 18.49.$$

   From the output only POL2 is significant. The odds are exp(-0.707)=0.49 for Independent vs Republicans. Note that the parameters in log-linear

model can be interpret as odds, rather than odds ratio. We get the estimated odds:

$$Low - Med : exp(\lambda_1^{INC} - \lambda_2^{INC}) = exp(-0.0279 - 0.3169) = 0.708$$
$$Low - High : exp(\lambda_1^{INC}) = exp(-0.0279) = 0.972$$
$$Med - High : exp(\lambda_2^{INC}) = exp(0.317) = 1.373$$
$$Dem - Inde : exp(\lambda_1^{POL} - \lambda_2^{POL}) = exp(0.0403 + 0.7069) = 2.11$$
$$Dem - Rep : exp(\lambda_1^{POL}) = exp(0.0403) = 1.041$$
$$Ind - Rep : exp(\lambda_2^{POL}) = exp(-0.7069) = 0.493$$

As the independent model holds there is no significant association between income and party affiliation.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4	6.7040	1.6760
Scaled Deviance	4	6.7040	1.6760
Pearson Chi-Square	4	6.6995	1.6749
Scaled Pearson X2	4	6.6995	1.6749
Log Likelihood		388.4050	
Full Log Likelihood		-24.7359	
AIC (smaller is better)		59.4719	
AICC (smaller is better)		79.4719	
BIC (smaller is better)		60.4580	

Algorithm converged

The SAS System

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	3.1476	0.1589	2.8246	3.4480	392.50	<.0001
Inc	<$20k	1	-0.2709	0.1979	-0.6634	0.1149	1.87	0.1711
Inc	$20k-$30k	1	0.3169	0.1712	-0.0164	0.6561	3.43	0.0641
Inc	>$30k	0	0.0000	0.0000	0.0000	0.0000	.	.
Pol	Dem	1	0.0403	0.1639	-0.2813	0.3625	0.06	0.8059
Pol	Ind	1	-0.7069	0.2037	-1.1165	-0.3155	12.05	0.0005
Pol	Repub	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Inc	2	10.60	0.0050
Pol	2	17.65	0.0001

(b) When you fitted the above model, did you ignore or control for all the other variables present in this dataset?

Solution: We ignored all the other variables present in the dataset.

(c) State the saturated log-linear model in a log-linear notation and with the same two variables as above. Fit the saturated log-linear model for income and party affiliation? Comment on the fit of this model. Does this model fit better or worse than the independence model from part (a). Interpret the association/interaction parameter estimates of this model. How do these estimates compare to the odds ratio(s) you calculated directly from the observed table. What do you learn from this model about associations between income and party affiliation.

Solution: The saturated model is:

$$log(\mu_{ij}) = \lambda + \lambda_i^{INC} + \lambda_j^{POL} + \lambda_{ij}^{INC*POL}.$$

Since this is the saturated model, it fits the data perfectly and thus it is clear that is better than the model of independence. From the LR statistics we can see that the interaction term INC*POL is not significant, confirming conclusion (a), that is the independent model fits. So, there is no statistically significant association between INC and POL. The estimated odds ratios would be exactly the same as the observed odds ratios from the data because the model fits the data perfectly.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	3.2958	0.1925	2.8934	3.6507	293.29	<.0001
Inc	<$20k	1	-0.5232	0.3155	-1.1631	0.0837	2.75	0.0972
Inc	$20k-$30k	1	0.1054	0.2653	-0.4151	0.6306	0.16	0.6912
Inc	>$30k	0	0.0000	0.0000	0.0000	0.0000	.	.
Pol	Dem	1	-0.0770	0.2776	-0.6262	0.4685	0.08	0.7816
Pol	Ind	1	-1.3499	0.4241	-2.2635	-0.5741	10.13	0.0015
Pol	Repub	0	0.0000	0.0000	0.0000	0.0000	.	.

The SAS System

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Inc*Pol	<$20k	Dem	1	0.3954	0.4301	-0.4438	1.2491	0.85	0.3579
Inc*Pol	<$20k	Ind	1	0.5232	0.6207	-0.7057	1.7601	0.71	0.3992
Inc*Pol	<$20k	Repub	0	0.0000	0.0000	0.0000	0.0000	.	.
Inc*Pol	$20k-$30k	Dem	1	0.0431	0.3806	-0.7044	0.7919	0.01	0.9099
Inc*Pol	$20k-$30k	Ind	1	1.0398	0.5086	0.0780	2.0936	4.18	0.0409
Inc*Pol	$20k-$30k	Repub	0	0.0000	0.0000	0.0000	0.0000	.	.
Inc*Pol	>$30k	Dem	0	0.0000	0.0000	0.0000	0.0000	.	.
Inc*Pol	>$30k	Ind	0	0.0000	0.0000	0.0000	0.0000	.	.
Inc*Pol	>$30k	Repub	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Inc	2	13.76	0.0010
Pol	2	21.22	<.0001
Inc*Pol	4	6.70	0.1524

(d) How do these analysis and interpretations differ, if at all, from what you did in Homework 5?

Solution: The results are exactly the same as those in homework 5.

(e) **Note: maybe you want to come back to this part after you do the next problem**. What can you say about associations between gender (SEX), income (INC) and party affiliation(POL)? Does the model of complete (mutual) independence fit? State three possible conditional independence models for these three variables. Do any of these conditional independence model fit? Do your conclusions from (a) change in any way?

Solution: The following Table gives the three possible conditional independence models for the three variables sex (S), income (I) and party affiliation(P). We can see that only (PI, IS) and (IS, PS) fit the data reasonably well. These results are consistent with part (a) which shows that the independent model of INC and POL fits the data.

Conditional independence models					
Model	DF	G^2	p-value	X^2	p-value
(IP, PS)	6	16.5444	0.0111	15.8254	0.0147
(IP, IS)	6	8.6209	0.1960	8.5961	0.1976
(IS, PS)	8	10.8018	0.2132	10.5136	0.2308

SAS code/Output

```
/*Here is yet another way of reading the datafile and then
 creating tables; /
please note you need to
 put your own correct path to import that data. */

proc import datafile = 'montana.xls' out=montana replace;
run;

proc format;
value IncFmt 1='Under $20K'
       2='$20-35K'
3='Over $35K';
value PolFmt 1='Democrat'
       2='Independent'
       3='Republican';
run;

/* Count the frequencies of every combination of INC and POL
 and  save them in data set montana1 */
proc freq data=montana;
table INC*POL /out=montana1 chisq;
format INC IncFmt. POL PolFmt.;
run;


data montana1;
set montana1;
if INC = . or POL=. then delete;
drop percent;
run;
```

```
proc print data=montana1;
run;



/*Then you can call GENMOD or CATMOD; e.g. */
/* Independence model using PROC GENMOD */
proc genmod data=montana1 order=data;
class INC POL;
model count = INC POL /link=log dist=poisson;
run;

/* Sat model using PROC GENMOD */
proc genmod data=montana1 order=data;
class INC POL;
model count = INC POL INC*POL /link=log dist=poisson;
run;
```

R code/Output

```
> ###Read data in
> montana=read.table("montana.csv", header=TRUE, sep=",")
> ###Attach the file to have direct access to the variables
> attach(montana)
>
> ### Create two-way table Age x Pol, excluding missing data
> age_pol=table(Age,Pol, exclude=".", dnn=list("Age","Pol"))
> ### Create two-way table Inc x Pol, excluding missing data
> inc_pol=table(Inc,Pol, exclude=".", dnn=list("Inc","Pol"))
>
> ### create a data frame
> age_pol_df=as.data.frame(age_pol)
> inc_pol_df=as.data.frame(inc_pol)
>
> ### fit model of independence
> ip_ind=glm(Freq~Inc+Pol, family=poisson(), data=inc_pol_df)
> summary(ip_ind)

Call:
glm(formula = Freq ~ Inc + Pol, family = poisson(),
 data = inc_pol_df)
```

```
Deviance Residuals:
      1        2        3        4        5        6        7
 0.7932  -0.7580   0.1540  -0.6154   1.4816  -1.4266  -0.4241
 -0.3507   0.7515

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.91704    0.17313  16.849  < 2e-16 ***
Inc2         0.58779    0.18592   3.161 0.001570 **
Inc3         0.27087    0.19792   1.369 0.171116
Pol2        -0.74721    0.20233  -3.693 0.000222 ***
Pol3        -0.04027    0.16388  -0.246 0.805873
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 34.955  on 8  degrees of freedom
Residual deviance:  6.704  on 4  degrees of freedom
AIC: 59.472

Number of Fisher Scoring iterations: 4

>
> ### fit saturated model
> ip_sat=glm(Freq~Inc+Pol+Inc*Pol, family=poisson(),
data=inc_pol_df)
> summary(ip_sat)

Call:
glm(formula = Freq ~ Inc + Pol + Inc * Pol, family = poisson(),
    data = inc_pol_df)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.0910     0.2132  14.498  < 2e-16 ***
Inc2          0.2763     0.2827   0.977  0.32853
Inc3          0.1278     0.2923   0.437  0.66190
Pol2         -1.1451     0.4339  -2.639  0.00832 **
```

```
Pol3         -0.3185     0.3286  -0.969  0.33243
Inc2:Pol2     0.8689     0.5179   1.678  0.09342 .
Inc3:Pol2    -0.1278     0.6092  -0.210  0.83380
Inc2:Pol3     0.3524     0.4193   0.840  0.40066
Inc3:Pol3     0.3954     0.4301   0.919  0.35792
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  3.4955e+01  on 8  degrees of freedom
Residual deviance: -1.7764e-15  on 0  degrees of freedom
AIC: 60.768

Number of Fisher Scoring iterations: 3

>
> ## get the 2x2 subtables and the observed odds-ratio
> library(vcd)
> oddsratio(inc_pol[1:2,1:2], log=FALSE)
[1] 2.384236
> exp(confint(oddsratio(inc_pol[1:2,1:2])))
           lwr      upr
[1,] 0.8842402 6.428778
> oddsratio(inc_pol[2:3,3:3], log=FALSE)
> exp(confint(oddsratio(inc_pol[2:3,2:3])))
          lwr      upr
[1,] 1.068044 7.491093
>
> detach(montana)
```

2. (35 pts) The table below lists graduate admissions information for the six largest departments at U.C. Berkeley in the fall of 1973. You saw this the example in the lecture and in the last homework.

Dept.	No. of men rejected	No. of men accepted	No. of women rejected	No. of women accepted
A	313	512	19	89
B	207	353	8	17
C	205	120	391	202
D	278	139	244	131
E	138	53	299	94
F	351	22	317	24

Let D = department, S =sex,and A = admission status (rejected or accepted).

(a) Fit all nine hierarchical log-linear models models. Summarize in a tabular form the results of these models (e.g. goodness-of-fit statistics, etc...), that is complete the Table 1 from the lecture notes under "Summary of Log-Linear Inference for 3-way Tables". Comment on the fits of these models. Which model would you choose to describe this data (e.g., which model do you think is best) and why?

Solution:

The nine log-linear models are summarized in the following table:

Model	DF	G^2	p-value	X^2	p-value
(DSA)	0	0		0	
(DS, DA, SA)	5	20.23	<0.01	18.83	<0.01
(DA, SA)	10	1148.35	<0.001	1015.21	<0.001
(DS, DA)	6	21.66	<0.01	19.87	<0.01
(DS, SA)	10	782.65	<0.001	714.3	<0.001
(DA, S)	11	1242.28	<0.001	1078	<0.001
(SA, D)	15	2003.26	<0.001	1746.72	<0.001
(DS, A)	11	876.6	<0.001	797.1	<0.001
(D,S,A)	16	2097.2	<0.001	1999.6	<0.001

Therefore, except for the saturated model, we reject all the other models at the 0.05 significance level. We choose the saturated model to describe this data.

(b) Perform the partial association tests by completing the rest of the Table 2 from the lecture notes under "Summary Inference for Admissions data", and make appropriate inference.

Solution:

Model	DF	G^2	ΔG^2	Δdf	p-value for differene
(DS, DA, SA)	5	20.23			
(DA, SA)	10	1148.35	1128.12	5	<0.001
(DS, DA)	6	21.66	1.43	1	0.23
(DS, SA)	10	782.65	762.42	5	<0.001

Therefore, we can not reject $H_0 : \lambda^{SA} = 0$; that is there seems to be no significant association between S and A given D. However, we can reject $H_0 : \lambda^{DA} = 0$, that is there is significant association between D and A given S. Similarly, we can reject $H_0 : \lambda^{DS} = 0$, that is there is significant association between D and S given A.

(c) Based on your results from the previous parts, interpret the model you have chosen as "the best" model. Comment on the fit of this model in more detail (e.g. goodness-of-fit statistics, resuduals, etc...). What does this model tell you about the relationship between gender, admission and the department?

Solution:

Observation Statistics

Observation	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual	Likelihood Residual
1	1.1880213	0.0828112	0.0827317	0.5032159	0.5036993	0.5036862
2	-1.188028	-0.063126	-0.063162	-0.503984	-0.503702	-0.503706
3	-1.188148	-0.391973	-0.400912	-0.515241	-0.503753	-0.510739
4	1.1878394	0.2987185	0.2950909	0.4975062	0.5036222	0.501479
5	-6.002223	-0.413208	-0.415191	-0.872238	-0.868073	-0.869018
6	6.0021259	0.5621558	0.557328	0.8606035	0.8680585	0.8649397
7	6.0021906	0.3059008	0.3051111	0.8658268	0.8680678	0.8677898
8	-6.002224	-0.416177	-0.418203	-0.872299	-0.868073	-0.869046
9	3.1590906	0.1905555	0.1901921	0.4733913	0.4742956	0.4741497
10	-3.159096	-0.264957	-0.265948	-0.476069	-0.474296	-0.47485
11	-3.159092	-0.200944	-0.201374	-0.475312	-0.474296	-0.474478
12	3.1590858	0.2794001	0.278261	0.4723613	0.4742948	0.4736248
13	-4.9231	-0.411801	-0.4142	-1.006392	-1.000564	-1.001554
14	4.9227494	0.7099666	0.6983401	0.9841088	1.0004929	0.9922764
15	4.9230078	0.2870781	0.2862826	0.9977731	1.0005455	1.0003175
16	-4.92322	-0.494994	-0.499188	-1.009065	-1.000589	-1.00267
17	2.0308118	0.1087117	0.1086065	0.6191506	0.6197502	0.6197318
18	-2.030995	-0.414308	-0.420359	-0.628859	-0.619806	-0.623868
19	-2.030812	-0.113698	-0.113819	-0.62041	-0.61975	-0.619772
20	2.0306454	0.4332369	0.4268074	0.6105028	0.6196995	0.6152217

We suggest to use model (DS, DA), although strictly speaking G^2 test shows the lack of fit for the model. First, partial association from part (b) shows that (DS, DA) is not significantly different from the saturated model. Second, from the following adjusted residuals of (DS, DA), we can see that except for the values in Deparment A, the other residuals are

withing +/-1, less than the acceptable threshold 2 or 3. Therefore, it is Department A that causes the lack of fit of (DS, DA). Assuming that this model is appropriate, Admission and Gender are independent for a given Department.

(d) In the last homework, you fitted some models without the department A. Refit the loglinear models by dropping the department A from the analysis. Which model would you chose now to describe this data? Is it the same or different from the model you have chosen in part (a)? Interpret the fit and the parameters of this model. What can you say about the relationship between gender, admission and the department?

Solution:

Model	DF	G^2	p-value	X^2	p-value
(DSA)	0	0		0	
(DS, DA, SA)	4	2.45	0.65	2.45	0.65
(DA, SA)	8	717.02	<0.001	610.49	<0.001
(DS, DA)	5	2.61	0.76	2.62	0.76
(DS, SA)	8	500.43	<0.001	446.56	<0.001
(DA, S)	9	756.05	<0.001	638.84	<0.001
(SA, D)	12	1253.98	<0.001	1207.07	<0.001
(DS, A)	9	539.56	<0.001	495.97	<0.001
(D,S,A)	13	1293	<0.001	1329.29	<0.001

It is clear that (DS, DA) has the best fit other than the saturated model. It is different from the conclusion in part (a) where (DS, DA) is not significant. The model tell us that Admission and Gender are conditionally independent given Department.

SAS code

```
/* Analysis of a 3-way table Berkeley Admissions data
 using PROC GENMOD*/
/* Fitting various log-linear models*/
/* For a related analysis via PROC FREQ see berkeley.sas*/
options nocenter nodate nonumber linesize=80;
data berkeley;
   input D $ S $ A $ count;
   cards;
DeptA  Male    Reject  313
DeptA  Male    Accept  512
DeptA  Female  Reject   19
DeptA  Female  Accept   89
DeptB  Male    Reject  207
```

```
DeptB  Male    Accept  353
DeptB  Female  Reject    8
DeptB  Female  Accept   17
DeptC  Male    Reject  205
DeptC  Male    Accept  120
DeptC  Female  Reject  391
DeptC  Female  Accept  202
DeptD  Male    Reject  278
DeptD  Male    Accept  139
DeptD  Female  Reject  244
DeptD  Female  Accept  131
DeptE  Male    Reject  138
DeptE  Male    Accept   53
DeptE  Female  Reject  299
DeptE  Female  Accept   94
DeptF  Male    Reject  351
DeptF  Male    Accept   22
DeptF  Female  Reject  317
DeptF  Female  Accept   24
;

/*saturated model via PROC FREQ*/
proc freq data=berkeley order=data;
weight count;
tables A*D*S/cmh chisq relrisk expected nocol norow;
tables D*S/chisq relrisk;
run;

/*saturated model DSA, two different ways of fitting it*/
proc genmod data=berkeley order=data;
class D S A;
model count = D*S*A / dist=poisson link=log;
model count= D S A D*S D*A S*A D*S*A/dist=poisson link=log;
run;

/*model of complete independence*/
proc genmod data=berkeley order=data;
class D S A;
model count = D S A / dist=poisson link=log;
run;

/* joint independence of D and S from A*/
```

```
proc genmod data=berkeley order=data;
class D S A;
model count = D S A D*S / dist=poisson link=log;
run;

/*you can fill in the rest of the models*/
```

R code/Output:

```
> #### To test the odds-ratios in the marginal table
 ## and each of the subtables
>
> library(vcd)
Loading required package: MASS
Loading required package: grid
Loading required package: colorspace
>
> #Two ways of fitting a log-linear model of complete independence
>
> ### Via loglin() function
> berk.ind<-loglin(UCBAdmissions, list(1,2,3), fit=TRUE, param=TRUE)
2 iterations: deviation 4.547474e-13
> berk.ind
>
> #### Via glm() function
> berk.data<-as.data.frame(UCBAdmissions)
> berk.ind<-glm(berk.data$Freq~berk.data$Admit+berk.data$Gender+
berk.data$Dept, family=poisson())
> summary(berk.ind)

Call:
glm(formula = berk.data$Freq ~ berk.data$Admit + berk.data$Gender +
    berk.data$Dept, family = poisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-18.170   -7.719   -1.008    4.734   17.153

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                5.37111    0.03964 135.498  < 2e-16 ***
berk.data$AdmitRejected    0.45674    0.03051  14.972  < 2e-16 ***
berk.data$GenderFemale    -0.38287    0.03027 -12.647  < 2e-16 ***
```

```
berk.data$DeptB         -0.46679    0.05274  -8.852  < 2e-16 ***
berk.data$DeptC         -0.01621    0.04649  -0.349 0.727355
berk.data$DeptD         -0.16384    0.04832  -3.391 0.000696 ***
berk.data$DeptE         -0.46850    0.05276  -8.879  < 2e-16 ***
berk.data$DeptF         -0.26752    0.04972  -5.380 7.44e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2650.1  on 23  degrees of freedom
Residual deviance: 2097.7  on 16  degrees of freedom
AIC: 2272.7

Number of Fisher Scoring iterations: 5

> fits<-fitted(berk.ind)
> resids <- residuals(berk.ind,type="pearson")
> h <- lm.influence(berk.ind)$hat
> adjresids <- resids/sqrt(1-h)
> round(cbind(berk.data$Freq,fits,adjresids),2)
         fits adjresids
1   512 215.10     24.88
2   313 339.63     -1.97
3    89 146.68     -5.52
4    19 231.59    -17.40
5   353 134.87     22.42
6   207 212.95     -0.54
7    17  91.97     -8.85
8     8 145.21    -13.76
9   120 211.64     -7.73
10 205 334.17     -9.63
11 202 144.32      5.56
12 391 227.87     13.44
13 138 182.59     -4.01
14 279 288.30     -0.74
15 131 124.51      0.67
16 244 196.59      4.16
17  53 134.64     -8.40
18 138 212.59     -6.75
19  94  91.81      0.26
20 299 144.96     15.46
```

```
21  22 164.61    -13.41
22 351 259.91      7.55
23  24 112.25     -9.51
24 317 177.23     12.83
>
> # Saturated log-linear model
> ## via loglin()
> berk.sat<-loglin(UCBAdmissions, list(c(1,2,3)), fit=TRUE,
param=TRUE)
2 iterations: deviation 5.684342e-14
> berk.sat
$lrt
> # via glm()
> berk.sat<-glm(berk.data$Freq~
berk.data$Admit*berk.data$Gender*berk.data$Dept, family=poisson())
> summary(berk.sat)
> ###/* joint independence of Dept and Gender from Admit*/
> berk.join=glm(berk.data$Freq~berk.data$Admit+ berk.data$Gender+
berk.data$Dept+
berk.data$Gender*berk.data$Dept,family=poisson(link=log))
> summary(berk.join)

Call:
glm(formula = berk.data$Freq ~ berk.data$Admit + berk.data$Gender +
    berk.data$Dept +
 berk.data$Gender * berk.data$Dept, family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-12.744   -3.208   -0.058    2.495    9.869

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            5.76801    0.03951 145.992  < 2e-16 ***
berk.data$AdmitRejected    0.45674    0.03051  14.972  < 2e-16 ***
berk.data$GenderFemale   -2.03325    0.10233 -19.870  < 2e-16 ***
berk.data$DeptB   -0.38745    0.05475  -7.076 1.48e-12 ***
berk.data$DeptC     -0.93156    0.06549 -14.224  < 2e-16 ***
berk.data$DeptD    -0.68230    0.06008 -11.356  < 2e-16 ***
berk.data$DeptE    -1.46311    0.08030 -18.221  < 2e-16 ***
berk.data$DeptF    -0.79380    0.06239 -12.722  < 2e-16 ***
```

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2650.10  on 23  degrees of freedom
Residual deviance:  877.06  on 11  degrees of freedom
AIC: 1062.1

Number of Fisher Scoring iterations: 5

>
> # /*conditional independence of D and A given S */
> berk.cind=glm(berk.data$Freq~berk.data$Admit+
 berk.data$Gender+
berk.data$Dept+berk.data$Gender*berk.data$Dept+
berk.data$Admit*berk.data$Gender,family=poisson(link=log))
> summary(berk.cind)

Call:
glm(formula = berk.data$Freq ~ berk.data$Admit +
berk.data$Gender +
berk.data$Dept + berk.data$Gender * berk.data$Dept +
berk.data$Admit *
berk.data$Gender, family = poisson(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-14.1129   -3.6826    0.2158    2.9871    9.0983

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.90612    0.04093 144.299  < 2e-16 ***
berk.data$AdmitRejected 0.22013    0.03879   5.675 1.38e-08 ***
berk.data$GenderFemale   -2.41623    0.11038 -21.889  < 2e-16 ***
berk.data$DeptB   -0.38745    0.05475  -7.076 1.48e-12 ***
berk.data$DeptC     -0.93156    0.06549 -14.224  < 2e-16 ***
berk.data$DeptD     -0.68230    0.06008 -11.356  < 2e-16 ***
berk.data$DeptE     -1.46311    0.08030 -18.221  < 2e-16 ***
berk.data$DeptF     -0.79380    0.06239 -12.722  < 2e-16 ***

---
```

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2650.10  on 23  degrees of freedom
Residual deviance:  783.61  on 10  degrees of freedom
AIC: 970.67

Number of Fisher Scoring iterations: 5

> anova(berk.cind)
Analysis of Deviance Table

Model: poisson, link: log

Response: berk.data$Freq

Terms added sequentially (first to last)


                              Df Deviance Resid. Df Resid. Dev
NULL                                             23    2650.10
berk.data$Admit                1   230.03        22    2420.07
berk.data$Gender               1   162.87        21    2257.19
berk.data$Dept                 5   159.52        16    2097.67
berk.data$Gender:berk.data$Dept   5  1220.61        11     877.06
berk.data$Admit:berk.data$Gender  1    93.45        10     783.61
>
> ### /*homogeneous associations */
> berk.hom=glm(berk.data$Freq~berk.data$Admit+
 berk.data$Gender+berk.data$Dept + berk.data$Gender+
berk.data$Dept*berk.data$Gender
+berk.data$Dept*berk.data$Admit+
berk.data$Admit*berk.data$Gender, family=poisson(link=log))
> summary(berk.hom)



> anova(berk.hom)
Analysis of Deviance Table

Model: poisson, link: log
```

```
Response: berk.data$Freq

Terms added sequentially (first to last)


                                  Df Deviance Resid. Df Resid. Dev
NULL                                                 23    2650.10
berk.data$Admit                    1   230.03        22    2420.07
berk.data$Gender                   1   162.87        21    2257.19
berk.data$Dept                     5   159.52        16    2097.67
berk.data$Gender:berk.data$Dept    5  1220.61        11     877.06
berk.data$Admit:berk.data$Dept     5   855.32         6      21.74
berk.data$Admit:berk.data$Gender   1     1.53         5      20.20
>
```

3. (15 pts) Consider log-linear model (WXZ, WYZ) for 4 random variables, X, Y, Z, W.

   (a) Draw its independence graph, and identify variables that are conditionally independent.

   Solution: X and Y are conditionally independent given W and Z.

   (b) Explain why this is the most general log-linear model for a four-way table for which X and Y are conditionally independent.

   Solution: The other models for which X and Y are conditionally independent are (XZ, YZ) and (XW, YW). Both models are contained in (WXZ, WYZ). Therefore, (WXZ, WYZ) is the most general log-linear model for which X and Y are conditionally independent .