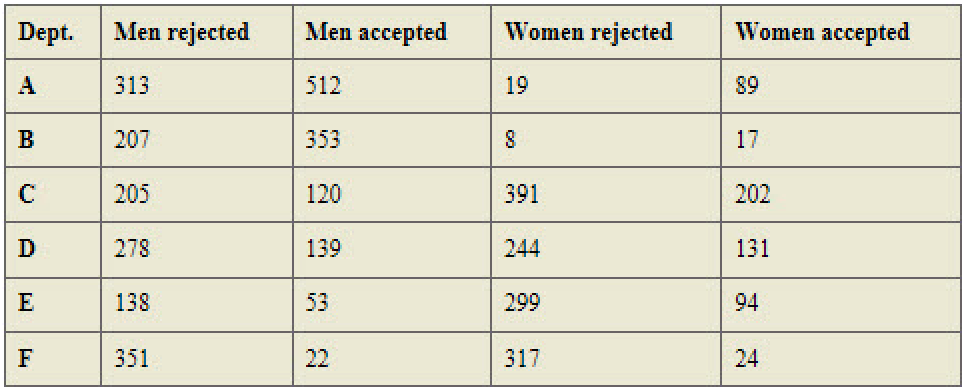
Lecture 4

1 Loglinear Model for 3-Way Table

This example will use the Berkeley Admissions data in R. Data was collected on graduate admissions at University of California Berkeley in 1973 for the six largest departments:



Dept.	Men rejected	Men accepted	Women rejected	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

This is a 6x2x2 table of three categorical variables: department, gender, and admission status. We can pull up this data automatically stored in R.

```r
rm(list=ls())
data = as.data.frame(UCBAdmissions)
data
```

```
##       Admit Gender Dept Freq
## 1  Admitted   Male    A  512
## 2  Rejected   Male    A  313
## 3  Admitted Female    A   89
## 4  Rejected Female    A   19
## 5  Admitted   Male    B  353
## 6  Rejected   Male    B  207
## 7  Admitted Female    B   17
## 8  Rejected Female    B    8
## 9  Admitted   Male    C  120
## 10 Rejected   Male    C  205
## 11 Admitted Female    C  202
## 12 Rejected Female    C  391
## 13 Admitted   Male    D  138
## 14 Rejected   Male    D  279
## 15 Admitted Female    D  131
## 16 Rejected Female    D  244
## 17 Admitted   Male    E   53
## 18 Rejected   Male    E  138
## 19 Admitted Female    E   94
## 20 Rejected Female    E  299
## 21 Admitted   Male    F   22
## 22 Rejected   Male    F  351
## 23 Admitted Female    F   24
## 24 Rejected Female    F  317
```

The data is read-in as one observation per cell with four columns corresponding to the cell count and the three categorical variables. There are two ways to fit a loglinear model in R: (1) the *loglin* command which operates on the 6x2x2 table, and (2) the *glm* command for a Poisson generalized linear model that operates on dataset in the format we have here.

1.1 The Saturated Model

Let's fit the saturated model using *glm* (which uses the Fisher Scoring algorithm for iterative fitting):

```
model = glm(Freq ~ Admit*Gender*Dept, data=data, family=poisson())
summary(model)
```

```
## 
## Call:
## glm(formula = Freq ~ Admit * Gender * Dept, family = poisson(), 
##     data = data)
## 
## Deviance Residuals: 
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [24]  0
## 
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                       6.23832    0.04419 141.157  < 2e-16 ***
## AdmitRejected                    -0.49212    0.07175  -6.859 6.94e-12 ***
## GenderFemale                     -1.74969    0.11484 -15.235  < 2e-16 ***
## DeptB                            -0.37186    0.06918  -5.375 7.65e-08 ***
## DeptC                            -1.45083    0.10142 -14.305  < 2e-16 ***
## DeptD                            -1.31107    0.09591 -13.669  < 2e-16 ***
## DeptE                            -2.26803    0.14430 -15.718  < 2e-16 ***
## DeptF                            -3.14728    0.21773 -14.455  < 2e-16 ***
## AdmitRejected:GenderFemale       -1.05208    0.26271  -4.005 6.21e-05 ***
## AdmitRejected:DeptB              -0.04163    0.11319  -0.368  0.71304    
## AdmitRejected:DeptC               1.02764    0.13550   7.584 3.34e-14 ***
## AdmitRejected:DeptD               1.19608    0.12641   9.462  < 2e-16 ***
## AdmitRejected:DeptE               1.44908    0.17681   8.196 2.49e-16 ***
## AdmitRejected:DeptF               3.26187    0.23120  14.109  < 2e-16 ***
## GenderFemale:DeptB               -1.28357    0.27358  -4.692 2.71e-06 ***
## GenderFemale:DeptC                2.27046    0.16270  13.954  < 2e-16 ***
## GenderFemale:DeptD                1.69763    0.16754  10.133  < 2e-16 ***
## GenderFemale:DeptE                2.32269    0.20663  11.241  < 2e-16 ***
## GenderFemale:DeptF                1.83670    0.31672   5.799 6.66e-09 ***
## AdmitRejected:GenderFemale:DeptB  0.83205    0.51039   1.630  0.10306    
## AdmitRejected:GenderFemale:DeptC  1.17700    0.29956   3.929 8.53e-05 ***
## AdmitRejected:GenderFemale:DeptD  0.97009    0.30262   3.206  0.00135 ** 
## AdmitRejected:GenderFemale:DeptE  1.25226    0.33032   3.791  0.00015 ***
## AdmitRejected:GenderFemale:DeptF  0.86318    0.40267   2.144  0.03206 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 2.6501e+03  on 23  degrees of freedom
## Residual deviance: 1.1191e-13  on  0  degrees of freedom
## AIC: 207.06
##
## Number of Fisher Scoring iterations: 3
```

R creates dummy variables for each of the levels in each categorical variable and the $*$ between variables in the *glm* call indicates to include ALL interactions. Note that one level for each categorical variable is left out (in the main effect dummy variables and all interactions) to satistfy the loglinear model constraints. The level that is left out is called the *reference group*: Male, Admitted and Dept A in this case.

The coefficient estimates provide information about the table association. For example, let's look at the effect of gender and admitted status together. The coefficient for the interaction *AdmitRejected:GenderFemale* is -1.05208. Thus the estimated odds ratio for gender and acceptance in department A (the reference group) is $exp(-1.05208) = .3492$. That is, women are less likely to be rejected than men in department A. The estimated odds ratio for gender and acceptance in department F (not the reference group) is $exp(-1.05208 + 0.86318) = .828$. Note that the three-way interaction coefficient *AdmitRejected:GenderFemale:DeptF* is statistically significant (p-value = .03206) indicating that the difference in the odds ratio for department A (.3492) and department F (.828) is statistically significant.

Wald hypothesis tests and confidence intervals for the odds ratio can be easily calculated using the standard error.

The *Residual Deviance* of approximately zero and associated zero degrees of freedom indicates the perfect fit of the saturated model. This is the likelihood ratio statistic, G^2, comparing this model (the saturated model) to the saturated model. The perfect fit is also evident by comparing the fitted values from the model to the observed values.

```r
cbind(model$fitted.values,data$Freq)
```

```
##    [,1] [,2]
## 1   512  512
## 2   313  313
## 3    89   89
## 4    19   19
## 5   353  353
## 6   207  207
## 7    17   17
## 8     8    8
## 9   120  120
## 10  205  205
## 11  202  202
## 12  391  391
## 13  138  138
## 14  279  279
## 15  131  131
## 16  244  244
## 17   53   53
## 18  138  138
## 19   94   94
## 20  299  299
## 21   22   22
## 22  351  351
## 23   24   24
## 24  317  317
```

The *Null Deviance* of 2650.1 is the likelihood ratio test statistic for the intercept only model. Because $2650.1 > 35.1724616 = \chi^2_{23}$ we can reject the intercept only model.

```r
qchisq(.95,model$df.null)
```

```
## [1] 35.17246
```

```r
1-pchisq(model$null.deviance,model$df.null)
```

```
## [1] 0
```

1.2 Complete Independence Model

Let's fit the complete independence model using *glm*:

```r
model1 = glm(Freq ~ Admit + Gender + Dept, data=data, family=poisson())
summary(model1)
```

```
##
## Call:
## glm(formula = Freq ~ Admit + Gender + Dept, family = poisson(),
##     data = data)
##
## 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 ***
## AdmitRejected  0.45674    0.03051  14.972  < 2e-16 ***
## GenderFemale  -0.38287    0.03027 -12.647  < 2e-16 ***
## DeptB         -0.46679    0.05274  -8.852  < 2e-16 ***
## DeptC         -0.01621    0.04649  -0.349 0.727355
## DeptD         -0.16384    0.04832  -3.391 0.000696 ***
## DeptE         -0.46850    0.05276  -8.879  < 2e-16 ***
## 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
```

All main effects are statistically significant with the exception of department C compared to department A. For example, the odds of being rejected are $exp(0.45674) = 1.579$.

The *Residual Deviance* of 2097.7 and associated 16 degrees of freedom indicates that the model with only main effects does not fit well compared to the saturated model because $2097.7 > 26.2962276 = \chi^2_{16}$. The lack of fit is also evident comparing the fitted values from the model to the observed values.

```r
qchisq(.95,model1$df.residual)
```

```
## [1] 26.29623
```

```r
1-pchisq(model$deviance,model$df.residual)
```

```
## [1] 0
```

```r
cbind(model1$fitted.values,data$Freq)
```

```
##         [,1] [,2]
## 1  215.10146  512
## 2  339.62744  313
## 3  146.67825   89
## 4  231.59285   19
## 5  134.87069  353
## 6  212.94968  207
## 7   91.96868   17
## 8  145.21095    8
## 9  211.64324  120
## 10 334.16719  205
## 11 144.32008  202
## 12 227.86949  391
## 13 182.59417  138
## 14 288.30110  279
## 15 124.51144  131
## 16 196.59328  244
## 17 134.64014   53
## 18 212.58566  138
## 19  91.81147   94
## 20 144.96272  299
## 21 164.61141   22
## 22 259.90781  351
## 23 112.24895   24
## 24 177.23182  317
```

We can look at the standardized Pearson residuals to see where the model fit is particularly poor - in most cells!

```r
resids = residuals(model1,type="pearson")
h = lm.influence(model1)$hat
adjresids = resids/sqrt(1-h)
cbind(data,model1$fitted.values,adjresids)
```

```
##       Admit Gender Dept Freq model1$fitted.values   adjresids
## 1  Admitted   Male    A  512            215.10146  24.8802909
## 2  Rejected   Male    A  313            339.62744  -1.9711664
## 3  Admitted Female    A   89            146.67825  -5.5209782
## 4  Rejected Female    A   19            231.59285 -17.4033457
## 5  Admitted   Male    B  353            134.87069  22.4162147
## 6  Rejected   Male    B  207            212.94968  -0.5380980
## 7  Admitted Female    B   17             91.96868  -8.8463067
## 8  Rejected Female    B    8            145.21095 -13.7636600
## 9  Admitted   Male    C  120            211.64324  -7.7321646
## 10 Rejected   Male    C  205            334.16719  -9.6255354
## 11 Admitted Female    C  202            144.32008   5.5601331
## 12 Rejected Female    C  391            227.86949  13.4448838
## 13 Admitted   Male    D  138            182.59417  -4.0072036
## 14 Rejected   Male    D  279            288.30110  -0.7371577
## 15 Admitted Female    D  131            124.51144   0.6674523
## 16 Rejected Female    D  244           196.59328   4.1600383
## 17 Admitted   Male    E   53           134.64014  -8.3962737
## 18 Rejected   Male    E  138           212.58566  -6.7507677
## 19 Admitted Female    E   94            91.81147   0.2584495
## 20 Rejected Female    E  299           144.96272  15.4633866
## 21 Admitted   Male    F   22           164.61141 -13.4083131
## 22 Rejected   Male    F  351           259.90781   7.5474457
## 23 Admitted Female    F   24           112.24895  -9.5092774
## 24 Rejected Female    F  317           177.23182  12.8305813
```

In your homework you will continue this example assessing the fit of all possible loglinear models to choose the *best* model…