Homework 5

Shun-Lung Chang, Dilip Hiremath

```
# import packages
library(vcd)
library(RcmdrMisc) # stepwise()
library(tidyr) # spread()
```

1. (2 points) Construct the two-way table for gender and whether admitted. Find the odds ratio for admission of males vs. females and interpret. For which gender is the probability of admission higher?

The table below shows the numbers of admitted students with repestive to genders, and the oddsratio for admission of males vs. females is 1.84108. Given that this ratio is greater than 1, male students are more likely to get admission than the females ones.

```
UCBAdmissions_m <- margin.table(UCBAdmissions, c(2, 1))
UCBAdmissions_m
```

```
        Admit
Gender   Admitted Rejected
  Male       1198     1493
  Female      557     1278
```

```
oddsratio(UCBAdmissions_m, log = FALSE)
```

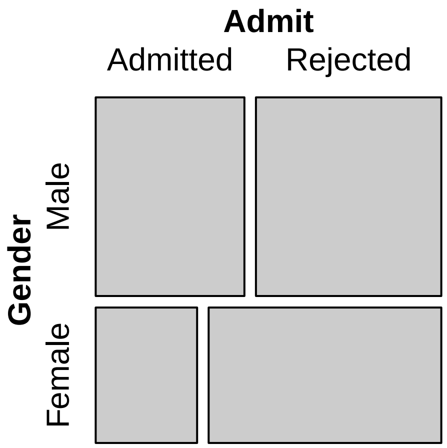
```
 odds ratios for Gender and Admit

[1] 1.84108
```

2. Draw a mosaic plot for the data aggregated over departments.

The following shows the mosaic plot for the aggregated data.

```
mosaic(UCBAdmissions_m)
```



3. (3 points) Fit a saturated log-linear model to the frequencies using A and G as predictors. Calculate the odds ratio for admission of males vs. females. Display the predicted frequencies in a table and compare them to the observed frequencies displayed in Question 1.

The log-linear model with frequencies as the response and A and G as predictors is shown as below. The odds ratio for admission of males vs. females can be obtained by taking exponential of the coefficient of the *AdmitRejected:GenderFemale* term, and the odd ratio is 1.84108, which is same as the ratio in task 1.

```
berkeley <- data.frame(UCBAdmissions)
mod_1 <- glm(Freq ~ Admit * Gender,
             family = 'poisson', data = berkeley)
summary(mod_1)
```

```
Call:
glm(formula = Freq ~ Admit * Gender, family = "poisson", data = berkeley)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-18.9074   -7.8908   -0.1399    5.6871   18.4285

Coefficients:
                           Estimate Std. Error z value     Pr(>|z|)
(Intercept)                 5.29665    0.02889 183.329      < 2e-16
AdmitRejected               0.22013    0.03879   5.675 0.0000000138
GenderFemale               -0.76584    0.05128 -14.933      < 2e-16
AdmitRejected:GenderFemale  0.61035    0.06389   9.553      < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2650.1  on 23  degrees of freedom
Residual deviance: 2163.7  on 20  degrees of freedom
AIC: 2330.8

Number of Fisher Scoring iterations: 5
```

```
exp(unname(mod_1$coefficients[4]))
```

```
[1] 1.84108
```

The predicted frequency table is same as the observed frequency table in task 1.

```
mod_table <- xtabs(mod_1$fitted.values ~ berkeley$Admit + berkeley$Gender)
mod_table
```

```
              berkeley$Gender
berkeley$Admit Male Female
      Admitted 1198    557
      Rejected 1493   1278
```

4. (2 points) Fit an independence log-linear model to the frequencies using A, D and G as predictors. Display the predicted frequencies in a table and compare them to the observed frequencies. Calculate the estimated odds ratios for admission of males vs. females for each department.

The requested model is shown as below. In the predicted frequency table, the ratio of the accepted to the rejected, regardless of gender, is always the same (0.6333454) in each department. The reason is that this ratio can be simplied to the ratio of total number of accepted students to total number of rejected students under the model settings. Given the same ratios, the estimated odds ratios are all equal to 1.

```r
mod_2 <- glm(Freq ~ Admit + Gender + Dept,
             family = 'poisson',
             data = berkeley)
summary(mod_2)
```

```

Call:
glm(formula = Freq ~ Admit + Gender + Dept, family = "poisson",
    data = berkeley)

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 0.0000000744

(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
```

```r
# predicted values (mod_2)
mod_2_table <- xtabs(mod_2$fitted.values ~ berkeley$Gender +
                         berkeley$Admit + berkeley$Dept)

# observed values
obs_table <- xtabs(berkeley$Freq ~ berkeley$Gender +
                         berkeley$Admit + berkeley$Dept)
```

```r
structable(obs_table)
```

```
                              berkeley$Admit Admitted Rejected
berkeley$Gender berkeley$Dept
```

```
Male            A                                 512      313
                B                                 353      207
                C                                 120      205
                D                                 138      279
                E                                  53      138
                F                                  22      351
Female          A                                  89       19
                B                                  17        8
                C                                 202      391
                D                                 131      244
                E                                  94      299
                F                                  24      317
```

```r
structable(mod_2_table)
```

```
                              berkeley$Admit  Admitted  Rejected
berkeley$Gender berkeley$Dept                                   
Male            A                            215.10146 339.62744
                B                            134.87069 212.94968
                C                            211.64324 334.16719
                D                            182.59417 288.30110
                E                            134.64014 212.58566
                F                            164.61141 259.90781
Female          A                            146.67825 231.59285
                B                             91.96868 145.21095
                C                            144.32008 227.86949
                D                            124.51144 196.59328
                E                             91.81147 144.96272
                F                            112.24895 177.23182
```

```r
# odds ratio for observed values
(obs_table[1, 1, ]/obs_table[1, 2, ])/(obs_table[2, 1, ]/obs_table[2, 2, ])
```

```
        A         B         C         D         E         F 
0.3492120 0.8025007 1.1330596 0.9212838 1.2216312 0.8278727 
```

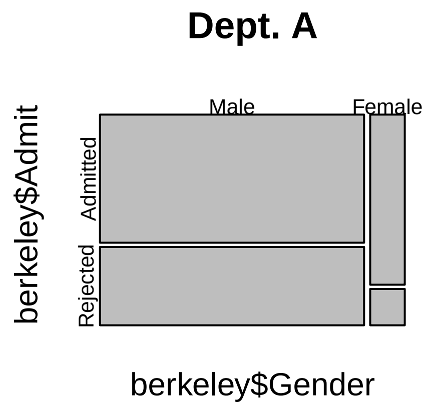
```r
# odds ratio for predicted values
(mod_2_table[1, 1, ]/mod_2_table[1, 2, ])/(mod_2_table[2, 1, ]/mod_2_table[2, 2, ])
```

```
A B C D E F 
1 1 1 1 1 1 
```

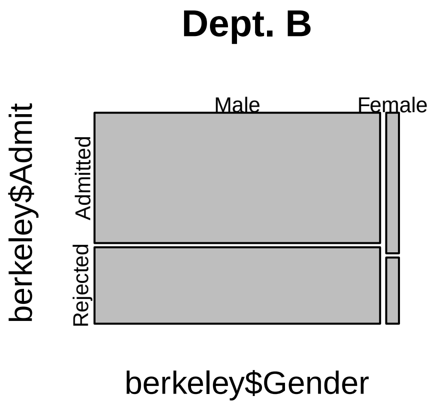
5. Draw mosaic plots of admission versus gender for each department separately.

The following plots demonstrate admission versus gender for each department separately.

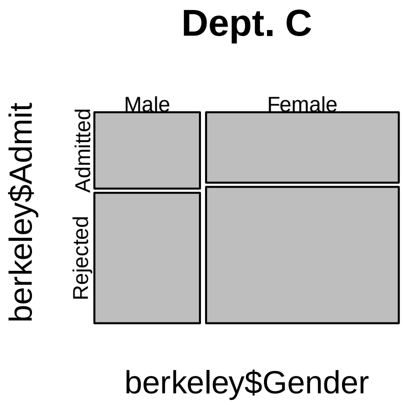
```r
mosaicplot(obs_table[1:2,1:2,1], main = "Dept. A")
```



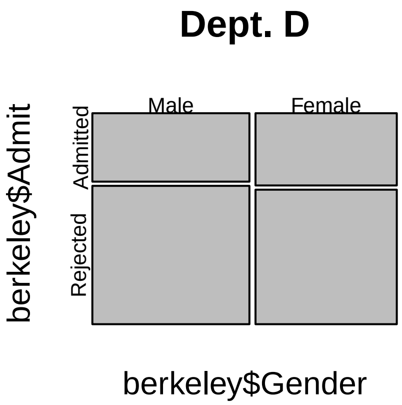
```r
mosaicplot(obs_table[1:2,1:2,2], main = "Dept. B")
```



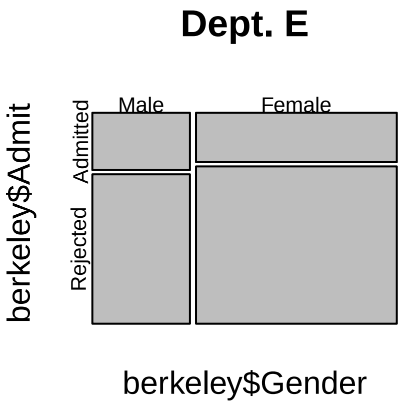
```r
mosaicplot(obs_table[1:2,1:2,3], main = "Dept. C")
```



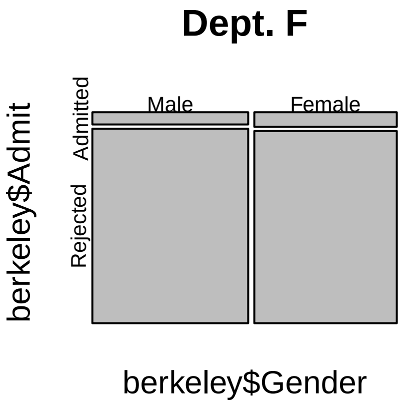
```r
mosaicplot(obs_table[1:2,1:2,4], main = "Dept. D")
```



```r
mosaicplot(obs_table[1:2,1:2,5], main = "Dept. E")
```



```r
mosaicplot(obs_table[1:2,1:2,6], main = "Dept. F")
```



6. (3 points) Fit a log-linear model to the frequencies using A, D and G as predictors that includes all two-way interactions but not the three-way interaction. Display the predicted frequencies in a table and compare them to the predicted frequencies displayed in Question 4 as well as the observed frequencies. Calculate the estimated odds ratio for admission of males vs. females in this model.

The request model and the predicted frequencies are shown below. We can see the ratios of the accepted to the rejected differ accross departments and genders, since the two-way interactions were included in this model. Besides, the predicted frequencies are closer to the observed frequencies. The estimated odds ratio for admission of males vs. females was also computed, and the value is 1.84108, which is the same as the odds ratio in task 1.

```
mod_3 <- glm(Freq ~ Admit + Gender + Dept +
                 Admit * Dept + Gender * Dept + Admit * Gender,
             family = 'poisson',
             data = berkeley)
summary(mod_3)
```

```

Call:
glm(formula = Freq ~ Admit + Gender + Dept + Admit * Dept + Gender * 
    Dept + Admit * Gender, family = "poisson", data = berkeley)

Deviance Residuals: 
       1         2         3         4         5         6         7  
-0.75481   0.99471   1.96454  -3.15768  -0.03402   0.04449   0.15709  
       8         9        10        11        12        13        14  
-0.22034   1.01273  -0.73839  -0.74367   0.54896   0.06760  -0.04741  
      15        16        17        18        19        20        21  
-0.06911   0.05080   1.05578  -0.61236  -0.73617   0.42678  -0.20117  
      22        23        24  
 0.05113   0.19803  -0.05370  

Coefficients:
                           Estimate Std. Error z value      Pr(>|z|)    
(Intercept)                 6.27150    0.04271 146.855       < 2e-16 ***
AdmitRejected              -0.58205    0.06899  -8.436       < 2e-16 ***
GenderFemale               -1.99859    0.10593 -18.866       < 2e-16 ***
DeptB                      -0.40322    0.06784  -5.944 0.00000000278 ***
DeptC                      -1.57790    0.08949 -17.632       < 2e-16 ***
DeptD                      -1.35000    0.08526 -15.834       < 2e-16 ***
DeptE                      -2.44982    0.11755 -20.840       < 2e-16 ***
DeptF                      -3.13787    0.16174 -19.401       < 2e-16 ***
AdmitRejected:DeptB         0.04340    0.10984   0.395         0.693    
AdmitRejected:DeptC         1.26260    0.10663  11.841       < 2e-16 ***
AdmitRejected:DeptD         1.29461    0.10582  12.234       < 2e-16 ***
AdmitRejected:DeptE         1.73931    0.12611  13.792       < 2e-16 ***
AdmitRejected:DeptF         3.30648    0.16998  19.452       < 2e-16 ***
GenderFemale:DeptB         -1.07482    0.22861  -4.701 0.00000258269 ***
GenderFemale:DeptC          2.66513    0.12609  21.137       < 2e-16 ***
GenderFemale:DeptD          1.95832    0.12734  15.379       < 2e-16 ***
GenderFemale:DeptE          2.79519    0.13925  20.073       < 2e-16 ***
GenderFemale:DeptF          2.00232    0.13571  14.754       < 2e-16 ***
AdmitRejected:GenderFemale -0.09987    0.08085  -1.235         0.217    

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2650.095  on 23  degrees of freedom
Residual deviance:   20.204  on  5  degrees of freedom
AIC: 217.26

Number of Fisher Scoring iterations: 4
```

```
mod_3_table <- xtabs(mod_3$fitted.values ~ berkeley$Gender +
                         berkeley$Admit + berkeley$Dept)
structable(mod_3_table)
```

```
                              berkeley$Admit   Admitted   Rejected
berkeley$Gender berkeley$Dept
Male            A                            529.269919 295.730081
                B                            353.639509 206.360491
                C                            109.245276 215.754724
                D                            137.207390 279.792610
                E                             45.680810 145.319190
                F                             22.957096 350.042904
Female          A                             71.730081  36.269919
                B                             16.360491   8.639509
                C                            212.754724 380.245276
                D                            131.792610 243.207390
                E                            101.319190 291.680810
                F                             23.042904 317.957096
```

```
(sum(mod_3_table[1, 1, ])/sum(mod_3_table[1, 2, ]))/(sum(mod_3_table[2, 1, ])/sum(mod_3_table[2, 2, ]))
```

```
[1] 1.84108
```

7. (2 points) Starting with the model in Question 6, use the stepwise method with the backward/forward option and BIC as criterion. Does this result in a simpler model? Interpret this model in plain English.

The model obtained by stepwise method is shown below, and it omitted the interaction term between *Admit* and *Gender*.

In a three-main-effect model, if we have already known that the interaction between A and B is independent of that between B and C (denoted as **(AB, BC)**) and the interaction between B and C is independent of that between A and C, which is **(BC, AC)**, we can safely conclude that the interaction between A and B is independent of that between A and C (**(AB, AC)**). Therefore, the redundant interaction term was omitted.

```
mod_4 <- stepwise(mod_3, direction = 'backward/forward', criterion = 'BIC')
```

```
summary(mod_4)
```

```

Call:
glm(formula = Freq ~ Admit + Gender + Dept + Admit:Dept + Gender:Dept,
    family = "poisson", data = berkeley)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4776  -0.4144   0.0098   0.3089   2.2321

Coefficients:
                    Estimate Std. Error z value      Pr(>|z|)
(Intercept)          6.27557    0.04248 147.744       < 2e-16
AdmitRejected       -0.59346    0.06838  -8.679       < 2e-16
GenderFemale        -2.03325    0.10233 -19.870       < 2e-16
DeptB               -0.40575    0.06770  -5.993 0.00000000206
DeptC               -1.53939    0.08305 -18.536       < 2e-16
```

```
DeptD               -1.32234    0.08159 -16.207       < 2e-16
DeptE               -2.40277    0.11014 -21.816       < 2e-16
DeptF               -3.09624    0.15756 -19.652       < 2e-16
AdmitRejected:DeptB  0.05059    0.10968   0.461         0.645
AdmitRejected:DeptC  1.20915    0.09726  12.432       < 2e-16
AdmitRejected:DeptD  1.25833    0.10152  12.395       < 2e-16
AdmitRejected:DeptE  1.68296    0.11733  14.343       < 2e-16
AdmitRejected:DeptF  3.26911    0.16707  19.567       < 2e-16
GenderFemale:DeptB  -1.07581    0.22860  -4.706 0.00000252480
GenderFemale:DeptC   2.63462    0.12343  21.345       < 2e-16
GenderFemale:DeptD   1.92709    0.12464  15.461       < 2e-16
GenderFemale:DeptE   2.75479    0.13510  20.391       < 2e-16
GenderFemale:DeptF   1.94356    0.12683  15.325       < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2650.095  on 23  degrees of freedom
Residual deviance:   21.736  on  6  degrees of freedom
AIC: 216.8

Number of Fisher Scoring iterations: 4
```

8. (2 points) Calculate the saturated model using all three predictors and compare this model to the one obtained in Question 7 using AIC, BIC and the deviance criterion. Give a verbal summary of your findings.

The saturated model is shown below. A decrease can be seen in AIC and BIC scores and deviance criterion if we compare the model to the one in task 7. Furthermore, the predicted frequency table is the same as the observed table, since all the interaction terms were all taken into account in this model.

```r
mod_5 <- glm(Freq ~ Admit * Gender * Dept,
             family = 'poisson',
             data = berkeley)
summary(mod_5)
```

```
Call:
glm(formula = Freq ~ Admit * Gender * Dept, family = "poisson",
    data = berkeley)

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

(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
# compare AIC scores
AIC(mod_4)
```

```
[1] 216.7952
```

```r
AIC(mod_5)
```

```
[1] 207.0597
```

```r
# compare BIC scores
BIC(mod_4)
```

```
[1] 238.0002
```

```r
BIC(mod_5)
```

```
[1] 235.333
```

```r
# Compute DIC = -2 * log(likelihood)
-2 * logLik(mod_4)
```

```
'log Lik.' 180.7952 (df=18)
```

```r
-2 * logLik(mod_5)
```

```
'log Lik.' 159.0597 (df=24)
```

```r
mod_5_table <- xtabs(mod_5$fitted.values ~ berkeley$Gender +
                         berkeley$Admit + berkeley$Dept)
structable(mod_5_table)
```

```
                              berkeley$Admit Admitted Rejected
berkeley$Gender berkeley$Dept
```

```
Male            A                                  512      313
                B                                  353      207
                C                                  120      205
                D                                  138      279
                E                                   53      138
                F                                   22      351
Female          A                                   89       19
                B                                   17        8
                C                                  202      391
                D                                  131      244
                E                                   94      299
                F                                   24      317
```

9. To run a logistic regression model with aggregated data it is best to create a data frame that comprises two frequency variables: one for the numbers of admitted students, and one for the numbers of rejected students.

```
  UCBAdmissions <- as.data.frame(UCBAdmissions)
  attach(UCBAdmissions)
  UCBAdmit <- cbind(UCBAdmissions[Admit=="Rejected",-1],
     A=UCBAdmissions[Admit=="Admitted", "Freq"])
  names(UCBAdmit)[3] <- "R"
  detach()
```

Treating cbind(A,R) as response and D and G as qualitative predictors, fit the logit model having main effects only.

```
berkeley_t <- spread(berkeley, Admit, Freq)

mod_6 <- glm(cbind(Rejected, Admitted) ~  Gender * Dept,
             family = binomial(link = "logit"),
             data = berkeley_t)
summary(mod_6)
```

```
Call:
glm(formula = cbind(Rejected, Admitted) ~ Gender * Dept, family = binomial(link = "logit"),
    data = berkeley_t)

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

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.49212    0.07175  -6.859 6.94e-12
GenderFemale       -1.05208    0.26271  -4.005 6.21e-05
DeptB              -0.04163    0.11319  -0.368  0.71304
DeptC               1.02764    0.13550   7.584 3.34e-14
DeptD               1.19608    0.12641   9.462  < 2e-16
DeptE               1.44908    0.17681   8.196 2.49e-16
DeptF               3.26187    0.23120  14.109  < 2e-16
```

```
GenderFemale:DeptB  0.83205    0.51039   1.630  0.10306
GenderFemale:DeptC  1.17700    0.29956   3.929 8.53e-05
GenderFemale:DeptD  0.97009    0.30262   3.206  0.00135
GenderFemale:DeptE  1.25226    0.33032   3.791  0.00015
GenderFemale:DeptF  0.86318    0.40267   2.144  0.03206

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 877.0564132197750951  on 11  degrees of freedom
Residual deviance:   0.0000000000001088  on  0  degrees of freedom
AIC: 92.94

Number of Fisher Scoring iterations: 3
```

(a) (1.5 points) Report the prediction equation. Interpret the coefficients of D, G and the interaction in this equation.

The prediction equation and the coefficients of each term are shown above. If we add the coefficient of *GenderFemale* and the coefficients of interaction terms between *Dept*, and *Gender*, the results are exactly the logarithmic odds ratios for admission of males vs. females in each department. And the exponent of the results are the observed odd ratios in task 4.

```
exp(c(mod_6$coefficients[2], mod_6$coefficients[2] + mod_6$coefficients[8:12]))
```

```
      GenderFemale GenderFemale:DeptB GenderFemale:DeptC 
         0.3492120          0.8025007          1.1330596 
GenderFemale:DeptD GenderFemale:DeptE GenderFemale:DeptF 
         0.9212838          1.2216312          0.8278727 
```

(b) (half a point) To which log-linear model is this model equivalent?

The model is equivalent to the saturated model using all three predictors in task 8.