Question 1:

Code in R:

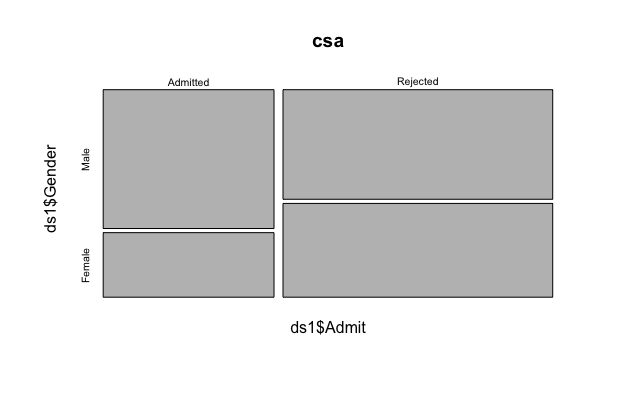
csa<-xtabs(ds1$Freq~ds1$Admit+ds1$Gender)

csa

(csa[1,1]/csa[2,1]) / (csa[1,2]/csa[2,2])

The probability of male gets admitted is 80,24% the probability of female gets admitted is 43.58%. The odd of male gets admitted to female ratio is 1.8:1.

Question 2:



Question 3:

Code in R:

temp1<-glm(Freq ~ Admit\*Gender, family=poisson(log), data=ds1)

summary(temp1)

ds1$pred<-temp1$fitted.values

csap<-xtabs(ds1$pred~ds1$Admit+ds1$Gender)

csap

The number is exactly the same with the frequency table in the question 1. The reason is because the saturated model make a perfect fit with the dataset we have and left no data for estimate variance. That’s why the result for the frequency table of predicted values has the same values as the table in question 1 with the same probability and odd ratio.

Question 4:

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.

Code in R:

temp2<-glm(Freq~Admit+Gender+Dept, family=poisson(log), data=ds1)

summary(temp2)

ds1$pred1<-temp2$fitted.values

csap1a<-xtabs(ds1$Freq~ds1$Gender+ds1$Admit+ds1$Dept)

csap1p<-xtabs(ds1$pred1~ds1$Gender+ds1$Admit+ds1$Dept)

library(vcd)

structable(csap1a)

structable(csap1p)

#using Admissions values

(csap1a[1,1,1]/csap1a[1,2,1])/(csap1a[2,1,1]/csap1a[2,2,1])

(csap1a[1,1,2]/csap1a[1,2,2])/(csap1a[2,1,2]/csap1a[2,2,2])

(csap1a[1,1,3]/csap1a[1,2,3])/(csap1a[2,1,3]/csap1a[2,2,3])

(csap1a[1,1,4]/csap1a[1,2,4])/(csap1a[2,1,4]/csap1a[2,2,4])

(csap1a[1,1,5]/csap1a[1,2,5])/(csap1a[2,1,5]/csap1a[2,2,5])

(csap1a[1,1,6]/csap1a[1,2,6])/(csap1a[2,1,6]/csap1a[2,2,6])

Odd ratios for department A: 0.349212:1

Odd ratios for department B: 0.8025007:1

Odd ratios for department C: 1.13306:1

Odd ratios for department D: 0.9212838:1

Odd ratios for department E: 1.221631:1

Odd ratios for department F: 0.8278727:1

#Using predicted Admissions values

(csap1p[1,1,1]/csap1p[1,2,1])/(csap1p[2,1,1]/csap1p[2,2,1])

(csap1p[1,1,2]/csap1p[1,2,2])/(csap1p[2,1,2]/csap1p[2,2,2])

(csap1p[1,1,3]/csap1p[1,2,3])/(csap1p[2,1,3]/csap1p[2,2,3])

(csap1p[1,1,4]/csap1p[1,2,4])/(csap1p[2,1,4]/csap1p[2,2,4])

(csap1p[1,1,5]/csap1p[1,2,5])/(csap1p[2,1,5]/csap1p[2,2,5])

(csap1p[1,1,6]/csap1p[1,2,6])/(csap1p[2,1,6]/csap1p[2,2,6])

All the odds ratios are 1.

Question 5:

Code in R:

par(mfrow=c(3,2))

mosaicplot(csap1a[1:2,1:2,1], main="Dept A")

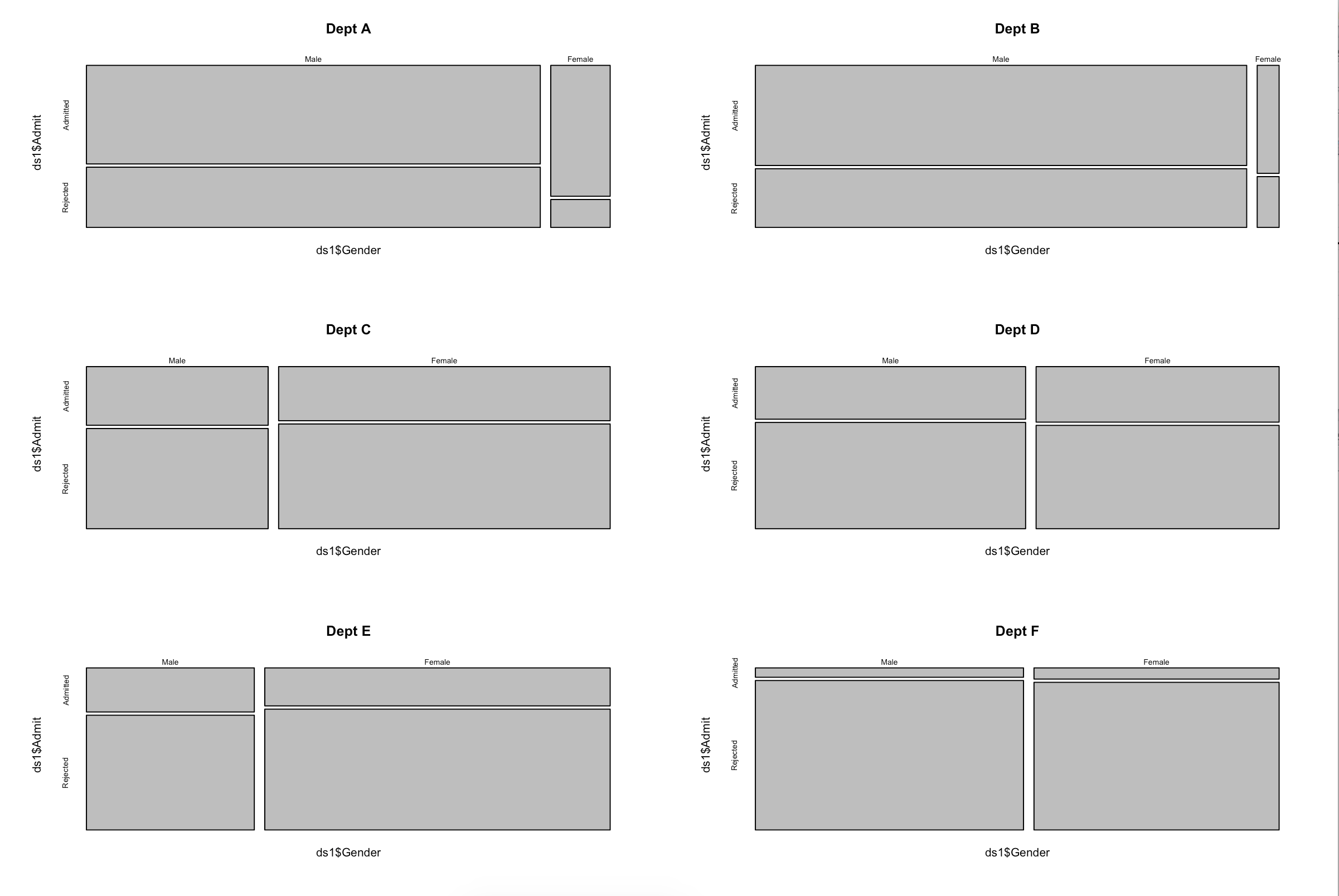
mosaicplot(csap1a[1:2,1:2,2], main="Dept B")

mosaicplot(csap1a[1:2,1:2,3], main="Dept C")

mosaicplot(csap1a[1:2,1:2,4], main="Dept D")

mosaicplot(csap1a[1:2,1:2,5], main="Dept E")

mosaicplot(csap1a[1:2,1:2,6], main="Dept F")



Question 6:

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.  
Code in R:

temp3<-glm(Freq ~ Admit + Gender + Dept + Admit\*Dept + Gender\*Dept +Admit\*Gender, family=poisson(log), data=ds1)

summary(temp3)

ds1$pred2<-temp3$fitted.values

csap2a<-xtabs(ds1$Freq~ds1$Gender+ds1$Admit+ds1$Dept)

csap2p<-xtabs(ds1$pred2~ds1$Gender+ds1$Admit+ds1$Dept)

structable(csap2a)

structable(csap2p)

Predicted frequencies table from : Admit+Gender+Dept

ds1$Admit Admitted Rejected

ds1$Gender ds1$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

Predicted frequencies table from : Admit + Gender + Dept + Admit\*Dept + Gender\*Dept +Admit\*Gender

ds1$Admit Admitted Rejected

ds1$Gender ds1$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

Frequencies table

ds1$Admit Admitted Rejected

ds1$Gender ds1$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

From the result we can clearly see that there is a big correlation between the interaction between the values of the features in the dataframe. The model with 2 ways interaction has closer value to the data values.

The odds ratios is: 1.84:1, which is almost the same as the frequencies from the source.

Question 7:

Code in R:

temp4 <- stepwise(temp3, "backward/forward", criterion="BIC")

summary(temp4)

The model doesn’t lead to simpler model. The model is still the same with all the predictors are relevant except for admitrejected in department B. The AIC number is 218.8

Question 8:

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 finding

Code in R:

temp5<-glm(Freq~Admit\*Gender\*Dept, family=poisson(log), data=ds1)

summary(temp5)

The result is almost the same. The weight for each predictors change a bit and you have 5 more parameters.AdmitRejected:GenderFemale:DeptB, AdmitRejected:GenderFemale:DeptC, AdmitRejected:GenderFemale:DeptD, AdmitRejected:GenderFemale:DeptE, AdmitRejected:GenderFemale:DeptF .

Question 9:

Code in R:

ds1 <- as.data.frame(ds1)

attach(ds1)

ds2 <- cbind(ds1[Admit=="Rejected",-1],

A=ds1[Admit=="Admitted", "Freq"])

names(ds2)[3] <- "R"

detach()

temp6<-glm(cbind(ds2$R, ds2$A)~Gender+Dept+Gender\*Dept, family=binomial(link="logit"), data=ds2)

summary(temp6)

1. The equation is:

Frequencies of female: -.49212-1.05208\*genderfemale - 0.04162\*DepartmentB + 1.02764\*DepartmentC + 1.19608\*DepartmentD + 1.44908\*DepartmentE + 3.26187\*DepartmentF + 0.83205\*GenderFemale:DeptB + 1.17700\*GenderFemale:DeptC + 0.97009\*GenderFemale:DeptD + 1.25226\*GenderFemale:DeptE + 0.86318\*GenderFemale:DeptF

This model has the AIC = 92,94

1. The model when you make bins for variable has the lowest AIC ⇒ this model is the most relevant.