R Notebook

Exercise 2, Chapter 6

The dataset melanoma gives data on a sample of patients suffering from melanoma (skin cancer) cross-classified by the type of cancer and the location on the body.

```
suppressMessages(library(faraway)); data(melanoma)
summary(melanoma)
```

```
##      count                  tumor          site
##  Min.   :  2.00   freckle      :3   extremity:4
##  1st Qu.: 14.75   indeterminate:3   head     :4
##  Median : 20.50   nodular      :3   trunk    :4
##  Mean   : 33.33   superficial  :3
##  3rd Qu.: 38.25
##  Max.   :115.00
```

```
head(melanoma)
```

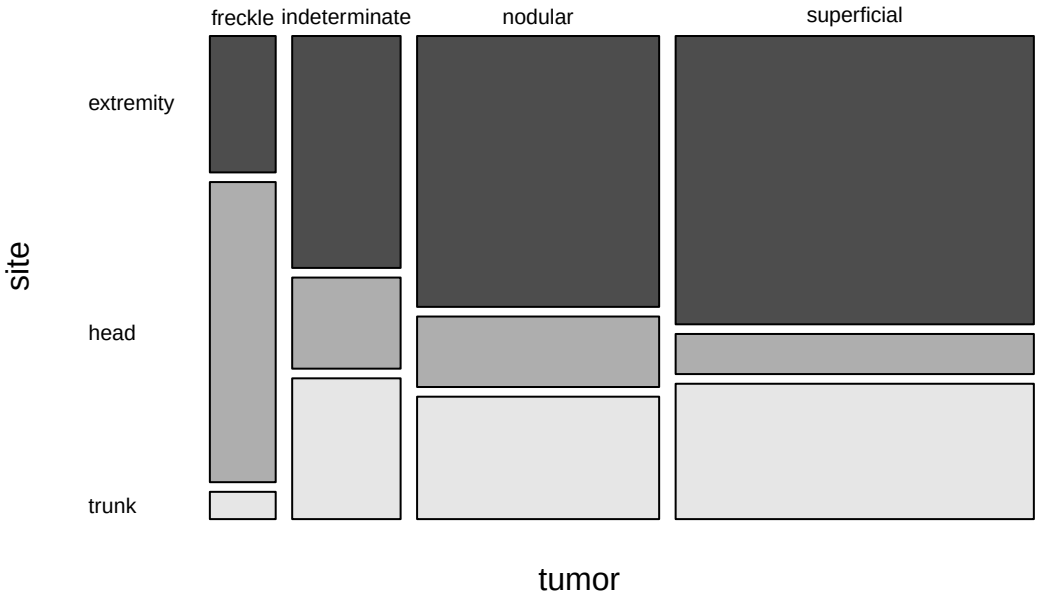
```
##   count         tumor  site
## 1    22       freckle  head
## 2    16   superficial  head
## 3    19       nodular  head
## 4    11 indeterminate  head
## 5     2       freckle trunk
## 6    54   superficial trunk
```

(a) Display the data in a two-way table. Make a mosaic plot and comment on the evidence for independence.

```
yep <- xtabs(formula = count ~ tumor + site, data = melanoma)
yep
```

```
##                site
## tumor           extremity head trunk
##   freckle              10   22     2
##   indeterminate        28   11    17
##   nodular              73   19    33
##   superficial         115   16    54
```

```
mosaicplot(yep, color=TRUE, main=NULL, las=1 )
```



It appears as though extremity increases and trunk increases as tumer goes from frackle to superficial. It does not appear that this these variables are independent.

(b) Check for independence between site and tumor type using a Chi-squared test.

```
summary(yep)
```

```
## Call: xtabs(formula = count ~ tumor + site, data = melanoma)
## Number of cases in table: 400
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 65.81, df = 6, p-value = 2.943e-12
```

Since the p-value for the chi-squared test is <0.05, we will reject the null hypothesis that there site and tumor are independent

(c) Fit a Poisson GLM model and use it to check for independence.

```
mods <- glm(count ~ tumor + site, data = melanoma, family=poisson)
pchisq(deviance(mods),df.residual(mods),lower=F)
```

```
## [1] 2.050453e-09
```

(d) Make a two-way table of the deviance residuals from the last model. Comment on the larger residuals.

```
xtabs(residuals(mods) ~ tumor + site, data = melanoma)
```

```
##                site
## tumor             extremity        head       trunk
##   freckle       -2.31583297  5.13537787 -2.82829426
##   indeterminate -0.66016102  0.46798432  0.54787007
##   nodular        0.28104581 -0.49711084 -0.02173229
##   superficial    1.00813975 -3.04533605  0.69899703
```

Exercise 10, Chapter 6

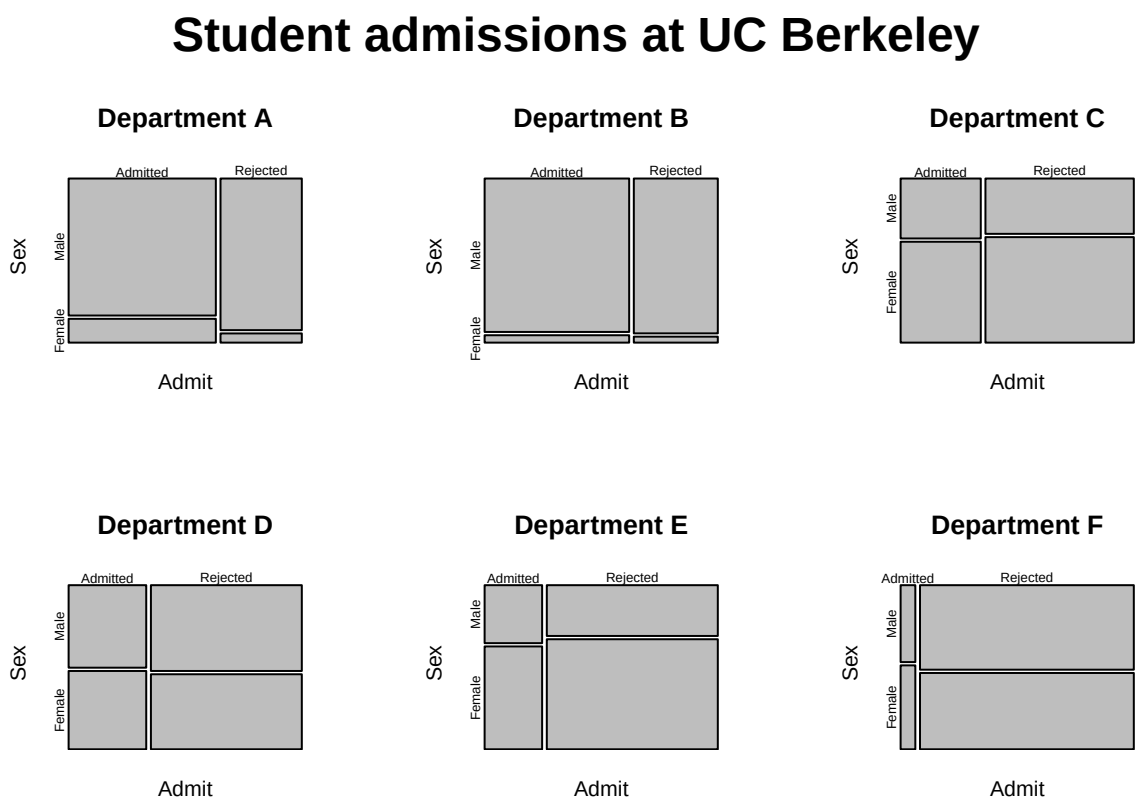
The UCB Admissions dataset presents data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

1. (a) Show that this provides an example of Simpson's paradox.

```
require(graphics)
## Data aggregated over departments
apply(UCBAdmissions, c(1, 2), sum)
```

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

```
## Data for individual departments
opar <- par(mfrow = c(2, 3), oma = c(0, 0, 2, 0))
for(i in 1:6)
  mosaicplot(UCBAdmissions[,,i],
    xlab = "Admit", ylab = "Sex",
    main = paste("Department", LETTERS[i]))
mtext(expression(bold("Student admissions at UC Berkeley")),
      outer = TRUE, cex = 1.5)
```




```
par(opar)
```

FROM rdrr.io "There were 2691 male applicants, of whom 1198 (44.5%) were admitted, compared with 1835 female applicants of whom 557 (30.4%) were admitted. This gives a sample odds ratio of 1.83, indicating that males were almost twice as likely to be admitted. In fact, graphical methods (as in the example below) or log-linear modelling show that the apparent association between admission and sex stems from differences in the tendency of males and females to apply to the individual departments (females used to apply more to departments with higher rejection rates)."

From a quick look at the over all data, it appears that women are more likely to be rejected than men. But when we partition the data by departmentand look at the gender vs. admit breakdowns, its clear that female students are applying to less of the departments with higher acceptance rates (ie, Dept A) and more of the

departments with higher rejection rates (ie, Dept F or C).

2. (b) Determine the most appropriate dependence model between the variables.

```
## ====================================
# Model Selection
ucb_df <- data.frame(UCBAdmissions)

modsat <- glm(Freq ~ Gender*Dept*Admit, ucb_df, family=poisson) # Most complicated

modu <- glm(Freq ~ (Gender+Dept+Admit)^2, ucb_df, family=poisson) # Simpler than above

1 - pchisq(q=(deviance(modu)- deviance(modsat)), df=(length(coef(modsat)) - length(coef(modu))))
```

```
## [1] 0.001144078
```

Since the pchisq test is significant, at this level, it seems that no further testing is neccesary. We have reached a dependence model that fits our data significantly, which is the uniform association Freq ~ (Gender+Dept+Admit)^2.

3. (c) Fit a binomial regression with admissions status as the response and show the relationship to your model in the previous question.

```
## Convert to dataframe & sort
ucb_df <- ucb_df[order(ucb_df$Admit),]
y_bin <- matrix(ucb_df$Freq,ncol=2) #First column is accepted, second column is rejected
portion <- ucb_df[1:12,]

modbin <- glm(y_bin ~ (Gender+Dept)^2, portion, family=binomial)

modbin
```

```
##
## Call:  glm(formula = y_bin ~ (Gender + Dept)^2, family = binomial, data = portion)
##
## Coefficients:
##        (Intercept)        GenderFemale               DeptB
##            0.49212             1.05208             0.04163
##              DeptC               DeptD               DeptE
##           -1.02764            -1.19608            -1.44908
##              DeptF  GenderFemale:DeptB  GenderFemale:DeptC
##           -3.26187            -0.83205            -1.17700
## GenderFemale:DeptD  GenderFemale:DeptE  GenderFemale:DeptF
##           -0.97009            -1.25226            -0.86318
##
## Degrees of Freedom: 11 Total (i.e. Null);  0 Residual
## Null Deviance:       877.1
## Residual Deviance: 1.159e-13     AIC: 92.94
```

This binomialmodel seems to fit the data terrifically. These results seem to confirm one another from the previous question. Department B stands out as a consistently weak predictor. There seems to be alot of dependence among the variables.