S632 HW5

Erik Parker

March 21st, 2018

1. The dataset *melanoma* in the library *faraway* 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.

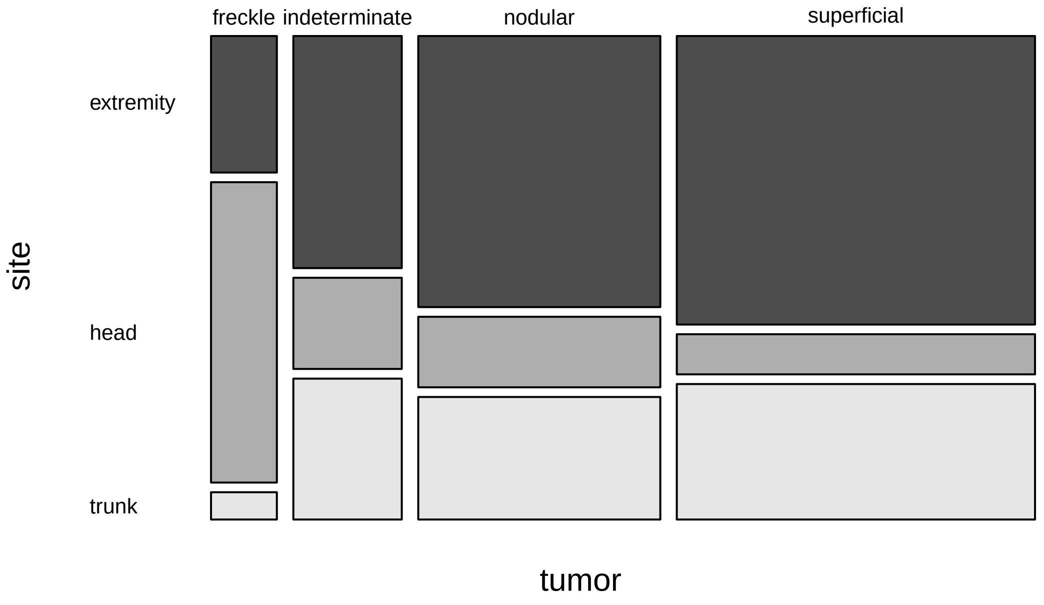
```r
rm(list = ls())

library(ggplot2)
library(faraway)
library(alr4)
library(dplyr)

cancer <- melanoma
```

a) Display the data in a contingency table and obtain a mosaic plot to visually check if *tumor* and *site* are independent. In addition, use a poisson model to determine if *tumor* and *site* are independent. Does your test agree with your visual conclusions?

```r
ct <- xtabs(count ~ tumor + site, cancer)
ct
```

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

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



```
m1 <- glm(count ~ tumor + site, cancer, family = poisson)

pchisq(deviance(m1), df.residual(m1), lower = FALSE)
```

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

> Based on the mosaic plot, it appears that *tumor* and *site* are not independent because the elements of the grid are not all proportional and symmetric. Furthermore, when a chi-square test based on a poisson model is performed to test for independence it results in very low p-value allowing rejection of the null hypothesis that the two variables are independent, and supporting the earlier visual interpretation.

b) Make a two-way table of the deviance residuals from your model in a). Are there any larger residuals? Comment.

```
round(xtabs(residuals(m1) ~ tumor + site, cancer), 3)
```

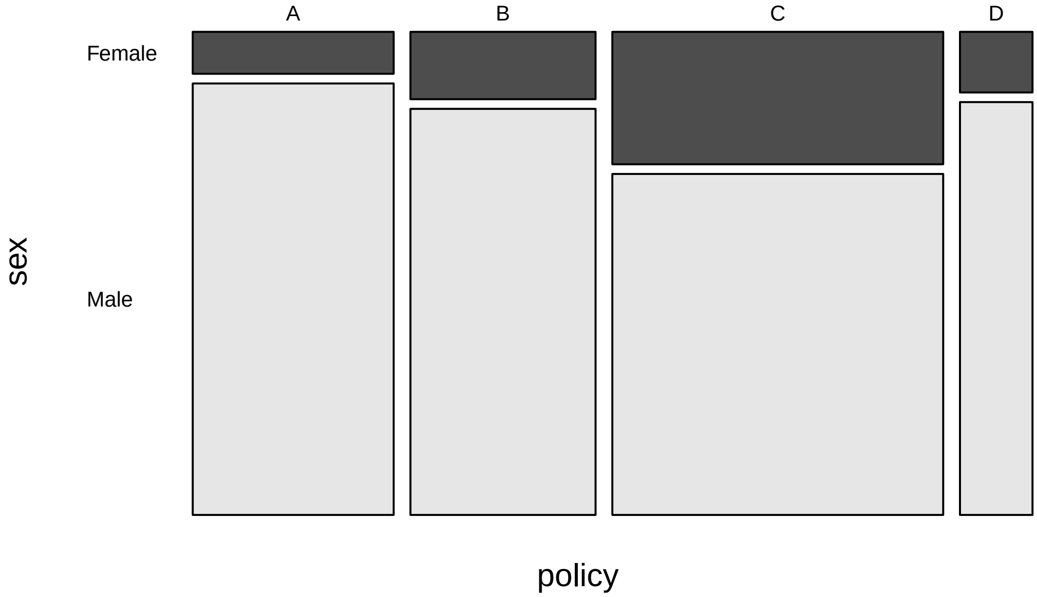
```
##                  site
## tumor             extremity   head  trunk
##   freckle            -2.316  5.135 -2.828
##   indeterminate      -0.660  0.468  0.548
##   nodular             0.281 -0.497 -0.022
##   superficial         1.008 -3.045  0.699
```

> Based on this two-way table of deviance residuals, we see that there are a number of larger residuals. All entries in the *freckle* row have quite large residuals, as do the *superficial* tumors found on the *head* and *extremity* regions. These factor levels with large residuals correspond to boxes in the mosaic plot from part a) that appeared quite different from the others - so I suppose it makes sense that they would be found to have relatively large deviance residuals.

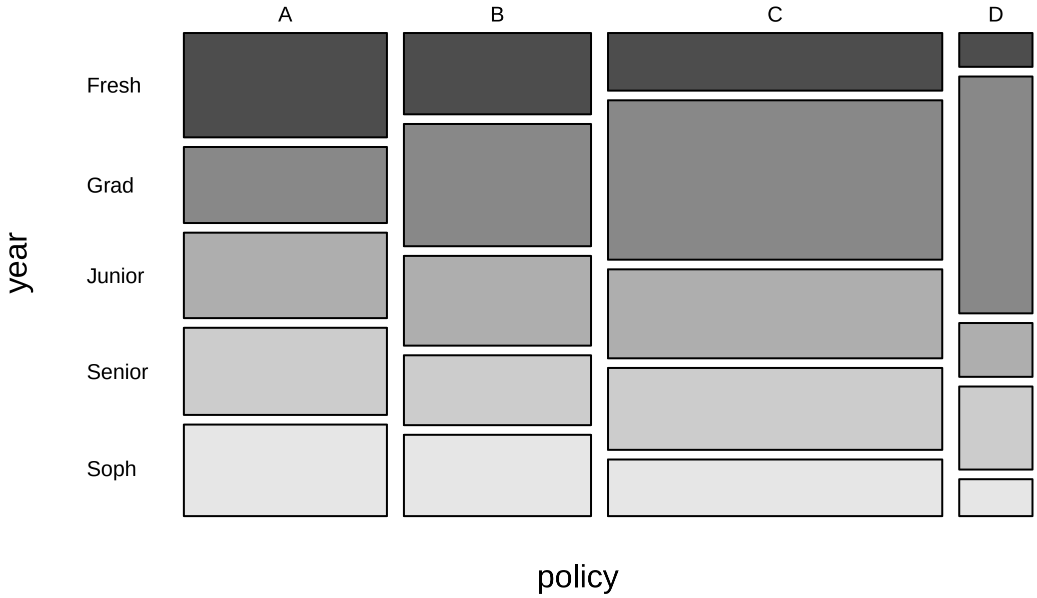
2. A student newspaper conducted a survey of student opinions about the Vietnam War in May 1967. Responses were classified by sex, year in the program and one of four opinions. The survey was voluntary. The data may be found in the dataset *uncviet* in the library *faraway*.

a) Conduct an analysis of the patterns of dependence in the data assuming that all variables are nominal.

```
war <- uncviet

ct1 <- xtabs(y ~ policy + sex, war)

ct2 <- xtabs(y ~ policy + year, war)

mosaicplot(ct1, color = TRUE, main = NULL, las = 1)
```



```r
mosaicplot(ct2, color = TRUE, main = NULL, las = 1)
```



```r
m2 <- glm(y ~ policy + sex, family = poisson, war)
pchisq(deviance(m2), df.residual(m2), lower = FALSE)
```

```
## [1] 2.351976e-112
```

```r
m3 <- glm(y ~ policy + year, family = poisson, war)
pchisq(deviance(m3), df.residual(m3), lower = FALSE)
```

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

From the mosaic plots, and the poisson tests for independence, we see clearly that the variables *sex* and *year* are not indepdendent from *policy*, when all variables are treated as nominal.

b) Assign scores to *year* and *policy* and fit an appropriate model. Interpret the trends in opinion over the years. Check the sensitivity of your conclusions to the assignment of the

scores (by trying other sensible alternatives).

```
war$yearo <- rep(c(1, 2, 3, 4, 5), each = 4)
war$policyo <- unclass(war$policy)


m1 <- glm(y ~ year + policy + sex, family = poisson, war)
Anova(m1)
```

```
## Analysis of Deviance Table (Type II tests)
## 
## Response: y
##        LR Chisq Df Pr(>Chisq)    
## year     216.56  4  < 2.2e-16 ***
## policy   718.59  3  < 2.2e-16 ***
## sex     1349.11  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m2 <- glm(y ~ year + policy + sex + I(yearo * policyo), family = poisson, war)

anova(m1, m2, test = "Chi")
```

```
## Analysis of Deviance Table
## 
## Model 1: y ~ year + policy + sex
## Model 2: y ~ year + policy + sex + I(yearo * policyo)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
## 1        31     423.83                          
## 2        30     246.13  1   177.69 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# So there is evidence that there is an association between year and policy
# when they are treated as ordinal variables.

summary(m2)$coef["I(yearo * policyo)", ]
```

```
##     Estimate   Std. Error      z value     Pr(>|z|) 
## 1.751092e-01 1.354426e-02 1.292867e+01 3.101548e-38
```

```
# This coefficient is positive, indicating based on how the data is coded,
# that a higher year in college is associated with a greater probability of
# supporting less involvement in the Vietnam war.

# now to check sensitivity of the scores, I will assign different ones.

ayear <- c(1, 2, 3, 4, 8)
apolicy <- c(1, 5, 6, 10)
# Larger difference between senior and grad, as there can be way more than
# one year seperation between those two levels.  Also, larger differences
# between the policy standings, as A seems really different from B and B and
# C are pretty similar, and D is quite different from C.

m2a <- glm(y ~ year + policy + sex + I(ayear[yearo] * apolicy[policyo]), family = poisson,
    war)
```

```r
anova(m1, m2a, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ year + policy + sex
## Model 2: y ~ year + policy + sex + I(ayear[yearo] * apolicy[policyo])
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        31     423.83
## 2        30     239.48  1   184.35 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Same conclusion using different scores.
```

Based on the above data and code, we see that a higher year in college is associated with a greater probability of supporting reduced US involvement in the Vietnam war among UNC students in 1967. This conclusion is not dependent on a particular assignment of ordinal scores, as an alternative assignment strategy also leads to the same conclusion that there is an association between *year* and *policy* (both coding schemes tested resulted in a LR test p-value of 2.2e-16).

4. The *hsb* data from the library *faraway* was collected as a subset of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. The variables are gender; race; socioeconomic status (SES); school type; chosen high school program type; scores on reading, writing, math, science, and social studies. We want to determine which factors are related to the choice of the type of program academic, vocational or general that the students pursue in high school. The response is multinomial with three levels.

a) Make a table showing the proportion of males and females choosing the three different programs. Comment on the difference. Repeat this comparison but for *SES* rather than gender.

```r
school <- hsb

schoolg <- group_by(school, gender, prog) %>% summarise(count = n()) %>% group_by(gender) %>%
    mutate(total = sum(count), proportion = count/total)


schools <- group_by(school, ses, prog) %>% summarise(count = n()) %>% group_by(ses) %>%
    mutate(total = sum(count), proportion = count/total)

xtabs(proportion ~ gender + prog, schoolg)
```

```
##         prog
## gender     academic   general  vocation
##   female 0.5321101 0.2201835 0.2477064
##   male   0.5164835 0.2307692 0.2527473
```

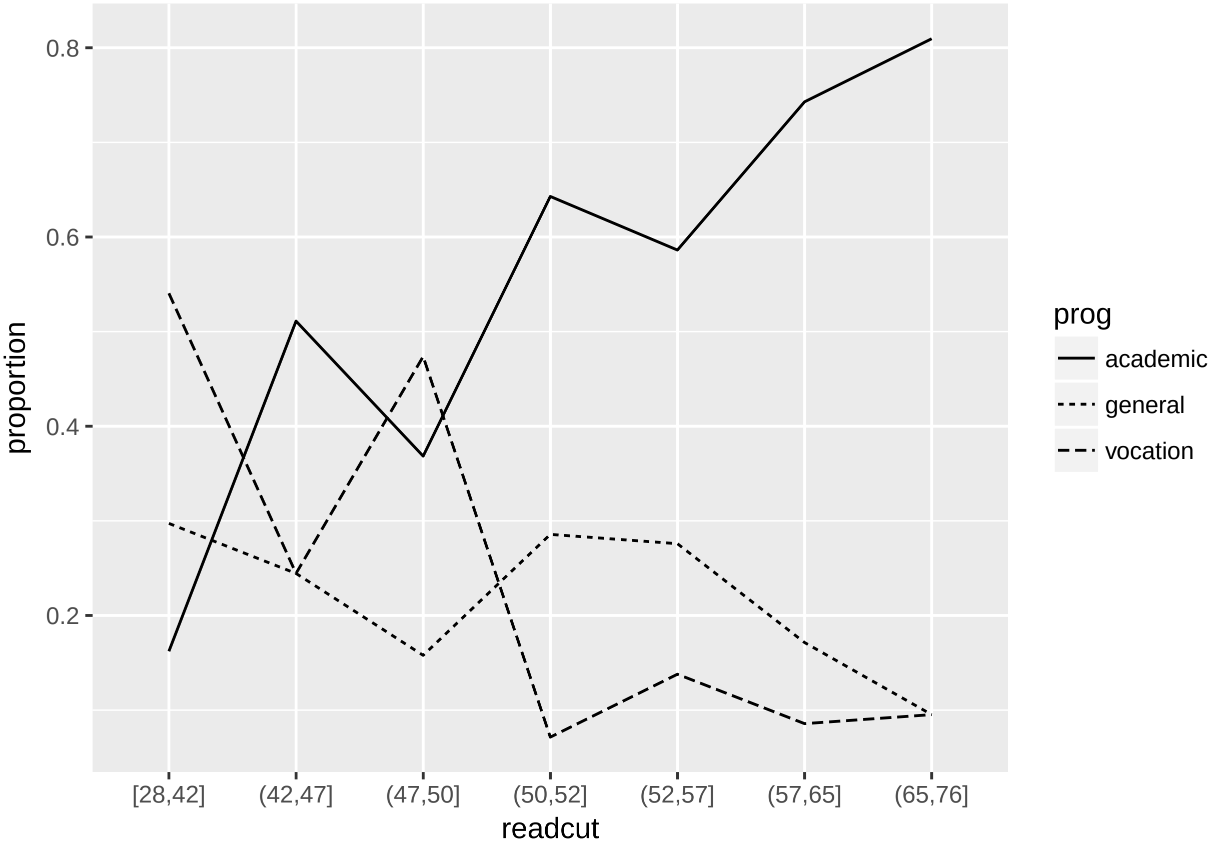
```r
xtabs(proportion ~ ses + prog, schools)
```

```
##         prog
## ses        academic   general  vocation
##   high   0.7241379 0.1551724 0.1206897
##   low    0.4042553 0.3404255 0.2553191
##   middle 0.4631579 0.2105263 0.3263158
```

There doesn't appear to be any large difference between the number of male and female students who choose different academic programs. On the other hand, there does seem to be a real difference between the academic program choices between different socioeconomic levels. Individuals from high ses backgrounds overwhelmingly choose to enroll in academic programs, while students from low and middle ses levels enroll more in general and vocational programs.

b) Construct a plot like the right panel of Figure 7.1 in ELM that shows the relationship between program choice and reading score. Comment on the plot.

```
schoolr <- mutate(school, readcut = cut_number(read, 7)) %>% group_by(readcut,
    prog) %>% summarise(count = n()) %>% group_by(readcut) %>% mutate(total = sum(count),
    proportion = count/total)

ggplot(schoolr, aes(x = readcut, y = proportion, group = prog, linetype = prog)) +
    geom_line()
```



The above plot shows that in general, as reading scores increase, the proportion of students choosing to enroll in academic programs also increases. Students with lower reading scores generally enroll in either general or vocational programs, though there is some variation to this pattern.

c) Fit a multinomial response model for the program choice and examine the fitted coefficients. Interpret at least two coefficients. In addition, observe that of the five subjects, one gives unexpected coefficients. Why do you think this happens?

```
library(nnet)
```

```r
mmod <- multinom(prog ~ gender + race + ses + schtyp + read + write + math +
    science + socst, data = school)
```

```
## # weights:  42 (26 variable)
## initial  value 219.722458
## iter  10 value 171.814970
## iter  20 value 153.793692
## iter  30 value 152.935260
## final  value 152.935256
## converged
```

```r
summary(mmod)
```

```
## Call:
## multinom(formula = prog ~ gender + race + ses + schtyp + read +
##     write + math + science + socst, data = school)
##
## Coefficients:
##          (Intercept)  gendermale raceasian racehispanic racewhite
## general     3.631901 -0.09264717  1.352739   -0.6322019 0.2965156
## vocation    7.481381 -0.32104341 -0.700070   -0.1993556 0.3358881
##              seslow sesmiddle schtyppublic        read       write
## general  1.09864111 0.7029621    0.5845405 -0.04418353 -0.03627381
## vocation 0.04747323 1.1815808    2.0553336 -0.03481202 -0.03166001
##                math    science       socst
## general  -0.1092888 0.10193746 -0.01976995
## vocation -0.1139877 0.05229938 -0.08040129
##
## Std. Errors:
##          (Intercept) gendermale raceasian racehispanic racewhite    seslow
## general     1.823452  0.4548778  1.058754    0.8935504 0.7354829 0.6066763
## vocation    2.104698  0.5021132  1.470176    0.8393676 0.7480573 0.7045772
##          sesmiddle schtyppublic       read      write       math
## general  0.5045938    0.5642925 0.03103707 0.03381324 0.03522441
## vocation 0.5700833    0.8348229 0.03422409 0.03585729 0.03885131
##             science      socst
## general  0.03274038 0.02712589
## vocation 0.03424763 0.02938212
##
## Residual Deviance: 305.8705
## AIC: 357.8705
```

In the above model, I will interpret the coefficients for the program levels *general* and *vocation* for the regressor *gender*. First, the log-odds of moving from the baseline program category of *academic* to the category *general* decrease by 0.093, or $(1 - e^{-0.093}) \times 100 = 8.85\%$, when we move from considering female, to male students. Similarly, the log-odds of moving from the category *academic* to *vocation* decrease by 0.321, or $(1 - e^{-0.321}) \times 100 = 27.5\%$, when we move from female to male students.

Furthermore, we can see from the summary above that of the five subject tests, *science* gives coefficients that are in the opposite direction of the others. Where the other four subjects have negative coefficients, saying that increasing scores on those tests results in lower odds of attending either a general or vocational school compared to an academic one, higher scores on the science test lead to *increased* odds of attending either general or vocational schools. One possible reason for this observation could be that students with parents who work in vocational, or otherwise "non-academic" jobs (like plumbers, carpenters, electricians, etc.) may have more exposure to real-life scientific concepts found in physics and chemistry and so find that those concepts are easier to understand, leading to them doing better on that subject test. These same students may then choose to enter a general or vocational high school, not because of their test scores, but because of their family history in those situations. Alternatively, maybe the science subject test was the last test taken, and the high-achieving students, who would later go on to enter academic programs, were tired from their high effort on the other tests and so did comparitively worse on that one.