HW 6

Ted Henson

3/4/2020

#Chapter 6, questions 2 and 4, pages 126-7. Omit question 2(e).

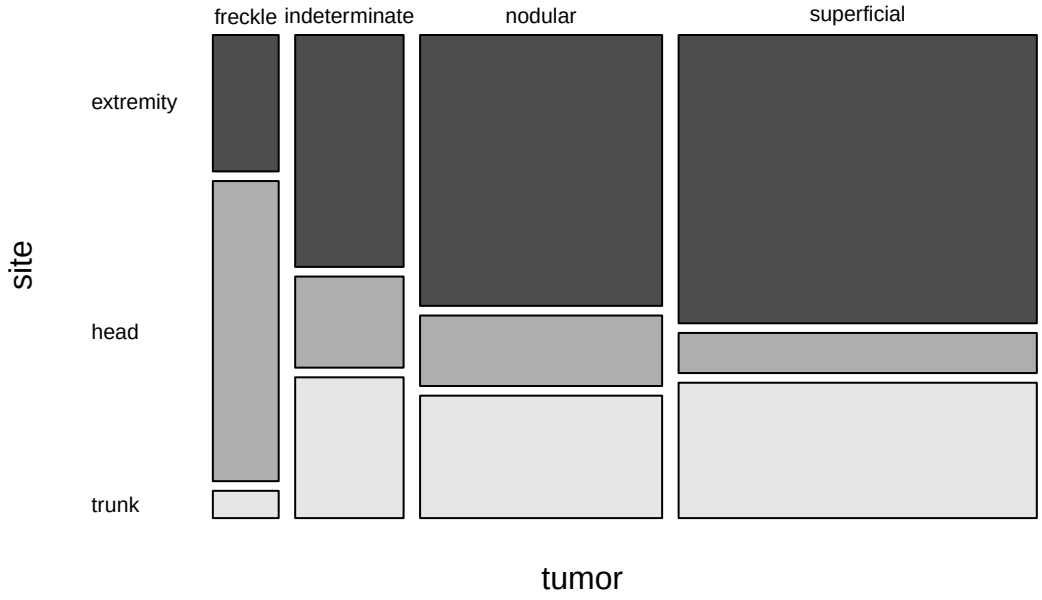
Question 2

a)

```
library(faraway)
data(melanoma)
cross.table=xtabs(count~tumor+site, data = melanoma)
cross.table
```

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

```
mosaicplot(cross.table,color=T,main=NULL,las=1)
```



The type of cancer and location of the cancer do not appear to be independent. As an example, superficial tumors are much more likely to be found in the extremities.

b)

```
summary(cross.table)
```

```
## 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
```

As shown by the chisq test statistic and corresponding p value, tumor and site are almost certainly not independent.

c)

```
mod1 = glm(count ~ site + tumor, data = melanoma, family = 'poisson')
summary(mod1)
```

```
##
## Call:
## glm(formula = count ~ site + tumor, family = "poisson", data = melanoma)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -3.0453  -1.0741   0.1297   0.5857   5.1354
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.9554     0.1770  16.696  < 2e-16 ***
## sitehead             -1.2010     0.1383  -8.683  < 2e-16 ***
## sitetrunk            -0.7571     0.1177  -6.431 1.27e-10 ***
## tumorindeterminate    0.4990     0.2174   2.295   0.0217 *
## tumornodular          1.3020     0.1934   6.731 1.68e-11 ***
## tumorsuperficial      1.6940     0.1866   9.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 295.203  on 11  degrees of freedom
## Residual deviance:  51.795  on  6  degrees of freedom
## AIC: 122.91
##
## Number of Fisher Scoring iterations: 5
```

```
mod = glm(count ~ site*tumor, data = melanoma, family = 'poisson')
sumary(mod)
```

```
##                               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                    2.30259    0.31623  7.2814 3.303e-13
## sitehead                       0.78846    0.38139  2.0674 0.0387009
## sitetrunk                     -1.60944    0.77460 -2.0778 0.0377300
## tumorindeterminate             1.02962    0.36839  2.7949 0.0051918
## tumornodular                   1.98787    0.33719  5.8954 3.738e-09
## tumorsuperficial               2.44235    0.32969  7.4080 1.282e-13
## sitehead:tumorindeterminate   -1.72277    0.52161 -3.3028 0.0009573
## sitetrunk:tumorindeterminate   1.11045    0.83339  1.3324 0.1827134
## sitehead:tumornodular         -2.13448    0.46020 -4.6381 3.516e-06
## sitetrunk:tumornodular         0.81549    0.80250  1.0162 0.3095409
## sitehead:tumorsuperficial     -2.76080    0.46546 -5.9314 3.004e-09
## sitetrunk:tumorsuperficial     0.85349    0.79197  1.0777 0.2811758
##
## n = 12 p = 12
## Deviance = 0.00000 Null Deviance = 295.20301 (Difference = 295.20301)
```

```
drop1(mod,test='Chi')
```

```
## Single term deletions
##
## Model:
## count ~ site * tumor
##            Df Deviance     AIC    LRT Pr(>Chi)
## <none>           0.000  83.111
## site:tumor  6   51.795 122.906 51.795 2.05e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown by the original poisson model, full model with interactions and dropped chi sq test, the tumor and site are not independent as shown by the significance of the interaction terms.

d)

```
round(xtabs(residuals(mod1,
                      type = 'deviance')~tumor+site,melanoma), 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
```

```
round(xtabs(residuals(mod,
                      type = 'deviance')~tumor+site,melanoma), 3)
```

```
##                site
## tumor           extremity head trunk
##   freckle               0    0     0
##   indeterminate         0    0     0
##   nodular               0    0     0
##   superficial           0    0     0
```

The largest residuals were for freckle tumors located in the head site. Head tumors were very uncommon overall, with most of them occuring with freckles, so this is not too suprising given the model generating these residuals did not consider interaction terms.

f)

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data.2 = melanoma %>% dplyr::filter(site != 'head')
data.2$site  = factor(data.2$site,
                      levels = c('extremity',
                                 'trunk'))
cross.table=xtabs(count~tumor+site, data = data.2)
summary(cross.table)
```

```
## Call: xtabs(formula = count ~ tumor + site, data = data.2)
## Number of cases in table: 332
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 2.0254, df = 3, p-value = 0.5671
##  Chi-squared approximation may be incorrect
```

As shown by the above chisq test statistic and corresponding p value, these factors are probably independent with the removal of the head observations.

Question 4

a)

```
ct.v=xtabs(y~penalty + victim,death)
ct.v
```

```
##         victim
## penalty   b   w
##     no  106 184
##     yes   6  30
```

```
ct.d=xtabs(y~penalty + defend,death)
ct.d
```

```
##         defend
## penalty   b   w
##     no  149 141
##     yes  17  19
```

```
ct.c=xtabs(y~ defend + victim + penalty, data = death)
ct.c
```

```
## , , penalty = no
##
##       victim
## defend   b   w
##      b  97  52
##      w   9 132
##
## , , penalty = yes
##
##       victim
## defend   b   w
##      b   6  11
##      w   0  19
```

```
summary(ct.c)
```

```
## Call: xtabs(formula = y ~ defend + victim + penalty, data = death)
## Number of cases in table: 326
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 122.4, df = 4, p-value = 1.642e-25
```

Yes this is an example of Simpson's paradox because when looking at the marginal frequencies, it appears that the defendant is much more likely to be charged with the death penalty when the victim is white as opposed to black. It also appears that black defendants are more likely to be charged with the death penalty than white defendants. The effect is more complex when considering all pairwise interactions. Black defendents are more likely to be charged with the death penalty when the victim is white. So the effect that black defendents are more likely to be charged still applies so this is not an example of Simpson's paradox. These differences are statistically significant according to the chisq statistic and corresponding p value.

b)

The most appropriate model should be a poisson model

```r
mod = glm(y ~ penalty + victim*defend, death, family = poisson)
summary(mod)
```

```
##
## Call:
## glm(formula = y ~ penalty + victim * defend, family = poisson,
##     data = death)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
##  0.5569  -0.2012  -1.4099   0.3443   1.4118  -0.5467  -1.7531   0.5561
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       4.5177     0.1004  44.976  < 2e-16 ***
## penaltyyes       -2.0864     0.1767 -11.807  < 2e-16 ***
## victimw          -0.4916     0.1599  -3.074  0.00212 **
## defendw          -2.4375     0.3476  -7.013 2.34e-12 ***
## victimw:defendw   3.3116     0.3786   8.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 395.9153  on 7  degrees of freedom
## Residual deviance:   8.1316  on 3  degrees of freedom
## AIC: 53.813
##
## Number of Fisher Scoring iterations: 4
```

c)

```r
death = death %>% group_by(victim, defend) %>% mutate(total = sum(y)) %>% group_by(victim, defend) %>% r


binom = glm(prop ~ victim + defend + victim:defend, data = death,
            family = binomial("logit"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
summary(binom)
```

```
##
## Call:
## glm(formula = prop ~ victim + defend + victim:defend, family = binomial("logit"),
##     data = death)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -0.7934  -1.1774  -0.6783  -0.9706   0.7934   1.1774   0.6783   0.9706
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.748e-16  1.414e+00       0        1
```

```
## victimw         -5.495e-16  2.000e+00       0        1
## defendw         -1.021e-15  2.000e+00       0        1
## victimw:defendw  1.099e-15  2.828e+00       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.8359  on 7  degrees of freedom
## Residual deviance: 6.8359  on 4  degrees of freedom
## AIC: 19.09
##
## Number of Fisher Scoring iterations: 2
```

The above created a binomial model with the proportion of death penalties assigned as the response.