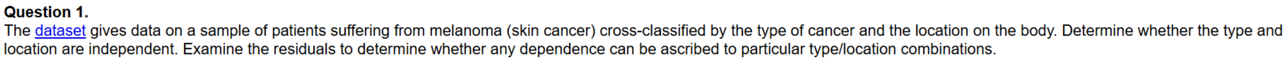
Discrete analysis_HW5

ID : 111024517 Name : 鄭家豪

Problem 1



Question 1.
The dataset 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. Determine whether the type and location are independent. Examine the residuals to determine whether any dependence can be ascribed to particular type/location combinations.

Sol.

```
data = read.table("melanoma.txt",header = T)
ct = xtabs(count ~ tumor + site, data=data)
ct
```

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

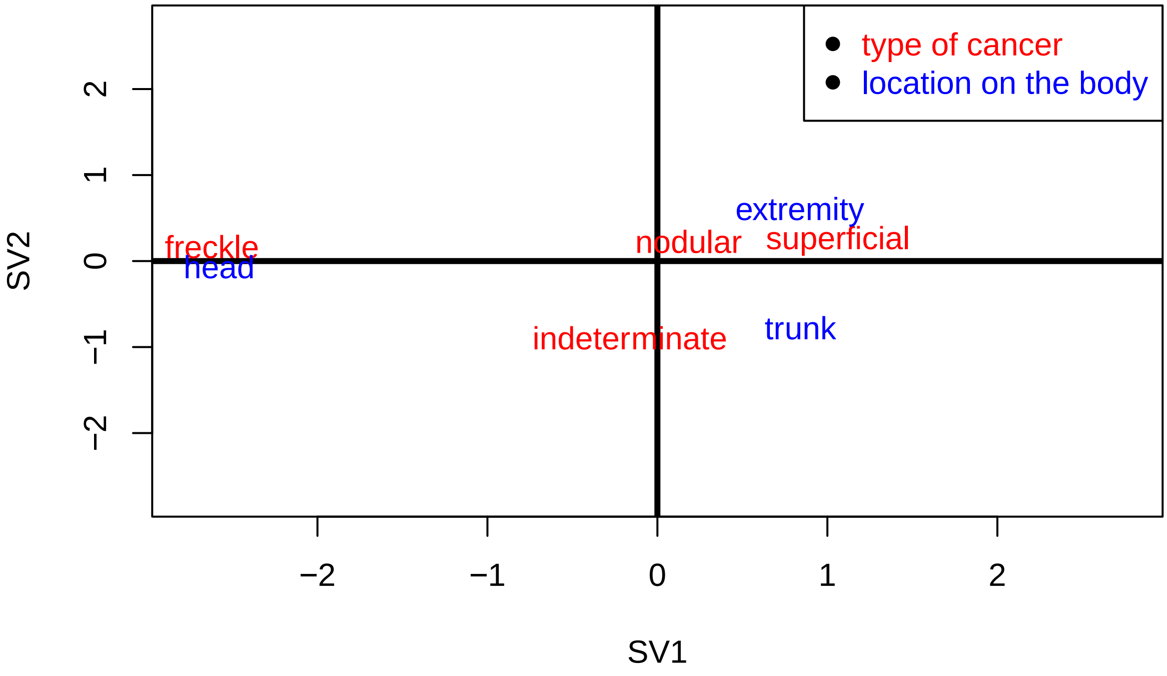
Directly do Chi-square test for testing H_0 : "tumor" and "site" are independent.

```
chi_test = summary(ct)
chi_test
```

```
Call: xtabs(formula = count ~ tumor + site, data = data)
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
```

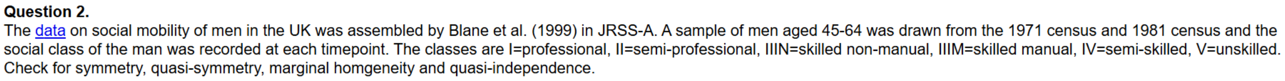
This indicates that we reject H_0 at level 0.05.
Next, the correspondence analysis is used to examine the residuals to find where the dependence is coming from.

```
fit = glm(count ~ tumor + site, family="poisson",data=data)
residual = residuals(fit, type="pearson")
rct = xtabs(residual ~ tumor + site, data=data)
svd_rct = svd(rct, 2, 2)
left_svd = svd_rct$u %*% diag(sqrt(svd_rct$d[1:2]))
right_svd = svd_rct$v %*% diag(sqrt(svd_rct$d[1:2]))
scale = 1.05 * max(abs(left_svd), abs(right_svd))
plot(rbind(left_svd, right_svd), xlim= c(-scale,scale), ylim=c(-scale,scale),
     xlab = "SV1", ylab="SV2", type="n")
abline(h=0,v=0, lwd=3)
text(left_svd, dimnames(rct)[[1]], col="red") ; text(right_svd, dimnames(rct)[[2]], col="blue")
legend("topright", legend=c("type of cancer", "location on the body"),
       text.col=c("red", "blue"),pch=16)
```



From the above, the distribution of tumor within the subgroup "head" is not typical, that is, "tumor:freckle" and "site:head" have strong association. Thus, the dependence comes from "freckle-head".

Problem 2



Question 2.
The data on social mobility of men in the UK was assembled by Blane et al. (1999) in JRSS-A. A sample of men aged 45-64 was drawn from the 1971 census and 1981 census and the social class of the man was recorded at each timepoint. The classes are I=professional, II=semi-professional, IIIN=skilled non-manual, IIIM=skilled manual, IV=semi-skilled, V=unskilled. Check for symmetry, quasi-symmetry, marginal homgeneity and quasi-independence.

Sol.

- Test for symmetry:

```
data = read.table("cmob.txt")
ct = xtabs(y ~ class71+class81, data=data)
ct
```

```
       class81
class71     I    II  IIIM  IIIN    IV     V
   I     1759   553   130   141    22     2
   II     541  6901   824   861   367    60
   IIIM   293  1409 12054   527  1678   586
   IIIN   248  1238   346  2562   308    56
   IV     132   419  1779   461  3565   461
   V       37    53   582    88   569   813
```

```
symfac <- factor(apply(data[,2:3],1,function(x) paste(sort(x),collapse="-")))
fit_sym <- glm(y ~ symfac, family="poisson", data=data)
pchisq(deviance(fit_sym), df.residual(fit_sym), lower = F)
```

```
[1] 9.053713e-105
```

It is statistically significant that social mobility in 1971 and 1981 are not symmetric.

- Test for quasi-symmetry:

```
fit_qsym <- glm(y ~ class71+class81+symfac, family="poisson", data=data)
pchisq(deviance(fit_qsym), df.residual(fit_qsym), lower = F)
```

```
[1] 2.167122e-22
```

It is statistically significant that social mobility in 1971 and 1981 are not quasi-symmetric.

- Test for marginal homgeneity:

By the result of testing quasi-symmetry, we can not test marginal homogeneity.

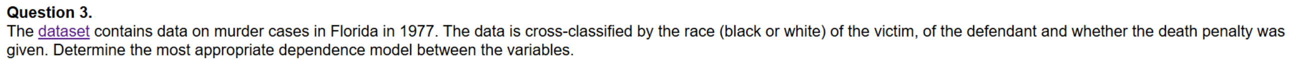
- Test for quasi-independence:

```
fit_qind <- glm(y ~ class71+class81+symfac, family="poisson", data=data,
                subset = - which(ct %in% diag(ct)))
pchisq(deviance(fit_qind), df.residual(fit_qind), lower = F)
```

```
[1] 2.167122e-22
```

It is statistically significant that social mobility in 1971 and 1981 are not quasi-independence.

Problem 3



Question 3.
The dataset contains data on murder cases in Florida in 1977. The data is cross-classified by the race (black or white) of the victim, of the defendant and whether the death penalty was given. Determine the most appropriate dependence model between the variables.

Sol.

```
data = read.table("death.txt",header=T)
ct3way = ftable(xtabs(y ~ penalty + victim + defend, data=data))
ct3way
```

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

Since the response variable is a count variable with no apparent upper limit, we fit a Poisson regression model. Starting with a saturated model that includes three variables: penalty, defend, victim, and interaction terms. Then, we iteratively remove the insignificant variable using deviance tests until all remaining variables are significant.

```r
fit_sat =  glm(y ~ penalty*defend*victim, family = "poisson",  data=data)
drop1(fit_sat, test = "Chi")
```

```
Single term deletions

Model:
y ~ penalty * defend * victim
                      Df Deviance    AIC     LRT Pr(>Chi)
<none>                     0.00000 51.682
penalty:defend:victim  1   0.70074 50.382 0.70074   0.4025
```

After removing the effect “penalty:defend:victim”, it is uniform association model.

```r
fit_1 = update(fit_sat, .~. - penalty:defend:victim)
drop1(fit_1, test = "Chi")
```

```
Single term deletions

Model:
y ~ penalty + defend + victim + penalty:defend + penalty:victim +
    defend:victim
               Df Deviance     AIC     LRT  Pr(>Chi)
<none>               0.701  50.382
penalty:defend  1    1.882  49.563   1.181  0.277121
penalty:victim  1    7.910  55.592   7.209  0.007252 **
defend:victim   1  131.458 179.140 130.757 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After removing the effect “penalty:defend”, it is conditional independence model.

```r
fit_2 = update(fit_1, .~. - penalty:defend)
drop1(fit_2, test = "Chi")
```

```
Single term deletions

Model:
y ~ penalty + defend + victim + penalty:victim + defend:victim
               Df Deviance     AIC    LRT Pr(>Chi)
<none>               1.882  49.563
penalty:victim  1    8.132  53.813   6.25  0.01242 *
defend:victim   1  131.680 177.361 129.80  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All variables of conditional independence model are significant, do goodness-of-fit via p-value.

```r
pchisq(fit_2$deviance, fit_2$df.residual, lower = F)
```

```
[1] 0.3902578
```

At level 0.05, we do not reject conditional independence model, which means that penalty and defend are independent conditioned on victim.