Discrete Analysis Homework 5

110024516 邱繼賢

Problem 1.

```
melanoma = read.table("melanoma.txt")
ct.1 = xtabs(count ~ tumor + site, melanoma)
ct.1
```

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

To test the whether the type and location are independent by Pearson's X^2 test :

$$X^2 = \sum_{ij} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \overset{a}{\sim} \chi^2_6$$

where

$$\hat{\mu}_{ij} = \frac{Y_{i+} Y_{+j}}{Y_{++}}$$

```
summary(ct.1)
```

```
## 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
```

Thus, the p-value $= P(\chi^2_6 > X^2) = P(\chi^2_6 > 65.81) = 2.943e-12 < 0.05$

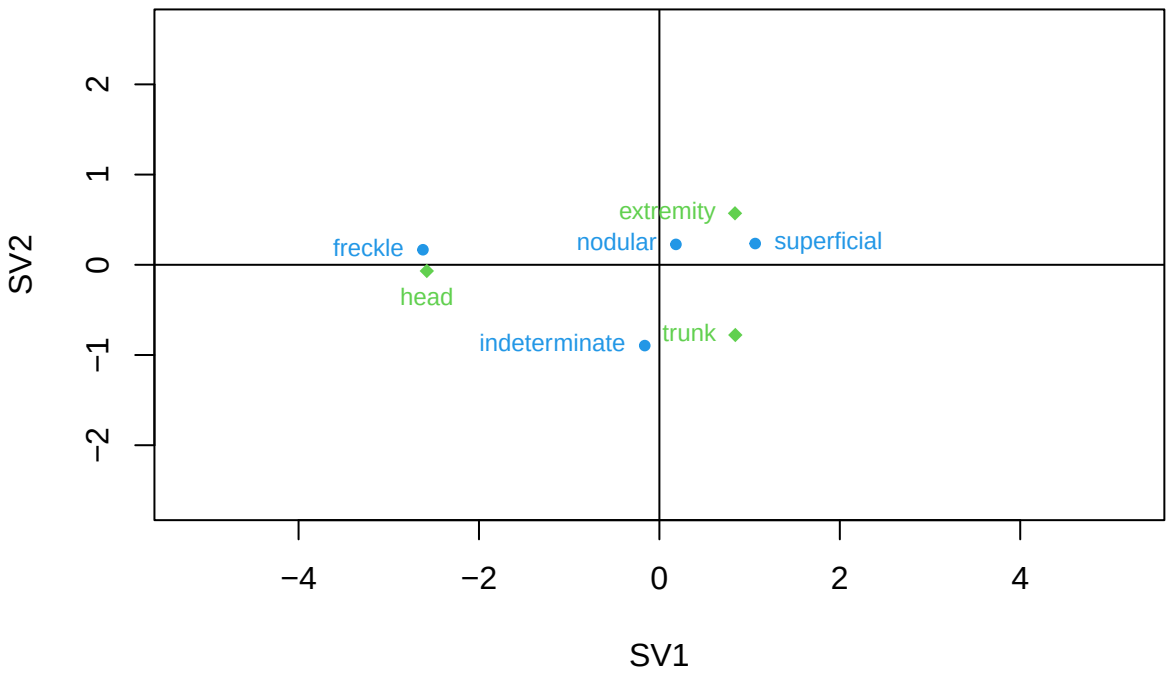
$\Rightarrow$ Reject that the type and location are independent.

And then fit the main effect only GLM

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$
$$\log(\mu_{ij}) = \eta_{ij} \sim \text{tumor} + \text{site}$$

Examine the residual of the model by Correspondence Analysis :

```
mod.1 = glm(count ~ tumor + site, family = poisson, melanoma)
z = xtabs(residuals(mod.1,type="pearson")~tumor+site, melanoma)
svdz = svd(z,2,2)
leftsv = svdz$u %*% diag(sqrt(svdz$d[1:2])) # U'
rightsv = svdz$v %*% diag(sqrt(svdz$d[1:2])) # V'
ll = max(abs(rightsv), abs(leftsv))
plot(rbind(leftsv,rightsv), asp = 1, xlim = c(-ll,ll), ylim = c(-ll,ll),
     xlab = "SV1", ylab = "SV2", type = "n")
abline(h=0,v=0)
points(leftsv, col = 4, pch = 20)
points(rightsv, col = 3, pch = 18)
text(leftsv[-4,], dimnames(z)$tumor[-4], col = 4, pos=2, cex = 0.75)
text(t(leftsv[4,]), dimnames(z)$tumor[4], col = 4, pos=4, cex = 0.75)
text(rightsv[-2,], dimnames(z)$site[-2], col = 3, pos=2, cex = 0.75)
text(t(rightsv[2,]), dimnames(z)$site[2], col = 3, pos=1, cex = 0.75)
```



先觀察單一變數 *tumor (site)* 的設定值，是否有特別遠離或是靠近原點的設定值：

(1) *tumor = freckle (site = head)* 特別遠離原點，代表這種設定值下的 conditional distribution 和 marginal distribution (也就是在 independence 成立下的分布)，相差很遠

(2) *tumor = nodular* 距離原點相對較近，代表這種設定值下的 conditional distribution 和 marginal distribution 相差不遠

再來觀察是否有兩變數的組合距離很近且同時遠離原點，或是距離很遠且落在原點的兩側：

(1) *(freckle,head)* , *(superficial,extremity)* , *(indeterminate,trunk)* 這三種組合的兩變數都是彼此距離較近且遠離原點，代表這三種組合的 residual 數值都較大且大於零，也就是這些組合發生的機率大於 independent asumption 下所估計出的機率

(2) *(freckle,extremity)* , *(freckle,trunk)* , *(superficial,head)* 這三種組合的兩變數都是彼此距離較遠且落在原點兩側，代表這三種組合的 residual 數值都較大且小於零，也就是這些組合發生的機率小於 independent asumption 下所估計出的機率

Problem 2.

```
cmob = read.table("cmob.txt")
ct.2 = xtabs(y ~ class71 + class81, cmob)
ct.2
```

```
##        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
```

(1) Check for symmetry :

Construct 21 levels symmetric factor

```
symfac = factor(apply(cmob[,2:3],1,function(x) paste(sort(x),collapse="-")))
matrix(symfac,6,6)
```

```
##      [,1]     [,2]      [,3]        [,4]        [,5]      [,6]
## [1,] "I-I"    "I-II"    "I-IIIN"    "I-IIIM"    "I-IV"    "I-V"
## [2,] "I-II"   "II-II"   "II-IIIN"   "II-IIIM"   "II-IV"   "II-V"
## [3,] "I-IIIN" "II-IIIN" "IIIN-IIIN" "IIIM-IIIN" "IIIN-IV" "IIIN-V"
## [4,] "I-IIIM" "II-IIIM" "IIIM-IIIN" "IIIM-IIIM" "IIIM-IV" "IIIM-V"
## [5,] "I-IV"   "II-IV"   "IIIN-IV"   "IIIM-IV"   "IV-IV"   "IV-V"
## [6,] "I-V"    "II-V"    "IIIN-V"    "IIIM-V"    "IV-V"    "V-V"
```

Fit GLM

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$\log(\mu_{ij}) = \eta_{ij} \sim \text{sym-factor}$$

```
mod_2.1 = glm(y ~ symfac, family = poisson, cmob)
summary(mod_2.1)
```

```
##
## Call:
## glm(formula = y ~ symfac, family = poisson, data = cmob)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -9.185  -2.078   0.000   1.951   8.408
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      7.47250    0.02384 313.400   <2e-16 ***
## symfacI-II      -1.16805    0.03850 -30.336   <2e-16 ***
## symfacI-IIIM    -2.11828    0.05415 -39.116   <2e-16 ***
## symfacI-IIIN    -2.20207    0.05603 -39.303   <2e-16 ***
## symfacI-IV      -3.12870    0.08404 -37.231   <2e-16 ***
## symfacI-V       -4.50209    0.16189 -27.809   <2e-16 ***
## symfacII-II      1.36692    0.02671  51.177   <2e-16 ***
## symfacII-IIIM   -0.45455    0.03188 -14.258   <2e-16 ***
## symfacII-IIIN   -0.51643    0.03233 -15.976   <2e-16 ***
## symfacII-IV     -1.49869    0.04290 -34.931   <2e-16 ***
## symfacII-V      -3.43826    0.09705 -35.429   <2e-16 ***
## symfacIIIM-IIIM  1.92465    0.02552  75.406   <2e-16 ***
## symfacIIIM-IIIN -1.39371    0.04140 -33.664   <2e-16 ***
## symfacIIIM-IV   -0.01749    0.02929  -0.597     0.55
## symfacIIIM-V    -1.10260    0.03774 -29.212   <2e-16 ***
## symfacIIIN-IIIN  0.37604    0.03096  12.144   <2e-16 ***
## symfacIIIN-IV   -1.52056    0.04323 -35.173   <2e-16 ***
## symfacIIIN-V    -3.19583    0.08668 -36.871   <2e-16 ***
## symfacIV-IV      0.70642    0.02914  24.244   <2e-16 ***
## symfacIV-V      -1.22833    0.03923 -31.307   <2e-16 ***
## symfacV-V       -0.77177    0.04241 -18.198   <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: 79869.78  on 35  degrees of freedom
## Residual deviance:   536.82  on 15  degrees of freedom
## AIC: 859.89
##
## Number of Fisher Scoring iterations: 5
```

and do the deviance-based goodness-of-fit test

```
pchisq(mod_2.1$deviance, mod_2.1$df.residual, lower.tail = F)
```

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

p-value $= P(\chi^2_{15} > D_S) = 9.053713e-105 < 0.05$
$\Rightarrow$ Symmetry is not hold.

(2) Check for quasi-symmetry :

Fit GLM

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$
$$\log(\mu_{ij}) = \eta_{ij} \sim \text{class71} + \text{class81} + \text{sym-factor}$$

```
mod_2.2 = glm(y ~ class71 + class81 + symfac, family = poisson, cmob)
summary(mod_2.2)
```

```
##
## Call:
## glm(formula = y ~ class71 + class81 + symfac, family = poisson,
##     data = cmob)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.495  -1.370   0.000   1.215   4.381
##
## Coefficients: (5 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      7.47250    0.02384 313.400  < 2e-16 ***
## class71II       -3.00884    0.09974 -30.168  < 2e-16 ***
## class71IIIM     -0.35206    0.04404  -7.993 1.31e-15 ***
```

```
## class71IIIN     -2.59938    0.09025 -28.803  < 2e-16 ***
## class71IV       -0.48153    0.04616 -10.431  < 2e-16 ***
## class71V         0.01856    0.03777   0.491    0.623
## class81II       -3.19218    0.09921 -32.177  < 2e-16 ***
## class81IIIM     -1.08290    0.04396 -24.636  < 2e-16 ***
## class81IIIN     -3.05207    0.08954 -34.088  < 2e-16 ***
## class81IV       -1.20518    0.04599 -26.206  < 2e-16 ***
## class81V        -0.79033    0.03777 -20.924  < 2e-16 ***
## symfacI-II       1.92826    0.10138  19.020  < 2e-16 ***
## symfacI-IIIM    -1.46613    0.06123 -23.945  < 2e-16 ***
## symfacI-IIIN     0.59826    0.10016   5.973 2.33e-09 ***
## symfacI-IV      -2.34942    0.08945 -26.264  < 2e-16 ***
## symfacI-V       -4.19585    0.16197 -25.905  < 2e-16 ***
## symfacII-II      7.56794    0.19235  39.345  < 2e-16 ***
## symfacII-IIIM    3.32643    0.10694  31.107  < 2e-16 ***
## symfacII-IIIN    5.40076    0.13280  40.670  < 2e-16 ***
## symfacII-IV      2.40912    0.11125  21.655  < 2e-16 ***
## symfacII-V            NA         NA      NA       NA
## symfacIIIM-IIIM  3.35960    0.06885  48.796  < 2e-16 ***
## symfacIIIM-IIIN  2.13985    0.10098  21.191  < 2e-16 ***
## symfacIIIM-IV    1.54333    0.05787  26.669  < 2e-16 ***
## symfacIIIM-V          NA         NA      NA       NA
## symfacIIIN-IIIN  6.02749    0.17170  35.104  < 2e-16 ***
## symfacIIIN-IV    2.13938    0.10231  20.911  < 2e-16 ***
## symfacIIIN-V          NA         NA      NA       NA
## symfacIV-IV      2.39313    0.07347  32.574  < 2e-16 ***
## symfacIV-V            NA         NA      NA       NA
## symfacV-V             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 79869.78  on 35  degrees of freedom
## Residual deviance:   126.73  on 10  degrees of freedom
## AIC: 459.8
```

```
##
## Number of Fisher Scoring iterations: 5
```

and do the deviance-based goodness-of-fit test

```
pchisq(mod_2.2$deviance, mod_2.2$df.residual, low = F)
```

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

p-value $= P(\chi^2_{10} > D_S) = 2.167122e-22 < 0.05$
$\Rightarrow$ Quasi-Symmetry is not hold.

(3) Check for marginal homogeneity :

Because there is no log-linear model that directly corresponds to marginal homogeneity and quasi-symmetry is also not hold, we do not have an appropriate and simple test for marginal homogeneity. We just check the two marginal distribution.

```
margin.table(ct.2, 1)
```

```
## class71
##     I    II  IIIM  IIIN    IV     V
##  2607  9554 16547  4758  6817  2142
```

```
margin.table(ct.2, 2)
```

```
## class81
##     I    II  IIIM  IIIN    IV     V
##  3010 10573 15715  4640  6509  1978
```

可以看出從 1971 年到 1981 年，社會階層較高的男性 (I & II) 數量皆有所提升，而社會階層較低的男性 (IIIN , IV & V) 數量則有所減少，推測可能同樣這一群男性隨著這十年的時間推進，他們整體的社會階層有所上升，故 Marginal Homogeneity 有可能不成立。

(3) Check for quasi-independence :

Omit the diagonal data and fit GLM

$$Y'_{ij} \sim \text{Poisson}(\mu_{ij})$$
$$\log(\mu_{ij}) = \eta_{ij} \sim \text{class71} + \text{class81}$$

```r
mod_2.3 = glm(y ~ class71 + class81, subset = -c(1,8,15,22,29,36), family = poisson, cmob)
summary(mod_2.3)
```

```
##
## Call:
## glm(formula = y ~ class71 + class81, family = poisson, data = cmob,
##     subset = -c(1, 8, 15, 22, 29, 36))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -20.138  -10.203   -6.933    9.632   22.535
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.15618    0.04539  91.571   <2e-16 ***
## class71II    1.34260    0.03982  33.714   <2e-16 ***
## class71IIIM  1.94437    0.03805  51.098   <2e-16 ***
## class71IIIN  1.01562    0.04061  25.007   <2e-16 ***
## class71IV    1.49475    0.03888  38.446   <2e-16 ***
## class71V     0.44620    0.04405  10.129   <2e-16 ***
## class81II    1.23267    0.03307  37.280   <2e-16 ***
## class81IIIM  1.43846    0.03350  42.934   <2e-16 ***
## class81IIIN  0.60161    0.03598  16.723   <2e-16 ***
## class81IV    1.04995    0.03411  30.779   <2e-16 ***
## class81V    -0.04211    0.04082  -1.032    0.302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 12525.8  on 29  degrees of freedom
## Residual deviance:  4747.7  on 19  degrees of freedom
## AIC: 4991.3
##
## Number of Fisher Scoring iterations: 5
```

and do the deviance-based goodness-of-fit test

```
pchisq(mod_2.3$deviance, mod_2.3$df.residual, low = F)
```

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

p-value = $P(\chi^2_{19} > D_S) = 0 < 0.05$
$\Rightarrow$ Quasi-Independence is not hold.

Problem 3.

Take a look at the 3×3 contingency table

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

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

We have three 2-level factors in the data. Let's do backward model selection start from the most complex model, saturated model :

$$\begin{aligned} Y_{ijk} &\sim \text{Poisson}(\mu_{ijk}) \\ \log(\mu_{ijk}) = \eta_{ijk} &\sim \text{penalty} * \text{victim} * \text{defend} \end{aligned}$$

and then do the deviance-based test comparing to other simpler models

```
mod3_sat = glm(y ~ penalty*victim*defend, death, family = poisson)
drop1(mod3_sat, test = "Chi")
```

```
## Single term deletions
##
## Model:
## y ~ penalty * victim * defend
```

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

We can see that the 3-factor interaction effect of *penalty:victim:defend* is not significant in the saturated model, so we can reduce to the uniform association model

$$\begin{aligned} Y_{ijk} &\sim \text{Poisson}(\mu_{ijk}) \\ \log(\mu_{ijk}) = \eta_{ijk} &\sim (\text{penalty} + \text{victim} + \text{defend})^2 \end{aligned}$$

and then do the deviance-based test comparing to other simpler models

```
mod3_ua = glm(y ~ (penalty+victim+defend)^2, death, family = poisson)
drop1(mod3_ua, test = "Chi")
```

```
## Single term deletions
##
## Model:
## y ~ (penalty + victim + defend)^2
##                Df Deviance     AIC     LRT  Pr(>Chi)
## <none>               0.701  50.382
## penalty:victim  1    7.910  55.592   7.209  0.007252 **
## penalty:defend  1    1.882  49.563   1.181  0.277121
## victim:defend   1  131.458 179.140 130.757 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the 2-factor interaction effect *penalty:defend* is not significant in the uniform association model, so we can reduce to the conditional independence model

$$\begin{aligned} Y_{ijk} &\sim \text{Poisson}(\mu_{ijk}) \\ \log(\mu_{ijk}) = \eta_{ijk} &\sim \text{penalty} + \text{victim} + \text{defend} + \text{penalty} : \text{victim} + \text{victim} : \text{defend} \end{aligned}$$

and then do the deviance-based test comparing to other simpler models

```
mod3_ci = glm(y ~ penalty*victim + victim*defend, death, family = poisson)
drop1(mod3_ci, test = "Chi")
```

```
## Single term deletions
##
```

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

We can see that all the effects in the model are significant. Let's check the deviance-based goodness-of-fit test for the model.

```
pchisq(mod3_ci$deviance,mod3_ci$df.residual, low = F)
```

```
## [1] 0.3902578
```

$\Rightarrow$ p-value $=$ 0.3902578 $>$ 0.05, so the model fits well to the data.

This conditional independence model means that *penalty* and *defend* are independent for given *victim*. The result can be seen in the below conditional probability table.

```
round(prop.table(ftable(xtabs(y ~ victim + defend + penalty, death)),1),3)
```

```
##               penalty    no   yes
## victim defend
## b      b              0.942 0.058
##        w              1.000 0.000
## w      b              0.825 0.175
##        w              0.874 0.126
```