P8131 HW4

Shihui Zhu, sz3029

1. Summarize the data using appropriate tables of percentages to show the pairwise associations between the levels of satisfaction and 1) contact with other residents and 2) type of housing. Comment on patterns in the associations.

```
value = c(65, 130, 67, 34, 141, 130, 54, 76, 48, 47, 116, 105, 100, 111, 62, 100, 191, 104)
data1 <- tibble(
  Contact = c(rep("Low", 3), rep("High", 3), rep("Low", 3), rep("High", 3), rep("Low", 3), rep("High", 3
  Satisfaction = c(rep("Low", 6), rep("Medium", 6), rep("High", 6)),
  HouseType = c("Tower Block", "Apartment", "House", "Tower Block", "Apartment", "House",
                    "Tower Block", "Apartment", "House", "Tower Block", "Apartment", "House",
                    "Tower Block", "Apartment", "House", "Tower Block", "Apartment", "House")
)
data1 = data1[rep(seq_len(nrow(data1)), value),]
```

Produce Summary Table:

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1681
##
##
##                     | data1$HouseType
## data1$Satisfaction |   Apartment |       House | Tower Block |   Row Total |
## -------------------|-------------|-------------|-------------|-------------|
##               High |         302 |         166 |         200 |         668 |
##                    |       0.013 |       7.437 |      10.600 |             |
##                    |       0.452 |       0.249 |       0.299 |       0.397 |
##                    |       0.395 |       0.322 |       0.500 |             |
##                    |       0.180 |       0.099 |       0.119 |             |
## -------------------|-------------|-------------|-------------|-------------|
##                Low |         271 |         197 |          99 |         567 |
##                    |       0.652 |       3.027 |       9.563 |             |
```

```
##                     |       0.478 |       0.347 |       0.175 |       0.337 |
##                     |       0.354 |       0.382 |       0.247 |             |
##                     |       0.161 |       0.117 |       0.059 |             |
## --------------------|-------------|-------------|-------------|-------------|
##              Medium |         192 |         153 |         101 |         446 |
##                     |       0.593 |       1.892 |       0.248 |             |
##                     |       0.430 |       0.343 |       0.226 |       0.265 |
##                     |       0.251 |       0.297 |       0.253 |             |
##                     |       0.114 |       0.091 |       0.060 |             |
## --------------------|-------------|-------------|-------------|-------------|
##       Column Total |         765 |         516 |         400 |        1681 |
##                     |       0.455 |       0.307 |       0.238 |             |
## --------------------|-------------|-------------|-------------|-------------|
##
##


##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1681
##
##
##                     | data1$Contact
## data1$Satisfaction |      High |       Low | Row Total |
## --------------------|-----------|-----------|-----------|
##                High |       395 |       273 |       668 |
##                     |     0.278 |     0.377 |           |
##                     |     0.591 |     0.409 |     0.397 |
##                     |     0.408 |     0.383 |           |
##                     |     0.235 |     0.162 |           |
## --------------------|-----------|-----------|-----------|
##                 Low |       305 |       262 |       567 |
##                     |     1.416 |     1.923 |           |
##                     |     0.538 |     0.462 |     0.337 |
##                     |     0.315 |     0.367 |           |
##                     |     0.181 |     0.156 |           |
## --------------------|-----------|-----------|-----------|
##              Medium |       268 |       178 |       446 |
##                     |     0.486 |     0.660 |           |
##                     |     0.601 |     0.399 |     0.265 |
##                     |     0.277 |     0.250 |           |
##                     |     0.159 |     0.106 |           |
## --------------------|-----------|-----------|-----------|
##       Column Total |       968 |       713 |      1681 |
##                     |     0.576 |     0.424 |           |
```

```
## -------------------|-----------|-----------|-----------|
##
##
```

From the table we see that among the high, medium and low satisfaction category, the percentage of residents live in apartment is the highest among others, and low contact higher comparing with high contact. Among apartment residents, most of them have high satisfaction, and among house residents, most of them have low satisfaction, then among tower block residents, most of them have high satisfaction.

Then among the high, medium and low satisfaction category, the percentage of residents having high contact with other residents is higher comparing with low contact. Among the high contact category, most residents have high satisfaction, and among the low contact category, most of them also have high satisfaction.

2. Nominal logistic regression model

Use nominal logistic regression model for the associations between response variable, the levels of satisfaction, and the other two variables.

Obtain a model that summarizes the patterns in the data.

Construct a nominal logistic regression model

```
data1.mult <- multinom(cbind(Sat.Low, Sat.Medium, Sat.High) ~ HouseType + Contact, data = data1.sat)
```

```
## # weights:  15 (8 variable)
## initial  value 1846.767257
## iter  10 value 1803.046285
## final  value 1802.740161
## converged
```

```
summary(data1.mult)
```

```
## Call:
## multinom(formula = cbind(Sat.Low, Sat.Medium, Sat.High) ~ HouseType +
##     Contact, data = data1.sat)
##
## Coefficients:
##            (Intercept) HouseTypeHouse HouseTypeTower Block ContactLow
## Sat.Medium  -0.2180364     0.06967922            0.4067631 -0.2959832
## Sat.High     0.2474047    -0.30402275            0.6415948 -0.3282264
##
## Std. Errors:
##            (Intercept) HouseTypeHouse HouseTypeTower Block ContactLow
## Sat.Medium  0.10930968      0.1437749            0.1713009  0.1301046
## Sat.High    0.09783068      0.1351693            0.1500774  0.1181870
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

The multinomial model is $log(\frac{\pi_{medium}}{\pi_{low}}) = \beta_{01} + \beta_{11}(HouseType = House) + \beta_{21}(HouseType = TowerBlock) + \beta_{31}(Contact = Low)$, and $log(\frac{\pi_{high}}{\pi_{low}}) = \beta_{02} + \beta_{12}(HouseType = House) + \beta_{22}(HouseType = TowerBlock) + \beta_{32}(Contact = Low)$.

So our fitted multinomial model is:

$$log(\frac{\pi_{medium}}{\pi_{low}}) = -0.2180364 + 0.06967922x_1 + 0.4067631x_2 - 0.2959832x_3$$

$$log(\frac{\pi_{high}}{\pi_{low}}) = 0.2474047 - 0.30402275x_1 + 0.6415948x_2 - 0.3282264x_3$$

Odds Ratios

Odds ratios with 95% confidence intervals:

```
invfisher.mult <- vcov(data1.mult) # inverse of fisher information matrix
CI.logit.medium = coef(data1.mult)[1,] + kronecker(t(c(0,qnorm(0.025),-qnorm(0.025))),
                                                   t(t(sqrt(diag(invfisher.mult)[1:4]))))
CI.logit.high = coef(data1.mult)[2,] + kronecker(t(c(0,qnorm(0.025),-qnorm(0.025))),
                                                   t(t(sqrt(diag(invfisher.mult)[5:8]))))

out.pi_medium <- exp(CI.logit.medium[2:4,])
out.pi_high <- exp(CI.logit.high[2:4,])

colnames(out.pi_medium)=c('Estimate of Odds Ratio','95% CI lower','95% CI upper')
rownames(out.pi_medium) = c("House", "Tower Block", "Contact.Low")
colnames(out.pi_high)=c('Estimate of Odds Ratio','95% CI lower','95% CI upper')
rownames(out.pi_high) = c("House", "Tower Block", "Contact.Low")

out.pi_medium %>% knitr::kable(digits = 3, caption = "For OR of Meidum over Low Satisfaction")
```

Table 1: For OR of Meidum over Low Satisfaction

	Estimate of Odds Ratio	95% CI lower	95% CI upper
House	1.072	0.809	1.421
Tower Block	1.502	1.074	2.101
Contact.Low	0.744	0.576	0.960

```
out.pi_high %>% knitr::kable(digits = 3, caption = "For OR of High over Low Satisfaction")
```

Table 2: For OR of High over Low Satisfaction

	Estimate of Odds Ratio	95% CI lower	95% CI upper
House	0.738	0.566	0.962
Tower Block	1.900	1.415	2.549
Contact.Low	0.720	0.571	0.908

Association

To test the association between levels of satisfaction and contact with others, we perform chi-squared test

Test of Homogeneity:

H_0 : the proportions of low/medium/high satisfaction levels among contact levels are equal

H_1 : not all proportions are equal

```
#data.sc <- data1 %>%
  #filter(Contact == 'High') %>%
  #group_by(Satisfaction) %>%
  #summarize(n = n())
data.sc <- tibble(
  contact.low = c(262, 178, 273, 262+178+273),
  contact.high = c(305, 268, 395, 305+268+395),
) %>%
  t()
chisq.test(data.sc)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.sc
## X-squared = 5.1398, df = 3, p-value = 0.1618
```

The test gives p-value of $0.1618 > 0.05$. So we fail to reject the null hypothesis and conclude that there is no enough evidence showing that there is association between contact with others and satisfaction levels.

To test the association between levels of satisfaction and housing types, we perform chi-squared test

Test of Homogeneity:

H_0 : the proportions of low/medium/high satisfaction levels among housing type are equal

H_1 : not all proportions are equal

```
#data.sc <- data1 %>%
  #filter(HouseType == 'Tower Block') %>%
  #group_by(Satisfaction) %>%
  #summarize(n = n())
data.sh <- tibble(
  house = c(197, 153, 166, 197+153+166),
  apartment = c(271, 192, 302, 271+192+302),
  tower = c(99, 101, 200, 99+101+200)
) %>%
  t()
chisq.test(data.sh)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data.sh
## X-squared = 34.024, df = 6, p-value = 6.657e-06
```

The test gives p-value of approximately 0. So we reject the null hypothesis and conclude that there is association between housing type and satisfaction levels.

Goodness of fit and Interaction

Then we calculate chi-squared value to evaluate the goodness of fit of this model:

H_0 : The model is close to the full model, H_1 : not close to full model, significant level is 0.05

```r
# goodness of fit
pihat=predict(data1.mult,type='probs')
m=rowSums(data1.sat[,3:5])
res.pearson=(data1.sat[,3:5]-pihat*m)/sqrt(pihat*m);res.pearson # pearson residuals
```

```
##      Sat.Low  Sat.Medium   Sat.High
## 1  0.6462082  0.01458006 -0.4986448
## 2  0.3770510  0.08967620 -0.4648120
## 3 -1.0575683 -0.12653898  1.4047956
## 4 -0.8014220 -0.01559243  0.5248140
## 5 -0.3508834 -0.07196683  0.3670803
## 6  0.8402535  0.08670506 -0.9471979
```

```r
G.stat=sum(res.pearson^2) # Generalized Pearson Chisq Stat
G.stat
```

```
## [1] 6.932341
```

```r
pval=1-pchisq(G.stat,df=(6-4)*(3-1))
pval# fit is good
```

```
## [1] 0.1395072
```

```r
# deviance
D.stat = sum(2*data1.sat[,3:5]*log(data1.sat[,3:5]/(pihat*m)))
D.stat
```

```
## [1] 6.893028
```

The Generalized Pearson Chisq Statistics is 6.932341. The Deviance is 6.893028. The p-value is 0.1395072 > 0.05, so we do not reject the null hypothesis and the model fit is good. Since the model fit is good, there is no interaction of contact level by house type in our model.

3. Ordinal categories

```r
# Order dataset
data1.grouped$Satisfaction = factor(data1.grouped$Satisfaction, levels = c("Low", "Medium", "High"), or
data1.grouped$Contact = factor(data1.grouped$Contact, levels = c("Low", "High"), ordered=T)
data1.grouped$ApartmentType = as.factor(data1.grouped$HouseType)

data1.polr=polr(Satisfaction ~ HouseType + Contact, data = data1.grouped, weights = Value)
summary(data1.polr)
```

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = Satisfaction ~ HouseType + Contact, data = data1.grouped,
##     weights = Value)
##
## Coefficients:
##                         Value Std. Error t value
## HouseTypeHouse        -0.2353     0.1052  -2.236
## HouseTypeTower Block   0.5010     0.1168   4.291
## Contact.L              0.1785     0.0658   2.713
##
## Intercepts:
##             Value   Std. Error t value
## Low|Medium  -0.6226  0.0721    -8.6347
## Medium|High  0.4899  0.0714     6.8575
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

The model shows the following relationships:

Let X_1 be type House, X_2 be type Tower Block, X_3 be low contact.

Since the ordinal logistic regression model is parameterized as $logit(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - ... - \eta_p x_p$ where $\eta_i = -\beta_i$, so the log odds are $logit(P(Y \leq j|x_i = 1)) - logit(P(Y \leq j|x_i = 0)) = -\eta_1 = -\beta_i$

$$logit(P(Sat \leq low)) = log(\frac{\pi_{low}}{\pi_{medium} + \pi_{high}}) = -0.6226 - 0.2353x_1 + 0.5010x_2 + 0.1785x_3$$

$$logit(P(Sat \leq medium)) = log(\frac{\pi_{low} + \pi_{medium}}{\pi_{high}}) = 0.4899 - 0.2353x_1 + 0.5010x_2 + 0.1785x_3$$

Since $\beta_i = -\eta_i$, $exp(\beta_i) = \frac{1}{exp(\eta_i)} = \frac{P(Y>j|x_i=1)/P(Y\leq j|x_i=1)}{P(Y>j|x_i=0)/P(Y\leq j|x_i=0)}$.

So the ORs are:

```
# 95% CI for OR
exp(cbind(coef(data1.polr),confint(data1.polr)))
```

```
##                                   2.5 %    97.5 %
## HouseTypeHouse        0.7903395 0.6429196 0.9711892
## HouseTypeTower Block  1.6502987 1.3136017 2.0762957
## Contact.L             1.1954228 1.0509003 1.3602010
```

```
exp(-0.6226)
```

```
## [1] 0.5365476
```

```
exp(0.4899)
```

```
## [1] 1.632153
```

The odds ratio across the all $J - 1$ categories are the same. The interpretation for $j = 1$ is: when holding the contact level at constant, the odds of having **high satisfaction** is 0.790 times the odds of having **low or medium satisfaction** if the residents live in **house** comparing with residents live in **other types of**

housing, and the odds of having **high satisfaction** is 1.650 times the odds of having **low or medium satisfaction** if the resident lives in **tower block** comparing with residents in **other types of housing**.

Holding the housing type at constant, the odds of having **high satisfaction** is 1.195 times the odds of having **low or medium satisfaction** if the resident has **low contact** with others.

Also, when the resident **lives in apartment** and has **high contact** with other residents, the odds of him having **low satisfaction** is 0.5365476 times the odds of having **medium and high satisfaction**.

When the resident **lives in apartment** and **has high contact** with other residents, the odds of him having **low and medium satisfaction** is 1.632153 times the odds of him having **high satisfaction**.

Goodness of fit and discrepancy:

```
pihat=predict(data1.polr,data1.sat,type='p')
m=rowSums(data1.sat[,3:5])
res.pearson=(data1.sat[,3:5]-pihat*m)/sqrt(pihat*m) # pearson residuals

G=sum(res.pearson^2)
G
```

```
## [1] 11.64205
```

```
numsamp=(3-1)*6 # degree of freedom for grouped data
numparam=2+3 # total num of param
pval=1-pchisq(G ,df=numsamp-numparam)
pval # fits well
```

```
## [1] 0.112962
```

The p-value is $0.112962 > 0.05$, so the model fits the data well.

The pearson residual tells us where is the largest discrepancy:

```
res.pearson
```

```
##      Sat.Low Sat.Medium    Sat.High
## 1  0.7794163 -0.3696759 -0.31516502
## 2  0.9176717 -1.0671397 -0.01522921
## 3 -1.1408504  0.1397991  1.24412460
## 4 -0.9946605  0.4549798  0.33539244
## 5 -0.2370110 -0.4051905  0.53781037
## 6  0.2742957  1.3678375 -1.47778315
```

```
max(abs(res.pearson))
```

```
## [1] 1.477783
```

The largest discrepancy is when the satisfaction is high, and the resident lives in house, has high contact with other residents.