P8131 HW4

Yihan Feng

2/19/2021

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

a. association between levels of satisfaction and contact

```
sat.contact = housing.df %>%
  group_by(contact, satisfaction) %>%
  summarize(n = sum(count)) %>%
  group_by(contact) %>%
  mutate(n_total = sum(n),
         percentage = n * 100/ n_total) %>%
  dplyr::select(-n_total, -n) %>%
  spread(key = satisfaction, value = percentage) %>%
  mutate("sat_low" = low,
         "sat_high" = high,
         "sat_med" = medium) %>%
  dplyr::select(-low, -high, -medium) %>%
  knitr::kable()
sat.contact
```

contact	sat_low	sat_high	sat_med
high	31.50826	40.80579	27.68595
low	36.74614	38.28892	24.96494

From the table of percentage of satisfaction and contact with other residents, I observed that people who highly contact with other residents have higher proportion of high satisfaction. People who barely contact (low contact) with other residents tend to have either high or low satisfaction.

b. association between levels of satisfaction and type of housing

```
sat.type = housing.df %>%
  group_by(type, satisfaction) %>%
  summarize(n = sum(count)) %>%
  group_by(type) %>%
  mutate(n_total = sum(n),
         percentage = n * 100/ n_total) %>%
  dplyr::select(-n_total, -n) %>%
  spread(key = satisfaction, value = percentage) %>%
  mutate("sat_low" = low,
         "sat_high" = high,
         "sat_med" = medium) %>%
  dplyr::select(-low, -high, -medium) %>%
  knitr::kable()
sat.type
```

type	sat_low	sat_high	sat_med
apartment	35.42484	39.47712	25.09804
house	38.17829	32.17054	29.65116
tower	24.75000	50.00000	25.25000

From the table of percentage of satisfaction and types of housing, I observed that people who live in the tower have highest proportion of high satisfaction. People who live in house have highest proportion of low satisfaction.

2. 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. Describe your fndings (the pattern in the associations, odds ratios with 95% confidence intervals, goodness-of-fit).

```
housing.nom = housing.df %>%
  pivot_wider(
    names_from = satisfaction,
    names_prefix = "sat_",
    values_from = count
  )
```

```
nom.model = multinom(cbind(sat_low, sat_medium, sat_high) ~ type + contact,
                     data = housing.nom)
```

```
## # weights:  15 (8 variable)
## initial  value 1846.767257
## iter  10 value 1803.278543
## final  value 1802.740161
## converged
```

```
summary(nom.model)
```

```
## Call:
## multinom(formula = cbind(sat_low, sat_medium, sat_high) ~ type +
##     contact, data = housing.nom)
##
## Coefficients:
##            (Intercept) typeapartment  typehouse contacthigh
## sat_medium  -0.1072644    -0.4067537 -0.3370771   0.2959803
## sat_high     0.5607737    -0.6415967 -0.9456177   0.3282263
##
## Std. Errors:
##            (Intercept) typeapartment typehouse contacthigh
## sat_medium   0.1524077     0.1713011 0.1803577   0.1301046
## sat_high     0.1329301     0.1500773 0.1644850   0.1181870
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

```
pihat = predict(nom.model, type = 'probs')
m = rowSums(housing.nom[,3:5])
res.pearson = (housing.nom[,3:5] - pihat*m)/sqrt(pihat*m)

G.stat = sum(res.pearson^2)
pval = 1 - pchisq(G.stat,df = (6-4)*(3-1))
D.stat = sum(2*housing.nom[,3:5] * log(housing.nom[,3:5]/(m*pihat)))
```

- Odds ratio with confidence interval:

Take low contact and housing type = tower as the reference group, the odds of **medium** satisfaction versus low satisfaction when comparing the reference group and people who:
* live in apartment is 0.666, with 95% CI (0.476, 0.931).
* live in house is 0.714, with 95% CI (0.501, 1.016).
* highly contact with other residents is 1.344, with 95% CI (1.042, 1.735).

Take low contact and housing type = tower as the reference group, the odds of **high** satisfaction versus low satisfaction when comparing the reference group and people who:
* live in apartment is 0.526, with 95% CI (0.392, 0.707).
* live in house is 0.388, with 95% CI (0.281, 0.536).
* highly contact with other residents is 1.388, with 95% CI (1.102, 1.751).

- Goodness of Fit:

The deviance of this model is 6.8930278. The deviance statistics should approximately follow chi-squared distribution with degree of freedom 4, with the critical value 9.487729. Therefore, it fails to reject the null hypothesis at 0.05 significance level, and we are able to conclude that the model fits the data well.

3. As the response has ordinal categories, fit proportional odds model to the data that include the same variables as used in the nominal logistic model obtained in (ii). What does the fitted model tell?

```
freq = c(housing.nom$sat_low, housing.nom$sat_medium, housing.nom$sat_high)
res = c(rep(c("sat_low","sat_medium","sat_high"), c(6,6,6)))
res = factor(res, levels = c("sat_low","sat_medium","sat_high"), ordered = T)
housing.ord = data.frame(res = res, type = rep(housing.nom$type, 3), contact = rep(housing.nom$contact,
  mutate(res = factor(res, levels = c("sat_low","sat_medium","sat_high"), ordered = TRUE))
housing.ord
```

```
##           res      type contact freq
## 1     sat_low     tower     low   65
## 2     sat_low apartment     low  130
## 3     sat_low     house     low   67
## 4     sat_low     tower    high   34
## 5     sat_low apartment    high  141
## 6     sat_low     house    high  130
## 7  sat_medium     tower     low   54
## 8  sat_medium apartment     low   76
## 9  sat_medium     house     low   48
## 10 sat_medium     tower    high   47
## 11 sat_medium apartment    high  116
## 12 sat_medium     house    high  105
## 13   sat_high     tower     low  100
## 14   sat_high apartment     low  111
## 15   sat_high     house     low   62
## 16   sat_high     tower    high  100
## 17   sat_high apartment    high  191
## 18   sat_high     house    high  104
```

```
housing.polr = polr(res ~ type + contact, data = housing.ord, weights = freq)
summary(housing.polr)
```

```
## Call:
## polr(formula = res ~ type + contact, data = housing.ord, weights = freq)
##
## Coefficients:
##                 Value Std. Error t value
## typeapartment -0.5009    0.11675  -4.291
## typehouse     -0.7362    0.12610  -5.838
## contacthigh    0.2524    0.09306   2.713
##
## Intercepts:
##                      Value   Std. Error t value
## sat_low|sat_medium  -0.9973  0.1075    -9.2794
## sat_medium|sat_high  0.1152  0.1047     1.1004
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

```
pihat.ord = predict(housing.polr, housing.nom, type = 'probs')
m.ord = rowSums(cbind(housing.nom$sat_high, housing.nom$sat_medium, housing.nom$sat_low))
res.pearson.ord = (housing.nom[,3:5] - pihat.ord*m.ord)/sqrt(pihat.ord*m.ord)


G.stat.ord = sum(res.pearson.ord^2)
pval.ord = 1 - pchisq(G.stat.ord,df = (6-4)*(3-1))
D.stat.ord = sum(2*housing.nom[,3:5] * log(housing.nom[,3:5]/(m.ord*pihat.ord)))
```

- Odds ratio with confidence interval:

Take low contact and housing type = tower as the reference group, the odds of people have high and medium satisfaction versus low satisfaction (or high satisfaction versus medium and low satisfactions) when comparing the reference group and people who:
* live in apartment is 1.65, with 95% confidence interval (1.313, 2.075).
* live in house is 2.088, with 95% confidence interval (1.631, 2.672).
* highly contact with other residents is1.287, with 95% confidence interval (0.647, 0.932).

- Goodness of Fit:

The deviance of this model is 11.6990865. The deviance statistics should approximately follow chi-squared distribution with degree of freedom 7, with the critical value 14.0671404. Therefore, it fails to reject the null hypothesis at 0.05 significance level, and we are able to conclude that the model fits the data well.

4. Calculate Pearson residuals from the proportional odds model for ordinal response to find where the largest discrepancies are between the observed frequencies and expected frequencies estimated from the model.

```
res.pearson.ord = (housing.nom[,3:5] - pihat.ord*m.ord)/sqrt(pihat.ord*m.ord)
res.pearson.ord
```

```
##      sat_low sat_medium    sat_high
## 1  0.7793957 -0.3697193 -0.31511792
## 2  0.9177560 -1.0671823 -0.01527344
## 3 -1.1407855  0.1397563  1.24407710
## 4 -0.9946852  0.4549302  0.33544295
## 5 -0.2369309 -0.4052334  0.53777345
## 6  0.2743817  1.3677881 -1.47782697
```

According to the table, the largest discrepancy is the housing type = House, with high contact and high satisfaction. The Pearson residuals value is -1.478.