Homework 4

Yuki Joyama

2024-02-29

```r
# data prep
df_house = data.frame(
  contact = c(rep(c("low", "high"), times = c(3, 3))),
  type = c(rep(c("tower", "apartment", "house"), length.out = 3)),
  sat.low = c(65, 130, 67, 34, 141, 130),
  sat.medium = c(54, 76, 48, 47, 116, 105),
  sat.high = c(100, 111, 62, 100, 191, 104)
)
```

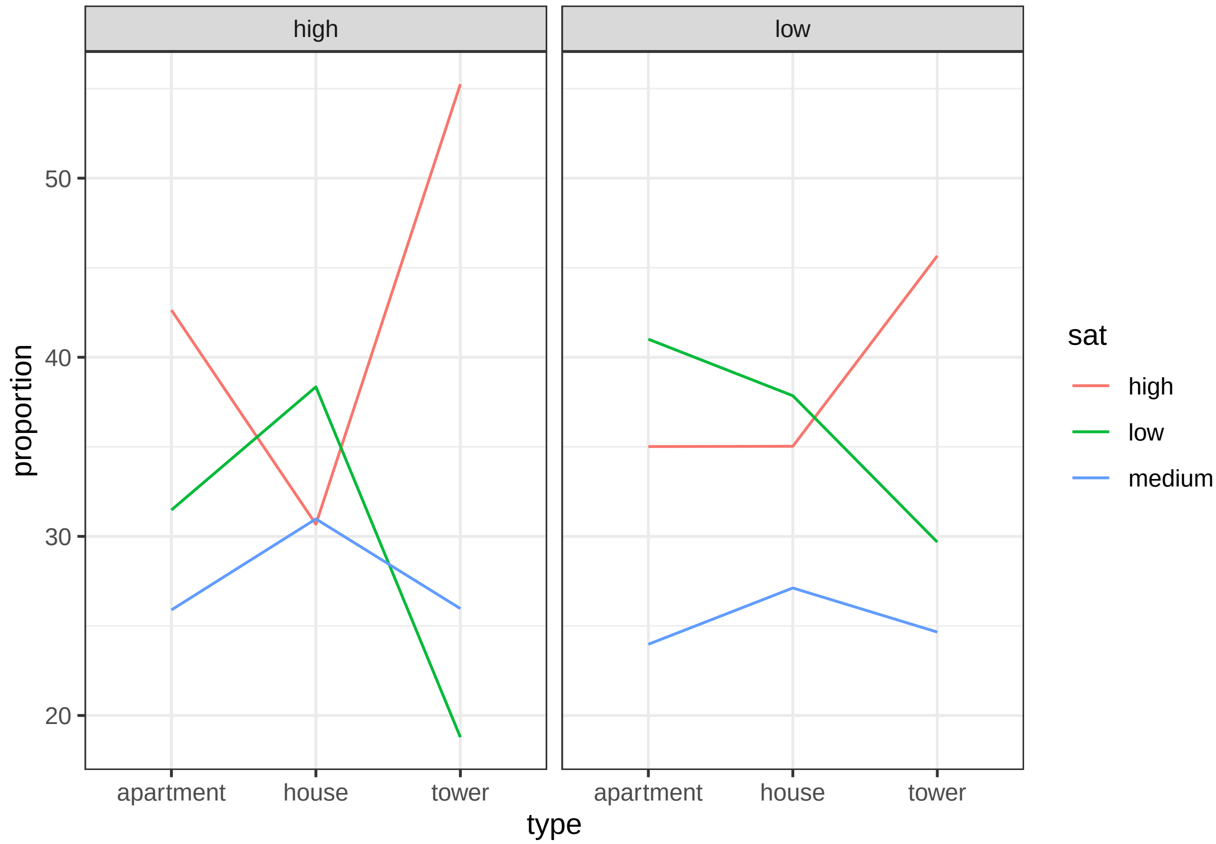
1

```r
# calculate row-wise percentages
df_house$sat.low_per <- (df_house$sat.low / rowSums(df_house[, c("sat.low", "sat.medium", "sat.high")])
df_house$sat.medium_per <- (df_house$sat.medium / rowSums(df_house[, c("sat.low", "sat.medium", "sat.hi
df_house$sat.high_per <- (df_house$sat.high / rowSums(df_house[, c("sat.low", "sat.medium", "sat.high")

# table of percentages
df_house[-(3:5)]
```

```
##   contact      type sat.low_per sat.medium_per sat.high_per
## 1     low     tower    29.68037       24.65753     45.66210
## 2     low apartment    41.00946       23.97476     35.01577
## 3     low     house    37.85311       27.11864     35.02825
## 4    high     tower    18.78453       25.96685     55.24862
## 5    high apartment    31.47321       25.89286     42.63393
## 6    high     house    38.34808       30.97345     30.67847
```

```r
# plot
df_house |>
  dplyr::select(contact, type, sat.low_per, sat.medium_per, sat.high_per) |>
  pivot_longer(cols = starts_with("sat."),
               names_to = "sat",
               values_to = "proportion") |>
  mutate(sat = str_remove(sat, "sat\\.") |>  str_remove("_per")) |>
# plot
  ggplot(aes(x = type, y = proportion, group = sat, color = sat)) +
  geom_line() +
  facet_grid(~contact) +
  theme_bw()
```



Percentages of responses in each category by contact with other residents and type of housing is summarized in the above table and plots. Top panel is the group that answered "low" for contact and bottom is the group that answered "high". We can see that tower residents are likely to have high satisfaction compared to other types of housing, and high contact tends to have high satisfaction expect for those who live in a house.

2

```r
# fit a nominal logistic regression model
house.mult <- multinom(cbind(sat.low, sat.medium, sat.high) ~ factor(contact) + factor(type), data = df_
```

```r
summary(house.mult)
```

```
## Call:
## multinom(formula = cbind(sat.low, sat.medium, sat.high) ~ factor(contact) +
##     factor(type), data = df_house)
##
## Coefficients:
##            (Intercept) factor(contact)low factor(type)house factor(type)tower
## sat.medium  -0.2180364         -0.2959832        0.06967922         0.4067631
## sat.high     0.2474047         -0.3282264       -0.30402275         0.6415948
##
## Std. Errors:
##            (Intercept) factor(contact)low factor(type)house factor(type)tower
```

```
## sat.medium  0.10930968          0.1301046         0.1437749         0.1713009
## sat.high    0.09783068          0.1181870         0.1351693         0.1500774
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

```r
# obtain odds ratio for each coefficient
exp(coef(house.mult))
```

```
##            (Intercept) factor(contact)low factor(type)house factor(type)tower
## sat.medium   0.8040962          0.7437999         1.0721642          1.501948
## sat.high     1.2806973          0.7201999         0.7378441          1.899508
```

The nominal logistic regression model is as follows:
$log(\frac{\pi_j}{\pi_1}) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3$, $j = 2, 3$

where
π_1 : the probability of low satisfaction
π_2 : the probability of medium satisfaction
π_3 : the probability of high satisfaction

$$x_1 = \begin{cases} 1 & \text{for low contact} \\ 0 & \text{for high contact} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{for house} \\ 0 & \text{for apartment} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{for tower} \\ 0 & \text{for apartment} \end{cases}$$

We observe that when the level of contact with other residents is low, the odds of experiencing medium and high satisfaction decrease compared to low satisfaction, while controlling for other variables. Similarly, for individuals residing in houses compared to apartments, the odds of experiencing medium satisfaction are higher relative to low satisfaction, whereas the odds of high satisfaction are lower holding other variables unchanged. Among those living in towers, the odds of experiencing both medium and high satisfaction are higher than those for low satisfaction, all else being equal.
For example, the odds ratio of falling into high satisfaction category (vs low satisfaction) for tower residents with high contact with other residents is 1.9

The 95% confidence intervals for each odds ratio is as follows:

```r
# 95%CI for odds ratio
exp(confint(house.mult))
```

```
## , , sat.medium
##
##                        2.5 %    97.5 %
## (Intercept)        0.6490280 0.9962138
## factor(contact)low 0.5763827 0.9598455
## factor(type)house  0.8088721 1.4211592
## factor(type)tower  1.0736021 2.1011960
##
## , , sat.high
```

```
##
##                          2.5 %    97.5 %
## (Intercept)          1.0572382 1.5513869
## factor(contact)low  0.5712840 0.9079335
## factor(type)house   0.5661197 0.9616586
## factor(type)tower   1.4154515 2.5491018
```

```r
# goodness of fit
pihat = predict(house.mult,type = 'probs')
pihat
```

```
##     sat.low sat.medium  sat.high
## 1 0.2739485  0.2460866 0.4799649
## 2 0.3967554  0.2372941 0.3659505
## 3 0.4306997  0.2761849 0.2931154
## 4 0.2154984  0.2602598 0.5242418
## 5 0.3241708  0.2606645 0.4151647
## 6 0.3562423  0.3071247 0.3366329
```

```r
m = rowSums(df_house[, 3:5])

# pearson residuals
res.pearson = (df_house[, 3:5] - pihat*m) / sqrt(pihat*m)
res.pearson
```

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

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

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

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

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

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

Generalized Pearson χ^2 statistic shows p-value of 0.14, indicating that the model has a good fit.

```r
# interaction
house.mult_int <- multinom(cbind(sat.low, sat.medium, sat.high) ~ factor(contact) + factor(type) + facto
```

```
## # weights:  21 (12 variable)
## initial  value 1846.767257
## iter  10 value 1800.128659
## final  value 1799.293647
## converged
```

```r
summary(house.mult_int)
```

```
## Call:
## multinom(formula = cbind(sat.low, sat.medium, sat.high) ~ factor(contact) +
##     factor(type) + factor(contact) * factor(type), data = df_house)
##
## Coefficients:
##            (Intercept) factor(contact)low factor(type)house factor(type)tower
## sat.medium  -0.1951677          -0.341634       -0.01840665         0.5189502
## sat.high     0.3035139          -0.461520       -0.52665690         0.7752913
##            factor(contact)low:factor(type)house
## sat.medium                            0.2217172
## sat.high                              0.6071035
##            factor(contact)low:factor(type)tower
## sat.medium                           -0.1675522
## sat.high                             -0.1865006
##
## Std. Errors:
##            (Intercept) factor(contact)low factor(type)house factor(type)tower
## sat.medium   0.1253510          0.1912147         0.1814635         0.2576842
## sat.high     0.1110307          0.1703794         0.1721496         0.2274631
##            factor(contact)low:factor(type)house
## sat.medium                            0.2992288
## sat.high                              0.2781928
##            factor(contact)low:factor(type)tower
## sat.medium                            0.3480726
## sat.high                              0.3063093
##
## Residual Deviance: 3598.587
## AIC: 3622.587
```

```r
# obtain odds ratio for each coefficient
exp(coef(house.mult_int))
```

```
##            (Intercept) factor(contact)low factor(type)house factor(type)tower
## sat.medium   0.8226967          0.7106083         0.9817617          1.680263
## sat.high     1.3546104          0.6303248         0.5905760          2.171224
##            factor(contact)low:factor(type)house
## sat.medium                             1.248218
## sat.high                               1.835108
##            factor(contact)low:factor(type)tower
## sat.medium                            0.8457325
## sat.high                              0.8298581
```

```r
# 95%CI for odds ratio
exp(confint(house.mult_int))
```

```
## , , sat.medium
##
##                                          2.5 %   97.5 %
## (Intercept)                          0.6434885 1.051814
## factor(contact)low                   0.4885038 1.033695
## factor(type)house                    0.6879297 1.401097
## factor(type)tower                    1.0139955 2.784315
## factor(contact)low:factor(type)house 0.6943629 2.243854
## factor(contact)low:factor(type)tower 0.4275167 1.673065
##
## , , sat.high
##
##                                          2.5 %    97.5 %
## (Intercept)                          1.0896950 1.6839294
## factor(contact)low                   0.4513747 0.8802207
## factor(type)house                    0.4214459 0.8275797
## factor(type)tower                    1.3902335 3.3909524
## factor(contact)low:factor(type)house 1.0638088 3.1656277
## factor(contact)low:factor(type)tower 0.4552740 1.5126373
```

From the output, we can see that there is a statistically significant interaction between low contact and house type for odds ratio of high satisfaction vs low satisfaction.

3

Now, I will treat the satisfaction categories as ordinal variable and fit a proportional odds model.

```r
# data prep
df_house2 <- df_house |>
  dplyr::select(contact, type, sat.low, sat.medium, sat.high) |>
  pivot_longer(cols = starts_with("sat."),
               names_to = "sat",
               values_to = "frequency") |>
  mutate(
    sat = str_remove(sat, "sat\\."),
    sat = factor(sat, levels = c("low", "medium", "high"))
  )

# fit an ordinal logistic regression model
house.mult2 <- polr(sat ~ contact + type, data = df_house2, weights = frequency)
summary(house.mult2)
```

```
## Call:
## polr(formula = sat ~ contact + type, data = df_house2, weights = frequency)
##
## Coefficients:
##              Value Std. Error t value
## contactlow -0.2524    0.09306  -2.713
```

```
## typehouse  -0.2353    0.10521  -2.236
## typetower   0.5010    0.11675   4.291
##
## Intercepts:
##              Value   Std. Error t value
## low|medium  -0.7488  0.0818    -9.1570
## medium|high  0.3637  0.0801     4.5393
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

```r
exp(-coef(house.mult2))
```

```
## contactlow  typehouse  typetower
##  1.2871583  1.2652791  0.6059505
```

Let Y be an ordinal outcome with J categories. $P(Y \leq j)$ is the cumulative probability of Y less than or equal to a specific category $j = 1, ..., J-1$
In polr, the ordinal logistic regression model is parameterized as
$logit(P(Y \leq j)) = log\frac{P(Y \leq j)}{P(Y > j)} = \beta_{j0} - \eta_1 x_1 - \eta_2 x_2 - \eta_3 x_3$
where $\eta_i = -\beta_i$

The output shows that the odds ratio of falling in lower category is 1.29 in people with low contact holding other variables constant. And the odds ratio of falling in lower category is 0.61 in people living in tower all else being equal.

4

```r
# Pearson residuals
pihat = predict(house.mult2, df_house2, type = 'p')
m = rowSums(cbind(df_house[, 3:5]))
res.pearson = (df_house[, 3:5] - pihat*m) / sqrt(pihat*m)
G = sum(res.pearson^2)
G
```

```
## [1] 162.3016
```

```r
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
```