hw4-Multinomial/Ordinal data case study

Cary Ni

2023-03-02

```
# build three-dimension array to store the data
raw_data = c(65, 130, 67,
             34, 141, 130,
             54, 76, 48,
             47, 116, 105,
             100, 111, 62,
             100, 191, 104)
row_name = c("tower", "apart", "house")
column_name = c("con.low", "con.high")
matrix_name = c("sat.low", "sat.med", "sat.high")
data_array = array(raw_data, dim = c(3, 2, 3),
                   dimnames = list(row_name,  column_name,matrix_name))
# check the structure of data
data_array
```

```
## , , sat.low
##
##       con.low con.high
## tower      65       34
## apart     130      141
## house      67      130
##
## , , sat.med
##
##       con.low con.high
## tower      54       47
## apart      76      116
## house      48      105
##
## , , sat.high
##
##       con.low con.high
## tower     100      100
## apart     111      191
## house      62      104
```

```
# create frequency table for con vs sat
table_1 = margin.table(data_array, margin = c(2, 3))
# turn it into proportion table for comparison
table_con = prop.table(table_1, margin = 1)
```

```r
# examine the independence based on chisquared test
chisq.test(table_1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_1
## X-squared = 5.1398, df = 2, p-value = 0.07654
```

1.1 The results from chi-square test shows that the level of satisfaction is not associated with the levels of contact with other residents as the test statistic is 5.14 with p value of 0.077, which fails to reject the null hypothesis of independence between the two variables.

```r
# create frequency table for house type vs sat
table_2 = margin.table(data_array, margin = c(1, 3))
# turn it into proportion table for comparison
table_type = prop.table(table_2, margin = 1)
# examine the independence based on chi-squared test
chisq.test(table_2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table_2
## X-squared = 34.024, df = 4, p-value = 7.369e-07
```

1.2 The results from chi-square test shows that the level of satisfaction is associated with the different housing type as the test statistic is 34.02 with p value 7.4e-7 , which rejects the null hypothesis of independence between the two variables. It can be seen from the proportion table that tower block has significantly higher satisfaction rate (high) and lower dissatisfaction rate (low) compared to the other housing types.

Fit multinomial logistic model

```r
# turn the variables into vectors
low_sat = as.vector(data_array[,,1])
med_sat = as.vector(data_array[,,2])
high_sat = as.vector(data_array[,,3])
con_level = c(rep("Low", 3), rep("High", 3)) %>% as.factor()
type_level = rep(c("tower", "apart", "house"), 2) %>% as.factor()
mult_model = multinom(cbind(low_sat, med_sat, high_sat)~con_level+type_level)
```

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

```r
summary(mult_model)
```

```
## Call:
## multinom(formula = cbind(low_sat, med_sat, high_sat) ~ con_level +
##     type_level)
##
## Coefficients:
##          (Intercept) con_levelLow type_levelhouse type_leveltower
## med_sat   -0.2180364   -0.2959832      0.06967922       0.4067631
## high_sat   0.2474047   -0.3282264     -0.30402275       0.6415948
##
## Std. Errors:
##          (Intercept) con_levelLow type_levelhouse type_leveltower
## med_sat   0.10930968    0.1301046       0.1437749       0.1713009
## high_sat  0.09783068    0.1181870       0.1351693       0.1500774
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

```
# use likelihood ratio test for goodness of fit for two terms
lrtest(mult_model, "con_level")
```

```
## # weights:  12 (6 variable)
## initial  value 1846.767257
## iter  10 value 1807.174032
## iter  10 value 1807.174031
## iter  10 value 1807.174031
## final  value 1807.174031
## converged

## Likelihood ratio test
##
## Model 1: cbind(low_sat, med_sat, high_sat) ~ con_level + type_level
## Model 2: cbind(low_sat, med_sat, high_sat) ~ type_level
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -1802.7
## 2   6 -1807.2 -2 8.8677    0.01187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(mult_model, "type_level")
```

```
## # weights:  9 (4 variable)
## initial  value 1846.767257
## final  value 1821.875901
## converged

## Likelihood ratio test
##
## Model 1: cbind(low_sat, med_sat, high_sat) ~ con_level + type_level
## Model 2: cbind(low_sat, med_sat, high_sat) ~ con_level
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   8 -1802.7
## 2   4 -1821.9 -4 38.272   9.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# CI for log odds for each of the terms
exp(confint(mult_model))
```

```
## , , med_sat
##
##                       2.5 %    97.5 %
## (Intercept)     0.6490280 0.9962138
## con_levelLow    0.5763827 0.9598455
## type_levelhouse 0.8088721 1.4211592
## type_leveltower 1.0736021 2.1011960
##
## , , high_sat
##
##                       2.5 %    97.5 %
## (Intercept)     1.0572382 1.5513869
## con_levelLow    0.5712840 0.9079335
## type_levelhouse 0.5661197 0.9616586
## type_leveltower 1.4154515 2.5491018
```

```r
mult_model_10 = multinom(cbind(low_sat, med_sat, high_sat)~con_level)
```

```
## # weights:  9 (4 variable)
## initial  value 1846.767257
## final  value 1821.875901
## converged
```

```r
deviance(mult_model_10) - deviance(mult_model)
```

```
## [1] 38.27148
```

2. It can be seen from the results of likelihood ratio test which tests the β of `con_level` and `type_level` that the null hypothesis that these two coefficients are zero are rejected, suggesting that both contact level and housing types are related to satisfaction level. Based on the CI of each coefficients, lower contact level decreases the odds of medium satisfaction to low satisfaction by factor of 0.74 with 95% CI (0.576, 0.960) and the odds of high to low satisfaction by factor of 0.72 with 95% CI (0.571, 0.908) for each fixed housing types. Compared to apartment, living in house increases the odds of medium satisfaction to low satisfaction by factor of 1.07 with 95% CI (0.809, 1.421) and the decrease the odds of high to low satisfaction by factor of 0.738 with 95% CI (0.566, 0.961) holding contact level fixed. Compared to apartment, living in tower increases the odds of medium satisfaction to low satisfaction by factor of 1.50 with 95% CI (1.074, 2.101) and the odds of high to low satisfaction by factor of 1.90 with 95% CI (1.415, 2.549) holding contact level fixed.

```r
# find deviance for goodness of fit
pihat = predict(mult_model, type = "probs")
data_df = data_frame(
  low = low_sat,
  med = med_sat,
  high = high_sat
)
m = rowSums(data_df)
# get d_stat to test goodness of fit
d_stat = sum(2*data_df*log(data_df/(pihat*m)))
1 - pchisq(d_stat, (6-4)*(3-1))
```

```
## [1] 0.1416504
```

```
# G_stat
res_pearson = (data_df-pihat*m)/sqrt(pihat*m)
g_stat = sum(res_pearson^2)
d_stat
```

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

The Deviance statistic is 6.89 with p value of 0.142, which suggests the good fit of this multinomial model.

```
# fit the model with interaction
mult_model_2 = multinom(cbind(low_sat, med_sat, high_sat)~con_level*type_level)
```

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

```
summary(mult_model_2)
```

```
## Call:
## multinom(formula = cbind(low_sat, med_sat, high_sat) ~ con_level *
##     type_level)
##
## Coefficients:
##          (Intercept) con_levelLow type_levelhouse type_leveltower
## med_sat   -0.1951677    -0.341634     -0.01840665       0.5189502
## high_sat   0.3035139    -0.461520     -0.52665690       0.7752913
##          con_levelLow:type_levelhouse con_levelLow:type_leveltower
## med_sat                     0.2217172                   -0.1675522
## high_sat                    0.6071035                   -0.1865006
##
## Std. Errors:
##          (Intercept) con_levelLow type_levelhouse type_leveltower
## med_sat    0.1253510    0.1912147       0.1814635       0.2576842
## high_sat   0.1110307    0.1703794       0.1721496       0.2274631
##          con_levelLow:type_levelhouse con_levelLow:type_leveltower
## med_sat                     0.2992288                    0.3480726
## high_sat                    0.2781928                    0.3063093
##
## Residual Deviance: 3598.587
## AIC: 3622.587
```

```
# compare the two models with lrt
lrtest(mult_model, mult_model_2)
```

```
## Likelihood ratio test
##
## Model 1: cbind(low_sat, med_sat, high_sat) ~ con_level + type_level
```

```
## Model 2: cbind(low_sat, med_sat, high_sat) ~ con_level * type_level
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   8 -1802.7
## 2  12 -1799.3  4 6.893     0.1417
```

```
# use wald test to measure each term in the multinomial model
# wald.test(Sigma = vcov(mult_model), b = coef(mult_model), Terms = 1:2)
# z = summary(mult_model)$coefficients/summary(mult_model)$standard.errors
# (1 - pnorm(abs(z), 0, 1)) * 2
# anova(mult_model, update(mult_model, ~ 1), test = "Chisq")
```

The result of likelihood ratio test gives test statistic of 6.9 and p value of 0.1417 when comparing the two models differentiated by interaction term. The null hypothesis that there is no interaction between `con_level` and `type_level` is not rejected, thus we could say the odds of medium and high satisfaction to low satisfaction level is not associated with the interaction between contact level and housing types.

Fit proportional odds model

```
# complile into a dataframe
por_data = tibble(
  res = factor(c(rep(c("sat.low", "sat.med","sat.high"), c(6, 6, 6))),
               levels = c("sat.low", "sat.med","sat.high"),
               ordered = TRUE),
  house_type = rep(c("tower", "apart", "house"), 6),
  cont_level = rep(rep(c("con.low", "con.high"), c(3, 3)), 3),
  freq = c(low_sat, med_sat, high_sat)
)
ord_model = polr(res ~ factor(house_type) + factor(cont_level), data = por_data, weights = freq)
summary(ord_model)
```

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

## Call:
## polr(formula = res ~ factor(house_type) + factor(cont_level),
##     data = por_data, weights = freq)
##
## Coefficients:
##                              Value Std. Error t value
## factor(house_type)house   -0.2353    0.10521  -2.236
## factor(house_type)tower    0.5010    0.11675   4.291
## factor(cont_level)con.low -0.2524    0.09306  -2.713
##
## Intercepts:
##                   Value   Std. Error t value
## sat.low|sat.med  -0.7488  0.0818    -9.1570
## sat.med|sat.high  0.3637  0.0801     4.5393
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

```r
exp(coef(ord_model))
```

```
##   factor(house_type)house   factor(house_type)tower factor(cont_level)con.low
##                 0.7903394                 1.6502997                 0.7769052
```

```r
exp(confint(ord_model))
```

```
## Waiting for profiling to be done...
##
## Re-fitting to get Hessian
```

```
##                               2.5 %    97.5 %
## factor(house_type)house   0.6429196 0.9711892
## factor(house_type)tower   1.3136017 2.0762957
## factor(cont_level)con.low 0.6472271 0.9321964
```

3. The coefficients indicates that lower level of contact decrease the cumulative odds for higher satisfaction level by 0.78 with 95% CI (0.647, 0.932) holding housing type fixed. Compared to apartment, living in house increases the cumulative odds for higher satisfaction level by factor of 1.65 with 95% CI (1.314, 2.076) and living in tower decrease cumulative odds for higher satisfaction level to factor of 0.79 with 95% CI (0.643, 0.971) holding contact level fixed.

```r
# make sure the prediction is for six groups
data_df_2 = por_data[-c(1, 4)] %>% unique()
pihat_2 = predict(ord_model,data_df_2, type = "probs")
# get d_stat to test goodness of fit for model
d_stat_2 = sum(2*data_df*log(data_df/(pihat_2*m)))
1 - pchisq(d_stat_2, (12-5))
```

```
## [1] 0.110899
```

```r
# G_stat for gof
res_pearson_2 = (data_df-pihat_2*m)/sqrt(pihat_2*m)
g_stat_2 = sum(res_pearson_2^2)
g_stat_2
```

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

```r
# lrt to test for coefficient
lrtest(ord_model, update(ord_model, ~ factor(house_type)))
```

```
## Likelihood ratio test
##
## Model 1: res ~ factor(house_type) + factor(cont_level)
## Model 2: res ~ factor(house_type)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -1805.1
## 2   4 -1808.8 -1 7.3757   0.006611 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lrtest(ord_model, update(ord_model, ~ factor(cont_level)))
```

```
## Likelihood ratio test
##
## Model 1: res ~ factor(house_type) + factor(cont_level)
## Model 2: res ~ factor(cont_level)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -1805.1
## 2   3 -1822.8 -2 35.322  2.138e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The assumption of this proportional odds model is met and the deviance statistic is 11.7 with p value of 0.11 which suggests good fit. The likelihood ratio tests of the predictor `house_type` and `cont_level` generates p value lower than 0.05, which suggests the response variable is indeed associated with the two predictors above.

Calculate Pearson residuals from proportional odds model

```
obs_df = data_df %>% as.matrix()
what = paste(data_df_2$house_type, data_df_2$cont_level, sep = "&")
rownames(obs_df) = what
expc_df = m*pihat_2
rownames(expc_df) = what
obs_df
```

```
##                low med high
## tower&con.low   65  54  100
## apart&con.low  130  76  111
## house&con.low   67  48   62
## tower&con.high  34  47  100
## apart&con.high 141 116  191
## house&con.high 130 105  104
```

```
expc_df
```

```
##                   sat.low   sat.med  sat.high
## tower&con.low   59.01254  56.78574 103.20172
## apart&con.low  119.94955  85.88991 111.16053
## house&con.low   77.01170  47.04117  52.94713
## tower&con.high  40.31555  43.98260  96.70185
## apart&con.high 143.84263 120.44691 183.71047
## house&con.high 126.90999  91.88815 120.20186
```

```
# show the pearson residual for each of the groups
(obs_df - expc_df)/sqrt(expc_df)
```

```
##                       low        med        high
## tower&con.low   0.7794178 -0.3696760 -0.31516596
```

```
## apart&con.low   0.9176690 -1.0671401 -0.01522607
## house&con.low  -1.1408527  0.1397992  1.24412782
## tower&con.high -0.9946598  0.4549796  0.33539209
## apart&con.high -0.2370150 -0.4051916  0.53781495
## house&con.high  0.2742913  1.3678370 -1.47777863
```

```
# show sd for residual
sqrt(g_stat_2/(12-5))
```

```
## [1] 1.289632
```

4. It can be seen from the summary table that the largest differences between observed and expected frequencies are seen in house living type for both low and high contact level, especially for the high satisfaction level. The largest residual is -1.48 with house, high contact level, and high satisfaction, which is not over 2 standard deviation which is 2.58.