P8131-hw4-rw2844

Renjie Wei

2022-03-02

```
library(tidyverse)
library(nnet)
library(MASS)
```

Question 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.

Answers

Table 1: Contact and satisfaction (%)

Contact with others	Low Satisfaction	Medium Satisfaction	High Satisfaction
low	36.746	24.965	38.289
high	31.508	27.686	40.806

The table above shows the pair-wise associations between the levels of satisfaction and contact with other residents. From the table, we can see that the levels of satisfaction is relatively higher in high contact group than that in low contact group. Residents who lived with a high level contacts with other residents tends to have higher satisfaction.

Table 2: Housing and satisfaction (%)

House Type	Low Satisfaction	Medium Satisfaction	High Satisfaction
apartment	35.425	25.098	39.477
tower block	24.750	25.250	50.000
house	38.178	29.651	32.171

The table above shows the pair-wise associations between the levels of satisfaction and types of housing. From the table, we can see that the levels of satisfaction is relatively higher in residents who lived in tower block. However, the difference in those who lived in apartment and house are not that obvious.

Question 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 findings (the pattern in the associations, odds ratios with 95% confidence intervals, goodness-of-fit). (Hint: use dummy variable for house types.) Is there interaction of contact level by house type?

Answers

Model Fitting

The multinomial logistic models to be fitted are:

$$\begin{cases} \log(\frac{\pi_{\text{satisfaction=medium}}}{\pi_{\text{satisfaction=low}}}) = \beta_{20} + \beta_{21} \times \text{House} + \beta_{22} \times \text{Tower Block} + \beta_{23} \times \text{High Contact} \\ \log(\frac{\pi_{\text{satisfaction=high}}}{\pi_{\text{satisfaction=low}}}) = \beta_{30} + \beta_{31} \times \text{House} + \beta_{32} \times \text{Tower Block} + \beta_{33} \times \text{High Contact} \end{cases}$$

```
set.seed(2022)
house.multinom <- multinom(cbind(sat.low, sat.medium, sat.high)~ house.type + contact, data = house.data
```

```
## # weights:  15 (8 variable)
## initial  value 1846.767257
## iter  10 value 1802.866981
## final  value 1802.740161
## converged
```

```
multinom.sum <- summary(house.multinom)
```

Findings

The following table shows the results of the multinomial models:

Table 3: Multinomial Model Results

Groups	Variable	$\hat{\beta}$	95% CI of $\hat{\beta}$	OR	95% of OR
High	High Contact	0.3282256	(0.097 0.56)	1.3885022	(1.101 1.75)
Medium	High Contact	0.2959796	(0.041 0.551)	1.3444428	(1.042 1.735)
High	House	-0.3040194	(-0.569 -0.039)	0.7378466	(0.566 0.962)
Medium	House	0.0696779	(-0.212 0.351)	1.0721628	(0.809 1.421)
High	Tower Block	0.6415915	(0.347 0.936)	1.8995016	(1.415 2.549)
Medium	Tower Block	0.4067570	(0.071 0.743)	1.5019391	(1.074 2.101)
High	Intercept	-0.0808231	(-0.292 0.131)	0.9223569	(0.746 1.14)
Medium	Intercept	-0.5140171	(-0.751 -0.277)	0.5980882	(0.472 0.758)

From this table, we can know that:

- Given the level of contact:
    1. The odds of having medium satisfaction level comparing to low satisfaction in the residents who live in house is 1.072 times the odds of that in the residents who live in apartment.
    And the CI of odds ratio is (0.809 1.421), which is not significant at our significance level (0.05).

2. The odds of having high satisfaction level comparing to low satisfaction in the residents who live in house is 0.738 times the odds of that in the residents who live in apartment.
   And the CI of odds ratio is (0.566 0.962).
3. The odds of having medium satisfaction level comparing to low satisfaction in the residents who live in tower block is 1.502 times the odds of that in the residents who live in apartment.
   And the CI of odds ratio is (1.074 2.101).
4. The odds of having high satisfaction level comparing to low satisfaction in the residents who live in tower block is 1.9 times the odds of that in the residents who live in apartment.
   And the CI of odds ratio is (1.415 2.549).

- Given the type of housing:

1. The odds of having medium satisfaction level comparing to low satisfaction in the residents who have high contacts with other residents is 1.344 times the odds of that in the residents who have low contacts.
   And the CI of odds ratio is (1.042 1.735).
2. The odds of having high satisfaction level comparing to low satisfaction in the residents who have high contacts with other residents is 1.389 times the odds of that in the residents who have low contacts.
   And the CI of odds ratio is (1.101 1.75).

Goodness-of-Fit

```
# goodness of fit
multinom.pihat = predict(house.multinom, type = "probs")
multinom.m = rowSums(cbind(house.data$sat.low, house.data$sat.medium, house.data$sat.high))
multinom.res.pearson = (cbind(house.data$sat.low, house.data$sat.medium, house.data$sat.high) - multinor
# Generalized Pearson Chisq Stat
multinom.G.stat = sum(multinom.res.pearson^2)
multinom.pval_pearson = 1 - pchisq(multinom.G.stat, 4)
# deviance
multinom.D.stat = sum(2 * cbind(house.data$sat.low, house.data$sat.medium, house.data$sat.high) * log(cl
multinom.pval_deviance = 1 - pchisq(multinom.D.stat, 4)
```

The goodness of fit tests shows that the pearson statistic is 6.932 with degree of freedom of 4, and the p-value is 0.14 greater than 0.05. The deviance is 6.893 with degree of freedom of 4, and the p-value is 0.142 greater than 0.05. Hence, the model fits the data well.

Interaction

The model with interaction terms is:

$$\begin{cases} \log(\frac{\pi_{\text{satisfaction=medium}}}{\pi_{\text{satisfaction=low}}}) = \beta_{20} + \beta_{21} \times \text{House} + \beta_{22} \times \text{Tower Block} + \beta_{23} \times \text{High Contact} + \beta_{24} \times \text{House} \cdot \text{High Contact} + \beta_{25} \\ \log(\frac{\pi_{\text{satisfaction=high}}}{\pi_{\text{satisfaction=low}}}) = \beta_{30} + \beta_{31} \times \text{House} + \beta_{32} \times \text{Tower Block} + \beta_{33} \times \text{High Contact} + \beta_{34} \times \text{House} \cdot \text{High Contact} + \beta_{35} \times \end{cases}$$

The previous multinomial model is a nest model of this one. And we want to do the hypothesis testing with
$H_0 : \beta_{j4} = \beta_{j5} = 0$
$H_0 : \beta_{j4} \neq 0, \beta_{j5} \neq 0$ $(j \in \{2, 3\})$

```
house.interaction <- multinom(cbind(sat.low, sat.medium, sat.high)~ house.type* contact, data = house.da
```

```
## # weights:  21 (12 variable)
## initial  value 1846.767257
## iter  10 value 1804.235447
## final  value 1799.293647
## converged
```

```r
interaction.sum <- summary(house.interaction)
# Deviance Analysis
multinom.dev = house.multinom$deviance
interaction.dev = house.interaction$deviance
nest.p2 = house.interaction$edf - house.multinom$edf
nest.p_val = pchisq(multinom.dev - interaction.dev, nest.p2, lower.tail = FALSE)
```

The deviance analysis shows that the difference in deviance between model without and with interaction term is 6.893, and the p-value is $0.142 > 0.05$. Thus, we fail to reject the null hypothesis and accept the smaller model, concluding that there is no interaction between type of housing and level of contact with other residents on the odds of medium or high level of satisfaction to low.

Question 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?

Answers

Model Fitting

The proportional logistic models to be fitted are:

$$\begin{cases} \log(\frac{\pi_{\text{satisfaction=low}}}{\pi_{\text{satisfaction=medium}}+\pi_{\text{satisfaction=high}}}) = \beta_{20} + \beta_1 \times \text{House} + \beta_2 \times \text{Tower Block} + \beta_3 \times \text{High Contact} \\ \log(\frac{\pi_{\text{satisfaction=low}}+\pi_{\text{satisfaction=medium}}}{\pi_{\text{satisfaction=high}}}) = \beta_{30} + \beta_1 \times \text{House} + \beta_2 \times \text{Tower Block} + \beta_3 \times \text{High Contact} \end{cases}$$

```r
house.polr=polr(satisfaction ~ contact + house.type ,data = house.ord, weights = freq)
polr.sum <- summary(house.polr)
```

Findings

The following table shows the results of the proportional models:

Table 4: Proportional Model Results

Variable	$\hat{\beta}$	95% CI of $\hat{\beta}$	OR	95% of OR
High Contact	0.2524400	(0.07 0.435)	1.2871623	(1.073 1.545)
Tower Block	0.5009525	(0.272 0.73)	1.6502924	(1.313 2.075)
House	-0.2352978	(-0.442 -0.029)	0.7903354	(0.643 0.971)
Low vs. Medium and High	-0.4963881	(-0.672 -0.321)	0.6087253	(0.511 0.726)
Low and Medium vs. High	0.6161200	(0.44 0.793)	1.8517294	(1.552 2.209)

From this table, we can know that:

- Given the level of contact:
    1. The odds of having an higher satisfaction level in the residents who live in house is 0.79 times the odds of that in the residents who live in apartment.
    And the CI of odds ratio is (0.643 0.971).

2. The odds of having an higher satisfaction level in the residents who live in tower block is 1.65 times the odds of that in the residents who live in apartment.
   And the CI of odds ratio is (1.313 2.075).

- Given the type of housing:

  1. The odds of having an higher satisfaction level in the residents who have high contact with other residents is 1.287 times the odds of that in the residents who live in apartment.
     And the CI of odds ratio is (1.073 1.545).

- The odds of having low satisfaction than having a medium or high satisfaction on average is 0.609
  And the CI of odds ratio is (0.511 0.726).

- The odds of having low or medium satisfaction than having high satisfaction on average is 1.852
  And the CI of odds ratio is (1.552 2.209).

Question 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.

Answers

The following table shows the discrepancies are between the observed frequencies and expected frequencies estimated from the model.

```
# residuals
plor.pihat = predict(house.polr, house.data, type='p')
plor.m = rowSums(cbind(house.data$sat.low, house.data$sat.medium, house.data$sat.high))
plor.res.pearson = (house.data[,1:3] - plor.pihat * plor.m)/sqrt(plor.pihat * plor.m)
plor.res.tab <- data.frame(house.data, plor.res.pearson)
#max(abs(plor.res.pearson))# low+apartment~high
plor.G = sum(plor.res.pearson^2)
numsamp = (3 - 1)*6 # degree of freedom for grouped data
numparam = 2+3 # total num of param
plor.pval = 1 - pchisq(plor.G ,df = numsamp - numparam)
plor.res.tab %>% dplyr::select(-c("sat.low", "sat.medium", "sat.high")) %>% knitr::kable(col.names = c(
```

Table 5: Discrepancies

House Type	Contact	Low satisfaction	medium satisfaction	High satisfaction
tower block	low	0.779	-0.370	-0.315
apartment	low	0.918	-1.067	-0.015
house	low	-1.141	0.140	1.244
tower block	high	-0.995	0.455	0.335
apartment	high	-0.237	-0.405	0.538
house	high	0.274	1.368	-1.478

From the table we can see that there is a largest discrepancy in residents live in house who highly contact with others and having a high satisfaction, the absolute value of discrepancy is 1.478.