Homework 4

Xinyi Lin

2/26/2019

Question 1

1) contact with other residents

Contact with others/Satisfaction	Low(%)	Medium(%)	High(%)
Low	36.746	24.965	38.289
High	31.508	27.686	40.806

In the medium and high satisfaction level, percentages of high contact level are larger than percentages of low contact level, which means people who contact with other residents regularly are more likely to be satisfied with housing conditions.

2) type of housing

Housing type/Satisfaction	Low(%)	Medium(%)	High(%)
Tower block	24.75	25.25	50
Apartment	35.425	25.098	39.477
House	38.178	29.651	32.171

Comparing to apartment and house, a higher percentage of people who live in tower block have a high satisfaction level, which means people who live in tower block are more likely to be satisfied with housing conditions. While among people who have low satisfaction, the percentage of people who live in house are largest, which means people living in house are more likely to be unsatisfied with housing conditions.

Question 2

Fit model

To fit a nominal logistic regression, we first choose the category "low" as the reference ctegory and the model is:

$$\log(\frac{\pi_2}{\pi_1}) = \beta_{02} + \beta_{12}x_1 + \beta_{22}x_2 + \beta_{32}x_3$$

$$\log(\frac{\pi_3}{\pi_1}) = \beta_{03} + \beta_{13}x_1 + \beta_{23}x_2 + \beta_{33}x_3$$

in which

x_1 is the indicator of contact frequency and $x_1 = 0$ means low contact level while $x_1 = 1$ means high contact level;

x_2 is the indicator of housing type house;

x_3 is the indicator of housing type tower block.

By using R, we get fitted model:

$$\log(\frac{\pi_2}{\pi_1}) = -0.514 + 0.296x_1 + 0.070x_2 + 0.407x_3$$

$$\log(\frac{\pi_3}{\pi_1}) = -0.081 + 0.328x_1 + -0.304x_2 + 0.642x_3$$

Odd ratios with 95% CI

The fitted model we get:

```
## Call:
## multinom(formula = cbind(sat_low, sat_medium, sat_high) ~ contact +
##     housing, data = house_data_nominal)
##
## Coefficients:
##            (Intercept) contacthigh housinghouse housingtowerblock
## sat_medium -0.51401706   0.2959796   0.06967794         0.4067570
## sat_high   -0.08082309   0.3282256  -0.30401939         0.6415915
##
## Std. Errors:
##            (Intercept) contacthigh housinghouse housingtowerblock
## sat_medium   0.1207955   0.1301045    0.1437749         0.1713008
## sat_high     0.1079357   0.1181870    0.1351693         0.1500773
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

model	term	estimated odd ratio	95% CI
medium	contact_high	1.3444702	1.0418612, 1.7349721
medium	house	1.0721865	0.8088492, 1.4212586
medium	towerblock	1.5020037	1.0736437, 2.1012698
high	contact_high	1.3884666	1.1013451, 1.750441
high	house	0.7378609	0.5660985, 0.9617383
high	towerblock	1.8995177	1.4153961, 2.5492281

The pattern in the associations

According to estimated odd ratios, we can find that in medium and high models, estimated odd ratios of contact high vs low are larger than 1, which means that people with high contact level are more likely to have higher satisfaction level. We can also find that in medium and high models, estimated odd ratios of towerblock vs apartment are larger than 1, which means that comparing with people living in apartment, people living in tower block are more likely to have higher satisfaction level. And this pattern meet what we find in previous tables.

Test the goodness of fit

Now, we need to test the goodness of fit.

By using R to calculate, we get generalized Pearson χ^2 statistic:

$$G = \sum_{i=1}^{n}\sum_{j=1}^{J}\frac{(y_{ij} - m_i\hat{\pi}_{ij})^2}{m_i\hat{\pi}_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{J}R^2_{p_{ij}} = 6.932$$

As it follows $\chi^2(4)$, corresponding p-value is 0.14 which is larger than 0.05, so this model fit the data well.

Question 3

We use following proportional odds model to fit data:

$$\log(\frac{\pi_1}{\pi_2+\pi_3}) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$\log(\frac{\pi_1+\pi_2}{\pi_3}) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

in which

x_1 is the indicator of contact frequency and $x_1 = 0$ means low contact level while $x_1 = 1$ means high contact level;

x_2 is the indicator of housing type house;

x_3 is the indicator of housing type tower block.

By using R, we get fitted model as follow:

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

## Call:
## polr(formula = satisfaction ~ contact + housing, data = house_data_ordinal,
##     weights = number)
##
## Coefficients:
##                      Value Std. Error t value
## contacthigh         0.2524    0.09306   2.713
## housinghouse       -0.2353    0.10521  -2.236
## housingtowerblock   0.5010    0.11675   4.291
##
## Intercepts:
##                      Value   Std. Error t value
## sat_low|sat_medium  -0.4964  0.0897    -5.5356
## sat_medium|sat_high  0.6161  0.0901     6.8381
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

which is:

$$\log(\frac{\pi_1}{\pi_2+\pi_3}) = -0.496 - 0.2524 x_1 + 0.2353 x_2 - 0.5010 x_3$$
$$\log(\frac{\pi_1+\pi_2}{\pi_3}) = 0.616 - 0.2524 x_1 + 0.2353 x_2 - 0.5010 x_3$$

According to results, as $\beta_1 = -0.2524$, people who have high contact level are more likely to have higher satisfaction level given house types are same. As $\beta_2 = 0.2353$, comparing to people who live in apartment, people who live in house are more likely to have lower satisfaction level given contact levels are same. As $\beta_3 = -0.5010$, comparing to people who live in apartment, people who live in tower block are more likely to have higher satisfaction level given contact levels are same.

Question 4

By using R, we get Pearson residuals as follow:

```
##      sat_low sat_medium    sat_high
## 1 -0.2369928 -0.4051744  0.53778021
## 2  0.2742748  1.3678506 -1.47777334
## 3 -0.9946675  0.4549867  0.33539286
## 4  0.9176664 -1.0671289 -0.01523393
## 5 -1.1408779  0.1398125  1.24414948
## 6  0.7793895 -0.3696724 -0.31514862
```

According to results, we can find the largest discrepancy is the discrepancy of frequency in people who live in house with high contact level and high satisfaction level and the discrepancy is -1.478.