Polytomous Responses (nominal, ordinal)

Hana Akbarnejad

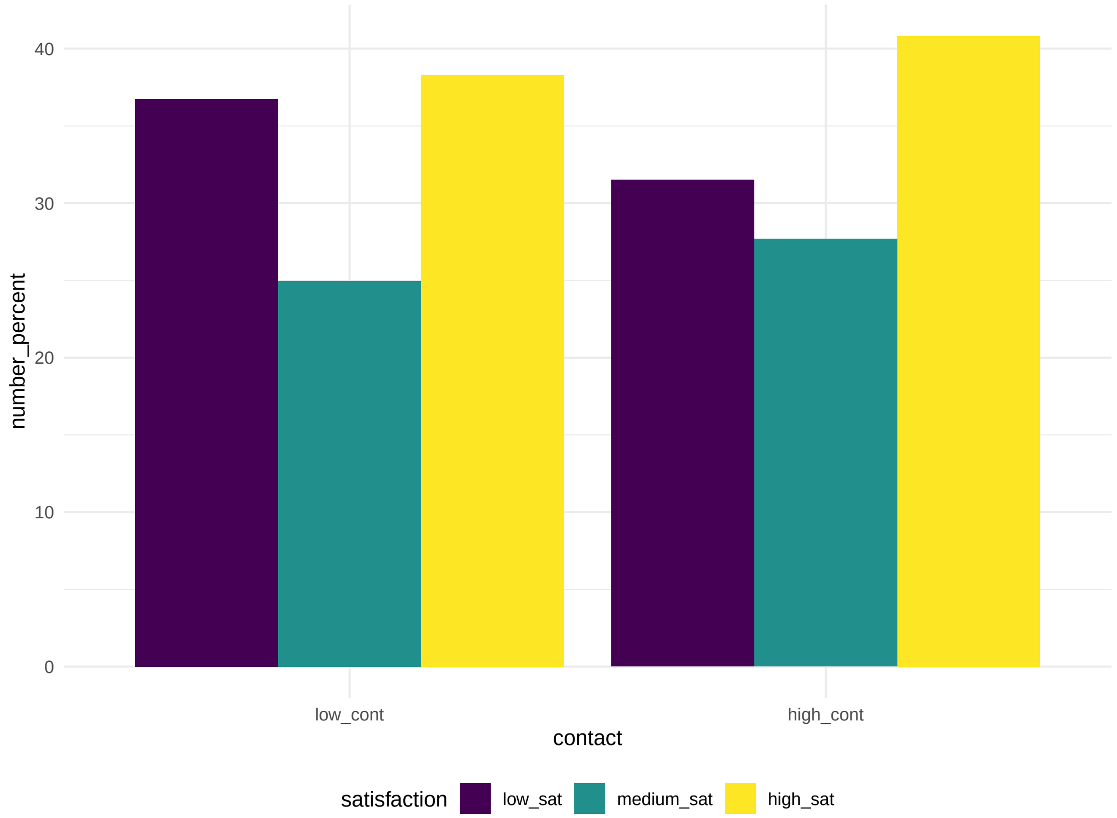
2/28/2020

We are given the data from an investigation into residents' satisfaction with their housing conditions and different types of housing and the degree of contact with other residents. We are interested in associations between the levels of satisfaction and contact with other residents and associations between the levels of satisfaction and type of housing.

Part i

We first look at association between satisfaction and contact with other residents. The table below shows the percentage of residents in each combination of contact-satisfaction level, and the bar chart below depicts the distribution:

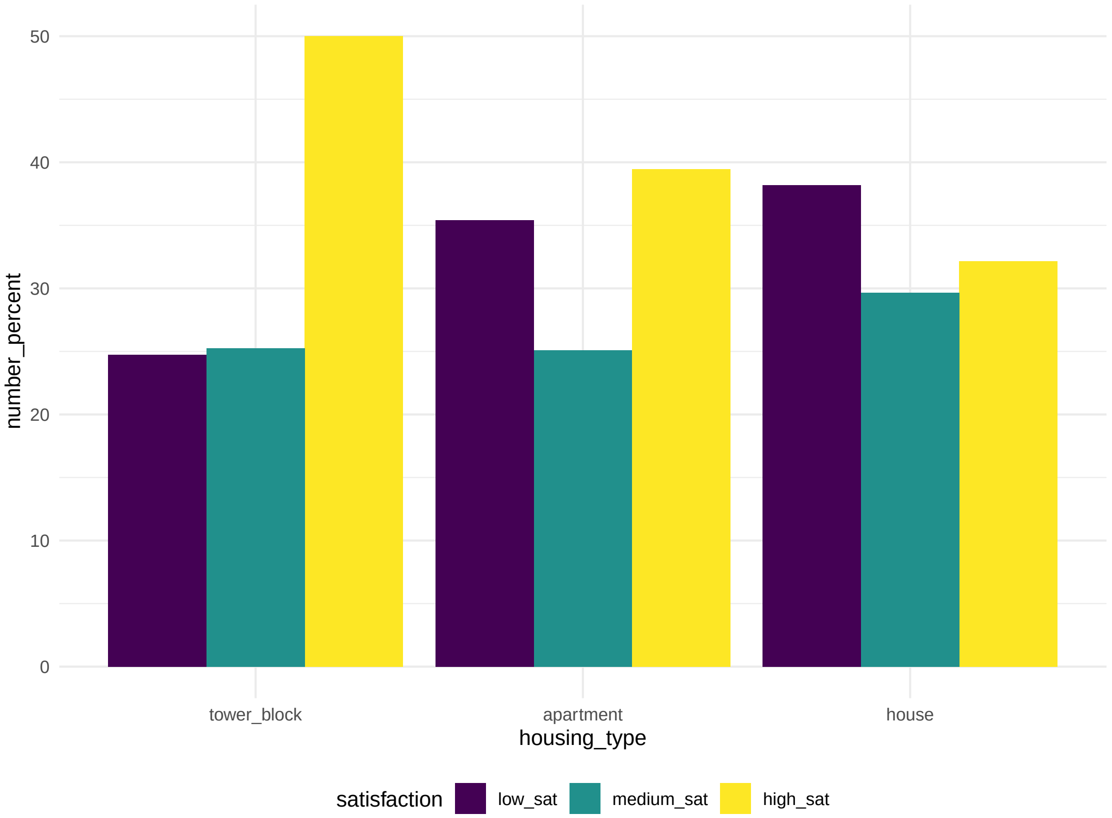
contact	low_sat	medium_sat	high_sat
low_cont	36.74614	24.96494	38.28892
high_cont	31.50826	27.68595	40.80579



We can observe that among people who have lower contact, the percentage of people with both high and low satisfaction is high. However, in people who have higher contact, the percentage of people with high satisfaction is obviously higher than medium and low satisfaction.

Now, we look at association between satisfaction and housing type that residents live in. The table below shows the percentage of residents in each combination of housing-satisfaction level, and the bar chart below depicts the distribution:

housing_type	low_sat	medium_sat	high_sat
tower_block	24.75000	25.25000	50.00000
apartment	35.42484	25.09804	39.47712
house	38.17829	29.65116	32.17054



We can observe that amon people who live in live in tower blocks, there is a high proportion of people who are highly satisfied with their living condition. In apartment housing, the percentage of people who have high and low satisfaction with their living condition is higher than medium satisfaction, and these eprcentages are relatively close. Finally, among people who live in houses, the percentage of people with low satisfaction is higher than the other two groups. In other words, when we go from tower block housing to apartment and to house, the proportion of low satisfied people increase and the percentage of medium satisfied people also increases, but leass sharply. However, the proportion of highly satisfied people drastically decreases.

Part ii

Fitting the nominal model and obtaining odds ratios with 95% confidence intervals

In this part, we are looking for the ssociations between the levels of satisfaction and the two other variables of interest using nominal logistic regression. To make the model, I use dummy variable for house types.

The first step is to choose a reference category. Here, the reference category is Low Satisfaction.

Then we need to choose the dummy variables:

x1: the indicator of high contact

x2: the indicator of apartment

x3: indicator of house

Our nominal models would look like this:

```r
# fitting a nominal model
nominal_housing_data = housing_data %>%
  pivot_wider(names_from = satisfaction, values_from = n)
```

```
sat_nominal = multinom(cbind(low_sat, medium_sat, high_sat) ~ housing_type + contact, data = nominal_hou
```

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

```
sat_nominal_fit = summary(sat_nominal)
sat_nominal_fit
```

```
## Call:
## multinom(formula = cbind(low_sat, medium_sat, high_sat) ~ housing_type +
##     contact, data = nominal_housing_data)
##
## Coefficients:
##            (Intercept) housing_typeapartment housing_typehouse contacthigh_cont
## medium_sat  -0.1072644            -0.4067537        -0.3370771        0.2959803
## high_sat     0.5607737            -0.6415967        -0.9456177        0.3282263
##
## Std. Errors:
##            (Intercept) housing_typeapartment housing_typehouse contacthigh_cont
## medium_sat   0.1524077             0.1713011         0.1803577        0.1301046
## high_sat     0.1329301             0.1500773         0.1644850        0.1181870
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

```
# computing 95% CI
ci_1 = round(exp(sat_nominal_fit$coefficients[1, 1] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
ci_2 = round(exp(sat_nominal_fit$coefficients[1, 4] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
ci_3 = round(exp(sat_nominal_fit$coefficients[1, 2] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
ci_4 = round(exp(sat_nominal_fit$coefficients[1, 3] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
ci_5 = round(exp(sat_nominal_fit$coefficients[2, 1] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
ci_6 = round(exp(sat_nominal_fit$coefficients[2, 4] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
ci_7 = round(exp(sat_nominal_fit$coefficients[2, 2] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
ci_8 = round(exp(sat_nominal_fit$coefficients[2, 3] + c(qnorm(0.025),-qnorm(0.025))* sat_nominal_fit$sta
```

Here are our two models and the interpretations of intercepts and coefficients of each:

First Model:

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

$\beta_{02} = -0.107$: The log odds of medium satisfaction versus low satisfaction between peoplewith low contact who live in tower is -0.107 (Equivalently, odds = 0.898 with $95\%CI = (0.666, 1.211)$).

$\beta_{12} = 0.296$: The log odds ratio of medium satisfaction versus low satisfaction between people with low contact and high contact is 0.296, keeping their housing condition constant (Equivalently, odds ratio = 1.344 with $95\%CI = (1.042, 1.735)$).

$\beta_{22} = -0.407$: The log odds ratio of medium satisfaction versus low satisfaction between people with live in apartment and people who live in tower is -0.407, keeping their contact level constant (Equivalently, odds ratio = 0.665 with $95\%CI = (0.476, 0.931)$).

$\beta_{32} = -0.337$: The log odds ratio of medium satisfaction versus low satisfaction between people with live in house and people who live in tower is -0.337, keeping their contact level constant (Equivalently, odds ratio $= 0.714$ with $95\%CI = (0.501, 1.017)$).

Second Model:

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

$\beta_{03} = 0.561$: The log odds of high satisfaction versus low satisfaction between people with low contact who live in tower is 0.561 (Equivalently, odds $= 1.752$ with $95\%CI = (1.35, 2.273)$).

$\beta_{13} = 0.328$: The log odds ratio of high satisfaction versus low satisfaction between people with low contact and high contact is 0.328, keeping their housing condition constant (Equivalently, odds ratio $= 1.388$ with $95\%CI = (1.101, 1.75)$).

$\beta_{23} = -0.642$: The log odds ratio of high satisfaction versus low satisfaction between people with live in apartment and people who live in tower is -0.642, keeping their contact level constant (Equivalently, odds ratio $= 0.526$ with $95\%CI = (0.392, 0.706)$).

$\beta_{33} = -0.946$: The log odds ratio of high satisfaction versus low satisfaction between people with live in house and people who live in tower is -0.946, keeping their contact level constant (Equivalently, odds ratio $= 0.388$ with $95\%CI = (0.281, 0.536)$).

Analyzing goodness-of-fit

```r
# computing pearson residuals
pihat = predict(sat_nominal,type='probs')
m = rowSums(nominal_housing_data[,3:5])
res_pearson=(nominal_housing_data[,3:5] - pihat * m)/sqrt(pihat * m)

G_stat = sum (res_pearson ^ 2) # computing G-stat

pval = 1-pchisq (G_stat, df = (6-4)*(3-1)) # pvalue >>> 0.1395076, fail to reject null >> fit is good

D_stat = sum(2*nominal_housing_data[,3:5]*log(nominal_housing_data[,3:5]/(m*pihat))) # deviance
```

It can be observed that **Generalized Pearson** χ^2 is 6.932 with **p-value** of 0.14, and the **Deviance statistics** is 6.893 which show that the model fits the data well.

Part iii

In this part, I am going to fit a proportional odds model to the housing data because the response is ordinal.

```r
sat_ordinal = polr(satisfaction ~ housing_type + contact, data = housing_data, weights = n)
summary(sat_ordinal)
```

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

## Call:
## polr(formula = satisfaction ~ housing_type + contact, data = housing_data,
##     weights = n)
##
## Coefficients:
##                            Value Std. Error t value
## housing_typeapartment -0.5009    0.11675  -4.291
## housing_typehouse     -0.7362    0.12610  -5.838
## contacthigh_cont       0.2524    0.09306   2.713
##
## Intercepts:
##                     Value   Std. Error t value
## low_sat|medium_sat  -0.9973  0.1075    -9.2794
## medium_sat|high_sat  0.1152  0.1047     1.1004
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

The ordinal models will look like this:

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

Or:

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

In this problem we have the same references and x variables: x1: the indicator of high contact

x2: the indicator of apartment

x3: indicator of house

$\beta_1 = -0.2524$ which is the log odds ratio of falling in lower category versus higher category for people with high contact versus low, holding their houisng condition constant.

$\beta_2 = 0.5009$ which is the log odds ratio of falling in lower category versus higher category between people who live in apartment versus in tower, holding their contact rate constant.

$\beta_3 = 0.7362$ which is the log odds ratio of falling in lower category versus higher category between people who live in house versus in tower, holding their contact rate constant.

We can observe that using proportinal odds model is easier than nominal model since we can interpret coefficients more easily by dividing them into lower and higher categories.

Looking at the coefficients, it can be observed that fixing housing condition, people who have higher contact are less likely to fall in the lower satisfaction groups. Also we can see that both groups of people who live in apartments and houses, are more likely to fall into lower categories of satisfaction compared to people who live in tower blocks, fixing their contact rate.

Part iv

In this part, the goal is to calculate Pearson residuals from the proportional odds model to identify the discrepancies between the observed frequencies and expected frequencies estimated from the model.

```
p_hat = predict(sat_ordinal, nominal_housing_data, type = 'p')

res_pearson2 = (nominal_housing_data[,3:5] - p_hat * m) / sqrt(p_hat * m)

cbind(housing = nominal_housing_data$housing_type, contact = nominal_housing_data$contact, res_pearson2)
  knitr::kable()
```

housing	contact	low_sat	medium_sat	high_sat
tower_block	low_cont	0.7793957	-0.3697193	-0.3151179
tower_block	high_cont	-0.9946852	0.4549302	0.3354430
apartment	low_cont	0.9177560	-1.0671823	-0.0152734
apartment	high_cont	-0.2369309	-0.4052334	0.5377735
house	low_cont	-1.1407855	0.1397563	1.2440771
house	high_cont	0.2743817	1.3677881	-1.4778270

We can observe that the highest value of Pearson residuals is -1.48. This shows that the highest discrepancies of data is observed when there is high satisfaction, high contact, and hous housing type. Also, the Pearson residuals representing people who live in houses and have low contacts and medium satisfaction is also high with the value of 1.368. The third highest residual is for people of high satisfaction who live in house and have low cantact. This shows that the model has not been as successful predictiong this portion of data and in general, we can observe some discrepencies in people who live in houses and have high satisfaction.