P8131_hw4

Shan Jiang

Problem 1 Copenhagen Residents sstisfactory survey

1. summarize the data

```
# import in the copenhagen data

copen_raw <- data.frame(res.low = c(65, 130, 67, 34, 141, 130),          res.med = c(54,
       res.hig = c(100, 111, 62, 100, 191, 104),
       contact = c(rep("low", 3), rep("high",3)),
       house = rep(c("Tower block", "Apartment", "House"), 2) )


copen_raw$house = factor(copen_raw$house,
                         levels = c("Tower block", "Apartment", "House"))
copen_raw$contact  = factor(copen_raw$contact,
                         levels = c("low", "high"))
```

Pair-wise Tables

```
## categorical tables by contact and house
copen.df1 = copen_raw %>%
  mutate(total = res.low + res.med + res.hig) %>%
  mutate(low.p =  percent(res.low / total),
         med.p =  percent(res.med / total),
         hig.p = percent(res.hig / total)) %>%
  dplyr::select(house, contact, res.low,
         low.p, res.med, med.p, res.hig, hig.p, total)

knitr::kable(copen.df1, caption = "")
```

house	contact	res.low	low.p	res.med	med.p	res.hig	hig.p	total
Tower block	low	65	29.68%	54	24.66%	100	45.66%	219
Apartment	low	130	41.01%	76	23.97%	111	35.02%	317
House	low	67	37.85%	48	27.12%	62	35.03%	177
Tower block	high	34	18.78%	47	25.97%	100	55.25%	181
Apartment	high	141	31.47%	116	25.89%	191	42.63%	448
House	high	130	38.35%	105	30.97%	104	30.68%	339

(1) Pair wise table of house

```
## Pair wise table of house
copen.df3 = copen.df1 %>%
  dplyr::select(-c(contact, low.p, med.p, hig.p)) %>%
  group_by(house) %>%
  summarise_each(funs(sum)) %>%
  mutate(low.p =  percent(res.low / total),
         med.p =  percent(res.med / total),
         hig.p = percent(res.hig / total))
```

```
## `summarise_each()` is deprecated.
## Use `summarise_all()`, `summarise_at()` or `summarise_if()` instead.
## To map `funs` over all variables, use `summarise_all()`
```

```
copen.df3 = copen.df3 %>%
 dplyr::select(house, res.low,
         low.p, res.med, med.p, res.hig, hig.p, total)
knitr::kable(copen.df3, caption = "")
```

house	res.low	low.p	res.med	med.p	res.hig	hig.p	total
Tower block	99	24.75%	101	25.25%	200	50.00%	400
Apartment	271	35.42%	192	25.10%	302	39.48%	765
House	197	38.18%	153	29.65%	166	32.17%	516

Based a single variate category analysis, it's clear that for Tower, people tend to have a higher satisfaction level while for the apartment and house type, the difference of satisfaction level is not that huge.

(2) By contact: Pair wise table

```
## Pair wise table of contact
copen.df2 = copen.df1 %>%
  dplyr::select(-c(house, low.p, med.p, hig.p)) %>%
  group_by(contact) %>%
  summarise_each(funs(sum)) %>%
  mutate(low.p =  percent(res.low / total),
         med.p =  percent(res.med / total),
         hig.p = percent(res.hig / total))
```

```
## `summarise_each()` is deprecated.
## Use `summarise_all()`, `summarise_at()` or `summarise_if()` instead.
## To map `funs` over all variables, use `summarise_all()`
```

```
copen.df2 = copen.df2 %>%
 dplyr::select(contact, res.low,
```

```
         low.p, res.med, med.p, res.hig, hig.p, total)
knitr::kable(copen.df2, caption = "")
```

contact	res.low	low.p	res.med	med.p	res.hig	hig.p	total
low	262	36.75%	178	24.96%	273	38.29%	713
high	305	31.51%	268	27.69%	395	40.81%	968

Based on the Pair wise table low and high category group seperation, high contact groups have a higher cluster in higher level satisfaction level instead of lower level.

From this table, we can find that there are three ordinal classes for satisfaction level, as we are curious about the causal relationship between degree of contact with other residents, housing conditions and satisfaction level, we can fit an ordinal logistic regression for modelling.

2.1 Nominal logistic Regression

The default reference level is the first row; The summary outputs two rows = two model results

Notation: *contact*: Reference level is 0; 0 is low, 1 is High;

```
## create an indicator variable for housing conditions
  library(nnet)
## Fit nominal model
copen.mult <- multinom(cbind(res.low, res.med, res.hig) ~
                     contact + house, data = copen_raw)
```

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

```
summary(copen.mult)
```

```
## Call:
## multinom(formula = cbind(res.low, res.med, res.hig) ~ contact +
##     house, data = copen_raw)
##
## Coefficients:
##         (Intercept) contacthigh houseApartment houseHouse
## res.med  -0.1072644   0.2959803     -0.4067537 -0.3370771
## res.hig   0.5607737   0.3282263     -0.6415967 -0.9456177
##
## Std. Errors:
```

```
##         (Intercept) contacthigh houseApartment houseHouse
## res.med   0.1524077   0.1301046      0.1713011  0.1803577
## res.hig   0.1329301   0.1181870      0.1500773  0.1644850
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

```
stargazer::stargazer(copen.mult, header = T )
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Tue, Mar 05, 2019 - 08:46:22

Table 4:

	Dependent variable:	
	res.med	res.hig
	(1)	(2)
contacthigh	0.296**	0.328***
	(0.130)	(0.118)
houseApartment	−0.407**	−0.642***
	(0.171)	(0.150)
houseHouse	−0.337*	−0.946***
	(0.180)	(0.164)
Constant	−0.107	0.561***
	(0.152)	(0.133)
Akaike Inf. Crit.	3,621.480	3,621.480
Note:	*p<0.1; **p<0.05; ***p<0.01	

Reference level is housing = tower;

$$ln(\frac{P(median)}{P(low)}) = \beta_{10} + \beta_{11}(contact = 1) + \beta_{12}(apartment = 1) + \beta_{13}(house = 1)$$

* β_{10}: The estimated log-odds for satisfactory level being median vs being low for residents with low contacts and lives in tower is 0.1887213.

- β_{11}: The estimated log-odds for satisfactory level being median vs being low for residents who have a high contact when keeping the housing the same is 0.2959.

- β_{12}: The estimated log-odds for satisfactory level being median versus vs being low for residents who live in an apartment vs tower is -0.4067.
- β_{13}: The estimated log-odds for satisfactory level being median vs being low for residents who live in a house vs tower is in the amount of -0.3371.

*For equation 2, the interpretation is similar:

$$ln(\frac{P(high)}{P(low)}) = \beta_{20} + \beta_{21}(contact = 1) + \beta_{22}(apartment = 1) + \beta_{23}(house = 1)$$

- β_{20}: The estimated log-odds for satisfactory level being high vs being low for residents with low contacts and lives in Tower is 0.561.
- β_{21}: The estimated log-odds for satisfactory level being high vs being low for residents who have a high contact keeping the housing the same is 0.328.
- β_{22}: The estimated log-odds for satisfactory level being high versus vs being low for residents who live in an apartment housing vs. tower in the amount of -0.946.
- β_{23}: The estimated log-odds for satisfactory level being high vs being low for residents who live in a house vs. tower in the amount of is -0.6416.

95% CI for odds ratio

```
## 95% confidence Interval of odds ratio
### model 1
### OR
exp(summary(copen.mult)$coefficients)[1,]
```

```
##    (Intercept)    contacthigh houseApartment     houseHouse
##      0.8982882      1.3444437      0.6658082      0.7138538
```

```
### lower bound:
exp(summary(copen.mult)$coefficients - 1.96 * summary(copen.mult)$standard.errors)[1,]
```

```
##    (Intercept)    contacthigh houseApartment     houseHouse
##      0.6663212      1.0418263      0.4759208      0.5012861
```

```
### higher bound:
exp(summary(copen.mult)$coefficients + 1.96 * summary(copen.mult)$standard.errors)[1,]
```

```
##    (Intercept)    contacthigh houseApartment     houseHouse
##      1.2110100      1.7349617      0.9314585      1.0165597
```

Model 1:

$$ln(\frac{P(median)}{P(low)}) = \beta_{10} + \beta_{11}(contact = 1) + \beta_{12}(apartment = 1) + \beta_{13}(house = 1)$$

- OR for contact: In this study, subjects with high contact 1.3443 times the odds of having median satisfaction vs low satisfaction compared to low-contact subjects.
- With 95% conficence the true odds ratio for median vs low satisfaction for high vs low contact lies in the range of [1.0418, 1.7349].
- OR for apartment: subjects who live in apartment have 0.6658 times the odds of having median satisfaction vs low satisfaction compared to those who live in Tower Block.
- We are 95% confident that the true odds ratio for median vs low satisfaction for house vs Tower Block lies in the range of [0.4759, 0.9315].
- OR for house: subjects who live in a house have 0.6658 times the odds of having median satisfaction vs low satisfaction compared to those who live in Tower Block.
- We are 95% confident that the true odds ratio for median vs low satisfaction for Apartment vs Tower Block lies in the range of [0.5013, 1.0166].

Model 2:

$$ln(\frac{P(high)}{P(low)}) = \beta_{20} + \beta_{21}(contact = 1) + \beta_{22}(apartment = 1) + \beta_{23}(house = 1)$$

- 95% CI for OR of model 2: high/low:

```
## model 2
### OR
exp(summary(copen.mult)$coefficients)[2,]
```

```
##    (Intercept)    contacthigh houseApartment     houseHouse
##      1.7520274      1.3885031      0.5264512      0.3884396
```

```
## lower bound
exp(summary(copen.mult)$coefficients - 1.96 * summary(copen.mult)$standard.errors)[2,]
```

```
##    (Intercept)    contacthigh houseApartment     houseHouse
##      1.3501703      1.1013975      0.3922923      0.2813915
```

```
## higher bound
exp(summary(copen.mult)$coefficients + 1.96 * summary(copen.mult)$standard.errors)[2,]
```

```
##    (Intercept)    contacthigh houseApartment     houseHouse
##      2.2734911      1.7504498      0.7064907      0.5362112
```

- OR for contact: In this study, subjects with high contact 1.3885 times the odds of having high satisfaction vs low satisfaction compared to low-contact subjects.
- With 95% conficence the true odds ratio for high vs low satisfaction for high vs low contact lies in the range of [1.1014, 1.7504].
- OR for Apartment: subjects who live in apartment have 0.5265 times the odds of having high satisfaction vs low satisfaction compared to those who live in Tower Block.

- We are 95% confident that the true odds ratio for high vs low satisfaction for house vs Tower Block lies in the range of [0.3923, 0.7065].
- OR for house: subjects who live in an house have 0.3884 times the odds of having high satisfaction vs low satisfaction compared to those who live in Tower Block.
- We are 95% confident that the true odds ratio for high vs low satisfaction for House vs Tower Block lies in the range of [0.2814, 0.5362].

Goodness of Fit

```
pihat = predict(copen.mult,type = 'probs')
## pihat: print out the pi(i, j) hats corresponding to the data obs.
m = rowSums(copen_raw[,1:3])
res.pearson = (copen_raw[,1:3] - pihat*m)/sqrt(pihat*m) # pearson residuals
res.pearson
```

```
##      res.low     res.med    res.hig
## 1  0.6461949  0.01462578 -0.4986673
## 2  0.3770499  0.08966220 -0.4647999
## 3 -1.0575696 -0.12653242  1.4047905
## 4 -0.8014370 -0.01553623  0.5247838
## 5 -0.3508946 -0.07196205  0.3670866
## 6  0.8402407  0.08673464 -0.9472127
```

```
# deviance
D.stat =  sum(2 * copen_raw[,1:3] * log(copen_raw[,1:3]/(m * pihat)))
D.stat
```

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

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

```
## [1] 6.932334
```

```
# deviance analysis
pval = 1 - pchisq(G.stat,df = (6 - 4) * (3 - 1))
# fit is good # not rejected, go with the smaller model
```

Based on the results, we can know that the model fits okay.

3. Proportional odds model

```
# fit proportional odds model
copen.polr = polr(res ~ contact + house,
```

```
                  data = copen.ord, weights = freq)

summary(copen.polr)
```

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

```
## Call:
## polr(formula = res ~ contact + house, data = copen.ord, weights = freq)
##
## Coefficients:
##                   Value Std. Error t value
## contacthigh      0.2524    0.09306   2.713
## houseApartment -0.5009    0.11675  -4.291
## houseHouse     -0.7362    0.12610  -5.838
##
## Intercepts:
##              Value   Std. Error t value
## low|median  -0.9973  0.1075    -9.2794
## median|high  0.1152  0.1047     1.1004
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

```
# pay attention to sign, read help doc of polr (-eta)
```

```
stargazer(copen.polr, type = 'latex', header = F )
```

Table 5:

	Dependent variable:
	res
contacthigh	0.252^{***}
	(0.093)
houseApartment	-0.501^{***}
	(0.117)
houseHouse	-0.736^{***}
	(0.126)
Observations	1,681
Note:	*p<0.1; **p<0.05; ***p<0.01

Output in R gives us negative results of linear predictors:

$$log(\frac{P(\pi_{low})}{P(\pi_{median}+\pi_{high})}) = -0.9973 - 0.2524 * Contact + 0.5009 * apartment + 0.7362 * house$$

$$log(\frac{P(\pi_{low}+\pi_{median})}{P(\pi_{high})}) = 0.1152 - 0.2524 * Contact + 0.5009 * apartment + 0.7362 * house$$

Model Interpretation: * Reference level is **apartment** for housing; * The two models have the same parameters for covariates while they have different intercepts;

- Log Odds ratio interpretation: The estimated log odds of prob. being low satisfaction over the prob. of (median + high) satisfaction [**or** prob. being low and median over the prob. of highly satisfied] is -0.2524 for high contact level vs. low contact;
- Log Odds ratio interpretation: The estimated log odds of prob. being low satisfaction over the prob. of (median + high) satisfaction [**or** prob. being low and median over the prob. of highly satisfied] for who lives in house vs tower is 0.7362;

*The log odds of in prob. being low satisfaction over the prob. of (median + high) satisfaction [**or** prob. being low and median over the prob. of highly satisfied] for for who lives in apartment vs. tower is 0.5009.

problem 4

```
# residuals for the dataset.
p_hat = predict(copen.polr, copen_raw, type = 'p')
m_2 = rowSums(cbind(copen_raw$res.low,
                    copen_raw$res.med,
                    copen_raw$res.hig))
res.pearson_2 = (copen_raw[,1:3] - p_hat*m_2)/sqrt(p_hat*m_2)

## Absolute value of res.pearson_2
abs(res.pearson_2)
```

```
##     res.low   res.med    res.hig
## 1 0.7793957 0.3697193 0.31511792
## 2 0.9177560 1.0671823 0.01527344
## 3 1.1407855 0.1397563 1.24407710
## 4 0.9946852 0.4549302 0.33544295
## 5 0.2369309 0.4052334 0.53777345
## 6 0.2743817 1.3677881 1.47782697
```

```
## The max is high contact level and housing type is house.
abs(res.pearson_2)[6,3]
```

```
## [1] 1.477827
```

From the model diagnostic results above, we can see that in the these cells, when housing type is **House** and contact level is **high**, this is the largest discrepancy that will be resulted in the model.