p8131_hw4_ps3194

Pangsibo Shen

2/19/2021

Contents

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

```
## Warning: package 'MASS' was built under R version 4.0.3

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
contact_level = c(rep("low", 3), rep("high",3))
house_type = rep(c("Tower_block", "Apartment", "House"), 2)
sat_low = c(65,130,67,34,141,130)
sat_med = c(54,76,48,47,116,105)
sat_high = c(100,111,62,100,191,104)

housing_df =
  tibble(contact_level,house_type,sat_low,sat_med,sat_high) %>%
```

```
  mutate(
    contact_level = as.factor(contact_level),
    house_type = as.factor(house_type)
         )


#Table of percentages showing the pair-wise associations between satisfaction and contact level
housing_df %>% dplyr::select(contact_level,sat_low,sat_med, sat_high) %>%
  tidyr::pivot_longer(cols = starts_with("sat"),
                names_to = "sat_level",
                names_prefix = "sat_",
                values_to = "freq") %>%
  group_by(contact_level,sat_level) %>%
  summarise(
    total = sum(freq)
  ) %>%
  tidyr::pivot_wider(
    names_from = sat_level,
    values_from = total,
    names_prefix = "sat_"
  ) %>%
  mutate(total_freq = sat_high + sat_low + sat_med,
         sat_low_pct = (sat_low/total_freq)*100,
         sat_med_pct = (sat_med/total_freq)*100,
         sat_high_pct = (sat_high/total_freq)*100
         ) %>%
  dplyr::select(contact_level,total_freq,sat_low_pct,sat_med_pct,sat_high_pct) %>%
  knitr::kable()
```

0.0.0.1 1)

```
## `summarise()` regrouping output by 'contact_level' (override with `.groups` argument)
```

contact_level	total_freq	sat_low_pct	sat_med_pct	sat_high_pct
high	968	31.50826	27.68595	40.80579
low	713	36.74614	24.96494	38.28892

```
#Table of percentages showing the pair-wise associations between satisfaction and house type
housing_df %>% dplyr::select(house_type,sat_low,sat_med, sat_high) %>%
  tidyr::pivot_longer(cols = starts_with("sat"),
                names_to = "sat_level",
                names_prefix = "sat_",
                values_to = "freq") %>%
  group_by(house_type,sat_level) %>%
  summarise(
    total = sum(freq)
  ) %>%
  tidyr::pivot_wider(
    names_from = sat_level,
    values_from = total,
    names_prefix = "sat_"
  ) %>%
  mutate(total_freq = sat_high + sat_low + sat_med,
```

```
        sat_low_pct = (sat_low/total_freq)*100,
        sat_med_pct = (sat_med/total_freq)*100,
        sat_high_pct = (sat_high/total_freq)*100
        ) %>%
  dplyr::select(house_type,total_freq,sat_low_pct,sat_med_pct,sat_high_pct) %>%
  knitr::kable()
```

```
## `summarise()` regrouping output by 'house_type' (override with `.groups` argument)
```

house_type	total_freq	sat_low_pct	sat_med_pct	sat_high_pct
Apartment	765	35.42484	25.09804	39.47712
House	516	38.17829	29.65116	32.17054
Tower_block	400	24.75000	25.25000	50.00000

From the table of percentages showing the pair-wise associations between satisfaction and contact level, we learned that among residents with high contact level, residents appear to have higher probability to have high satisfaction and among residents with low contact level, residents appear to have the lowest probability to have medium satisfaction.

From table of percentages showing the pair-wise associations between satisfaction and house type, we learned that among residents living in the apartments, residents appear to have higher probability to have high satisfaction;among residents living in the houses, residents appear to have the highest probability to have low satisfaction; among residents living in the Tower block, residents appear to have the highest probability to have high satisfaction.

```
housing_nominal = multinom(cbind(sat_low, sat_med, sat_high)~contact_level+house_type, data = housing_df
```

0.0.0.2 2)

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

```
# fit nominal logistic regression
summary(housing_nominal)
```

```
## Call:
## multinom(formula = cbind(sat_low, sat_med, sat_high) ~ contact_level +
##     house_type, data = housing_df)
##
## Coefficients:
##          (Intercept) contact_levellow house_typeHouse house_typeTower_block
## sat_med   -0.2180364       -0.2959832      0.06967922             0.4067631
## sat_high   0.2474047       -0.3282264     -0.30402275             0.6415948
##
## Std. Errors:
##          (Intercept) contact_levellow house_typeHouse house_typeTower_block
## sat_med   0.10930968        0.1301046       0.1437749             0.1713009
```

```
## sat_high  0.09783068        0.1181870       0.1351693             0.1500774
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

```
# exponentiate the coefficients for interpretations
exp(summary(housing_nominal)$coefficients)
```

```
##          (Intercept) contact_levellow house_typeHouse house_typeTower_block
## sat_med    0.8040962        0.7437999       1.0721642              1.501948
## sat_high   1.2806973        0.7201999       0.7378441              1.899508
```

The odds for having medium satisfaction among residents with high contact level and living in the apartments is 0.8040962 and The odds for having high satisfaction among residents with high contact level and living in the apartments is 1.2806973.

the odds ratio of low satisfaction vs medium satisfaction between residents with low contact level and residents with high contact level is 0.7437999; the odds ratio of low satisfaction vs medium satisfaction between residents living in the house and residents not living in the house is 1.0721642; the odds ratio of low satisfaction vs medium satisfaction between residents living in the tower block and residents not living in the tower block is 1.501948.

the odds ratio of low satisfaction vs high satisfaction between residents with low contact level and residents with high contact level is 0.7201999; the odds ratio of low satisfaction vs high satisfaction between residents living in the house and residents not living in the house is 0.7378441; the odds ratio of low satisfaction vs high satisfaction between residents living in the tower block and residents not living in the tower block is 1.899508.

```
#odds ratio with 95% confidence interval
coef_df =
  summary(housing_nominal)$coefficients

std_err_df=
  summary(housing_nominal)$standard.errors %>%
  as_tibble() %>%
  dplyr::select(-1) %>%
  tidyr::pivot_longer(cols = c(contact_levellow, house_typeHouse, house_typeTower_block),
                names_to = "covariates",
                values_to = "std_err") %>%
  dplyr::select(std_err)

or_ci =
  tibble(
    term = c("med_sat_contactLow","med_sat_typehouse","med_sat_typetowerblock","high_sat_contactLow","hi
    point_est = c(coef_df[3],coef_df[5],coef_df[7],coef_df[4],coef_df[6],coef_df[8]),
    std_err_df
  ) %>%
  mutate(
    exp_point_est = exp(point_est),
    lower_exp_ci = exp(point_est + qnorm(0.025)*std_err),
    higher_exp_ci = exp(point_est - qnorm(0.025)*std_err)
  )
or_ci %>%
  knitr::kable()
```

term	point_est	std_err	exp_point_est	lower_exp_ci	higher_exp_ci
med_sat_contactLow	-0.2959832	0.1301046	0.7437999	0.5763827	0.9598455
med_sat_typehouse	0.0696792	0.1437749	1.0721642	0.8088721	1.4211592
med_sat_typetowerblock	0.4067631	0.1713009	1.5019482	1.0736021	2.1011960
high_sat_contactLow	-0.3282264	0.1181870	0.7201999	0.5712840	0.9079335
high_sat_typehouse	-0.3040227	0.1351693	0.7378441	0.5661197	0.9616586
high_sat_typetowerblock	0.6415948	0.1500774	1.8995078	1.4154515	2.5491018

The table with exponential point estimates and confidence intervals for odds ratios is shown above.

```
# goodness of fit
pihat=predict(housing_nominal,type='probs')
m=rowSums(housing_df[,3:5])
res_pearson=(housing_df[,3:5]-pihat*m)/sqrt(pihat*m) # pearson residuals

G_stat=sum(res_pearson^2) # Generalized Pearson Chisq Stat
G_stat
```

```
## [1] 6.932341
```

```
pval=1-pchisq(G_stat,df = (6-4)*(3-1))
pval# fit is good
```

```
## [1] 0.1395072
```

Since the p-value for Generalized Pearson Chisq Statistic is greater than 0.05, we can conclude that our model is a good fit.

```
#create a dataframe in long format which can be used in proportional odds model
housing_df_long =
  housing_df %>%
  dplyr::select(contact_level,house_type,sat_low,sat_med, sat_high) %>%
  tidyr::pivot_longer(cols = starts_with("sat"),
                      names_to = "sat_level",
                      names_prefix = "sat_",
                      values_to = "freq") %>%
  mutate(sat_level = factor(sat_level,levels = c("low", "med", "high")))

# fit proportional odds model
housing_prop = polr(sat_level~contact_level+house_type,data = housing_df_long,weights = freq)
summary(housing_prop)
```

0.0.0.3 3

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

## Call:
## polr(formula = sat_level ~ contact_level + house_type, data = housing_df_long,
```

```
##     weights = freq)
##
## Coefficients:
##                          Value Std. Error t value
## contact_levellow      -0.2524    0.09306  -2.713
## house_typeHouse       -0.2353    0.10521  -2.236
## house_typeTower_block  0.5010    0.11675   4.291
##
## Intercepts:
##          Value   Std. Error t value
## low|med  -0.7488  0.0818    -9.1570
## med|high  0.3637  0.0801     4.5393
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

```
# exponentiate the coefficients for interpretations
exp(0.2524) #contact_levellow
```

```
## [1] 1.287111
```

```
exp(0.2353) #house_typeHouse
```

```
## [1] 1.265288
```

```
exp(-0.5010) #house_typeTower_block
```

```
## [1] 0.6059244
```

The fitted model told us that: the odds ratio of low satisfaction vs medium/high satisfaction between residents with low contact level and residents with high contact level is 1.2871108, while holding other variable fixed;the odds ratio of low satisfaction vs medium/high satisfaction between residents living in house and residents not living in house is 1.2652883, while holding other variable fixed;the odds ratio of low satisfaction vs medium/high satisfaction between living in tower block and residents not living in tower block is 0.6059244, while holding other variable fixed.

```
pi_prop = predict(housing_prop,housing_df, type = "p")
m_prop = rowSums(housing_df[,3:5])
res_pearson_prop = (housing_df[,3:5]-pi_prop*m_prop)/sqrt(pi_prop*m_prop)
res_pearson_prop %>%
  mutate(
    contact_level = c(rep("low", 3), rep("high",3)),
    house_type = rep(c("Tower_block", "Apartment", "House"), 2)
  ) %>%
  knitr::kable()
```

		sat_low	sat_med	sat_high	contact_level	house_type
		0.7794178	-0.3696760	-0.3151660	low	Tower_block
		0.9176690	-1.0671401	-0.0152261	low	Apartment
0.0.0.4	**4**	-1.1408528	0.1397992	1.2441278	low	House
		-0.9946598	0.4549796	0.3353921	high	Tower_block
		-0.2370150	-0.4051916	0.5378150	high	Apartment
		0.2742913	1.3678370	-1.4777786	high	House

The largest discrepancy between the observed frequencies and expected frequencies estimated from the model happened when the contact level is high, housing type is house, and satisfactory level is high. It has the largest pearson residual of -1.48.