HW4

Anna Ma

3/5/2022

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

Problem 1

```
prob1_df = tibble(
  contact = rep(rep(c("low","high"), each = 3),3),
  home_type = rep(c("tower_block", "apartment", "house"), each = 6),
  sat_level = rep(c("low satisfaction", "medium satisfaction", "high satisfaction"), 6),
  sat_value = c(c(65, 54, 100, 34, 47, 100),c(130, 76, 111, 141, 116, 191), c(67, 48, 62, 130, 105, 104
) %>%
  mutate(sat_level = factor(sat_level, levels = c("low satisfaction", "medium satisfaction", "high satis
         home_type = factor(home_type, levels = c("apartment","house","tower_block")),
         contact = factor(contact, levels = c("high","low")))
```

1. Association between satisfaction and contact with others

```
sat_contact = prob1_df %>%
  group_by(contact, sat_level) %>%
  summarize(n = sum(sat_value)) %>%
  group_by(contact) %>%
  mutate(n_total = sum(n),
         percentage = paste(round((n * 100 / n_total),3),"%")) %>%
  dplyr::select(-n_total, -n) %>%
  pivot_wider(names_from = sat_level, values_from = percentage)
```

```
## `summarise()` has grouped output by 'contact'. You can override using the `.groups` argument.
```

```
sat_contact %>% knitr::kable()
```

contact	low satisfaction	medium satisfaction	high satisfaction
high	31.508 %	27.686 %	40.806 %
low	36.746 %	24.965 %	38.289 %

From the table we can see that people who have high contact with others has a higher proportion in high satisfaction. On the other hand, the proportion of satisfaction for people who has low contact with others spreads in low, medium, and high. The low contact group has a higher proportion in low satisfaction and a lower proportion in high satisfaction compare to high contact group. Meanwhile, both group has its lowest proportion in medium satisfaction. But generally, it appears that the association between satisfaction and contact with others does not vary too much much.

2. Association between satisfaction and type of housing

```
sat_type = prob1_df %>%
  group_by(home_type, sat_level) %>%
  summarize(n = sum(sat_value)) %>%
  group_by(home_type) %>%
  mutate(n_total = sum(n),
         percentage = paste(round((n * 100 / n_total),3),"%")) %>%
  dplyr::select(-n_total, -n) %>%
  pivot_wider(names_from = sat_level, values_from = percentage)
```

```
## `summarise()` has grouped output by 'home_type'. You can override using the `.groups` argument.
```

```
sat_type %>% knitr::kable()
```

home_type	low satisfaction	medium satisfaction	high satisfaction
apartment	35.425 %	25.098 %	39.477 %
house	38.178 %	29.651 %	32.171 %
tower_block	24.75 %	25.25 %	50 %

From the table, we can see that the satisfaction level varies only slightly for those who lives in apartments and houses. However, tower block residents generally has a high satisfaction and a smaller proportion of them has low satisfaction compare to the other two home types.

Problem 2

1. Model

```
nom_data = prob1_df %>% pivot_wider(names_from = sat_level, values_from = sat_value)

nom_mod = multinom(cbind(nom_data$`low satisfaction`,
nom_data$`medium satisfaction`, nom_data$`high satisfaction`) ~ home_type + contact, data = nom_data)
```

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

```
nom_mod_result = summary(nom_mod)

nom_mod_result
```

```
## Call:
## multinom(formula = cbind(nom_data$`low satisfaction`, nom_data$`medium satisfaction`,
##     nom_data$`high satisfaction`) ~ home_type + contact, data = nom_data)
##
## Coefficients:
##   (Intercept) home_typehouse home_typetower_block contactlow
## 2  -0.2180364     0.06967922            0.4067631 -0.2959832
## 3   0.2474047    -0.30402275            0.6415948 -0.3282264
##
## Std. Errors:
```

```
##   (Intercept) home_typehouse home_typetower_block contactlow
## 2  0.10930968      0.1437749            0.1713009  0.1301046
## 3  0.09783068      0.1351693            0.1500774  0.1181870
##
## Residual Deviance: 3605.48
## AIC: 3621.48
```

The multinomial model is

$log(\frac{\pi_2(X)}{\pi_1(X)}) = \beta_2 + \beta_{21}(HomeType = House) + \beta_{22}(HomeType = TowerBlock) + \beta_{23}(Contact = Low) = -0.218 + 0.0697x_h + 0.407x_t - 0.296x_l$

$log(\frac{\pi_3(X)}{\pi_1(X)}) = \beta_3 + \beta_{31}(HomeType = House) + \beta_{32}(HomeType = TowerBlock) + \beta_{33}(Contact = Low) = 0.247 - 0.304x_h + 0.642x_t - 0.328x_l$.

1 = low, 2 = medium, 3 = high

2. Odds ratios and 95% confidence interval

```
nom_mod %>%
  broom::tidy() %>%
  filter(term != '(Intercept)') %>%
  mutate(odds_ratio =exp(estimate),
         Lower_bound = exp(estimate + qnorm(0.025)*std.error),
         Higher_bound = exp(estimate - qnorm(0.025)*std.error)) %>%
  dplyr::select(y.level,term, odds_ratio, Lower_bound, Higher_bound)
```

```
## # A tibble: 6 x 5
##   y.level term                 odds_ratio Lower_bound Higher_bound
##   <chr>   <chr>                     <dbl>       <dbl>        <dbl>
## 1 2       home_typehouse            1.07        0.809        1.42
## 2 2       home_typetower_block      1.50        1.07         2.10
## 3 2       contactlow                0.744       0.576        0.960
## 4 3       home_typehouse            0.738       0.566        0.962
## 5 3       home_typetower_block      1.90        1.42         2.55
## 6 3       contactlow                0.720       0.571        0.908
```

From this result, we can conclude that given a reference group of people who lives in an apartment and has high contact with others:

- people who lives in a house are 1.07 times more likely to have a medium satisfaction other than low satisfaction with a confidence interval of (0.809,1.42)
- people who lives in a tower block are 1.5 times more likely to have a medium satisfaction other than low satisfaction with a confidence interval of (1.07,2.1)
- people who have low contact with others are 0.744 times likely to have a medium satisfaction other than low satisfaction with a confidence interval of (0.576, 0.96)

Given a reference group of people who lives in an apartment and has high contact with others:

- people who lives in a house are 0.738 times more likely to have a high satisfaction other than low satisfaction with a confidence interval of (0.566,0.962)
- people who lives in a tower block are 1.9 times more likely to have a high satisfaction other than low satisfaction with a confidence interval of (1.42,2.55)
- people who have a low contact with others are 0.72 times likely to have a high satisfaction other than low satisfaction with a confidence interval of (0.571,0.908)

3. Goodness of Fit

```
pihat = predict(nom_mod, type = 'probs')
m = rowSums(nom_data[,3:5])
res.pearson = (nom_data[,3:5] - pihat*m)/sqrt(pihat*m)

G.stat = sum(res.pearson^2)
pval.G = 1 - pchisq(G.stat,df = (6-4)*(3-1))

D.stat = sum(2*nom_data[,3:5] * log(nom_data[,3:5]/(m*pihat)))
pval.D = 1 - pchisq(D.stat, df = (6 - 4) * (3 - 1))
```

The pvalue we got from Pearson chi-square analysis is 0.14, and the p value we got from Deviance analysis is 0.142. Since both p values are larger than 0.05, we failed to reject the null hypothesis. We can conclude that the model fits the data well.

4. Interaction $m_{int} = log(\frac{\pi_j(X)}{\pi_1(X)}) = \beta_j + \beta_{j1}(HomeType = House) + \beta_{j2}(HomeType = TowerBlock) + \beta_{j3}(Contact = Low) + \beta_{j4}(Contact = Low * HomeType = House) + \beta_{j5}(Contact = Low * HomeType = TowerBlock)$

$H_0 : \beta_{j4} = \beta_{j5} = 0, j = 2, 3$

```
nom_mod_int = multinom(cbind(nom_data$`low satisfaction`,
nom_data$`medium satisfaction`, nom_data$`high satisfaction`) ~ home_type + contact + home_type * conta
```

```
## # weights:  21 (12 variable)
## initial  value 1846.767257
## iter  10 value 1800.128659
## final  value 1799.293647
## converged
```

```
dev0 = nom_mod$deviance
dev1 = nom_mod_int$deviance
diff = dev0 - dev1
p = pchisq(diff, 4, lower.tail = FALSE)
```

Using the deviance analysis, we see that the difference between the models are 6.8930277, and the p value is 0.1416504, which is larger than 0.05. Therefore, we fail to reject the null hypothesis and conclude that there is interaction between contact and home types.

Problem 3

1. Fit the model

```
ord_mod = polr(sat_level ~ home_type + contact, data = prob1_df, weights = sat_value)

ord_mod_result = summary(ord_mod)
```

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

```
ord_mod_result
```

```
## Call:
## polr(formula = sat_level ~ home_type + contact, data = prob1_df,
##     weights = sat_value)
##
## Coefficients:
##                          Value Std. Error t value
## home_typehouse       -0.2353    0.10521  -2.236
## home_typetower_block  0.5010    0.11675   4.291
## contactlow           -0.2524    0.09306  -2.713
##
## Intercepts:
##                                         Value   Std. Error t value
## low satisfaction|medium satisfaction  -0.7488  0.0818    -9.1570
## medium satisfaction|high satisfaction  0.3637  0.0801     4.5393
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

The ordinal model is

$log(\frac{\pi_1(X)}{\pi_2(X)+\pi_3(X)}) = \beta_1 + \beta_{11}(HomeType = House) + \beta_{12}(HomeType = TowerBlock) + \beta_{13}(Contact = Low) = -0.7488 - 0.2353x_h + 0.501x_t - 0.2524x_l$

$log(\frac{\pi_1(X)+\pi_2(X)}{\pi_3(X)}) = \beta_2 + \beta_{21}(HomeType = House) + \beta_{22}(HomeType = TowerBlock) + \beta_{23}(Contact = Low) = -0.3637 - 0.2353x_h + 0.501x_t - 0.2524x_l$

1 = low, 2 = medium, 3 = high

2. Goodness of Fit

```
ord_pihat = predict(ord_mod, nom_data, type = 'probs')
ord_m = rowSums(nom_data[,3:5])
ord_res_pearson = (nom_data[,3:5] - ord_pihat*ord_m)/sqrt(ord_pihat*ord_m)

ord_G.stat = sum(ord_res_pearson^2)
ord_pval.G = 1 - pchisq(ord_G.stat,df = 7)

ord_D.stat = sum(2*nom_data[,3:5] * log(nom_data[,3:5]/(ord_m*ord_pihat)))
ord_pval.D = 1 - pchisq(ord_D.stat, df = 7)
```

The pvalue we got from Pearson chi-square analysis is 0.113, and the p value we got from Deviance analysis is 0.111. Since both p values are larger than 0.05, we failed to reject the null hypothesis. We can conclude that the model fits the data well.

3. Estimations and CIs

```
exp(cbind(coef(ord_mod), confint(ord_mod)))
```

```
## Waiting for profiling to be done...

##
## Re-fitting to get Hessian

##                                 2.5 %    97.5 %
## home_typehouse       0.7903394 0.6429196 0.9711892
## home_typetower_block 1.6502997 1.3136017 2.0762957
## contactlow           0.7769052 0.6472271 0.9321964
```

From this result, we can conclude that holding the contact level constant, comparing to those who lives in apartments:

- the odds of high satisfaction is 0.79 times the odds of low or medium satisfaction for residents who lives in a house with an confidence interval of (0.64,0.97)
- the odds of high satisfaction is 1.65 times the odds of low or medium satisfaction for residents who lives in a tower block with an confidence interval of (1.31,2.08)

Holding the home-type constant:

- the odds of high satisfaction is 0.78 times the odds of low or medium satisfaction for residents who has low contact with others with an confidence interval of (0.64,0.93)

Problem 4

1. Pearson Residuals

```
ord_pihat = predict(ord_mod, nom_data, type = 'probs')
ord_m = rowSums(nom_data[,3:5])
ord_res_pearson = (nom_data[,3:5] - ord_pihat*ord_m)/sqrt(ord_pihat*ord_m)

cbind(home_type = nom_data$home_type, contact = nom_data$contact, ord_res_pearson) %>% knitr::kable(dig
```

home_type	contact	low satisfaction	medium satisfaction	high satisfaction
tower_block	low	0.779	-0.370	-0.315
tower_block	high	-0.995	0.455	0.335
apartment	low	0.918	-1.067	-0.015
apartment	high	-0.237	-0.405	0.538
house	low	-1.141	0.140	1.244
house	high	0.274	1.368	-1.478

2. Largest discrepancy

From the table above we can see that the largest discrepancy is when home type is house, contact level is high, and satisfaction level is high, the residual is -1.478