p8105_hw3_mp3745

Matthew Parker

2020-03-02

Problem 1

i)

Enter in data

```r
housing_df = tibble(
  contact = c(rep("low", 3), rep("high", 3)),
  home_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)
)
```

Table to compare satisfaction with contact level of residents

```r
# Table
sat_level_contact = housing_df %>%
  dplyr::select(-home_type) %>%
  group_by(contact) %>%
  summarize(
    sat_low = sum(sat_low),
    sat_med = sum(sat_med),
    sat_high = sum(sat_high),
    total = sum(sat_low, sat_med, sat_high),
    sat_low_perc = round((sat_low * 100 / total), 2),
    sat_med_perc = round((sat_med * 100 / total), 2),
    sat_high_perc = round((sat_high * 100 / total), 2)
  ) %>%
  dplyr::select(contact, sat_low, sat_low_perc, sat_med, sat_med_perc, sat_high, sat_high_perc, total)

# View
sat_level_contact %>%
  knitr::kable()
```

contact	sat_low	sat_low_perc	sat_med	sat_med_perc	sat_high	sat_high_perc	total
high	305	31.51	268	27.69	395	40.81	968
low	262	36.75	178	24.96	273	38.29	713

From the above table, we can see that they may be a slight association with degree of contact with other residents and their satisfaction. Of those with a high degree of contact with other residents 40.81% have a high level of satisfaction, whereas among those with a low degree of contact, 38.29% have a high level of satisfaction. Of those with a high degree of contact with other residents 31.51% have a low level of satisfaction, whereas among those with a low degree of contact, 36.75% have a low level of satisfaction.

Table to compare satisfaction with type of housing

```r
# Table
sat_level_housing = housing_df %>%
  dplyr::select(-contact) %>%
  group_by(home_type) %>%
  summarize(
    sat_low = sum(sat_low),
    sat_med = sum(sat_med),
    sat_high = sum(sat_high),
    total = sum(sat_low, sat_med, sat_high),
    sat_low_perc = round((sat_low * 100 / total), 2),
    sat_med_perc = round((sat_med * 100 / total), 2),
    sat_high_perc = round((sat_high * 100 / total), 2)
  ) %>%
  dplyr::select(home_type, sat_low, sat_low_perc, sat_med, sat_med_perc, sat_high, sat_high_perc, total

# View
sat_level_housing %>%
  knitr::kable()
```

home_type	sat_low	sat_low_perc	sat_med	sat_med_perc	sat_high	sat_high_perc	total
apartment	271	35.42	192	25.10	302	39.48	765
house	197	38.18	153	29.65	166	32.17	516
tower_block	99	24.75	101	25.25	200	50.00	400

From the above table, we can see that they may be an association with housing type and residents satisfaction. Of those with a tower block 50% have a high level of satisfaction. Among those with a house, 32.17% have a high level of satisfaction. And among those with an apartment, 39.48% have a high level of satisfaction. Of those with a tower block 24.75% have a low level of satisfaction. Among those with a house, 38.18% have a low level of satisfaction. And among those with an apartment, 35.42% have a low level of satisfaction.

ii)

Fit nominal logistic regression model

```r
# Response matrix
resp = housing_df %>%
  dplyr::select(sat_low, sat_med, sat_high) %>%
  as.matrix()

# Nominal model
housing_fit_nom = multinom(resp ~ contact + home_type, data = housing_df)
```

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

```r
# View summary
summary(housing_fit_nom)
```

```
## Call:
## multinom(formula = resp ~ contact + home_type, data = housing_df)
```

```
##
## Coefficients:
##          (Intercept) contactlow home_typehouse home_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) contactlow home_typehouse home_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
```

Check goodness of fit

```r
pihat = predict(housing_fit_nom, type = 'probs')
m = rowSums(housing_df[,3:5])

# Pearson residuals
res_pearson = (housing_df[,3:5] - pihat * m) / sqrt(pihat * m)

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

# P-value
pval = 1 - pchisq(g_stat, df = (6 - 4) * (3 - 1))
```

Since $0.1395072 > 0.05$, we fail to reject our null hypothesis that the model does a good job fitting the data.

Based on the signs of the coefficients in the above model, it appears there is a negative association between both low contact and renting a house with high satisfaction vs low satisfaction. There is a positive association between renting a tower block with high satisfaction vs low satisfaction. There is a negative association between low contact with medium satisfaction vs low satisfaction. There may be a slight positive association between renting a house and medium satisfaction vs low satisfaction. And there is a positive association between renting a tower block and medium satisfaction vs low satisfaction.

Odds ratios with 95% CIs

```r
# ORs with 95% CIs
or_ci = cbind(summary(housing_fit_nom)$coefficients, summary(housing_fit_nom)$standard.errors) %>%
  as_tibble() %>%
  janitor::clean_names() %>%
  rename(
    coef_int = intercept,
    coef_contact_low = contactlow,
    coef_home_type_house = home_typehouse,
    coef_home_type_tower_block = home_typetower_block,
    std_err_int = v5,
    std_err_contact_low = v6,
    std_err_home_type_house = v7,
    std_err_home_type_tower_block = v8
  ) %>%
  mutate(
    model = c("sat_med", "sat_high")
  ) %>%
  dplyr::select(model, coef_int:std_err_home_type_tower_block) %>%
  mutate(
    int_lower = coef_int - std_err_int,
    int_higher = coef_int + std_err_int,
    contact_low_lower = coef_contact_low - std_err_contact_low,
    contact_low_higher = coef_contact_low + std_err_contact_low,
    home_type_house_lower = coef_home_type_house - std_err_home_type_house,
    home_type_house_higher = coef_home_type_house + std_err_home_type_house,
    home_type_tower_block_lower = coef_home_type_tower_block - std_err_home_type_tower_block,
    home_type_tower_block_higher = coef_home_type_tower_block + std_err_home_type_tower_block,
  ) %>%
  mutate(
    exp_int = exp(coef_int),
    exp_int_lower = exp(int_lower),
    exp_int_higher = exp(int_higher),
    exp_contact_low = exp(coef_contact_low),
    exp_contact_low_lower = exp(contact_low_lower),
    exp_contact_low_higher = exp(contact_low_higher),
    exp_home_type_house = exp(coef_home_type_house),
    exp_home_type_house_lower = exp(home_type_house_lower),
    exp_home_type_house_higher = exp(home_type_house_higher),
    exp_home_type_tower_block = exp(coef_home_type_tower_block),
    exp_home_type_tower_block_lower = exp(home_type_tower_block_lower),
    exp_home_type_tower_block_higher = exp(home_type_tower_block_higher)
  ) %>%
  mutate(
    exp_contact_low_ci =
      str_c(round(exp_contact_low, 2),
            " (", round(exp_contact_low_lower, 2), ", ", round(exp_contact_low_higher, 2), ")"),
    exp_home_type_house_ci =
      str_c(round(exp_home_type_house, 2),
            " (", round(exp_home_type_house_lower, 2), ", ", round(exp_home_type_house_higher, 2), ")")
    exp_home_type_tower_block =
      str_c(round(exp_home_type_tower_block, 2),
            " (", round(exp_home_type_tower_block_lower, 2), ", ", round(exp_home_type_tower_block_high
  ) %>%
  dplyr::select(model, exp_contact_low_ci, exp_home_type_house_ci, exp_home_type_tower_block)
```

The odds ratio (with 95% CI) between medium satisfaction and low satisfaction for:

- degree of contact low vs high is 0.74 (0.65, 0.85)
- home type house vs home type apartment is 1.07 (0.93, 1.24)
- home type tower block vs home type aparment is 1.5 (1.27, 1.78)

The odds ratio (with 95% CI) between high satisfaction and low satisfaction for:

- degree of contact low vs high is 0.72 (0.64, 0.81)
- home type house vs home type apartment is 0.74 (0.64, 0.84)
- home type tower block vs home type aparment is 1.9 (1.63, 2.21)

iii)

Put data frame together

```
freq = c(housing_df$sat_low, housing_df$sat_med, housing_df$sat_high)

housing_ord = tibble(
  res = c(rep(c("sat_low", "sat_med", "sat_high"), c(6, 6, 6))),
  contact = rep(housing_df$contact, 3),
  home_type = rep(housing_df$home_type, 3),
  freq = freq
) %>%
  mutate(
    res = factor(res, levels = c("sat_low", "sat_med", "sat_high"), ordered = TRUE)
  )
```

Fit proportional odds model

```
# Fit
housing_polr = polr(res ~ contact + home_type, data = housing_ord, weights = freq)

# Summary
summary(housing_polr)
```

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

## Call:
## polr(formula = res ~ contact + home_type, data = housing_ord,
##     weights = freq)
##
## Coefficients:
##                        Value Std. Error t value
## contactlow           -0.2524    0.09306  -2.713
## home_typehouse       -0.2353    0.10521  -2.236
## home_typetower_block  0.5010    0.11675   4.291
##
## Intercepts:
##                  Value   Std. Error t value
## sat_low|sat_med  -0.7488  0.0818    -9.1570
## sat_med|sat_high  0.3637  0.0801     4.5393
##
## Residual Deviance: 3610.286
## AIC: 3620.286
```

From the results of the proportional odds model, based on the signs of the coefficients, we can tell that a low degree of contact with residents is associated with a lower level of satisfaction compared to a high degree of contact. In addition, renting a house is associated with a lower level of satisfaction compared to renting an apartment. Finally, renting a tower block is associated with a higher level of satisfaction compared to renting an apartment.

iv)

Calculate Pearson residuals

```
pihat = predict(housing_polr, housing_df, type = 'p')
m = rowSums(cbind(housing_df$sat_low, housing_df$sat_med, housing_df$sat_high))
res_pearson = (housing_df[,3:5] - pihat * m) / sqrt(pihat * m)

# table of pearson residuals
```

```r
res_pearson_table = as_tibble(res_pearson) %>%
  mutate(
    contact = housing_df$contact,
    home_type = housing_df$home_type
  ) %>%
  dplyr::select(contact, home_type, sat_low:sat_high)

# View table
res_pearson_table %>%
  knitr::kable()
```

contact	home_type	sat_low	sat_med	sat_high
low	tower_block	0.7794178	-0.3696760	-0.3151660
low	apartment	0.9176690	-1.0671401	-0.0152261
low	house	-1.1408527	0.1397992	1.2441278
high	tower_block	-0.9946598	0.4549796	0.3353921
high	apartment	-0.2370150	-0.4051916	0.5378150
high	house	0.2742913	1.3678370	-1.4777786

The above table show the pearson residuals from the proportional odds model. The largest discrepancies are for:

- contact high, home type house, satisfaction high (pearson_residual = -1.48)
- contact high, home type house, satisfaction medium (pearson_residual = 1.37)
- contact low, home type house, satisfaction high (pearson_residual = 1.24)
- contact low, home type house, satisfaction low (pearson_residual = -1.14)