HW4_answer

Guojing Wu

2/28/2019

Problem

Read data:

```
data_copen = tibble(type = rep(c("Tower", "Apartment", "House"), each = 6),
                    contact = rep(rep(c("Low", "High"), each = 3), 3),
                    satisfaction = rep(c("Low satisfaction", "Medium satisfaction", "High satisfaction")
                    n = c(c(65, 54, 100, 34, 47, 100),
                          c(130, 76, 111, 141, 116, 191),
                          c(67, 48, 62, 130, 105, 104))) %>%
  mutate(contact = factor(contact, levels = c("Low", "High")),
         type = factor(type, levels = c("Tower", "Apartment", "House")),
         satisfaction = factor(satisfaction, levels = c("Low satisfaction", "Medium satisfaction", "High
```

i) Summarize the data

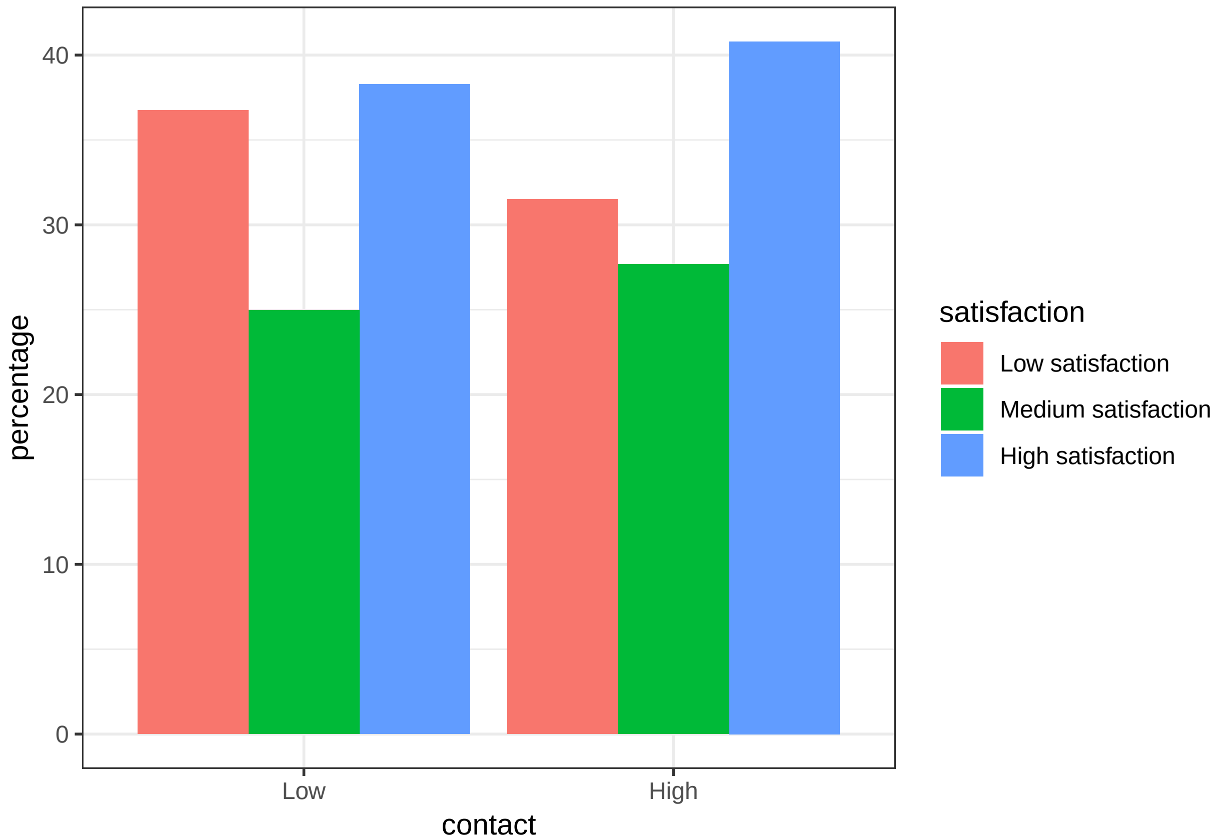
1) association between satisfaction and contact

```
data_SC = data_copen %>%
  group_by(contact, satisfaction) %>%
  summarise(n = sum(n)) %>%
  group_by(contact) %>%
  mutate(n_total = sum(n),
         percentage = n * 100 / n_total) %>%
  select(-n_total, -n)

data_SC %>%
  spread(key = satisfaction, value = percentage) %>% knitr::kable()
```

contact	Low satisfaction	Medium satisfaction	High satisfaction
Low	36.74614	24.96494	38.28892
High	31.50826	27.68595	40.80579

```
data_SC %>%
  ggplot(aes(x = contact, y = percentage, fill = satisfaction)) +
  geom_bar(stat = "identity", position = position_dodge())
```



From the table and barplot, we can see that 'Low' contact is associated with more 'Low satisfaction', while 'High' contact is associated with more 'Medium satisfaction' and 'High satisfaction'.

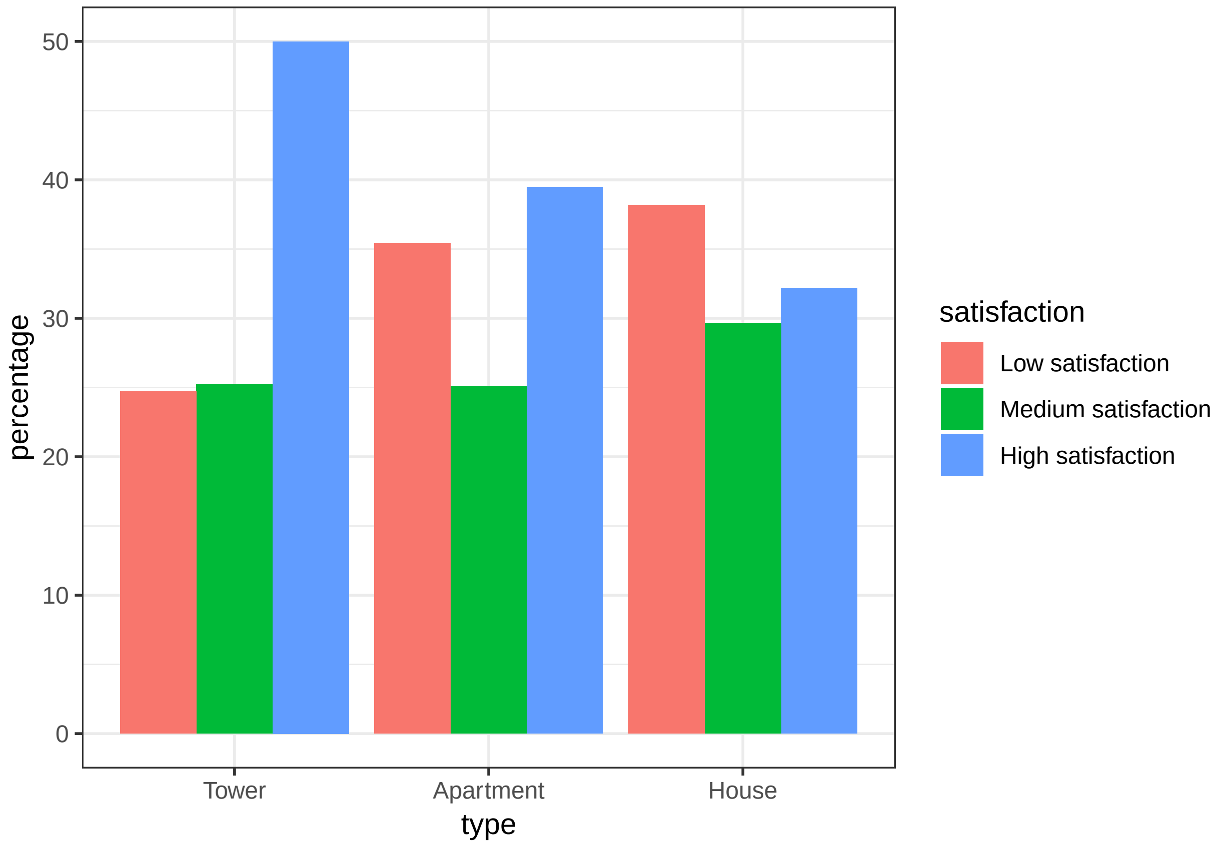
2) association between satisfaction and type of housing

```
data_ST = data_copen %>%
  group_by(type, satisfaction) %>%
  summarise(n = sum(n)) %>%
  group_by(type) %>%
  mutate(n_total = sum(n),
         percentage = n * 100 / n_total) %>%
  select(-n_total, -n)

data_ST %>%
  spread(key = satisfaction, value = percentage) %>% knitr::kable()
```

type	Low satisfaction	Medium satisfaction	High satisfaction
Tower	24.75000	25.25000	50.00000
Apartment	35.42484	25.09804	39.47712
House	38.17829	29.65116	32.17054

```
data_ST %>%
  ggplot(aes(x = type, y = percentage, fill = satisfaction)) +
  geom_bar(stat = "identity", position=position_dodge())
```



From the table and barplot, we can see that 'Tower' is associated with more 'High satisfaction', while 'House' is associated with more 'Low satisfaction' and 'Medium satisfaction'.

ii) Nomial logistic regression

We use multinomial model to fit the data:

- the reference response is 'Low satisfaction'.
- the reference housing type is 'Tower'.
- the reference contact is 'Low'.

```
data_nom = data_copen %>%
  spread(key = satisfaction, value = n)

fit.mult = multinom(cbind(`Low satisfaction`, `Medium satisfaction`, `High satisfaction`) ~ type + conta
```

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

```
res.mult = summary(fit.mult)
res.odds = tibble("type=Apartment" = rep(0,2),
                  "type=House" = rep(0,2),
                  "contact=High" = rep(0,2))
rownames(res.odds) = c("Medium satisfaction", "High satisfaction")
```

```r
for (i in 1:nrow(res.odds)) {
  for (j in 1:ncol(res.odds)) {
    res.odds[i,j] = paste(round(exp(res.mult$coefficients[i,j+1]), 3),
                         ", CI = (",
                         round(exp(res.mult$coefficients[i,j+1] + qnorm(0.025) * res.mult$standard.erro
                         ", ",
                         round(exp(res.mult$coefficients[i,j+1] - qnorm(0.025) * res.mult$standard.erro
                         ")", sep = "")
  }
}

res.odds %>% knitr::kable()
```

	type=Apartment	type=House	contact=High
Medium satisfaction	0.666, CI = (0.476, 0.931)	0.714, CI = (0.501, 1.017)	1.344, CI = (1.042, 1.735)
High satisfaction	0.526, CI = (0.392, 0.706)	0.388, CI = (0.281, 0.536)	1.389, CI = (1.101, 1.75)

From the odds ratio table above, we could interpret that:

- The odds ratio between number of Medium satisfaction and number of Low satisfaction is 0.666 given housing type change from Tower to Apartment.
- The odds ratio between number of Medium satisfaction and number of Low satisfaction is 0.714 given housing type change from Tower to House.
- The odds ratio between number of Medium satisfaction and number of Low satisfaction is 1.344 given contact change from Low to High.
- The odds ratio between number of High satisfaction and number of Low satisfaction is 0.526 given housing type change from Tower to Apartment.
- The odds ratio between number of High satisfaction and number of Low satisfaction is 0.388 given housing type change from Tower to House.
- The odds ratio between number of High satisfaction and number of Low satisfaction is 1.389 given contact change from Low to High.

```r
pihat = predict(fit.mult, type = 'probs')
m = rowSums(data_nom[,3:5])
res.pearson = (data_nom[,3:5] - pihat * m) / sqrt(pihat * m) # pearson residuals

G.stat = sum(res.pearson ^ 2) # Generalized Pearson Chisq Stat
pval.G = 1 - pchisq(G.stat, df = (6 - 4) * (3 - 1)) # n = 6, p = 4, J = 3

D.stat = sum(2 * data_nom[,3:5] * log(data_nom[,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
- The pvalue we got from Deviance analysis is 0.142

which all shows that we failed to reject the null hypothesis, meaning these isn't much of a difference between this model and the full model, so the model fits the data well.

iii) Ordinal logistic regression

We use proportional odds model to fit the data:

- the reference housing type is 'Tower'.
- the reference contact is 'Low'.

```
fit.ord = polr(satisfaction ~ type + contact, data = data_copen, weights = n)

res.ord = summary(fit.ord)
res.ord$coefficients %>% knitr::kable()
```

	Value	Std. Error	t value
typeApartment	-0.5009409	0.1167538	-4.290575
typeHouse	-0.7362314	0.1261027	-5.838347
contactHigh	0.2524351	0.0930579	2.712667
Low satisfaction|Medium satisfaction	-0.9973417	0.1074788	-9.279429
Medium satisfaction|High satisfaction	0.1151734	0.1046627	1.100424

From the estimated β_p above, we could interpret that:

- The log odds ratio of lower categories vs. higher categories is -0.501, given housing type change from Tower to Apartment.
- The log odds ratio of lower categories vs. higher categories is -0.736, given housing type change from Tower to House
- The log odds ratio of lower categories vs. higher categories is 0.252, given contact level change from Low to High

iv) Pearson residuals

```
pihat = predict(fit.ord, data_nom, type = 'p')
m = rowSums(cbind(data_nom$`Low satisfaction`, data_nom$`Medium satisfaction`, data_nom$`High satisfact
res.pearson = (data_nom[,3:5] - pihat * m) / sqrt(pihat * m)
cbind(type = data_nom$type, contact = data_nom$contact, res.pearson) %>% knitr::kable()
```

type	contact	Low satisfaction	Medium satisfaction	High satisfaction
Tower	Low	0.7793957	-0.3697193	-0.3151179
Tower	High	-0.9946852	0.4549302	0.3354430
Apartment	Low	0.9177560	-1.0671823	-0.0152734
Apartment	High	-0.2369309	-0.4052334	0.5377735
House	Low	-1.1407855	0.1397563	1.2440771
House	High	0.2743817	1.3677881	-1.4778270

From the table, we could see that the largest discrepancy is when given housing type = `House`, contact level = `High` and `High satisfaction`, the Pearson residual is -1.478