Homework 12

Applied Mutlivariate Analysis

Emorie Beck

November 27, 2018

1 Workspace

1.1 Packages

```
library(car)
library(knitr)
library(kableExtra)
library(psych)
library(MASS)
library(Rmisc)
library(mlogit)
library(broom)
library(plyr)
library(tidyverse)
```

1.2 data

The file, Set_10.csv, contains the following data from the job search study: number of publications while in graduate school, length of time to complete the Ph.D. (in years), sex of candidate (1 = men, 2 = women), quality of the degree-granting institution (1 = top-tier research institution, 2 = middle-tier research institution, and 3 = lower-tier research institution), and the outcome of the job search (1 = no interviews, 2 = interviewed but not hired, 3 = hired).

Conduct a multinomial logistic regression on these data, predicting job search outcome from the other variables. Use the "no interviews" outcome as the reference for the dependent variable. Use the lower-tier category as the reference for the quality of the degree-granting institution predictor. Use women as the reference for the sex of candidate predictor.

```
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework12"

dat <- sprintf("%s/Set_10.csv", wd) %>% read.csv(., stringsAsFactors = F) %>% tbl_df %>%
  mutate(outcome = factor(outcome, levels = c(1,2,3), labels = c("no interview", "interview", "hired"))
         sex = as.numeric(mapvalues(sex, 1:2, c(1,0))),
         Institution = factor(mapvalues(Institution, 1:3, 0:2), levels = 0:2, labels = c("Tier 1", "Tie
  mutate_at(vars(pubs, years), funs(c = as.numeric(scale(., scale = F))))

head(dat)

## # A tibble: 6 x 8
##      ID Institution   sex years  pubs outcome      pubs_c years_c
##   <int> <fct>       <dbl> <int> <int> <fct>         <dbl>   <dbl>
```

```
## 1   122 Tier 3          1     5     0 interview     -4.30 -1.09
## 2     1 Tier 3          1     6     0 no interview  -4.30 -0.0900
## 3   191 Tier 3          0     6     0 no interview  -4.30 -0.0900
## 4   194 Tier 2          0     6     0 no interview  -4.30 -0.0900
## 5     4 Tier 3          1     7     0 no interview  -4.30  0.91
## 6     6 Tier 2          1     7     0 no interview  -4.30  0.91
```

2 Question 1

1. When the "interviewed but not hired" outcome is compared to the reference outcome:

```
jobs <- dat %>%
  mutate(NI = ifelse(outcome == "no interview", 1, 0),
         I = ifelse(outcome == "interview", 1, 0),
         H = ifelse(outcome == "hired", 1, 0)) %>%
  gather(key = outcome.ids, value = outcome, NI:H) %>%
  mutate(T1vT2 = ifelse(Institution == "Tier 2", 1, 0),
         T1vT3 = ifelse(Institution == "Tier 3", 1, 0)) %>%
  select(ID, sex:pubs, outcome.ids:T1vT3) %>%
  mutate(outcome.ids = factor(outcome.ids, levels = c("NI", "I", "H"))) %>%
  arrange(ID, outcome.ids) %>% data.frame

J <- mlogit.data(jobs,shape="long",choice="outcome",alt.var="outcome.ids")

Ref_Level <- "NI"
fit_1 <- mlogit(outcome ~ 0 | 1 + sex + T1vT2 + T1vT3 + years + pubs, data = J, reflevel = Ref_Level)
```

```
cbind(data.frame(b = coef(fit_1)), confint(fit_1)) %>% data.frame() %>%
  mutate(term = rownames(.)) %>%
  tbl_df %>%
  select(term, everything()) %>%
  setNames(c("term", "b", "lower", "upper")) %>%
  mutate(sig = ifelse(sign(lower) == sign(upper), "sig", "ns")) %>%
  mutate_at(vars(b, lower, upper), funs(exp)) %>%
  mutate(CI = sprintf("[%.2f, %.2f]", lower, upper), b = sprintf("%.2f", b)) %>%
  mutate_at(vars(b, CI), funs(ifelse(sig == "sig", sprintf("\\textbf{%s}", .), .))) %>%
  select(term, b, CI) %>%
  kable(., "latex", booktabs = T, escape = F) %>%
  kable_styling(full_width = F)
```

term	b	CI
I:(intercept)	**368880.06**	**[1299.26, 104730785.79]**
H:(intercept)	**21544.17**	**[57.57, 8062497.58]**
I:sex	1.82	[0.46, 7.25]
H:sex	2.86	[0.63, 12.90]
I:T1vT2	0.26	[0.03, 1.98]
H:T1vT2	0.17	[0.02, 1.55]
I:T1vT3	0.50	[0.05, 4.50]
H:T1vT3	0.58	[0.05, 6.28]
I:years	**0.07**	**[0.03, 0.18]**
H:years	**0.03**	**[0.01, 0.07]**
I:pubs	**9.90**	**[4.47, 21.92]**
H:pubs	**36.95**	**[15.89, 85.89]**

2.1 Part A

What are the significant predictors?

Both years and publications are significant predictors of the outcome.

2.2 Part B

How should the significant predictors be interpreted?

Years: An additional year in graduate school multiplies the odds associated with being interviewed by .07.

Publications: Each additional publication multiplies the odds of being interviewed by 9.90.

3 Question 2

When the "hired" outcome is compared to the reference outcome:

3.1 Part A

What are the significant predictors?

Both years and publications are significant predictors of the outcome.

3.2 Part B

How should the significant predictors be interpreted?

Years: An additional year in graduate school multiplies the odds associated with being hired by .03.

An additional year in graduate school is associated with a .03 increase in odds of being hired.

Publications: Each additional publication multiplies the odds of being hired by 36.95.

4 Question 3

What is the probability that a man will be hired if he completes his degree in 5 years at a third-tier institution and enters the job market with 5 publications?

$Y_H = b_{0H} + b_{1H} * sex + b_{2H} * years + b_{3H} * pubs$

```
# get cases that match this because I'm too lazy to  create a data frame
dat %>% filter(years == 5 & pubs == 5 & sex == 0 & Institution == "Tier 3" & outcome == "hired")
```

```
## # A tibble: 0 x 8
## # ... with 8 variables: ID <int>, Institution <fct>, sex <dbl>,
## #   years <int>, pubs <int>, outcome <fct>, pubs_c <dbl>, years_c <dbl>

nd <- crossing(sex = 1,
         years = 5,
         pubs = 5,
         outcome.ids = c("NI", "I", "H"),
         T1vT2 = 0,
         T1vT3 = 1
         ) %>%
  mutate(outcome = c(0, 0, 1))

O.Q3 <- predict(fit_1, newdata = nd)
# e^L_M / (1 + sigma_{j=2}^{M}{e^L_j})
(p.Q3 <- exp(O.Q3["H"]) / (1 + sum(exp(O.Q3[c("I", "H")]))))

##         H
## 0.3492526
```

The probability would be 34.93%.

5 Question 4

4. How do his odds of getting hired change if he gets 2 more publications but takes a year longer to finish?

```
# get cases that match this because I'm too lazy to  create a data frame
nd <- crossing(sex = 1,
         years = 6,
         pubs = 7,
         outcome.ids = c("NI", "I", "H"),
         T1vT2 = 0,
         T1vT3 = 1
         ) %>%
  mutate(outcome = c(0, 0, 1))

O.Q4 <- predict(fit_1, newdata = nd)
(O.diff <- exp(O.Q4["H"])) - exp(O.Q3["H"])

##         H
## 0.6969054
```

The difference in odds will be .70 higher odds of being hired.