# 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, 0:1))) %>%
  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>       <int> <dbl> <int> <int> <fct>         <dbl>   <dbl>
## 1   122           3     0     5     0 interview     -4.30 -1.09
```

```
## 2     1           3     0     6     0 no interview  -4.30 -0.0900
## 3   191           3     1     6     0 no interview  -4.30 -0.0900
## 4   194           2     1     6     0 no interview  -4.30 -0.0900
## 5     4           3     0     7     0 no interview  -4.30  0.91
## 6     6           2     0     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)) %>%
  select(ID:pubs, NI:H) %>%
  gather(key = outcome.ids, value = outcome, NI:H) %>%
  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 + Institution + 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) | **272925.12** | **[1096.35, 67941793.17]** |
| H:(intercept) | **15268.19** | **[43.09, 5410328.25]** |
| I:sex | 0.54 | [0.14, 2.11] |
| H:sex | 0.34 | [0.08, 1.52] |
| I:Institution | 0.85 | [0.31, 2.33] |
| H:Institution | 0.98 | [0.32, 2.95] |
| I:years | **0.08** | **[0.03, 0.19]** |
| H:years | **0.03** | **[0.01, 0.08]** |
| I:pubs | **8.97** | **[4.26, 18.88]** |
| H:pubs | **33.44** | **[15.09, 74.11]** |

### 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 is associated with a .08 increase in odds of being interviewed.
**Publications**: Each additional publication is assocatied with a 9.19 increase in odds of being interviewed.

## 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 is associated with a .03 increase in odds of being hired.
**Publications**: Each additional publication is assocatied with a 33.82 increase in odds of being hired.

## 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 == 3 & outcome == "interview")
```

```
## # A tibble: 3 x 8
##      ID Institution   sex years  pubs outcome   pubs_c years_c
##   <int>       <int> <dbl> <int> <int> <fct>      <dbl>   <dbl>
## 1   124           3     0     5     5 interview  0.698   -1.09
## 2   140           3     0     5     5 interview  0.698   -1.09
## 3   141           3     0     5     5 interview  0.698   -1.09
```

```
O1 <- exp(predict(fit_1, newdata = jobs %>% filter(ID == 124)))["H"]
O1 / (1 + O1)
```

```
##        H
## 0.590306
```

The probability would be 59%.

# 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
dat %>% filter(years == 6 & pubs == 7 & sex == 0 & Institution == 3 & outcome == "interview")

## # A tibble: 1 x 8
##      ID Institution   sex years  pubs outcome   pubs_c years_c
##   <int>       <int> <dbl> <int> <int> <fct>      <dbl>   <dbl>
## 1   152           3     0     6     7 interview   2.70 -0.0900

O2 <- exp(predict(fit_1, newdata = jobs %>% filter(ID == 152)))["H"]
```

The difference in odds will be -0.69 lower odds of being hired.