Homework 9

Applied Mutlivariate Analysis

Emorie Beck

November 1, 2018

1 Workspace

1.1 Packages

```
library(car)
library(knitr)
library(psych)
library(gridExtra)
library(knitr)
library(kableExtra)
library(MASS)
library(vegan)
library(smacof)
library(scatterplot3d)
library(ape)
library(ade4)
library(ecodist)
library(cluster)
library(factoextra)
library(candisc)
library(ggdendro)
library(lme4)
library(plyr)
library(tidyverse)
```

1.2 data

The file, Set_8.csv, contains the data from a follow-up to the job search study. The file contains GRE scores (Verbal + Quantitative) upon entering graduate school, number of publications while in graduate school, length of time to complete the Ph.D. (in years), and the outcome of the job search (1=no interviews, 2=got a job, 3=interviews but no job). The variable, sample, divides the sample into two random halves. Analyze the data from sample=1 using discriminant analysis to determine how best to predict job search outcome. Use sample=2 for cross- validation. Answer the following questions.

```
wd <- "https://github.com/emoriebeck/homeworks/raw/master/multivariate/homeworks/homework10"

dat <- sprintf("%s/Set_8.csv", wd) %>% read.csv(., stringsAsFactors = F)

dat1 <- dat %>% filter(sample == 1)
dat2 <- dat %>% filter(sample == 2)
```

2 Question 1

How many discriminant functions are significant?

```
# lda
LDA_1 <- lda(outcome ~ gre + pubs + years,  data = dat1)

# cda
MLM_1 <- lm(cbind(gre, pubs, years)~as.factor(outcome),data=dat1)
CDA_1 <- candisc(MLM_1, data=dat1)

CDA_1

##
## Canonical Discriminant Analysis for as.factor(outcome):
##
##     CanRsq Eigenvalue Difference Percent Cumulative
## 1 0.791158   3.788299     3.6939 97.5698      97.57
## 2 0.086221   0.094356     3.6939  2.4302     100.00
##
## Test of H0: The canonical correlations in the
## current row and all that follow are zero
##
##   LR test stat approx F numDF denDF   Pr(> F)
## 1      0.19084   105.28     6   490 < 2.2e-16 ***
## 2      0.91378              2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

One discriminant function is significant.

3 Question 2

Comment on the irrelative ”importance.”

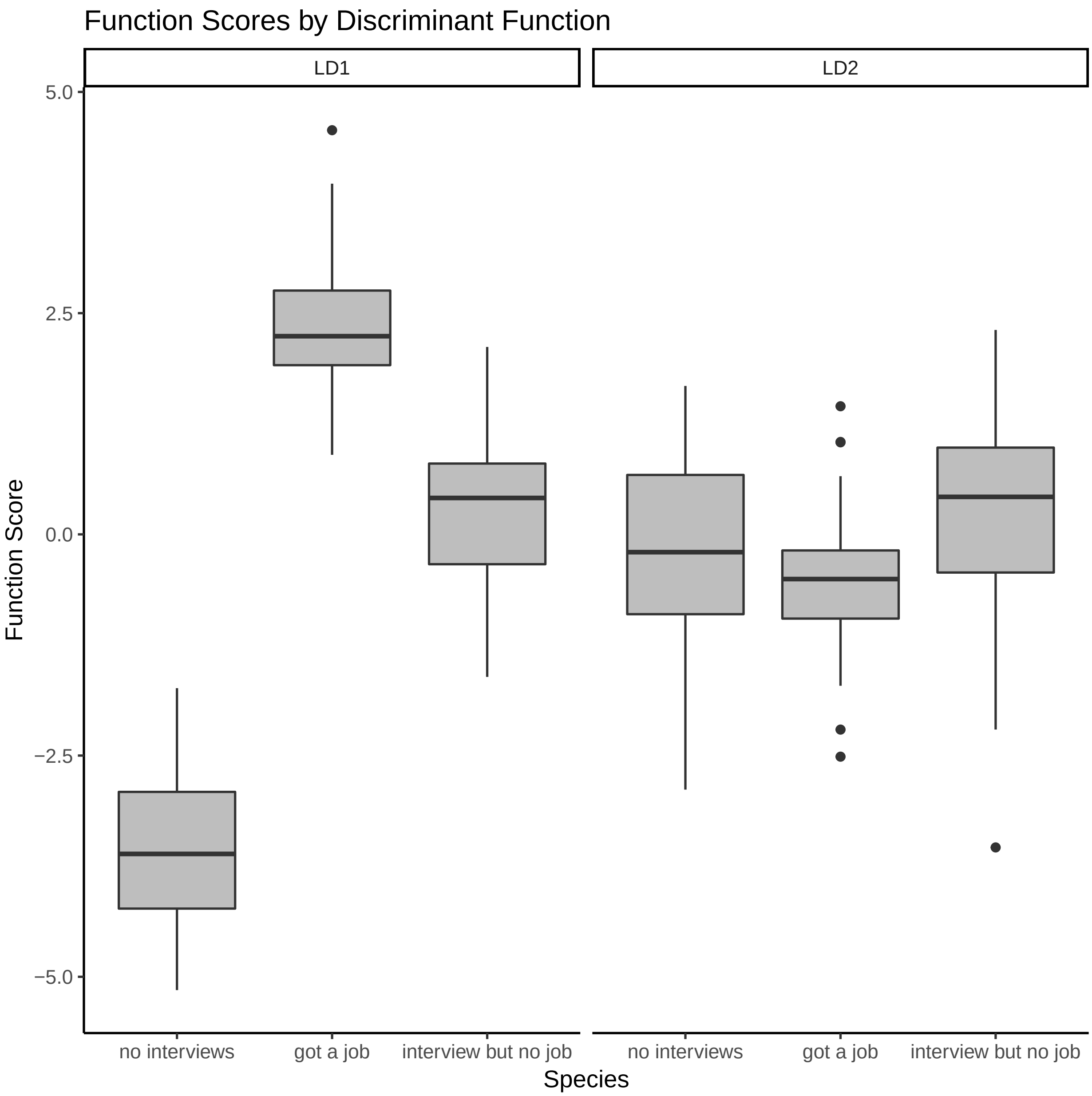
There is only one function, so it is all that is important.

4 Question 3

How would you interpret the(se) function(s)?

```
LDA_Values <- predict(LDA_1)
as.data.frame(LDA_Values$x) %>%
  bind_cols(data.frame(class = LDA_Values$class)) %>%
  group_by(class) %>%
  summarize_at(vars(-class), funs(mean), na.rm = T)

## # A tibble: 3 x 3
##   class    LD1    LD2
##   <fct>  <dbl>  <dbl>
## 1 1     -3.53  -0.213
## 2 2      2.42  -0.479
## 3 3      0.307  0.292
```

```
as.data.frame(LDA_Values$x) %>%
  bind_cols(data.frame(Class = LDA_Values$class)) %>%
  gather(key = Function, value = Score, -Class) %>%
  mutate(Class_long = mapvalues(Class, from = c(1, 2, 3), to = c("no interviews", "got a job", "intervi
  ggplot(aes(x = Class_long, y = Score)) +
  geom_boxplot(fill = "gray") +
  ylab("Function Score") +
  xlab("Species") +
  ggtitle("Function Scores by Discriminant Function") +
  facet_grid(~Function) +
  theme_classic()
```



Function 1 appears to be discriminating between people who got jobs, interviews, or neither.

5 Question 4

How well are the original cases classified?

5.1 Part A

Calculate a significance test that compares the classification to what would be expected by chance.

```
(Class_T <- table(Original = dat1$outcome, Predicted = LDA_Values$class))

##         Predicted
## Original   1   2   3
##        1  51   0   3
##        2   0  48  23
##        3   2  12 111

# Total observations
N <- nrow(dat1)

# Observed agreement
O <- sum(diag(Class_T))

# Marginals (O = Observed, P = Predicted)
MO1 <- MP1 <- sum(Class_T[1, ])
MO2 <- MP2 <- sum(Class_T[2, ])
MO3 <- MP3 <- sum(Class_T[3, ])

# Expected agreement
E <- (MO1 * MP1/N) + (MO2 * MP2/N) + (MO3 * MP3/N)

t <- (O - E)/sqrt(N * (E/N) * (1 - E/N))
t

## [1] 15.0929

chi_squared <- (((O - E)^2)/E) + ((((250 - O) - (250 - E))^2)/(250 -E))
chi_squared

## [1] 227.7956
```

5.2 Part B

Calculate Kleckas tau.

```
Tau <- (O - E)/(N - E)
Tau

## [1] 0.7430495
```

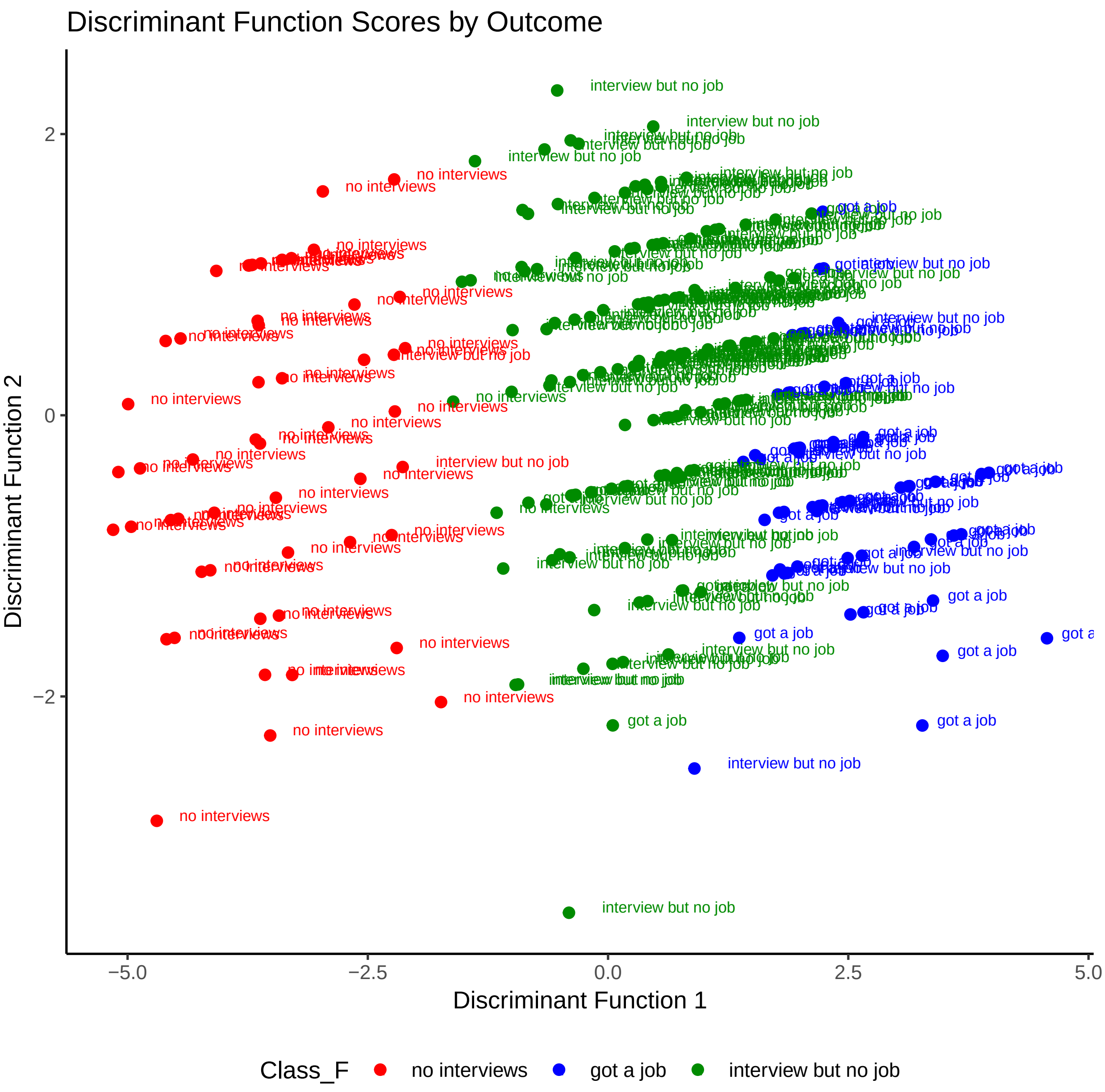
6 Question 5

6.1 Part A

What is the most common type of misclassification?

The most common misclassification was for people who were classified as interviewed but not hired who were actually hired.

```
plot_data <- cbind(LDA_Values$x[, 1], LDA_Values$x[, 2], LDA_Values$class, dat1$outcome)
plot_data <- as.data.frame(plot_data)
names(plot_data) <- c("DS1", "DS2", "Class", "Outcome")
plot_data$Class_F <- factor(plot_data$Class, levels = c(1, 2, 3), labels = c("no interviews", "got a jol
plot_data$Outcome_F <- factor(plot_data$Outcome, levels = c(1, 2, 3), labels = c("no interviews", "got a

ggplot(plot_data, aes(x = DS1, y = DS2, color = Class_F)) +
  geom_point(shape = 19, size = 2, na.rm = TRUE) +
  scale_color_manual(values =c("red", "blue", "green4")) +
  geom_text(aes(label = Outcome_F), hjust = -.25, vjust = 0, size = 2.5) +
  xlab("Discriminant Function 1") +
  ylab("Discriminant Function 2") +
  theme_classic() +
  ggtitle("Discriminant Function Scores by Outcome") +
  theme(legend.position = "bottom")
```



6.2 Part B

Speculate about what might account for this misclassification?

There are likely a number of additional variables that factor into whether someone gets a job interview or job (or not).

6.3 Part C

What additional predictor(s) might this suggest for future analysis? (There is no correct answer here; speculate about what else might determine job search outcome beyond the variables included in the present data.)

Prestige of granting institution, prestige of mentor, conference networking, blog/social media presence, teaching experience

7 Question 6

How well are the cases classified using the jackknife (leave-one-out) procedure?

```
jackknife_1 <- lda(outcome ~ gre + pubs + years, data = dat1, CV = TRUE)
Jack_T <- table(Original = dat1$outcome, Predicted = jackknife_1$class)
Proportion_of_Correct_Classification <- sum(diag(Jack_T))/sum(Jack_T)
Proportion_of_Correct_Classification

## [1] 0.836
```

8 Question 7

How well are cases in the cross-validation sample classified?

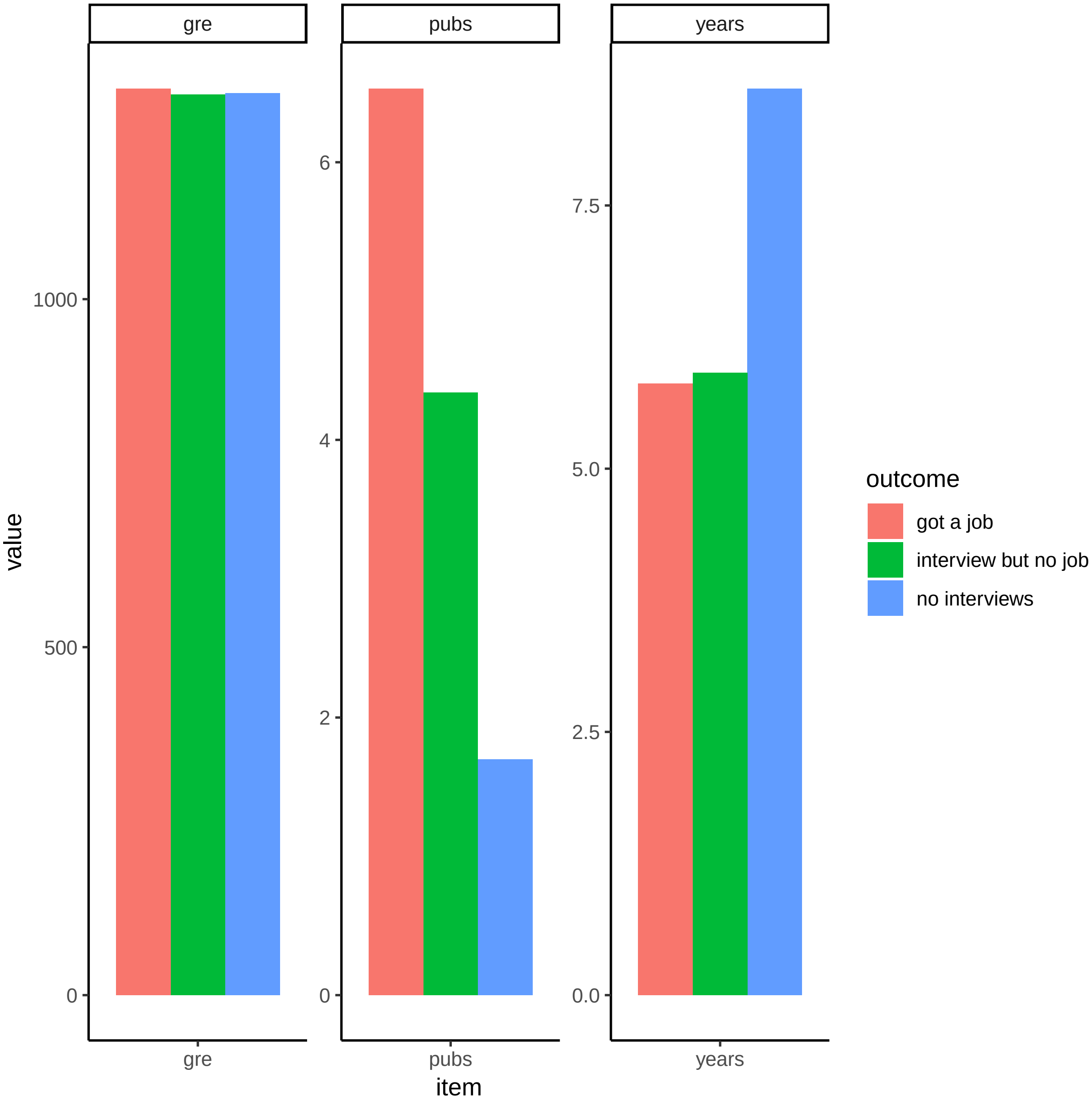
```
Train_1 <- lda(outcome ~ gre + pubs + years, data = dat1, CV = FALSE)

Predict_1 <- predict(Train_1, newdata = dat2)

(tab <- table(Original = dat2$outcome, Predicted = Predict_1$class))

##          Predicted
## Original   1   2   3
##        1  59   0   7
##        2   0  44  21
##        3   3  13 103

sum(diag(tab))/sum(tab)

## [1] 0.824
```

Decently well (Percent classified correct was 82.4)%.

9 Question 8

Based on the analysis, what advice would you give to a student thinking about a career in academia?

```
dat %>% gather(key = item, value = value, gre:years) %>%
  mutate(outcome = mapvalues(outcome, from = c(1, 2, 3), to = c("no interviews", "got a job", "intervie
  group_by(outcome, item) %>%
  summarize(value = mean(value, na.rm = T)) %>%
  ggplot(aes(x=item, y = value, fill = outcome)) +
    geom_bar(stat = "identity", position = "dodge") +
    facet_wrap(~item, scale = "free") +
    theme_classic()
```



Publish or perish. And get out fast.