Homework #1

Dustin Leatherman

April 6, 2019

```r
# pca_analysis
# Parameters:
# dataset - data.frame containing all items that will be run against pca. Note: filter factors beforeha
# type - enumeration for "original" or "standard". orginal applies pca to the orginal values while stan
# returns a tibble containing the original data, a PCA object, and an augmented set of the data with fi
pca_analysis <- function(dataset, type) {
  if(type == "original") {
      dataset %>%
      # calculate mean eigenvalue from covariance matrix to determine floor for principle component sel
      mutate(avg_eigen = eigen(var(.))$values %>% mean) %>%
      nest() %>%
      mutate(
        pca = map(data, ~ prcomp(.x %>% select(-avg_eigen))),
        # add values from pca output onto original data
        pca_aug = map2(pca, data, ~augment(.x, data = .y))
      )
  } else if(type == "standard") {
      dataset %>%
      # calculate mean eigenvalue from correlation matrix to determine floor for principle component se
      mutate(avg_eigen = eigen(cor(.))$values %>% mean) %>%
      nest() %>%
      mutate(
        pca = map(data, ~ prcomp(.x %>% select(-avg_eigen), scale. = TRUE, center = TRUE)),
        # add values from pca output onto original data
        pca_aug = map2(pca, data, ~augment(.x, data = .y))
      )
  }
}

# pca_tidy
# Parameters:
# pca_analysis - a tibble returned from the pca_analysis function
# Returns: a tibble containing summarized pca information
pca_summary <- function(pca_analysis) {
  pca_analysis %>%
    unnest(pca_aug) %>%
    # calculate variance for all PC variables
    summarize_at(.vars = vars(contains("PC")), .funs = funs(var)) %>%
    # pivot columns to rows
    gather(
      key = pc,
      value = variance
    ) %>%
    mutate(
      # variance explained
      var_exp = variance / sum(variance),
      # cumulative sum of variance explained so far
      cum_var_exp = cumsum(var_exp),
      # name of principle component
      pc = str_replace(pc, ".fitted", "")
    )
}

# pca_kable
# Taking all the above parameters, produce a kable which highlights Principle Components where
# the variance exceeds the average eigenvalue
pca_kable <- function(pca_analysis, pca_summary, type) {
  caption_txt <- ifelse(type == "standard", "Standardized", "Original")
  pca_summary %>%
  kable(
    digits = 4,
    col.names = c("PC", "Variance", "Var. Explained", "Cumulative  Var. Explained"),
    caption = paste(caption_txt, "Principle Components. Bolded rows indicate where variance is greater
    ) %>%
  kable_styling("striped", full_width = FALSE, latex_options = "hold_position") %>%
  # bold rows that are greater than the average eigenvalue.
  row_spec(which(pca_summary$variance >= (pca_analysis$pca_aug[[1]]$avg_eigen) %>% mean), bold = T)

}

# kable_pc_list
# Parameters:
# pca_analysis - output from pca_analysis function
# type - "original"/"standard" for type of analysis
# pc_list - vector of Principle Components to retrieve. i.e. c("PC1", "PC2")
# This function is a utility function to write out the PC coefficients in a formatted way
kable_pc_list <- function(pca_analysis, type, pc_list){
  caption_txt <- ifelse(type == "standard", "Standardized", "Original")

  pca_analysis$pca[[1]]$rotation[, pc_list] %>%
  kable(
    caption = paste("Coefficients for selected ", caption_txt, " Principle Components"),
    col.names = pc_list
  ) %>%
  kable_styling("striped", full_width = FALSE, latex_options = "hold_position")
}


# return correlation for a given principle component
pc_corr <- function(dataset, pc_num, type) {
  if(type == "standard") {
    R <- cor(dataset)
    pca <- prcomp(dataset, scale. = TRUE, center = TRUE)
    lambda <- eigen(R)$values
    diag <- diag(R)
  } else if (type == "original") {
    S <- var(dataset)
    pca <- prcomp(dataset)
    lambda <- eigen(S)$values
    diag <- diag(S)
  } else {
    stop(paste("Invalid type specified:", type))
  }
  PC1 <- pca$rotation[,pc_num]
  PC1 * sqrt(lambda[1] / diag)
}
```

1

Carry out a principal components analysis of the diabetes data (diabetes.csv on D2L). Use all five variables, including y's and x's. Use both S and R. Which do you think is more appropriate here? Show the precent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either S or R? Note that: x1 = glucose intolerance x2 = insulin response to oral glucose x3 = insulin resistance y1 = relative weight y2 = fasting plasma glucose

```
diabetes <- read.csv("~/Downloads/diabetes.csv")
diabetes.pca.orig <- pca_analysis(diabetes %>% select(-Patient), "original")
diabetes.pca.tidy.orig <- pca_summary(diabetes.pca.orig)

pca_kable(diabetes.pca.orig, diabetes.pca.tidy.orig, "original")
```

Table 1: Original Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**3467.4640**	**0.6077**	**0.6077**
PC2	**1273.0306**	**0.2231**	**0.8308**
PC3	895.7416	0.1570	0.9878
PC4	69.5111	0.0122	1.0000
PC5	0.0114	0.0000	1.0000

```
# get coefficients from first two principle components since these are the winners per above
kable_pc_list(diabetes.pca.orig, "original", c("PC1", "PC2"))
```

Table 2: Coefficients for selected Original Principle Components

	PC1	PC2
y1	0.0003998	-0.0007613
y2	-0.0081562	0.0152213
x1	0.1567243	0.6484187
x2	0.7430467	0.4194744
x3	0.6505785	-0.6351080

```
diabetes.pca.std <- pca_analysis(diabetes %>% select(-Patient), "standard")
diabetes.pca.tidy.std <- pca_summary(diabetes.pca.std)

pca_kable(diabetes.pca.std, diabetes.pca.tidy.std, "standard")
```

```
# get coefficients from first two principle components since these are the winners per above
kable_pc_list(diabetes.pca.std, "standard", c("PC1", "PC2"))
```

The average eigenvalue for the original scale is 1141.152 which means the Principle Components of Interest are PC1 and PC2. The average eigenvalue for the standardized data is always 1 indicating the same results.

Table 3: Standardized Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**1.7165**	**0.3433**	**0.3433**
PC2	**1.2304**	**0.2461**	**0.5894**
PC3	0.9605	0.1921	0.7815
PC4	0.7918	0.1584	0.9398
PC5	0.3009	0.0602	1.0000

Table 4: Coefficients for selected Standardized Principle Components

	PC1	PC2
y1	0.4131436	-0.5418329
y2	0.0669611	-0.6846833
x1	0.3623932	-0.1766914
x2	0.5447779	0.4240333
x3	0.6298545	0.1631007

Interpretation

Using the Original variables allows for a bit easier interpretation.

PC1

The coefficients are much higher for x2 (insulin response to oral glucose) and x3 (insulin resistance). Without knowing more about the data points, I posit that PC1 is a measure of insulin impact to the body.

PC2

The largest magnitudes for the coefficients are x1 (glucose intolerance), x2 (insulin response to oral glucose), and x3 (insulin resistance). x3 is negative while x1 and x2 are positive. This indicates that this variable may reference an interaction effect of oral glucose and insulin. This is because x1 and x3 are nearly the same magnitude with opposite signs and x3 is a measure involving some combonation of them both.

2

Carry out a principal components analysis of the word probe data (wordprobe.csv on D2L). Use all five variables, including y's and x's. Use both S and R. Which do you think is more appropriate here? Show the precent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either S or R? Note that:

y_i *= response time for the i-th probe word*

where i = 1, 2, 3, 4, 5.

```r
wordprobe <- read.csv("~/Downloads/wordprobe.csv")

wordprobe.pca.orig <- pca_analysis(wordprobe %>% select(-Subject), "original")
wordprobe.pca.tidy.orig <- pca_summary(wordprobe.pca.orig)

pca_kable(wordprobe.pca.orig, wordprobe.pca.tidy.orig, "original")
```

Table 5: Original Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**200.3754**	**0.6841**	**0.6841**
PC2	36.0908	0.1232	0.8074
PC3	34.0721	0.1163	0.9237
PC4	14.9673	0.0511	0.9748
PC5	7.3853	0.0252	1.0000

```r
# get coefficients from first two principle components since these are the winners per above
kable_pc_list(wordprobe.pca.orig, "original", c("PC1"))
```

Table 6: Coefficients for selected Original Principle Components

	PC1
y1	-0.4727831
y2	-0.3918187
y3	-0.4875471
y4	-0.4677199
y5	-0.4080320

```r
wordprobe.pca.std <- pca_analysis(wordprobe %>% select(-Subject), "standard")
wordprobe.pca.tidy.std <- pca_summary(wordprobe.pca.std)

pca_kable(wordprobe.pca.std, wordprobe.pca.tidy.std, "standard")
```

Table 7: Standardized Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**3.4165**	**0.6833**	**0.6833**
PC2	0.6144	0.1229	0.8062
PC3	0.5723	0.1145	0.9206
PC4	0.2712	0.0542	0.9749
PC5	0.1256	0.0251	1.0000

```r
# get coefficients from first two principle components since these are the winners per above
kable_pc_list(wordprobe.pca.std, "standard", c("PC1"))
```

Table 8: Coefficients for selected Standardized Principle Components

	PC1
y1	-0.4418394
y2	-0.4535595
y3	-0.4727808
y4	-0.4536224
y5	-0.4120276

```
# calculate correlation between a principle component and the original variables
# pc_corr(wordprobe %>% select(-Subject), 1, "standard")
```

The average eigenvalue based on the original values is 58.5782. Only the first principle component exceeds this value so it is the only one that is chosen according to this method. The 80% rule would include the second principle component as well but it is outside this analysis. The same results are seen when looking at the Standardized Principle Components.

Interpretation

The coefficients in both analysis appear to be generally close so this PC can be interpreted as a weighted average of response times.

3

Carry out a principal components analysis separately for males and females in the psyc.csv dataset. Compare the results for the two groups. Use S.

Men

```
psych <- read.csv("~/Downloads/psych.csv")

psych.pca.men <- pca_analysis(
  psych %>% filter(Sex == "Male") %>% select(-Sex),
  "standard"
  )


psych.pca.men.tidy <- pca_summary(psych.pca.men)

pca_kable(psych.pca.men, psych.pca.men.tidy, "standard")
```

Table 9: Standardized Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**2.5498**	**0.6375**	**0.6375**
PC2	0.7209	0.1802	0.8177
PC3	0.3983	0.0996	0.9173
PC4	0.3310	0.0827	1.0000

```
# get coefficients from first principle component
kable_pc_list(psych.pca.men, "standard", c("PC1"))
```

Table 10: Coefficients for selected Standardized Principle Components

	PC1
Pic	0.5224837
Paper	0.4420448
Tool	0.5043526
Vocab	0.5265316

Women

```
psych.pca.women <- pca_analysis(
  psych %>% filter(Sex == "Female") %>% select(-Sex),
  "standard"
  )
psych.pca.women.tidy <- pca_summary(psych.pca.women)

pca_kable(psych.pca.women, psych.pca.women.tidy, "standard")
```

Table 11: Standardized Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**2.1425**	**0.5356**	**0.5356**
PC2	0.9240	0.2310	0.7666
PC3	0.5642	0.1410	0.9077
PC4	0.3693	0.0923	1.0000

```
# get coefficients from first principle component
kable_pc_list(psych.pca.women, "standard", c("PC1"))
```

Table 12: Coefficients for selected Standardized Principle Components

	PC1
Pic	-0.5048449
Paper	-0.5408625
Tool	-0.5121513
Vocab	-0.4362344

Interpretation

The coefficients for both groups are roughly similar in magnitude to the coefficients within and between groups. The women have all negative loadings while the men are positive. Positive loadings of approximately the same size indicate a weighted average. All negative loadings of approximately the same indicate weighted average as well.

Comparison

The loadings are similar in magnitude and spread for both groups so there does not appear to be much of a difference between men and women.

4

Carry our a principal components analysis separately for the two species in the beetle data set (beetles.csv). Compare the results for the two groups. Use S.

Oleracea

```
beetles <- read.csv("~/Downloads/beetles.csv")
```

```
beetles.pca.oler <- pca_analysis(
  beetles %>% filter(Species == "Oleracea") %>% select(-Species),
  "standard"
  )

beetles.pca.oler.tidy <- pca_summary(beetles.pca.oler)

pca_kable(beetles.pca.oler, beetles.pca.oler.tidy, "standard")
```

Table 13: Standardized Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**2.3953**	**0.5988**	**0.5988**
PC2	0.8869	0.2217	0.8205
PC3	0.4300	0.1075	0.9280
PC4	0.2878	0.0720	1.0000

```
# get coefficients from first principle component
kable_pc_list(beetles.pca.oler, "standard", c("PC1"))
```

Table 14: Coefficients for selected Standardized Principle Components

	PC1
y1	0.5723723
y2	0.5625399
y3	0.4190465
y4	0.4246632

Carduorum

```
beetles.pca.card <- pca_analysis(
  beetles %>% filter(Species == "Carduorum") %>% select(-Species),
  "standard"
  )
beetles.pca.card.tidy <- pca_summary(beetles.pca.card)

pca_kable(beetles.pca.card, beetles.pca.card.tidy, "standard")
```

Table 15: Standardized Principle Components. Bolded rows indicate where variance is greater than the average eigenvalue. These are the chosen PCs

PC	Variance	Var. Explained	Cumulative Var. Explained
PC1	**2.3143**	**0.5786**	**0.5786**
PC2	0.8280	0.2070	0.7856
PC3	0.6404	0.1601	0.9457
PC4	0.2172	0.0543	1.0000

```
# get coefficients from first principle component
kable_pc_list(beetles.pca.card, "standard", c("PC1"))
```

Table 16: Coefficients for selected Standardized Principle Components

	PC1
y1	-0.4746835
y2	-0.5938453
y3	-0.5121195
y4	-0.3996960

Interpretation

The coefficients for both groups are roughly similar in magnitude to the coefficients within and between groups. Carduorum beetles have all negative loadings while Oleracea beetles are positive. Positive loadings of approximately the same size indicate a weighted average. All negative loadings of approximately the same indicate weighted average as well.

Comparison

The loadings are similar in magnitude and spread for both groups so there does not appear to be much of a difference between men and women.

5

The weekly rates of return for five stocks were recorded for the period January 1995 through March 1995. The data provided here are for 16 weekly rates of return. The stocks observed are Allied Chemical, Dow Chemical, Union Chemical, the Standard Oil Company, and Texas oil. Data can be found under the Week #1 course content in the file stocks.csv.

```
stocks <- read.csv("~/Downloads/stocks.csv")

# (a) Calculate the mean and standard deviation for each variable.
stocks %>% summarize_all(.funs = c(Mean=mean, Stdev=sd))
```

```
##   Allied_Mean Dow_Mean Union_Mean Standard_Mean Texas_Mean Allied_Stdev
## 1    2.633333 2.671333   2.536667      4.732667      5.142    0.9379359
##   Dow_Stdev Union_Stdev Standard_Stdev Texas_Stdev
## 1 0.8428596     0.30591      0.4722358   0.6364432
```

```
# (b) Calculate the correlation structure between and within the chemical companies and the oil compani
chemical <- stocks %>% select(Allied,Dow,Union)
oil <- stocks %>% select(Standard,Texas)
cca.res <- cca(chemical, oil, xscale = TRUE, yscale = TRUE)
summary(cca.res)
```

```
##
## Canonical Correlation Analysis - Summary
##
##
## Canonical Correlations:
##
##      CV 1      CV 2
## 0.8810089 0.4459326
##
## Shared Variance on Each Canonical Variate:
##
```

```
##      CV 1      CV 2
## 0.7761766 0.1988559
##
## Bartlett's Chi-Squared Test:
##
##         rho^2    Chisq df   Pr(>X)
## CV 1  0.77618 18.90474  6 0.004328 **
## CV 2  0.19886  2.43886  2 0.295399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Canonical Variate Coefficients:
##
##  X Vars:
##              CV 1       CV 2
## Allied 0.03158698  1.1609907
## Dow    0.15758801 -0.2035019
## Union  0.92246655 -0.3503606
##
##  Y Vars:
##                 CV 1         CV 2
## Standard  0.06825228 -1.000161166
## Texas    -1.00248457  0.002327606
##
##
## Structural Correlations (Loadings):
##
##  X Vars:
##             CV 1       CV 2
## Allied 0.3851902  0.9191516
## Dow    0.5013262  0.5440600
## Union  0.9852173 -0.1244170
##
##  Y Vars:
##                  CV 1        CV 2
## Standard -0.002321831 -0.99999730
## Texas    -0.997679664 -0.06808294
##
##
## Fractional Variance Deposition on Canonical Variates:
##
##  X Vars:
##             CV 1      CV 2
## Allied 0.1483715 0.8448396
## Dow    0.2513280 0.2960013
## Union  0.9706531 0.0154796
##
##  Y Vars:
##                  CV 1        CV 2
## Standard 5.390898e-06 0.999994609
## Texas    9.953647e-01 0.004635287
##
##
```

```
## Canonical Communalities (Fraction of Total Variance
## Explained for Each Variable, Within Sets):
##
##  X Vars:
##    Allied       Dow     Union
## 0.9932111 0.5473293 0.9861327
##
##  Y Vars:
## Standard    Texas
##        1        1
##
##
## Canonical Variate Adequacies (Fraction of Total Variance
## Explained by Each CV, Within Sets):
##
##
##  X Vars:
##      CV 1      CV 2
## 0.4567842 0.3854402
##
##  Y Vars:
##      CV 1      CV 2
## 0.4976851 0.5023149
##
##
## Redundancy Coefficients (Fraction of Total Variance
## Explained by Each CV, Across Sets):
##
##
##  X | Y:
##       CV 1       CV 2
## 0.35454521 0.07664704
##
##  Y | X:
##       CV 1       CV 2
## 0.38629150 0.09988827
##
##
## Aggregate Redundancy Coefficients (Total Variance
## Explained by All CVs, Across Sets):
##
##  X | Y: 0.4311923
##  Y | X: 0.4861798
```

(c) State the hypotheses and test for the significance of the canonical correlations.

There is convincing evidence that the first or second canonical correlation is non-zero (Bartlett Chi-Squared Test. p-value = 0.0043). There is no evidence that the second canonical correlation is non-zero (Bartlett Chi-Squared Test. p-value = 0.2954). The first canonical correlation will be used going forward.

(d) Give the values for the significant canonical correlations and the squared canonical correlations. Interpret the squared canonical correlations in the context of the problem.

Significant Canonical Correlation

```
cca.res$corr[1]
```

```
##      CV 1
## 0.8810089
```

Canonical Correlation Squared

```
cca.res$corr[1]^2
```

```
##      CV 1
## 0.7761766
```

77.62% of the variance in the stock price of the Oil companies is explained by the stock price of the Chemical companies.

(e) Write the equations of the significant canonical variables (based on the method, unstandardized or standardized) that you choose. Comment on their relative importance of the original variables to the canonical variables.

$U_1 = 0.0316(\frac{Allied - \bar{Allied}}{Allied_s}) + 0.1576(\frac{Dow - \bar{Dow}}{Dow_s}) + 0.9225(\frac{Union - \bar{Union}}{Union_s})$

$V_1 = 0.0683(\frac{Standard - \bar{Standard}}{Standard_s})$ - $1.0025(\frac{Texas - \bar{Texas}}{Texas_s})$

```
cca.res$xstructcorr[,1]
```

```
##    Allied       Dow     Union
## 0.3851902 0.5013262 0.9852173
```

```
cca.res$ystructcorr[,1]
```

```
##     Standard        Texas
## -0.002321831 -0.997679664
```

The standardized value of Union dominates U_1 while the standardized value of Texas dominates V_1. Inspecting at the correlations between the canonical loadings and the standardized variables show that Union and U_1 have correlation of 0.985 while Texas and V_1 have a correlation of -0.998. This means that $r_1 = 0.881$ may be a surrogate of the relationship between the stock price of Union Chemical and the stock price of Texas Oil.

(f) Comment and interpret the redundancy analysis.

```
##      CV 1
## 0.4567842
```

```
##      CV 1
## 0.4976851
```

```
##      CV 1
## 0.3545452
```

```
##      CV 1
## 0.3862915
```

U_1 explains 45.7% of the variance in the chemical variables and 35.5% of the variance in the oil variables. V_1 explains 49.8% of the variance in the oil variables and 38.6% of the variance in the chemical variables.

6

Height and weight data were collected on a sample of 19 fathers and their adult sons. Height is given in inches and weight is given in pounds. Data can be found under the Week #1 course content in the file fatherson.csv.

```
fatherson <- read.csv("~/Downloads/fatherson.csv")

# (a) Calculate the mean and standard deviation for each variable.
fatherson %>% summarize_all(.funs = c(Mean=mean, Stdev=sd))
```

```
##   FatherHT_Mean FatherWT_Mean SonHT_Mean SonWT_Mean FatherHT_Stdev
## 1      72.15789      211.6842   71.52632   205.2632       2.141214
##   FatherWT_Stdev SonHT_Stdev SonWT_Stdev
## 1       26.00652    1.954153    22.79094
```

```
# (b) Calculate the correlation structure between and within the fathers’ characteristics and the sons’
father <- fatherson %>% select(FatherHT, FatherWT)
son <- fatherson %>% select(SonHT, SonWT)
cca.res <- cca(son, father, xscale = TRUE, yscale = TRUE)
summary(cca.res)
```

```
##
## Canonical Correlation Analysis - Summary
##
##
## Canonical Correlations:
##
##      CV 1      CV 2
## 0.8776164 0.2009508
##
## Shared Variance on Each Canonical Variate:
##
##       CV 1       CV 2
## 0.77021052 0.04038121
##
## Bartlett's Chi-Squared Test:
##
##         rho^2     Chisq df    Pr(>X)
## CV 1  0.770211 23.433068  4 0.0001037 ***
## CV 2  0.040381  0.638897  1 0.4241105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Canonical Variate Coefficients:
##
##  X Vars:
##             CV 1       CV 2
## SonHT -0.1028398 -1.0223841
## SonWT -0.9713278  0.3352129
##
##  Y Vars:
##                 CV 1       CV 2
## FatherHT -0.09156018 -1.0716054
## FatherWT -0.96266764  0.4795752
##
```

```
##
## Structural Correlations (Loadings):
##
##  X Vars:
##             CV 1       CV 2
## SonHT -0.3262275 -0.9452913
## SonWT -0.9949791  0.1000832
##
##  Y Vars:
##                 CV 1        CV 2
## FatherHT -0.4459050 -0.89508031
## FatherWT -0.9963697  0.08513189
##
##
## Fractional Variance Deposition on Canonical Variates:
##
##  X Vars:
##            CV 1       CV 2
## SonHT 0.1064244 0.89357565
## SonWT 0.9899834 0.01001665
##
##  Y Vars:
##               CV 1        CV 2
## FatherHT 0.1988312 0.801168766
## FatherWT 0.9927526 0.007247439
##
##
## Canonical Communalities (Fraction of Total Variance
## Explained for Each Variable, Within Sets):
##
##  X Vars:
## SonHT SonWT
##     1     1
##
##  Y Vars:
## FatherHT FatherWT
##        1        1
##
##
## Canonical Variate Adequacies (Fraction of Total Variance
## Explained by Each CV, Within Sets):
##
##
##  X Vars:
##      CV 1      CV 2
## 0.5482039 0.4517961
##
##  Y Vars:
##      CV 1      CV 2
## 0.5957919 0.4042081
##
##
## Redundancy Coefficients (Fraction of Total Variance
## Explained by Each CV, Across Sets):
```

```
##
##
##  X | Y:
##       CV 1       CV 2
## 0.42223237 0.01824407
##
##  Y | X:
##       CV 1       CV 2
## 0.45888519 0.01632241
##
##
## Aggregate Redundancy Coefficients (Total Variance
## Explained by All CVs, Across Sets):
##
##  X | Y: 0.4404764
##  Y | X: 0.4752076
```

(c) State the hypotheses and test for the significance of the canonical correlations.

There is convincing evidence that the first or second canonical correlation is non-zero (Bartlett Chi-Squared Test. p-value = 0.0001). There is no evidence that the second canonical correlation is non-zero (Bartlett Chi-Squared Test. p-value = 0.4241). The first canonical correlation will be used going forward.

(d) Give the values for the significant canonical correlations and the squared canonical correlations. Interpret the squared canonical correlations in the context of the problem.

Significant Canonical Correlation

```
cca.res$corr[1]
```

```
##      CV 1
## 0.8776164
```

Canonical Correlation Squared

```
cca.res$corr[1]^2
```

```
##      CV 1
## 0.7702105
```

77.02% of the variance in the measurements for sons is explained by the measurements of their fathers.

(e) Write the equations of the significant canonical variables (based on the method, unstandardized or standardized) that you choose. Comment on their relative importance of the original variables to the canonical variables.

U_1 = -0.1028($\frac{SonHT - \bar{SonHT}}{SonHT_s}$) - 0.9713($\frac{SonWT - \bar{SonWT}}{SonWT_s}$)

V_1 = -0.0916($\frac{FatherHT - \bar{FatherHT}}{FatherHT_s}$) - 0.9627($\frac{FatherWT - \bar{FatherWT}}{FatherWT_s}$)

```
cca.res$xstructcorr[,1]
```

```
##      SonHT      SonWT
## -0.3262275 -0.9949791
```

```r
cca.res$ystructcorr[,1]
```

```
##   FatherHT   FatherWT 
## -0.4459050 -0.9963697
```

The standardized value of Son Weight dominates U_1 while the standardized value of Father Weight dominates V_1. Inspecting at the correlations between the canonical loadings and the standardized variables show that Son Weight and U_1 have correlation of -0.995 while Father Weight and V_1 have a correlation of -0.996. This means that $r_1 = 0.8776$ may be a surrogate of the relationship between the Father and Son's Weight.

(f) Comment and interpret the redundancy analysis.

```
##      CV 1 
## 0.5482039
```

```
##      CV 1 
## 0.5957919
```

```
##      CV 1 
## 0.4222324
```

```
##      CV 1 
## 0.4588852
```

U_1 explains 54.8% of the variance in the Son variables and 42.2% of the variance in the Father variables. V_1 explains 59.6% of the variance in the Father variables and 45.9% of the variance in the Son variables.

7

The data set student.csv (found under the Week #1 course content) contains information on six predictor measures of student success in a statistics class:

- socioeconomic status (SES) (1=high, 2=middle, 3=low)
- sex (0=male, 1=female)
- grade point average (GPA)
- Scholastic Aptitude Test (SAT)
- previous statistics class (0=no, 1=yes)
- pretest score

and three statistics course measures:

- Exam 1 score
- Exam 2 score
- Exam 3 score

```r
students <- read.csv("~/Downloads/student.csv")

# (a) Calculate the mean and standard deviation for each variable.
students %>% summarize_all(.funs = c(Mean=mean, Stdev=sd))
```

```
##   SES_Mean  sex_Mean GPA_Mean SAT_Mean stats_Mean pretest_Mean exam1_Mean
## 1 1.958333 0.5833333  3.37625    532.5      0.375     19.41667   48.58333
##   exam2_Mean exam3_Mean SES_Stdev sex_Stdev GPA_Stdev SAT_Stdev
## 1   56.58333     85.625 0.7506036 0.5036102 0.3715047  107.3495
##   stats_Stdev pretest_Stdev exam1_Stdev exam2_Stdev exam3_Stdev
## 1   0.4945354      8.198179    13.30223    20.67327    25.46748
```

```
# (b) Calculate the correlation structure between and within the predictor measures of student success
success <- students %>% select(SES, sex, GPA, SAT, stats, pretest)
measures <- students %>% select(exam1, exam2, exam3)
cca.res <- cca(measures, success, xscale = TRUE, yscale = TRUE)
summary(cca.res)
```

```
##
## Canonical Correlation Analysis - Summary
##
##
## Canonical Correlations:
##
##      CV 1      CV 2      CV 3
## 0.8982980 0.3949529 0.3515344
##
## Shared Variance on Each Canonical Variate:
##
##      CV 1      CV 2      CV 3
## 0.8069393 0.1559878 0.1235764
##
## Bartlett's Chi-Squared Test:
##
##         rho^2    Chisq df   Pr(>X)
## CV 1  0.80694 35.03240 18 0.009364 **
## CV 2  0.15599  5.42689 10 0.860900
## CV 3  0.12358  2.37430  4 0.667276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Canonical Variate Coefficients:
##
##  X Vars:
##             CV 1       CV 2       CV 3
## exam1 -0.3299346 -0.2147327  1.0882967
## exam2  0.1888477 -1.0342310 -0.2426904
## exam3 -0.8514550  0.3861439 -0.7597453
##
##  Y Vars:
##                CV 1        CV 2        CV 3
## SES      0.24776044  0.03738834 -0.77085000
## sex     -0.16124501  0.07408239 -0.02375038
## GPA     -0.03154107 -0.05721031  0.32221347
## SAT     -1.02510871  0.10949961 -0.03815904
## stats   -0.17276011  0.94269425  0.21210653
## pretest -0.23276356 -0.43132980  0.65311271
##
##
## Structural Correlations (Loadings):
##
##  X Vars:
##             CV 1       CV 2       CV 3
## exam1 -0.7067719 -0.2813665  0.6490812
## exam2 -0.2066129 -0.9461244 -0.2493186
## exam3 -0.9464152 -0.1008167 -0.3068131
##
##  Y Vars:
##                CV 1        CV 2       CV 3
## SES     -0.27938176 -0.11623247 -0.5938428
## sex     -0.17743778  0.05980504 -0.0142325
## GPA     -0.01840454 -0.20031406  0.5233973
## SAT     -0.92445692 -0.10765007 -0.3115605
## stats   -0.03752215  0.90685763  0.3010500
## pretest -0.36894574 -0.33698895  0.4555249
##
##
## Fractional Variance Deposition on Canonical Variates:
##
##  X Vars:
##            CV 1       CV 2       CV 3
## exam1 0.4995265 0.07916712 0.42130641
## exam2 0.0426889 0.89515136 0.06215974
## exam3 0.8957017 0.01016400 0.09413431
##
##  Y Vars:
##                 CV 1        CV 2         CV 3
## SES     0.0780541684 0.013509987 0.3526492574
## sex     0.0314841672 0.003576643 0.0002025641
## GPA     0.0003387272 0.040125722 0.2739447337
## SAT     0.8546205942 0.011588538 0.0970699350
## stats   0.0014079118 0.822390765 0.0906310990
## pretest 0.1361209576 0.113561551 0.2075029059
##
##
## Canonical Communalities (Fraction of Total Variance
## Explained for Each Variable, Within Sets):
##
##  X Vars:
## exam1 exam2 exam3
##     1     1     1
##
##  Y Vars:
##        SES        sex        GPA        SAT      stats    pretest
## 0.44421341 0.03526337 0.31440918 0.96327907 0.91442978 0.45718541
##
##
## Canonical Variate Adequacies (Fraction of Total Variance
## Explained by Each CV, Within Sets):
##
##
##  X Vars:
##      CV 1      CV 2      CV 3
## 0.4793057 0.3281608 0.1925335
##
##  Y Vars:
##      CV 1      CV 2      CV 3
## 0.1836711 0.1674589 0.1703334
##
```

```
##
## Redundancy Coefficients (Fraction of Total Variance
## Explained by Each CV, Across Sets):
##
##
##  X | Y:
##       CV 1       CV 2       CV 3
## 0.38677059 0.05118907 0.02379260
##
##  Y | X:
##       CV 1       CV 2       CV 3
## 0.14821142 0.02612153 0.02104920
##
##
## Aggregate Redundancy Coefficients (Total Variance
## Explained by All CVs, Across Sets):
##
##  X | Y: 0.4617523
##  Y | X: 0.1953821
```

(c) State the hypotheses and test for the significance of the canonical correlations.

There is convincing evidence that at least one canonical correlation is non-zero (Bartlett Chi-Squared Test. p-value = 0.0094). There is no evidence that the second or third canonical correlation is non-zero (Bartlett Chi-Squared Test. p-value = 0.8609). The first canonical correlation will be used going forward.

(d) Give the values for the significant canonical correlations and the squared canonical correlations. Interpret the squared canonical correlations in the context of the problem.

Significant Canonical Correlation

```
cca.res$corr[1]
```

```
##     CV 1
## 0.898298
```

Canonical Correlation Squared

```
cca.res$corr[1]^2
```

```
##      CV 1
## 0.8069393
```

80.69% of the variance in the measurements for the exams are explained by the student success measurements.

(e) Write the equations of the significant canonical variables (based on the method, unstandardized or standardized) that you choose. Comment on their relative importance of the original variables to the canonical variables.

U_1 = -0.3299($\frac{exam1-\bar{exam1}}{exam1_s}$) + 0.1889($\frac{exam2-\bar{exam2}}{exam2_s}$) - 0.8515($\frac{exam3-\bar{exam3}}{exam3_s}$)

V_1 = 0.2478($\frac{SES-\bar{SES}}{SES_s}$) - 0.1612($\frac{sex-\bar{sex}}{sex_s}$) - 0.0315($\frac{GPA-\bar{GPA}}{GPA_s}$) - 1.025($\frac{SAT-\bar{SAT}}{SAT_s}$) - 0.1728($\frac{stats-\bar{stats}}{stats_s}$) - 0.2328($\frac{pretest-\bar{pretest}}{pretest_s}$)

```
cca.res$xstructcorr[,1]
```

```
##      exam1      exam2      exam3
## -0.7067719 -0.2066129 -0.9464152
```

```
cca.res$ystructcorr[,1]
```

```
##         SES         sex         GPA         SAT       stats     pretest
## -0.27938176 -0.17743778 -0.01840454 -0.92445692 -0.03752215 -0.36894574
```

The standardized value of exam 3 dominates U_1 while the standardized value of the SAT score dominates V_1. Inspecting at the correlations between the canonical loadings and the standardized variables show that exam 3 and U_1 have correlation of -0.9464 while SAT score and V_1 have a correlation of -0.9245. This means that $r_1 = 0.8983$ may be a surrogate of the relationship between the standardized third exam score and the standardized SAT score.

(f) Comment and interpret the redundancy analysis.

```
##      CV 1
## 0.4793057
```

```
##      CV 1
## 0.1836711
```

```
##      CV 1
## 0.3867706
```

```
##      CV 1
## 0.1482114
```

U_1 explains 47.9% of the variance in the Measurement variables and 38.7% of the variance in the Success variables. V_1 explains 18.4% of the variance in the Success variables and 14.8% of the variance in the Measurement variables.