Brad Tweardy
March 25, 2016
MSDS 6372 – 401
Project 2 – PCA

Introduction

The data being analyzed contains individual test scores in ten different subjects. The goal is to determine if a particular combination, if any, of subjects can be used to reasonably explain a linear relationship to future professional success. Success is determined by future income.

Descriptive Statistics

The data comes from a study in which psychologist Richard Hernstein and political scientist Charles Murray had disputed that intelligence was a better predictor to future success than was a person's education and family's socioeconomic status. They published their findings in the book *The Bell Curve: Intelligence and Class Structure*.

The analysis involved data that contained test scores of individuals from the Armed Services Vocational Battery of tests: General Sciences (Science), Arithmetic Reasoning (Arith), Word Knowledge (Word), Paragraph Comprehension (Parag), Numerical Operations (Numer), Coding Speed (Coding), Automotive and Shop Information (Auto), Mathematics Knowledge (Math), Mechanical Comprehension (Mechanic), and Electronics Knowledge (Elec). The test result data includes results from 2,584 Americans selected by the National Longitudinal Study of Youth in 1979. The success indicator is the income of these 2,584 individuals in 2005, 26 years after taking these tests, reported in a survey taken in 2006.

The descriptive statistics can be seen below in Figure 1.

fig. 1

	Mean	Std. Dev.	Median	Min	Max
Science	16.252	4.771	17	0	25
Arith	18.515	7.157	19	0	30
Word	26.557	7.047	28	0	35
Parag	11.202	3.161	12	0	15
Numer	35.323	10.195	36	0	50
Coding	46.880	15.282	48	0	84
Auto	14.238	5.295	14	0	25
Math	14.195	6.284	13	0	25
Mechanic	14.378	5.096	14	0	25
Elec	11.575	4.088	12	0	20
Income 2005	$49417.00	$46727.93	$38231	$63	$703637

Analysis

The dataset is multivariate and each observation is independent from the others. We can also mention that it is a complete dataset without any missing values.

The first step that I took to analyze the data included looking for the number of primary components needed to reduce the model into only necessary components. In order to do so, I can look at a scree plot and the Importance of Coefficients output to make this determination.

We can see with the scree plot in Figure 2 that where the elbow is on the curve before it levels off, we would definitely want to use the first two principal components. Principal Component 1 explains most of the variance, but Principal Component 2 also shows to be worth including as it explains a moderate proportion. It can also be argued that Principal Component three might be worth including.

Figure 3 confirms that it is worth having Principal Component 3 included in order to properly analyze this data. We see that only including the first two principal components only explains about 75% of the variance. By including Principal Component 3, we can explain 80.97% of the variance with this model.

fig. 2

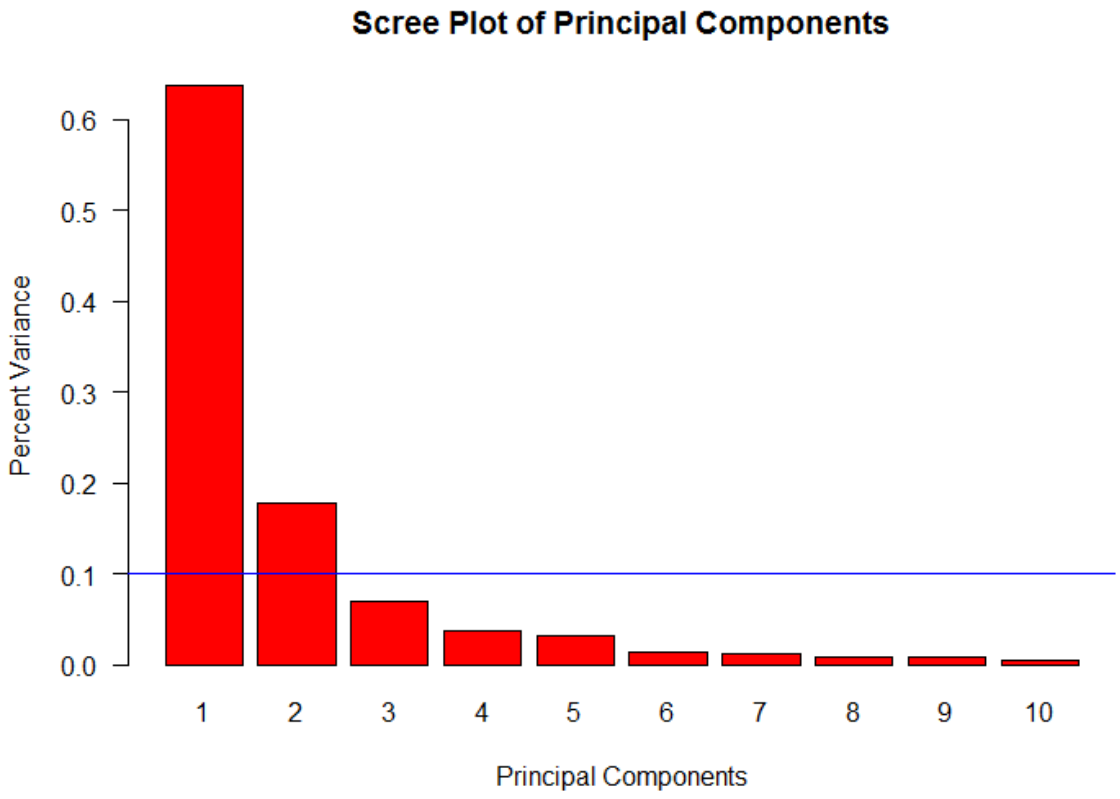

Fig. 3

```
Importance of Components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Std. Dev.                2.471841   1.192511   0.751516   0.705905   0.557774
Proportion of Variance   0.611000   0.142208   0.056477   0.049830   0.031111
Cumulative Proportion    0.611000   0.753208   0.809686   0.859516   0.890627

                          Comp.6     Comp.7     Comp.8     Comp.9     Comp.10
Std. Dev.                0.542537   0.484515   0.475775   0.419593   0.402738
Proportion of Variance   0.029434   0.023475   0.022636   0.017605   0.016219
Cumulative Proportion    0.920062   0.943538   0.966174   0.983780   1.000000
```

The next step in the analysis is to look at the loadings in order to see how each variable contributes to each principal component. I have calculated what the loadings would be if all of them contributed equally to be 0.316. Therefore, any loading with an absolute value larger than 0.316 is considered large and to have a strong effect on the principal component. In Figure 4 we can see the following:

Principal Component 1 is mostly related to:

- Coding Speed
- Numerical Operations

Principal Component 2 is mostly related to:

- Coding Speed
- Arithmetic Reasoning
- Automotive and Shop Information
- Word Knowledge

Principal Component 3 is mostly related to:

- Coding Speed
- Numerical Operations

fig. 4

```
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
Science  -0.157 -0.276 -0.124        -0.196  0.214 -0.141  0.750 -0.463
Arith    -0.274 -0.364        -0.360  0.412 -0.690
Word     -0.265 -0.321 -0.116 -0.173 -0.763 -0.125        -0.323         0.264
Parag    -0.117 -0.102               -0.188                              -0.958
Numer    -0.444         0.838  0.311
Coding   -0.722  0.580 -0.372
Auto     -0.105 -0.324 -0.255  0.597  0.144        -0.505 -0.373 -0.210
Math     -0.237 -0.273        -0.507  0.285  0.653 -0.184 -0.257
Mechanic -0.136 -0.312 -0.198  0.283  0.260  0.153  0.811
Elec     -0.109 -0.252 -0.152  0.202               -0.106  0.333  0.853
```

After reviewing the loadings, we can see that Principal Components 1 and 3 are may be related systematic tasks. On the other hand, Principal Component 2 may be related to cognitive functions.

Looking at a regression analysis, I first see that only Principal Components 1 and 2 show to be significant. This may be the reason on the scree plot that it is arguable on whether to include Principal Component 3 and why it strongly affected by the same variables as Principal Component 1. With that said, it is still necessary to keep it in the model so that we can explain 80% of the variance. I can also see in the Correlation of the Coefficients Matrix in Figure 5 that there is no correlation among the three principal components with all values of 0.00. As mentioned above, only Principal Components 1 and 2 show to be significant as they produce an extremely low p-value of <2e-16 declaring them to be statistically significant which means we would want to include both of their intercepts. Lastly, we have a R^2 of 0.1218 which is a relatively low number and is a problem when generating precise predictions.

fig. 5

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49417.00     861.95  57.331   <2e-16 ***
pca1         -612.94      44.80 -13.680   <2e-16 ***
pca2        -1105.62      84.77 -13.042   <2e-16 ***
pca3          -96.79     134.77  -0.718    0.473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43820 on 2580 degrees of freedom
Multiple R-squared:  0.1218,   Adjusted R-squared:  0.1208
F-statistic: 119.3 on 3 and 2580 DF, p-value: < 2.2e-16

Correlation of Coefficients:
     (Intercept) pca1 pca2
pca1 0.00
pca2 0.00        0.00
pca3 0.00        0.00 0.00
```

To follow this analysis I incorporated the original data set and created a linear model to compare to the Principal Component Analysis performed above. I removed the Math and Auto variables due to high correlations with other variables. What I found was that some of the coefficients are correlated to each other while others are not. We see that Arithmetic Reasoning, Word Knowledge, Numerical Operations, Mechanical Comprehension and Electronics Knowledge show to be significant. As evidence of model being weak with these variables as predictors of success, the model produced a R^2 of 0.1296. This is extremely close to the Principal Component Analysis provided evidence that the analysis was performed correctly. The results can be seen in Figure 6.

fig. 6

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   752.21    3937.90   0.191  0.84853
Science       611.22     344.70   1.773  0.07632 .
Arith        1340.80     208.83   6.421 1.61e-10 ***
Word         -583.70     242.25  -2.409  0.01604 *
Parag        -590.23     461.13  -1.280  0.20067
Numer         354.69     128.55   2.759  0.00584 **
Coding        -63.53      78.26  -0.812  0.41700
Mechanic      596.57     268.94   2.218  0.02662 *
Elec         1545.53     354.02   4.366 1.32e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43660 on 2575 degrees of freedom
Multiple R-squared:  0.1296,   Adjusted R-squared:  0.1269
F-statistic: 47.94 on 8 and 2575 DF,  p-value: < 2.2e-16

Correlation of Coefficients:
         (Intercept) Science Arith Word  Parag Numer Coding Mechanic
Science  -0.11
Arith     0.30       -0.15
Word     -0.17       -0.39   -0.08
Parag    -0.12       -0.07   -0.15 -0.43
Numer    -0.27        0.00   -0.26 -0.05 -0.13
Coding   -0.14        0.04   -0.05 -0.11 -0.11 -0.49
Mechanic -0.18       -0.18   -0.30  0.06 -0.01  0.05  0.01
Elec     -0.04       -0.28   -0.09 -0.15  0.02  0.03  0.07  -0.40
```

Conclusion

It is not perfectly clear on whether there is a "best" model to use to determine future success with the given data. The principal component model does not give a convincing conclusion on whether there is a combination or set of combinations that are better than another to predict success via future income. A very weak model shows that Coding Speed and Numerical Operations may be the best indicators of success, but again it is a weak model. It may be fair to say without further research and based on the analysis performed, factors such as education and socioeconomic status may be better, or at worst worthy, indicators of success.

APPENDIX 1

R CODE

```
## Load in dataset
demographics = read.csv("C:/Users/Brad_2/Desktop/DemographicsData2.csv")

## Display desciptive statistics
sapply(demographics, mean, na.rm = FALSE)
sapply(demographics, sd, na.rm = FALSE)
sapply(demographics, median, na.rm = FALSE)
sapply(demographics, min, na.rm = FALSE)
sapply(demographics, max, na.rm = FALSE)

##
X=demographics[, -11]
Y=demographics[, 11]
pca.cor=princomp(X, cor = T)
summary(pca.cor)


pca.cor <- princomp(X, retx=TRUE, center=TRUE, scale.=TRUE)
sd <- pca.cor$sdev
loadings <- pca.cor$rotation
scores <- pca.cor$X

## Determines cutoff and prints loadings with cutoff requirement to remove
"unimporant" loadings.
cut <- sqrt(1/ncol(X))
print(cut)
loadings(pca.cor, digits = 3, cutoff = cut || cutoff < -1*cut, sort = TRUE)

## Plots bar chart with red line that intercepts bars on bar chart.
## The line is the point at which all ten principal components would
## equally contribute to the variance.
var <- sd^2
var.percent <- var/sum(var)
dev.new()
barplot(var.percent, main = "Scree Plot of Principal Components",
xlab="Principal Components",
        ylab="Percent Variance", names.arg=1:length(var.percent), las=1,
ylim=c(0,max(var.percent)),
        col="red")
abline(h=1/ncol(X), col="blue")


## Principal components 1 and 2 individual % variance explained and combined
varinace explained.
var.percent[1:3]
sum(var.percent[1:3])

## Perform regression analysis on PC1, PC2, and PC3
dim(pca.cor$scores)
pca1 = pca.cor$scores[,1]
pca2 = pca.cor$scores[,2]
pca3 = pca.cor$scores[,3]
```

```
pclm = lm(Y ~ pca1 + pca2 + pca3)
summary.lm(pclm, correlation = T)

## Took out Math and Auto because of high correlations with other predictors
lm2 = lm(Y ~ Science + Arith + Word + Parag + Numer + Coding + Mechanic +
Elec, data = demographics)
summary.lm(lm2, correlation = T)

## Added Math and Auto back in to see what happens
lm3 = lm(Y ~ Science + Arith + Word + Parag + Numer + Coding + Mechanic +
Elec + Math + Auto, data = demographics)
summary.lm(lm3, correlation = T)

## Here Income2005 is modeled by everything but Income2005
lm4 = lm(Income2005 ~ ., data = demographics)
summary.lm(lm4, correlation = T)
```