Hw 6

r12631055

2024-11-04

Problem 1

L. R. Iannaccone argues that the explanation for differences between denominations has predictive power and is “unidimensional.” In support of the theory, he supplied data on 17 Protestant denominations plus Catholics. To get a measure of “distinctiveness” (strictness of discipline), he averaged scores from 16 experts on a seven-point scale for the criteria: “Does the denomination emphasize maintaining a separate and distinctive life style or morality in personal and family life, in such areas as dress, diet, drinking, entertainment, uses of time, marriage, sex, child rearing, and the like? Or does it affirm the current American mainline life style in these respects?” These scores and fou survey-based measures appear in Display 17.22. Attendance is the average percentage of weeks that individuals attended a church meeting. Nonchurch is the average number of secular organizations to which members belong. Strong is the average percentage of members that describe themselves as being strong church members. Income is the average annual income of members. The last four variables come from an extensive survey of church members.

Treat the final four variables as responses characterizing different aspects of the churches’ memberships. To what extent does a single dimension describe differences on all four scales? (Use PCA.) Draw a scatterplot of the first versus the second principal components, and label the points with the church denomination. Interpret the first principal component by comparing the denominations at the ends of this scale. How does that single dimension relate to the author’s distinctiveness scale? To what extent do these data support his theory?

Church deonmination	Distinctiveness	Church attendance (% weekly)	Number of nonchurch memberships	Strong member (%)	Annual Income ($US)
American Baptist	2.5	25.6	1.01	50.6	24,000
Aseemblies of God	4.8	35.4	0.68	58.6	27,100
Catholic	3.0	26.4	1.43	40.0	32,900
Disciples of Christ	2.1	24.3	2.58	47.0	28,600
Episopal	1.1	17.3	1.93	32.0	39,000
Evaneglical Lutheran	2.7	23.0	1.71	41.5	33,700
Jhovah’s Witness	6.0	33.6	0.38	60.6	26,300
Methodist	1.8	19.1	1.56	30.6	32,800
Missouri Synod Lutheran	3.6	27.5	1.76	47.7	35,100
Mormon	5.4	37.8	1.73	70.2	31,600

Church deonmination	Distinctiveness	Church attendance (% weekly)	Number of nonchurch memberships	Strong member (%)	Annual Income ($US)
Nazarene	4.5	33.1	0.86	48.1	31,600
Presbyterian	1.6	21.2	1.88	32.4	37,100
Quaker	4.1	29.6	1.89	58.3	32,500
Reformed Church	2.8	36.7	1.12	61.4	30,400
Seventh Day Adventist	5.8	28.5	0.61	58.7	29,700
Southern Baptist	4.0	25.0	1.13	44.8	30,400
Unitarian	1.6	13.2	2.79	40.8	42,700
United Church of Christ	1.3	19.2	1.56	33.6	40,200

First, loading the data:

```
religion <- read.csv("religion.csv", header = TRUE)
```

Let's take a look:

```
str(religion)
```

```
## 'data.frame':    18 obs. of  6 variables:
##  $ Denomination: chr  "American Baptist" "Assemblies of God" "Catholic" "Disciples of Chri
st" ...
##  $ Distinct    : num  2.5 4.8 3 2.1 1.1 2.7 6 1.8 3.6 5.4 ...
##  $ Attend      : num  25.6 35.4 26.4 24.3 17.3 23 33.6 19.1 27.5 37.8 ...
##  $ NonChurch   : num  1.01 0.68 1.43 2.58 1.93 1.71 0.38 1.56 1.76 1.73 ...
##  $ PctStrong   : num  50.6 58.6 40 47 32 41.5 60.6 30.6 47.7 70.2 ...
##  $ AnnInc      : int  24000 27100 32900 28600 39000 33700 26300 32800 35100 31600 ...
```

Use it to do PCA, we want to know the four variables as responses characterizing different aspects of the churches' memberships.

```
response_vars <- religion[, c("Attend", "NonChurch", "PctStrong", "AnnInc")]
```

Use prcomp function:

```
pca_result <- prcomp(response_vars, scale. = TRUE)
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7113 0.8002 0.57198 0.32260
## Proportion of Variance 0.7321 0.1601 0.08179 0.02602
## Cumulative Proportion  0.7321 0.8922 0.97398 1.00000
```

See also table of loadings

```
pca_result$rotation
```

```
##                  PC1        PC2        PC3         PC4
## Attend     0.5432183 -0.2864479  0.2879025 -0.73482897
## NonChurch -0.4496003 -0.7147684 -0.4783436 -0.24114934
## PctStrong  0.5016987 -0.5849047  0.0839476  0.63177341
## AnnInc    -0.5010706 -0.2548335  0.8253801  0.05230436
```

Extract the principal component scores and convert them into a data frame

```
pca_scores <- as.data.frame(pca_result$x)
str(pca_scores)
```

```
## 'data.frame':    18 obs. of  4 variables:
##  $ PC1: num  1.258 2.271 -0.334 -0.556 -2.35 ...
##  $ PC2: num  0.847 0.252 0.416 -0.892 0.317 ...
##  $ PC3: num  -1.1062 0.1187 0.0392 -1.5704 0.2665 ...
##  $ PC4: num  0.336 -0.104 -0.38 -0.257 0.022 ...
```

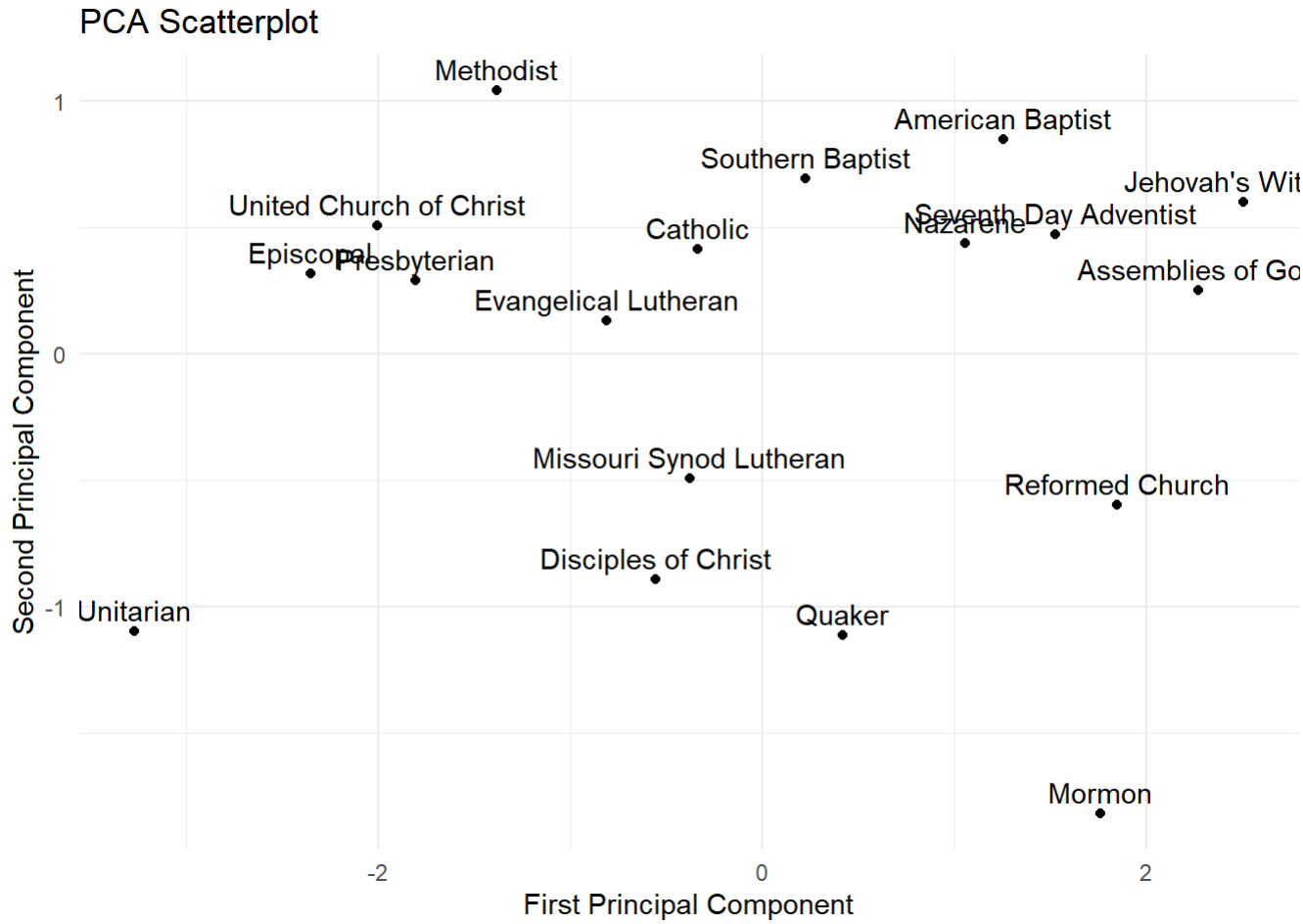
Ensure that the denomination names are added to the principal component data frame

```
pca_scores$Denomination <- religion$Denomination
str(pca_scores)
```

```
## 'data.frame':    18 obs. of  5 variables:
##  $ PC1         : num  1.258 2.271 -0.334 -0.556 -2.35 ...
##  $ PC2         : num  0.847 0.252 0.416 -0.892 0.317 ...
##  $ PC3         : num  -1.1062 0.1187 0.0392 -1.5704 0.2665 ...
##  $ PC4         : num  0.336 -0.104 -0.38 -0.257 0.022 ...
##  $ Denomination: chr  "American Baptist" "Assemblies of God" "Catholic" "Disciples of Chri
st" ...
```

PCA scatterplot:

```
library(ggplot2)
ggplot(pca_scores, aes(x = PC1, y = PC2, label = Denomination)) +
  geom_point() +
  geom_text(vjust = -0.5) +
  theme_minimal() +
  labs(title = "PCA Scatterplot", x = "First Principal Component", y = "Second Principal Comp
onent")
```



Interpretation of the First Principal Component (PC1) Proportion of Variance

From the PCA results:

```
The first principal component (PC1) explains 73.21% of the variance in the data.
PC1 is the most significant dimension for summarizing differences among denominations.
```

Interpretation of PC1 from the Scatterplot

From the scatterplot:

```
On the positive end of PC1, denominations like Assemblies of God, Jehovah's Witness, Mormon,
and Reformed Church are located. These groups may be associated with higher levels of church
attendance, stronger member identification, and lower involvement in secular organizations.
On the negative end of PC1, denominations like Unitarian, Episcopal, and Presbyterian are pre
sent. These groups might have lower levels of religious participation and stronger ties to se
cular activities or higher income levels.
```

Relation to Distinctiveness Scale

```
PC1 correlates strongly with the distinctiveness scale because it captures behaviors like:
    Church attendance (higher for more distinctive denominations),
    Fewer secular memberships (lower for more distinctive denominations),
    Strong self-identification with the church.
Denominations on the positive end of PC1 likely have higher scores on the distinctiveness sca
le, emphasizing a separate and disciplined lifestyle. For example, Jehovah's Witness and Morm
ons are known for their strict religious and moral practices, aligning with higher distinctiv
eness.
```

Support for the Author’s Theory

```
Author's Theory: Distinctiveness predicts differences between denominations in terms of parti
cipation and discipline.
Support from Data:
    The data strongly supports this theory because denominations that score higher on PC1 (in
dicating distinctiveness) also exhibit higher levels of religious engagement (e.g., attendanc
e, strong membership, reduced secular involvement).
    PC1 effectively aligns with distinctiveness, explaining 73.21% of the variation.
```

Conclusion

The single dimension (PC1) strongly reflects the distinctiveness scale proposed by the author. This alignment validates the theory that stricter, more disciplined denominations emphasize higher levels of religious engagement while discouraging secular activities.

Problem 2

Magnetic resonance imaging (MRI) can be used to obtain cross-sectional images through living tissue. One application of MRI is in estimating pig fatness. MRI images were col- lected at 13 equal spacings (approximately 8 cm apart) along the full-body lengths of 12 pigs. The pigs were subsequently killed and dissected to determine their actual fat percentages. It is desired to use MRI for predicting pig fat prior to killing a pig. Some redundancies exist in the MRI measurements, and not all 13 measurements are needed. Two general approaches are possible for developing a prediction equation: Use a variable selection procedure, such as forward selection, to identify a subset of measurements and a regression model for predict- ing pig fat; or use principal components analysis on the MRI measurements to identify a few linear combinations that explain most of the variability, and then regress actual pig fat percentage on the meaningful combinations suggested by these. (a) What are the relative advantages and disadvantages of these two approaches? (b) As a practical matter, it is time-consuming to interpret all 13 images. If the goal is to develop a prediction model that requires the fewest MRI measurements, which of the two strategies is preferable? (c) use principal components analysis to find two or three linear combinations of the MRI measurements of pig fat that explain most of the variability in the measurements (M1 through M13). Try to find meaningful linear combinations (for example, the average of all 13) that these suggest.

a. Advantages and Disadvantages of the Two Methods Method 1: Variable Selection (e.g., Stepwise Selection)

   Advantages: Identifies the MRI measurements with the greatest influence on the prediction model. The resulting model is interpretable, as the selected variables directly correspond to specific MRI measurements. Simple and straightforward to compute. Disadvantages: Susceptible to multicollinearity. If some measurements are highly correlated, the results may become unstable. Ineffective at handling redundancy among variables, potentially ignoring the overall structure of the measurements.

Method 2: Principal Component Analysis (PCA)

```
Advantages:
    Effectively handles multicollinearity among multiple variables.
    Reduces redundant data, extracting a few linear combinations that explain most of the var
iability.
    Results in a more compact model, suitable for large datasets.
Disadvantages:
    Principal components are linear combinations, making their practical interpretation diffi
cult.
    Requires additional steps to incorporate principal components into a regression model.
```

b. Prediction Model with the Fewest MRI Measurements

If the goal is to develop a prediction model requiring the fewest MRI measurements, the following is recommended:

```
Stepwise Selection is preferable:
    Stepwise selection can directly identify the most important MRI measurements for predicti
ng pig fat.
    The resulting model will involve fewer variables, making it simpler and more interpretabl
e.

In the case of PCA:
    Although PCA reduces redundancy, the principal components are linear combinations of all
MRI measurements and do not directly reduce the number of measurements needed.
```

Conclusion: If the goal is to minimize the number of MRI measurements, stepwise selection is more appropriate. (c) Using PCA to Identify Meaningful Linear Combinations

Here is how to use R to perform PCA and address the problem: 1. Perform Principal Component Analysis

```
pig <- read.csv("pig.csv", header = TRUE)
```

Let's take a look:

```
str(pig)
```

```
## 'data.frame':    12 obs. of  14 variables:
##  $ Fat: num  17.1 17.3 23.1 23.3 27.2 ...
##  $ M1 : num  29.2 18.8 35.5 26 39.5 ...
##  $ M2 : num  29.09 9.65 14.16 17.22 26.16 ...
##  $ M3 : num  10 11.9 12.5 15.2 19.9 ...
##  $ M4 : num  15.1 14.5 19.2 19.7 25 ...
##  $ M5 : num  14.2 13.9 21.3 17.7 28 ...
##  $ M6 : num  18.1 13.7 18 20.9 19.9 ...
##  $ M7 : num  14.2 14.3 14.4 21.2 18 ...
##  $ M8 : num  10.8 12.4 18.4 18 21.8 ...
##  $ M9 : num  17.5 17.3 18.6 17.5 25.2 ...
##  $ M10: num  19.3 19.6 28.5 20.8 31.5 ...
##  $ M11: num  13 20.2 26 13.8 32.8 ...
##  $ M12: num  12.2 18.7 14.1 16.9 17.7 ...
##  $ M13: num  10.7 16.9 18.9 17.9 22.4 ...
```

2. Select 2 Principal Components that Explain Most Variance

Choose the top components that cumulatively explain 80%-90% of the variance, as indicated in the summary output. 3. Interpret the Principal Components

Analyze the principal component loadings (pca_result$rotation) to determine the contributions of each MRI measurement to the components. For example:

```
If a principal component has similar weights for all measurements, it may represent the "aver
age of all measurements."
If a principal component is heavily influenced by M1, M2, and M3, it may represent specific r
egional characteristics of the pig body.
```

4. Build a Regression Model

Use the selected principal components to predict pig fat percentages:

```
# extract data from M1 to M13
mri_data <- pig[, 2:14]

# PCA
pca_result <- prcomp(mri_data, scale. = TRUE)

summary(pca_result)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.3250 0.82659 0.72324 0.54481 0.45290 0.33979 0.26030
## Proportion of Variance 0.8505 0.05256 0.04024 0.02283 0.01578 0.00888 0.00521
## Cumulative Proportion  0.8505 0.90301 0.94325 0.96608 0.98186 0.99074 0.99595
##                           PC8    PC9    PC10    PC11      PC12
## Standard deviation     0.17813 0.1249 0.07178 0.01322 1.834e-16
## Proportion of Variance 0.00244 0.0012 0.00040 0.00001 0.000e+00
## Cumulative Proportion  0.99839 0.9996 0.99999 1.00000 1.000e+00
```

```
principal_components <- as.data.frame(pca_result$x[, 1:2])
str(principal_components)
```

```
## 'data.frame':    12 obs. of  2 variables:
##  $ PC1: num  4.436 4.428 2.913 3.361 0.994 ...
##  $ PC2: num  1.6336 -1.1855 -0.3891 -0.0658 0.035 ...
```

```
principal_components$Fat <- pig$Fat
str(principal_components)
```

```
## 'data.frame':    12 obs. of  3 variables:
##  $ PC1: num  4.436 4.428 2.913 3.361 0.994 ...
##  $ PC2: num  1.6336 -1.1855 -0.3891 -0.0658 0.035 ...
##  $ Fat: num  17.1 17.3 23.1 23.3 27.2 ...
```

Regresion:

```
model <- lm(Fat ~ PC1 + PC2, data = principal_components)
summary(model)
```

```
##
## Call:
## lm(formula = Fat ~ PC1 + PC2, data = principal_components)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.66587 -1.15960 -0.03254  1.27114  2.53631
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.1533     0.5121  60.837 4.42e-13 ***
## PC1          -2.8177     0.1609 -17.517 2.91e-08 ***
## PC2          -1.0788     0.6471  -1.667     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.774 on 9 degrees of freedom
## Multiple R-squared:  0.9718, Adjusted R-squared:  0.9655
## F-statistic: 154.8 on 2 and 9 DF,  p-value: 1.07e-07
```

Regression Results Interpretation

Based on the output of the linear regression model, here is a detailed interpretation of the results: 1. Regression Model Formula

The regression equation is: $Fat = 31.1533 - 2.8177 \times PC1 - 1.0788 \times PC2$

```
Intercept: 31.1533
    When both PC1 and PC2 are 0, the predicted fat percentage for pigs is 31.1533.
PC1 Coefficient (Estimate): -2.8177
    PC1 has a significant negative effect on fat percentage. For each 1-unit increase in PC1,
the fat percentage decreases by an average of 2.8177.
PC2 Coefficient (Estimate): -1.0788
    PC2 has a smaller effect on fat percentage and is not statistically significant (p-value
> 0.05).
```

2. Significance Testing

   PC1: p-value: 2.91e-08 (highly significant, $p < 0.001$). This indicates that PC1 is an important predictor of fat percentage. PC2: p-value: 0.13 (not significant, $p > 0.05$). This suggests that PC2 has a limited and possibly unnecessary effect on fat percentage.

3. Model Fit

   Multiple R-squared: 0.9718 This indicates that the model explains 97.18% of the variability in fat percentage, demonstrating excellent fit. Adjusted R-squared: 0.9655 After accounting for the number of predictors, the model still explains 96.55% of the variability. F-statistic: 154.8 (p-value = 1.07e-07) The overall model is highly significant ($p < 0.001$).

4. Residual Analysis

   Residual Range: Minimum -2.66587, Maximum 2.53631. Residual Standard Error: 1.774 The model's prediction error is small, with most predicted values deviating within ±1.774 of the observed values.

Conclusion and Recommendations

```
PC1 is the Key Variable:
    PC1 is the primary explanatory variable for fat percentage and is highly significant (p < 0.001). This aligns with the PCA results, where PC1 explains 85.05% of the variance.
    The negative coefficient of PC1 indicates that as PC1 increases (possibly representing a reduction in overall fat distribution), fat percentage decreases.

PC2 is Not Significant:
    PC2 is not statistically significant (p = 0.13) and likely contributes minimally to the model.
    Consider using only PC1 as a predictor to further simplify the model.

Excellent Model Fit:
    The high R-squared value and low residual standard error indicate that the model predicts fat percentage accurately.

Recommendations:
    If model simplicity is a priority, use only PC1.
    If higher accuracy is needed, retain PC2 but evaluate its practical importance and cost.
```

This model demonstrates the feasibility of using PCA to simplify MRI measurements while constructing an accurate and effective prediction model for pig fat percentage.