Homework6

Sherine George

11 December, 2023

The cities.csv dataset is a subset of the 500 Cities Project of the Centers for Disease Control and Prevention (CDC). It contains population of 123 cities of the US. In particular, these are the columns available in the cities.csv dataset:

- City: the name of the city
- Arthritis among adults aged >=18 Years: the prevalence of arthritis in the adult population of the city
- Chronic kidney disease among adults aged >=18 Years: the preva- lence of chronic kidney disease in the adult population of the city
- Chronic obstructive pulmonary disease among adults aged >=18 Years: the prevalence of chronic obstructive pulmonary disease in the adult pop- ulation of the city
- Coronary heart disease among adults aged >=18 Years: the preva- lence of chronic heart disease in the adult population of the city
- Current lack of health insurance among adults aged 18-64 Years: the proportion of the adult population in the city that is not covered by health insurance
- Diagnosed diabetes among adults aged >=18 Years: the prevalence of diabetes in the adult population of the city
- High cholesterol among adults aged >=18 Years: the prevalence of high colesterol in the adult population of the city
- No leisure-time physical activity among adults aged >=18 Years: the proportion of the adult population in the city that does not participate in any physical activity.

Question 1 - 5 points

Load the cities.csv dataset in R. Drop the City column, since we will not need it in our analysis. Also, rename the other columns as follows:

- Arthritis among adults aged >=18 Years → arthritis
- Chronic kidney disease among adults aged >=18 Years→kidney_disease
- Chronic obstructive pulmonary disease among adults aged >=18 Years → copd
- Coronary heart disease among adults aged >=18 Years→heart_disease
- Current lack of health insurance among adults aged 18-64 Years no_health_insurance
- Diagnosed diabetes among adults aged >=18 Years → diabetes
- High cholesterol among adults aged >=18 Years→high_colesterol
- No leisure-time physical activity among adults aged >=18 Years → no_exercise

Solution:

```
df <- read.csv("D:/Downloads/cities.csv")
df <- df[, !(names(df) %in% c("City"))]
# 1:arthritis
colnames(df)[1] <- "arthritis"
# 2:kidney_disease
colnames(df)[2] <- "kidney_disease "
# 3:copd
colnames(df)[3] <- "copd"
# 4:heart_disease
colnames(df)[4] <- "heart_disease"
# 5:no_health_insurance
colnames(df)[5]<- "no_health_insurance"
# 6:diabetes
colnames(df)[6]<- "diabetes"
# 7:high_colesterol
colnames(df)[7] <- "high_colesterol"
# 8:no_exercise
colnames(df)[8] <- "no_exercise"
```

Question 2 - 5 points

Apply Principal Component Analysis (PCA) to the dataset that you just cre- ated. Make sure to specify that the variables are centered (i.e., their empirical mean is set to 0) and also scaled (i.e., their empirical standard deviation is set to 1) in the prcomp function.

Solution:

```
pca_df <- prcomp(df, scale = TRUE)
pca_df$center
```

```
##           arthritis      kidney_disease                copd       heart_disease
##          0.22444715          0.02968293          0.06443089          0.06149593
## no_health_insurance            diabetes     high_colesterol         no_exercise
##          0.20389431          0.11730081          0.31866667          0.28982927
```

```
pca_df$scale
```

```
##           arthritis      kidney_disease                copd       heart_disease
##         0.047626312         0.005279588         0.019329787         0.012530863
## no_health_insurance            diabetes     high_colesterol         no_exercise
##         0.089303756         0.029596171         0.024780236         0.077105390
```

Question 3 - 15 points

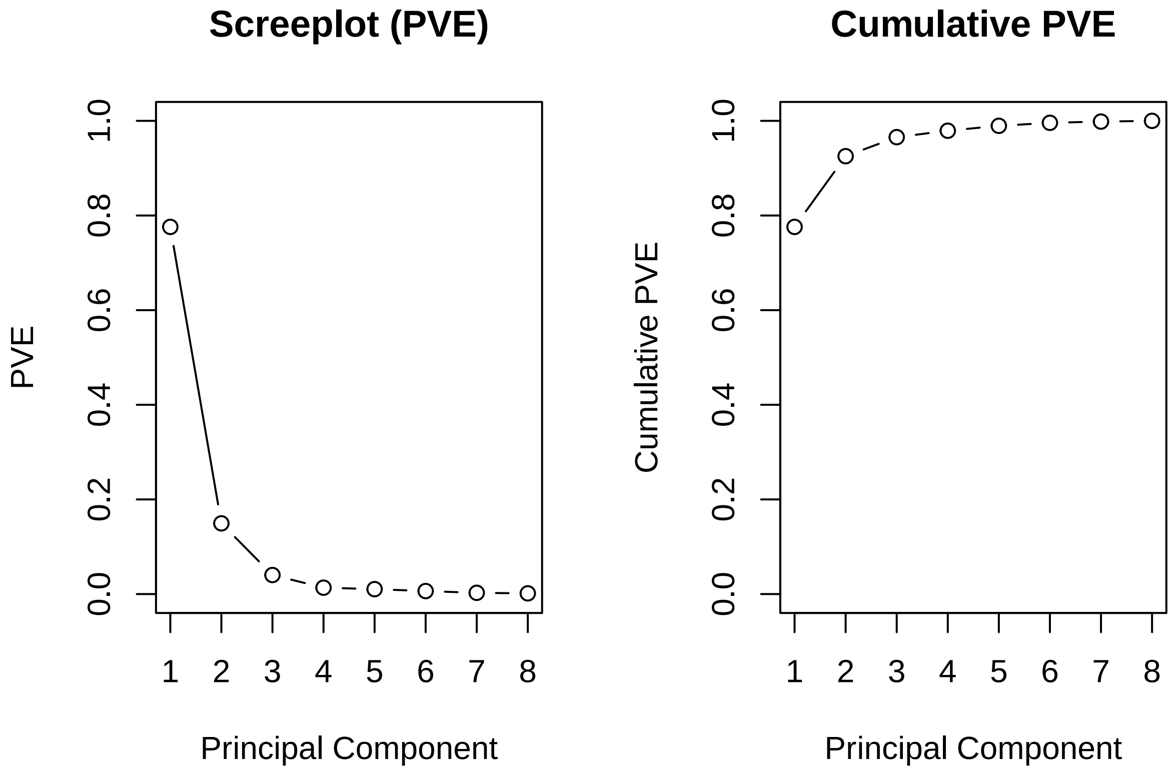
Compute and plot the proportion of variance explained by the principal com- ponents and the cumulative proportion of variance explained by the principal components.

Solution:

```
pca_df_vars <- pca_df$sdev^2

pca_df_pve <- pca_df_vars / sum(pca_df_vars)

pca_df_cum_pve <- cumsum(pca_df_pve)
pca_df_cum_pve
```

```
## [1] 0.7759188 0.9253555 0.9656183 0.9791321 0.9895207 0.9958772 0.9985143
## [8] 1.0000000
```

```
pca_df_cum_pve <- cumsum(pca_df_pve)

par(mfrow = c(1, 2))
plot(
    pca_df_pve,
    xlab = "Principal Component",
    ylab = "PVE",
    main = "Screeplot (PVE)",
    type = "b",
    ylim = c(0, 1)
)
plot(
    pca_df_cum_pve,
    xlab = "Principal Component",
    ylab = "Cumulative PVE",
    main = "Cumulative PVE",
    type = "b",
    ylim = c(0, 1)
)
```



Question 4 - 15 points

The nominal dimension of this dataset is 8 (i.e., we have 8 variables available in total). Based on the plot of the cumulative proportion of variance explained by the principal components that you just produced, what do you think is the effective dimensionality of this dataset (i.e., are the observations in these data concentrated on a smaller subspace and what is the dimension of this subspace)? Explain.

Solution:

The plot reveals that the number 3 signifies the ‘elbow’ point, which holds significant importance as it indicates the optimal dimensionality of the dataset. This observation is rooted in the structure of principal components, which are arranged based on the variance they capture.

Typically, the initial principal components are the most crucial because they capture the majority of the important information within the data. These components represent the directions where the variance is maximized.

Consequently, the point just before this ‘elbow’ on the plot is generally considered the effective dimensionality of the dataset. At this juncture, a balance is achieved where enough information is retained while avoiding the risks of overfitting or introducing excessive noise into the model.

Question 5 - 10 points

Compute the correlation matrix for the variables of the cities.csv dataset. Af- ter inspecting the correlation matrix, are you surprised that PCA was successful in reducing the dimensionality of this dataset? Explain.

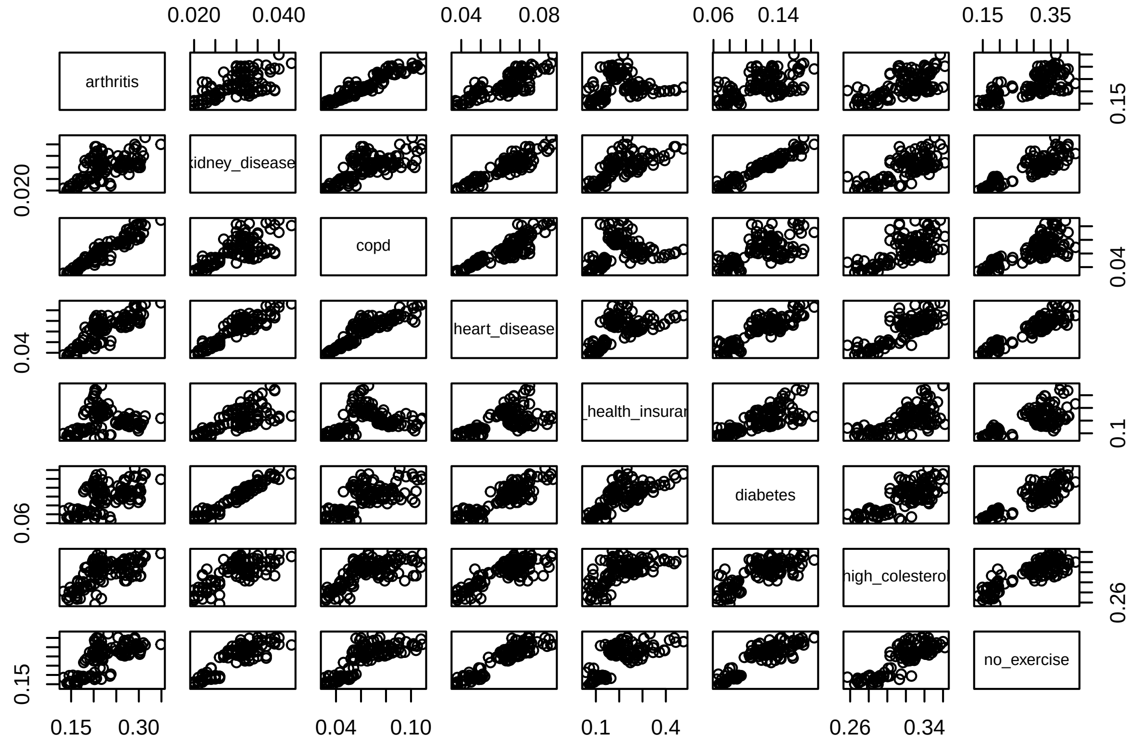
Solution:

```
round(cor(df), 2)
```

```
##                     arthritis kidney_disease copd heart_disease
## arthritis                1.00           0.65 0.94          0.83
## kidney_disease           0.65           1.00 0.73          0.93
## copd                     0.94           0.73 1.00          0.90
## heart_disease            0.83           0.93 0.90          1.00
## no_health_insurance      0.13           0.72 0.25          0.59
## diabetes                 0.58           0.97 0.67          0.88
## high_colesterol          0.63           0.75 0.69          0.82
## no_exercise              0.72           0.87 0.78          0.93
##                     no_health_insurance diabetes high_colesterol no_exercise
## arthritis                          0.13     0.58            0.63        0.72
## kidney_disease                     0.72     0.97            0.75        0.87
## copd                               0.25     0.67            0.69        0.78
## heart_disease                      0.59     0.88            0.82        0.93
## no_health_insurance                1.00     0.74            0.67        0.69
## diabetes                           0.74     1.00            0.76        0.88
## high_colesterol                    0.67     0.76            1.00        0.85
## no_exercise                        0.69     0.88            0.85        1.00
```

The outcome doesn’t come as a surprise. Upon reviewing the correlation matrix, it becomes evident that correlations equal to or exceeding 0.7 can be classified as highly correlated within this context. With this criterion in consideration, it is noticeable that three of the components exhibit a more pronounced relationship with other components.

```
pairs(df)
```

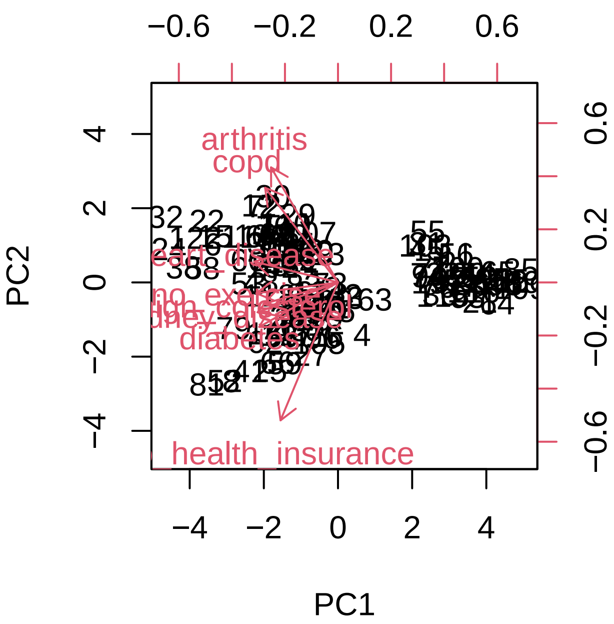


Question 6 - 15 points

Let's focus on the first 2 principal components found for the cities.csv dataset. Produce the biplot for the first 2 principal components and interpret it.

Solution:

```r
# Biplot
par(mfrow = c(1, 2))
biplot(pca_df, scale = 0)
```



The red vectors in the chart signify the influence of each variable on the principal components. Notably, variables like “no_exercise,” “high_cholesterol,” “kidney_disease,” and “diabetes” appear to be closely interrelated, aligning with our hypothesis from Question 5. Additionally, the variable “no_health_insurance” exhibits a weaker correlation compared to the others.

The numerical values on the chart correspond to the scores assigned to the first two Principal Components, and the orange vectors represent their loading vectors. It’s essential to emphasize the significance of the magnitude of these loading vectors.

On the chart, the x-axis represents the first principal component (PC1), which accounts for approximately 77.59% of the data’s variance. Conversely, the y-axis represents the second principal component (PC2), explaining roughly 14.94% of the data’s variance.

We observe that the following variables have a more substantial loading on the second principal component:

Arthritis (indicating the prevalence of arthritis in the adult population of the city) COPD (reflecting the prevalence of chronic obstructive pulmonary disease in the adult population of the city) No_health_insurance (indicating the proportion of the adult population in the city lacking health insurance coverage) In contrast, the remaining variables demonstrate a higher loading on the first principal component, indicating their stronger association with it.

Question 7 - 5 points

In the last month, the Product organization of your web company ran 100 exper- iments to evaluate ideas to improve the User Experience (UX) of its customers. In each experiment, a Product Engineering team would be responsible to enable a different UX for a randomly selected group of users. For instance, randomly selected users would see different colors for some of the navigation buttons, different positioning of the search bar on the page, modified text for different components of the page, etc. At the end of each experiment, the Product Man- ager in charge of the experiment would use a tool to compute the p-value for the one-sided t-test associated with following statistical hypothesis test: (H0 : user engagement is not higher with the new user experience H1 : user engagement higher with the new user experience. The experiments.csv file contains the p-values of the 100 experiments that were run in the last month. Load the dataset in R.

Solution:

```r
exp <- read.csv("D:/Downloads/experiments.csv")
```

Question 8 - 5 points

The Product organization of your web company has an internal policy by which the default significance level that should be used when evaluating the results of UX experiments for the company's website is = 0.10. How many experiments were found to generate a statistically significant UX improvement at the = 0.10 level over the last month?

Solution:

```r
sum(exp$p <= 0.1)
```

```
## [1] 47
```

47 experiments were found to generate a statistically significant UX improvement at the = 0.10 level over the last month.

Question 9 - 10 points

As we learned in class, the Family-Wise Error Rate (FWER) across 100 statis- tical tests - each carried out at the = 0.10 significance level - is much larger than 0.10. Assuming that these statistical tests were independent, what is the effective FWER that the Product team incurred into by not accounting for the problem of multiple testing?

Solution:

```r
# Number of tests
number_of_tests <- 100

# Significance level
alpha <- 0.10


# Calculate effective FWER
effective_fwer <- 1 - (1 - alpha)^number_of_tests

# Print the result
cat("Effective FWER:", effective_fwer)
```

```
## Effective FWER: 0.9999734
```

0.9999734 is the effective FWER that the Product team incurred into by not accounting for the problem of multiple testing.

Question 10 - 15 points

Using the Benjamini-Hochberg method to account for the problem of multiple testing, provide the list of experiment ids that likely resulted in an improvement of the user experience. Control the False Discovery Rate (FDR) at the level q = 0.10. You can take a look at chapter 13.6.3 of ISL to learn how to use the p.adjust function to perform different types of multiple hypothesis tests, including the Benjamini-Hochberg method. Alternatively, feel free to provide your own implementation of the Benjamini-Hochberg method.

Solution:

```r
# Perform Benjamini-Hochberg adjustment
adjusted_p_values <- p.adjust(exp$p, method = "BH")

# Identify significant experiments based on adjusted p-values
significant_experiments <- which(adjusted_p_values <= 0.1)

# Print the list of experiment ids likely resulting in an improvement
cat("Experiment IDs:", significant_experiments, "\n")
```

```
## Experiment IDs: 1 10 12 27 29 33 34 38 58 66 67 68 76 78 82 83
```