Statistical Foundations of Business Analytics
46-883

Carnegie Mellon University
Tepper School of Business

Homework 6

The `cities.csv` dataset is a subset of the 500 Cities Project of the Centers for Disease Control and Prevention (CDC). It contains information about the prevalence of various medical conditions in the 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 prevalence 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 population of the city
- `Coronary heart disease among adults aged >=18 Years`: the prevalence 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`

Question 2 - 5 points

Apply Principal Component Analysis (PCA) to the dataset that you just created. 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.

Question 3 - 15 points

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

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.

Question 5 - 10 points

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

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.

Question 7 - 5 points

In the last month, the Product organization of your web company ran 100 experiments 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 Manager 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:

$$\begin{cases} H_0 : \text{user engagement is not higher with the new user experience} \\ H_1 : \text{user engagement higher with the new user experience.} \end{cases}$$

The `experiments.csv` file contains the p-values of the 100 experiments that were run in the last month. Load the dataset in R.

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 $\alpha = 0.10$. How many experiments were found to generate a statistically significant UX improvement at the $\alpha = 0.10$ level over the last month?

Question 9 - 10 points

As we learned in class, the Family-Wise Error Rate (FWER) across 100 statistical tests - each carried out at the $\alpha = 0.10$ 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?

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.