PCA Assignment
Behavioral Research: Statistical Methods
Spring 2024

Instructions

Please read the following instructions carefully before attempting.

1. If you are coding in R, make sure to submit only the pdf file after knitting. File name format should be <Roll Number>_PCAFA_Class_Activity.pdf
2. If you are using any other language/software, submit the appropriate file following the same nomenclature.

1 Wine Dataset

For the purposes of this assignment, we will be using the Wine Dataset. The dataset for all questions are available at this link. The required sheet name is mentioned in each question. Download the dataset and load in your code accordingly.

The dataset is used to classify types of wine based on their chemical properties, containing 13 features (chemical properties) and 1 target variable (type of wine).

2 Principal Component Analysis

1. Generate a correlation heatmap for all the features in the dataset.
2. Perform parallel analysis on the wine dataset. How does this method's recommendation for the number of principal components to retain compare to the proportion of variance explained criterion?
3. Given the wine dataset, perform PCA and report the proportion of variance explained by the first three principal components. What does this tell you about the dataset?
4. Create a scatter plot of the first two principal components of the wine dataset. How well do these components separate the different types of wine? Provide an interpretation of your findings.

3 Chi-Square Test - Goodness of Fit

The dataset containing information on various properties of wines, includes a categorical variable 'Type' that classifies wines into different categories. The dataset includes measurements for 178 wines, classified into 3 distinct types. A preliminary analysis has revealed the following distribution of wine types:

- Type 1: 59 occurrences
- Type 2: 71 occurrences
- Type 3: 48 occurrences

Conduct a Chi-Square Goodness of Fit Test to determine whether the observed distribution of wine types significantly deviates from an expected distribution, where each type of wine is equally represented in the dataset, by following the steps:

1. Formulate the null and alternative hypotheses for the Chi-Square Goodness of Fit Test in the context of this study. Consider the null hypothesis to propose that the observed distribution of wine types does not significantly differ from an equal distribution amongst the three types.
2. Using the observed frequencies of wine types and the assumption of equal distribution as the expected frequency, perform the Chi-Square Goodness of Fit Test. Calculate the chi-square statistic and the corresponding p-value.

3. Interpret the results of the Chi-Square Goodness of Fit Test. Discuss whether the null hypothesis can be rejected based on the p-value and the significance level (α). Assume $\alpha = 0.05$ for this analysis.

We want to check if the distribution of wine types in the dataset matches an expected distribution of an equal division amongst the types.

4 Chi-Square Test - Independence

There is a general conception that people like to choose their type of wine based on alcohol content. In our dataset,

1. For this question, categorize the 'Alcohol' content into three levels: 'Low', 'Medium', and 'High', based on percentile divisions (0-33rd percentile as "Low", 34-66th percentile as "Medium", and 67-100th percentile as "High")
2. Create a contingency table (cross-tabulation) showing the frequencies of each combination of 'Type' and 'Alcohol Category' .
3. Formulate and state the null and alternate hypothteses.
4. Perform the Chi-Squared Test for Independence using the contingency table. Calculate the chi-square statistic, degrees of freedom, and the corresponding p-value.
5. Interpret the results of the Chi-Squared Test for Independence. Discuss whether the null hypothesis can be rejected based on the p-value and a significance level (α) of 0.05.