CSC 5825 Term Project Progress Report Obesity Risk Factor Analysis

Chaowei Liu
dept. CSC of Wayne State University
Detroit, Michigan
cliu@wayne.edu

I. Introduction

In the United States, obesity is a problem that cannot be ignored. According to the computed statistics from the Behavioral Risk Factor Surveillance System (BRFSS), 31.10% of the respondents of the survey are obese. This figure is increasing significantly over decades. More reports from medical institutes indicated that obesity has many negative effects on human health. For example, obese people have much higher chances to get cardiovascular disease, diabetes disease, osteoarticular disease, and other kinds of cancers. It is critical to predict and prevention the risk for people of being obese by developing scientific methods of identifying potential possibility of obesity.

In this project, methods of risk factor analysis will be conducted to identify and understand the underlying relationships of human body indicators and health. The goal of this project is to find and describe the factors, also called latent variables, which are quite meaningful but can be only inferred, rather than being directly observable.

The rest of this progress report will be presented as follows. Section II introduces methods that will be used in the project. Section III describes the dataset used in the project, demonstrates the procedures of conduction of the project, results obtained, and corresponding discussion. In the end, a brief conclusion is drawn in Section IV.

II. Methods

A. Methods for Analysis

As the dataset is pre-processed and cleaned, it is ready to analyze it. In this project, we performed principal component analysis (PCA) and factor analysis (FA) on the dataset to figure out the latent variables, i.e. the most important features that have the greatest impacts on health. These two kinds of analysis methods are always mentioned together, but they are not identical. FA aims to seek the latent variables that are behind those surface observable representations. However, on the other hand, PCA selects the features that have the most significant variance.

1) Principle Component Analysis: Principle Component Analysis (PCA) is actually a tool for dimensional reduction. However, the perspective of PCA to reduce the dimension is exactly what we are looking for in this project, since we are aimed to find out the most principle components in our dataset and discard those irrelevant ones which have no or very little impact. The main idea of PCA is to find out the principle components (eigenvectors) and their weights (eigenvalues) by decomposing the covariance matrix.

2) Factor Analysis: The target of Factor Analysis is to figure out the latent components which are hard to be observed, but have a large impact on the results, decisions or behaviors that are difficult to explain directly. This is just like our dataset, among hundreds of features and more than 450,000 instances, it is hard to tell which of them make the greatest contribution to our health. Fortunately, FA is able to do the job well.

III. Experiment, Results, and Discussion

A. Dataset: BRFSS

BRFSS is used in this project. The government has been conducting a survey-like activity that ask people in all over the country a series of health related questions every year since 1984. This is the biggest database, storing information about health and relevant behaviors of more than 400 thousand adults from all the 50 states.

However, the BRFSS dataset is not ready to use directly. Its structure keeps changing year by year. In this project, the latest (2017) version of BRFSS is selected as the target dataset. The dataset file is extremely large (around 1GB) and the data inside is messy. It is really inconvenient to process such a large dataset (it takes approximately 50 seconds to just load). In order to make it simple, tidy and ready to analysis, a lot of filtering and cleaning work needs to be done in advance.

B. Data Filtering and Cleaning

The raw BRFSS dataset contains 358 features, corresponding to very detailed health conditions and behaviors of the respondents. However, not all of these features are useful or suitable for analysis. The website of BRFSS also provides the detailed codebook for BRFSS dataset, with which it is easier to see the meaning of each

Identify applicable funding agency here. If none, delete this.

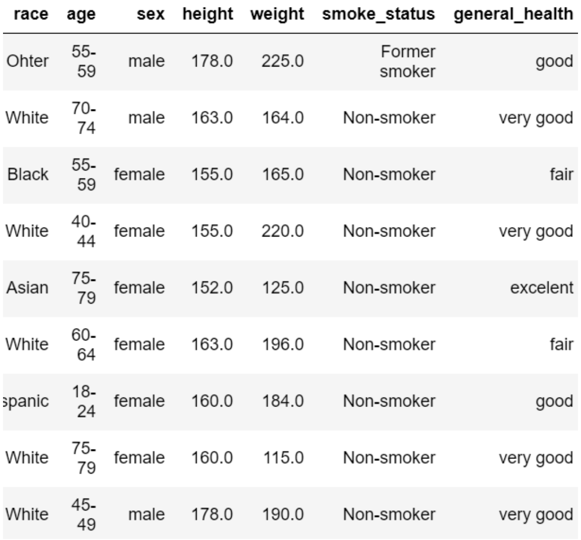
feature and decide which features are helpful for the analysis. There are a lot of features that are irrelevant to human health, for example, the interview date, state, an so on, which have very limited impact on the research of relationship between obesity and health problems.

1) Filtering: After carefully examination of the codebook, 14 out of 358 features are kept for the analysis. They are:

- **race**: Respondent's race, including White, Black, Asian, American / Alaskan Native, Hispanic, and others.
- **age**: Age group in every 5-year categories, e.g. 18 to 24, 25 to 29, so on and so forth.
- **sex**: Sex of the respondent.
- **height**: Height in centimeters.
- **weight**: Weight in pounds.
- **smoking_status**: Computed categories indicating whether a respondent is an everyday smoker, someday smoker, former smoker or non-smoker.
- **general_health**: General health condition of the respondent. The categories are excellent, very good, good, fair and poor.
- **has_cancer**: Number of different types of cancers the respondent has. A blank record indicates that the respondent does not have cancer.
- **has_heart_disease**: This feature indicates that whether or not this respondent has cardiovascular problems.
- **mental_health**: Number of days that a particular respondent feels bad on his or her mental condition.
- **binge_drinker**: This feature indicates that whether this respondent is a binge drinker.
- **has_diabetes**: This feature indicates that whether this respondent has diabetes problems.
- **bmi_label**: Shape of the respondent according to his or her Body mass index (BMI).

It is easy to use Python's Pandas library to extract these features from the whole dataset, and then concatenate the extracted features into a new dataframe. Finally the new dataframe is written to a file "brfss_short.csv".

2) Cleaning: The information in the dataset is still messy and hard to understand. The dataset only stores numbers, instead of any actual meanings of the data. To understand the dataset, one needs to check the codebook variable by variable, which is time consuming since there are 358 variables in the codebook. For example, the feature "race" uses number 1, 2, 3, 4, 5 and 6 to represent White, Black, Asian, American Indian / Alaskan Native, Hispanic and Ohter, respectively. By taking advantage of the replace function in Pandas, we constructed dictionaries for these features and replace the numbers by their actual meanings. In addition, there are a few blanks in the dataset. For example, most respondents' responses to the "**has_cancer**" question is null because only very small number of people have cancer. We fill the blanks with 0, simply indicating that one has no cancer. The cleaned dataset is shown in Figure 1.



race	age	sex	height	weight	smoke_status	general_health
Ohter	55-59	male	178.0	225.0	Former smoker	good
White	70-74	male	163.0	164.0	Non-smoker	very good
Black	55-59	female	155.0	165.0	Non-smoker	fair
White	40-44	female	155.0	220.0	Non-smoker	very good
Asian	75-79	female	152.0	125.0	Non-smoker	excelent
White	60-64	female	163.0	196.0	Non-smoker	fair
spanic	18-24	female	160.0	184.0	Non-smoker	good
White	75-79	female	160.0	115.0	Non-smoker	very good
White	45-49	male	178.0	190.0	Non-smoker	very good

Fig. 1. Filtered and cleaned dataset.

Also, the filtered and cleaned dataset is stored in a "brfss_clean.csv" file for further use.

C. Results and discussion

We load the pre-processed dataset and apply FA and PCA on it. Since we do not know how many latent variables are important, we leave the number of components as default, which will output the same number of features of results, i.e. 13. The result shows that there are 5 out of 13 factors have significant connections with the features, they are: **weight**, **age**, **race**, **mental_health**, **has_cancer** and **smoking_status**. Now let's look at the result returned by PCA. Again, the number of principle components are set to 13, which equals to the total number of features. The purpose of such a setting is to view the importance order of the principle components generated by PCA. Consequently, PCA returns 13 results. However, the order of the first 6 most important features are just the same with the order generated by FA.

IV. Conclusion

In both sets of result, it can be seen that the feature **weight** dominates the importance of all the features. It is not expected that some of the features we selected among all the 358 features are less important than our expectation, such as **binge_drinker**, **bmi_label**, **has_diabetes**, and so on. From the result, we need to reconsider the procedure of data filtering and cleaning, because the method is too subjective. There are a lot more features which are meaningful and have great contributions to our health but are ignored.