Interim Report

Team 08

Department of Statistics	2015-15347	Wonhyeong Choe
Department of Mathematical Sciences	2016-11797	Kihyun Han

1. Introduction

1.1. Research Background

Since the rate of population who completed COVID-19 vaccine injection exceeded 78% in South Korea, South Korea has become one of the countries having the highest vaccination rate. Accordingly, the public health authority has alleviated the social distancing restriction and implemented "With-Corona" policy, thus encouraging many people to participate in social activities, which was reduced due to COVID.

However, there are controversies concerning the vaccine's side effects. Although there are a lot of cases of diseases reported through the press, the authority does not acknowledge the causal relation and correlation of these instances and vaccination. In fact, according to Korea Disease Control and Prevention Agency, a lot of side effect instances such as thrombosis (혈전증), anaphylaxis (아나필락시스), myocarditis (심근염), and pericarditis (심낭염) occurred and the deceased instances has recorded 894, but there are only 6-7 cases where its correlation is admitted.

We proceed the classification by analyzing the factors that have an influence on deaths or life-threatening illnesses with respect to factors such as age, sex, allergies, and COVID vaccine manufacturers – Pfizer, Moderna, and Janssen. We expect that the prediction model makes it possible for people to understand and be aware of the potential danger of vaccination in a statistical and quantitative fashion.

1.2. Data Description

We obtained data from Kaggle, which was collected from VAERS (Vaccine Adverse Event Report System). We analyzed the cases of COVID vaccine side-effects reported from January to September 2021.

2. Methods

2.1. Data Pre-processing

First, we selected variables that can affect the occurrence of death and life-threatening illnesses from the data. Then, we removed observations in which some variables were not recorded and duplicated data. This made the data reduced from to 331589 instances.

In the data, symptoms were recorded as a type of MedDRA (Medical Dictionary for Regulatory Activities) terms. Generally, the MedDRA dictionary contains a five-level hierarchy (Figure 1), and the symptoms in the data are written as “Preferred Term” (PT), which accounts for the fourth hierarchy. We assigned the corresponding “System Organ Class” (SOC), the highest hierarchy with 27 elements, for each of the symptoms. Then we transformed into 27 dummy variables representing whether each individual suffered the symptom in a particular SOC. We discovered that one SOC does not include any individuals. Thus, we removed it, and 26 dummy variables remain.

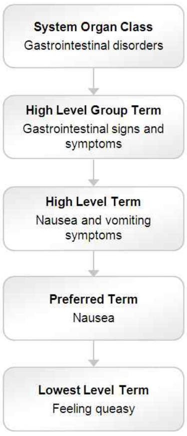


Figure 1 MedDRA symptom hierarchy example

Since ‘History’ and ‘Allergies’ variables were entered as sentences by individuals, which is unstructured data, we applied natural language processing to detect keywords of each variable. To be specific, for the ‘History’ variable, we selected the terms and prefixes that appeared more than 500 times in data. Then we classified the keywords with respect to their medical interpretation into 23 classes. For instance, 4 keywords “kidney”, “nephritis”, “nephr-”, “CKD” are related to kidney diseases, so we put these as same class. Then we develop 23 dummy variables representing whether an individual had history with associated classes. In the similar manner, we make ‘Allergy’ dummy variables to represent whether each included individual has the specific allergy.

In addition, we created two dummy variables representing the vaccine manufacturer, and 11 dummy variables representing each month of vaccination. Also, we wanted to find whether the temperature of the state where a person is vaccinated would affect the rate of illnesses, so we added "MEAN_TEMP" variable, which indicates the mean temperature of a state by month.

For the response variable, we made a dummy variable "THREAT" to represent whether a person has death or life-threatening illnesses. The explanatory and response variables are summarized as follows.

- Explanatory variables: 97 (Continuous: 4, Discrete: 93)

Variables	Description	Variable Type
SYMPTOM1 - SYMPTOM26	Adverse symptoms in MedDRA Term	Transformed into 26 dummy variables representing highest hierarchy.
MODERNA - PFIZER	Vaccine manufacturer	Transformed into 2 dummy variables representing Pfizer, Moderna, and Janssen(default)
VAX_DOSE_SERIES	Number of doses administered	Discrete variable representing the first and the second dose
MEAN_TEMP	Mean temperature of the state by month	Continuous variable
AGE	Age in years	Continuous variable
SEX	Sex	Dummy variable
HOSPITAL	Hospitalized	Dummy variable
HOSPDAYS	Number of days hospitalized	Continuous variable
DISABLE	Disability	Dummy variable
JAN - NOV	Month of vaccination	Transformed into 11 dummy variables representing each month (default: December)
NUMDAYS	Number of days from vaccination date to onset date	Continuous variable
HISTORY1 - HISTORY23	Chronic or long-standing health conditions	Transformed into 23 dummy variables representing history
BIRTH_DEFECT	Congenital anomaly or birth defect	Dummy variable
ALLERGY1 - ALLERGY26	Allergies to medications, food, or other products	Transformed into 26 dummy variables representing history

- Response variable: 1 (Discrete: 1)

Variables	Description	Variable Type
THREAT	Death or life-threatening illnesses	Dummy variable

2.2. Model Design

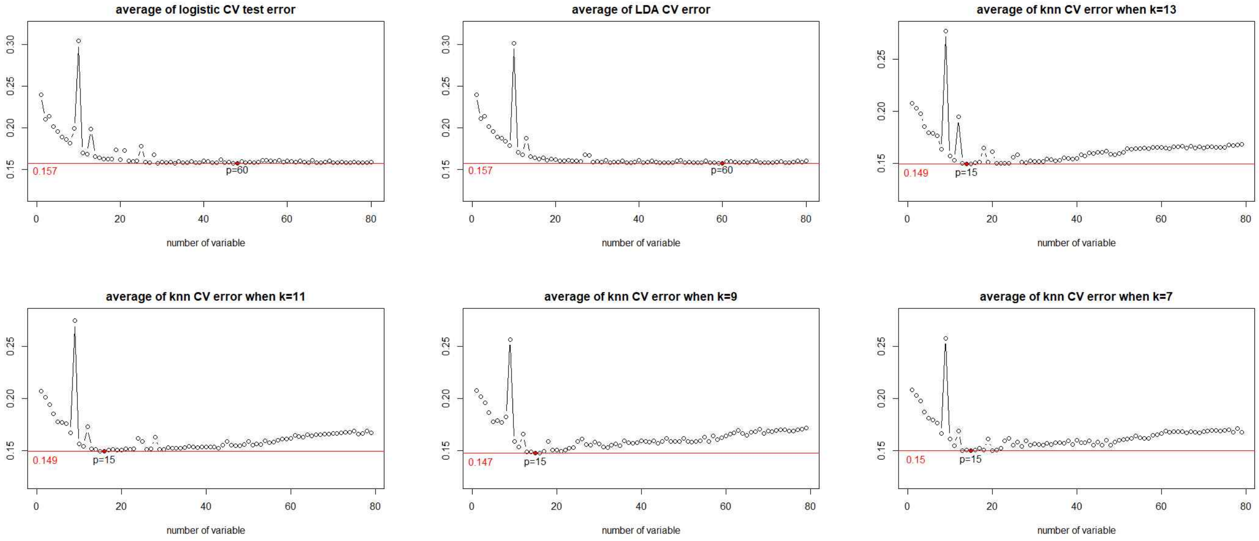
We applied random undersampling in order to deal with the imbalance discovered in the response variable. The cases of death or life-threatening illnesses were 8690, which was relatively smaller than the rest (about 300,000). Therefore, we sampled 8690 observations from which did not result in death or life-threatening illnesses in order to make balanced data, and analyzed with 17380 observations.

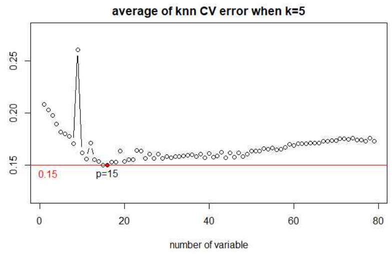
The methods we applied are logistic regression (LR), linear discriminant analysis (LDA), and K-nearest neighbors (KNN) for different K's. (1, 3, 5, 7, 9, 11, 13). Also, we implemented 10-fold cross-validation (CV) and found the best subset of predictor variables that minimizes the CV error, which is the accuracy in this case. We note that the accuracy is identical to the balanced accuracy (BA) since the numbers of the case and the control are equal.

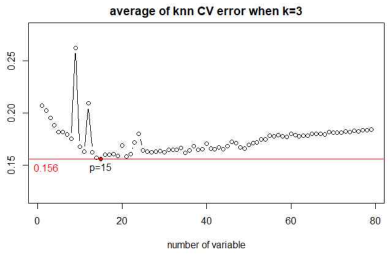
3. Results and Plans

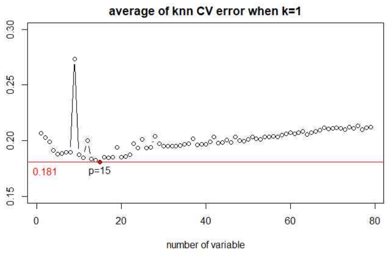
3.1. Results

For LR and LDA, we developed the prediction model and set the threshold value as 0.5. As the number of predictor variables increases, the CV error of LR and LDA models exhibit the decreasing trend. The best predictors are attained when the number of variables p is 60. On the other hand, KNN models see their CV error minimized at $p = 15$. The CV errors of all methods are around 0.15, and the CV error for KNN is minimized at K = 9.









3.2. Further Plans

The undersampling method is, in fact, affected by which observation was chosen as a sample in the larger class (no death or no life-threatening illnesses). So, the prediction model has to increase its accuracy by applying undersampling several times and taking the mean of CV errors. Since different undersampling implies different data, we need to analyze the hierarchy or the linear correlation, if exists, between the predictor variables for the new samples.

Since the symptom variables are posterior variables, we will consider the prediction model which does not have corresponding variables too.

Also, we are going to draw ROC curves and compute AUC for LR and LDA models by changing the threshold value. Using the best number of variables above, we are planning to use random forest method to develop a prediction model that can accurately explain the response variable. Finally, we will find out the best predictor variables and interpret the degree of contribution of each variable to the rate of death or life-threatening illnesses.

[References]

[1] COVID-19 Vaccine Adverse Reactions (VAERS) Dataset, https://www.kaggle.com/landfallmotto/covid19-vaccine-adverse-reactions-vaers-dataset?select=vaers_jan_sep_2021.csv

[2] VAERS Data Use Guide, https://vaers.hhs.gov/docs/VAERSDataUseGuide_November2020.pdf

[3] MedDRA Hierarchy, https://www.meddra.org/how-to-use/basics/hierarchy

[4] 박근우 and 정인경. "이분형 자료의 분류문제에서 불균형을 다루기 위한 표본재추출 방법 비교" 응용통계연구 32, no.3 (2019) : 349-374.