Chapter 5

Generalized Linear Model for Automobile Fatality Rate Prediction in R

Gao Niu
Bryant University, USA

Alan Olinsky
Bryant University, USA

ABSTRACT

This chapter demonstrates the descriptive and statistical modeling function in R. The automobile fatal accident data of the United States is extracted from the Fatality Analysis Reporting System (FARS). The model will be used to understand significant contributing factors of automobile accident death when a fatal crash happens. First, descriptive analysis is performed by basic R functions and packages. Then, generalized linear model (GLM) with logit link function is explored and constructed. Finally, multiple validation metrics are introduced and calculated to ensure the reasonability and accuracy of the predictions. The focus of this chapter is to demonstrate the power and flexibility of the most popular Open Source Statistical Software (OSSS) through a real data analysis.

INTRODUCTION

A fatal automobile accidental crash is defined as any crash that is involved with at least one death. This chapter analyzes all fatal automobile accidental crashes of the year 2015 in the United States. Data is collected by Fatality Analysis Reporting
Generalized Linear Model for Automobile Fatality Rate Prediction in R

System. The data has been used for Big Data and visualization research (Segall & Niu, 2018). This chapter uses the data for Open Source Statistical Software application. All parties’ information involved in a fatal death is collected and analyzed. Analysis starts with descriptive analysis, which includes creating graphs and charts to check the relationship and significance of the variables. Next, in the model construction phase, a step-wise variable selection process is performed. A finalized generalized linear model guarantees that all of its selected variables are statistically significant. Once the model is constructed, holdout data is used to validate the predicting power to ensure the validity of the model.

This chapter intends to demonstrate a typical data analysis process through various R functions and packages. Although the sample data demonstrated in this chapter is not large enough to be defined as Big Data, R’s calculation efficiency does allow Big Data to be processed with its strong programming environment. There are also publicly available sources that enhance its Big Data process capability such as parallel programing. Parallel programming distributes calculation tasks into multiple jobs and execute them separately. R working environment introduction and detailed R code along with code comments are included in the body of the chapter. They ensure that readers gain enough knowledge in the R working environment, syntax, and statistical capability.

BACKGROUND

In 2015, there were 35,092 deaths due to fatal automobile accidents, and 80,588 people were involved in the fatal accidents. Fatality Analysis Reporting System captures key factors behind fatalities in motor vehicle traffic crashes. Factors such as State, Age, Gender, Injury Severity, Seat Position, Drinking Status, Owner Status, Automobile Model Year, Travel Speed, Deformation, Day Week, Route, Location (Latitude and Longitude), Light Condition, and Weather Condition are analyzed and modeled.

This chapter first introduces the R working environment. After that, it follows with several commonly used descriptive analysis functions. Then, generalized linear model construction topics, such as model selection and variable selection, will be discussed. Once the model is created, multiple validation approaches, for instance, calendar year, 75/25, 80/20, and k-fold cross validation, will be discussed. All steps will be introduced and demonstrated with coding and statistical background examples.
What Is Open Source Software (OSS) and What Is Big Data?
www.igi-global.com/chapter/what-is-open-source-software-oss-and-what-is-big-data/248872?camid=4v1a