Accident Analysis and Prevention 115 (2018) 79–88

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Latent class analysis of accident risks in usage-based insurance: Evidence from Beijing

Wen Jin[a], Yinglu Deng[b], Hai Jiang[a,*], Qianyan Xie[c], Wei Shen[d], Weijian Han[c]

[a] Department of Industrial Engineering, Tsinghua University, Beijing 100084, China
[b] PBC School of Finance, Tsinghua University, Beijing 100084, China
[c] Research and Advanced Engineering, Ford Motor Company, 2101 Village Road MD-2149, Dearborn, MI 48121, United States
[d] Asia Pacific Research,Ford Motor Company, Unit 4901, Tower C, Beijing Yintai Center, No. 2 Jianguomenwai Street, Beijing 100022, China

ARTICLE INFO

Keywords:
Usage-based insurance
Vehicle telematics
Accident risk
Latent class model

ABSTRACT

Car insurance is quickly becoming a big data industry, with usage-based insurance (UBI) poised to potentially change the business of insurance. Telematics data, which are transmitted from wireless devices in car, are widely used in UBI to obtain individual-level travel and driving characteristics. While most existing studies have introduced telematics data into car insurance pricing, the telematics-related characteristics are directly obtained from the raw data. In this study, we propose to quantify drivers' familiarity with their driving routes and develop models to quantify drivers' accident risks using the telematics data. In addition, we build a latent class model to study the heterogeneity in travel and driving styles based on the telematics data, which has not been investigated in literature. Our main results include: (1) the improvement to the model fit is statistically significant by adding telematics-related characteristics; (2) drivers' familiarity with their driving trips is critical to identify high risk drivers, and the relationship between drivers' familiarity and accident risks is non-linear; (3) the drivers can be classified into two classes, where the first class is the low risk class with 0.54% of its drivers reporting accidents, and the second class is the high risk class with 20.66% of its drivers reporting accidents; and (4) for the low risk class, drivers with high probability of reporting accidents can be identified by travel-behavior-related characteristics, while for the high risk class, they can be identified by driving-behavior-related characteristics. The driver's familiarity will affect the probability of reporting accidents for both classes.

1. Introduction

Usage-based insurance (UBI), also known as Pay-As-You-Drive (PAYD), is an innovative concept and recently commercialized in the car insurance industry (Karapiperis et al., 2015). While conventional car insurance pricing quantifies drivers' accident risks using their demographic characteristics (such as age, occupation, income and so on) and vehicle characteristics (such as vehicle price, vehicle size and so on), UBI can directly observe the travel and driving behavior of the drivers and make individual-level car insurance possible (Tselentis et al., 2017). The travel and driving behavior characteristics of the drivers are derived from the telematics data, which are transmitted from wireless devices in real time back to an organization (SAS, 2013). Examples of telematics data include the global position system (GPS) data and the in-vehicle sensor data. With more and more vehicles equipped with telematics devices, it is becoming a very promising practice with significant potential impact on accident analysis and many other aspects, such as the mitigation of congestion and the improvement in fuel efficiency. UBI, as one of the applications of the telematics data, is getting increasingly popular among insurance companies across the world.

While existing studies have introduced UBI, drivers' familiarity with their driving routes, that can be obtained from the telematics data, is not studied. Most existing studies use average speed, rates of hard accelerations or hard brakes, and other characteristics that can be directly measured. Consider a daily commuter who drives every weekday between home and work, and another driver who rarely drives. Although the former driver has a higher annual mileage, he/she may be at a lower risk since most of the trips are driven on the same route. The acquaintance of the route shall actually decrease the driver's risk. This is not in line with the concept of UBI, that the driver's accident risk is positively correlated with the annual mileage (Paefgen et al., 2014; Elvik, 2015; Lemaire et al., 2016). In this study, we propose to measure drivers' familiarity with their driving routes using the telematics data and introduce it into car insurance models.

Another limitation of existing studies is that the heterogeneity in

* Corresponding author.
E-mail address: haijiang@tsinghua.edu.cn (H. Jiang).

https://doi.org/10.1016/j.aap.2018.02.023
Received 24 November 2017; Received in revised form 23 January 2018; Accepted 22 February 2018

travel and driving styles based on the telematics data have not been thoroughly investigated. Failure to understand the behavioral heterogeneity may result in failure to identify accident risks at the individual level so that the estimation of accident risks would be biased. For car insurance companies, for example, the estimation of only average accident risk for a group of driver, as in the widely used generalized multivariate regression models in the conventional car insurance pricing, may lead one to miss that a driver with particular travel and driving behavior is at a great risk for accidents. Modelling heterogeneity is a major research problem in marketing (Allenby and Rossi, 1998; Fiebig et al., 2010). Existing studies have found that the heterogeneity in the travel and driving behavior does exist (Paleti et al., 2010; Tu et al., 2010; Eluru et al., 2012; Behnood et al., 2014; Ayuso et al., 2014; Yasmin and Eluru, 2016; Behnood and Mannering, 2016; Yu et al., 2017). In these studies, road characteristics, vehicle characteristics and other environmental factors such as weather conditions are investigated. However, telematics data, that can directly describe drivers' behavior, are not used.

In this study, we use telematics data collected in Beijing, China, to study the effects of drivers' familiarity with the driving routes and the behavioral heterogeneity in their accident risks based on telematics data. Specifically, we build models to estimate their probability of reporting accidents. We first build two binary logistic models to investigate the effects of the telematics-related variables on predicting the probability of reporting accidents. The first binary logistic model uses the conventional car insurance variables, that is, the driver-related and the vehicle-related variables; while the second binary regression model introduces telematics-related variables in addition to those already included in the first model. To capture the travel and driving behavioral heterogeneity among drivers, we develop a latent class model, which first probabilistically classifies the drivers into two latent classes according to the driver-related and vehicle-related variables. Then a class-specific binary logistic model is applied to predict the probability of reporting accidents for each driver.

We summarize the contributions of this research as follows:

- We are the first to measure drivers' familiarity with their driving routes based on the telematics data. We find that drivers' familiarity with their driving trips is critical to identify high risk drivers, and the relationship between drivers' familiarity and accident risks is non-linear.
- We are the first to study the heterogeneity in travel and driving styles based on the telematics data. We find that the drivers can be classified into two classes, where the first class is the low risk class with 0.54% of its drivers reporting accidents, and the second class is the high risk class with 20.66% of its drivers reporting accidents. For the low risk class, drivers with high probability of reporting accidents can be identified by travel-behavior-related characteristics, while for the high risk class, they can be identified by driving-behavior-related characteristics.

The remainder of this paper is organized as follows. Section 2 reviews existing literature on conventional car insurance and UBI. Section 3 describes the binary logistic regression model and the latent class model used in this research. Section 4 describes the data, the dependent and independent variables. Section 5 presents the estimation of the parameters. Finally, Section 6 concludes the paper.

2. Literature review

The related literature is reviewed in this section. We first introduce the variables in conventional car insurance pricing and UBI in Section 2.1. We then summarize recent researches on UBI in Section 2.2.

2.1. Car insurance variables

Conventional car insurance pricing incorporates a broad range of categorial variables in coordination with various non-linear regression models. These variables distinguish between driver-related and vehicle-related characteristics (Weidner et al., 2016; Tselentis et al., 2017). The driver-related variables include, for example, the gender, age, occupation, and regional classification of the driver. The vehicle-related variables comprise the price, age, and type of vehicle and other vehicle characteristics. Annual mileage performance is also included in the conventional car insurance pricing. It can be classified as a driver-related variable (Weidner et al., 2016), or defined as an exposure variable (Paefgen et al., 2014). The annual mileage used in conventional car insurance is usually reported by the drivers. Studies have shown that the reported values are usually lower than the actual values, resulting in inaccurate premium calculation (White, 1976). In addition to these driver-related and vehicle-related variables, there are discount factors in the calculation of the final premium, such as no claims discount (No-Claim Discount, NCD), which measures the historical accident reports of the drivers. If a driver did not report accidents in his/her previous insurance periods, the value of NCD will be low. The final premium of the driver is obtained by multiplying the value of NCD with the premium calculated from the pricing models.

The variables in UBI pricing are generated from the telematics data, including the actual annual mileage, rates of hard accelerations or hard brakes, speed-related variables, and exposure variables. The annual mileage is the core in the concept of UBI that the premium is positively correlated with the annual mileage (Elvik, 2015; Lemaire et al., 2016; Litman, 2005; Paefgen et al., 2014). The rates of hard accelerations or hard brakes are defined as the number of hard accelerations or hard brakes per unit mile or time (Handel et al., 2014). Handel et al. (2014) find that the rate of hard brakes is a very significant variable in insurance pricing. The speed-related variables are derived from the driving speed with the driving time and location information. For example, Bagdadi and Várhelyi (2011) define critical jerks as the rate of change of acceleration with a threshold level of 9.9 m/s^3. The exposure variables describe the characteristics of travel time and geographical distribution (Paefgen et al., 2013, 2014). Paefgen et al. (2014) use GPS data to generate exposure variables that represent the fraction of mileage exposure under specified driving conditions. In addition to these variables, Tselentis et al. (2017) distinguish between travel and driving behaviour variables. They refer to travel behaviour of the driver as her/his strategic choices concerning, for example, the type of road network and the time, and driving behaviour of the driver as her/his operational choices at real time in handling the vehicle within the existing traffic conditions, including hard brakes and travel speed, and so on.

2.2. Recent researches on UBI

Existing studies provide different insights into telematics-based accident risk studies and UBI solutions. In the field of transportation, researchers use various regression models and statistic tests to quantify the possibility of reporting accidents for a particular driver. In the field of insurance, researchers focus on introducing the telematics data into conventional insurance problems to supplement the conventional car insurance research.

In the field of transportation, Bagdadi and Várhelyi (2011) study the relationship between the rate of hard brakes and the self-reported accident involvement of the drivers. They use driving data from 166 passenger cars to derive the rate of hard brakes. A count regression model is estimated where the dependent variable is the number of reported accidents, and the independent variables are the gender and the rate of hard brakes of the drivers. They find that the number of accidents increases as the rate of hard brakes increases. Paefgen et al. (2013, 2014) use location data from 1600 vehicles to study the effect of