8th National Congress on Civil Engineering, 7-8 May 2014
Babol Noshirvani University of Technology, Babol, Iran

Cluster-Based Modeling of Crash Frequency

Pooya Najaf[1], Venkata R. Duddu[2], Srinivas S. Pulugurtha[3]

1- Research and Teaching Assistant, INES Ph.D. Candidate, The University of North Carolina at Charlotte, NC, USA

2- Assistant Research Professor, Department of Civil & Environmental Engineering, The University of North Carolina at Charlotte, NC, USA

3- Associate Professor and Graduate Program Director, Department of Civil & Environmental Engineering, The University of North Carolina at Charlotte, NC, USA

1- pnajaf@uncc.edu
2- vduddu@uncc.edu
3- sspulugurtha@uncc.edu

Abstract

Several recent studies have tried to use new techniques to increase the accuracy of crash frequency models. The objective of this manuscript is to evaluate interpretability and predictive ability of Cluster-based Negative Binomial Regression (CNBR) in comparison with basic conventional Negative Binomial Regression (NBR) model. First, the crash data is clustered into different homogenous categories using Two-Step Cluster Analysis (TSCA) and then NBR is developed separately for each category. The results from comparison of the modeling procedures indicate that CNBR has higher fitting ability, more predictive accuracy, and better interpretability. In addition, TSCA generates homogeneous categories which facilitate the interpretation of effective factors across each category. It can be helpful for operators to consider significant factors in each category separately. However, the combination of TSCA and NBR makes it a time consuming procedure. On the other hand, NBR model for the entire database is quick and easy to develop, but has a lower predictive ability.

Keywords: Crash Frequency, Two-step Cluster Analysis, Cluster-Based Negative Binomial Regression

1. INTRODUCTION

Researchers try to have better understanding about the role of various significant factors (including roadway characteristics, traffic flow conditions, environmental characteristics, and users' and vehicles' features) on traffic safety. In addition, it is vital for practitioners and operators to predict crash frequencies more accurately. Therefore, interpretability and predictive ability of crash frequency models have gained a lot of attention from researchers and practitioners over the last two decades. In this research, a Two-Step Cluster Analysis (TSCA) method is used to segregate the crash data into different clusters and then develop Negative Binomial Regression (NBR) model separately for each cluster. Also, a general NBR model is developed for the entire database as well. The results of these two modeling procedures (i.e., NBR for the entire dataset and CNBR) are compared to assess their fitting and predictive accuracy as well as the ability of interpretation. Data for 1,354 segments in the city of Charlotte, North Carolina is used to develop the models, while data for 396 segments is used to validate the models.

2. LITERATURE REVIEW

Clustering methods try to divide the dataset into different homogenous groups to determine possible patterns [1, 2]. Sohn and Lee [3] suggested clustering the crash dataset into different homogenous clusters before developing a model. In fact, they suggested a cluster-based modeling procedure, especially when the variables are dealing with large variations, rather than an overall homogenous assumption for the safety model. "The clustering