Big Data Methodologies for Simplifying Traffic Safety Analyses (01-001) Dataset

Dataset available at: https://doi.org/10.15787/VTT1/QQEZOP

(This dataset supports report **Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data,** http://hdl.handle.net/10919/95171**)**

This U.S. Department of Transportation-funded dataset is preserved by the Virginia Tech Transportation Institute (VTTI) in their data repository (https://dataverse.vtti.vt.edu/), and is available at https://doi.org/10.15787/VTT1/QQEZOP

The related final report **Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data**, is available from the National Transportation Library's Digital Repository at https://rosap.ntl.bts.gov/view/dot/61488

Metadata from the VTTI Repository record:

Dataset Persistent ID: doi:10.15787/VTT1/QQEZOP
Publication Date: 2019-07-18
Title: Big Data Methodologies for Simplifying Traffic Safety Analyses (01-001)
Author:

- Lord, Dominique (TTI) - ORCID: 0000-0002-7434-6886
- Guo, Feng (VTTI) - ORCID: 0000-0002-3161-3235
- Deng, Xinwei - ORCID: 0000-0002-1560-2405
- Mao, Huiying - ORCID: 0000-0003-4721-2767

Description:

- Decision-adjusted Modeling
    - Product Description:
        - Predicting crash risk and identifying high-risk drivers are critical for developing appropriate safety countermeasures, driver education programs, and user-based insurance. However, predicting driver risk is a challenging task because crashes are rare events and many factors contribute to individual crash risk. As in-vehicle data collection becomes more prevalent and cost-effective, it has become more feasible to improve risk prediction by utilizing kinematics information. Currently, there are several challenges to implementing kinematics-based driver risk prediction models. We focus on two primary issues: (1) the decision rule and (2) the optimal threshold values for kinematics predictors.
    - Data Scope:
        - The naturalistic driving data collected from the second Strategic Highway Research Program (SHRP 2) is used to identify optimal prediction models for high-risk drivers by kinematic signatures. The dataset includes 3440 rows of drivers and 200 columns of features. Each row represents one driver, and the columns represent the characteristics of each driver.

    - Data Specification:
        - The specific data description is shown in Table 1.
- Adjust finite-sample bias for traffic safety modeling
    - Product Description:
        - The Poisson and NB models are generally estimated using the maximum likelihood method. When the sample size is small and/or when the number of events is limited (e.g., small number of crashes), the maximum likelihood estimators (MLEs) are biased and the bias could be substantial. This finite sample bias could lead to incorrect estimation of the impacts of risk factors and jeopardize traffic safety improvement efforts. This project addresses this gap by studying the finite sample bias for the parameter estimation of Poisson and NB regression models in the context of traffic safety modeling.
    - Data Scope:
        - To illustrate the benefit of bias correction and examine the magnitude of bias, we applied the bias-correction procedure to an infrastructure safety evaluation dataset. This dataset includes information from 5,238 short road segments, which are collected from 2012 to 2014 in the State of Washington. The length for each segment is 0.1 mile. The covariates used in the analysis include route type, whether the road segment is an entrance/exit, whether it is an intersection, whether it is a ramp, whether it is a wye connection, whether it is a divided highway, rural/urban, number of lanes, pavement type, friction, gradient, and horizontal curvature.
    - Data Specification:
        - The specific data description are listed below.
            - In order to comply with participant informed consent and IRB requirements, a portion of this data set is governed by a Data Use License (DUL). Additionally, the other portion of the data will not be available for re-use. Please send inquiries to datasharing@vtti.vt.edu.

Subject: Engineering; Other
Keyword: driver risk, decision-adjusted modeling, predictive modeling, telematics data, finite sample bias, maximum likelihood estimate, Poisson regression, negative binomial regression, traffic safety
Deposit Date: 2019-07-18

Recommended citation:
Lord, Dominique; Guo, Feng; Deng, Xinwei; Mao, Huiying, 2019, "Big Data Methodologies for Simplifying Traffic Safety Analyses (01-001)", https://doi.org/10.15787/VTT1/QQEZOP, VTTI, V1

Dataset description:
This dataset contains 1 file collection, described below.

Analyzing Highway Safety Datasets Simplifying Statistical Analyses from Sparse to Big Data_Data.zip

- 01-001_Decision-adjusted Modeling.pdf
- 01-001_Adjust finite-sample bias for traffic safety modeling.pdf

The .pdf file format is an Adobe Acrobat Portable Document Format (PDF) file and can be opened with the Adobe Acrobat software.

National Transportation Library (NTL) Curation Note:
As this dataset is preserved in a repository outside U.S. DOT control, as allowed by the U.S. DOT's Public Access Plan (https://doi.org/10.21949/1503647) Section 7.4.2 Data, the NTL staff has performed ***NO*** additional curation actions on this dataset. NTL staff last accessed this dataset at https://doi.org/10.15787/VTT1/QQEZOP on 2022-04-27. If, in the future, you have trouble accessing this dataset at the host repository, please email NTLDataCurator@dot.gov describing your problem. NTL staff will do its best to assist you at that time.