基于车载记录仪数据的车辆风险模型实证①

胡　伊[1], 王天梅[1], 崔　鹏[2]

[1](中央财经大学信息学院, 北京 100081)
[2](中央财经大学科研处, 北京 100081)

摘　要: 随着计算机技术的发展,越来越多的车载记录仪(in-vehicle data recorders,简称 IVDR)因为商业目的被安装到了车辆上.在获取更多实时车辆行驶数据的前提下, 也为车辆事故原因的分析提供了更加丰富的数据基础.本文通过对车载记录仪的数据进行分析并结合不同车辆的驾驶条件与环境因素开展车辆风险因素研究具有一定的理论价值与实践创新. 研究运用大量的行驶数据、地理位置等数据信息,建立多元线性回归模型对车辆风险与事故关系进行研究并对所得结果进行了实证检验. 过研究发现车辆风险因素与事故之间的影响关系:行驶里程与发生事故的概率存在非线性关系等. 望通过本研究能对车辆风险量化、交通出行规划、车险个性化定价提供一定的借鉴作用.

关键词: GPS 轨迹, 载记录仪, 辆风险, 元线性回归

Empirical Study of Vehicle Risk Model Based on IVDR Data

HU Yi[1], WANG Tian-Mei[1], CUI Peng[2]

[1](School of Information, Central University of Finance and Economics, Beijing 100081, China)
[2](Scientific Research Office, Central University of Finance and Economics, Beijing 100081, China)

Abstract: As the development of computer science, the commercial applications of In-Vehicle Data Recorders (IVDR) are growing rapidly. With the real-time vehicle travelling data, the implications of IVDR facilitate the rich data base of traffic accident causal analysis. Combining IVDR data analysis with driving conditions and environmental factors of different vehicles, this paper carries out the study on vehicle risk factors. On the basis of a large amount of driving and geographic location information, this study builds multiple linear regression models to analyze the relationship between vehicle risk and traffic accident, and applies empirical test to the research results. This research indicates the interact relationship between vehicle risk factors and traffic accidents, for instance, that the nonlinear relationship between travelling distance and accident probability and etc. It is expected that this study would provide useful preference for vehicle risk quantification, transport planning, and vehicle personalized pricing.

Key words: GPS trajectory; in-vehicle data recorder; vehicle risk; multiple linear regressions

1　引言

目前基于物联网技术的车载记录仪被越来越多的应用到车辆安全驾驶分析上, 为开展车辆风险因素识别与事故调查提供了丰富的数据支持. 早在 1976 年便出现了外形笨重的驾驶数据记录仪器(磁带机), 但是受制于仪器设备与分析工具的落后, 严重制约了对驾驶数据的研究与应用, 直到近几年才取得了较大突破. 随着现代信息技术的快速发展, 已经能够生产出性能更好、外形更小巧的车载数据记录仪应用到各类车辆上, 同时为开展科学研究提供海量的车辆驾驶相关数据[1]. IVDR 提供的数据比传统的人工记录数据具有更高的有效性和实用性, 此外还能够模拟出一部分人的驾驶行为, 从而提高了通过总体样本来分析研究车辆风险因素的准确率[2]. 为了得到更加趋近于自然

① 基金项目:教育部人文社科重点研究基地重大项目(14JJD790013);中央财经大学博士研究生创新基金重点项目(201418)
收稿时间:2015-09-17;收到修改稿时间:2015-11-02

驾驶人的行为方式, 研究人员还在 IVDR 设备上增加了卫星定位, 平均速度, 视频采集等的新功能[3].

通过丰富的数据能够获得更多的驾驶行为细节, 但是在多源数据获取、分析、处理上却进入了一个困惑的领域, 一些科研部门为了满足研究需要, 已经开始设计自己的数据采集系统(DAS)[4]. 在我国商业车险改革并全面放开的市场背景下国内几家保险公司已经率先开始尝试使用 IVDR 数据来分析驾驶行为并个性化定价车险产品(按驾驶行为付费 PSYD). 这些保险产品将按照驾驶人的驾驶行为与车辆的实际使用情况来收取保费, 这有别于传统车险事前一次性支付的方式, 能够更加精确核算保费并且有效的减少因为信息不对称所带来的道德风险与逆向选择问题, 从驾驶人的主观行为上鼓励保险客户为更低的保费而自身降低车辆风险[2,5]. 国外学者 Desyllas 与 Sako 在 2012 年通过大量传统文本记录的保险数据分析研究显示出行驶里程与事故发生之间存在正向的关联关系[6]. 随着研究数据的日渐丰富, 本研究从车载记录仪的数据分类聚合出发, 分析比较了不同类型的车辆风险因素, 通过多元回归模型对车辆风险因素与事故进行相关性研究, 最后通过对比实际出险情况加以验证, 为进一步开展相关研究提出意见和建议.

2 相关研究

早期受制于数据技术处理的瓶颈以及数据采集的复杂性, 相关行业协会的关注和参与程度不高, 对 IVDR 数据的应用研究虽然已有几十年, 但一直没有突破性的进展, 直到近几年物联网技术在我国的快速发展与应用才取得了较大突破. 通过梳理国内此类研究文献, 发现基于 IVDR 研究的数据样本直到最近几年才有了比较代表性的实际使用案例. 鉴于当前相关理论研究主要来源于英文文献, 本研究在变量选取和指标确定上主要参考了英文研究文献. 其中, Jolliffe 等人在 2007 年通过一个基本的线性回归模型, 回归分析年度总里程和发生保险损失车辆之间的函数关系, 得到的拟合优度指标较好(R^2= 0. 82). 该研究中除了使用年度总里程, 并没有研究其他变量, 或者对样本进行详细分类与描述[7]. 虽然这项研究没有提供太多新颖的见解, 但对车辆保险产品的创新与个性化却带来了新的契机.

2.1 行驶里程与速度对车辆风险的影响

自然驾驶又称为最大程度上的趋近与真实驾驶数据值, 表示通过基于行驶里程数据收集研究日常车辆习惯的驾驶风格和行为方式. 这一领域出版的研究成果是美国国家公路交通安全管理局与弗吉尼亚理工大学交通研究所通过对 100 辆车进行 2001 年至 2006 年的 5 年跟踪行驶里程数据分析研究. 在实验中, 研究人员通过利用运动学的一些指标来识别主要风险因素, 通过对相关行车数据、视频的筛选、清洗, 为研究提供了前所未有的详细测量[4]. Helander 等人在 2010 年针对车辆行驶里程和车辆排放等外部性风险变量进行了描述, 得出车辆行驶里程和自身性能对于事故风险的正向显著影响. 主要分析了速度因素对车辆风险提高正相关, 得出驾驶环境对车辆风险的影响显著; 同时也指出驾驶速度是交通事故风险重要的预测、判定指标[8].

2.2 驾驶环境对车辆风险的影响

通过对车载记录仪中 GPS 日志信息分析可以获得车辆行为偏向, 包括驾驶者的兴趣地点、驾驶习惯、实时或历史的监控信息等. 根据行驶地理位置、车主个人信息等隐私数据的暴露分级, 结合车载传感装置(包括光照传感器、天气传感器、速度传感器等)、GPS 等设备(地理位置信息等)监测车辆的仪表盘信息、驾驶者的驾驶模式(包括驾驶时段、紧急刹车、突然加速等)等空间信息, 可以有效的分析各种道路情况可能存在的风险因素. Jun 等人 2010 年对亚特兰大公交系统中 IVDR 数据部分功能进行分析, 通过个案对比研究的方法比较了 167 辆安装有 IVDR 设备的汽车, 在固定的观察期内肇事车辆与正常车辆之间数据的异同. 研究发现了在两组样本在不同道路类型下(高速公路、干道或本地道路)、运行时间内, 车辆速度与风险的影响关系[9]. 我国学者吴义与宁洪在 2014 年着重指出了驾驶环境、驾驶行为是车辆风险的重要因素, 并提出了 PSYS(按驾驶速度支付保险)将是车险产品的创新[10].

综上所述, 车辆风险因素的分析已逐渐从事前因素到事后因素、从静态风险因素到动态风险因素的过渡阶段. 国内外学者从不同角度和风险因素种类进行了相关研究, 本文尝试通过大量的车载数据、地理位置等信息, 加入更多与车辆风险相关的因素来建立多元回归模型加以分析研究并对所得结果进行实证验证. 本研究在方法和风险因素种类上有所创新, 期望通过本研究能对车辆风险进行量化分析处理, 并对未来的车辆风险识别与车险个性化定价提供一定的借鉴作用.

3　数据收集与整理

本研究采用某保险公司 PSYD 类型车辆保险产品测试 IVDR 数据, 在使用过程中对涉及个人隐私的内容进行保护隐去. 当车辆在运行期间, 地理位置只要有变动, IVDR 设备将自动更新记录一次数据内容, 当地理位置没有发生变动或者发动机处于关闭状态将停止更新数据. 通过计算不同地理位置的变动距离与更新的时间, 可以获得驾驶过程的平均速度及行驶总里程, 再配合 GPS 电子地图的数据还可以获得行驶区域和道路类型的数据. 通过随机抽样程序从 IVDR 数据库提取研究数据, 抽样将通过个案对照研究设计开展. 随机从2012年发生交通事故的600辆肇事车作为样本, 收集他们近 6 个月的 IVDR 数据记录, 事故类别分为: 普通事件, 人身伤害事件, 致死事件三个类别. 按照 1/12 的分层抽样, 使得每 50 辆车的平均事故率更趋向于整体样本均值. 在研究中按每个月的平均值来抽取, 避免季节性变化带来的影响. 没有地理信息的风险事故数据将被排除在样本之外, 后续可以作为驾驶行为对风险事故的影响研究. 在参照组中, 随机选出了在 2012 年 7 月至 2012 年 12 月间地理位置数据完整的 1000 辆无肇事车数据库信息作为另一个样本组. 在处理数据时将排除以下情况的数据:

(1) 在研究期间车辆卷入其他事故中.

(2) 在数据记录过程中出错或者在存储时, 无法生成日志的车辆数据

(3) GPS 传感器长时间没响应导致没有地理位置信息及数据更新的车辆数据

(4) 不连续的 GPS 数据, 导致的行驶过程不明确车辆数据.

最终这些研究车辆数量发生了如下变动, 肇事车辆样本组从 600 减少了 17 辆, 变为 583 辆(减少 2.9%), 无肇事车辆样本组从 1000 减少 16 辆变为 984(减少 1.6%). 因为大样本的原因, 没有进一步消除异常值, 可以认为对于回归的影响结果可以忽略不计.

合并后的数据集包含了 2679425 个数据点, 行驶里程, 平均每车有 958.5 个数据点. 为了便于后续的统计分析, 本文把不同片段的数据内容整合到单辆车的特征表现级别内. 定义一个从 N 到 M 的特征矩阵 E, n 行代表 1567 辆车当中的第 n 辆车数据, m 列代表不同行驶里程下的特征值. 在矩阵 E 当中, 聚合算法将会自动识别驾驶里程增加的情况下其他数据对应的变化值, 下图 1 是对数据聚合过程的概述.

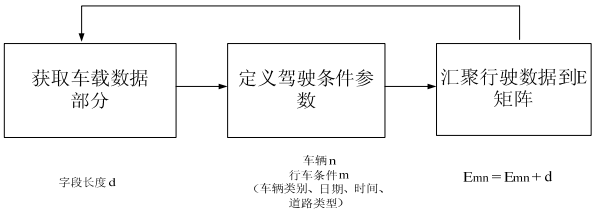

图 1　数据聚合过程

4　模型建立与修正

为了系统的分析实验组的数据与真实事件之间的误差情况, 本研究将使用多元逻辑回归模型. 在线性逻辑回归模型 $g=\beta_0+\sum i\beta_i x_i$ 中 x_i 为自变量函数, 预测事件发生的概率性.

$$\hat{P}(case|x_i)=1-\hat{P}(control|x_i)=\frac{e^g}{1+e^g} \quad (1)$$

使用最大似然法, β_0 为截距系数, β_i 系数为预测概率与真实观察值之间的误差(0 至 1 之间). 考虑到样本量包含 1567 辆, 在进行模型估计时的影响因素可能将达到 43 个.

4.1 融合区间变量设定

本文主要考虑以下数据信息来进行设定. 选择合理的约束条件, 对于确定矩阵 E.

① 一天当中的什么时候

② 一周当中的哪一天

③ 主要在什么类型的道路行驶

④ 平均行驶速度

⑤ 模型对影响因素总量拟合的效果如何等

把不同驾驶情况下的数据分类聚合, 需要考察一段连续的离散时间、速度变量. 处理时间变量方面, 以每小时作为一个间隔; 日期将按照每天作为一个间隔; 道路类型分城市、市郊、高速公路共 3 类; 速度每提高 30km/h 作为一个间隔, 最高 120km/h 及以上间隔是开发的. 如果上述条件同时考虑将会在矩阵 E 中产生多达 2520 种的组合条件, 而有的组合变量并不适合建模的研究目的, 所以做了适当的调整, 矩阵中列的数量也减少到了 39 项, 如表 1 所示.

表 1　矩阵 E 的聚合结构分析表

Time of day	Day of week	Road type	Velocity interval	Σ
24 类	7 类	3 类	5 类	39 类

对于一个初始模型, 本文考虑整个变量集包含在风险矩阵 E 中. 预计模型中可能将会涉及到 30-40 个变量, 它们包括:每天的时间间隔、每周的日期, 月平均里程, 道路类型(城市、市郊、高速公路)和速度间隔 0-30 km/h, 60-90 km/h 等等. 然而, 模型表现出较高的误差项(例如时间变量>2800), 这可能是较高的共线性所造成的. 本文选择合并一些相邻的特征属性来减少这个问题带来的误差.

4.2 合并区间变量

作为一个初步的间隔指标合并, 本文将在时间间隔与日期变量组间进行探索性因素分析, 为了获得目标估计数量合并的间隔, 将采用主成分分析的方法[7]. 通过 principal component analysis(PCA)主成分分析法得出的结果(如图 2、图 3 所示), 本文调整合并时间间隔为: (0:00-5:00, 5:00-18:00, 18:00-21:00, 21:00-24:00)共 4 类, 合并的日期间隔为: (周一至周四, 周五至周日)共 2 类, 道路类型和速度类型维持不变.

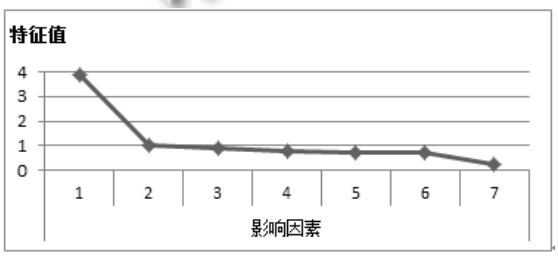

图 2 日期主成分因素分析

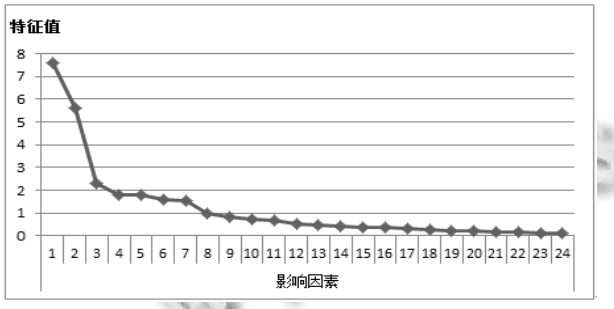

图 3 时间主成分因素分析

5 数据分析

通过对以上处理数据的回归分析, 得到如下表 2 中的统计结果.

表 2 回归统计结果(变量、系数值、标准误差、统计意义 P 值<0. 001)

Variable	Coefficient	Standard error	Significance
Time of day (LN-transformed)			
05-18 h	-.398	123	.001
18-21 h	.697	.169	<.001
Day of week (LN-transformed)			
Friday through Sunday	-.267	.134	.047
Road type (LN-transformed)			
Urban	.738	.164	<.001
Highway	-.288	.143	.043
Velocity interval (LN-transformed)			
0-30 km/h	.381	.153	.013
60-90 km/h	-.734	.138	<.001
90-120 km/h	.305	.178	.085
Average monthly mileage (categorical)			
< 677 km	0 (reference)	-	-
677-1020 km	.360	.575	.530
1021-1274 km	1.024	.545	.062
1275-1584 km	1.395	.528	. 007
1585-1999 km	2.624	.507	<.001
2000-2519 km	4.076	.515	<.001
2520-3187 km	4.671	.524	<.001
3188-3964 km	6.538	.578	<.001
3965-5700 km	7.192	.638	<.001
> 5700 km	7.033	.722	<.001
Constant	-4.418	.476	<.001

① 车辆事故风险系数在 5:00 至 18:00 之间相对较低, 而在 18:00 至 21:00 则存在相对较高的车辆事故风险系数.

② 周末由于出行时间和目的地点的分散性, 事故风险系数有所降低.

③ 城市内驾驶相对比郊区风险系数更大, 如果按里程数和事故率来比较, 高速公路则相对较低.

④ 从平均时速来看, 60-90 km/h 风险系数相对较低, 0-30 km/h 其次, 90-120 km/h 风险系数最高.

直觉意识是随着速度越快意味着车辆风险系数将会更高, 但实际情况是对于固定行驶里程的情况下, 速度与驾驶时间是成反比的, 即在固定行驶里程的情况下车辆速度风险系数提高的同时驾驶时间内遇到的风险系数却可能减少. 在实际生活中, 随着车辆平均速度的下降, 意味着到达目的地将会需要更长的驾驶时间, 车辆暴露在行驶环境中的风险几率却在增加[8].

6 实际出险情况验证

通过对比某保险公司 2012 年机动车出险情况统计数据验证分析结果(图 4).

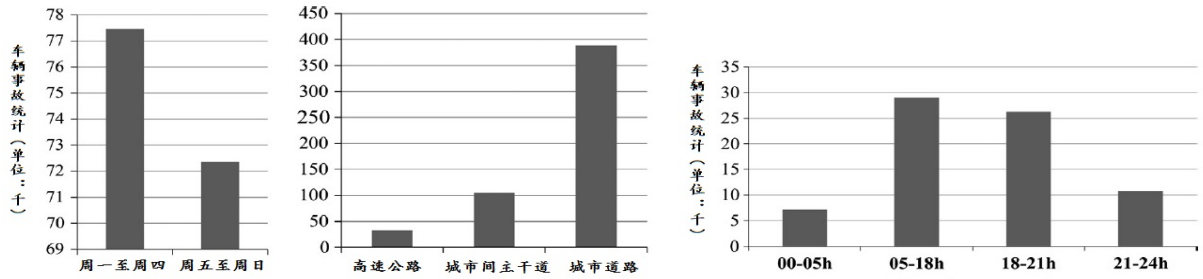

图 4 2013 年某保险公司机动车辆出险情况统计

本文根据 IVDR 数据分析的车辆出险情况, 基本符合实际出险统计. 另外, 每月平均里程数, 类别指标变量显示出非线性关系, 函数单调递增行驶里程达到4000公里后略有下降. 当每月平均里程数低于1600公里时, 本文研究的线性函数高估了风险系数; 当高于1600公里这个临界值后又低估了风险系数. 虚变量数如图 5 所示.

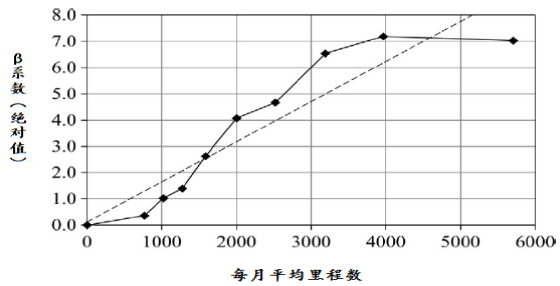

图 5 β系数与每月平均行驶里程关系图

7 研究结论与讨论

本研究通过建立多元回归的模型来处理 IVDR 数据, 分析了车辆事故与风险因素之间的关系. 从 1567 辆行驶地理位置变化的车辆 IDVR 数据中, 采用主成分分析的方法优化减少影响因素的多重共线性问题, 并通过个体比较研究区分了肇事车辆与无事故车辆数据的差异性[9]. 讨论了结合行驶里程测量数据的几组变量程度, 比如:时间点, 日期, 路况类型, 速度等的风险情况. 从多元回归模型得到的结果, 能够看到各风险因素与环境变量对风险事故存在的显著影响. 另外, 当车辆行驶的其他条件固定不变的时, 行驶里程与风险事故之间存在偏离线性函数的关系存在[10]. 本文最后拓展讨论了技术应用这部分的潜在应用可能性, 这些技术将会促进商业行为的 IVDR 数据收集与分析研究.

参考文献

1 Huang H. META: A mobility model of MEtropolitan Taxis extracted from GPS traces. Proc. of the China IEEE. 2010.

2 Bolderdijk JW. Effects of pay-as-you-drive vehicle insurance on young drivers'speed choice: Results of a Dutchfield experiment. Accid. Anal. Prev. 2011, 43 (3): 1181–1186.

3 Jovanis PP, Aguero-Valverde J, Wu KF, Shankar V. Analysis of naturalistic driving event data. Trans. Res. Rec, 2011, 2236: 49–57.

4 华荣晖.美国强制车险费率制度的特点与启示——以马萨诸赛州为例.上海金融学院学报,2009,(1).

5 黄永波.车险费率市场化改革及中小保险公司应对策略.金融会计,2011,(2).

6 Desyllas P, Sako M. Profiting from business model innovation: Evidence from pay-as-you-drive auto insurance. Res. Policy, 2012, 42(1): 101–116.

7 Jolliffe I. Principal component analysis. Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd., Hoboken, NJ, 2007, 13(4): 311−315.

8 Helander M, Hagvall B. An instrumented vehicle for studies of driver behavior. Accid. Anal. Prev, 2010, (8): 271–277.

9 Jun J, Guensler R, Ogle J. Differences in observed speed patterns between crash-involved and crash-not-involved drivers: Application of in-vehicle. IEEE. 2010.

10 吴义,宁洪.车载导航大数据在车险行业的应用.计算机光盘软件与应用,2014(7):23−24,27.