A COHORT-BASED DATA STRUCTURE DESIGN FOR ANALYZING CRASH RISK USING NATURALISTIC DRIVING DATA

Paul P. Jovanis[1] & Kun-Feng (Ken) Wu[2]
[1]Pennsylvania State University, University Park, Pennsylvania, USA
[2]Federal Highway Administration, McLean, Virginia, USA
Email: ppj2@engr.psu.edu

Summary: Although naturalistic driving studies (NDS) have become more prevalent in recent years, many challenges remain in analyzing the data. One challenge is inclusion of exposure in modeling crash risk. While this is a potential strength of NDS, comparatively few studies have emphasized exposure-based analyses. A second challenge is the formulation of analysis methods that include driver attributes, event attributes, and driving environment in a structured formulation. A third challenge is the formulation of baseline hazard to frequently accompany the identification of NDS "events" (e.g. crashes, near crashes and/or safety critical events). This paper reports on a cohort-based data structure design to address these three challenges. Collision warning alert frequency data from University of Michigan Transportation Institute (UMTRI)'s Roadway Departure and Curve Warning System (RDCW) Field Operation Test (FOT) are used to demonstrate this approach. The paper concludes with a discussion of applications which include crash and other NDS-observed events, including potential applications to road safety management through the development of enhanced safety performance functions.

INTRODUCTION

Naturalistic driving studies (NDS) have been shown to have great potential to provide more insight into traffic safety analyses (e.g. Dingus et al., 2005; Shankar et al., 2008; McGehee et al., 2010; Wu and Jovanis 2012a; 2012b), but many challenges are to be overcome to exploit the utility of such studies. The challenges discussed in this paper include: the inclusion of exposure in modeling crash risk; the formulation of a flexible analysis structure to allow inclusion of many types of predictor variables; and, the formulation of events other than crashes that can be used for comparison purposes and identification of baseline risk.

Comparatively few NDS studies have emphasized exposure-based analyses; exceptions include Shankar et al., 2008 and Jovanis et al., 2012. It is much more common for NDS analyses to include detailed study of crash or near-crash events without explicit regard to exposure (e.g. Dingus et al., 2005). Without considering driving exposure, one would expect the safety-related events occur more frequently with tasks and activities that drivers perform more frequently (Hanowski et al, 2005). One goal of this paper is to explore the use of an NDS data formulation that facilitates the inclusion of exposure in the data analyses.

A second challenge is the formulation of analysis methods that include driver attributes, event attributes, and driving environment in a structured formulation; Jovanis et al. (2011) showed that modeling of NDS event data should include all three of these variable types to reduce the

likelihood of bias. The research indicated a substantial omitted-variable bias for estimation of the effect of context variables but little difference for driver variables.

A third challenge is the formulation of baseline hazard to be used in comparison with NDS events. Inclusion of baseline events would allow an estimation of the odds of a crash as compared with baseline non-crashes (Dingus et al., 2005; Jovanis et al., 2011). But non-safety-related events are costly to obtain and difficult to define, particularly if an array of typical predictors is to be included in a data set. Although in current practice these data are obtained at random from a large driving file, it remains an open question as to how the observations will be selected and how many will be needed.

This study proposes a cohort-based data structure design to address these three challenges.

METHODOLOGY

A cohort study enrolls subjects into a particular cohort (study group) based upon their current risk factor status; the prior outcome status (e.g. crash or non-crash) is then tabulated for each individual cohort. Our proposed cohort-based data structure begins with a driver as a unit of analysis. The driver is followed over multiple trips throughout the course of the study (Shankar et al., 2008; Jovanis et al., 2012). Each driver is associated with specific attributes that are constant such as age, gender, driver attitudinal measures, and vehicle type/characteristics. Other variables can change throughout the course of the study and within each trip (e.g., roadway type, roadway characteristics, environmental factors, driver distraction, driver impairment, and driving speed). A subset of these variables can be used to define a cohort – a trip segment that is homogeneous with respect to the variables of interest. Note: travel time and/or distance may thus be accumulated during the study for individual drivers in each defined cohort (i.e., homogeneous trip segment). Travel undertaken in each homogeneous trip segment would then be aggregated to determine total exposure and total number of events within a cohort. A cohort thus represents a set of drivers, by type, who experience travel over defined homogeneous trip segments characterized by time or distance of travel. The number of events of interest (e.g., crashes or other events) occurring for a cohort is thus accumulated across drivers, retaining the number of events and/or the time between events for each driver.

This concept is illustrated in Table 1 and Table 2. Table 1 contains the initial cohort-structured data in which a particular outcome (i.e., an event or non-event) occurs after some period of time or length of travel. The context and driver attributes are selected by the researcher depending on the issues to be explored. Table 2 shows how the individual outcomes can be grouped, if needed, for each cohort. Each unique combination of driver and context variables is now listed with the cumulative time or distance – a measure of *exposure to risk*. Notice that each cohort includes the sum of individual trip segments and their outcomes. Each driver's outcomes are aggregated and matched to context. The sum of the "1" values in the "Outcome" column in Table 1 are the number of events of interest for that cohort. The length and time variables from Table 1 are also summed to derive the total time and total distance for each driver in each context. Note that the trips without an event of interest (i.e., outcome zero) are summed and included in the corresponding total distance and time for each cohort. A dummy variable designation is employed for the context variables and driver attributes.

Table 1: Initial Cohort-Based Data Structure

Outcome (0/1)	Length	Time	Event Attributes (as many as needed)	Context Variables (as many as needed)	Driver Attributes (as many as needed)

Table 2: Summed Event Outcomes by Context and Driver Attributes with Exposure Measures

# of Outcomes (count)	Total Length (veh. mi.)	Total Time (Veh.hours)	Event Attributes (as many as needed)	Context Variables (as many as needed)	Driver Attributes (as many as needed)

One may now aggregate the data as in Table 2 and use a count regression approach to estimate the number of surrogate events in each cohort. Count regression formulations developed using the cohort structure thus readily include the amount of travel within each cohort including non-crashes as well as crashes (or other events of interest such as alert warnings). By aggregating the exposure units in column 2 or three, one can estimate exposure to risk within each cohort.

A multilevel specification can also be considered to improve model precision. The unit of the first level is the context combination, and the unit of the second level is the individual driver. Concerning multilevel model random effect covariance, we can either assume the second-level predictors are independent of each other or that they are correlated to each other. Random-effects negative binomial (RENB) models have been applied previously in NDS data analyses (Jovanis et al., 2011). Notice that the context variables include those normally associated with the development of the safety performance function (SPF) in engineering studies of road safety. The enhancement with naturalistic data is the explicit inclusion of driver-related variables which provide an enhanced understanding of driver-related factors and crash (or in this case, alert risk). Exposure in this NDS application is derived from the cohort formulation while in engineering studies in typically comes from road traffic counting programs.

One of the immediate applications of this approach is to statistically distinguish different event types. To group events with similar contributing factors and etiologies, a counterpart to the Chow test as suggested by Greene (2003), is proposed to undertake this task. The procedure tests whether the log-likelihood for a pooled-dataset model is significantly different from the sum of log-likelihoods for reduced dataset models.

DATA DESCRIPTION

The UMTRI data consist of NDS-measured driving for a set of drivers who experienced a series of alerts about potential crashes from on-board safety systems (Leblanc et al., 2006; Sayer et al., 2005). The dependent variables used in the analyses were derived from a system designed to detect excessive speed entering a curve (the Curve Speed Warning System or CSW) and an alert triggered when the subject vehicle deviated from the lane or road edge (i.e., the Lateral Drift Warning System or LDW). The data include the number of miles and length of time driven on road segments of particular geometry and environmental conditions by each driver as summarized in Table 1 and as later processed into Table 2.

The data were collected in 11 instrumented 2003 Nissan Altima 3.5 SE sedans. A total of 87 drivers were enrolled during 2004 to 2005, and each driver drove the car for four weeks with the first week with the system disabled. To demonstrate this approach can be used to statistically distinguish different event types, the response variables used in this study are the number of CSW alerts, LDW alerts, and the total number of alerts. The predictors included the following:

- Context variables: road functional class, ramp, urban/rural, day/night.
- Event variables: wet/dry (based on windshield wiper use), system disabled/enabled status (i.e. either the data were collected in the first week (system disabled) or in weeks 2-4).
- Driver variables: gender, education, years of driving experience, last year's mileage driven, use of glasses or contacts, smoker.

A final sample size of 72 drivers is included in this study. On average, there were 32 cohorts experienced by each driver with the minimum of 9 and the maximum of 52.

DATA ANALYSIS

Table 3 contains the estimation results of the three RENB models with gamma distribution for the random effect. Models were estimated using distance and time as exposure, but the results are consistent, so only distance is shown here. The likelihood-ratio tests (the row in the bottom) for all the models indicate that RENB models fit better than its counterpart using pure negative binomial model, suggesting significant differences between drivers. The first model considers the number of total alerts (pool model), and it was broken down into models for LDW and CSW separately (reduced models). As shown in Table 3, the magnitudes and signs between the pool and reduced models are quite different, and the Chow test confirms the structural difference between CSW and LDW models (p-value = 0.000).

Although all driver, event, and context variables were included in the models, only predictors achieving statistical significance in LDW and CSW models are discussed here. LDW alerts are 30 percent more likely to be triggered on minor arterial than on freeways ((exp(0.263)-1)*100 = 30), but are 95 percent less likely to be triggered on local roads ((1-exp(-3.05)*100 = 95). Although the former is intuitive, the later may reflect the fact that there is no lane marker on local roads in many occasions. The variable "system disabled" shows that the LDW triggered were reduced by 11 percent after the system was enabled. The variable of miles driven in last year is a self-reported variable, and is used as a proxy to reflect the amount of driving for a driver in a year. Drivers with more miles driven in a year were found to less likely to trigger LDW alerts. Compared to freeways, CSW alerts are more likely to be triggered on every other functional class, in particular, on ramps. The "system disabled" variable is significant and negative in the CSW model, indicated more alerts with system on than with system off. We interpret this is as possibly being related to driver adaptation to the device, resulting in more alerts with the device activated. Interestingly, the variable of years of driving is considered to be an indicator of driving experience, and it was found to be beneficial in reducing frequency of CSW alerts. Consistent with the findings in Jovanis et al. (2011), some correlation between event attributes and driving environment were found. As an example, the correlation coefficient of speed differentials on a homogeneous trip segment and minor arterial is as high as 0.2. Failure to include either one of them would lead to biased estimates.

By including exposure measures, RENB regression enables the comparison between safety-related events and non-safety-related events, the baseline. The relationship between event frequency and miles-traveled is still non-linear, same as identified in most current SPFs. The magnitude of miles-traveled for LDW model is greater than that for the CSW model. Although this finding indicates LDW alters occurs more frequently than CSW alters in terms of the same distance travel, it may be simply because CSW alters could only be triggered on horizontal curves, but LDW could be triggered everywhere there are lane markings.

Table 3. Random-effects negative binomial models

	CSW and LDW	LDW	CSW
Logarithm of Miles Traveled on a Homogeneous Trip Segment	0.901***	0.905***	0.812***
S.E.	0.016	0.019	0.035
Major Arterial with Limited Access (Baseline: freeway)	0.284	0.275	0.575
S.E.	0.243	0.254	0.723
Major Arterial (Baseline: freeway)	0.179***	0.043	1.218***
S.E.	0.05	0.057	0.132
Minor Arterial (Baseline: freeway)	0.362***	0.263***	1.239***
S.E.	0.044	0.048	0.13
Local (Baseline: freeway)	-0.832***	-3.051***	1.671***
S.E.	0.098	0.26	0.158
Ramp (Baseline: freeway)	1.668***	0.284*	3.019***
S.E.	0.074	0.117	0.142
Urban (Baseline: Rural)	0.162**	0.186**	0.183
S.E.	0.055	0.061	0.126
System Disabled	0.066	0.108*	-0.134
S.E.	0.038	0.044	0.08
Miles Driven in Last Year	-0.113	-0.174*	0.048
S.E.	0.061	0.071	0.08
Years of driving	0.001	0.004	-0.011**
S.E.	0.003	0.003	0.004
Male	0.169	0.137	0.12
S.E.	0.094	0.106	0.124
Speed Differential on a Homogeneous Trip Segment	-0.006	-0.004	-0.006
S.E.	0.004	0.005	0.008
Constant	-2.320***	-2.426***	-4.609***
S.E.	0.171	0.197	0.288
Sample size (72 drivers)	2325	2325	2325
Log-likelihood	-3576.935	-2818.68	-1933.15
P-value for likelihood-ratio test again pure NB model	0.0000	0.0000	0.0000

*significant at 10% level; **significant at 5% level; ***significant at 1% level

SUMMARY AND DISCUSSION

The cohort model formulation takes advantage of the trip-by-trip information in the UMTRI data set, along with additional GIS-related factors coded by UMTRI (such as road type and environmental conditions) to derive the alert frequency in each trip segment. The issue of interest is the ability to truly capitalize on not only the naturalistic driver behavior data, but detailed GIS roadway data. Further, availability of detailed GPS readings as part of NDS will allow for more

detailed roadway descriptors possibly at a finer scale. This would reduce the level of aggregation of speed along a route and undoubtedly yield more precise findings for cohorts with promise of safety improvement. Lastly, if the duration of the NDS is a year or more (as in SHRP 2 Safety) then the additional predictor for annual mileage would not be needed; the precise measured driving from the NDS GPS and GIS should suffice.

This approach has many potentially useful applications. Even though the models are estimated with *alerts*, there is a direct parallel to the modeling of *crashes* or other events of interest such as well-behaved surrogates similar to crashes as described in Wu and Jovanis (2012a and b). This formulation is also superior for evaluating the effectiveness of in-vehicle safety devices in FOTs; compared to a naïve before-after design the cohort formulation includes exposure while also considering all predictors simultaneously. Moreover, the creation and inclusion of interactions terms between system on/off and context variables can be used to better understand when the systems work better for driver safety.

This approach is not without limitations. First, although the cohort may be defined quite flexibly, little is known about the minimal number of required cohorts and how should the cohorts be defined This issue is related to a sampling zero problem, as well as whether the decomposition or aggregation is meaningful. As an example, it may not be appropriate or useful to include kinematic variables in a specification or model of this type because it would lead to a sampling zero problem. On the one hand, the aggregation of average values for each kinematic variable may be problematic because they may be affected by factors that could be used to redefine homogenous trip segments, but they are not included in the data set. For example, suppose additional variables were included in the dataset, including curve/tangent presence, presence of an intersection, traffic volume, and grade. These could be used to redefine homogeneous trip segments. Once we redefine the segments, average speeds would more accurately reflect travel speeds on each segment. The averages of kinematic values over homogeneous trip segments may lack resolution about event occurrence during the course of the traversal of the entire segment.

REFERENCES

Cameron, A. C., Trivedi, P. K., (2005). Microeconometrics: Methods and Applications, Cambridge University Press.

Dingus, T.A., S.G. Klauer, V.L. Neale, A. Petersen, S.E. Lee, J. Sudweeks, M.A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z.R. Doerzaph, J. Jermeland, and R.R. Knipling (2005). *The 100-Car Naturalistic Driving Study, Phase II—Results of the 100-Car Field Experiment*, National Highway Traffic Safety Admin (DOT HS 810 593).

Greene, W.H., (2003). Econometric Analysis, 5th ed. Prentice Hall, New York.

Hanowski, R.J., M. A. Perez, Dingus, T. A. (2005). Driver distraction in long-haul truck drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 8(6), pp. 441-458.

Hauer, E. (1982). Traffic Conflicts and Exposure. Accident Analysis and Prevention, 14(5), pp. 359-364.

Hilbe, J.M. (2010). Negative Binomial Regression, second edition. Cambridge University Press.

Jovanis, P.P., Aguero-Valverde, J., Wu, K., Shankar, V. (2011). Naturalistic Driving Event Data Analysis: Omitted Variable Bias and Multilevel Modeling Approaches. Transportation Research Record: Journal of the Transportation Research Board, No. 2236, pp. 49-57.

Jovanis, P.P., Shankar, V., Aguero-Valverde, J., Wu, K., Greenstein, A. (2012). Analysis of Existing Data: Prospective Views on Methodological Paradigms. The Strategic Highway Research Program 2, Transportation Research Board of The National Academies of Sciences.

Jovanis, P.P., Agüero Valverde, J, Wu, K.F. and Shankar, S., Naturalistic Driving Event Data Analysis: Omitted Variable Bias and Multilevel Modeling Approaches , Journal of the Transportation Research Board No. 2236, p 49-57, 2011

Leblanc, D., J. Sayer, C. Winkler, R. Ervin, S. Bogard, J. Devonshire, M. Mefford, M. Hagan, Z. Bareket, R. Goodsell, and T. Gordon. (2006). Road Departure Crash Warning System Field Operational Test: Methodology and Results. The University of Michigan Transportation Research Institute.

McGehee, D. V., Boyle, L. N., Hallmark, S. J., Lee, D., Neyens, D. M., Ward, N. J. (2010). S02 Integration of Analysis Methods and Development of Analysis Plan Phase II Report. The Strategic Highway Research Program 2, Transportation Research Board of The National Academies of Sciences.

Sayer, J.R., Mefford, M. L., Shirkey, K., Lantz. L. (2005). Driver Distraction: Naturalistic Observation of Secondary Behaviors with the Use of Driver Assistance Systems. Driver Assessment 2005: Third International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.

Shankar, V., Jovanis, P.P., Aguero, J., Gross, F. (2008). Analysis of naturalistic driving data: a prospective view on methodological paradigms. Transportation Research Record of the Transportation Research Board, No. 2061, pp. 1-8.

Wu, K., Jovanis, P.P. (2012a). The Relationships of Crashes and Crash-Surrogate Events in Naturalistic Driving. Accident Analysis and Prevention, 45, pp. 507-516.

Wu, K., Jovanis, P.P. (2012b). Defining, Screening, and Validating Crash Surrogate Events Using Naturalistic Driving Data. Accident Analysis and Prevention, in press.

Wu, K., Jovanis, P.P. (2013). Screening Naturalistic Driving Study Data for Safety-critical Events. Transportation Research Record: Journal of the Transportation Research Board, in press.