# Exploring the Factors that Impact Injury Severity using Hierarchical Linear Modeling (HLM)

## Introduction

Injury Severity describes the severity of the injury to the person involved in the crash. Understanding the factors that influence injury severity can be helpful in designing mechanisms to reduce accident fatalities. Several research studies have examined the relationship between injury severity and the possible factors that can be related to the severity of the injury. These include driver-related factors, vehicle characteristics, and road and weather conditions at the time of the accident. Various data mining techniques were used to identify the possible factors that can be associated with the occurrence of the crashes. Examples of the techniques include classification algorithms, network models, and exploratory models [4][5][6].

However, in analyzing the impact of various factors on injury severity, past research ignores the hierarchical nature of the data related to the phenomena. Weather, road, vehicle and driver factors exhibit a hierarchical structure when analyzing their impact on injury severity. When the hierarchical nature of the data is not taken into account, data mining methods or traditional statistical methods such as ordinary least squares are not sufficient to isolate the impact of each level on injury severity since the observations cannot be considered as fully independent. In this research, in contrast, we model and analyze the data as a hierarchy with three levels to isolate the impact of each level on injury severity. At the highest level of the hierarchy (level-3) are road level variables such as route and road surface condition. Situated at the middle level of the hierarchy (level-2) are vehicle variables such as vehicle make. Variables at the lowest level of the hierarchy (level-1) are nested within level-2 groups. Estimating the impact of each level on injury severity allows transportation policy makers to design customized mechanisms specific to each level to reduce the occurrence of fatal accidents.

Because our study involves different hierarchy levels, we used hierarchical linear model (HLM), for analyzing nested data. HLM can handle data in different levels of hierarchy to separate the influence of the variables in these levels on the dependent variable. Moreover, HLM is a common statistical technique and has been used in many domains such as education, health, business, and social work sectors. This technique is known by several names, including multilevel mixed level, mixed linear, mixed effects, random effects, and complex covariance component modeling [3].

## Data

Data were drawn from Fatality Analysis Reporting System (FARS)[1]. This dataset represents U.S. road accident information for 50 state from the year 2011 to 2013. It consists of 190835 records and 244 attributes. For our study, the attributes came from (PARKWORK, PERSON, VEHICLE, ACCIDENT, VIOLATION, MANEUVER) tables. These tables were merged based on attribute ST_CASE. The study variables and description are shown in Table 1 (Appendix 1) .

## Problem

The problem for this study is to investigate whether level-4, level-3, level-2 and level-1 affect injury severity. The study question involves a hierarchy with many levels with the highest level of the hierarchy being weather conditions. Level-3 variables are nested within level-4 (weather) groups and are impacted by level-4 variables. For example, the areas that have bad weather condition would probably have bad road surface conditions than the areas in good weather condition. Level-2 variables are nested within level-3 groups and are impacted by level-3 variables. For example, the road with bad surface condition would probably increase vehicle failures or damage than a road with good surface condition. Variables at the lowest level of the hierarchy (level-1) are nested with level-2 groups. For example, person level factors (level-1) are situated within a vehicle related factors (level-2). For an instance, certain vehicle make for sports car, for example, would influence driver's ability to have a control with a high speed. Table 2 shows the hierarchical levels and Table 3 presents the study hypotheses.

**Table 2. Factors at each Hierarchical Level that Affect Injury Severity.**

| **Hierarchical level** | **Variables** |
|---|---|
| Weather (Level-4) | Atmospheric Conditions |
| Road (Level-3) | Road Surface Condition, Total Lanes in Roadway, Route Signing |
| Vehicle (Level-2) | Vehicle Make |
| Person* (Level-1) | Speed Related, Alcohol Involvement, Driver Maneuvered to Avoid, Injury Severity* |
| *In person level, we only examined driver-related factors.<br>*The outcome variable (Injury Severity) is always at the lowest level variable. | |

**Table 3. Research Hypotheses**

| |
|---|
| **$H_1$**: Person level factors are related to the injury severity. |

[1] http://www.nhtsa.gov/FARS

| |
|---|
| **$H_2$**: Vehicle make moderates the relationship between person level factors and injury severity. |
| **$H_3$**: Road related factors moderate the relationship between person level factors and injury severity. |
| **$H_4$**: Weather condition moderates the relationship between person level factors and injury severity. |

**Data Cleaning**

In our data set, all attributes values are in continuous scale. We replaced these values with its equivalent categorical information as per the guidelines shown in [2]. Therefore, the continuous variables are converted into categorical ones.
For these categorical attributes, we created several dummy variables as follows.

1) SPEEDREL: For this variable we created one dummy variable (SPEEDING) with values 0/1, where 1 represents speeding_yes otherwise 0.
2) MDRMANAV: For this variable we created one dummy variable (MANV_ATTEMPT) with values 0/1, where 1 represents driver did something to avoid the accident otherwise 0.
3) DRINKING: We created one dummy variable for this variable (DR_DRINK) with values 0/1, where 1 represents driver is drunk otherwise 0.

We did not create dummy variables for weather, road, and vehicle related variables because these variables will be used as class variables [1].
In this data cleaning process, we also checked for multicollinearity and we found that the Variance Inflation Factors (VIF) were within the acceptable threshold (less than 10).

**Analysis**

The analysis was performed with SAS (version), using PROC GLIMMIX. The dependent variable is injury severity which describes the severity of injury for the individuals involved in the crash. Unlike past studies, we converted this variable into polytomous outcomes and include four categories (1 = severe fatality, 2 = severe incapacitating injury, 3 = possible injury, 4 = no injury). Figure A and Table A presents the frequency of different injury severity categories (Appendix 1). The purpose of this categorization is to identify the multiple levels of injury severity. Because the outcome variable is ordinal, we performed PROC GMLMIX procedure as this procedure works for both dichotomous and polytomous outcomes [1].

In the first iteration of our analysis, we tested the effect of each of the four hypothetical hierarchical levels, using PROC GLIMMIX. We found driver related factors have a significant impact on injury severity. On the other hand, based on the Inter-class Correlation Coefficient

(ICC)[2] value and p-values, we observed that the amount of variability by our hypothetical level (weather) is not statistically significant. Moreover, the variability by road surface conditions and total lanes in roadway are not statistically significant. However, the hypothesis of using vehicle make from vehicle level and route type from road level have significant variability and therefore we further explore these effects on injury severity.

Therefore, using person-level data (level-1), vehicle-level data (level-2) and road-level data (level-3), we build two hierarchical models to investigate the relationship between the predictor variables and the injury severity at the three levels. We build the model with one of the class variable using person level variables as fixed effects and the class variable with random effect (see code in Appendix 1). For each class variable, we computed the ICC that indicates how much the total variation in the probability in the injury severity is accounted for by the given class. The solutions from the fixed effect table generated by PROC GLIMMIX is shown below.

**Table 4. Moderation Effect on Injury Severity of Vehicle Make with Predictors**

| Solutions for Fixed Effects | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Effect | INJ_SEVR | Estimate | Standard Error | DF | t Value | Pr > \|t\| | Alpha | Lower | Upper |
| Intercept | 1 | -0.7287 | 0.1183 | 69 | -6.16 | <.0001 | 0.05 | -0.9647 | -0.4927 |
| Intercept | 2 | -0.2024 | 0.1183 | 69 | -1.71 | 0.0915 | 0.05 | -0.4384 | 0.03354 |
| Intercept | 3 | 0.9261 | 0.1183 | 69 | 7.83 | <.0001 | 0.05 | 0.6900 | 1.1621 |
| MANV_ATTEMPT | | -0.5336 | 0.02539 | 60885 | -21.02 | <.0001 | 0.05 | -0.5834 | -0.4838 |
| SPEEDING | | 0.7798 | 0.02010 | 60885 | 38.79 | <.0001 | 0.05 | 0.7404 | 0.8192 |
| DRINKING | | 0.7438 | 0.02075 | 60885 | 35.84 | <.0001 | 0.05 | 0.7032 | 0.7845 |

In Table 4 we observed that maneuver attempt is negatively associated with the Injury Severity. The coefficient is negative and significant (coeff. = -0.5336, p <0.0001). That means if the driver did something to avoid the accident prior to the crash event, this will reduce the fatality of the Injury Severity by (0.5336). We also observed that the speeding is positively related with Injury Severity (coeff. = 0.7798, p<0.0001). The coefficient is positive and significant. Similarly, drinking is positively associated with Injury Severity with significant coefficient (coeff. = 0.7438, p<0.0001).

**Table 5. The Estimate of Moderation Effects (MAKE) on Injury Severity**

| Covariance Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Cov Parm | Subject | Estimate | Standard Error | Z Value | Pr > Z |
| Intercept | MAKE | 0.8513 | 0.1656 | 5.14 | <.0001 |

[2] ICC indicates how much the total variation in the probability in the injury severity is accounted for by the given class.

As we hypothesized that vehicle characteristic moderate the relationship between person level factors and injury severity, we found the moderation effect of vehicle make has statistically significant amount of variability in Injury Severity between different types of vehicle make of our sample (coeff. = 0.8513, p<0.0011). As shown in Table 5, we calculated ICC and found that approximately 20.6% of the variability in the Injury Severity that is accounted for by the vehicle MAKE in our study, leaving 79.4% of the variability by person related factors (MDRMANAV, SPEEDING, DRINKING) or other unknown factors.

$$ICC = \boldsymbol{\tau}_{00} \ / \ (\boldsymbol{\tau}_{00} + 3.29^{3})$$
$$= 0.8513 \ / \ (0.8513 + 3.29)$$
$$= 0.2056$$

**Table 6. Moderation Effect on Injury Severity of ROUTE with Predictors**

| Solutions for Fixed Effects | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Effect | INJ_SEVR | Estimate | Standard Error | DF | t Value | Pr > \|t\| | Alpha | Lower | Upper |
| Intercept | 1 | -0.8650 | 0.07115 | 8 | -12.16 | <.0001 | 0.05 | -1.0291 | -0.7009 |
| Intercept | 2 | -0.3696 | 0.07108 | 8 | -5.20 | 0.0008 | 0.05 | -0.5335 | -0.2057 |
| Intercept | 3 | 0.7089 | 0.07112 | 8 | 9.97 | <.0001 | 0.05 | 0.5449 | 0.8729 |
| MANV_ATTEMPT | | -0.5489 | 0.02485 | 60946 | -22.09 | <.0001 | 0.05 | -0.5976 | -0.5002 |
| SPEEDING | | 0.8692 | 0.01977 | 60946 | 43.96 | <.0001 | 0.05 | 0.8305 | 0.9080 |
| DRINKING | | 0.8167 | 0.02036 | 60946 | 40.12 | <.0001 | 0.05 | 0.7768 | 0.8566 |

Similar to the results shown in Table 5, in Table 6, we observed that maneuver attempt is negatively associated with the Injury Severity. The coefficient is negative and significant (coeff. = -0.5489, p <0.0001). That means if the driver did something to avoid the accident prior to the crash event, this will reduce the fatality of the Injury Severity. We also observed that the speeding is positively related with Injury Severity (coeff. = 0.8692, p<0.0001). The coefficient is positive and significant. Similarly, drinking is positively associated with Injury Severity with significant coefficient (coeff. = 0.8167, p<0.0001).

**Table 8. The Estimate of Moderation Effects (ROUTE) on Injury Severity**

| Covariance Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Cov Parm | Subject | Estimate | Standard Error | Z Value | Pr > Z |
| Intercept | ROUTE | 0.04241 | 0.02096 | 2.02 | 0.0215 |

[3] Stawski, R. S. (2013). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 541-550.

As we hypothesized that road conditions moderate the relationship between person level factors and injury severity, we found here the moderation effect of road ROUTE has statistically significant amount of variability in Injury Severity between different types of vehicle make of our sample (coeff. = 0.04241, $p<0.05$). As shown in the table 8, we calculated ICC and found that approximately 2.15% of the variability in the Injury Severity that is accounted for by the vehicle ROUTE in our study, leaving 97.85% of the variability by driver related factors (MDRMANAV, SPEEDING, DRINKING) or other unknown factors.

$$ICC = \boldsymbol{\tau}_{00} / (\boldsymbol{\tau}_{00} + 3.29)$$
$$= 0.04241 / (0.04241 + 3.29)$$
$$= 0.0215$$

**Generalization**

We considered multi-years (2011 - 2013) across all U.S. states to examine the influence of multi hierarchy level of predictor variable on injury severity. Therefore, we claimed that our result are sufficiently generalizable.

**Findings and Implications**

In this paper, we analyzed the impact of driver, vehicle, and road related factors on injury severity. The results show that speed, drinking, and maneuver variables from the person level data are directly related to injury severity. Route from road level and vehicle make from vehicle level have significant moderation impact on injury severity. Also, compared with route class, vehicle make explains more variability in injury severity variable.

The findings have important implications for driver-related factors (speed, drinking, and maneuver) effects on injury severity. This provides understanding about the importance of speed and drinking policies as well as policies related to educating a driver about accident avoidance/maneuver behavior on reducing fatal accidents.

The model also provides understanding about the role played by vehicle design/make in affecting injury severity. The findings provide some important insights about the standards that should be given attention when designing or manufacturing vehicle as this can reduce injury severity levels. The finding on the significant moderation effect of route variable on injury severity also has important implication for road design and speed limit determination policies on different road types.

**Suggestion for Future Studies**

In this research, data collected from 2013 across 50 US states were used to build HLM models with road, vehicle, and person level factors and injury severity. The conclusion was reached that HLM has the best performance in explaining the effect of nested variables on injury severity. However, this research has some drawbacks which are stated as follow. The quality of the research is dependent on the quality and availability of the data. It is recommended to obtain more valuable data especially for weather conditions from other sources and merge it with FARS dataset as this will be very helpful in improving future research. Moreover, in person level, we only examined driver-related factors. Therefore it is also recommended to explore other uninvestigated person level variables.

## Conclusion

This paper explores the impact of driver, vehicle, and road related factors on injury severity. This project also provides a use case illustrating that data mining/HLM technique is beneficial in studying the influence of multi hierarchy levels of predictor variables on injury severity. The results support our hypotheses regarding the direct impact of driver-related factors on injury severity and the moderation impact of vehicle and road related variables on injury severity.
We believe that this study has important policy implications for designing mechanisms to reduce accident severity.

## References

[1] Ene, M., Leighton, E. A., Blue, G. L., & Bell, B. A. Multilevel Models for Categorical Data Using SAS® PROC GLIMMIX: The Basics.

[2] United States Department of Transportation (2014). *Fatality Analysis Reporting System (FARS) Analytical User's Manual 1975-2012*. Retrieved from http://www-nrd.nhtsa.dot.gov/Pubs/811855.pdf

[3] Raudenbush, S. W. Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, second edition*. Newbury Park, CA: Sage.

[4] Shanthi, S., & Ramani, R. G. (2012). Feature relevance analysis and classification of road traffic accident data through data mining techniques. In*Proceeding of WCECSC2012 Conference, San Francisco (October 2012).*

[5] Singh, Santokh. "Identification of driver and vehicle characteristics through data mining the highway crash data." *2003 Conference Federal Committee on Statistical Methodology, Arlington, Virginia*. 2003.

[6] Tseng, W. S., Nguyen, H., Liebowitz, J., & Agresti, W. (2005). Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files. *Industrial Management & Data Systems*, *105*(9), 1188-1205.

## Appendix 1

**Table 1. Study Variables**

| **Variable (SAS Name)** | **Description** |
|---|---|
| Injury Severity (INJ_SEV) | This variable describes the severity of the injury of person and crash. |
| Vehicle Make (MAKE) | This variable identifies the make (manufacturer) of this vehicle. |
| Roadway Surface Condition (VSURCOND) | This variable identifies the attribute that best represents the roadway surface condition prior to this vehicle's critical pre-crash event. |
| Total Lanes in Roadway (VNUM_LAN) | This variable identifies the attribute that best describes the number of travel lanes just prior to this vehicle's critical pre-crash event. |
| Route Signing (ROUTE) | This variable identifies the route signing of the traffic way on which the crash occurred. |
| Atmospheric Conditions (WEATHER) | This variable records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the case material. |
| Speed Related (SPEEDREL) | This variable records whether the driver's speed was related to the crash as indicated by law enforcement. |
| Alcohol Involvement (DRINKING) | This variable records whether alcohol was involved for this person and reflects the judgment of law enforcement. |
| Driver Maneuvered to Avoid (MDRMANAV ) | This variable identifies the thing(s) this driver attempted to avoid while the vehicle was on the road portion of the traffic way, just prior to the first harmful event for this vehicle. |

```
PROC GLIMMIX DATA=WORK.sample METHOD=LAPLACE NOCLPRINT;
        CLASS MAKE;
        MODEL INJ_SEVR = MANV_ATTEMPT SPEEDING DRINKING / CL
DIST=MULTI LINK=CLOGIT SOLUTION ODDSRATIO (DIFF=FIRST LABEL);
        RANDOM INTERCEPT / SUBJECT = MAKE TYPE=VC SOLUTION CL;
        COVTEST / WALD;

PROC GLIMMIX DATA=WORK.sample METHOD=LAPLACE NOCLPRINT;
        CLASS ROUTE;
        MODEL INJ_SEVR = MANV_ATTEMPT SPEEDING DRINKING / CL
DIST=MULTI LINK=CLOGIT SOLUTION ODDSRATIO (DIFF=FIRST LABEL);
        RANDOM INTERCEPT / SUBJECT = ROUTE TYPE=VC SOLUTION CL;
```

```
COVTEST / WALD;
```

**Figure A. Bar Plot of Injury Severity level**

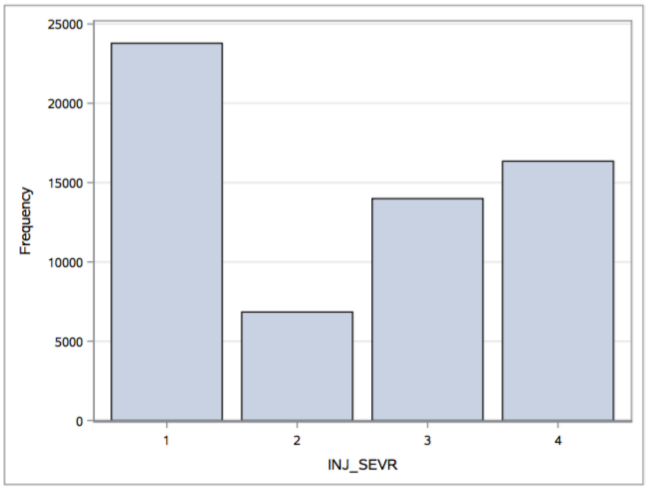


**Table A. Different categories of Injury Severity (INJ_SEV).**

| INJ_SEVR | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 2 | 6839 | 11.22 | 6839 | 11.22 |
| 1 | 23787 | 39.02 | 30626 | 50.24 |
| 3 | 13988 | 22.95 | 44614 | 73.19 |
| 4 | 16346 | 26.81 | 60960 | 100.00 |