Master's Degree in Civil Engineering

Infrastructure and transport systems

Master of Science Degree Thesis

Injury Severity Analysis Based on Police Crash Report

Supervisors

prof. Marco Bassani

prof. Cinzia Cirillo

Candidate

Riccardo Nali

Academic year 2018/2019

Contents

ABSTRACT

Objective. The current research aims to develop a statistical method to analyze police-reported Injury Severity data looking for the most relevant variable, and which condition are more prone to increase the severity level as a crash output. This study tries to contribute to the knowledge about crash data analysis in literature, thanks to the fact of the large amount of data available in the Maryland Crash Database (two years with more than 300K crashes per year).

Methods. Since Injury Severity (IS) level is the dependent variable in the model and being this consisting of discrete ordered value, a Discrete Choice Modelling approach was considered as the most suitable option in this scenario. IS contains discrete and ordered hierarchically value, so the model selected for this study is an Ordered Logit model. Computations were carried out through the Python language using two libraries: Biogeme and Pandas (package PandasBiogeme).

Results. Several independent variables play a role in the IS magnitude. Although the result was found to be overall consistent with literature, in few cases the sign and the significance contrasted with expectations and previous investigations (e.g. gender and age different outcome are interpreted by different point of view and some attribute resulting unexpectedly not significant such as the *Alcohol* effect and *Wet surface* effect). As a result, some considerations were carried out to explain such discrepancies.

Conclusion. The calibrated model provided a good statistical fit. The strong point of this study is the huge availability of data that allow a solid statistical consistency. This could be a starting point to many different insights studies regarding the variables here analyzed, in order to limit the loss of lives on roads and improve road safety.

Keywords:
Ordered Logit Model, Crash Data, Injury Severity, Police Report, ACRS, PandasBiogeme.

1 INTRODUCTION

The issue of road safety is often underestimated. The numbers are worrying. Road traffic crashes now represent the eighth leading cause of death in the world. They count more than 1.35 million lives each year and cause up to 50 million injuries (World Health Organization, 2018). The number of deaths and injuries could be drastically reduced acting in areas such as legislation, vehicle standards, infrastructure design and maintenance, road users education, safety technologies, and so on. Furthermore, the consequences of a crash could be also decreased providing a more efficient access to care. Death and injuries resulting from road crashes represent a serious problem and current trends suggest that it will continue to be the case in the future. Unfortunately, the underestimation of deaths from road accidents is common in many parts of the world and has a lower priority for road safety than other public health challenges. The number of people dying annually in traffic crashes is significantly greater than that due to HIV/AIDS and tuberculosis.

Respectively, in 2016/2017 around 7,277,000/6,452,000 police-reported motor vehicle crashes in the United States were counted, resulting in 37,461/37,133 fatalities and 3,144,000/2,746,000 people injured. Among these crashes, less than 1% (34,439)/(34,247) were fatal, around 30% (2,177,000)/(1,889,000) resulted in at least one injury, and almost 70% (5,065,000)/(4,530,000) were property-damage-only crashes (National Center for Statistics and Analysis, 2018) (National Center for Statistics and Analysis, 2019).

In the United States, all traffic crashes are investigated by police officers who fill the *Police Accident Report* (PAR) immediately after the crash. The report contains information about driver characteristics, vehicle attributes, traffic and environmental conditions, and crash characteristics according to the *Model Minimum Uniform Crash Criteria* (MMUCC), a guideline that defines the minimum set of uniform variables or data elements useful to describe a motor vehicle crash (National Highway Traffic Safety Administation, 2017). The use of MMUCC data elements generates data that can be employed to make more informed decisions, and to finally improve the road safety at the national, State and local scales. States in the US are encouraged to adopt the MMUCC data elements to fill the PARs.

Road safety agencies as *NHTSA* (National Highway Traffic Safety Administration) need high-quality data to develop policies and programs to improve safety and operations in the Nation's roadway network. The improvement of motor vehicle traffic crash data allows the identification of specific traffic safety problems, the communication of safety issues to the public and media, and support better programming and resource allocation decisions, and enable better monitoring.

This study focuses on the injury severity evaluation of road crash data from Maryland state (US) during the years 2016 and 2017. The *Automated Crash Reporting System* (ACRS) is adopted

by the state of Maryland. This reporting system was developed to replace the existing *Maryland State Police* (MSP 1) accident report form used by law enforcement agencies in the state. This new form conforms to the MMUCC standards established by the United States Department of Transportation (Maryland State Police Information Thechnology Division, 2013).

One of the most investigated variable for safety analysis is the injury severity. The crash injury severity level is the unit of measurement of damages occurred at people involved in a crash. There are two different assessments about the injury severity after a crash: the first is provided by the police in the location where the crash occurred; the second is from the Hospital on the basis the medical cares (hospital-assigned). In both cases, the evaluations follow an ordinal scoring system (Burch, 2014).

Generally, the injury severity scale in the PARs report is recorded on the KABCO five-point ordinal scale (**Table 1)**. The KABCO scale is not the same for all US states, in fact it varies across the states depending on the jurisdictions. On the other hand, the hospital scale for the assessment of injury severity is anatomically based on the *Abbreviated Injury Scale* (AIS) that is a six-point ordinal scale (**Table 2**).

Table 1. KABCO scale (FHWA, 2008)

KABCO scale adopted by the Maryland State		
05 = Fatal	K	A fatal injury is any injury that results in death.
04 = Disabled (incapacitating)	A	An incapacitating injury is any injury, other than a fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities he was capable of performing before the injury occurred.
03 = Non-Incapacitating	B	A no incapacitating evident injury is any injury, other than a fatal injury or an incapacitating injury, which is evident to observers at the scene. of the accident in which the injury occurred
02 = Possible	C	A possible injury is any injury reported or claimed which is not a fatal injury, incapacitating injury, or non-incapacitating evident injury.
01 = Not Injured	O	No injury was evident, or the person in question departed from the scene (but was not transported by EMS as an injured person).

Table 2. AIS scale (TRAUMA.ORG, n.d.)

AIS scale	
1	Minor
2	Moderate
3	Serious
4	Severe
5	Critical
6	Survivable

AIS is a global severity scoring system that classifies an individual injury by body region according to its relative severity. AIS represents the base for the *Injury Severity Score* (ISS) calculation of the multiply injured patient (Farmer, 2003) (for more information refer to

Appendix 1). Previous studies found some relevant differences between the KABKO and ASI scoring system suggesting that hospital-base recordings are more reliable compared with the KABCO scale police one (Paleti, 2018) (Isabelle Aptel, 1999).

Other studies (Burch, 2014) found that these two measurements are reasonably consistent: (Compton, 2005) concluded that the police injury scale appears to be an appropriate tool to discriminate the more serious crashes from the multitude of minor ones. Anyway, there is no agreement about the level of discordance and the factors that lead to discordance between KABCO and ASI (Amoros, 2007).

The objective of this research (as part of a wider project, see **Appendix 2**) is to develop a statistical method to analyze police-reported injury severity looking at the most relevant variable concerning road crashes. The empirical analysis in this work was undertaken using the 2016 and 2017 police -recorded crash data in the whole state of Maryland. In particular, the aim is to find out the conditions more prone to increase the injury severity level. Variables such as gender, age, drive under influence, use of seatbelt, vehicle body type, ejection case, rolled over case, pavement condition, weather, collision type, crash in proximity of an intersection, lighting condition, fix object involved in the crash and traffic control type have been investigated. A comparison with the results of other studies is provided in the Results and Discussion section.

An ordered logit model was used in this study to estimate the effect of the most significant factors on accident severity. The code used to calibrate the model adopted the maximum Likelihood inference method for *Generalized Extreme Value* (GEV) *Model.* The dependent variable is injury severity, while a wide variety of independent variables were considered, those include socio-demographics, influence of alcohol, drug or medicaments, vehicle and road conditions, and crash dynamics.

In the literature, discrete response models have been widely used to explore the relationship between driver, environment, road, vehicle characteristics and the severity of injuries suffered by drivers and passengers of vehicles. Many type of discrete models have been considered such as *logistic model* (Al-Ghamdi, 2002)*, Multinomial logit model* (Gudmundur F.Ulfarsson, 2004), *Ordered Probit* model (Kara Maria Kockelman, 2002), (Abdel-Aty, 2003) (Mohammed Quddus, 2002), *Log-linear model* (J.R Schott, 1998), *Mixed logit model* (Joon-Ki Kim, 2013).

In the thesis, comparisons between results with other studies are reported and commented.

2 DATABASE

The data collection process generally requires a large amount of time and resources. In many cases the samples are characterized by a limited number of observations that have to be representative of the whole population. Another important issue is represented by the presence of invalid or not relevant value that should be deleted from the database. This decreases the number of the available and reliable observations and may lead to issues estimation during the model calibration.

The analysis was carried out on data from the ACRS reporting system filled by the Maryland Police agents in 2016 and 2017. As a result, the outputs of the investigation were obtained on the whole population and not on a representative sample. Police recorded all crashes that involved at least one motor vehicle traveling on a road resulting in property damage only (PDO), injury, or death.

The size of this database is of 316.820 observations in 2016, and 309.758 in 2017, which refer to the individual involved in a road crash. The database includes 230 variables grouped in the following categories:

1. people involved: (i) gender, (ii) age, (iii) seat belt use, (iv) drive under effect of alcohol, drugs, medicine or combined effect, (v) occupant ejected or trapped inside the vehicle, (vi) injury severity level;
2. attributes of the vehicles involved: (i) vehicle body type, (ii) vehicle rolled over;
3. roadway geometry: (i) sign/control at the accident location, (ii) intersection type, (iii) road division;
4. environmental factors: (i) lighting conditions, (ii) surface conditions;
5. crash dynamic: (i) type of collision, (ii) collision occurred in an intersection, (iii) fix object involved in the crash.

Before the model computation, a cleaning step is needed for the goodness of the process data. In this study, the data cleaning consists of the elimination of observations with invalid value or missing value in specific “control variable”, generally variable containing basic information. Age and *sex* have been selected as “control variable”, in this case all the observation containing Age equal to “999” and *sex* equal to “99” (values used to fill the field with missing information or having useless values such as “unknown” or “not applicable”) were deleted from the database. This process is possible only if the elimination of them is random, in order to uniformly affect the dependent variable. In fact, if after the cleaning data process all the invalid value regard only a subset of the dependent variable, the outcome will be significantly distorted. At the end of the cleaning process, the size of the dataset was reduced to 267.187 observations for the 2016, and 268.705 observations

for the 2017, with a selection of only 28 out of original 230 attributes considered the most relevant for this study (the whole database is described in the **Appendix 3**). After the data cleaning, the change in percentage of the injury severity distribution is quite uniform as show the **Table 3** and **Table 4**. Thus, this cleaning process did not introduce errors in further computations.

Table 3. Global IS Overview 2016, pre and post data cleaning. Where Δ(U-S) is the difference in number between the original and sanitized database

Global IS overview 2016					
	Original observations	[%]	Sanitized observations	[%]	Δ(U-S)
Fatal injury	542	0.17%	533	0.20%	-0.03%
Incapacitating injury	3660	1.16%	3638	1.36%	-0.21%
Non-incapacitating injury	26594	8.39%	26362	9.87%	-1.47%
Possible injury	25486	8.04%	25143	9.41%	-1.37%
No injury	260538	82.24%	211511	79.16%	3.07%
Total observation	316820		267187		

Table 4. Global IS Overview 2017, pre and post data cleaning. Where Δ(U-S) is the difference in number between the original and sanitized database

Global IS overview 2017					
	Original observations	%	Sanitized observations	%	Δ(U-S)
Fatal injury	567	0.18%	555	0.21%	-0.02%
Incapacitating injury	3770	1.22%	3744	1.39%	-0.18%
Non-incapacitating injury	18737	6.05%	18578	6.91%	-0.86%
Possible injury	33164	10.71%	32619	12.14%	-1.43%
No injury	253520	81.84%	213209	79.35%	2.50%
Total observation	309758		268705		

Some statistics are showed in the **Figure 1**, **2** and **3** for descriptive purpose to have a general database overview about the distribution followed by some attributes.

Figure 1a show the injury severity levels (in percentage) in the years analyzed. The only noticeable differences between the years considered are about the percentages regarding "Possible injury" and "Non-incapacitating injury"; the first one is greater in 2017 and the second one in the 2016. The other percentage remain similar , this small difference between the two years could be attributed to the stochastic nature of the events.

Figure 1b and **1c** show that, after excluding the cases in which the seat-belts were not properly working, the use of seat-belts, decreases significantly the possibility to get involved in a road crash with a high Injury Severity Score. Seat-belts reduce the probability to have any type of injury, with the two most severe levels presenting in proportion the biggest decrease. The trend is easily justified considering that seat-belt are helpful especially in severe accidents, representing one of the most reliable safety tools to use inside a vehicle.

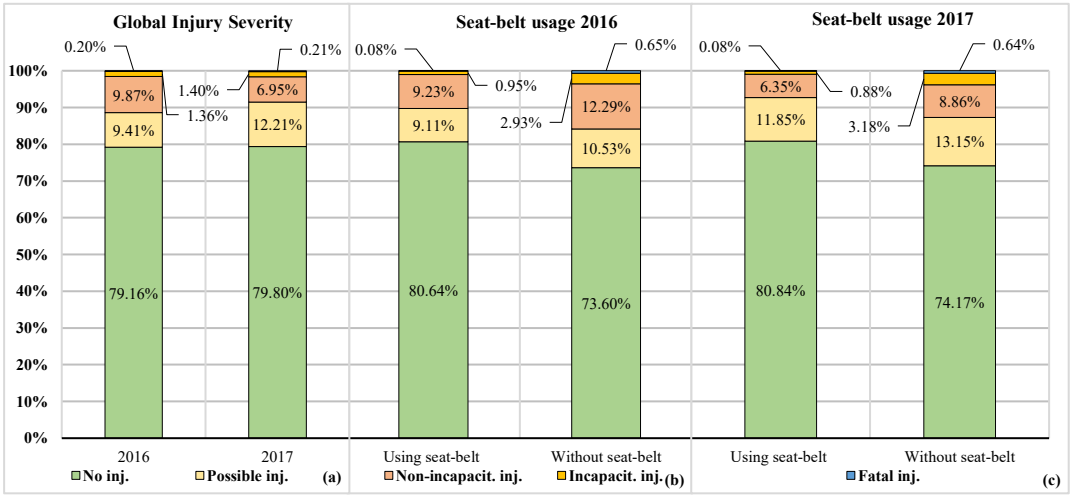


Figure 1. Statistical information about: (a) Global Injury Severity in 2016 and 2017, (b)(c) Seat-belt usage in 2016 and 2017 respectively

This variable can also be studied in relation to the age class as in **Figure 2a** (about 2016) and 2b (about 2017) where it is considered the usage of seat-belts by drivers and occupants into a vehicle. An interesting result is shown in the first age class (0-15 years) where most of the individuals tend to not use the seatbelt.

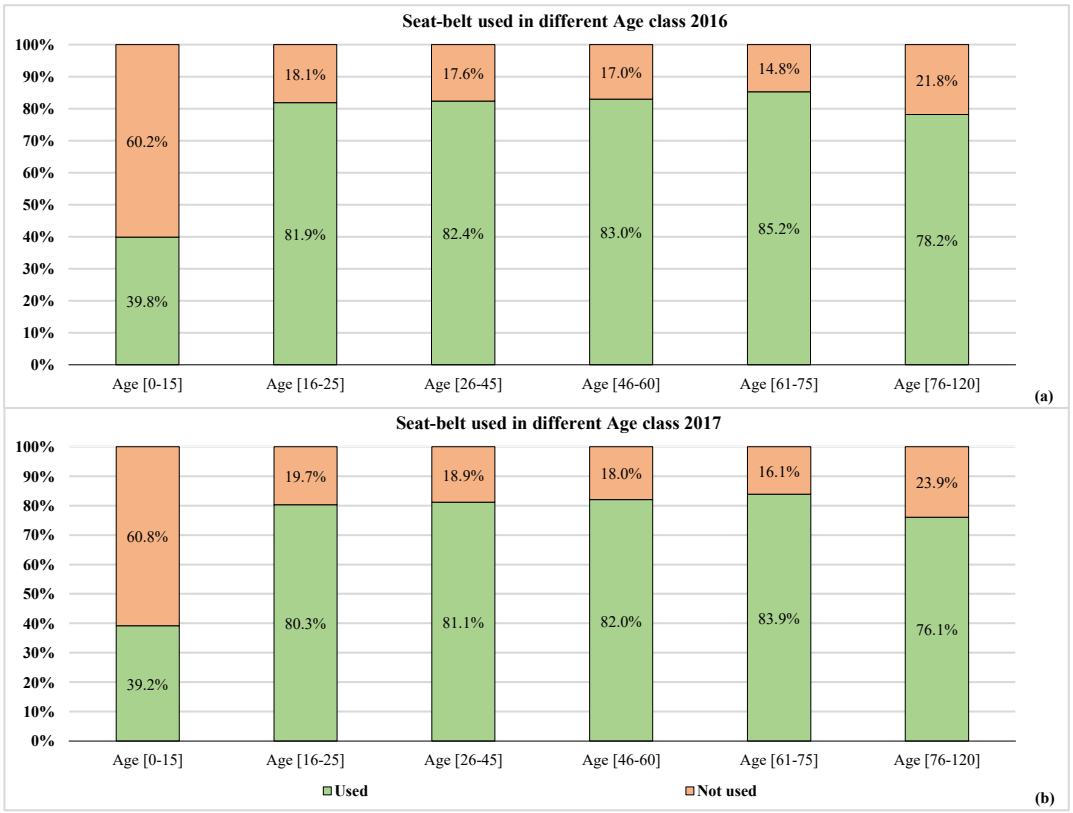


Figure 2. Seat-belt usage in different Age class (a) for 2016 and (b) for 2017

In **Figure 3,** which reports the influence of Alcohol, Drugs, Medicine, Combined effect and Not under influence condition on injury severity, it is interesting to note that driving under the influence of alcohol produces effects that are less severe than those produced by other substances.

Finally **Figure 3c** shows a comparison of the feature between the categories under/over 65, it underline a trend in which people younger than 65 are less prone to get a higher Injury Severity score, both in 2016 and in 2017 with comparable percentages, justified by the decrease in the resilience of the human body to traumatic events with advancing age.

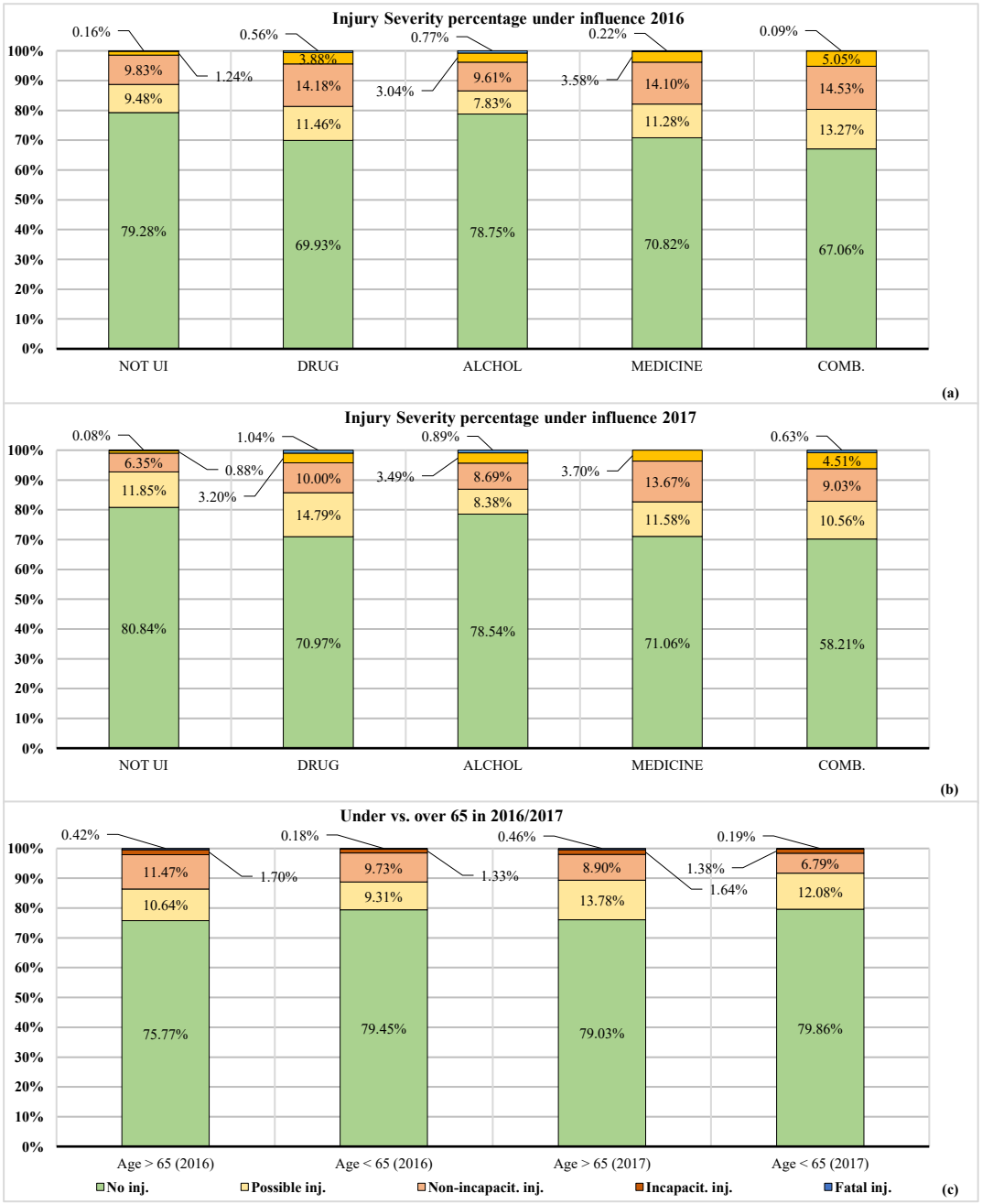


Figure 3. Statistical information about: (a)(b) Under influence effect in 2016 and 2017 respectively, (c) under vs. over 65 in 2016/2017

In order to study the effect of each level in a categorical variable and to make the results of the modeling exercise easier to interpret, the variables have been grouped and reorganized in new dummy variables. For each dummy variable, a *base* is defined as a reference condition. The procedure then transforms the *n* levels of the categorical variable into *n-1* dummy variables *(not considering the variable refereed to the base)*.

Example.

Consider the independent variable called "vehicle" (**Figure 4**) which is defined on three levels that are identified as follows: 1 - car, 2 – van, and 3 – SUV. If one coefficient is estimated for the vehicle variable it is not possible to distinguish the effect of van or SUV on the injury severity score. Therefore, for the vehicle variable we define three dummy variables, one for each level in the original variable specification.

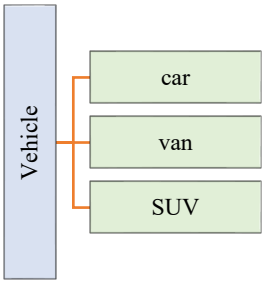


Figure 4. Example of independent variable

The example in **Table 5** further illustrates the way the dummy variables are constructed; in the first observation the variable vehicle assumes the value 1 which corresponds to "car"; then for the first observation the dummy variable for the VEHICLE1 (car) variable is one while the other two dummy variables VEHICLE2 (van) and VEHICLE3 (SUV) are zero.

In the model estimation phase, we will estimate the coefficients for VEHICLE2 (van) and VEHICLE3 (SUV) and we assume VEHICLE1 to be the base; all the estimates are interpreted with respect to the base. In other words, if the sign of these coefficients is positive, this means that the attribute tends to increase IS, and a negative sign goes in the opposite condition. In general, the *base* is chosen considering that most common condition in the database.

Table 5. Example of Dummy transformation

Observations	VEHICLE (original Attribute)
1	car
2	van
3	SUV
4	car
5	SUV
6	SUV
1 coefficient for 1 attribute	

Base		
VEHICLE1 (car)	VEHICLE2 (van)	VEHICLE3 (SUV)
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1
0	0	1
3 coefficients for 3 attributes		

VEHICLE2 (van)	VEHICLE3 (SUV)
0	0
1	0
0	1
0	0
0	1
0	1
2 coefficients for 2 attributes	

Table 6 and **Table 6bis** contain the complete list of variables used for model estimation. These Tables are composed of 3 columns: the *first* shows the list of variables selected from the original database for the estimation process; the *second* column describes the attribute levels and their meaning; the *third* column lists all the new dummy variables used in the computation as a composition of the attribute in the *second* column. After this transformation, the final database will contain only dummy variables.

Table 6. List of dummy variables obtained starting from the variables included in the original database

Variable management		
Original Database	**Value considered**	**New Dummy variable**
Sex	1= Male 2=Female	SEX1 (base = male) SEX2
Age	Numerical value	AGECLASS0 [0-15 years] (= base) AGECLASS1 [16-25 years] AGECLASS2 [26-45 years] AGECLASS3 [46-60 years] AGECLASS4 [61-75 years] AGECLASS5 [76-120 years]
Contrib_code1 Contrib_code2 Contrib_code3 Contrib_code4	1 = Under influence of Drugs (a) 2 = Under influence of Alcohol (b) 3 = Under influence of Medication (c) 4 = Under combined influence (d)	DRUGSEFFECT0 (base = absence of all the variables a, f, i, m, n) DRUGSEFFECT1 (presence of at least one of the variables a, f, i, m, n)
Condition_code	2 = Had been Drinking (e) 3 = Using Drugs (f) 10 = Influenced by Medications/Drugs/Alcohol (g)	ALCHOLEFFECT0 (base = absence of all the variables b, e, h, o, p) ALCHOLEFFECT1 (presence of at least one of the variables b, e, h, o, p)
Alco_drug_impaired_flag	A = Alcohol (h) D = Drug (i) B = Both Alcohol and Drug (l)	
Drug_test_result_flag	P = Positive drug test (m)	MEDICINEEFFECT0 (base = absence of all the variable c) MEDICINEEFFECT1 (presence of at least one of the variables c)
Drug_test_code	2 = Positive preliminary drug test (n)	
Alcol_test_code	2 = Positive preliminary test (o)	DAMEFFECT0 (base = absence of all the variable d, g, l) DAMEFFECT1 (presence of at least one of the variables d, g, l)
BAC	BAC > 0.08 (p)	
Ejection_code	2 = Fully Ejected (a) 3 = Partially Ejected (b) 4 = Trapped (c)	EJTRAP0 (base = absence of all the variable a, b, c) EJTRAP1 (presence of at least one of the variables a, b, c)
Safe_equip_code	11 = Lap belt only (a) 12 = Shoulder belt only (b) 13 = Shoulder/Lap belt(s) (c) 32 = Air Bag & Belts(s) (d)	SEATBELT0 (base = absence of all the variable a, b, c, d or presence at least of the e, f, g, h, i, l, m) SEATBELT1 (presence at least of one of the variable a, b, c, d and absence of all the variable e, f, g, h, i, l, m)
Equip_prob_code	11 = Belts/Anchors Broken (e) 13 = Belt(s) Misused (f) 42 = Facing Wrong Way (g) 43 = Not Anchored Right (h) 44 = Anchor Not Secure (i) 45 =Not Strapped Right (l) 46 =Strap/Tether Loose (m)	
Veh_body_type_code	2 = Car 20 = Pickup Truck 21 = Van 22.05 = Other Light Truck 23.08 = SUV	VEHBODY1 (base = Car) VEHBODY2 (SUV) VEHBODY3 (Pickup Truck) VEHBODY4 (Van) VEHBODY5 (OLT)
Persontype	D = Driver O = Occupant P = Pedestrian	PERSONTYPE0 (base = driver) PERSONTYPE1 (Occupant) PERSONTYPE2 (Pedestrian)

Table 6bis. List of dummy variables obtained starting from the variables included in the original database

Variable management		
Original Database	**Value considered**	**New Dummy variable**
Harm_event_code Harm_event_code_1 Harm_event_code_2	11 = Overturn	ROLLEDOVER0 (base = absence of overturn) ROLLEDOVER1 (presence of overturn)
Rd_div_code	1 = Two-way, Not Divided (a) 2 = One-way Trafficway (b) 3 = Two-way, Divided, unprotected (painted >4 feet) Median (c) 4 = Two-way, Divided, Positive Median Barrier (d) 5 = Two-way, Not Divided with a Continuous Left Turn Lane (e)	ROADTYPE1 (base = presence of at least one of the variables a, b, e) ROADTYPE2 (presence of the variable c) ROADTYPE3 (presence of the variable d)
Surface_code	1 = Wet 2 = Dry 3 = Snow 4 = Ice	SURFCOND1 (base = dry) SURFCOND2 (wet) SURFCOND3 (snow and ice)
Collision_type_code	1=Head On (a) 2=Head On Left Turn (b) 3=Same Direction Rear End (c) 4=Same Direction Rear End Right Turn (d) 5=Same Direction Rear End Left Turn (e) 6=Opposite Direction Sideswipe (f) 7=Same Direction Sideswipe (g) 12=Angle Meets Right Turn (h) 13=Angle Meets Left Turn (i) 14=Angle Meets Left Turn Head On (l)	COLLTYPE1 (base = presence of the variable a and b) COLLTYPE2 (presence of the variable c, d, e) COLLTYPE3 (presence of the variable h, i, l) COLLTYPE4 (presence of the variable g) COLLTYPE5 (presence of the variable f)
Intersection_type_code	1=Four-Way intersection 2=T- intersection 3=Y- intersection 4=Traffic Circle 5=Roundabout 6=Five-point or more	INTERSECTIONTYPE0 (base = absence of all the variables) INTERSECTIONTYPE1 (presence at least of one of the variables listed)
Light_code	1=Daylight (a) 3=Dark Lights On (b) 4=Dark No Lights (c) 5.02=Dawn (d) 6.02=Dusk (e)	LIGHTCOND1 (base = presence at least of one variable between a and b) LIGHTCOND2 (presence of the variable c) LIGHTCOND3 (presence of the variable d) LIGHTCOND4 (presence of the variable e)
Fix_obj_code	1=Bridge or Overpass 2=Building 3=Culvert or Ditch 4=Curb 5=Guardrail or Barrier 6=Embankment 7=Fence 8=Light Support Pole 9=Sign Support Pole 10=Other Pole 11=Tree Shrubbery 12=Construction Barrier 13=Crash Attenuator 14=Guardrail End 15=Concrete Traffic Barrier 16=Other Traffic Barrier 17=Traffic Signal Support4 18=Mailbox 19=Bridge Overhead Structure 20=Bridge Pier Support 21=Bridge Ra	FIXOBJECT0 (base = absence of all variable listed) FIXOBJECT1 (presence at least one of the variables listed)
Traffic_control_code	1=No controls 3=Traffic sign 6=Stop sign 7=Yield sign	TRAFFICCONTROL1 (base = No control) TRAFFICCONTROL2 (Traffic sign) TRAFFICCONTROL3 (Stop sign) TRAFFICCONTROL4 (Yield sign)

3 METHODOLOGY

The injury severity (IS) is a discrete ordered variable, so a Discrete Choice model (Ben-Akiva, 1985) was assumed to be the most suitable model structure for the problem.

A linear model is not the solution for this type of computation because the independent variable is not normally distributed, continue, unbounded and measured on an interval or ratio scale.

Discrete Choice models are used to predict choices between two or more discrete alternatives. It is largely employed in analysis of choice data such as those related to the selection of the residential location or modes of transport (Kenneth E Train, 2007).

Unlike the continuous case, discrete choice analysis examines situation in which the potential outcomes are discrete, such that the optimum is not characterized by standard first-order conditions. Instead of examining "how much" as in problems with continuous choice variables, Discrete Choice models are aimed at examining "which one". This type of analysis is also used to study situation when only a limited quantity must be chosen from (for instance, the number of vehicles owned by a household) (Train, 1986).

In the literature there are numerous examples that use discrete choice techniques to assess the impact of different factors on injury severity data using disaggregate data. Methodological advances have enabled the development of accurate models capable of determining the influence of these factors (Savolainen, 2011) on different problems related to road accidents. Models used for the analysis of accident data include: Binary outcome models, Ordered discrete outcome models, Unordered multinomial discrete outcome models and some other methods.

An accurate review of research on driver injury severity analysis using ordered and unordered response models can be found in (Shamsunnahar Yasmin, 2013).

The most common mechanisms to study driver injury severity are *logistic regression* and ordered response models. The number of studies employing unordered models has been steadily increasing in recent years and the most prevalent unordered response structure considered is the *multinomial logit model*.

In this study the Injury Severity variable is considered as the independent variable. This variable is characterized by an internal hierarchically ordered structure. For this reason, an Ordered Logit model from the Discrete Choice model family was chosen to model injury severity.

The theoretical explanation about Discrete Choice models that follows is mainly inspired by the book Ben Akiva, Lerman, (1985), except in cases where a different source is indicated.

Discrete Choice analysis models individual behavior theory that is:

-*Descriptive* because it is based on the reproduction of human behavior and therefore does not define how people should behave.

-*Abstract* because it can be defined in not specific circumstance.

-*Operational* because allows the measurement of parameters in models where parameters and variables can be estimated.

Alternative theories differ mainly in the level of detail in which they idealize the thought processes (unobservable variables) that produce observed behaviors. Starting from the basis of these theories, the goal is to propose a probabilistic choice model that is the basis of empirical models of discrete choice.

3.1 Fundamentals of Discrete Choice analysis

In general terms a choice could be considered as a result of a decision-making procedure that includes the following fundamentals steps:

- Choice problem definition
- Generation of alternatives
- Evaluation of attributes of the alternatives
- Choice
- Implementation.

In other words, this specific theory of choice is a collection of procedures that defines the following elements: (i) decision maker, (ii) alternatives and the choice set, (iii) decision rule.

It is fundamental to underline that not all the observed choice behavior is a result of a so explicit decision-making process. An individual could be used to making a choice, assume conventional behavior, make a choice by intuition or imitate someone behavior considered an expert in that field.

3.2 The decision maker

The entity to decide can be represented by an individual or a group of people, like a household. To reduce the level of complexity of the interactions among a group of people or an organization it can considered as a single decision maker.

Although the goal is to predict the demand, consider the decision maker as an individual can explicitly determine the differences in the decision-making processes between individuals because they behave in different way depending by the choice conditions.

Due to variation of within group interactions may arise differences among groups decision processes that affect the outcomes.

3.3 The alternatives and the choice set

Each choice come from a non-empty set of alternatives. The environment where the decision maker is placed determines what can be called the universal set of alternatives. Every decision maker considers a subset from the universal set, defined as choice set. This last set contain the alternative that are both available and known to the decision maker during the decision procedure (Swite, 1984) has discussion about the role of environmental and personal constraints on the decision of the choice set.

It could be helpful to distinguish between two general categories of choice sets. In the first category the choice set is continuous and the second one is where the alternatives are naturally discontinuous, as in this study. The evaluation of the attractiveness of an alternative it is considered as a vector of attribute value. The measurement is carried on through an attractiveness scale which can be ordinal or cardinal.

In this study the attention falls in alternatives not-continuous, so discrete, and ordinal.

3.4 The decision rules

To reach a single choice, the decision maker must process the information available from a set of choices containing one or more alternatives. The internal mechanism that regulates this choice is described by the decision rule. The literature (e.g., see the lists given by (P Slovic, 1977) and (Svenson, 1979)). proposes various decision rules that can be classified into four categories.

1. *Dominance rule*. This is the case which an alternative is better for at least one attribute and not worse for all the others, it means that the alternative is dominant even if, in most cases, this does not lead to a single choice. This rule can be used to delete inferior alternatives from a choice set. In the **Figure 5** the alternative 1 and 4 satisfy the dominance condition. In most of the cases there are many attributes of relevance to each decision maker, for this reason is very rare to find an alternative that is dominant over all attributes. To consider that one alternative is better than another, the difference in attribute values must exceed a threshold. This will lead to an additional degree of complexity.

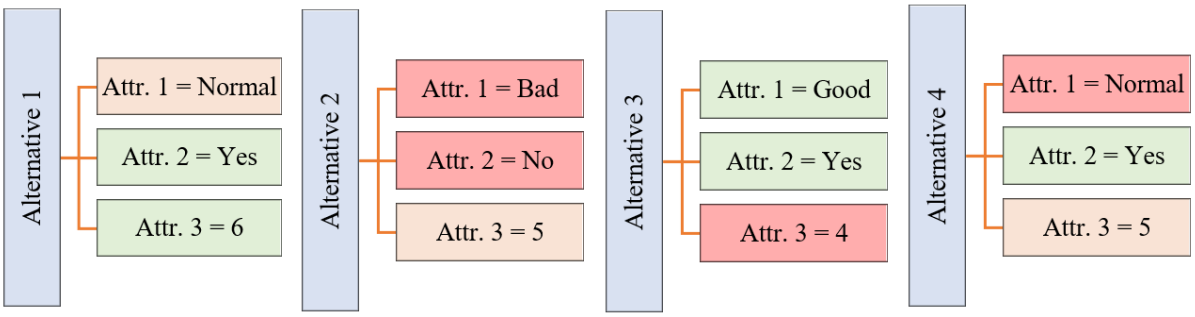


Figure 5. Set of alternative where the worst attributes are colored in red and the best in green.

2. *Satisfaction rule*. It is based on the current information or previous decision maker experience which has a level of aspiration where each attribute taken serves as a criterion of satisfaction. For example, the minimum value admissible for the attribute 3 must be greater than 5 (the only alternative that satisfy this condition is the number 1) see **Figure 5**. For this reason, an alternative can be eliminated if it does not meet the criterion of at least one attribute. This rule will not necessarily lead to a univocal choice. could be helpful to combine this rule with others, such as dominance, it can be more effective.
3. *Lexicographic rule*. If the attributes are sorted by level of importance, the decision maker will choose the most interesting alternative. When the attributes are qualitative, the alternatives that have this quality will be maintained. If this procedure does not return a single choice, the decision maker will pass to the second most important attribute as long as the process does not exclude all alternatives except one. For instance, in **Figure 5**, the decision maker order hierarchy the attribute from the most relevant to the least one as in **Figure 6:**

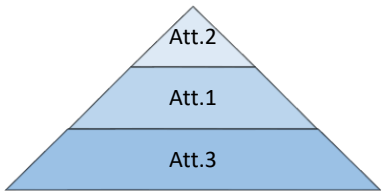


Figure 6. Hypothetical hierarchy of attribute

Following this scheme, the resulting alternative that will be choose by the decision maker is the number 3. The merge between lexicographic and satisfaction rules is also known as "elimination by aspects" (Tversky, 1972).

4. *Utility*. In this rule class the attractiveness of an alternative is expressed by a vector of attributes values by scalars. This defines a single function expressing the attraction of an alternative considering its attributes. It is possible to refer to this index of attractiveness as

utility, that the decision maker tries to maximize through the choice. The consideration of a single index is based on the notion of compensatory offsets, used by a decision maker comparing different attributes (the three previous decision rules are non-compensatory).

There is a distinction to consider, ordinal and cardinal utilities. The ranking of alternative preferences is expressed mathematically by the ordinal utility which is unique only up to an order that preserves the transformation. The numerical comparison of utility values has no meaning except for the relation greater than, less than and equal to. Different is a cardinal utility that implies some uniqueness of its numerical value and is for this reason more restrictive than the ordinal one.

3.5 The Random Utility Theory

It is assumed that the individuals choose the alternative with the maximum utility, but the analyst is not able to know with certainty the value of it and thus they will be treated as random variables. It is possible forecast the choice of an individual *n* among a finite discrete set of alternatives C_n as follow:

$$P(i|C_n) = \Pr\left[U_{in} \geq U_{jn}, all\ j \in C_n\right] \quad [1]$$

In a general Multinomial Logit model, the basic assumption is that each individual associate a quantity "utility" to each alternative inside the alternatives set, selecting the alternative with the highest utility. From this point of view, it is possible to represent the choice probability of alternative *i* with the probability that the utility of alternative *i*, U_{in}, is greater than or equal to the utilities of all other alternatives in the choice set.

(Manski, 1977) identified four distinct sources of randomness: unobserved attributes, taste variations, measurement errors and instrumental (or proxy) variables.

In general, it is possible write the random utility of an alternative as a sum of observable (or systematic) and unobservable components of the total utilities as follows:

$$U_{ni} = V_{in} + \varepsilon_{in} \quad [2]$$

where $V_{in} \in \mathbb{R}$ is the deterministic or *systematic* (*or representative*) components of the utility and ε_{in} is a random term called *disturbances* (or random components).

3.6 The Logit Model

The general form of a logit model to express the probability for choosing an alternative than the other is represented by the formula:

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}} \quad [3]$$

but it defines a proper probability function only if:

$0 \leq P_n(i) \leq 1$ for $i \in C_n$

and

$\sum_{i \in C_n} P_n(i)=1$

3.6.1 The hypothesis on the error term

The assumptions about the *systematic* part of the Utility function expressed as $U_{in} = V_{in} + \varepsilon_{in}$ *for* $i \in C_n$ are that ε are: independently distributed, identically distributed and Gumbel-distributed with a scale parameter $\mu > 0$, in this case the choice probability formula is:

$$P_n(i) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}} \quad [4]$$

3.6.2 Independence from Irrelevant Alternatives (IIA) property

For any two alternatives *i* and *k*, the ratio of the logit probabilities is (Train, 2009):

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{in}} / \sum_j e^{V_{jn}}}{e^{V_{kn}} / \sum_j e^{V_{jn}}} = \frac{e^{V_{in}}}{e^{V_{kn}}} = e^{V_{in} - V_{kn}} \quad [5]$$

This ration does not depend on alternatives different from *i* and *k,* this represent the relative odds of choosing *i* than *k* is the same independently by the other alternatives available. When this ratio has this feature, it is said that it is independent from "irrelevant" alternative and the Logit model have this "independence from irrelevant alternative" (IIA) property.

3.6.3 The Ordered Logit Model for Injury Severity

Ordered Logit model is widely used to analyze ranking responses from a survey. Injury Severity based on the KABCO scale (Table 1) is defined on five levels; given the ordinal nature of the independent variable, an *Ordered Logit* structure is proposed for the analysis. The Ordered Logit framework assumes a single utility function (U) mapped into one of J ordered outcomes by J-1 threshold parameters. Let q (1, 2, . . . Q) be the index for the observations, the utility function is specified as the sum of a linear in parameters deterministic component $(\beta_q{}^* x_q)$ and a random component (ε_q) or unobserved factors that influence the ordered outcome. Basically, the utility is a latent (or unobserved) variable that depends on some observable variables x_q and some coefficients (β_q) to be estimated through inference on a sample of observations.

$$U = \beta_q{}^* x_q + \varepsilon_q \qquad [6]$$

The latent utility function is mapped into ordinal outcomes by threshold parameters τ. This cutoff must be considered to understand in which category the evaluation is made, these are labelled τ1, τ2, τ3 and τ4 respectively, so that it is possible to represent the decision regarding the IS in the following way:

Choice "1" if $U < \tau_1$
Choice "2" if $\tau_1 < U < \tau_2$
Choice "3" if $\tau_2 < U < \tau_3$
Choice "4" if $\tau_3 < U < \tau_4$
Choice "5" if $\tau_4 < U$

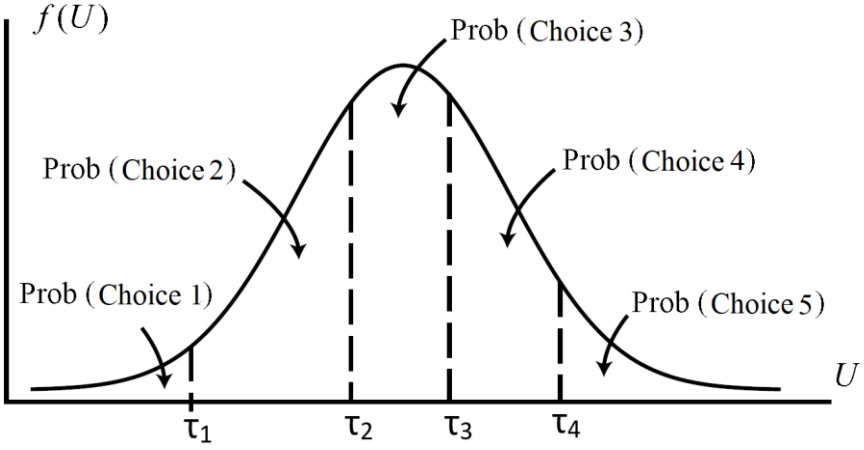


Figure 7. Example of a distribution of general choice

Figure 7 illustrates in a clear way that depending on the value assumed by U it is possible to determine the probability that the choice follows in a certain interval. For instance, the probability that a policeman reports the Choice (or Rank) 1 in the KABCO scale which corresponds to "No injury" is the probability that U is less than τ_1, which is the area subtended by the curve in the distribution left tail. The probability that the policeman chooses the Choice 2 "Possible injury" is the probability that U is above τ_1 and below τ_2; this probability is the area between τ_1 and τ_2 (Train, Discrete choice Methods With Simulation, 2009). The remaining ranks in the KABCO scale can be derived following the same reasoning.

3.7 The estimation with Biogeme

BIOGEME was adopted for model estimation (Bierlaire, 2003), this is an open source Python package made available by Michael Bierlaire, it uses **Maximum Likelihood Methods** to estimate Generalized Extreme Values (GEV) models' parameters. The PandasBiogeme library is used for data management and cleaning, it relies on the package Python data analysis library called Pandas (Bierlaire, 2018).

The GEV family definition provides an interesting theoretical framework to develop closed-form **random utility models**. This models family consist of a large type of model that include the Multinomial Logit, the Nested Logit, the Cross-Nested Logit and the Ordered Logit models (Ordered Generalized Extreme Value) (Small, 1987). The theory relative to the GEV models was developed by McFadden (McFadden, 1978).

This formulation shows that the probability of choosing alternative "i" within the choice set C for a given choice maker is the following:

$$P(i|C) = \frac{y_i \frac{\partial G}{\partial y_i}(y_i, \dots, y_j)}{\mu G(y_i, \dots, y_j)} \qquad [7]$$

Where j is the number of available alternatives, $y_i = e^{V_i}$, V_i is the deterministic part of the utility function associated with alternative i, and G is a μ-GEV function, this is a differentiable function defined on $\mathbb{R}_+^J$ with the following properties:

1. $G(y) \geq 0$ for all $y \in \mathbb{R}_+^J$
2. G is homogeneous of degree $\mu > 0$, that is $G(\alpha y) = \alpha^{\mu} G(y)$, for $\alpha > 0$
3. $\lim_{y_i \to \infty} G(y_i, \dots, y_j) = +\infty$, for each i = 1,…,J

4. The mixed partial derivatives of G exist and are continuous. Furthermore, the kth partial derivative with respect to k distinct y_i is non-negative if k is odd and non-positive if k is even that is, for any distinct indices $i_1, \ldots, i_J \in J = \{1, \ldots, J\}$, we have:

$$(-1)^k \frac{\partial^k G}{\partial x_{i_1} \ldots \partial x_{i_{1k}}}(x) \leq 0, \forall x \in \mathbb{R}_+^J \quad [7]$$

It is also required that $G(x) \neq 0$.

3.7.1 Maximum likelihood estimation

Maximum likelihood (Bierlaire, 2003) is commonly used for the estimation of unknown parameters. A k observation is characterized by a set of value belonging to the set of attributes x_n, indicate as x_n^k and an observed choice. The probability for the model to reproduce the observed choice is $P^k(i|C) = P^k(\beta, \gamma)$, where β is the unknown parameters associated with the utility function and γ the unknown parameters associated with a specific GEV model. If a sample is composed by K observation, the probability of the model to reproduce the whole sample is called likelihood, and it is given by:

$$\mathcal{L}^*(\beta, \gamma) = \prod_{k=1}^{K} P^k(\beta, \gamma) \quad [9]$$

The maximum likelihood estimators $\hat{\beta}$ and $\hat{\gamma}$ are given by the formulation

$$(\hat{\beta}, \hat{\gamma}) = argmax_{\beta,\gamma}\, \mathcal{L}(\beta, \gamma) \quad [10]$$

where

$$\mathcal{L}(\beta, \gamma) = \ln \mathcal{L}^*(\beta, \gamma) = \sum_{k=1}^{K} \ln P^k(\beta, \gamma) \quad [11]$$

$\mathcal{L}$ is the log-likelihood function. In some cases, a weight could be added to underline the relative importance of a group in the population, obtaining the following formula:

$$\mathcal{L}(\beta, \gamma) = \sum_{k=1}^{K} \omega_k \ln P^k(\beta, \gamma) \quad [12]$$

4 RESULTS AND DISCUSSION

In this Chapter the outcome from model estimation are presented and discussed. Several indicators are used to give a behavioral interpretation of the model parameters obtained and to assess their statistical significance:

- **The sign of the parameters**; model parameters can be positive or negative. A parameter with a positive sign means that an increase in the corresponding variable is expected to produce a higher Injury Severity level; the opposite is true for negative coefficients. The sign obtained are usually compared to the general expectations or to results available in the literature.
- The asymptotic t-test is mainly used to check if a given parameter in the model differs from a known constant, often considered equal to zero. It is used in the same way as the t-test in linear regression, but for the case of non-linear models this test is valid only asymptotically, it is valid only for large samples. Critical values for the test statistics are percentiles of a standard normal distribution, which for two-tailed tests with a significance level (α) of 0.05 is ± 1.96.

  If the null hypothesis for a parameter β_n is to be equal to zero (no contribute to the IS evaluation) it is possible to consider a normal standardize distribution N(0,1) with the mine equal to zero (μ_0).

  Extracting a sample from the population and considering the distribution probability of the parameter, it is easy to calculate the sample variance as follow

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad [13]$$

  where: s^2 is the sample variance, $\bar{x}$ is the sample mean, n is the numerosity of the sample and x_i is the generic value of a "i" component of the sample.

  T test is given by the formulation:

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad [14]$$

  Here, μ_0 represent the mean condition expressed by the null hypothesis (**Figure 8)**.

- **p-value**, which indicate if the variable is significant or not (if it has an impact on the output of the model which belong to); when the significance level is set at α, the p-value must be lower than α to indicate the significance of that variable (i.e., the empirical evidence is strongly opposed to the null hypothesis which therefore must be rejected. In this case it is said that the observed data are statistically significant). A value lower than α = 0.05 is required for the validation of the estimated parameter (**Figure 8)**:

$$p = P(T) \qquad \text{it is a one-sided test}$$
$$p = P(T) \cdot 2 \qquad \text{it is a two-sided test}$$

The number of tales depend on the alternative hypothesis (hypothesis alternative to the null one). For instance:

HP_a> or < than a certain value, it is considered a one-sided test

HP_a= or ≠ to a certain value, it is considered a two-sided test

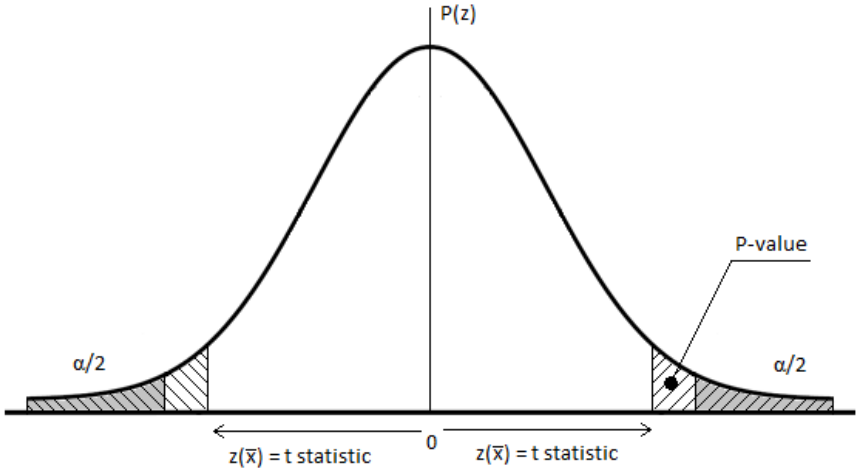


Figure 8. Normal standardized Probability distribution with P-value and T-statistic visualization

From the results reported in **Table 7** and **Table 8** (see **Appendix 4** for the model estimation with the software), it is possible to understand how the factors considered contribute to explain the Injury Severity (IS) of accidents that happened respectively in 2016 and 2017. The results should be interpreted with reference to the base value of the variables as defined in **Table 5**. It worth highlighting that a positive sign for the variable of interest means that the variable is prone to increase the IS level and a negative sign tend to decrease it.

In addition to the model parameters, the ordered logit structural parameters should be estimated, those are: $\tau_1, \Delta_2, \Delta_3, \Delta_4$; the other τ-parameters can be calculated considering the following relations:

$\tau_2 = \tau_1 + \Delta_2$
$\tau_3 = \tau_2 + \Delta_3$
$\tau_4 = \tau_3 + \Delta_4$

$\Delta_2, \Delta_3, \Delta_4$ are respectively the intervals between τ_1 - τ_2, τ_2 - τ_3 and τ_3 - τ_4 (see **Figure 7**).

With reference to the results obtained from the 2016 data and presented in **Table 7** the following comments can be made.

Table 7. Model outcome 2016 containing the value of parameter estimated for each variable, standard error, t-test and p-value. Coefficients not significant are highlights in grey. (See Table 6 for the variable meaning).

2016	Value	Std err	t-test	p-value
B_SEX2	0.422697	0.01013	41.744887	0.00E+00
B_AGECLASS1	0.230257	0.02225	10.350109	0.00E+00
B_AGECLASS2	0.323123	0.02188	14.769922	0.00E+00
B_AGECLASS3	0.457586	0.02322	19.708609	0.00E+00
B_AGECLASS4	0.525231	0.02608	20.138525	0.00E+00
B_AGECLASS5	0.540735	0.03486	15.510019	0.00E+00
B_DRUGEFFECT1	0.408964	0.05194	7.873664	3.55E-15
B_ALCHOLEFFECT1	0.008278	0.02339	0.353934	7.23E-01
B_MEDICINEEFFECT1	0.046277	0.08385	0.551883	5.81E-01
B_DAMEFFECT1	0.351347	0.07934	4.428383	9.49E-06
B_EJTRAP1	3.377048	0.03492	96.704608	0.00E+00
B_SEATBELT1	-0.02716	0.01358	-1.99993	4.55E-02
B_VEHBODY2	-0.07511	0.01555	-4.829129	1.37E-06
B_VEHBODY3	-0.24542	0.0225	-10.907851	0.00E+00
B_VEHBODY4	-0.15954	0.02748	-5.804587	6.45E-09
B_VEHBODY5	-0.3882	0.05824	-6.665816	2.63E-11
B_PERSONTYPE1	0.429113	0.01322	32.462678	0.00E+00
B_PERSONTYPE2	2.862117	0.02938	97.43277	0.00E+00
B_ROLLEDOVER1	1.029977	0.0319	32.290418	0.00E+00
B_ROADTYPE2	0.141494	0.01648	8.586301	0.00E+00
B_ROADTYPE3	0.183323	0.01119	16.383811	0.00E+00
B_SURFCOND2	0.010877	0.0134	0.811919	4.17E-01
B_SURFCOND3	0.224228	0.03295	6.805271	1.01E-11
B_COLLTYPE2	-0.09949	0.01187	-8.383909	0.00E+00
B_COLLTYPE3	-0.03476	0.04113	-0.845145	3.98E-01
B_COLLTYPE4	-0.7325	0.02569	-28.512862	0.00E+00
B_COLLTYPE5	-0.30669	0.04876	-6.289984	3.17E-10
B_INTERSECTIONTYPE1	0.192669	0.0133	14.491466	0.00E+00
B_LIGHTCOND2	0.106237	0.02045	5.195874	2.04E-07
B_LIGHTCOND3	0.101322	0.03419	2.963514	3.04E-03
B_LIGHTCOND4	-0.08766	0.03074	-2.852003	4.34E-03
B_FIXOBJECT1	0.552539	0.01334	41.42877	0.00E+00
B_TRAFFICCONTROL2	0.129948	0.01457	8.920894	0.00E+00
B_TRAFFICCONTROL3	0.122646	0.02044	5.999987	1.97E-09
B_TRAFFICCONTROL4	0.069645	0.05082	1.370404	1.71E-01
delta2	0.804405	0.00493	163.29588	0.00E+00
delta3	2.340841	0.01605	145.82428	0.00E+00
delta4	2.18788	0.04138	52.874859	0.00E+00
tau1	2.310857	0.02439	94.743811	0.00E+00

Five AGE related parameters are estimated, one for each of the five intervals over which the AGE variable has been discretized (AGE 0-15 being the base). It is interesting to note that all the coefficients estimated are positive and their value increases with the individuals' age. This

means that the likelihood of higher IS increases with age, which is consistent with the general expectations. Elderly people tend to have higher response time in the case of danger, and they tend to be more fragile.

The parameter about the variable *Alcohol effect* results to be positive as expected but surprisingly it is not statistically significant. However, *Medicine effect*, *Drug/Alcohol/Medicine combined effect* and *Drug effect*, which have been hierarchically ordered from the least to the most dangerous for the IS level evaluation, are all positive and statistically significant. In particular, drugs and their combined effect with medicaments highly affect the severity of an accident.

All the collision type variables are likely to decrease the IS compared when compared to the *Base* (*Head On collision*), which represents the worst possible condition. The second worst collision type is *Angle meets collisions* (which is not significant), followed in order by *Same direction collision (rear end, rear end right turn, rear end left turn), Opposite direction collision* and *Same direction sideswipe*.

The light condition variables (*Dark Lights On* and *Dark No Lights*), having a positive sign, are likely to increase the IS compared to the *Base* condition (*Daylight*). The condition *Dark no Lights* is the most severe condition while *Dawn* condition is the less severe. *Dusk* has a negative sign meaning that possibly Dusk conditions ensure good enough visibility to the driver and that it is not a critical factor when evaluation IS.

The road type variables considered, having all a positive sign, are likely to increase the IS compared to the *Base* (*one/two way undivided*); the following two conditions: *Two way divided unprotected* and *Two way divided positive median barrier* are more dangerous. This could be explained considering that for the second type of road the speed limit could be greater than for the first one and therefore accidents are of a higher severity.

Surface condition parameters are compared to the base value of *Dry* condition. The *Wet* case has a positive sign, but it is not significant. Regarding the *Snow* and *Ice* conditions, those road conditions definitively increase the likelihood of accidents of high severity.

Concerning the traffic control variable, the BASE has been set to *No controls*. The conditions considered from the most to the least dangerous are as follows: *Traffic sign*, *Stop sign* and *Yield sign*. Their coefficients are all positive, the first two have similar values, while the third one is about half (and not very significant), which seems to attest that accidents at a *Yield sign* are less severe that those that happens at Traffic lights and Stop signs.

Another interesting aspect to analyze is the type of vehicle involved in the crash. The base vehicle is *Car*; estimation results for vehicle types are all negative attesting that heavier vehicles tend to lower the IS score. This could be explained considering that probably inside a *Pickup Truck,*

Van, Light Truck or *SUV* people are more protected than in a common car and consequently that the crash results into no injuries or less severe consequences.

Consistently with intuition, the parameter regarding the situation in which a person is E*jected/Trapped* in a crash at an intersection, the presence of *Fix Object*, and rolled over condition, all increase the probability to have an accident with a higher IS score.

The parameter related to the use of the *Seatbelt* is negative, meaning that the use of the seatbelt (working in the properly way) safes life.

Finally, our results attest that *Females* are more likely to be involved in more severe accidents.

Results obtained for 2016 and 2017 are very similar, except for a few points that are summarized below:

- *Dawn light* and *Dusk light* are not significant;
- *Medicine effect* turns out to be significant;
- *Dusk light* is positive but not significant;
- *Wet surface* is negative but barely significant;
- *Snow/ice on the surface* is negative.

Table 8 shows the results of model estimation obtained on the 2017 dataset.

Table 8. Model outcome 2017 containing the value of parameter estimated for each variable, standard error, t-test and p-value. Coefficients not significant are highlights in grey. See (See Table 6 for the variable meaning).

2017	Value	Std err	t-test	p-value
B_SEX2	0.412602	0.010118	40.77721	0.00E+00
B_AGECLASS1	0.219371	0.022037	9.954554	0.00E+00
B_AGECLASS2	0.330453	0.021649	15.26436	0.00E+00
B_AGECLASS3	0.490517	0.022979	21.34612	0.00E+00
B_AGECLASS4	0.570815	0.025637	22.26517	0.00E+00
B_AGECLASS5	0.47707	0.034611	13.78387	0.00E+00
B_DRUGEFFECT1	0.428037	0.049851	8.586275	0.00E+00
B_ALCHOLEFFECT1	0.011093	0.023211	0.477925	6.33E-01
B_MEDICINEEFFECT1	0.223061	0.097051	2.298385	2.15E-02
B_DAMEFFECT1	0.382267	0.083037	4.603556	4.15E-06
B_EJTRAP1	3.692685	0.036455	101.2957	0.00E+00
B_SEATBELT1	-0.04821	0.013227	-3.645	2.67E-04
B_VEHBODY2	-0.07548	0.01585	-4.7622	1.91E-06
B_VEHBODY3	-0.27816	0.02397	-11.6046	0.00E+00
B_VEHBODY4	-0.06201	0.026617	-2.32979	1.98E-02
B_VEHBODY5	-0.62539	0.070978	-8.811	0.00E+00
B_PERSONTYPE1	0.328751	0.013188	24.92803	0.00E+00
B_PERSONTYPE2	2.990793	0.029123	102.6955	0.00E+00
B_ROLLEDOVER1	1.080302	0.031508	34.28662	0.00E+00
B_ROADTYPE2	0.168738	0.016216	10.40588	0.00E+00
B_ROADTYPE3	0.160428	0.011299	14.19787	0.00E+00
B_SURFCOND2	-0.02321	0.013313	-1.74371	8.12E-02
B_SURFCOND3	-0.29086	0.048147	-6.04112	1.53E-09
B_COLLTYPE2	-0.18263	0.011982	-15.2426	0.00E+00
B_COLLTYPE3	-0.05184	0.041471	-1.24995	2.11E-01
B_COLLTYPE4	-0.73751	0.024931	-29.5824	0.00E+00
B_COLLTYPE5	-0.26983	0.047561	-5.67332	1.40E-08
B_INTERSECTIONTYPE1	0.18861	0.01331	14.17037	0.00E+00
B_LIGHTCOND2	0.126726	0.020679	6.128166	8.89E-10
B_LIGHTCOND3	0.019313	0.034843	0.554281	5.79E-01
B_LIGHTCOND4	0.035056	0.031209	1.123261	2.61E-01
B_FIXOBJECT1	0.558858	0.013266	42.12609	0.00E+00
B_TRAFFICCONTROL2	0.17759	0.014627	12.14127	0.00E+00
B_TRAFFICCONTROL3	0.183743	0.02013	9.127709	0.00E+00
B_TRAFFICCONTROL4	0.002663	0.051563	0.051649	9.59E-01
delta2	1.177504	0.006398	184.047	0.00E+00
delta3	2.010365	0.015839	126.9286	0.00E+00
delta4	2.217881	0.040871	54.26491	0.00E+00
tau1	2.266865	0.024138	93.91134	0.00E+00

4.1 Validation

Model validation is necessary to test the ability of the model to correctly predict Injury Severity scores when applied to a new set of data that has not been used for model estimation. This procedure is known as out-of-sample validation. The models proposed are probabilistic and therefore they are not able to reproduce with certainty the output of the model in a deterministic way, instead they provide the probability that given certain circumstances a crash has an IS on the scale from 1 to 5. In order to validate the 2016 and 2017 models, it is necessary to divide the whole database in two sets of data, one for the parameters' estimation and the other for model application. Sample enumeration is used to aggregate single accident probability to be on a certain scale and to simulate "Market Share" for each IS level. The difference between the observed IS scores and the simulated "Market Shares" is an indicator of the ability of the model to predict accident severity. The estimation has been carried out on about 80% of the original database and the remaining 20% has been set aside for simulation purpose.

If the model is properly working the resulting "Market Shares" should be comparable to the IS in the original database. As show in **Table 9** the model is able to produce aggregate forecasts that are very close to the IS in the original database for both 2016 and 2017 (see **Appendix 5** for the Market Share evaluation with the software).

Table 9. Out-of-Sample Validation 2016-2017, comparison between the original market-share and the simulated one.

2016			
Original Database		**Parameters from the 80% applied to the 20%**	
IS	Market Share	Market Share	Confidence interval
1	79.3%	79.2%	[78.5%,79.9%]
2	9.3%	9.3%	[9.0%,9.7%]
3	9.8%	9.8%	[9.4%,10.2%]
4	1.4%	1.5%	[1.4%,1.5%]
5	0.20%	0.2%	[0.2%,0.2%]

2017			
Original Database		**Parameters from the 80% applied to the 20%**	
IS	Market Share	Market Share	Confidence interval
1	79.4%	79.5%	[78.8%,80.2%]
2	12.2%	12.0%	[11.6%,12.4%]
3	6.90%	6.8%	[6.6%,7.1%]
4	1.40%	1.4%	[1.4%,1.5%]
5	0.2%	0.2%	[0.2%,0.2%]

4.2 Forecast of IS in different Hypothetical scenarios

A sensitivity analysis was carried out in order to test how model predictions change after certain variables are varied based on hypothetical scenario conditions; the original ISs are then compared to model predictions.

The procedure consists in the following four steps: 1) select the attribute of interest; 2) apply the model; 3) calculate the difference between predicted IS (F) and observed IS (O); calculate the percentage increase between (F) and (O) . We test the effect of one variable at a time, although it would be interesting to test the combined effect of several variables. Results are shown in Table 10.

The variation of a factor (e.g. *Age*) produces an uneven shift of the effects in the IS levels, which means that when a factor is altered a different scenario of consequences of the accident is produced in different way across the IS level.

The most interesting scenarios to be analyzed are the ones that most differ from the unperturbed situation e.g. all pedestrians, all ejected/trapped, and all rolled over.

- **All pedestrian**: the scenario in which all the individuals are pedestrian is not so realistic also because no other condition was not changed, however it is interesting to note how high the pedestrian's vulnerability is compared to drivers or vehicle occupants (see **Scenario10**).
- **All ejected/trapped**: the evidence concerns a general increase of the IS level particularly about the *non-incapacitating* and *incapacitating* injury (level *3* and *4*).
- **All rolled over**: as for the previous scenario but affecting more the *possible* and *non-incapacitating* injury (level *2* and *3*).

The list of all the forecasted scenario is provided in the below tables. The outcomes showed are about the “Market Share”, and relative confidence interval, of IS level in that specific scenario; the difference in percentage Δ(F-O) and the net percentage increase (F / O) – 1.

Brief considerations are provided for each group of scenarios belonging to the same variable. The considerations are focused more on the level 2,3,4 and 5 neglecting the 1th level except for few cases.

Table 10. Scenarios 1, 2 and 3: growth of average age, respectively +5, +10 and +15 years.

Scenario 1: all people are 5 years older								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	78.80%	[78.2%,79.5%]	-0.50%	-0.63%	78.90%	[78.3%,79.6%]	-0.50%	-0.63%
2	9.50%	[9.2%,9.8%]	0.20%	2.15%	12.30%	[11.9%,12.7%]	0.20%	1.65%
3	10.00%	[9.6%,10.3%]	0.20%	2.04%	7.00%	[6.8%,7.3%]	0.10%	1.45%
4	1.50%	[1.4%,1.5%]	0.10%	7.14%	1.50%	[1.4%,1.6%]	0.10%	7.14%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 2: all people are 10 years older								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	78.40%	[77.7%,79.1%]	-0.90%	-1.13%	78.40%	[77.7%,79.1%]	-1.00%	-1.26%
2	9.70%	[9.4%,10.0%]	0.40%	4.30%	12.60%	[12.2%,13.0%]	0.50%	4.13%
3	10.20%	[9.9%,10.6%]	0.40%	4.08%	7.20%	[7.0%,7.5%]	0.30%	4.35%
4	1.50%	[1.4%,1.6%]	0.10%	7.14%	1.50%	[1.5%,1.6%]	0.10%	7.14%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.3%]	0.00%	0.00%
Scenario 3: all people are 15 years older								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	78.00%	[77.3%,78.7%]	-1.30%	-1.64%	78.00%	[77.3%,78.7%]	-1.40%	-1.76%
2	9.80%	[9.5%,10.1%]	0.50%	5.38%	12.80%	[12.4%,13.2%]	0.70%	5.79%
3	10.40%	[10.0%,10.8%]	0.60%	6.12%	7.40%	[7.1%,7.7%]	0.50%	7.25%
4	1.50%	[1.4%,1.6%]	0.10%	7.14%	1.60%	[1.5%,1.6%]	0.20%	14.29%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.3%]	0.00%	0.00%

Table 10: in 2016, only for the worst IS level (5th level), the growth of the average age has no effect. In both years the 2th and 3th level uniformly increased their percentage [*(F/O)-1*], and 4th level increasing more than the others.

Table 11. Scenario 4: drug usage.

Scenario 4: all people are drugged during the crash								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	72.70%	[70.8%,74.4%]	-6.60%	-8.32%	72.50%	[70.8%,74.0%]	-6.90%	-8.69%
2	11.80%	[11.1%,12.6%]	2.50%	26.88%	15.70%	[14.8%,16.6%]	3.60%	29.75%
3	13.20%	[12.2%,14.2%]	3.40%	34.69%	9.40%	[8.8%,10.2%]	2.50%	36.23%
4	2.00%	[1.8%,2.2%]	0.60%	42.86%	2.10%	[1.9%,2.2%]	0.70%	50.00%
5	0.30%	[0.3%,0.4%]	0.10%	50.00%	0.30%	[0.3%,0.4%]	0.10%	50.00%

Table 11: the use of drugs shows an increase, in percentage difference *[Δ(F - O)],* of the 2th (*possible injury*) and 3th (*non-incapacitating injury*) IS level. The lasts two more severe levels present the big growth in percentage around 50% in both years.

Table 12. Scenario 5: alcohol usage.

Scenario 5: all people are drunk during the crash								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	79.20%	[78.4%,80.0%]	-0.10%	-0.13%	79.20%	[78.4%,80.0%]	-0.20%	-0.25%
2	9.40%	[9.0%,9.7%]	0.10%	1.08%	12.10%	[11.7%,12.6%]	0.00%	0.00%
3	9.80%	[9.4%,10.2%]	0.00%	0.00%	6.90%	[6.6%,7.3%]	0.00%	0.00%
4	1.40%	[1.3%,1.5%]	0.00%	0.00%	1.50%	[1.4%,1.5%]	0.10%	7.14%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%

Table 12: From the model outcome the variable related to the *alcohol* usage was not significant. The scenario was forecasted anyway to understand how it could be affected by a not significant variable.
For the 2016 the percentage remain quite undisturbed and there is a decrease in the 2th IS level. In 2017 instead, the growth in percentage is referred to the incapacitating injuries (4th level).

Table 13. Scenario 6: medicine usage.

Scenario 6: all people use medicine during the crash								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	78.60%	[76.0%,80.4%]	-0.70%	-0.88%	76.00%	[72.9%,78.5%]	-3.40%	-4.28%
2	9.60%	[8.8%,10.7%]	0.30%	3.23%	13.90%	[12.4%,15.6%]	1.80%	14.88%
3	10.10%	[9.2%,11.5%]	0.30%	3.06%	8.10%	[7.2%,9.4%]	1.20%	17.39%
4	1.50%	[1.3%,1.7%]	0.10%	7.14%	1.70%	[1.5%,2.0%]	0.30%	21.43%
5	0.20%	[0.2%,0.3%]	0.00%	0.00%	0.30%	[0.2%,0.3%]	0.10%	50.00%

Table 13: the coefficients differences between 2016 and 2017 are reflected also in this forecasted scenario.

In 2016 the coefficient *B_MEDICINEEFFECT1= 0.046277* (close to zero) and the percentage remain approximately the same with the exceptions of a slight increase in levels 2, 3 and 4, instead in the 2017 where *B_MEDICINEEFFECT1= 0.223061* a different pattern in percentage is shown, with an emphasis for the increase in percentage of all the levels in 2017 most of all the 5th level (*fatal injury*).

This difference in prediction (and in the parameters estimated) suggest a more serious situation regarding the drive under influence of drugs in 2017. The growth in percentage in 2017 is considerably greater than in 2016.

Table 14. Scenario 7: drug/medicine/alcohol usage at the same time.

Scenario 7: all people drive under combined effect during the crash								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	73.70%	[71.6%,76.2%]	-5.60%	-7.06%	73.30%	[70.9%,75.7%]	-6.10%	-7.68%
2	11.50%	[10.4%,12.3%]	2.20%	23.66%	15.30%	[13.9%,16.5%]	3.20%	26.45%
3	12.60%	[11.3%,13.8%]	2.80%	28.57%	9.10%	[8.2%,10.1%]	2.20%	31.88%
4	1.90%	[1.7%,2.1%]	0.50%	35.71%	2.00%	[1.8%,2.2%]	0.60%	42.86%
5	0.30%	[0.2%,0.3%]	0.10%	50.00%	0.30%	[0.3%,0.4%]	0.10%	50.00%

Table 14: in both years the percentage trend is comparable. The 2016 presents a great difference in percentage in the 3th IS level compared to 2017 where the greater difference in percentage is in the 2th IS level.

About the growth in percentage all the levels are interested, particularly the 5th in both years.

Table 15. Scenarios 8, 9 and 10: people person type among drivers, occupants and pedestrians.

Scenario 8: all people are drivers								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	81.80%	[81.2%,82.4%]	2.50%	3.15%	81.70%	[81.1%,82.3%]	2.30%	2.90%
2	8.60%	[8.3%,8.9%]	-0.70%	-7.53%	11.20%	[10.8%,11.5%]	-0.90%	-7.44%
3	8.30%	[8.0%,8.6%]	-1.50%	-15.31%	5.80%	[5.6%,6.0%]	-1.10%	-15.94%
4	1.10%	[1.1%,1.2%]	-0.30%	-21.43%	1.10%	[1.1%,1.2%]	-0.30%	-21.43%
5	0.20%	[0.1%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 9: all people are occupants								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	75.20%	[74.3%,75.9%]	-4.10%	-5.17%	76.70%	[75.9%,77.4%]	-2.70%	-3.40%
2	11.30%	[11.0%,11.6%]	2.00%	21.51%	14.00%	[13.5%,14.4%]	1.90%	15.70%
3	11.70%	[11.3%,12.1%]	1.90%	19.39%	7.60%	[7.3%,7.9%]	0.70%	10.14%
4	1.60%	[1.5%,1.7%]	0.20%	14.29%	1.50%	[1.4%,1.6%]	0.10%	7.14%
5	0.20%	[0.2%,0.3%]	0.00%	0.00%	0.20%	[0.2%,0.3%]	0.00%	0.00%
Scenario 10: all people are pedestrians								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	23.20%	[22.1%,24.4%]	-56.10%	-70.74%	21.00%	[20.1%,22.0%]	-58.40%	-73.55%
2	16.30%	[15.8%,16.7%]	7.00%	75.27%	24.00%	[23.4%,24.5%]	11.90%	98.35%
3	45.90%	[45.0%,46.8%]	36.10%	368.37%	39.20%	[38.4%,40.1%]	32.30%	468.12%
4	12.30%	[11.6%,13.0%]	10.90%	778.57%	13.20%	[12.5%,13.9%]	11.80%	842.86%
5	2.30%	[2.2%,2.6%]	2.10%	1050.00%	2.50%	[2.3%,2.7%]	2.30%	1150.00%

Table 15: these scenarios have the role of comparing the vulnerability of the pedestrian respect the drivers and occupants condition during a road crash. Pedestrians are more exposed to be more severely injured, in fact driver and the occupants are more protected inside the vehicle especially if it is a SUV or a Van. In the *Scenario 10* the great difference in percentage with the *Scenario 0* concern the 1st (*no injury*) and the 3th IS level (*non-incapacitating injury*) in both years.

More relevant is the growth in percentage for the 5th level, more of the 1000%. In this scenario, the probability by a pedestrian to have a fatal injury is 10 times higher than in the original one.

Table 16. Scenarios 11, 12: seat-belts usage.

Scenario 11: nobody uses seat-belt								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	79.00%	[78.3%,79.7%]	-0.30%	-0.38%	78.80%	[78.2%,79.5%]	-0.60%	-0.76%
2	9.50%	[9.1%,9.8%]	0.20%	2.15%	12.40%	[12.0%,12.8%]	0.30%	2.48%
3	9.90%	[9.5%,10.3%]	0.10%	1.02%	7.10%	[6.8%,7.4%]	0.20%	2.90%
4	1.40%	[1.4%,1.5%]	0.00%	0.00%	1.50%	[1.4%,1.6%]	0.10%	7.14%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 12: all people use seat-belt								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	79.40%	[78.7%,80.0%]	0.10%	0.13%	79.60%	[79.0%,80.2%]	0.20%	0.25%
2	9.30%	[9.0%,9.6%]	0.00%	0.00%	12.00%	[11.6%,12.3%]	-0.10%	-0.83%
3	9.70%	[9.4%,10.1%]	-0.10%	-1.02%	6.80%	[6.6%,7.1%]	-0.10%	-1.45%
4	1.40%	[1.3%,1.5%]	0.00%	0.00%	1.40%	[1.4%,1.5%]	0.00%	0.00%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%

Table 16: if nobody used seatbelts (*Scenario 11*) the IS level generally tend to grow or stay stable.

The *Scenario 12* presents similar features in 2016 only in the 3th IS level there is a small decrease in percentage. In the 2017 instead level 2 and 3 present decreases in percentage. This

scenario let understand that seat-belt is more effective for low IS level and less helpful for the more severe crash.

Table 17. Scenarios 13, 14: ejection/entrapment dynamics.

Scenario 13: no people ejected/trapped								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	80.20%	[79.5%,80.8%]	0.90%	1.13%	80.30%	[79.6%,80.9%]	0.90%	1.13%
2	9.30%	[9.0%,9.6%]	0.00%	0.00%	12.10%	[11.7%,12.5%]	0.00%	0.00%
3	9.20%	[8.9%,9.6%]	-0.60%	-6.12%	6.40%	[6.2%,6.7%]	-0.50%	-7.25%
4	1.10%	[1.0%,1.2%]	-0.30%	-21.43%	1.10%	[1.0%,1.2%]	-0.30%	-21.43%
5	0.10%	[0.1%,0.2%]	-0.10%	-50.00%	0.10%	[0.1%,0.2%]	-0.10%	-50.00%
Scenario 14: all people ejected/trapped								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	14.30%	[13.3%,15.2%]	-65.00%	-81.97%	11.00%	[10.3%,11.8%]	-68.40%	-86.15%
2	12.30%	[11.7%,12.8%]	3.00%	32.26%	16.80%	[16.1%,17.6%]	4.70%	38.84%
3	50.00%	[49.1%,50.6%]	40.20%	410.20%	43.90%	[43.3%,44.6%]	37.00%	536.23%
4	19.60%	[18.6%,20.9%]	18.20%	1300.00%	23.40%	[22.2%,24.5%]	22.00%	1571.43%
5	3.80%	[3.5%,4.2%]	3.60%	1800.00%	4.80%	[4.4%,5.2%]	4.60%	2300.00%

Table 17: similar trend is presented in both years. In the *Scenario 13* the IS level that difference in percentage is more than others is the 3th followed by the 4th and 5th.

On the other hand, the decrease in percentage grow with the IS level growth. In the complementary scenario, similar trend is followed in the opposite side.

The *Scenario 14* present one of the most growth in percentage compared to the original scenario and the highest is reached, in 2017, with the 5th level with a probability up to more than 20 times bigger than the *Scenario 0*.

Table 18. Scenarios 15, 16: fix objects involvement.

Scenario 15: no fix objects involved in crashes								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	80.70%	[80.1%,81.4%]	1.40%	1.77%	80.90%	[80.3%,81.5%]	1.50%	1.89%
2	8.80%	[8.5%,9.1%]	-0.50%	-5.38%	11.30%	[10.9%,11.6%]	-0.80%	-6.61%
3	9.00%	[8.7%,9.3%]	-0.80%	-8.16%	6.30%	[6.1%,6.6%]	-0.60%	-8.70%
4	1.30%	[1.2%,1.3%]	-0.10%	-7.14%	1.30%	[1.2%,1.4%]	-0.10%	-7.14%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 16: fix objects involved in all the crashes								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	71.80%	[70.9%,72.7%]	-7.50%	-9.46%	71.90%	[71.1%,72.7%]	-7.50%	-9.45%
2	12.20%	[11.9%,12.5%]	2.90%	31.18%	16.00%	[15.6%,16.5%]	3.90%	32.23%
3	13.60%	[13.1%,14.1%]	3.80%	38.78%	9.70%	[9.3%,10.0%]	2.80%	40.58%
4	2.10%	[2.0%,2.2%]	0.70%	50.00%	2.10%	[2.0%,2.2%]	0.70%	50.00%
5	0.30%	[0.3%,0.3%]	0.10%	50.00%	0.30%	[0.3%,0.4%]	0.10%	50.00%

Table 18: the involvement of fix object in a crash lead to have more severe level of injury. In the *Scenario 15* the differences in percentage more evident are showed in the 1th, 2th and 3th, the most growth in percentage concern level 2,3 and 4.

Scenario 16 presents the same behavior about the difference in percentage, instead regarding the growth in percentage it uniformly growth with the IS levels progress.

Table 19. Scenarios 17, 18: intersection influence.

Scenario 17: no crashes occur at an intersection								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	80.40%	[79.8%,81.1%]	1.10%	1.39%	80.50%	[79.9%,81.2%]	1.10%	1.39%
2	8.90%	[8.6%,9.1%]	-0.40%	-4.30%	11.40%	[11.1%,11.8%]	-0.70%	-5.79%
3	9.20%	[8.8%,9.5%]	-0.60%	-6.12%	6.50%	[6.2%,6.7%]	-0.40%	-5.80%
4	1.30%	[1.3%,1.4%]	-0.10%	-7.14%	1.40%	[1.3%,1.4%]	0.00%	0.00%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 18: all crashes occur at an intersection								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	77.60%	[76.8%,78.3%]	-1.70%	-2.14%	77.80%	[77.1%,78.5%]	-1.60%	-2.02%
2	10.00%	[9.7%,10.3%]	0.70%	7.53%	12.90%	[12.5%,13.3%]	0.80%	6.61%
3	10.60%	[10.2%,11.0%]	0.80%	8.16%	7.50%	[7.2%,7.7%]	0.60%	8.70%
4	1.60%	[1.5%,1.6%]	0.20%	14.29%	1.60%	[1.5%,1.7%]	0.20%	14.29%
5	0.20%	[0.2%,0.3%]	0.00%	0.00%	0.20%	[0.2%,0.3%]	0.00%	0.00%

Table 19: these two scenarios indicate a greater probability of having a higher IS level if the crash occurs at an intersection. The 1th, 3th and 2th level are the more influenced in percentage difference (in both years and scenarios). The growth in percentage is more about levels 2, 3 and 4 in 2016 and 2 and 3 in 2017.

Scenario 18 present, in addition to homogeneous growth in level 2 and 3, the biggest growth in percentage at 4th (incapacitating injury) level.

Table 20. Scenarios 19, 20: rollover dynamic.

Scenario 19: all crashes present no rolled-over								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	79.60%	[78.9%,80.2%]	0.30%	0.38%	79.70%	[79.1%,80.3%]	0.30%	0.38%
2	9.20%	[9.0%,9.6%]	-0.10%	-1.08%	11.90%	[11.6%,12.3%]	-0.20%	-1.65%
3	9.60%	[9.3%,9.9%]	-0.20%	-2.04%	6.80%	[6.5%,7.0%]	-0.10%	-1.45%
4	1.40%	[1.3%,1.4%]	0.00%	0.00%	1.40%	[1.3%,1.4%]	0.00%	0.00%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 20: all crashes present rolled-over								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	60.50%	[58.7%,62.0%]	-18.80%	-23.71%	59.60%	[58.1%,61.2%]	-19.80%	-24.94%
2	15.60%	[15.2%,16.1%]	6.30%	67.74%	21.50%	[20.8%,22.1%]	9.40%	77.69%
3	20.10%	[19.1%,21.2%]	10.30%	105.10%	15.00%	[14.2%,15.8%]	8.10%	117.39%
4	3.30%	[3.1%,3.5%]	1.90%	135.71%	3.40%	[3.2%,3.6%]	2.00%	142.86%
5	0.50%	[0.5%,0.6%]	0.30%	150.00%	0.60%	[0.5%,0.6%]	0.40%	200.00%

Table 20: the dynamics of overturning is one of the most dangerous for the IS level. In the *Scenario 20* all the levels (from 2 to 5) are affected by a strong increase in percentage, in the 5th level percentage grows up to 1.5 times in 2016 and almost 2 times in 2017.

This confirms the importance of this phenomenon and the huge impact it can have on human lives.

Table 21. Scenarios 21, 22, 23, 24: light effect.

Scenario 21: presence of enough light								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	79.40%	[78.8%,80.0%]	0.10%	0.13%	79.50%	[78.9%,80.1%]	0.10%	0.13%
2	9.30%	[9.0%,9.6%]	0.00%	0.00%	12.00%	[11.6%,12.4%]	-0.10%	-0.83%
3	9.70%	[9.4%,10.0%]	-0.10%	-1.02%	6.80%	[6.6%,7.1%]	-0.10%	-1.45%
4	1.40%	[1.3%,1.5%]	0.00%	0.00%	1.40%	[1.4%,1.5%]	0.00%	0.00%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 22: presence of no light								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	77.80%	[77.0%,78.6%]	-1.50%	-1.89%	77.60%	[76.7%,78.4%]	-1.80%	-2.27%
2	9.90%	[9.6%,10.3%]	0.60%	6.45%	13.00%	[12.6%,13.5%]	0.90%	7.44%
3	10.50%	[10.1%,11.0%]	0.70%	7.14%	7.50%	[7.2%,7.9%]	0.60%	8.70%
4	1.50%	[1.5%,1.6%]	0.10%	7.14%	1.60%	[1.5%,1.7%]	0.20%	14.29%
5	0.20%	[0.2%,0.3%]	0.00%	0.00%	0.20%	[0.2%,0.3%]	0.00%	0.00%
Scenario 23: presence of dawn light								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	77.90%	[76.8%,78.8%]	-1.40%	-1.77%	79.30%	[78.2%,80.3%]	-0.10%	-0.13%
2	9.90%	[9.5%,10.3%]	0.60%	6.45%	12.10%	[11.5%,12.7%]	0.00%	0.00%
3	10.50%	[9.9%,11.0%]	0.70%	7.14%	6.90%	[6.5%,7.3%]	0.00%	0.00%
4	1.50%	[1.4%,1.6%]	0.10%	7.14%	1.50%	[1.4%,1.6%]	0.10%	7.14%
5	0.20%	[0.2%,0.3%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%
Scenario 24: presence of dusk light								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	80.60%	[79.7%,81.5%]	1.30%	1.64%	79.00%	[78.1%,80.0%]	-0.40%	-0.50%
2	8.80%	[8.4%,9.2%]	-0.50%	-5.38%	12.30%	[11.7%,12.8%]	0.20%	1.65%
3	9.10%	[8.7%,9.6%]	-0.70%	-7.14%	7.00%	[6.6%,7.4%]	0.10%	1.45%
4	1.30%	[1.2%,1.4%]	-0.10%	-7.14%	1.50%	[1.4%,1.6%]	0.10%	7.14%
5	0.20%	[0.2%,0.2%]	0.00%	0.00%	0.20%	[0.2%,0.2%]	0.00%	0.00%

Table 21: in optimal condition of *enough light* the probability to have a crash is very poor and the probability to have a no injury is slightly higher than the original scenario.

In the *no light* case, a more evident growth in percentage concern the 2, 3 and 4 IS level. In the *dawn light* case, in 2016 have a similar trend to the *no light* case and for the 2017 only the 4th level is considerable affected by a growth in percentage. *Dusk light* presents low decrease in percentage about all the level differently by 2017 that present an opposite trend lightly increasing the percentages in all levels except for the 5th.

Table 22. Scenario 25: female gender effect.

Scenario 25: all female								
	2016				2017			
IS	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth	Market Share	Confidence interval	Δ(F - O) Difference	(F / O) – 1 % Growth
1	75.70%	[74.9%,76.4%]	-3.60%	-4.54%	75.90%	[75.2%,76.6%]	-3.50%	-4.41%
2	10.80%	[10.5%,11.1%]	1.50%	16.13%	13.90%	[13.5%,14.4%]	1.80%	14.88%
3	11.60%	[11.2%,12.0%]	1.80%	18.37%	8.10%	[7.8%,8.4%]	1.20%	17.39%
4	1.70%	[1.7%,1.8%]	0.30%	21.43%	1.80%	[1.7%,1.8%]	0.40%	28.57%
5	0.30%	[0.2%,0.3%]	0.10%	50.00%	0.30%	[0.3%,0.3%]	0.10%	50.00%

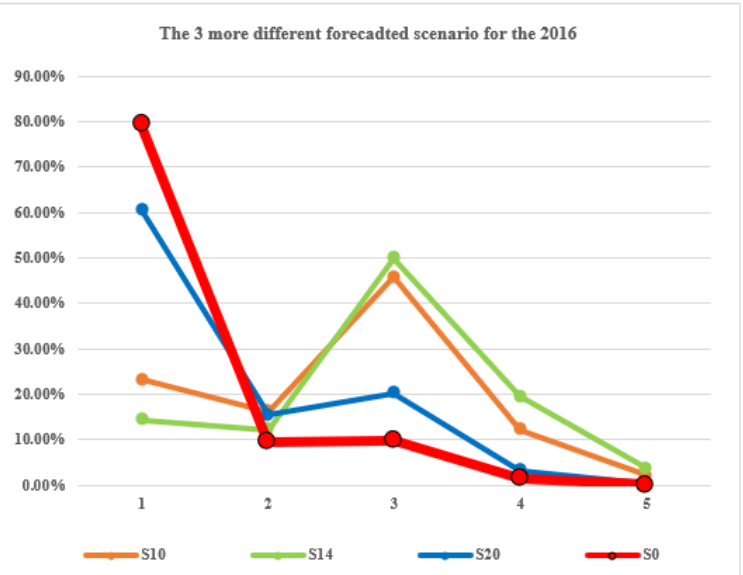
Table 22: this scenario put in evidence the tendency of women to get an IS level higher than men. With the growth of IS level also percentage for each level follow this trend as well.

The 5th level is the more evident case where, in both of years, where the increase of probability is around 50% than the original scenario.

This could be caused by a more vulnerability of the female body or a different driving habits such as the right use of seat-belts caused by breast discomfort.

Printing the results in a single graph too many lines result overlapped, for this reason only the three already mentioned scenario are shown in **Figure 9.**

S0	Real Scenario
S1	Scenario 1: all people are 5 years older
S2	Scenario 2: all people are 10 years older
S3	Scenario 3: all people are 15 years older
S4	Scenario 4: all people are drugged
S5	Scenario 5: all people are drunk
S6	Scenario 6: all people use medicine
S7	Scenario 7: all people drive under combined effect
S8	Scenario 8: all people are driver
S9	Scenario 9: all people are occupant
S10	Scenario 10: all people are pedestrian
S11	Scenario 11: nobody uses seat-belt
S12	Scenario 12: all people use seat-belt
S13	Scenario 13: no people ejected/trapped
S14	Scenario 14: all people ejected/trapped
S15	Scenario 15: no fix objects involved in crashes
S16	Scenario 16: fix objects involved in all the crashes
S17	Scenario 17: no crashes occur at an intersection
S18	Scenario 18: all crashes occur at an intersection
S19	Scenario 19: all crashes present no rolled-over
S20	Scenario 20: all crashes present rolled-over
S21	Scenario 21: presence of enough light
S22	Scenario 22: presence of no light
S23	Scenario 23: presence of dawn light
S24	Scenario 24: presence of dusk light
S25	Scenario 25: all female



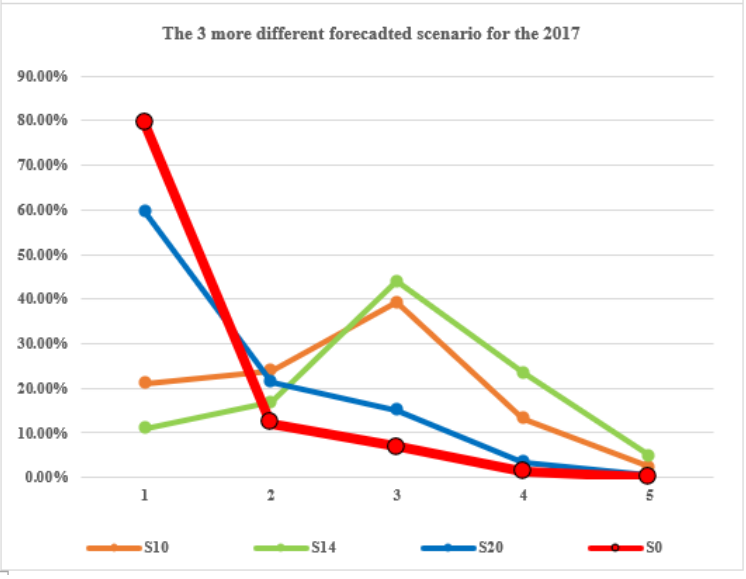


Figure 9. The three most divergent scenarios for the 2016 and 2017

4.3 Discussion

In this Chapter, the results from model estimation are compared to those available in the literature, this allows to understand if the model outcomes are consistent with those obtained in existing studies and to assess possible differences.

This present study is based on a large database that contains a high number of explanatory variables; the majority of the existing studies are usually based on a limited amount of data and focus on a few attributes.

For example, Ulfarsson (2004) used data from the Master Accident Record System (MARS) of the Washington State Department of Transportation (from January 1,1993 to July 31, 1996 collected in the northwest region of Washington State) to evaluate injury severity across female and male drivers in single and two vehicle crashes involving pickups, SUV or minivan. The analysis based on a multinomial logit concluded that vehicle type has a considerable effect on accident type and injury severity. and that female drivers are more likely to be subjected to injures during an accident than male drivers. Women are also more likely to suffer from fatal and incapacitating injuries in accidents involving a single SUV or minivan vehicle and as drivers in a vehicle that have collided with a pick-up. These results are consistent with those obtained in this study; the parameter relative to gender is positive (see **Table 7** and **Table 8**), which attests that females involved in a crash are more likely to suffer severe injuries than male in the same conditions. This is also confirmed by the scenario 25 in Table 9, where by assuming that the entire population is composed of women the number of "no injury" cases decreases, while the rates for all the other injury severity categories increase. There are no evident differences across 2016 and 2017.

Schott (1998) used *log-linear models* estimated on crashes recorded by the Florida Department of Highway Safety and Motor Vehicles (DHSMV) in 1994 and 1995 to study the relation among age and injury severity. Three IS were studied, and the results confirmed that IS is strongly linked to age; middle-aged drivers are more likely to be involved in specific crashes scenario where very young and young drivers tend to speed in curve, and that older drivers have the opposite trend and they are more likely to be involved in fatal accidents. Again these results are consistent with the findings reported in **Table 7** and **Table 8**, age related coefficients are all positive and their values increase with age, attesting that elderly suffer from more severe injuries when involved in a crash. Scenario 1, 2 and 3 in **Table 10** show how IS change when the average age of the population increases by 5, 10 and 15 years, respectively.

Kim (2013) studied driver-injury severity in single-vehicle crashes using a *mixed logit model* using data from California. The random parameter specification found heterogeneity related to age, for drivers older than 65 about half the population had a higher probability of incurring in a fatal injury and the other half had a lower probability. In addition, males on average had a higher probability of fatal injury in a newer vehicle compared to females. This result is not in line with the one presented by Ulfarsson (2004) and with the results of our analysis based on Maryland data.

Kockelman (2002) used a sample of police-reported crash from the 1998 National Automotive Sampling System (GES) collected in the US. The study based on an *ordered Probit model* shows that the type of collision, vehicle type, driver gender, number of vehicles involved, and the use of alcohol play a fundamental role in determining the IS. Rollover and head-on collisions are particularly serious. In addition, males tend to go much better than females . On the contrary, *daylight* / *dusk* conditions had rather negligible effects on injuries sustained by drivers. All these results are consistent with the outcomes of the models in **Table 7** and **Table 8**, not all the variables are present in both studies but the common ones have similar effect on IS. A few differences exist about the significance of some coefficients, like the one related to the use of alcohol which resulted not very significant on the Maryland data for both 2016 and 2017.

Abdel-Aty (2003) used crash data relative to Central Florida and collected in 1996 and 1997 to estimate an *ordered Probit* model. Results show the significance of driver's age, gender, seat belt use, point of impact and vehicle type on the injury severity level; in addition, driving under the effect of alcohol and lighting conditions affected considerably the probability of injuries in the roadway. Again, the results obtained for Maryland are comparable to those from central Florida exept for the significance of the alcohol related variable and the *Dawn light* and *Dusk light* conditions.

Ho-Yin (2003) studied the effect of sets belts use on the risk of suffering fatal injuries; the results show that 54% reduction in fatal injuries can be achieved by using properly seat belts. It was also found that the age group 16-19 has the highest percentage of fatal accidents due to the failure to use the seat belts. This finding is consistent with the model results obtained with Maryland data; however, the present study does not consider interaction among age and the use of seat belts and therefore nothing can be said about the age group most affected by the misuse of seatbelts.

A *logistic model* was proposed by Al-Ghamdi (2002) to analyze serious accidents reported in traffic police records in Riyadh, the capital of Saudi Arabia; the study only considers urban roads from August 1997 to November 1998 and is based on a sample of 560 individuals. One of the most significant results reported is that a fatal accident is less likely to happen near an intersection.

In the model developed on Maryland data the sign of the parameter related to the intersection location is positive, attesting that accidents happening near an intersection increase the IS score. In addition, from the scenarios 17 and 18 in **Table 19** it is possible to note that the increase/decrease of IS level affected by the intersection location does not influence considerably the fatal injury cases (score 5), but the presence of an intersection tends to decrease the percentage of accidents in the lowest severity level and to increase those in level 2 and 3, although all these effects are very minor.

5 CONCLUSION

This manuscript aimed at exploring crash data from police reports, modeling Injury Severity (IS), and understanding which factors have a major impact on the accidents' outcomes. To this scope, data from police reports and relative to the entire State of Maryland were used to estimate an ordered logit. The models were estimated using a state-of-the-art software for discrete choice models, called Biogeme.

The study is based on a very comprehensive database, that includes all accidents reported in Maryland for the years 2016 and 2017. Unlike most of the works available in the literature based on a sample, this research is based on the entire accident population. Furthermore, the different sections of the surveys have been linked and managed using Panda (a Python library), which allows the analysis of a large number of variables. The model forecasts have then been applied to study the IS sensitivity to a number of significant variables; this is rarely done in accident analysis literature. The outcomes of these analyses can be used to create measures aimed at reducing the number of accidents and their severity.

From the analysis of the results obtained the following conclusions can be drawn:

- The variables *age, gender, alcohol, drug, medicine, DAM (*combined effect of drug, alcohol and medicine*), collision type, ejected/trapped, fix objects, intersection, light conditions, person type, road type, rolled over, seatbelt, surface conditions, traffic control* and *vehicle body*, they all play a major role in crash IS.
- From the sensitivity analysis it can be said that attributes that describe the involvement of *pedestrians*, or that describe *rollover* and *ejected/trapped* conditions are particularly serious, contributing to more severe injury levels.
- Males tend to get involved in accident with higher IS scores, when compared to females.
- The majority of the variable estimated were found to be significant at 5% level of significance, which suggests that these variables were indeed good explanatory variables. Some of them resulted to be not significant but were kept in the final model for their importance in the attempt to understand all the factors that determines IS. For example, surprisingly the *alcohol* variable was found to be not significant. Several reasons might explain this result. Perhaps, the effect of the alcohol related variable gets diluted due to the presence of the variables that describe the combined effect of drug, alcohol and medicaments (*DAM*). Police agents might wrongly report that variable, or that in recent time people are under the influence of several factors and that alcohol does not represent one of the most dangerous substance for drivers.

- In addition, the following variables were found to be not significant: *angle meet, collision type*, *wet surface* and *yield sign* for both years analyzed. These results are not consistent with those reported in the literature, but might depend on the local conditions, especially for the variable *wet surface.*

This study has demonstrated how quantitative analyses can be used to study complex problems, and how data driven methods can support policy makers called at making important decisions about safety and peoples' life.

Several avenues for further research are possible. The IS score reported by the police and used in this study can be biased because it is based on just partial information. The police report database can be integrated with hospital records to quantify this bias and to compare the KABCO scale and AIS scale. Although this is technically feasible, hospital data are protected for privacy issues and their access is limited, also the data integration is a huge task and requires the application of probabilistic data linkage techniques. The database can be used to study particular classes of accidents, for example those involving pedestrians, bikers and in the future e-scooters that are gaining popularity in very recent years. Also, due to the aging of the population, all questions related to elderly can be explored in much more detail using the same database and similar techniques. Finally, the database can be integrated with real time traffic characteristics for a better management of accidents by both highway response teams and emergency rooms in the hospitals.

6 APPENDICES

6.1 APPENDIX 1 - AIS score calculation

ISS is an established medical score to assess trauma severity (Baker SP, O'Neill B, Haddon W, Long WB (1974). "The Injury Severity Score: a method for describing patients with multiple injuries and evaluating emergency care". The Journal of Trauma. Lippincott Williams & Wilkins.)

The Abbreviated Injury Scale (AIS) is an anatomically based consensus-derived global severity scoring system that classifies each injury in every body region according to its relative severity on a six-point ordinal scale:

0. No injury
1. Minor
2. Moderate
3. Serious
4. Severe
5. Critical
6. Maximal (currently untreatable).

There are nine AIS chapters corresponding to nine body regions:

- Head
- Face
- Neck
- Thorax
- Abdomen
- Spine
- Upper Extremity
- Lower Extremity
- External and other

The ISS is based (see below) upon the Abbreviated Injury Scale (AIS). To calculate an ISS for an injured person, the body is divided into six ISS body regions. These body regions are:

- Head or neck – including cervical spine
- Face – including the facial skeleton, nose, mouth, eyes and ears
- Chest – thoracic spine and diaphragm
- Abdomen or pelvic contents – abdominal organs and lumbar spine

- Extremities or pelvic girdle – pelvic skeleton
- External

To calculate an ISS, take the highest AIS severity code in each of the three most severely injured ISS body regions, square each AIS code and add the three squared numbers for an ISS

$$ISS = A^2 + B^2 + C^2$$

where A, B, C are the AIS scores of the three most injured ISS body regions). The ISS scores range from 1 to 75 (i.e. AIS scores of 5 for each category). If any of the three scores is a 6, the score is automatically set at 75. Since a score of 6 ("survivable") indicates the futility of further medical care in preserving life, this may mean a cessation of further care in triage for a patient with a score of 6 in any category (TRAUMA.ORG, n.d.).

ISS could be reinterpreted as a classification in 6 classes related to the AIS score.

The following picture could better explain the AIS score evaluation with an example:

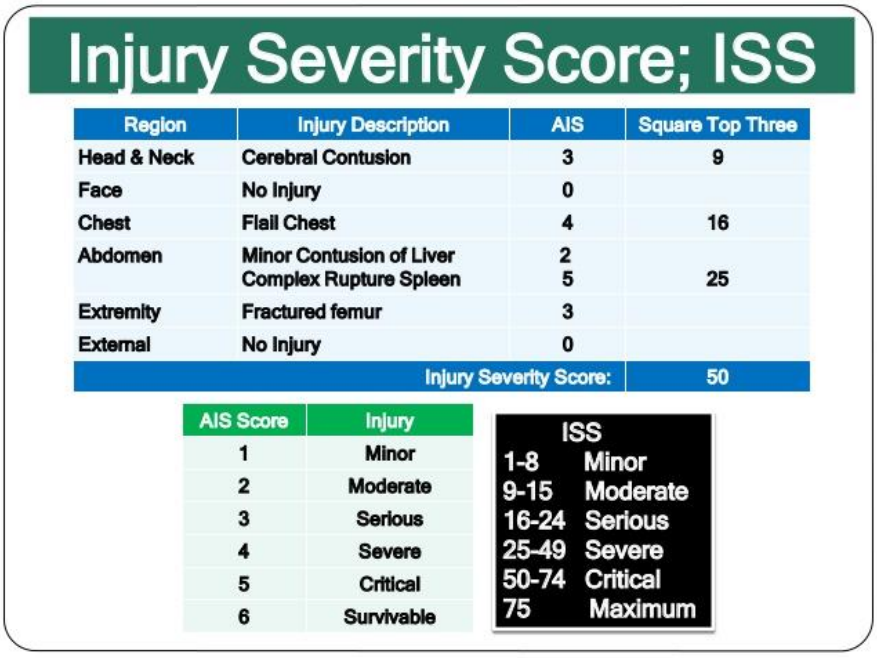


Figure 10. IS score ISS

As you can see the two scale have different number of possible values, 5 for the KABCO and 6 for the ASI one. This difference in the number between the two scale represent an additional issue faced in the modeling procedure after explained.

Parameter comparisons for these scaled models can help identify potential misclassifications on the KABCO scale by the police.

Ordered Logit Models are used to capture the fidelity of injury severity recorded on the KABCO scale. Model parameters are estimated using Maximum Likelihood Estimation.

6.2 APPENDIX 2 – Contextualization of the Project

The study for this thesis was carried out at the University of Maryland in College Park.

This study was possible thanks to an agreement between the Univrsity of Maryland (UMD) and the University of Maryland School of Medicine (SOM) during the period of my thesis in the United States. My American supervisor, Cinzia Cirillo, Professor at UMD in College Park, has been contacted for a collaboration in Baltimore at the Shock, Trauma and Anesthesiology Research organized research center (STAR-ORC) in order to improve data management hospital system. Prof. Cirillo provided to assemble a Team of students which comprises me and others tow component, Kartik Kaushik (Research Assistant at University of Maryland) and Darshan Pandit (Graduate Assistant, University of Maryland), experts in Phyton language and computer programming.

The collaboration aim was to make as easy as possible the accessibility to the many Database hold by SOM for research purpose, in the way to save time during this operation.

The data were available in multiple SAS file coming from different body administration like

- Police for the Road **Crash data**, collected coherently with the ACRC manual
- Health Service Cost Review Commission (HSCRC) for the **Patient data**
- Court for the **Citation data**. All the data are about the whole state of Maryland.

Having the data different source, they are collecting and treated with different Software so using different configuration and stored in different Database, all these features basically represents the issues for the purpose.

In this thesis only the first one has been investigated for reason of time because this was the first step of a huge project that has as goal the realization of a unique big database containing all the information available regarding the individuals so to have a unique source of information from which to draw and eventually to understand eventually misclassification in one database rather than in an another one. This study is focused on the years IS and IS, some differences were found about the data collecting procedure due to some slightly difference about the value assumed by some attribute, this inconsistency represents a non-negligible obstacle, this make the research very expensive in terms of time and energy by the researchers.

Since the information inside the database available were sensible because related to Age, date of birth, license number, home address and other personal information like that, some online-

courses data treatment and to identify and protect the researcher and SOM/FPI (Faculty Physicians Inc.) from cyber-attacks:

Courses:

- CITI program. Course in The Protection of Human Subjects. Group 2. Social / Behavioral Research Investigators and Key Personnel. 1 - Basic Course
- 2018 Creating Strong Passwords
- 2018 KnowBe4 Security Awareness Training
- 2018 Mobile Device Security
- IS Micro-module - Ransomware
- IS Micro-module - Safe Web browsing
- IS Micro-module - USB Attack - IS Handling Sensitive Information
- Significant New Findings: What They Are and What To Do About Them
- HIPAA 203 Information Technology
- HIPAA 201 Human Research
- HIPAA 125 Privacy & Security Awareness Training
- Department or Entity Scientific and Feasibility Review of Research

The works took place in the Shock, Trauma and Anesthesiology Research organized research center (STAR-ORC) that is a world-class, multi-disciplinary research and educational center focusing on critical care and organ support, resuscitation, surgical outcomes, patient safety and injury prevention (SOM, n.d.).

Little overview of the whole project:

The first idea for the whole project was to merge all the information from the different database in order to have a unique database where every observation (each refer to an individual) can show all the information regarding the story line of each individual from the road crash through the hospital to the citation scenario if available.

The purpose for the team was to provide for a new organization of the data in order to:

- Reduce time spent managing and manipulating data, for all the research purpose that SOM's staff need to carry on.
- Allow to respond to more complicated questions and further research interests querying the database and interact with analysts and modelers

- Increase productive uses of data in research and producing quality results
- Automate getting basic statistics for various conditions using a dashboard
- Provide for the realization of a model that can analyze the data, in particular way studying the Injury Severity assessment from different body as Police and Hospital to understand how the first one misclassifies the evaluation during the first evaluation in the place of the accident.

The mine focus of the team was to provide solution for:

- Data Preprocessing
- Implementation of Python packages for data management
- Storage & Retrieval solution
- Automated Codebook and Documentation
- Enabling Analytics
- Data Reporting for each year
- Realization of a Model to study the Injury Severity (topic of this thesis)
- Future Planned Steps
    - Working on loading HSCRC data, and citation data

Data Preprocessing

At first was loaded the unsanitized crash data from IS and IS.

Some value was substitute for an easier reading through an application of python scripts per dataset to:

- Apply standard data sanitization techniques eg: Remove special charecters, consecutive and trailing whitespaces, etc.
- Perform DateTime field transformations
- Handle special values based upon taxonomy eg: Interpret values ’A9.99’, ’A8.98’ as NaN values for the Crash Dataset
- Handle Geo-coordinates and special field values

Python package

- Package to generate customized data profiling reports for a collection of Pandas database
- Compares table schema against a Reference schema to track fields and field-types

- Provides reports on field-values, their interpretation and their distributions
- Facilitate quick lookups and helps identify issues in data

Storage & Retrieval

It is needed a Relational Database Management System (RDMS). For this purpose, the choice is fell on PostgreSQL as primary datastore:

- Efficient storage and retrieval of data
- Includes support for Geospatial computations
- Universal connectors allowing access from multiple tools and languages (SAS, ArcGIS, Python, R, Tableau) Retrieve PII/PHI free data from combined datasets.

Automated Codebook and Documentation

Quick summaries for new datasets received each year

- Allows for quick checks over taxonomy of categorical values or major field changes while loading the data itself
- Reduced data validation time
- Identify and allow taxonomy changes year-on-year

Standardized table-column-mapping facilitates:

- Data consistency against master dataset
- Tractable changes

Generate two standardized files for each table

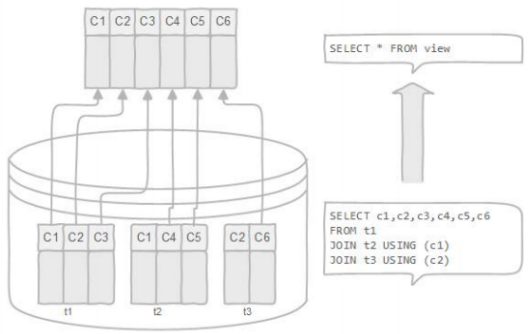
- A table summary file
- A detailed profile for each column, including distinct data

Use a standardized reference schema

Generate a mapping file for each new dataset

- Ensure tractability of schema and codes

A view is a window into the data, and can include joins



Row counts by table

Table	2016	2017	Total
Accident	120 124	114 846	234 970
Citation	58 136	58 606	116 742
Contributing Circumstances	697 719	666 268	1 363 987
EMS	35 099	34 573	69 672
Person	291 389	284 981	576 370
Road	120 122	114 846	234 968
Trailer Unit	3 791	3 571	7 362
Vehicle	226 257	215 793	442 050
Master View	316 820	309 758	626 578

Figure 11. Data managing

Enabling Analytics

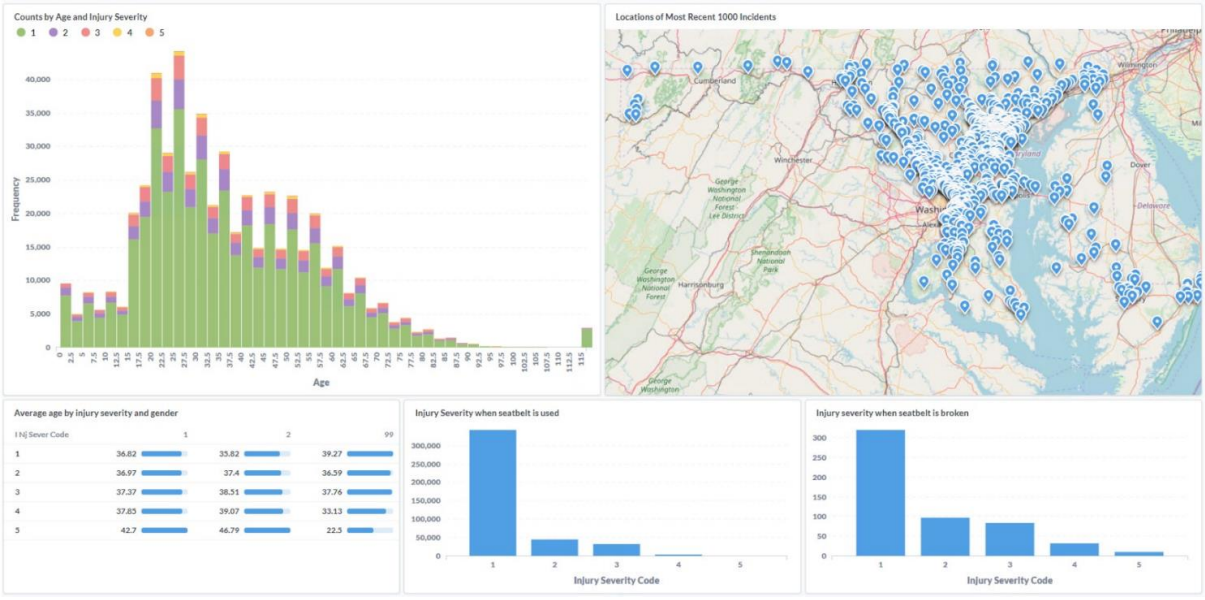


Figure 12. Screenshot from a dashboard tool

Data Reporting for each year

- Generate documentation geared towards end-users and data administrators for dataset updates received each year.

Future Planned Steps

Working on loading HSCRC data, and citation data

- Exploring merging options with crash, and MD CHART data

- Eventually produce a single dataset to track people right from crash, hospital, treatments, and court
  Together with the real-time traffic speed data
- Faster access to and evacuation of victims
  Would swallow any other datasets you might have
- We are used to handling all types of data, including big data
- Possibility to explore other storage solutions based on datasets and progress

MODEL Sanitation STEP and first goal to reach

At first, only the Crash data about the IS and IS data were available for the data management, which were subjected to a sanitized treatment regarding unknown and not reliable value.

As the first users of the resulting sanitized database, my role was underlining possible issue faced during the data acquisition from this new experimental database by the model.

The initial purposes were:

- Fosters better care and response
- Positive impacts on victim outcomes
- Injury severity modeled using Ordered Logit models

After the preliminary step about the sanitation of the Crash data, it starts the merging procedure to have all the Crash date into a single Dataset, in the meantime, an Ordered Logit Model was providing to analyze the Injury Severity as the dependent variable for crash data about IS and IS years separately.

6.3 APPENDIX 3 - Database description

At first there were not a complete Database able to explain the whole set of variables listed in it. A database is been composed though a merging procedure using different source as excel file from the (MODP, n.d.), access files provided by the STAR-ORC and from the ACRS manual.

Nevertheless, it was not possible to identify some variables.

Variable description:

Table 23. Whole database Dictionary available

ACCDATETIME: crash time and date	
DSKEY: (no dictionary, codes)	
REPORTNO: report number for the data	
RAMPMOVEMENTCODE: no information (all NaN value)	
LIGHTCODE: describe the light conditions	
00	Not Applicable
01	Daylight
03	Dark Lights On
04	Dark No Lights
05	Dawn
06	Dusk
07	Dark - Unknown Lighting
88	Other
99	Unknown
COUNTYNO: each number correspond to a County where the crash happened	
28	County in DISTRICT OF COLUMBIA
1	ALLEGANY
2	ANNE ARUNDEL
3	BALTIMORE
4	CALVERT
5	CAROLINE
6	CARROLL
7	CECIL
8	CHARLES
9	DORCHESTER
10	FREDERICK
11	GARRETT
12	HARFORD
13	HOWARD
14	KENT
15	MONTGOMERY

16	PRINCE GEORGES
17	QUEEN ANNES
18	ST MARYS
19	SOMERSET
20	TALBOT
21	WASHINGTON
22	WICOMICO
23	WORCESTER
24	BALTIMORE CITY
99	UNKNOWN
88	UNK-FOR-MASTER
25	County in DELAWARE
27	County in VIRGINIA
26	County in PENNSYLVANIA

MUNICODE: municipal code

JUNCTIONCODE: describe the type of junction

00	Not Applicable
01	Non Intersection
02	Intersection
03	Intersection Related
04	Driveway Alley Access Related
5.01	Interchange Related
6.01	Crossover Related
7.01	Railway Grade Crossing
8.04	Residential Driveway
9.04	Commercial Driveway
10.04	Alley

COLLISIONTYPECODE: describe the type of collision

00	Not Applicable
01	Head On
02	Head On Left Turn
03	Same Direction Rear End
04	Same Direction Rear End Right Turn
05	Same Direction Rear End Left Turn
06	Opposite Direction Sideswipe
07	Same Direction Sideswipe
08	Same Direction Right Turn
09	Same Direction Left Turn
10	Same Direction Both Left Turn
11	Same Movement Angle
12	Angle Meets Right Turn
13	Angle Meets Left Turn
14	Angle Meets Left Turn Head On
15	Opposite Direction Both Left Turn

17	Single Vehicle
88	Other
99	Unknown

SURFCONDCODE: describe the surface condition during the crash

00	Not Applicable
01	Wet
02	Dry
03	Snow
04	Ice
05	Mud, Dirt, Gravel
06	Slush
07	Water (standing/moving)
08	Sand
09	Oil
88	Other
99	Unknown

LANECODE: describe the area on the roadway where the crash is located

00	Not Applicable
01	Right Turn Lane
02	Left Turn Lane
03	Acceleration Lane
04	Deceleration Lane
05	Shoulder Area
06	Crossover Area
07	Off Road
08	Gore Area
09	Median Area
10	Parking Lot
11	Separator
12	Outside Right of Way
13	On Ramp
88	Other
99	Unknown

RDCONDCODE: describe eventually defect of the road where in correspondence of the crash

00	Not Applicable
01	No Defects
02	Shoulder Defect
03	Holes, Ruts, Etc.
04	Foreign Material
05	Loose Surface Material
06	Obstruction Not Lighted
07	Obstruction Not Signaled
08	View Obstructed

88	Other
99	Unknown

FIXOBJCODE: describe which fixed objects was eventually involved in the crash

00	Not Applicable
01	Bridge or Overpass
02	Building
03	Culvert or Ditch
04	Curb
05	Guardrail or Barrier
06	Embankment
07	Fence
08	Light Support Pole
09	Sign Support Pole
10	Other Pole
11	Tree Shrubbery
12	Construction Barrier
13	Crash Attenuator
14	Guardrail End
15	Concrete Traffic Barrier
16	Other Traffic Barrier
17	Traffic Signal Support
18	Mailbox
19	Bridge Overhead Structure
20	Bridge Pier Support
21	Bridge Rail
22	Culvert

WEATHERCODE: describe the weather condition during the crash

00	Not Applicable
02	Foggy
03	Raining
05	Severe Winds
6.01	Clear
7.01	Cloudy
8.04	Snow
9.04	Sleet
10.04	Blowing Snow
11.88	Blowing Sand, Soil, Dirt
12.04	Wintry Mix
88	Other
99	Unknown

LOCCODE: local codes

RAMPFLAG: describe if the crash happened on a ramp or not

Y	Crash located on a ramp
N	Crash not located on a ramp
X	Not applicable

SIGNALFLAG: describe if there were some traffic control sign close to the crash place

Y	Yes
N	No

CMZONEFLAG: describe if there were some construction site close to the crash place

Y	Yes
N	No
U	Unknown

INTERNUM: (no dictionary, all NaN value)

OFFICERINFO: (no dictionary, all NaN value)

AGENCYCODE: agency code who authored the report place

AREACODE: code for local area information no place (all NaN value)

HARMEVENTCODE1: describe the first main event that caused the harm event

00	Not Applicable
01	Other Vehicle
02	Parked Vehicle
03	Pedestrian
04	Bicycle
05	Other Pedalcycle
06	Other Conveyance
07	Railway Train
08	Animal
09	Fixed Object
10	Other Object
11	Overturn
12	Spilled Cargo
13	Jackknife
14	Units Separated
15	Other Non Collision
16	Off Road
17	Downhill Roadway
18	Explosion or Fire
19	Backing
20	U-turn
21.15	Immersion
22.15	Fell Jumped from Motor Vehicle
23.15	Thrown or Falling Object
88	Other
99	Unknown

HARMEVENTCODE2: describe the second main event that caused the harm event

00	Not Applicable
01	Other Vehicle
02	Parked Vehicle
03	Pedestrian
04	Bicycle
05	Other Pedalcycle

06	Other Conveyance
07	Railway Train
08	Animal
09	Fixed Object
10	Other Object
11	Overturn
12	Spilled Cargo
13	Jackknife
14	Units Separated
15	Other Non Collision
16	Off Road
17	Downhill Roadway
18	Explosion or Fire
19	Backing
20	U-turn
21.15	Immersion
22.15	Fell Jumped from Motor Vehicle
23.15	Thrown or Falling Object
88	Other
99	Unknown

LOCCASENO: (no dictionary, codes)

OFFICERID: officer identification code (all NaN value)

OFFICERNAME: officer name

REPORTTYPECODE: report type code (no dictionary, codes 1, 2, 3)

PHOTOSFLAG: Photography available

Y	Yes
N	No

LANENUMBER: number of lanes

LANEDIRECTIONCODE: describe the line direction

LANETYPECODE: describe the type of lane where the crash happened

00	NOT APPLICABLE
01	RIGHT TURN LANE
02	LEFT TURN LANE
03	ACCELERATION LANE
04	DECELERATION LANE
05	SHOULDER AREA
06	CROSSOVER AREA
07	OFF ROAD
08	GORE AREA
09	MEDIAN AREA
10	PARKING LOT
11	SEPERATOR
12	OUTSIDE RIGHT OF WAY
88	OTHER

99	UNKNOWN
13	ON RAMP
NaN	BLANK VALUE FROM ACRS

INTERSECTIONTYPECODE: describe the type of intersection where the crash happened

00	NOT APPLICABLE
01	FOUR-WAY INTERSECTION
02	T-INTERSECTION
03	Y-INTERSECTION
04	TRAFFIC CIRCLE
05	ROUNDABOUT
06	FIVE-POINT OR MORE
99	UNKNOWN
88	OTHER
NaN	BLANK VALUE FROM ACRS

TRAFFICCONTROLCODE: describe the type of traffic control where the crash happened

0	NOT APPLICABLE
1	NO CONTROLS
2	PERSON
3	TRAFFIC SIGNAL
4	FLASHING TRAFFIC SIGNAL
5	SCHOOL ZONE SIGN DEVICE
6	STOP SIGN
7	YIELD SIGN
8	WARNING SIGN
9	RAILWAY DEVICE
88	OTHER
99	UNKNOWN

TRAFFICCONTROLFUNCTIONFLAG: (no dictionary, codes X, Y, U, N)

NUMLANES: (no dictionary, different from LANENUMBER)

INTERAREACODE: describe the area where the crash happened

00	NOT APPLICABLE
02	INTERSECTION
03	INTERSECTION RELATED
04	ON RAMP ENTRANCE AREA
01	THRU ROADWAY
88	OTHER
99	UNKOWN
05	ON RAMP EXIT AREA
06	ON RAMP MID AREA
NaN	BLANK VALUE FROM ACRS

SCHOOLBUSINVOLVEDCODE: describe the scenario in which a school bus was involved in the crash

00	NOT APPLICABLE
01	NOT INVOLVED
02	DIRECTLY INVOLVED

03	INDIRECTLY INVOLVED
99	UNKNOWN

CMLOCATIONCODE: (no dictionary, almost NaN value)

CMCLOSURECODE: (no dictionary, almost NaN value)

CMWORKERSPRESENTFLAG: (no dictionary, codes X, Y, U, N)

REVIEWDATE: (no dictionary, all NaN value)

REVIEWOFFICERID: (no dictionary all NaN value)

REVIEWOFFICERNAME: (no dictionary, all NaN value)

SUPERDATE: (no dictionary, all NaN value)

SUPEROFFICERID: (no dictionary, all NaN value)

SUPEROFFICERNAME: (no dictionary, all NaN value)

NARRATIVE: description by the police about the dynamics of the crash

GOVPROPERTYTXT: description of damage to government property

CITATIONID: identification for the citation code

CITATION: citation code

CIRCUMSTANCEID: (no dictionary, code)

CONTRIBCODE1: first contribute that caused the accident

00	Not Applicable
01	Under Influence of Drugs
02	Under Influence of Alcohol
03	Under Influence of Medication
04	Under Combined Influence
05	Physical/Mental Difficulty
06	Fell Asleep, Fainted, Etc.
07	Failed to Give Full Time and Attention
08	Did Not Comply with License Restrictions
10	Improper Right Turn on Red
11	Failed to Yield Right of Way
12	Failed to Obey Stop Sign
13	Failed to Obey Traffic Signal
14	Failed to Obey Other Traffic Control
15	Failed to Keep Right of Center
16	Failed to Stop for School Bus
17	Wrong Way on One Way Road
18	Exceeded the Speed Limit
19	Operator Using Cellular Phone
20	Stopping in Lane/Roadway
21	Too Fast for Conditions
22	Followed Too Closely
23	Improper Turn
24	Improper Lane Change

25	Improper Backing
26	Improper Passing
27	Improper Signal
28	Improper Parking
29	Interference/Obstruction by Passenger
70.88	Ran Off the Road
71.88	Disregarded Other Road Markings
72.88	Operated Motor Vehicle in Erratic Reckless Manner
73.88	Swerved or Avoided Vehicle or Object in Road
74.88	Over Correcting Over Steering
75.88	Other Improper Action
76.88	Inattentive
77.88	Failure to Obey Traffic Signs Signals or Officer
78.88	Wrong Side of Road
NaN	BLANK VALUE FROM ACRS

CONTRIBCODE2: second contribute that caused the accident

00	Not Applicable
01	Under Influence of Drugs
02	Under Influence of Alcohol
03	Under Influence of Medication
04	Under Combined Influence
05	Physical/Mental Difficulty
06	Fell Asleep, Fainted, Etc.
07	Failed to Give Full Time and Attention
08	Did Not Comply with License Restrictions
10	Improper Right Turn on Red
11	Failed to Yield Right of Way
12	Failed to Obey Stop Sign
13	Failed to Obey Traffic Signal
14	Failed to Obey Other Traffic Control
15	Failed to Keep Right of Center
16	Failed to Stop for School Bus
17	Wrong Way on One Way Road
18	Exceeded the Speed Limit
19	Operator Using Cellular Phone
20	Stopping in Lane/Roadway
21	Too Fast for Conditions
22	Followed Too Closely
23	Improper Turn
24	Improper Lane Change
25	Improper Backing
26	Improper Passing
27	Improper Signal
28	Improper Parking
29	Interference/Obstruction by Passenger

70.88	Ran Off the Road
71.88	Disregarded Other Road Markings
72.88	Operated Motor Vehicle in Erratic Reckless Manner
73.88	Swerved or Avoided Vehicle or Object in Road
74.88	Over Correcting Over Steering
75.88	Other Improper Action
76.88	Inattentive
77.88	Failure to Obey Traffic Signs Signals or Officer
78.88	Wrong Side of Road
NaN	BLANK VALUE FROM ACRS

CONTRIBCODE3: third contribute that caused the accident

00	Not Applicable
01	Under Influence of Drugs
02	Under Influence of Alcohol
03	Under Influence of Medication
04	Under Combined Influence
05	Physical/Mental Difficulty
06	Fell Asleep, Fainted, Etc.
07	Failed to Give Full Time and Attention
08	Did Not Comply with License Restrictions
10	Improper Right Turn on Red
11	Failed to Yield Right of Way
12	Failed to Obey Stop Sign
13	Failed to Obey Traffic Signal
14	Failed to Obey Other Traffic Control
15	Failed to Keep Right of Center
16	Failed to Stop for School Bus
17	Wrong Way on One Way Road
18	Exceeded the Speed Limit
19	Operator Using Cellular Phone
20	Stopping in Lane/Roadway
21	Too Fast for Conditions
22	Followed Too Closely
23	Improper Turn
24	Improper Lane Change
25	Improper Backing
26	Improper Passing
27	Improper Signal
28	Improper Parking
29	Interference/Obstruction by Passenger
70.88	Ran Off the Road
71.88	Disregarded Other Road Markings
72.88	Operated Motor Vehicle in Erratic Reckless Manner
73.88	Swerved or Avoided Vehicle or Object in Road
74.88	Over Correcting Over Steering

75.88	Other Improper Action
76.88	Inattentive
77.88	Failure to Obey Traffic Signs Signals or Officer
78.88	Wrong Side of Road
NaN	BLANK VALUE FROM ACRS

CONTRIBCODE4: fourth contribute that caused the accident

00	Not Applicable
01	Under Influence of Drugs
02	Under Influence of Alcohol
03	Under Influence of Medication
04	Under Combined Influence
05	Physical/Mental Difficulty
06	Fell Asleep, Fainted, Etc.
07	Failed to Give Full Time and Attention
08	Did Not Comply with License Restrictions
10	Improper Right Turn on Red
11	Failed to Yield Right of Way
12	Failed to Obey Stop Sign
13	Failed to Obey Traffic Signal
14	Failed to Obey Other Traffic Control
15	Failed to Keep Right of Center
16	Failed to Stop for School Bus
17	Wrong Way on One Way Road
18	Exceeded the Speed Limit
19	Operator Using Cellular Phone
20	Stopping in Lane/Roadway
21	Too Fast for Conditions
22	Followed Too Closely
23	Improper Turn
24	Improper Lane Change
25	Improper Backing
26	Improper Passing
27	Improper Signal
28	Improper Parking
29	Interference/Obstruction by Passenger
70.88	Ran Off the Road
71.88	Disregarded Other Road Markings
72.88	Operated Motor Vehicle in Erratic Reckless Manner
73.88	Swerved or Avoided Vehicle or Object in Road
74.88	Over Correcting Over Steering
75.88	Other Improper Action
76.88	Inattentive
77.88	Failure to Obey Traffic Signs Signals or Officer
78.88	Wrong Side of Road
NaN	BLANK VALUE FROM ACRS

CONTRIBFLAG: (no dictionary)

EMSUNITLABEL: Alph Character related to EMS Unit

EMSID: (no dictionary, code)

RUNREPNO: (no dictionary, code)

EMSUNITTAKENBY: Text of unit

EMSUNITTAKENTO: Text of taken to

EMSSNO: (no dictionary)

EMSTRANSPORTTYPEFLAG: (no dictionary, all NaN value)

PERSONID: identification code for an individual

SEX: describe the gender of the person

F	Female
M	Male
U	Unknown

CONDITIONCODE: describe in which condition was the individual immediately after the accident

00	Not Applicable
01	Apparently Normal
02	Had Been Drinking
03	Using Drugs
04	Physical Defects
05	Other Handicaps
06	Ill
07	Fatigued Fainted
08	Apparently Asleep
09	Emotional Depressed Angry Disturbed
10	Influenced by Medications and/or Drugs and/or Alcohol
88	Other Handicaps
99	Unknown

DRUNIT: (no dictionary)

INJSEVERCODE: describe the injury severity assessed by the police to the victim of the crash

01	No Injury
02	Non-incapacitating Injury
03	Possible Incapacitating Injury
04	Incapacitating/Disabled Injury
05	Fatal Injury

PEDUNIT: (no dictionary)

OCCUNIT: (no dictionary)

OCCNUM: (no dictionary)

FIRSTNAME: Name of the individual involved in the accident

MIDDLEINITIAL: middle initial of the individual involved in the accident

LASTNAME: Last name of the individual involved in the accident

OCCSEATPOSCODE: describe the position of the occupant during the accident

00	Not Applicable
01	In Vehicle
02	Center Front Seat
03	Right Front Seat
04	Left Rear/Motorcycle Passenger
05	Center Rear Seat
06	Right Rear Seat
07	Other in Vehicle
08	Cargo Area
09	Riding on Motor Vehicle Exterior
12.88	Sleeper Section of Cab
13.08	Other Enclosed Cargo Area
14.08	Unenclosed Cargo Area
15.88	Trailer Unit
88	Other
NaN	BLANK VALUE FROM ACRS
99	Unknown

PEDVISIBLECODE: describe the visibility of the pedestrian in the moment of the accident

00	Not Applicable
01	Light Clothing
02	Dark Clothing
03	Mixed Clothing
04	Reflective Material
05	Head Light
06	Rear Reflector
07	Head Light and Reflectors
88	Other
99	Unknown

PEDLOCATIONCODE: describe the position of the pedestrian during the accident

00	NOT APPLICABLE
01	IN VEHICLE
02	Center Front Seat
03	Right Front Seat
04	Left Rear/MC Pass
05	Center Rear Seat
06	Right Rear Seat
07	Other In Vehicle
08	Cargo Area
09	RIDING ON MOTOR VEHICLE EXTERIOR
88	Other
99	Unknown
12.88	SLEEPER SECTION OF CAB
13.08	OTHER ENCLOSED CARGO AREA

14.08	UNENCLOSED CARGO AREA
15.88	TRAILER UNIT
A9.99	BLANK VALUE FROM ACRS

PEDOBEYCODE: describe the behavior of the pedestrian about the signals during the accident

00	Not Applicable
01	No Pedestrian Signal
02	Obeyed Pedestrian Signal
03	Disobeyed Pedestrian Signal
04	Pedestrian Signal Malfunction
88	Other
99	Unknown

PEDTYPECODE: describe the type of pedestrian involved in the accident

00	Not Applicable
01	Pedestrian
02	Bicyclist
03	Other Pedalcyclist
04	Rider of Animal
05	In Animal-Drawn Vehicle
06	Machine Operator/Rider
07	Other Conveyance
88	Other
99	Unknown

PERMOVEMENTCODE: (no dictionary)

PERSONTYPE: describe the type of individual

D	Driver
O	Occupant
P	Pedestrian

DEATHNUM: (no dictionary, almost NaN value)

AGE: age of the person

SUBSTTESTCODE: describe a substance test (only NaN and 99 value)

00	NOT APPLICABLE
01	TEST REFUSED
02	POSITIVE PRELIMINARY TEST
03	EVIDENCE TEST GIVEN
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

SUBSTUSECODE: describe substance used by the person

00	NOT APPLICABLE
01	NONE DETECTED
11	ALCOHOL PRESENT
12	ILLEGAL DRUG PRESENT
13	MEDICATION PRESENT

14	COMBINED SUBSTANCE PRESENT
21	ALCOHOL CONTRIBUTED
22	ILLEGAL DRUG CONTRIBUTED
23	MEDICATION CONTRIBUTED
24	COMBINATION CONTRIBUTED
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

BAC: describe the amount of alcohol in the blood with a numerical value (not permitted by the law a value over than 0.08)

FAULTFLAG: describe if the fault is of the person described or not

Y	Yes
N	No
U	Unknown
X	Not applicable

EQUIPPROBCODE: describe the safety equipment problem during the accident

00	Not Applicable
01	Report Number for the data
11	Belts/Anchors Broken
13	Belt(s) Misused
31	Air Bag Failed
42	Facing Wrong Way
43	Not Anchored Right
44	Anchor Not Secure
45	Not Strapped Right
46	Strap/Tether Loose
47	Size/Type Improper
88	Other
99	Unknown

SAFEQUIPCODE: describe the safety equipment used during the accident

00	Not Applicable
01	None
11	Lap Belt Only
12	Shoulder Belt Only
13	Local Codes
14	Traffic Control Signal Flag
15.14	Child Restraint System Forward Facing
16.14	Child Restraint System Rear Facing
17.14	Construction Zone Related Flag
18.14	Agency Code who authored the report
21	Code for local area information
22	MC/Bike Eye Protection Only
23	MC/Bike Helmet and Eye Protection
24.88	Protective Pads
25.88	Reflective Clothing

26.88	Lighting
31	Air Bag (Only)
32	Air Bag and Belt(s)
88	Other
99	Unknown

WOULDHAVELIVEDFLAG: (no dictionary, all NaN value)

EJECTCODE: describe if driver or occupant were ejected from de vehicle

00	Not Applicable
01	Not Ejected/Trapped
02	Fully Ejected
03	Partially Ejected
04	Trapped
88	Other
99	Unknown

DRIVERDOB: describe the date of birth of the individual

LICNUM: describe the license number of the driver

PERSTATECODE: describe the state of pertinence were the accident happen

CLASS: describe the License Class - State Specific

CDLFLAG: describe if the driver has a Commercial Driver's License

Y	Yes
N	No
X	Not applicable

ALCODRUGIMPAIREDFLAG:

A	Alcohol
D	Drug
B	Both Alcohol and Drug
N	Not applicable
None	Unknown

DEATHSUFFIX: (no dictionary, code A, B, None)

PERSONPHONENUMBER: describe the phone number of every person involved in the accident

PERSONOTHERPHONE: describe the alternative phone number of every person involved in the accident

PERSONSTREETADDRESS: describe the address of every person involved in the accident

PERSONCITY: describe the city of provenience of every person involved in the accident

PERSONSTATECODE: describe the state of provenience of every person involved in the accident

PERSONZIPCODE: describe the zip code of every person involved in the accident

NONMOTORACTIONTIMECODE1: (no dictionary)

NONMOTORACTIONTIMECODE2: (no dictionary)

NONMOTORPRIORCODE: (no dictionary)

UNITFIRSTSTRIKE: (no dictionary, code NaN,1)

AIRBAGCODE: (no dictionary)

DISTRACTEDBYCODE: describe by what the person was distracted during the accident

00	NOT DISTRACTED
01	LOOKED BUT DID NOT SEE
03	BY OTHER OCCUPANTS
04	BY MOVING OBJECT IN VEHICLE
99	UNKNOWN
88	OTHER DISTRACTION
05	TALKING OR LISTENING TO CELLULAR PHONE
06	DIALING CELLULAR PHONE
07	ADJUSTING AUDIO AND OR CLIMATE CONTROLS
09	USING OTHER DEVICE CONTROLS INTEGRAL TO VEHICLE
10	USING DEVICE OBJECT BROUGHT INTO VEHICLE
12	DISTRACTED BY OUTSIDE PERSON OBJECT OR EVENT
13	EATING OR DRINKING
14	SMOKING RELATED
15	OTHER CELLULAR PHONE RELATED
16	NO DRIVER PRESENT
17	INATTENTIVE OR LOST IN THOUGHT
02	OTHER ELECTRONIC DEVICE NAVIGATIONAL
18	TEXTING FROM CELLULAR PHONE
A9.99	BLANK VALUE FROM ACRS

ALCOTESTCODE: describe if an alcohol test was provided after the accident

00	Not Applicable
01	Test Refused
02	Positive Preliminary Test
03	Evidence Test Given
88	Other
99	Unknown

ALCOTESTTYPECODE: describe the type of alcohol test done

00	Not Applicable
01	Breath
02	Blood
03	Urine
88	Other
99	Unknown

DRUGTESTCODE: describe if a drug test was provided after the accident

00	Not Applicable
01	Test Refused
02	Positive Preliminary Test
03	Evidence Test Given
88	Other
99	Unknown

DRUGTESTRESULTFLAG: describe the outcome of the drug test

A	Not Applicable

P	Positive
N	Negative
U	Unknown

OCCSEATLOCATION: describe the occupant seat location of the person during the accident

00	NOT APPLICABLE
01	IN VEHICLE
02	Center Front Seat
03	Right Front Seat
04	Left Rear/MC Pass
05	Center Rear Seat
06	Right Rear Seat
07	Other In Vehicle
08	Cargo Area
09	RIDING ON MOTOR VEHICLE EXTERIOR
88	Other
99	Unknown
12.88	SLEEPER SECTION OF CAB
13.08	OTHER ENCLOSED CARGO AREA
14.08	UNENCLOSED CARGO AREA
15.88	TRAILER UNIT
NaN	BLANK VALUE FROM ACRS

OCCSEATROW: (no dictionary)

OCCPOSINROWCODE: describe the lane position of the car in the roadway

01	LEFT
02	MIDDLE
03	RIGHT
88	OTHER
99	UNKNOWN
00	NOT APPLICABLE
NaN	BLANK VALUE FROM ACRS

LICSTATUSFLAG: (no dictionary, code None, X)

CITATIONISSUEDFLAG: describe if there were citation issue related to the accident (no dictionary, code N, Y, X, None)

ROUTENUMBER: describe the number of the road where the accident happened

ROUTETYPECODE: describe the Route Types - Maintained by SHA

ROUTESUFFIX: describe the Route Suffix - Maintained by SHA

LOGMILE: describe the Log Mile - Maintained by SHA

RDCHARCODE: describe the road characteristics (contain only code NaN, 99)

00	NOT APPLICABLE
01	STRAIGHT & LEVEL
02	STRAIGHT & GRADE
03	STRAIGHT & HILL
04	CURVE & LEVEL
05	CURVE & GRADE

06	CURVE & HILL
07	ON BRIDGE
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

RDDIVCODE: describe the type of division between one or two ways trafficway

00	NOT APPLICABLE
01	TWO-WAY, NOT DIVIDED
02	ONE-WAY TRAFFICWAY
03	TWO-WAY, DIVIDED, UNPROTECTED (PAINTED>4 FEET)
04	TWO-WAY, DIVIDED, POSITIVE MEDIAN BARRIER
88	OTHER
99	UNKNOWN
5.01 TURN I	TWO-WAY, NOT DIVIDED WITH A CONTINUOUS LEFT
NaN	BLANK VALUE FROM ACRS

LOGMILEDIRFLAG: describe the direction of the road (no dictionary, code N,S,E,W,U,X, None)

ROADNAME: describe the Road Name - Maintained by SHA where the accident happened

DISTANCE: describe the Numeric distance from reference

FEETMILESFLAG: describe the Measurement for distance

DISTANCEDIRFLAG: describe the Numeric distance from reference

FINALLOGMILE: (no dictionary)

REFERENCENUMBER: describe the Numeric ID for Reference Point - Maintained by SHA

REFERENCETYPECODE: (no info)

REFERENCESUFFIX: (no info)

REFERENCEROADNAME: (no info)

COORDINATES: describe the coordinates of the accident place

RDALIGNMENTCODE: describe the geometric features of the road track in the place of the accident

00	NOT APPLICABLE
01	STRAIGHT
02	CURVE LEFT
03	CURVE RIGHT
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

RDGRADECODE: describe the level of grade in the road where the accident happened

00	NOT APPLICABLE
01	LEVEL
02	HILL CREST
03	HILL UPHILL
04	GRADE DOWNHILL
05	DIP SAG

06	ON BRIDGE
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

OFFROADTXT: describe the place of the accident when it happened off road

FINALXCOORDINATES: (no dictionary, all NaN value)

FINALYCOORDINATES: (no dictionary, all NaN value)

TRAILERRECORDID: describe the ID of the trailer involved in the accident

TOWEDVEHICLEUNITNO: describe the number of trailer towed by the vehicle involved in the accident (no dictionary, code 1, 2, 3, NaN)

TUOWNERFIRSTNAME: (no dictionary)

TUOWNERMIDDLEINIT: (no dictionary)

TUOWNERLASTNAME: (no dictionary)

OWNERSTREETADDRESS: describe the address of the vehicle owner

OWNERCITY: describe the city of the vehicle owner

TUOWNERSTATECODE: describe the city of the trailer owner

OWNERZIPCODE: describe the zip code of the vehicle owner

TUOWNERPHONENUMBER: describe the phone number of the trailer owner

TUOWNEROTHERPHONE: describe the phone number of the trailer owner

TUVIN: describe the trailer identification number

TUVEHYEAR: describe the trailer year

TUVEHMAKE: (no dictionary)

TUVEHMODEL: describe the model of the trailer

TUBODYTYPECODE: describe the body type of the trailer

TUPLATENUM: describe the plate number of the trailer

TUPLATESTATE: describe the plate state of the trailer

TUPLATEYEAR: describe the plate year of the trailer (no dictionary, all NaN value)

TUINSURER: (no dictionary)

TUPOLICYNUMBER: describe the policy number of the trailer

VEHICLEID: described the ID of the Vehicle

HARMEVENTCODE: describe the event that caused the harm in the accident

00	NOT APPLICABLE
01	OTHER VEHICLE
02	PARKED VEHICLE
03	PEDESTRIAN
04	BICYCLE
05	OTHER PEDALCYCLE
06	OTHER CONVEYANCE
07	RAILWAY TRAIN

08	ANIMAL
09	FIXED OBJECT
10	OTHER OBJECT
11	OVERTURN
12	SPILLED CARGO
13	JACKKNIFE
14	UNITS SEPARATED
15	OTHER NON COLLISION
16	OFF ROAD
17	DOWNHILL RUNAWAY
18	EXPLOSION OR FIRE
19	BACKING
20	U-TURN
88	OTHER
99	UNKNOWN
21.15	IMMERSION
22.15	FELL JUMPED FROM MOTOR VEHICLE
23.15	THROWN OR FALLING OBJECT
NaN	BLANK VALUE FROM ACRS

VEHOWNERFIRSTNAME: describe the first name of the owner of the vehicle involved in the accident

VEHOWNERMIDDLEINIT: describe the middle name of the owner of the vehicle involved in the accident

VEHOWNERLASTNAME: describe the last name of the owner of the vehicle involved in the accident

STREETADDRESS: describe the street address of the owner of the vehicle involved in the accident

CITY: describe the name of the city of the owner of the vehicle involved in the accident

VEHSTATECODE: describe the state code of the address owner of the vehicle involved in the accident

ZIP: describe the zip code of the owner of the vehicle involved in the accident

CONTIDIRECTIONCODE: (dictionary in excel file with but with conflicting value)

DAMAGECODE: describe the severity of the damage after the accident

00	Not Applicable
01	No Damage
02	Superficial
03	Functional
04	Disabling
05	Destroyed
88	Other
NaN	BLANK VALUE FROM ACRS
99	Unknown

VEHMOVEMENTCODE: describe the type of movement done by the vehicle during the accident

00	Not Applicable
01	Moving Constant Speed
02	Accelerating
03	Slowing or Stopping
04	Starting From Lane

05	Starting From Parked
06	Stopped in Traffic Lane
07	Changing Lanes
08	Passing
09	Parking
10	Parked
11	Backing
12	Making Left Turn
13	Making Right Turn
14	Right Turn on Red
15	Making U Turn
16	Skidding
17	Driverless Moving Vehicle
18.07	Leaving Traffic Lane
19.07	Entering Traffic Lane
20.03	Negotiating a Curve
88	Other
99	Unknown

VEHVIN: describe the Vehicle Identification Number

CVBODYTYPECODE: describe the commercial vehicle body type involved in the accident

00	Not Applicable
01	Bus
02	Van/Enclosed Box
03	Truck Tractor
04	Cargo Tank
05	Flatbed
06	Dump
07	Concrete Mixer
08	Auto Transporter
09	Garbage/Refuse
10.88	Hopper
11.88	Pole Trailer
12.88	Grain Chips Gravel
13.88	Log
14.88	Intermodal Container Carrier
15.88	Vehicle Towing Another Vehicle
88	Other
99	Unknown
NaN	Blank value

VEHVEHYEAR: describe the year of the vehicle involved during the accident

VEHVEHMAKE: describe the make of the vehicle involved during the accident

COMMERCIALFLAG: (dictionary in excel file with but with conflicting value, code None, N)

VEHVEHMODEL: describe the Model of Vehicle

DOTNUM: (no dictionary)

ICCNUM: (no dictionary)

HZMNUM: describe the Hazmat Number

CARRIER: describe the type of carrier involved in the accident

TOWEDAWAYFLAG: describe if a vehicle was towed away or not (no dictionary, value N, Y, X, U, None)

NUMAXLES: describe a Numeric value (no dictionary)

GVW: (no dictionary, all NaN value)

GOINGDIRECTIONCODE: describe the travel direction of the vehicle involved in the accident

1	North
2	South
3	East
4	West
99	Unknown
NaN	Blank value

VEHBODYTYPECODE: describe the vehicle body type involved in the accident

00	NOT APPLICABLE
01	MOTORCYCLE
02	PASSENGER CAR
03	STATION WAGON
04	LIMOUSINE
05	CARGO VAN/LIGHT TRUCK 2 AXLES (10,000LBS (4,536 KG) OR LESS)
06	MEDIUM/HEAVY TRUCKS 3 AXLES (MORE THAN 10,000LBS (4,536KG)
07	TRUCK TRACTOR
08	RECREATIONAL VEHICLE
09	FARM VEHICLE
10	TRANSIT BUS
11	CROSS COUNTRY BUS
12	SCHOOL BUS
13	AMBULANCE/EMERGENCY
14	AMBULANCE/NON EMERGENCY
15	FIRE VEHICLE/EMERGENCY
16	FIRE VEHICLE/NON EMERGENCY
17	POLICE VEHICLE/EMERGENCY
18	POLICE VEHICLE/NON EMERGENCY
19	MOPED
20	PICKUP TRUCK
21	VAN
88	OTHER
99	UNKNOWN
22.05	OTHER LIGHT TRUCKS (10,000LBS (4,536KG))
23	(SPORT) UTILITY VEHICLE
24.88	LOW SPEED VEHICLE
25.88	OTHER BUS

26.88	ALL TERRAIN VEHICLE (ATV)
27.88	SNOWMOBILE
NaN	BLANK VALUE FROM ACRS

DRIVERLESSFLAG: describe the case in which the vehicle was without driver

Y	Yes
N	No
None	Blank value

FIREFLAG: describe the case in which there was fire in the place of the accident

Y	Yes
N	No
None	Blank value

NUMOCC: describe the number of occupants of the vehicle involved in the accident

PARKEDFLAG: describe if the vehicle involved in the accident was parked or not

Y	Yes
N	No
None	Blank value

SPEEDLIMIT: Numeric Speed Limit (no dictionary)

HITANDRUNFLAG: describe if the accident was a hit and run type

Y	Yes
N	No
None	Blank value

HAZMATSPILLFLAG: describe if the accident was a case of Hazardous Materials Incidents

Y	Yes
N	No
None	Blank value

TOWEDVEHICLECODE1: describe what the vehicle was towing while the accident (first)

00	Not Applicable
01	1 Semi Trailer
02	Local Codes
03	Traffic Control Signal Flag
04	2 Full Trailers
05	3 Trailers
06	Construction Zone Related Flag
07	Agency Code who authored the report
08	Code for local area information
09	Camper
10	Travel/Home Trailer
11	Mobile Home
12	Farm Equipment
88	Other
99	Unknown
NaN	Blank value

TOWEDVEHICLECODE2: describe what the vehicle was towing while the accident (second)

00	Not Applicable
01	1 Semi Trailer
02	Local Codes
03	Traffic Control Signal Flag
04	2 Full Trailers
05	3 Trailers
06	Construction Zone Related Flag
07	Agency Code who authored the report
08	Code for local area information
09	Camper
10	Travel/Home Trailer
11	Mobile Home
12	Farm Equipment
88	Other
99	Unknown
NaN	Blank value

TOWEDVEHICLECODE3: describe what the vehicle was towing while the accident (third)

00	Not Applicable
01	1 Semi Trailer
02	Local Codes
03	Traffic Control Signal Flag
04	2 Full Trailers
05	3 Trailers
06	Construction Zone Related Flag
07	Agency Code who authored the report
08	Code for local area information
09	Camper
10	Travel/Home Trailer
11	Mobile Home
12	Farm Equipment
88	Other
99	Unknown
NaN	Blank value

VEHPLATENUM: describe the plate code of the vehicle

VEHPLATESTATE: describe the plate state of the vehicle

VEHPLATEYEAR: describe the plate year of the vehicle

AREADAMAGEDCODEIMP1: describe the first more important damage area

00	Not Applicable
01	One o'clock
02	Two o'clock
03	Three o'clock
04	Four o'clock
05	Five o'clock
06	Six o'clock

07	Seven o'clock
08	Eight o'clock
09	Nine o'clock
10	Ten o'clock
11	Eleven o'clock
12	Twelve o'clock
13	License Class - State Specific
14	Commercial Drivers License
15	Left Side Front Quarter
16	Front Left Corner
17	Hood
18	Roof Top
19	Trunk
20	Windshield
21	Route Types - Maintained by SHA
22	Underside
23	Route Suffix - Maintained by SHA
88	Other
98	Log Mile - Maintained by SHA
99	Unknown
NaN	Blank value

AREADAMAGEDCODE2: describe the second more important damage area

00	Not Applicable
01	One o'clock
02	Two o'clock
03	Three o'clock
04	Four o'clock
05	Five o'clock
06	Six o'clock
07	Seven o'clock
08	Eight o'clock
09	Nine o'clock
10	Ten o'clock
11	Eleven o'clock
12	Twelve o'clock
13	License Class - State Specific
14	Commercial Drivers License
15	Left Side Front Quarter
16	Front Left Corner
17	Hood
18	Roof Top
19	Trunk
20	Windshield
21	Route Types - Maintained by SHA
22	Underside

23	Route Suffix - Maintained by SHA
88	Other
98	Log Mile - Maintained by SHA
99	Unknown
NaN	Blank value

AREADAMAGEDCODE3: describe the therd more important damage area

00	Not Applicable
01	One o'clock
02	Two o'clock
03	Three o'clock
04	Four o'clock
05	Five o'clock
06	Six o'clock
07	Seven o'clock
08	Eight o'clock
09	Nine o'clock
10	Ten o'clock
11	Eleven o'clock
12	Twelve o'clock
13	License Class - State Specific
14	Commercial Drivers License
15	Left Side Front Quarter
16	Front Left Corner
17	Hood
18	Roof Top
19	Trunk
20	Windshield
21	Route Types - Maintained by SHA
22	Underside
23	Route Suffix - Maintained by SHA
88	Other
98	Log Mile - Maintained by SHA
99	Unknown
NaN	Blank value

AREADAMAGEDCODEIMP1: describe the damage area (not specified in the dictionary the difference from the other one)

00	Not Applicable
01	One o'clock
02	Two o'clock
03	Three o'clock
04	Four o'clock
05	Five o'clock
06	Six o'clock
07	Seven o'clock
08	Eight o'clock
09	Nine o'clock

10	Ten o'clock
11	Eleven o'clock
12	Twelve o'clock
13	License Class - State Specific
14	Commercial Drivers License
15	Left Side Front Quarter
16	Front Left Corner
17	Hood
18	Roof Top
19	Trunk
20	Windshield
21	Route Types - Maintained by SHA
22	Underside
23	Route Suffix - Maintained by SHA
88	Other
98	Log Mile - Maintained by SHA
99	Unknown
NaN	Blank value

AREADAMAGEDCODEMAIN: describe the damage area (not specified in the dictionary the difference from the other one)

00	Not Applicable
01	One o'clock
02	Two o'clock
03	Three o'clock
04	Four o'clock
05	Five o'clock
06	Six o'clock
07	Seven o'clock
08	Eight o'clock
09	Nine o'clock
10	Ten o'clock
11	Eleven o'clock
12	Twelve o'clock
13	License Class - State Specific
14	Commercial Drivers License
15	Left Side Front Quarter
16	Front Left Corner
17	Hood
18	Roof Top
19	Trunk
20	Windshield
21	Route Types - Maintained by SHA
22	Underside
23	Route Suffix - Maintained by SHA
88	Other
98	Log Mile - Maintained by SHA

99	Unknown
NaN	Blank value

SEQEVENTCODE1: describe the first event happened during the accident

00	NOT APPLICABLE
01	OVERTURN ROLLOVER
02	FIRE EXPLOSION
03	IMMERSION
04	JACKKNIFE
05	CARGO EQUIPMENT LOSS OR SHIFT
06	EQUIPMENT FAILURE
07	SEPERATION OF UNITS
08	RAN OFF ROAD RIGHT
09	RAN OFF ROAD LEFT
10	CROSS MEDIAN
11	CROSS CENTERLINE
12	DOWNHILL RUNAWAY
13	FELL JUMPED FROM MOTOR VEHICLE
14	THROWN OR FALLING OBJECT
15	OTHER NON COLLISON
16	REENTERING ROADWAY
30	STRUCK PEDESTRAIN
31	STRUCK PEDALCYCLE
32	STRUCK RAILWAY VEHICLE
33	STRUCK ANIMAL
34	STRUCK MOTOR VEHICLE IN TRANSPORT
35	STRUCK PARK MOTOR VEHICLE
36	STRUCK BY FALLING SHIFT CARGO
37	STRUCK OTHER NON FIXED OBJECT
38	STRUCK BY MOTOR VEHICLE DEBRIS
60	IMPACT ATTENUATOR CRASH CUSHION
61	BRIDGE OVERHEAD STRUCTURE
62	BRIDGE PEIR OR SUPPORT
63	BRIDGE RAIL
64	CULVERT
65	CURB
66	DITCH
67	EMBANKMENT
68	GUARDRAIL FACE
69	GUARDRAIL END
70	CONCRETE TRAFFIC BARRIER
71	OTHER TRAFFIC BARRIER
72	TREE STANDING
73	UTLITY POLE LIGHT SUPPORT
74	TRAFFIC SIGN SUPPORT
75	TRAFFIC SIGNAL SUPPORT

76	OTHER POST POLE OR SUPPORT
77	FENCE
78	MAILBOX
79	BUILDING
80	OTHER FIXED OBJECT
81	CABLE BARRIER
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

SEQEVENTCODE2: describe the second event happened during the accident

00	NOT APPLICABLE
01	OVERTURN ROLLOVER
02	FIRE EXPLOSION
03	IMMERSION
04	JACKKNIFE
05	CARGO EQUIPMENT LOSS OR SHIFT
06	EQUIPMENT FAILURE
07	SEPERATION OF UNITS
08	RAN OFF ROAD RIGHT
09	RAN OFF ROAD LEFT
10	CROSS MEDIAN
11	CROSS CENTERLINE
12	DOWNHILL RUNAWAY
13	FELL JUMPED FROM MOTOR VEHICLE
14	THROWN OR FALLING OBJECT
15	OTHER NON COLLISON
16	REENTERING ROADWAY
30	STRUCK PEDESTRAIN
31	STRUCK PEDALCYCLE
32	STRUCK RAILWAY VEHICLE
33	STRUCK ANIMAL
34	STRUCK MOTOR VEHICLE IN TRANSPORT
35	STRUCK PARK MOTOR VEHICLE
36	STRUCK BY FALLING SHIFT CARGO
37	STRUCK OTHER NON FIXED OBJECT
38	STRUCK BY MOTOR VEHICLE DEBRIS
60	IMPACT ATTENUATOR CRASH CUSHION
61	BRIDGE OVERHEAD STRUCTURE
62	BRIDGE PEIR OR SUPPORT
63	BRIDGE RAIL
64	CULVERT
65	CURB
66	DITCH
67	EMBANKMENT
68	GUARDRAIL FACE

69	GUARDRAIL END
70	CONCRETE TRAFFIC BARRIER
71	OTHER TRAFFIC BARRIER
72	TREE STANDING
73	UTLITY POLE LIGHT SUPPORT
74	TRAFFIC SIGN SUPPORT
75	TRAFFIC SIGNAL SUPPORT
76	OTHER POST POLE OR SUPPORT
77	FENCE
78	MAILBOX
79	BUILDING
80	OTHER FIXED OBJECT
81	CABLE BARRIER
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

SEQEVENTCODE3: describe the third event happened during the accident

00	NOT APPLICABLE
01	OVERTURN ROLLOVER
02	FIRE EXPLOSION
03	IMMERSION
04	JACKKNIFE
05	CARGO EQUIPMENT LOSS OR SHIFT
06	EQUIPMENT FAILURE
07	SEPERATION OF UNITS
08	RAN OFF ROAD RIGHT
09	RAN OFF ROAD LEFT
10	CROSS MEDIAN
11	CROSS CENTERLINE
12	DOWNHILL RUNAWAY
13	FELL JUMPED FROM MOTOR VEHICLE
14	THROWN OR FALLING OBJECT
15	OTHER NON COLLISON
16	REENTERING ROADWAY
30	STRUCK PEDESTRAIN
31	STRUCK PEDALCYCLE
32	STRUCK RAILWAY VEHICLE
33	STRUCK ANIMAL
34	STRUCK MOTOR VEHICLE IN TRANSPORT
35	STRUCK PARK MOTOR VEHICLE
36	STRUCK BY FALLING SHIFT CARGO
37	STRUCK OTHER NON FIXED OBJECT
38	STRUCK BY MOTOR VEHICLE DEBRIS
60	IMPACT ATTENUATOR CRASH CUSHION
61	BRIDGE OVERHEAD STRUCTURE

62	BRIDGE PEIR OR SUPPORT
63	BRIDGE RAIL
64	CULVERT
65	CURB
66	DITCH
67	EMBANKMENT
68	GUARDRAIL FACE
69	GUARDRAIL END
70	CONCRETE TRAFFIC BARRIER
71	OTHER TRAFFIC BARRIER
72	TREE STANDING
73	UTLITY POLE LIGHT SUPPORT
74	TRAFFIC SIGN SUPPORT
75	TRAFFIC SIGNAL SUPPORT
76	OTHER POST POLE OR SUPPORT
77	FENCE
78	MAILBOX
79	BUILDING
80	OTHER FIXED OBJECT
81	CABLE BARRIER
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

SEQEVENTCODE4: describe the fourth event happened during the accident

00	NOT APPLICABLE
01	OVERTURN ROLLOVER
02	FIRE EXPLOSION
03	IMMERSION
04	JACKKNIFE
05	CARGO EQUIPMENT LOSS OR SHIFT
06	EQUIPMENT FAILURE
07	SEPERATION OF UNITS
08	RAN OFF ROAD RIGHT
09	RAN OFF ROAD LEFT
10	CROSS MEDIAN
11	CROSS CENTERLINE
12	DOWNHILL RUNAWAY
13	FELL JUMPED FROM MOTOR VEHICLE
14	THROWN OR FALLING OBJECT
15	OTHER NON COLLISON
16	REENTERING ROADWAY
30	STRUCK PEDESTRAIN
31	STRUCK PEDALCYCLE
32	STRUCK RAILWAY VEHICLE
33	STRUCK ANIMAL

34	STRUCK MOTOR VEHICLE IN TRANSPORT
35	STRUCK PARK MOTOR VEHICLE
36	STRUCK BY FALLING SHIFT CARGO
37	STRUCK OTHER NON FIXED OBJECT
38	STRUCK BY MOTOR VEHICLE DEBRIS
60	IMPACT ATTENUATOR CRASH CUSHION
61	BRIDGE OVERHEAD STRUCTURE
62	BRIDGE PEIR OR SUPPORT
63	BRIDGE RAIL
64	CULVERT
65	CURB
66	DITCH
67	EMBANKMENT
68	GUARDRAIL FACE
69	GUARDRAIL END
70	CONCRETE TRAFFIC BARRIER
71	OTHER TRAFFIC BARRIER
72	TREE STANDING
73	UTLITY POLE LIGHT SUPPORT
74	TRAFFIC SIGN SUPPORT
75	TRAFFIC SIGNAL SUPPORT
76	OTHER POST POLE OR SUPPORT
77	FENCE
78	MAILBOX
79	BUILDING
80	OTHER FIXED OBJECT
81	CABLE BARRIER
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

REMOVEDBY: describe the place from where the vehicle was removed by

REMOVEDTO: describe the place from where the vehicle was removed to

VEHOWNERPHONENUMBER: describe the phone number of the vehicle owner involved in the accident

VEHOWNEROTHERPHONE: describe the other phone number of the vehicle owner involved in the accident

VEHINSURER: describe the insurer of the vehicle involved in the accident

VEHPOLICYNUMBER: describe the policy number of the vehicle involved in the accident

VEHSPECIALFUNCTIONCODE: describe the function of the vehicle involved in the accident

00	NOT APPLICABLE
01	TAXI
02	VEHICLE USED AS SCHOOL BUS
03	VEHICLE USED AS OTHER BUS
04	MILITARY
05	POLICE

06	AMBULANCE
07	FIRE TRUCK
99	UNKNOWN
88	OTHER
08	WORK ZONE EQUIPMENT
NaN	BLANK VALUE FROM ACRS

EMERGENCYUSEFLAG: describe if emergency vehicle went to the accident place (no dictionary, code N, Y, U, None)

CARRIERCLASSCODE: describe the carrier class involved in the accident

00	NOT APPLICABLE
01	INTERSTATE CARRIER
02	INTRASTATE CARRIER
03	NOT IN COMMERCE GOVERNMENT
04	NOT IN COMMERCE OTHER TRUCK
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

CARRIERSTREETADDRESS: describe the carriers street address

CARRIERCITY: describe the carriers city

CARRIERZIPCODE: describe the carriers zip code

CVCONFIGCODE: describe the commercial vehicle configuration involved in the accident

01	VEHICLE 10000 POUNDS OR LESS PLACARDED FOR HAZ MAT
02	SINGLE-UNIT TRUCK 2-AXLE GREATER THAN 10,000
03	SINGLE-UNIT TRUCK 3 OR MORE AXLES
04	TRUCK PULLING TRAILERS
05	TRUCK TRACTOR BOBTAIL
06	TRUCK TRACTOR SEMI-TRAILER
07	TRUCK TRACTOR DOUBLE
08	TRUCK TRACTOR TRIPLE
09	TRUCK MORE THAN 10000 CANNOT CLASSIFY
10	BUS LARGE VAN SEATS 9 TO 15 OCCUPANTS INCL DRIVER
11	BUS SEATS FOR MORE THAN 15 OCCUPANTS INCL DRIVER
99	UNKNOWN
00	NOT APPLICABLE
88	OTHER
NaN	BLANK VALUE FROM ACRS

BUSUSECODE: describe the type of bus involved in the accident

00	NOT APPLICABLE
01	SCHOOL
02	TRANSIT
03	COMMUTER
04	INTERCITY
05	CHARTER
06	TOUR

07	SHUTTLE
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

HZMNAME: describe the Hazmat Number (no dictionary, all NaN value)

PLACARDVISIBLEFLAG: (no dictionary)

VEHICLEWEIGHTCODE: describe the weight of the vehicle involved in the accident

01	10000 LBS OR LESS
02	10001 TO 26000 LBS
03	MORE THAN 26000 LBS
00	NOT APPLICABLE
88	OTHER
99	UNKNOWN
NaN	BLANK VALUE FROM ACRS

VEHOWNERSTATECODE: describe the state code of the vehicle owner involved in the accident

6.4 APPENDIX 4 – Ordered Logit Model code

Model: Ordered Logit (2016)

References

- Biogeme [manual] (Bierlaire, M. (2018). PandasBiogeme: a short introduction. Technical report TRANSP-OR 181219. Transport and Mobility Laboratory, ENAC, EPFL.)
- Train, K. E. (2009). Discrete choice methods with simulation. Cambridge university press

Libraries

Libraries needed

- Pandas for data manipulation
- Biogeme to use model estimation

In [1]:
```
import pandas as pd
import biogeme.database as db
import biogeme.biogeme as bio
import biogeme.distributions as dist
```

Comand to show more row and columns

In [2]:
```
pd.set_option('display.max_columns', 999999)
pd.set_option('display.max_rows', 999999)
```

In [3]:
```
PGEng = sal.create_engine('postgresql://*******Private Address information*******')

with PGEng.connect() as PGConn, PGConn.begin():
    select_stmnt = 'SELECT COALESCE("COORDINATES",ST_GeomFromText(\'POINT EMPTY\',4326)) as "SANITIZED_COORDINATES",* FROM "Clea
    CrashDF = gpd.GeoDataFrame.from_postgis(select_stmnt, PGConn, geom_col='SANITIZED_COORDINATES' )
```

Data manipulation

1. Delete the underscore from the columns name to avoid issues during the computation

In [4]:
```
CrashDF.columns = CrashDF.columns.str.replace('_', '')
```

In [5]:
```
CrashDF.shape
```

Out[5]:
```
(316820, 231)
```

2. New dataset with only a first selection of variables replacing also the null value with "-999"

In [6]:
```
dfmyfloat = CrashDF.fillna(-999).loc[:, ['INJSEVERCODE','SEX','AGE','CONTRIBCODE1','CONTRIBCODE2','CONTRIBCODE3',
                                         'CONTRIBCODE4','CONDITIONCODE','BAC','ALCOTESTCODE','DRUGTESTCODE','DRUGTESTRESULTFLAG',
                                         'EQUIPPROBCODE','ALCODRUGIMPAIREDFLAG','SAFEQUIPCODE','VEHBODYTYPECODE','EJECTCODE',
                                         'HARMEVENTCODE','HARMEVENTCODE1','HARMEVENTCODE2','RDDIVCODE','LANENUMBER',
                                         'SURFCONDCODE','TRAFFICCONTROLCODE','SIGNALFLAG','COLLISIONTYPECODE',
                                         'INTERSECTIONTYPECODE','JUNCTIONCODE','LIGHTCODE','FIXOBJCODE','WEATHERCODE']]
```

3. Procedure to replace alphanumeric in numerical value

In [7]:
```
print(dfmyfloat['ALCODRUGIMPAIREDFLAG'].unique())
print(dfmyfloat['SIGNALFLAG'].unique())
print(dfmyfloat['DRUGTESTRESULTFLAG'].unique())
```

```
['N' -999 'A' 'D' 'B']
['N' 'Y']
['X' -999 'U' 'A' 'P' 'N']
```

In [8]:
```
dfmyfloat = dfmyfloat.replace({'ALCODRUGIMPAIREDFLAG':{'A':1,'B':3,'D':2,'N':0}})
dfmyfloat = dfmyfloat.replace({'SIGNALFLAG':{'N':0,'Y':1}})
dfmyfloat = dfmyfloat.replace({'DRUGTESTRESULTFLAG':{'P':1,'N':0,'X':2,'U':3,'A':4}})
```

4.Creation of new variables

In [9]:
```
df = dfmyfloat

temp1 = df.CONTRIBCODE1.isin ([1])|df.CONTRIBCODE2.isin ([1])|df.CONTRIBCODE3.isin ([1])|df.CONTRIBCODE4.isin ([1])
temp2 = df.CONDITIONCODE.isin ([3]) | df.ALCODRUGIMPAIREDFLAG.isin ([2]) | df.DRUGTESTCODE.isin ([2]) |
        df.DRUGTESTRESULTFLAG.isin ([1])
temp3 = temp1 | temp2
df['DRUGEFFECT']= temp3*1

temp4 = df.CONTRIBCODE1.isin ([2])|df.CONTRIBCODE2.isin ([2])|df.CONTRIBCODE3.isin ([2])|df.CONTRIBCODE4.isin ([2])
temp5 = df.CONDITIONCODE.isin ([2]) | df.ALCODRUGIMPAIREDFLAG.isin ([1]) | df.ALCOTESTCODE.isin ([2])
#Alcohol Blood Content greater than 5
temp6 = df.BAC>6
```

```
temp7 = temp4 | temp5 |temp6
df['ALCHOLEFFECT']= temp7*1

temp8 = df.CONTRIBCODE1.isin ([3])|df.CONTRIBCODE2.isin ([3])|df.CONTRIBCODE3.isin ([3])|df.CONTRIBCODE4.isin ([3])
df['MEDICINEEFFECT']= temp8*1

temp9 = df.CONTRIBCODE1.isin ([4])|df.CONTRIBCODE2.isin ([4])|df.CONTRIBCODE3.isin ([4])|df.CONTRIBCODE4.isin ([4])
temp10 = df.CONDITIONCODE.isin ([10]) | df.ALCODRUGIMPAIREDFLAG.isin ([3])
temp11 = temp9 | temp10
df['DAMEFFECT']= temp11*1



df['EJTRAP']= ((dfmyfloat.EJECTCODE == 2) |(dfmyfloat.EJECTCODE == 3) |(dfmyfloat.EJECTCODE == 4))
df.EJTRAP= df.EJTRAP*1

t1 = df.SAFEQUIPCODE.isin ([11,12,13,32])
t2 = df.EQUIPPROBCODE.isin ([11,13,42,43,44,45,46])
t3 = ~t2
df['SEATBELT'] = t1 & t3
df.SEATBELT = df.SEATBELT*1

te1 = df.VEHBODYTYPECODE.isin ([2])*1
te2 = df.VEHBODYTYPECODE.isin ([23])*2
te3 = df.VEHBODYTYPECODE.isin ([20])*3
te4 = df.VEHBODYTYPECODE.isin ([21])*4
te5 = df.VEHBODYTYPECODE.isin ([22.05])*5
te6 = ~df.VEHBODYTYPECODE.isin ([2,23,20,21,22.05])*6
df['VEHBODY'] = te1 + te2 + te3 + te4 + te5 + te6

tem1 = df.HARMEVENTCODE.isin ([11])*1
tem2 = df.HARMEVENTCODE1.isin ([11])*1
tem3 = df.HARMEVENTCODE2.isin ([11])*1
df['ROLLEDOVER'] = tem1 | tem2 | tem3

a1 = df.RDDIVCODE.isin ([1,2,5])*1
a2 =df.RDDIVCODE.isin ([3])*2
a3 =df.RDDIVCODE.isin ([4])*3
a4 =~df.RDDIVCODE.isin ([1,2,3,4,5])*4
df['ROADTYPE'] = a1 + a2 + a3 + a4

b1 = df.SURFCONDCODE.isin ([2])*1
b2 = df.SURFCONDCODE.isin ([1])*2
b3 = df.SURFCONDCODE.isin ([3,4])*3
b4 =~df.SURFCONDCODE.isin ([1,2,3,4])*4
df['SURFCOND'] = b1 + b2 + b3 + b4
```

```
g1=df.TRAFFICCONTROLCODE.isin ([1])*1 # no control
g2=df.TRAFFICCONTROLCODE.isin ([3])*2 # traffic signal
g3=df.TRAFFICCONTROLCODE.isin ([6])*3 # stop signal
g4=df.TRAFFICCONTROLCODE.isin ([7])*4 # yeld signal
g5=~df.TRAFFICCONTROLCODE.isin ([1,3,6,7])*5 # other
df['TRAFFICCONTROL'] = g1 + g2 + g3 + g4 + g5

c1 = df.COLLISIONTYPECODE.isin ([1,2])*1
c2 = df.COLLISIONTYPECODE.isin ([3,4,5])*2
c3 = df.COLLISIONTYPECODE.isin ([12,13,14])*3
c4 = df.COLLISIONTYPECODE.isin ([7])*4
c5 = df.COLLISIONTYPECODE.isin ([6])*5
c6 =~df.COLLISIONTYPECODE.isin ([1,2,3,4,5,6,7,12,13,14])*6
df['COLLTYPE'] = c1 + c2 + c3 + c4 +c5 +c6

d1 = df.INTERSECTIONTYPECODE.isin ([1,2,3,4,5,6])*1
df['INTERSECTIONTYPE'] = d1

e1 = df.LIGHTCODE.isin ([1,3])*1
e2 = df.LIGHTCODE.isin ([4])*2
e3 = df.LIGHTCODE.isin ([5.02])*3
e4 = df.LIGHTCODE.isin ([6.02])*4
e5 = ~df.LIGHTCODE.isin ([1,3,4,5.02,6.02])*5
df['LIGHTCOND'] = e1 + e2 + e3 + e4 + e5

f1 = df.FIXOBJCODE.isin ([1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22])*1
df['FIXOBJECT'] = f1

df.head()
```

Out[9]:

	INJSEVERCODE	SEX	AGE	CONTRIBCODE1	CONTRIBCODE2	CONTRIBCODE3	CONTRIBCODE4	CONDITIONCODE	BAC	ALCOTESTCODE	DRUGTESTC
0											
1											
2											
3											
4											

Sensible data censored

4. Final step selection of variables in a new Dataset

In [10]:
```
dfmy = df.loc[:,['INJSEVERCODE','SEX','AGE','DRUGEFFECT','ALCHOLEFFECT','MEDICINEEFFECT','DAMEFFECT','EJTRAP','SEATBELT',
                 'VEHBODY','ROLLEDOVER','ROADTYPE','TRAFFICCONTROL','SURFCOND','COLLTYPE','INTERSECTIONTYPE','LIGHTCOND',
                 'FIXOBJECT']]
dfmy.head(10)
```

Out[10]:

	INJSEVERCODE	SEX	AGE	DRUGEFFECT	ALCHOLEFFECT	MEDICINEEFFECT	DAMEFFECT	EJTRAP	SEATBELT	VEHBODY	ROLLEDOVER	ROADTYPE
0	Sensible data censored											
1												
2												
3												
4												
5												
6												
7												
8												
9												

5. We can count how many time a value is present in a particular variable

In [11]:
```
dfmy['INJSEVERCODE'].value_counts()
```

Out[11]:
```
1    260538
3     26594
2     25486
4      3660
5       542
Name: INJSEVERCODE, dtype: int64
```

6. How to delete useless value

In [12]:
```
dfmy = dfmy[~dfmy['SEX'].isin([99])]
dfmy = dfmy[~dfmy['AGE'].isin([999])]
```

9. Definition of dummy variables

In [13]:
```
dfdummy = pd.get_dummies(dfmy, columns=['SEX'], drop_first=True) #True instead 0 if i want to base the Dummy with the male type

dfdummy['AGECLASS'] =pd.cut(dfdummy['AGE'], [0, 15, 25,45,60,75,120,999], labels=['0', '1', '2','3','4','5','6'])
dfdummy = pd.get_dummies(dfdummy, columns=['AGECLASS'], drop_first=True) #based in class 0

dfdummy = pd.get_dummies(dfdummy, columns=['DRUGEFFECT'], drop_first=True) #based on not under effect
dfdummy = pd.get_dummies(dfdummy, columns=['ALCHOLEFFECT'], drop_first=True) #based on not under effect
dfdummy = pd.get_dummies(dfdummy, columns=['MEDICINEEFFECT'], drop_first=True) #based on not under effect
dfdummy = pd.get_dummies(dfdummy, columns=['DAMEFFECT'], drop_first=True) #based on not under effect

dfdummy = pd.get_dummies(dfdummy, columns=['EJTRAP'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['SEATBELT'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['VEHBODY'], drop_first=True) #based on car

dfdummy = pd.get_dummies(dfdummy, columns=['ROLLEDOVER'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['ROADTYPE'], drop_first=True) #based on 1 way 2 way undevided

dfdummy = pd.get_dummies(dfdummy, columns=['TRAFFICCONTROL'], drop_first=True) #based on No control

dfdummy = pd.get_dummies(dfdummy, columns=['SURFCOND'], drop_first=True) #based on dry

dfdummy = pd.get_dummies(dfdummy, columns=['COLLTYPE'], drop_first=True) #based on head on

dfdummy = pd.get_dummies(dfdummy, columns=['INTERSECTIONTYPE'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['LIGHTCOND'], drop_first=True) #based on day/dark lighting

dfdummy = pd.get_dummies(dfdummy, columns=['FIXOBJECT'], drop_first=True) #based on NO


dfdummy.columns = dfdummy.columns.str.replace('_', '')
dfdummy.head(10)  #1=male 2 =female
```

Out[13]:

	INJSEVERCODE	AGE	SEX2	AGECLASS1	AGECLASS2	AGECLASS3	AGECLASS4	AGECLASS5	AGECLASS6	DRUGEFFECT1	ALCHOLEFFECT1	MEDICII
0	Sensible data censored											
1												
2												
3												
4												
5												
6												
7												
8												

Model estimation

10. We need to transform the data frame into a database for Biogeme

In [14]:
```
database = db.Database('MasterView', dfdummy)
```

11. We need to storage the name of the variable into a file in order to be available for the Software computations

In [15]:
```
from headers import *
```

12. We define the main parameters that will be estimate by the model

In [16]:
```
B_SEX2 = Beta('B_SEX2',0,None,None,0)

B_AGE1 = Beta('B_AGE1',0,None,None,0)
B_AGE2 = Beta('B_AGE2',0,None,None,0)
B_AGE3 = Beta('B_AGE3',0,None,None,0)
B_AGE4 = Beta('B_AGE4',0,None,None,0)
B_AGE5 = Beta('B_AGE5',0,None,None,0)
B_AGE6 = Beta('B_AGE6',0,None,None,1)

B_DRUGEFFECT1 = Beta('B_DRUGEFFECT1',0,None,None,0)
B_ALCHOLEFFECT1 = Beta('B_ALCHOLEFFECT1',0,None,None,0)
B_MEDICINEEFFECT1 = Beta('B_MEDICINEEFFECT1',0,None,None,0)
B_DAMEFFECT1 = Beta('B_DAMEFFECT1',0,None,None,0)

B_EJTRAP1 = Beta('B_EJTRAP1',0,None,None,0)

B_SEATBELT1 = Beta('B_SEATBELT1',0,None,None,0)

B_VEHBODY2 = Beta('B_VEHBODY2',0,None,None,0)
B_VEHBODY3 = Beta('B_VEHBODY3',0,None,None,0)
B_VEHBODY4 = Beta('B_VEHBODY4',0,None,None,0)
B_VEHBODY5 = Beta('B_VEHBODY5',0,None,None,0)
B_VEHBODY6 = Beta('B_VEHBODY6',0,None,None,1)

B_ROLLEDOVER1 = Beta('B_ROLLEDOVER1',0,None,None,0)

B_ROADTYPE2 = Beta('B_ROADTYPE2',0,None,None,0)
B_ROADTYPE3 = Beta('B_ROADTYPE3',0,None,None,0)
B_ROADTYPE4 = Beta('B_ROADTYPE4',0,None,None,1)

B_TRAFFICCONTROL2 = Beta('B_TRAFFICCONTROL2',0,None,None,0)
B_TRAFFICCONTROL3 = Beta('B_TRAFFICCONTROL3',0,None,None,0)
B_TRAFFICCONTROL4 = Beta('B_TRAFFICCONTROL4',0,None,None,0)
B_TRAFFICCONTROL5 = Beta('B_TRAFFICCONTROL5',0,None,None,1)

B_PERSONTYPE1 = Beta('B_PERSONTYPE1',0,None,None,0)
B_PERSONTYPE2 = Beta('B_PERSONTYPE2',0,None,None,0)
B_SURFCOND2 = Beta('B_SURFCOND2',0,None,None,0)
B_SURFCOND3 = Beta('B_SURFCOND3',0,None,None,0)
B_SURFCOND4 = Beta('B_SURFCOND4',0,None,None,1)

B_COLLTYPE2 = Beta('B_COLLTYPE2',0,None,None,0)
B_COLLTYPE3 = Beta('B_COLLTYPE3',0,None,None,0)
B_COLLTYPE4 = Beta('B_COLLTYPE4',0,None,None,0)
B_COLLTYPE5 = Beta('B_COLLTYPE5',0,None,None,0)
B_COLLTYPE6 = Beta('B_COLLTYPE6',0,None,None,1)

B_INTERSECTIONTYPE1 = Beta('B_INTERSECTIONTYPE1',0,None,None,0)

B_LIGHTCOND2 = Beta('B_LIGHTCOND2',0,None,None,0)
B_LIGHTCOND3 = Beta('B_LIGHTCOND3',0,None,None,0)
B_LIGHTCOND4 = Beta('B_LIGHTCOND4',0,None,None,0)
B_LIGHTCOND5 = Beta('B_LIGHTCOND5',0,None,None,1)

B_FIXOBJECT1 = Beta('B_FIXOBJECT1',0,None,None,0)
```

13. Also the model need other kind of parameters like the follow:

In [17]:
```
tau1 = Beta('tau1',-1,None,None,0)
delta2 = Beta('delta2',2,None,None,0)
delta3 = Beta('delta3',2,None,None,0)
delta4 = Beta('delta4',2,None,None,0)
```

In [18]:
```
tau2 = tau1 + delta2
tau3 = tau2 + delta3
tau4 = tau3 + delta4
```

14. Definition of the Utility functions

In [19]:
```
U = B_SEX2*SEX2+B_AGE1*AGECLASS1+B_AGE2*AGECLASS2+B_AGE3*AGECLASS3+B_AGE4*AGECLASS4+B_AGE5*AGECLASS5+B_AGE6*AGECLASS6+
    B_DRUGEFFECT1*DRUGEFFECT1+B_ALCHOLEFFECT1*ALCHOLEFFECT1+B_MEDICINEEFFECT1*MEDICINEEFFECT1+B_DAMEFFECT1*DAMEFFECT1+
    B_EJTRAP1*EJTRAP1+B_SEATBELT1*SEATBELT1+B_VEHBODY2*VEHBODY2+B_VEHBODY3*VEHBODY3+B_VEHBODY4*VEHBODY4+B_VEHBODY5*VEHBODY5+
    B_VEHBODY6*VEHBODY6+B_ROLLEDOVER1*ROLLEDOVER1+B_ROADTYPE2*ROADTYPE2+B_ROADTYPE3*ROADTYPE3+B_ROADTYPE4*ROADTYPE4+
    B_TRAFFICCONTROL2*TRAFFICCONTROL2+B_TRAFFICCONTROL3*TRAFFICCONTROL3+B_TRAFFICCONTROL4*TRAFFICCONTROL4+
    B_TRAFFICCONTROL5*TRAFFICCONTROL5+B_SURFCOND2*SURFCOND2+B_SURFCOND3*SURFCOND3+B_SURFCOND4*SURFCOND4+
    B_COLLTYPE2*COLLTYPE2+B_COLLTYPE3*COLLTYPE3+B_COLLTYPE4*COLLTYPE4+B_COLLTYPE5*COLLTYPE5+B_COLLTYPE6*COLLTYPE6+
    B_INTERSECTIONTYPE1*INTERSECTIONTYPE1+B_LIGHTCOND2*LIGHTCOND2+B_LIGHTCOND3*LIGHTCOND3+B_LIGHTCOND4*LIGHTCOND4+
    B_LIGHTCOND5*LIGHTCOND5+B_FIXOBJECT1*FIXOBJECT1
```

15. Assigning the probability for each choice

In [20]:
```
ChoiceProba = {
    1: 1-dist.logisticcdf(U-tau1),
    2: dist.logisticcdf(U-tau1)- dist.logisticcdf(U-tau2),
    3: dist.logisticcdf(U-tau2)- dist.logisticcdf(U-tau3),
    4: dist.logisticcdf(U-tau3)- dist.logisticcdf(U-tau4),
    5: dist.logisticcdf(U-tau4) }
```

16.The comands used by Biogeme for the Ordered Logit are the following

In [21]:
```
logprob = log(Elem(ChoiceProba,INJSEVERCODE))
```

In [22]:
```
biogeme = bio.BIOGEME(database,logprob,numberOfThreads=12)
biogeme.modelName = "OrderedLogitTEST"
results = biogeme.estimate()
print("Results=",results)
```

```
Results=
Results for model [OrderedLogi2016]
Output file (HTML):                     OrderedLogi2016.html
Nbr of parameters:              39
Sample size:                    267187
Excluded data:                  0
Init log likelihood:            -351225.3
Final log likelihood:           -174199.2
Likelihood ratio test:          354052.4
Rho square:                     0.504
Rho bar square:                 0.504
Akaike Information Criterion:   348476.3
Bayesian Information Criterion: 348885.6
Final gradient norm:            6.236117
B_AGE1          : 0.23[0.0222 10.4 0][0.022 10.5 0]
B_AGE2          : 0.323[0.0219 14.8 0][0.0216 14.9 0]
B_AGE3          : 0.458[0.0232 19.7 0][0.0231 19.8 0]
B_AGE4          : 0.525[0.0261 20.1 0][0.026 20.2 0]
B_AGE5          : 0.541[0.0349 15.5 0][0.035 15.4 0]
B_ALCHOLEFFECT1: 0.00828[0.0234 0.354 0.723][0.0243 0.341 0.733]
```

In [29]:
```
results.getEstimatedParameters()
```

Out[29]:

	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value
B_AGE1	0.230257	0.022247	10.350109	0.000000e+00	0.021962	10.484268	0.000000e+00
B_AGE2	0.323123	0.021877	14.769922	0.000000e+00	0.021626	14.941146	0.000000e+00
B_AGE3	0.457586	0.023218	19.708609	0.000000e+00	0.023071	19.833952	0.000000e+00
B_AGE4	0.525231	0.026081	20.138525	0.000000e+00	0.025987	20.211599	0.000000e+00
B_AGE5	0.540735	0.034864	15.510019	0.000000e+00	0.035028	15.437088	0.000000e+00
B_ALCHOLEFFECT1	0.008278	0.023389	0.353934	7.233885e-01	0.024256	0.341288	7.328865e-01
B_COLLTYPE2	-0.099492	0.011867	-8.383909	0.000000e+00	0.011864	-8.386359	0.000000e+00
B_COLLTYPE3	-0.034757	0.041125	-0.845145	3.980299e-01	0.041755	-0.832408	4.051785e-01
B_COLLTYPE4	-0.732503	0.025690	-28.512862	0.000000e+00	0.025600	-28.612884	0.000000e+00
B_COLLTYPE5	-0.306694	0.048759	-6.289984	3.174998e-10	0.049178	-6.236446	4.476239e-10
B_DAMEFFECT1	0.351347	0.079340	4.428383	9.494216e-06	0.085336	4.117231	3.834513e-05
B_DRUGEFFECT1	0.408964	0.051941	7.873664	3.552714e-15	0.056200	7.276982	3.412826e-13
B_EJTRAP1	3.377048	0.034921	96.704608	0.000000e+00	0.042300	79.836078	0.000000e+00
B_FIXOBJECT1	0.552539	0.013337	41.428770	0.000000e+00	0.013816	39.991528	0.000000e+00
B_INTERSECTIONTYPE1	0.192669	0.013295	14.491466	0.000000e+00	0.013370	14.411065	0.000000e+00
B_LIGHTCOND2	0.106237	0.020446	5.195874	2.037600e-07	0.021015	5.055357	4.295860e-07
B_LIGHTCOND3	0.101322	0.034190	2.963514	3.041479e-03	0.034282	2.955511	3.121515e-03
B_LIGHTCOND4	-0.087656	0.030735	-2.852003	4.344466e-03	0.030877	-2.838859	4.527512e-03
B_MEDICINEEFFECT1	0.046277	0.083853	0.551883	5.810283e-01	0.091518	0.505660	6.130951e-01
B_PERSONTYPE1	0.429113	0.013219	32.462678	0.000000e+00	0.013251	32.383282	0.000000e+00
B_PERSONTYPE2	2.862117	0.029375	97.432770	0.000000e+00	0.028196	101.508547	0.000000e+00
B_ROADTYPE2	0.141494	0.016479	8.586301	0.000000e+00	0.016626	8.510300	0.000000e+00
B_ROADTYPE3	0.183323	0.011189	16.383811	0.000000e+00	0.011317	16.199430	0.000000e+00
B_ROLLEDOVER1	1.029977	0.031897	32.290418	0.000000e+00	0.035002	29.426483	0.000000e+00
B_SEATBELT1	-0.027158	0.013580	-1.999930	4.550781e-02	0.013824	-1.964507	4.947131e-02
B_SEX2	0.422697	0.010126	41.744887	0.000000e+00	0.010146	41.661938	0.000000e+00
B_SURFCOND2	0.010877	0.013397	0.811919	4.168380e-01	0.013580	0.800983	4.231417e-01
B_SURFCOND3	0.224228	0.032949	6.805271	1.008593e-11	0.034017	6.591587	4.351497e-11

B_TRAFFICCONTROL2	0.129948	0.014567	8.920894	0.000000e+00	0.014634	8.879782	0.000000e+00
B_TRAFFICCONTROL3	0.122646	0.020441	5.999987	1.973338e-09	0.020478	5.989150	2.109400e-09
B_TRAFFICCONTROL4	0.069645	0.050821	1.370404	1.705607e-01	0.050913	1.367914	1.713390e-01
B_VEHBODY2	-0.075109	0.015553	-4.829129	1.371316e-06	0.015658	-4.796779	1.612374e-06
B_VEHBODY3	-0.245419	0.022499	-10.907851	0.000000e+00	0.022713	-10.805214	0.000000e+00
B_VEHBODY4	-0.159536	0.027484	-5.804587	6.452469e-09	0.027513	-5.798588	6.687549e-09
B_VEHBODY5	-0.388195	0.058237	-6.665816	2.631984e-11	0.058861	-6.595126	4.248957e-11
delta2	0.804405	0.004926	163.295875	0.000000e+00	0.004960	162.179111	0.000000e+00
delta3	2.340841	0.016052	145.824276	0.000000e+00	0.015376	152.237931	0.000000e+00
delta4	2.187880	0.041378	52.874859	0.000000e+00	0.040821	53.596905	0.000000e+00
tau1	2.310857	0.024391	94.743811	0.000000e+00	0.023850	96.889594	0.000000e+00

6.5 APPENDIX 5 - Example of Market Share code

Scenario 14 related to the prediction where all people are Ejected/tapped 2017.

In [1]:
```
import pandas as pd
import biogeme.database as db
import biogeme.biogeme as bio
import biogeme.distributions as dist
```

In [2]:
```
pd.set_option('display.max_columns', 999999)
pd.set_option('display.max_rows', 999999)
```

In [3]:
```
PGEng = sal.create_engine('postgresql://*******Private Address information*******')

with PGEng.connect() as PGConn, PGConn.begin():
    select_stmnt = 'SELECT COALESCE("COORDINATES",ST_GeomFromText(\'POINT EMPTY\',4326)) as "SANITIZED_COORDINATES",* FROM "Clea
    CrashDF = gpd.GeoDataFrame.from_postgis(select_stmnt, PGConn, geom_col='SANITIZED_COORDINATES' )
```

In [4]:
```
CrashDF.columns = CrashDF.columns.str.replace('_', '')
```

In [5]:
```
CrashDF.shape
```

Out[5]:
```
(316820, 231)
```

In [6]:
```
dfmyfloat = CrashDF.fillna(-999).loc[:, ['INJSEVERCODE','SEX','AGE','CONTRIBCODE1','CONTRIBCODE2','CONTRIBCODE3',
                                         'CONTRIBCODE4','CONDITIONCODE','BAC','ALCOTESTCODE','DRUGTESTCODE','DRUGTESTRESULTFLAG',
                                         'EQUIPPROBCODE','ALCODRUGIMPAIREDFLAG','SAFEQUIPCODE','VEHBODYTYPECODE','EJECTCODE',
                                         'HARMEVENTCODE','HARMEVENTCODE1','HARMEVENTCODE2','RDDIVCODE','LANENUMBER',
                                         'SURFCONDCODE','TRAFFICCONTROLCODE','SIGNALFLAG','COLLISIONTYPECODE',
                                         'INTERSECTIONTYPECODE','JUNCTIONCODE','LIGHTCODE','FIXOBJCODE','WEATHERCODE']]
```

In [7]:
```
print(dfmyfloat['ALCODRUGIMPAIREDFLAG'].unique())
print(dfmyfloat['SIGNALFLAG'].unique())
print(dfmyfloat['DRUGTESTRESULTFLAG'].unique())
```

```
['N' -999 'A' 'D' 'B']
['N' 'Y']
['X' -999 'U' 'A' 'P' 'N']
```

In [8]:
```
dfmyfloat = dfmyfloat.replace({'ALCODRUGIMPAIREDFLAG':{'A':1,'B':3,'D':2,'N':0}})
dfmyfloat = dfmyfloat.replace({'SIGNALFLAG':{'N':0,'Y':1}})
dfmyfloat = dfmyfloat.replace({'DRUGTESTRESULTFLAG':{'P':1,'N':0,'X':2,'U':3,'A':4}})
```

In [9]:
```
df = dfmyfloat

temp1 = df.CONTRIBCODE1.isin ([1])|df.CONTRIBCODE2.isin ([1])|df.CONTRIBCODE3.isin ([1])|df.CONTRIBCODE4.isin ([1])
temp2 = df.CONDITIONCODE.isin ([3]) | df.ALCODRUGIMPAIREDFLAG.isin ([2]) | df.DRUGTESTCODE.isin ([2]) |
        df.DRUGTESTRESULTFLAG.isin ([1])
temp3 = temp1 | temp2
df['DRUGEFFECT']= temp3*1

temp4 = df.CONTRIBCODE1.isin ([2])|df.CONTRIBCODE2.isin ([2])|df.CONTRIBCODE3.isin ([2])|df.CONTRIBCODE4.isin ([2])
temp5 = df.CONDITIONCODE.isin ([2]) | df.ALCODRUGIMPAIREDFLAG.isin ([1]) | df.ALCOTESTCODE.isin ([2])
#Alcohol Blood Content greater than 5
temp6 = df.BAC>6
temp7 = temp4 | temp5 |temp6
df['ALCHOLEFFECT']= temp7*1

temp8 = df.CONTRIBCODE1.isin ([3])|df.CONTRIBCODE2.isin ([3])|df.CONTRIBCODE3.isin ([3])|df.CONTRIBCODE4.isin ([3])
df['MEDICINEEFFECT']= temp8*1

temp9 = df.CONTRIBCODE1.isin ([4])|df.CONTRIBCODE2.isin ([4])|df.CONTRIBCODE3.isin ([4])|df.CONTRIBCODE4.isin ([4])
temp10 = df.CONDITIONCODE.isin ([10]) | df.ALCODRUGIMPAIREDFLAG.isin ([3])
temp11 = temp9 | temp10
df['DAMEFFECT']= temp11*1



df['EJTRAP']= ((dfmyfloat.EJECTCODE == 2) |(dfmyfloat.EJECTCODE == 3) |(dfmyfloat.EJECTCODE == 4))
df.EJTRAP= df.EJTRAP*1

t1 = df.SAFEQUIPCODE.isin ([11,12,13,32])
t2 = df.EQUIPPROBCODE.isin ([11,13,42,43,44,45,46])
t3 = ~t2
df['SEATBELT'] = t1 & t3
df.SEATBELT = df.SEATBELT*1
```

```
te1 = df.VEHBODYTYPECODE.isin ([2])*1
te2 = df.VEHBODYTYPECODE.isin ([23])*2
te3 = df.VEHBODYTYPECODE.isin ([20])*3
te4 = df.VEHBODYTYPECODE.isin ([21])*4
te5 = df.VEHBODYTYPECODE.isin ([22.05])*5
te6 = ~df.VEHBODYTYPECODE.isin ([2,23,20,21,22.05])*6
df['VEHBODY'] = te1 + te2 + te3 + te4 + te5 + te6

tem1 = df.HARMEVENTCODE.isin ([11])*1
tem2 = df.HARMEVENTCODE1.isin ([11])*1
tem3 = df.HARMEVENTCODE2.isin ([11])*1
df['ROLLEDOVER'] = tem1 | tem2 | tem3

a1 = df.RDDIVCODE.isin ([1,2,5])*1
a2 =df.RDDIVCODE.isin ([3])*2
a3 =df.RDDIVCODE.isin ([4])*3
a4 =~df.RDDIVCODE.isin ([1,2,3,4,5])*4
df['ROADTYPE'] = a1 + a2 + a3 + a4

b1 = df.SURFCONDCODE.isin ([2])*1
b2 = df.SURFCONDCODE.isin ([1])*2
b3 = df.SURFCONDCODE.isin ([3,4])*3
b4 =~df.SURFCONDCODE.isin ([1,2,3,4])*4
df['SURFCOND'] = b1 + b2 + b3 + b4


g1=df.TRAFFICCONTROLCODE.isin ([1])*1 # no control
g2=df.TRAFFICCONTROLCODE.isin ([3])*2 # traffic signal
g3=df.TRAFFICCONTROLCODE.isin ([6])*3 # stop signal
g4=df.TRAFFICCONTROLCODE.isin ([7])*4 # yeld signal
g5=~df.TRAFFICCONTROLCODE.isin ([1,3,6,7])*5 # other
df['TRAFFICCONTROL'] = g1 + g2 + g3 + g4 + g5

c1 = df.COLLISIONTYPECODE.isin ([1,2])*1
c2 = df.COLLISIONTYPECODE.isin ([3,4,5])*2
c3 = df.COLLISIONTYPECODE.isin ([12,13,14])*3
c4 = df.COLLISIONTYPECODE.isin ([7])*4
c5 = df.COLLISIONTYPECODE.isin ([6])*5
c6 =~df.COLLISIONTYPECODE.isin ([1,2,3,4,5,6,7,12,13,14])*6
df['COLLTYPE'] = c1 + c2 + c3 + c4 +c5 +c6

d1 = df.INTERSECTIONTYPECODE.isin ([1,2,3,4,5,6])*1
df['INTERSECTIONTYPE'] = d1

e1 = df.LIGHTCODE.isin ([1,3])*1
e2 = df.LIGHTCODE.isin ([4])*2
e3 = df.LIGHTCODE.isin ([5.02])*3
e4 = df.LIGHTCODE.isin ([6.02])*4
e5 = ~df.LIGHTCODE.isin ([1,3,4,5.02,6.02])*5
df['LIGHTCOND'] = e1 + e2 + e3 + e4 + e5

f1 = df.FIXOBJCODE.isin ([1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22])*1
df['FIXOBJECT'] = f1

df.head()
```

Out[9]:

	INJSEVERCODE	SEX	AGE	CONTRIBCODE1	CONTRIBCODE2	CONTRIBCODE3	CONTRIBCODE4	CONDITIONCODE	BAC	ALCOTESTCODE	DRUGTESTC
0	Sensible data censored										
1											
2											
3											
4											

In [10]:
```
dfmy = df.loc[:,['INJSEVERCODE','SEX','AGE','DRUGEFFECT','ALCHOLEFFECT','MEDICINEEFFECT','DAMEFFECT','EJTRAP','SEATBELT',
                 'VEHBODY','ROLLEDOVER','ROADTYPE','TRAFFICCONTROL','SURFCOND','COLLTYPE','INTERSECTIONTYPE','LIGHTCOND',
                 'FIXOBJECT']]
dfmy.head(10)
```

Out[10]:

	INJSEVERCODE	SEX	AGE	DRUGEFFECT	ALCHOLEFFECT	MEDICINEEFFECT	DAMEFFECT	EJTRAP	SEATBELT	VEHBODY	ROLLEDOVER	ROADTYPE
0	Sensible data censored											
1												
2												
3												
4												
5												
6												
7												
8												
9												

In [11]:
```
dfmy['INJSEVERCODE'].value_counts()
```

Out[11]:
```
1    260538
3     26594
2     25486
4      3660
5       542
Name: INJSEVERCODE, dtype: int64
```

6. How to delete useless value

In [12]:
```
dfmy = dfmy[~dfmy['SEX'].isin([99])]
dfmy = dfmy[~dfmy['AGE'].isin([999])]
```

In [18]:
```
dfdummy = pd.get_dummies(dfmy, columns=['SEX'], drop_first=True) #True instead 0 if i want to base the Dummy with the male type

dfdummy['AGECLASS'] =pd.cut(dfdummy['AGE'], [0, 15, 25,45,60,75,120,999], labels=['0', '1', '2','3','4','5','6'])
dfdummy = pd.get_dummies(dfdummy, columns=['AGECLASS'], drop_first=True) #based in class 0
dfdummy = pd.get_dummies(dfdummy, columns=['DRUGEFFECT'], drop_first=True) #based on NO
#dfdummy = pd.get_dummies(dfdummy, columns=['PERSONTYPE'], drop_first=True) #based on NO

dfdummy['EJTRAP0']=0
dfdummy['EJTRAP1']=1

dfdummy = pd.get_dummies(dfdummy, columns=['ALCHOLEFFECT'], drop_first=True) #based on not under effect
dfdummy = pd.get_dummies(dfdummy, columns=['MEDICINEEFFECT'], drop_first=True) #based on not under effect
dfdummy = pd.get_dummies(dfdummy, columns=['DAMEFFECT'], drop_first=True) #based on not under effect
dfdummy = pd.get_dummies(dfdummy, columns=['PERSONTYPE'], drop_first=True)

#dfdummy = pd.get_dummies(dfdummy, columns=['EJTRAP'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['SEATBELT'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['VEHBODY'], drop_first=True) #based on car

dfdummy = pd.get_dummies(dfdummy, columns=['ROLLEDOVER'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['ROADTYPE'], drop_first=True) #based on 1 way 2 way undevided

dfdummy = pd.get_dummies(dfdummy, columns=['TRAFFICCONTROL'], drop_first=True) #based on No control

dfdummy = pd.get_dummies(dfdummy, columns=['SURFCOND'], drop_first=True) #based on dry

dfdummy = pd.get_dummies(dfdummy, columns=['COLLTYPE'], drop_first=True) #based on head on

dfdummy = pd.get_dummies(dfdummy, columns=['INTERSECTIONTYPE'], drop_first=True) #based on NO

dfdummy = pd.get_dummies(dfdummy, columns=['LIGHTCOND'], drop_first=True) #based on day/dark lighting

dfdummy = pd.get_dummies(dfdummy, columns=['FIXOBJECT'], drop_first=True) #based on NO
```

In [20]:
```
database = db.Database('MasterView', dfdummy)
```

In [21]:
```
from headers import *
```

In [22]:
```
B_SEX2 = Beta('B_SEX2',0,None,None,0)

B_AGE1 = Beta('B_AGE1',0,None,None,0)
B_AGE2 = Beta('B_AGE2',0,None,None,0)
B_AGE3 = Beta('B_AGE3',0,None,None,0)
B_AGE4 = Beta('B_AGE4',0,None,None,0)
B_AGE5 = Beta('B_AGE5',0,None,None,0)
B_AGE6 = Beta('B_AGE6',0,None,None,1)

B_DRUGEFFECT1 = Beta('B_DRUGEFFECT1',0,None,None,0)
B_ALCHOLEFFECT1 = Beta('B_ALCHOLEFFECT1',0,None,None,0)
B_MEDICINEEFFECT1 = Beta('B_MEDICINEEFFECT1',0,None,None,0)
B_DAMEFFECT1 = Beta('B_DAMEFFECT1',0,None,None,0)

B_PERSONTYPE1 = Beta('B_PERSONTYPE1',0,None,None,0)
B_PERSONTYPE2 = Beta('B_PERSONTYPE2',0,None,None,0)

B_EJTRAP1 = Beta('B_EJTRAP1',0,None,None,0)

B_SEATBELT1 = Beta('B_SEATBELT1',0,None,None,0)

B_VEHBODY2 = Beta('B_VEHBODY2',0,None,None,0)
B_VEHBODY3 = Beta('B_VEHBODY3',0,None,None,0)
B_VEHBODY4 = Beta('B_VEHBODY4',0,None,None,0)
B_VEHBODY5 = Beta('B_VEHBODY5',0,None,None,0)
B_VEHBODY6 = Beta('B_VEHBODY6',0,None,None,1)

B_ROLLEDOVER1 = Beta('B_ROLLEDOVER1',0,None,None,0)

B_ROADTYPE2 = Beta('B_ROADTYPE2',0,None,None,0)
B_ROADTYPE3 = Beta('B_ROADTYPE3',0,None,None,0)
B_ROADTYPE4 = Beta('B_ROADTYPE4',0,None,None,1)

B_TRAFFICCONTROL2 = Beta('B_TRAFFICCONTROL2',0,None,None,0)
B_TRAFFICCONTROL3 = Beta('B_TRAFFICCONTROL3',0,None,None,0)
B_TRAFFICCONTROL4 = Beta('B_TRAFFICCONTROL4',0,None,None,0)
B_TRAFFICCONTROL5 = Beta('B_TRAFFICCONTROL5',0,None,None,1)

B_SURFCOND2 = Beta('B_SURFCOND2',0,None,None,0)
B_SURFCOND3 = Beta('B_SURFCOND3',0,None,None,0)
B_SURFCOND4 = Beta('B_SURFCOND4',0,None,None,1)
```

```
B_COLLTYPE2 = Beta('B_COLLTYPE2',0,None,None,0)
B_COLLTYPE3 = Beta('B_COLLTYPE3',0,None,None,0)
B_COLLTYPE4 = Beta('B_COLLTYPE4',0,None,None,0)
B_COLLTYPE5 = Beta('B_COLLTYPE5',0,None,None,0)
B_COLLTYPE6 = Beta('B_COLLTYPE6',0,None,None,1)

B_INTERSECTIONTYPE1 = Beta('B_INTERSECTIONTYPE1',0,None,None,0)

B_LIGHTCOND2 = Beta('B_LIGHTCOND2',0,None,None,0)
B_LIGHTCOND3 = Beta('B_LIGHTCOND3',0,None,None,0)
B_LIGHTCOND4 = Beta('B_LIGHTCOND4',0,None,None,0)
B_LIGHTCOND5 = Beta('B_LIGHTCOND5',0,None,None,1)

B_FIXOBJECT1 = Beta('B_FIXOBJECT1',0,None,None,0)
```

In [23]:
```
tau1 = Beta('tau1',-1,None,None,0)
delta2 = Beta('delta2',2,None,None,0)
delta3 = Beta('delta3',2,None,None,0)
delta4 = Beta('delta4',2,None,None,0)
```

In [24]:
```
tau2 = tau1 + delta2
tau3 = tau2 + delta3
tau4 = tau3 + delta4
```

In [25]:
```
U = B_SEX2*SEX2+B_AGE1*AGECLASS1+B_AGE2*AGECLASS2+B_AGE3*AGECLASS3+B_AGE4*AGECLASS4+B_AGE5*AGECLASS5+B_AGE6*AGECLASS6+B_DRUGEFFE
```

In [27]:
```
simulate = {
    1: 1-dist.logisticcdf(U-tau1),
    2: dist.logisticcdf(U-tau1)- dist.logisticcdf(U-tau2),
    3: dist.logisticcdf(U-tau2)- dist.logisticcdf(U-tau3),
    4: dist.logisticcdf(U-tau3)- dist.logisticcdf(U-tau4),
    5: dist.logisticcdf(U-tau4)
          }
```

In [28]:
```
biogeme  = bio.BIOGEME(database,simulate)
biogeme.modelName = "ORDERLogit2017_AllEjected_simulation"
```

In [29]:
```
betas = biogeme.freeBetaNames
```

In [30]:
```
results = res.bioResults(pickleFile = 'OrderedLogi2017.pickle')
```

In [31]:
```
betaValues = results.getBetaValues()
```

In [32]:
```
simulatedValues = biogeme.simulate(betaValues)
```

In [33]:
```
b = results.getBetasForSensitivityAnalysis(betas,size=100)
```

In [34]:
```
left,right = biogeme.confidenceIntervals(b,0.9) #90% confidence intervals
```

In [35]:
```
# alt0
marketShare_alt1 = simulatedValues[1].mean()
marketShare_alt1_left = left[1].mean()
marketShare_alt1_right = right[1].mean()
print(f"Market share for 1: {100*marketShare_alt1:.1f}% [{100*marketShare_alt1_left:.1f}%,{100*marketShare_alt1_right:.1f}%]")

# alt1
marketShare_alt2 = simulatedValues[2].mean()
marketShare_alt2_left = left[2].mean()
marketShare_alt2_right = right[2].mean()
print(f"Market share for 2: {100*marketShare_alt2:.1f}% [{100*marketShare_alt2_left:.1f}%,{100*marketShare_alt2_right:.1f}%]")
# alt2
marketShare_alt3 = simulatedValues[3].mean()
marketShare_alt3_left = left[3].mean()
marketShare_alt3_right = right[3].mean()
print(f"Market share for 3: {100*marketShare_alt3:.1f}% [{100*marketShare_alt3_left:.1f}%,{100*marketShare_alt3_right:.1f}%]")
# alt3
marketShare_alt4 = simulatedValues[4].mean()
marketShare_alt4_left = left[4].mean()
marketShare_alt4_right = right[4].mean()
print(f"Market share for 4: {100*marketShare_alt4:.1f}% [{100*marketShare_alt4_left:.1f}%,{100*marketShare_alt4_right:.1f}%]")
# alt4
marketShare_alt5 = simulatedValues[5].mean()
marketShare_alt5_left = left[5].mean()
marketShare_alt5_right = right[5].mean()
print(f"Market share for 5: {100*marketShare_alt5:.1f}% [{100*marketShare_alt5_left:.1f}%,{100*marketShare_alt5_right:.1f}%]")
```

```
Market share for 1: 11.0% [10.3%,11.8%]
Market share for 2: 16.8% [16.1%,17.6%]
Market share for 3: 43.9% [43.3%,44.6%]
Market share for 4: 23.4% [22.2%,24.5%]
Market share for 5: 4.8% [4.4%,5.2%]
```

Figures index

Tables index

Bibliography

Abdel-Aty, M. (2003). Analisys of drivers injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research, 34*, 597-603.

Al-Ghamdi, A. S. (2002). Using logic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention, 34*, 729-741.

Amoros, E. J.-L. (2007). Road Crash Causalties: Characteristics of Police Injury Severity Misclassification. *Journal of Trauma and Acute Surgery, 62*(482-490).

Anas, A. (1982). *RESIDENTIAL LOCATION MARKETS AND URBAN TRANSPORTATION. ECONOMIC THEORY, ECONOMETRICS AND POLICY ANALYSIS WITH DISCRETE CHOICE MODELS.* Chicago (Illinois): New York : Academic Press.

B. Edwards, J. (1998). The Relationship Between Road Accident Severity and Recorded Weather. *Journal of Safety Research, 29*(4), 249-262.

Ben-Akiva, M. E. (1985). *Discrete Choice Analisys: Theory and Application to Travel Demand.* Cambridge: MIT Press.

Bierlaire, M. (2003). BIOGEME: a free package for the estimation of choice models. *STRC.*

Bierlaire, M. (2018). PandasBiogeme: a short introduction. *Transport and Mobility Laboratory*.

Brian Ho-Yin, J. S. (2003). Restraint use and age and sex characteristics of persons involved in fatal motor vehicle crashes. *1830 issue:1*, 10-17.

Burch, C. L. (2014). A comparison of KABCO and ais Injury Severity Metrics Using CODES Linked Data. *Traffic Injury Prevention, 15*(6), 627-630.

Compton, C. P. (2005). Injury Severity Codes: A Comparison of Police Injury Codes and Medical Outcomes as Determined by NASS CDS Investigators. *Journal of Safety Research, 36*(5), 483-484.

Farmer, C. M. (2003). Reliability of Police-Report Information for Determining Crash and Injury Severity. *Traffic Injury Prevention, 4*, 38-44.

FHWA. (2008). *https://safety.fhwa.dot.gov/hsip/spm/conversion_tbl/pdfs/kabco_ctable_by_state.pdf*. Retrieved from Federal Highway Administration.

Gudmundur F.Ulfarsson, F. L. (2004). Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. *Accident Analysis & Prevention, 36*(2), 135-147.

Hassan T. Abdelwahab, M. A.-A. (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Journal of the Transportation Research Board, 1746*(1), 6-13.

Isabelle Aptel, L. R. (1999). Road accident statistics: discrepancies between police and hospital data in a French island. *Accident Analysis and Prevention, 31*, 101–108.

J.R Schott, M. A.-A. (1998). An assessment of the effect of driver age on traffic accident involvement using log-linear models. *Accident Analysis & Prevention, 30*(6), 851-861.

Joon-Ki Kim, G. F. (2013). Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. *Accident Analysis & Prevention, 50*, 1073-1081.

Kara Maria Kockelman, Y.-J. K. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis & Prevention, 34*(3), 313-321.

Kenneth E Train, C. W. (2007). *Vehicle choice behavior and the declining market share of US automakers* (Vol. 48). Blackwell Publishing Inc.

Manski, C. F. (1977). THE STRUCTURE OF RANDOM UTILITY MODELS. *Theory and Decision, 8*(3), 229–254.

Maryland State Police Information Thechnology Division. (2013). *ACRS (Automated Crash Report System) User Manual*.

McFadden, D. (1978). Modeling the choice of residencial location. *Transportation Research Record 672* (pp. 72-77). Washington DC: TRB.

MODP. (n.d.). *https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crash-Data-Dictionary/7xpx-5fte*. Retrieved from Marylands Open Data Portal.

Mohammed Quddus, R. B. (2002). An analysis of motorcycle injury and vehicle damage severity using ordered probit models. *Journal of Safety Research, 33*(4), 445-462.

National Center for Statistics and Analysis. (2018). Police-report motor vehicle traffic crash in 2016. *National Highway Trafic Administration*.

National Center for Statistics and Analysis. (2019). Police-report motor vehicle traffic crash in 2017. *National Highway Trafic Safety Administration*.

National Highway Traffic Safety Administation. (2017). Model Minimum Uniform Crash Criteria. *MMUCC Guideline*.

P Slovic, B. F. (1977). Behavioral Decision Theory. *Annual Review of Psychology, 28*, 1-39.

Paleti, L. B. (2018). Modified Mixed Generalized Ordered Response Model to Handle Misclassification in Injury Severity Data. *Transportation Research Record, 2672*(30), 53–63.

Peter T. Savolainen, F. L. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis and Prevention, 43*, 1666–1676.

Rodríguez. (2007). *Lecture Notes on Generalized Linear Models, Chapter 3, page 45* . Retrieved from http://data.princeton.edu/wws509/notes/

Shamsunnahar Yasmin, N. E. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention, 59*, 506-521.

Small, K. A. (1987). A Discrete Choice Model for Ordered Alternatives. *Econometrica, 55*(2), 409-424.

SOM. (n.d.). *www.medschool.umaryland.edu /ORC_trauma_anes/*. Retrieved from University of Maryland School of Medicine.

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance, 23*(1), 86-112.

Swite, J. D. (1984). Probabilistic choice set generation in transportation demand models. *Ph.D degree*.

Train, K. E. (1986). *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand.* MIT Press.

Train, K. E. (2009). Discrete choice Methods With Simulation.

TRAUMA.ORG. (n.d.). *http://www.trauma.org/archive/scores/ais.html*. Retrieved from www.trauma.org.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review, 79*(4), 281-299.

World Health Organization. (2018). *Geneva: World Health Organization;2018*.