Term Paper

Prediction and validation of motorcycle drivers' behavior

Present to

Assoc. Prof. Dr. Saksith Chalermpong

By

Watcharapong Wongkaew 6230481521

This report is the part of

2101550 Statistical Method of Transportation Analysis

2nd Semester 2021

Department of Civil Engineering, Chulalongkorn University

Abstract

This work investigated the self-reported questionnaire data from CUTI and created severity index and predicted on said index which from the results of analysis, we can summarize and verify hypotheses as follows, most of socioeconomic variables except education, annual tax, and life insurance are not significant. Most of motorcycles related variables except for win experience, no training, extra equipment, and modification equipment are not significant. The severity index model have badness of fit at R^2 near 0.03 and we may choose optimized model II to regress on severity index. The predictive model have badness of fit at AIC around 3000 and accuracy at 82.5 %. The likelihood model have badness of fit at AIC around 6000 and accuracy around 81 %.

Keyword : Severity Index, Accident, Socioeconomic, Logit, Ordered Logit, Predictive crashes

Watcharapong Wongkaew
Researcher

Dr. Patanapong Sanghatawatana
Advisor

Table of Contents

Introduction

There are inevitable facts that motorcycles compiled fatalities from accident up to 50 percent[1] by that of 2016, moreover up to more than 55 per cent of registered vehicles are 2 wheelers or motorcycles. It is undeniable that motorcycles' users' accidents must be investigated and prevented in the future.

Since 2019, the rise of food delivery application[2] such as Grab, LineMan, FoodPanda and Robinhood pushed even more population into using motorcycles in delivery and put them in much more vulnerable positions. Even more shockingly, the COVID-19 pandemic had worsen the situation as these application profited from the lockdown[3] and grow even more exponentially.

Therefore, it is utmost critical that the accident from those on motorcycles be investigated and hypothesized further since most of the data are unreported and underestimated. The data that the researcher use in this work are the work of Chulalongkorn University Transportation Institution (CUTI), which is self-reported accident data.

The objective of this work is to validate and predict motorcycle users' behavior and accidents encounters from the self-reported data from CUTI in 2021, also, determining most influencing variables to the results that obtain through statistical analysis.

[1] Global Road Safety Facility, World Bank Group. (2019). Retrieved from : https://www.roadsafetyfacility.org/country/thailand

[2] Brand Inside. (2019). Retrieved from : https://brandinside.asia/grab-thailand-6th-birthday/

[3] Vulcan Post. (2021). Retrieved from : https://vulcanpost.com/760782/covid-19-grab-new-services-record-revenues/

Literature Review

1. Severity Index (NCDOT, 2014)

The severity index defined by Department of Transportation (US. DOT) is a measure of a property damage only crash (PDO) which means that there were no injuries or fatalities. Therefore, the equivalent property damage only (EPDO) is a way of comparing severity types among each other.

With that, North Carolina Department of Transportation (NC. DOT., 2014) developed formula and weight to substitute for injury types from the accident. That being said, this severity index was only used on crash investigation in certain location, the researchers would want to apply this method to this work by using same formula as below

$$SI = \frac{76.8 \cdot (F) + 8.4 \cdot (PI) + 1(N)}{F + PI + N} \quad \text{- Eq. 1}$$

With descriptions below

The severity index (SI) of a crash is equal to the total equivalent property damage only (EPDO) divided by the number of crashes.

- A non-injury crash or non crashes (N) are equivalent to 1.0 PDO crashes (i.e. EPDO = 1.0)
- An evident injury crash and a possible injury crash (PI) are equivalent to 8.4 PDO crashes (i.e. EPDO = 8.4)
- A fatal crash and a disabling injury crash (F) are equivalent to 76.8 PDO crashes (i.e. EPDO = 76.8)

DOT also specified that A severity index of 8.4 is the threshold for locations that have more serious crashes, which the researcher would want to use in the same way as the formula.

2. Urban Traffic Crash Severity (Cao, Li, Fu, 2020)

Cao, and his team specified and assess urban traffic crashes severity through economic losses and third parties which data provided by PRC government. Therefore, there is some changes in definitions and slight differences between the original definition and classification in China. Although the definition and classification changes, most of the classifications are still usable.

Cao, and his team developed a comprehensive index with divided into 4 grades I through IV depending with crash consequences, with grade I being most serious and IV is not serious at all.

Table 1 : The classification of crash severity by the proposed approach

Crash severity	Index	Crashes consequences
I	> 9	casualties are very serious and lead to very severe congestion
II	7 < X < 9	The crash casualties are serious and produce severe traffic jams
III	5 < X < 7	The crash causes a part of economic losses and disturbs surrounding traffic to some extent
IV	< 5	Economic loss caused by the crash is little; there are no serious casualties and congestion.

Using that comprehensive index, we can use that as weighted index to classify and validate the data.

3. Abbreviated Injury Score (AIS) (UNECE, 2015)

AIS was defined by AAAM – Association for Advancement of Automotive Medicine and it was dedicated to limiting injuries from motor vehicle crashes. AIS is internationally accepted scale for injury severity scoring based on anatomic disruption It is assignment assumes single injury which is consensus based. It contains multiple dimensions of severity as listed

- Threat to life
- Tissue injury
- Cost
- Length of stay
- Temporary or permanent impairment/disability

The Abbreviated Injury Scale (AIS) severity score is on an ordinal scale of 1-6, with one indicating a minor injury and six being maximal (currently untreatable). Abbreviates description of injury severity to a number below as listed

- 1 = minor
- 2 = moderate
- 3 = serious
- 4 = severe
- 5 = critical
- 6 = maximal
- (9 = unknown)

4. Factors relating to motorcycles accident

There are many factors and many researches that conclude on socioeconomic factors and road conditions factors, but since we have data of much more magnitude, the wider the researcher must review the factors, therefore, the researcher only selected a few that contain context to the situation in Thailand and Bangkok

Chumpawadee, 2015 investigated about motorcycle accident risk behavior, and found that factor contributed is gender, experience, and perception. The team did not find a significant correlation between environmental conditions.

Champahom et. Al., 2021 investigated factors affecting severity of motorcycle accidents on Thailand's arterial roads. It was found that age and gender played a role in the accident.

Oltaye, 2021 investigated associated factors among road traffic accident patients. The team use Multiple logistic regression analyses and factored in age, gender, speed, place of residence and types of road which mostly played a significant role in accident.

Baral, 2015 investigated factors affecting the severity of motorcycles accidents and casualties in Thailand by using probit and logit model. The team did identify the main factors that affect the injury severity of motorcycle accident and motorcycle casualties which were ages, time and day, angle of crashes, and traffic violation behavior also played a role.

Data Overview

The researcher proposed that summary of overview of the data in the form of table would be easier to comprehend. The data, as said, was from self-reported accident questionnaire, the courtesy from Chulalongkorn University Transportation Institute.

The data was mostly cleaned and digitized in the form of XLSX workbook beforehand, the researcher then used Python 3.10 to clean up data and try to fill missing data as 0 and ignore any missing categorical data.

For the nature of the data, the data is mostly categorical in the form of dummy variable while there are some numeric variables, it was not much as it will be summarized below

Table 2.1 : Data overview

<table>
<tr><th>Name</th><th>Content</th><th>Specifics</th><th>Data Type</th><th>Notes</th></tr>
<tr><td>RiderType</td><td>Motorcycle Rider Type</td><td>3 types</td><td rowspan="2">Categorical</td><td>Pub, Win</td></tr>
<tr><td>Zone</td><td>Zone of operation within Bangkok</td><td>3 zones</td><td>Inner, Middle</td></tr>
<tr><td>Age</td><td>Age of rider</td><td>None</td><td rowspan="8">Numerical</td><td></td></tr>
<tr><td>Exp_Gen</td><td>General experiences</td><td rowspan="4">>0</td><td></td></tr>
<tr><td>Exp_Win</td><td>Motorcycle Win experiences</td><td></td></tr>
<tr><td>Exp_App</td><td>Application Rider experiences</td><td></td></tr>
<tr><td>Total_Ridehour</td><td>Total ride-hour within weeks</td><td></td></tr>
<tr><td>SumFatality</td><td>Total severe injuries and fatalities caused by accident</td><td rowspan="3">None</td><td></td></tr>
<tr><td>SumInjured</td><td>Total major and minor injuries caused by accident</td><td></td></tr>
<tr><td>SumNear</td><td>Total near accidents occurances</td><td></td></tr>
<tr><td>Gender</td><td>Gender of rider</td><td>2 Types</td><td rowspan="2">Categorical</td><td></td></tr>
<tr><td>MaritalStatus</td><td>Marital status of rider</td><td>4 Types</td><td></td></tr>
<tr><td>NoNurture</td><td>Number of child/ children nuture</td><td>None</td><td>Numerical</td><td></td></tr>
<tr><td>Education</td><td>Education level of rider</td><td>4 Levels</td><td rowspan="2">Ordinal</td><td></td></tr>
<tr><td>PersonalIncome</td><td>Income of rider</td><td>6 Level</td><td></td></tr>
</table>

AnnualTax	Annual tax paid by rider	2 Levels	Categorical	
Compul_Insurance	Rider have compulsory insurance			
Vol_Insurance	Rider have voluntary insurance			
HealthInsurance	Rider have health insurance			
AccidentInsurance	Rider have accident insurance			
LifeInsurance	Rider have life insurance			
SelfPractice	Rider have practiced by themselves			
NoTraining	Rider have no training			
License Personal	Rider have personal license			
License Public	Rider have public license			
License Temp	Rider have temporary personal license			
NoneLicense	Rider have no license			
CCSize	Engine Cylinder size	3 Levels	Ordinal	
Mod_Eq	Number of modification equipment	None	Numerical	
Ext_Eq	Number of extra safety equipment			

Table 2.2 : Dependent variables overview

Name	**Content**	**Data type**	Notes
SI	Adjusted Severity Index	Numerical with limits	
PSC	Predictive Severe Crashes	Binary Categorical	
QPSC	Quaternary Predictive Severe Crashes	Quaternary Ordinal	3 most likely 0 least likely

Hypotheses

The researcher want to emphasize the prediction and validation part of the work therefore, the hypotheses that formulated here would be relevant to the goals of prediction and validation from the data.

With that assumption holds, most of hypotheses would be relevant, or correlated to those of variables within the dataset.

- Socioeconomic variables such as age, education have significant effects on severity index
- Motorcycles related variables such as training, modification have signficant effects on severity index
- The more restricted model is, the more accuracy and distinction it will hold

Research Methodology

Mostly from the data overview which we will discuss and cover in next part, the researcher ran a preliminary exploration from the model which it seems that most of variables did not correlate with each other and have no correlation whatsoever. This will be discussed further.

1. Hypotheses
    i. Socioeconomic variables such as age, education have significant effects on severity index
    ii. Motorcycles related variables such as training, modification have significant effects on severity index
    iii. The more restricted model is, the more accuracy and distinction it will hold

2. Data Acquisition
    i. Data obtained from Self-Reported Questionnaire from CUTI
        i. Secondary data
        ii. Sources : CUTI
    ii. Weights and references from SciDirect and Google Scholar
        i. Secondary data

3. Data Cleaning
    i. Cleaning with Python 3.10 and excel
        i. Column cleaning
    ii. Using R to clean up columns and missing data
        i. Summary statistics

        ii. F
   iii. Handling of missing data
        i. Deletion
        ii. Filling with 0

4. Relevant variables
   i. All variables listed in **Table 2**
      i. Many of them are dummy variables
      ii. Some are ordinal and numerical
   ii. Since we have data of much more magnitude, the researcher wanted to try and regress all of the variables first, and then taking literature review suggestion after that.

5. Method of analysis
   i. Validation of the behavior
      i. Using **multiple regression** to validate the severity index (SI) that used the weight from Cao, 2020 with forms of **Eq.1**
      ii. The equation is **Eq. 2** which is describe below

$$SI = \frac{9 \cdot (F) + 5 \cdot (PI) + 1(N)}{F + PI + N}$$ **- Eq. 2**

With descriptions below

The adjusted severity index (ASI) of a crash is adjusted of weight from the equation to estimate index from Cao, 2020 which pertains to the consequences of the crashes divided by the number of crashes.

- A non-injury crash or non crashes (N), use weight of 1.0
- An evident injury crash and a possible injury crash (PI) are equivalent to type I, using weight of 5.0
- A fatal crash and a disabling injury crash (F) are equivalent type IV, using weight of 5.0

      iii. Create 5 models
         1. I Regress all variables
         2. II Regress with pre-processed data
         3. III Take only significant variable from I model
         4. IV Take suggestions from the literature review
         5. V Use only suggestions from literature review
   ii. Prediction of the behavior
      i. Using **logistic regression** to predict and validate the severity index (SI) calculated from
         1. **SI > 1.0** means that they are susceptible to crashes/ accident (X = 1)

        2. **SI < 1.0** means that they are not susceptible to crashes/ accident (X = 0)
    ii. Using **ordered logistic regression** to predict and validate the severity index (SI) calculated from
        1. **SI > 9** means that they are most likely susceptible to severe crashes/ accident (X = 3)
        2. **5 < SI < 9** means that they are more likely susceptible to severe crashes/ accident (X = 2)
        3. **1 < SI < 5** means that they are less likely susceptible to crashes/ accident (X = 1)
        4. **SI < 1** means that they are least likely susceptible to crashes/ accident (X = 0)
    iii. Create 2 models
        1. I Regress all variables
        2. II Improvement of variables

6. Testing hypotheses
    i. Validation of the behavior
        i. Testing the results of **multiple regression** with R^2
        ii. Testing the coefficients of regression with t-test
        iii. Testing different models with different type of variables with f-test (ANOVA)
    ii. Prediction of the behavior
        i. Testing the results of **logistic regression** with confusion matrix, deviance and AIC (hypothesis III)
        ii. Testing the coefficient of regression with t-test (hypotheses I and II)
        iii. Testing different models with different type of variables with Likelihood-Ratio Test (LR Test) (hypothesis III)

7. Results and Discussion
    i. Results of validation using severity index
    ii. Results of prediction using accident binary predictor
    iii. Comparing with the model prediction with machine learning method using Extreme Gradient Boosting (XGBoost)
    iv. Discussion and limitation of the work

Analysis Results

1. Summary Statistics

Using following R-code, we create new variable and delete old variable from many columns which are Mod_Eq and Ext_Eq which contain old variables inside and then delete them, then obtain summary statistics of all categorical data and N.A. (Not Available) data which are 2, we then fill those 2 as 0 altogether since it is very small compared to the size of data.

Sample code

```
sq_all$Mod_Eq <- sq_all$Modify_Engine + sq_all$Modify_intake +
sq_all$Modify_Wheel + sq_all$Modify_ColorBody #Creation of new variables#
sq_all$Modify_Engine <- sq_all$Modify_intake <- sq_all$Modify_Wheel <-
sq_all$Modify_ColorBody <- sq_all$Modify_None<- NULL #Delete old variables#
summary(sq_all) #Summary Statistics#

sq_all$Inner<-ifelse(sq_all$Zone=="Inner",1,0)
sq_all$Middle<-ifelse(sq_all$Zone=="Middle",1,0)
sq_all$Outer<-ifelse(sq_all$Zone=="Outer",1,0) #Dummy variables creation#
sq_all$RiderType <- sq_all$Zone <- NULL #Delete old variables#

sum(is.na(sq_all)) #find NA#
sq_all[is.na(sq_all)] <- 0) #Fill NA = 0#
```

Summary Statistics as follows : (in the next page)

From the data which presented as a table in next page, we can see that most of the columns are categorized and sorted into categorical data. There are still some data that needed to be sorted as dummy variables as RiderType and Zone.

With dummy variables creation done, we proceed in summarization of relevant variables and usage of each variable in the model. With these many dummy and categorical data, it should be done as in severity scoring which we will combine all columns with accident data by using the method of Abbreviated Injury Score (AIS) combined with accident severity index from DOT that covered in literature review.

The accident scoring system that we created is as follows

$$SI = \frac{9 \cdot (F) + 5 \cdot (PI) + 1(N)}{F + PI + N} \quad \textbf{- Eq. 2}$$

The description will not be repeated here.

Summary Statistics as follows :

```
Age             Exp_Gen         Exp_Win          Exp_App        Total_Ridehour   SumFatality_Adj   SumInjured_Adj
 Min.   :18.00   Min.   : 1.00   Min.   : 0.000   Min.   :0.000   Min.   :  0.40   Min.   :0.0000   Min.   :0.000
 1st Qu.:30.00   1st Qu.:12.00   1st Qu.: 0.000   1st Qu.:0.000   1st Qu.:  2.65   1st Qu.:0.0000   1st Qu.:0.000
 Median :39.00   Median :18.00   Median : 0.000   Median :0.000   Median : 56.00   Median :0.0000   Median :0.000
 Mean   :39.51   Mean   :19.68   Mean   : 3.537   Mean   :0.798   Mean   : 45.67   Mean   :0.0081   Mean   :0.058
 3rd Qu.:48.00   3rd Qu.:25.00   3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 72.00   3rd Qu.:0.0000   3rd Qu.:0.000
 Max.   :78.00   Max.   :57.00   Max.   :46.000   Max.   :9.000   Max.   :115.00   Max.   :6.6667   Max.   :9.412

  SumNear_Adj          Gender         MaritalStatus     NoNurture        Education       PersonalIncome     AnnualTax
 Min.   : 0.00000   Min.   :0.00000   Min.   :1.000   Min.   :0.000   Min.   :1.000   Min.   :1.000   Min.   :0.0000
 1st Qu.: 0.00000   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.0000
 Median : 0.00000   Median :0.00000   Median :2.000   Median :1.000   Median :2.000   Median :3.000   Median :1.0000
 Mean   : 0.12254   Mean   :0.07339   Mean   :1.677   Mean   :1.197   Mean   :2.449   Mean   :2.583   Mean   :0.8212
 3rd Qu.: 0.01587   3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1.0000
 Max.   :11.76471   Max.   :1.00000   Max.   :4.000   Max.   :8.000   Max.   :5.000   Max.   :6.000   Max.   :1.0000

 Compul_Insurance Vol_Insurance    Health_Insurance Accident_Insurance Life_Insurance  Self_Practice     NoTraining
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000     Min.   :0.000   Min.   :1.000   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.0000
 Median :1.000   Median :0.0000   Median :0.0000   Median :0.0000     Median :0.000   Median :1.000   Median :0.0000
 Mean   :0.931   Mean   :0.1077   Mean   :0.1983   Mean   :0.3016     Mean   :0.132   Mean   :1.218   Mean   :0.3963
 3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000     3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:1.0000
 Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000     Max.   :1.000   Max.   :3.000   Max.   :6.0000

  Licence_Temp     Licence_Personal Licence_Public    NoneLicence        CCSize          Ext_Eq          Mod_Eq
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000   Min.   :1.000   Min.   :0.00   Min.   :0.0000
 1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.0000
 Median :0.000   Median :1.0000   Median :0.0000   Median :0.000000   Median :1.000   Median :1.00   Median :0.0000
 Mean   :0.049   Mean   :0.8426   Mean   :0.3022   Mean   :0.002959   Mean   :1.498   Mean   :1.57   Mean   :0.0675
 3rd Qu.:0.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000   3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:0.0000
 Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000   Max.   :3.000   Max.   :6.00   Max.   :4.0000
 NA's   :1                                                                                           NA's   :1

 Inner           Middle           Outer             Pub              Win              App              Traffic Violation
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.356   Mean   :0.3945   Mean   :0.2495   Mean   :0.3471   Mean   :0.3107   Mean   :0.3421   Mean   :0.2299
 3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :50.0000
```

2. Histogram and data representation

Since there are many variables, therefore researcher would only show the sample of data representation and histogram by the category of the data listed below

1. Categorical data (Dummy Variable)

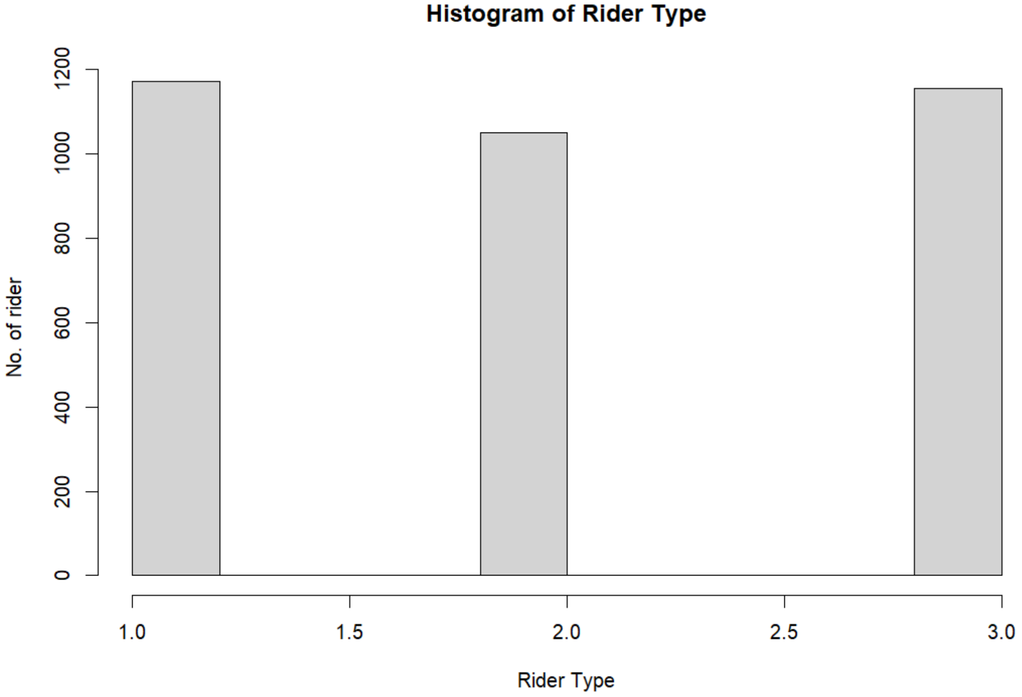

Figure 1 : Rider type before dummifying

Figure 2 : Riders' gender (Male Base)

2. Numerical data

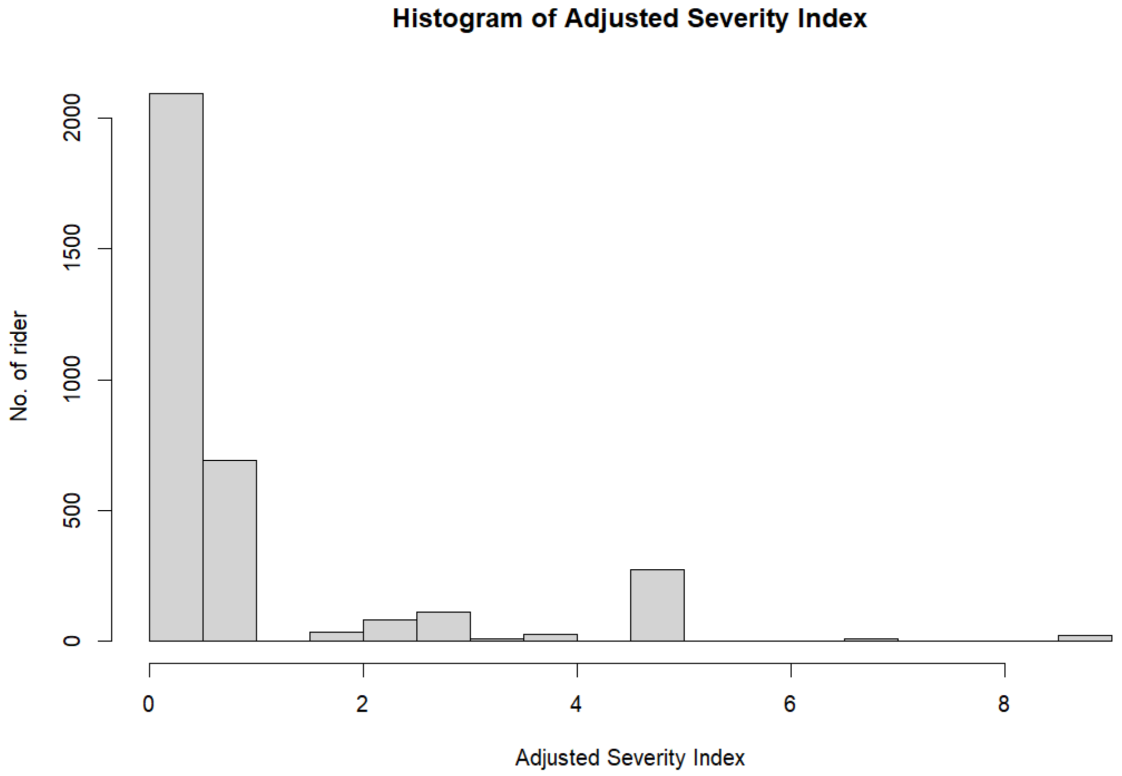

Figure 3 : Adjusted Severity Index

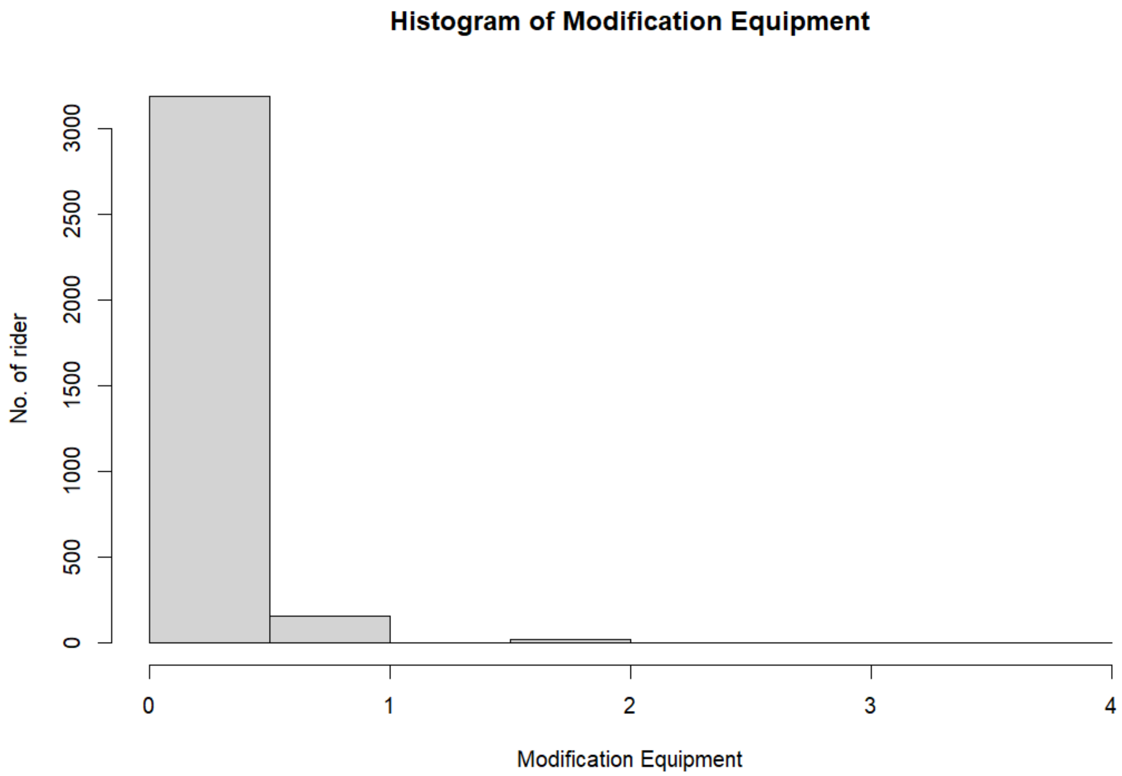

Figure 4 : Modification Equipment

3. Ordinal Data

Figure 5 : Riders' Age

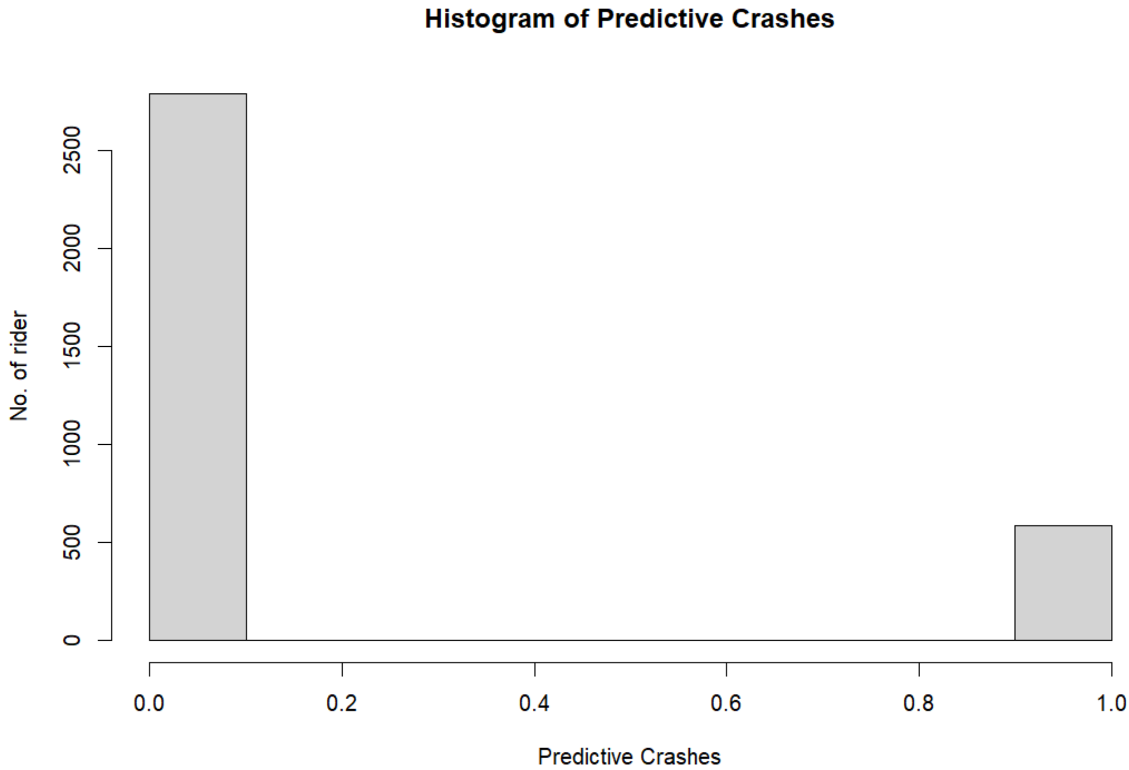

Figure 6 : Predictive Crashes

3. Severity Index Interpretation (SI)

With severity scoring which we will combine all columns with accident data by using the method of Abbreviated Injury Score (AIS) combined with accident severity index from DOT and Cao, 2020 that covered in literature review.

The accident scoring system that we created is as follows

$$SI = \frac{9\cdot(F) + 5\cdot(PI) + 1(N)}{F + PI + N}$$ **- Eq. 2**

The interpretation will be based on DOT interpretation of their severity index, thus we can interpret in 2 ways

1. Binary Interpretation
   The interpretation will be based on considering that have the rider been in the accident before, therefore it will be interpreted as below
   SI > 1.0 means that they likely been in the accident before, and they are susceptible to crash in the future
   SI < 1.0 means that they likely **had not** been in the accident before, and they are less susceptible to crash in the future
2. Quaternary Interpretation
   The interpretation will be based on Cao, 2020 with considering their comprehensive index in **Table 1** that how the riders' injury been in the accident before, therefore it will be interpreted as below
   SI >= 9 means that they are most likely been in severe accident before and susceptible to severe crashes/ accident
   5 =< SI < 9 means that they are more likely been in severe accident before and susceptible to severe crashes/ accident
   1 =< SI < 5 means that they are less likely been in accident before and susceptible to crashes/ accident
   SI < 1 means that they are least likely been in accident before and susceptible to crashes/ accident

With that we can see the results above briefly in **Figure 3** and we can see summary statistics below

```
> summary(sq_all$si) #Adjusted Severity Index
   Min.     1st Qu.    Median    Mean     3rd Qu.    Max.
 0.0000     0.0000     0.0000    0.9259   1.0000     9.0000

> summary(sq_all$psc) #Predictive Crashes
   Min.     1st Qu.    Median    Mean     3rd Qu.    Max.
 0.0000     0.0000     0.0000    0.1743   0.0000     1.0000

> summary(sq_all$qpsc) #Quaternary Predictive Severe Crashes
   Min.     1st Qu.    Median    Mean      3rd Qu.    Max.
 0.0000     0.0000     0.0000    0.4788    1.0000     3.0000
```

And we can see the results briefly in **Figure 7** below

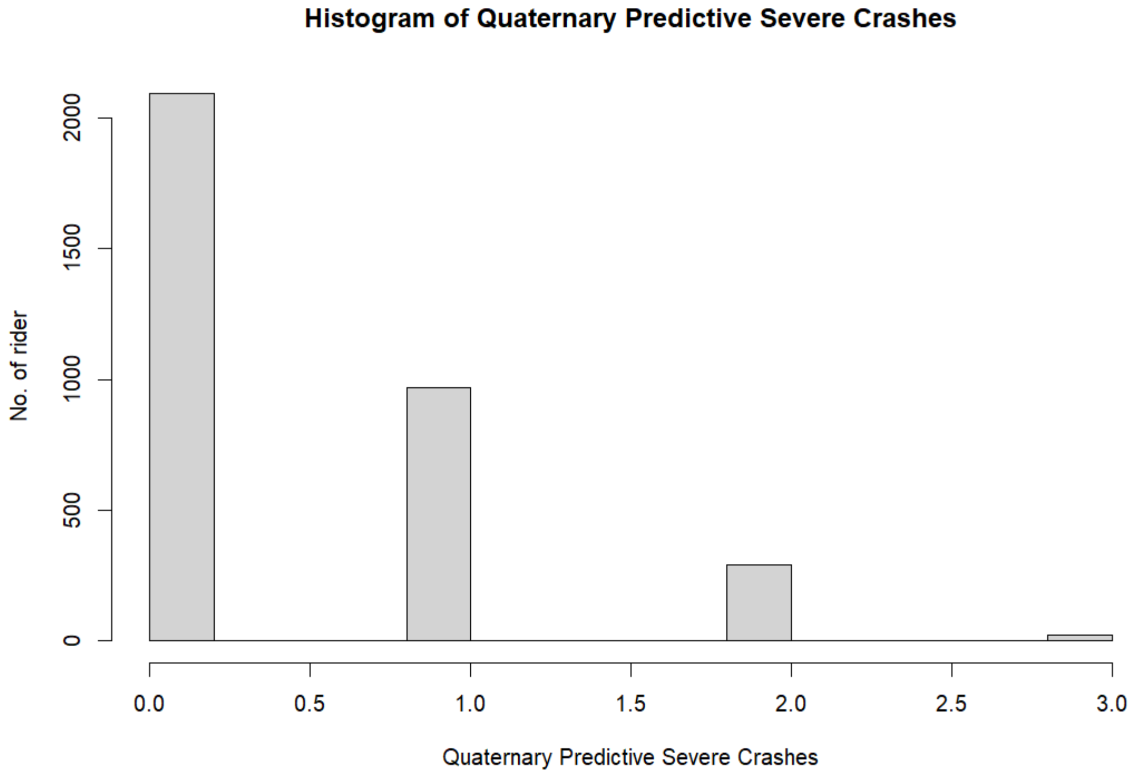

Figure 7 : Quaternary Predictive Severe Crashes

With all of that we can produce **multiple regression, logit model both of logit and ordered logit** with the command in R below

Sample code

```
#Multiple regression analysis
val <- lm(si ~ . -psc , data = sq_all )
summary(val)

val2 <- lm(si ~ . -psc, data = sq_all)
summary(val2)
anova(val, val2)

#Logistic Regression analysis
pre <- glm(psc ~ .-si, family = binomial(link = "logit"), data =
sq_all )
lrtest(pre, pre2)
summary(pre)

#Ordered Logistic Regression analysis
pre3 <- polr(as.factor(qpscf) ~ . - psc - si, data = sq_all,
Hess=TRUE, method = c("logistic"))
```

4. Preliminary Analysis

With severity scoring which we combined all columns with accident data by using the method of Abbreviated Injury Score (AIS) combined with accident severity index from DOT and Cao, 2020 that covered in literature review, we can estimate and regress on that result.

Before that, we explored the dependent variables separately with multiple regression model, with summary of results below

Table 3.1 : Preliminary Results

Model	R^2	ANOVA(F)
SumFatality	0.013	4.72 with 0.001 confidence
SumInjured	0.036	5.35 with 0.001 confidence
SumNear	0.039	5.81 with 0.001 confidence

Table 3.2 : Converted Preliminary Results (taking total ride hours into account)

Model	R^2	ANOVA(F)
SumFatality	0.005	1.71 with 0.05 confidence
SumInjured	0.071	10.26 with 0.001 confidence
SumNear	0.091	13.17 with 0.001 confidence

We can see that with table 3.1 and 3.2, when we take account directly from the variables, we can see that it does not fit that good, with this result the researcher conduct another test with taking exposure into account and it was better, but still distinct and have strange distribution. Therefore, the researcher decided to use severity index that compound these variables together and it may create better results.

Deep Analysis Results

0. Correlation matrix

We may need more input to better judge the model performance therefore, use function `corrplot` from package `corrplot` which called below

```
sq_all.cor = cor(sq_all)
corrplot(sq_all.cor)
```

And we have correlation matrix of decision below,

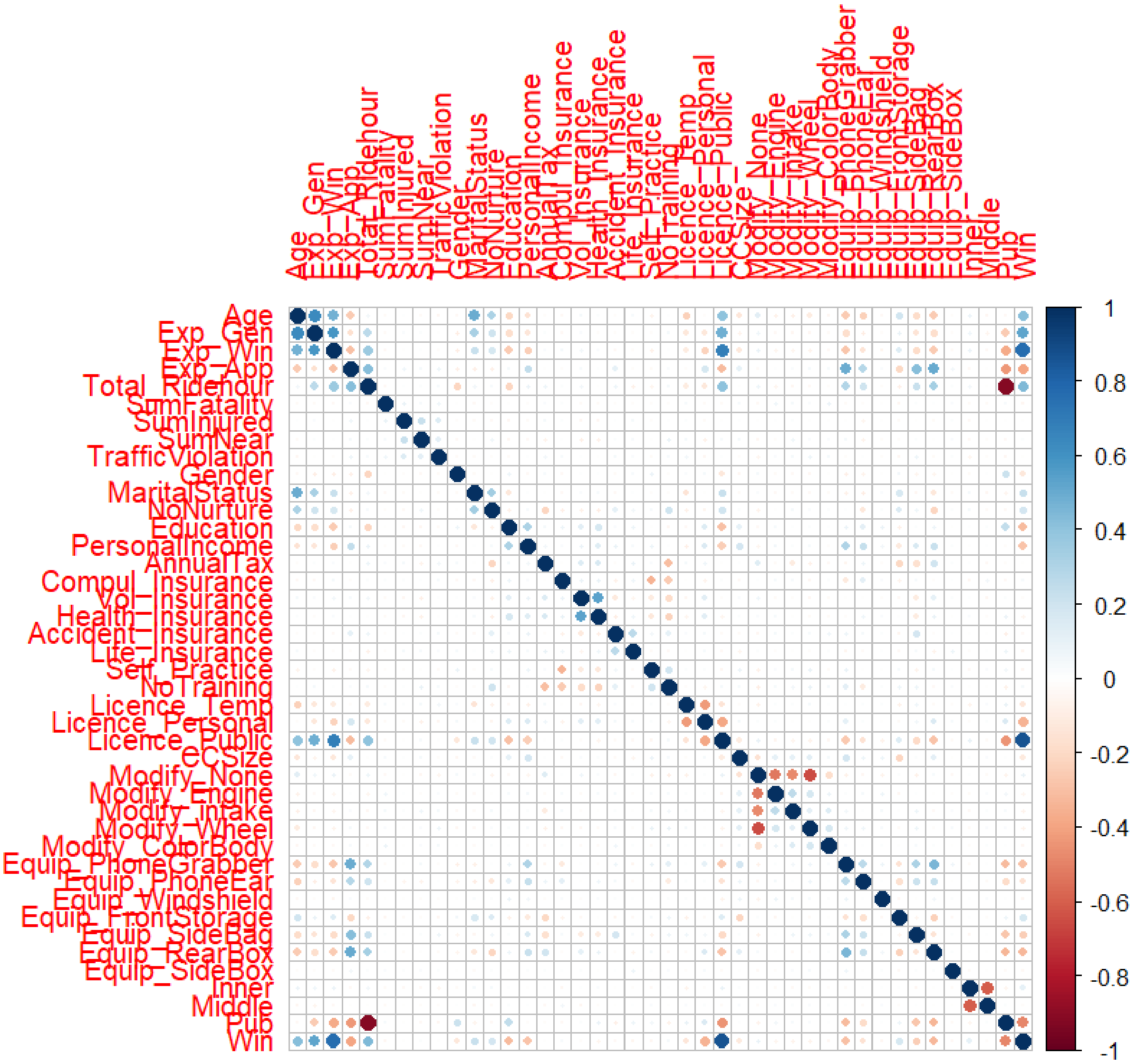

Figure 7.1 : Correlation Matrix of variables

This correlation matrix gives us a brief correlation between variables and dummy variables and how they interact with each other. With these results, we can see that most have no correlation with each other, but some that have correlations we may need to reduce that which we list as

- Equipment as extra equipment
- Modification as modifications equipment

Moreover, we think that our target/ dependent variables did not fit that good, thus we would try to use regression and compound dependent variable.

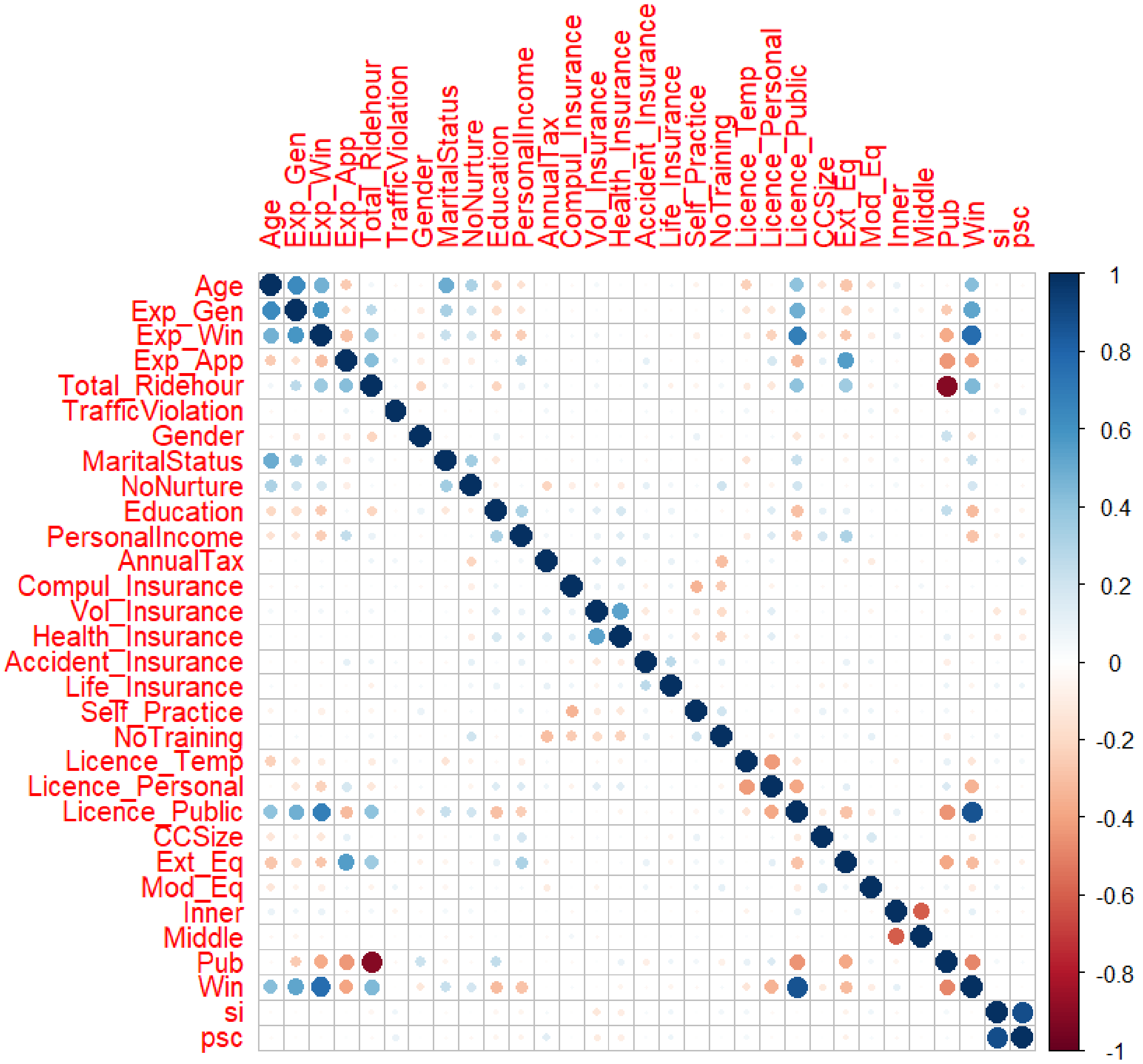

Figure 7.2 : Correlation Matrix

From the matrix we can see that we cannot rule out most of categorical data because it tied to the dummy variables, and it may cause unintended consequences. With that we may decease only 2 categories of data which are

- Insurances which compounded into total insurances
- Experiences which compounded into total experience

For other unsignificant variables we may rule out as listed below

- Marital status
- Children nurtured
- Self-Practice

These 3 variables show next to no correlation, and it would be redundant to put in the model, thus in model II, which is the model that we preprocess the data, we will cut off these variables.

1. Validation of behavior

Using code above in analysis results section, we have summary of 4 multiple regression models below

- First model (I : val)

```
Call:
lm(formula = si ~ . - psc, data = sq_all)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7825 -0.9742 -0.5983  0.1797  8.6935

Coefficients:
                     Estimate  Std. Error  t value Pr(>|t|)
(Intercept)         0.9409677  0.3292868   2.858 0.004295 **
Age                -0.0032538  0.0037339  -0.871 0.383582
Exp_Gen             0.0003023  0.0041074   0.074 0.941340
Exp_Win            -0.0016380  0.0069466  -0.236 0.813599
Exp_App             0.1342695  0.0387636   3.464 0.000539 ***
Total_Ridehour      0.0016115  0.0021004   0.767 0.442982
TrafficViolation    0.0445342  0.0182113   2.445 0.014520 *
Gender             -0.0232980  0.1115527  -0.209 0.834577
MaritalStatus      -0.0288876  0.0577580  -0.500 0.617003
NoNurture           0.0158138  0.0254769   0.621 0.534833
Education          -0.0748220  0.0364282  -2.054 0.040057 *
PersonalIncome     -0.0683493  0.0470570  -1.452 0.146462
AnnualTax           0.3429401  0.0820174   4.181 2.97e-05 ***
Compul_Insurance    0.1071064  0.1231249   0.870 0.384417
Vol_Insurance      -0.5692527  0.1104206  -5.155 2.68e-07 ***
Health_Insurance   -0.1321906  0.0873946  -1.513 0.130483
Accident_Insurance  0.0212883  0.0658795   0.323 0.746609
Life_Insurance      0.2310272  0.0886698   2.605 0.009215 **
Self_Practice      -0.0779564  0.0711050  -1.096 0.273002
NoTraining          0.1633041  0.0587130   2.781 0.005443 **
Licence_Temp        0.1953593  0.1600347   1.221 0.222274
Licence_Personal   -0.0321828  0.1005908  -0.320 0.749036
Licence_Public     -0.0751120  0.1283284  -0.585 0.558378
CCSize              0.0031545  0.0583908   0.054 0.956919
Ext_Eq             -0.0792304  0.0355252  -2.230 0.025796 *
Mod_Eq              0.3184952  0.0951348   3.348 0.000823 ***
Inner              -0.2009273  0.0746232  -2.693 0.007126 **
Middle             -0.1460904  0.0727254  -2.009 0.044639 *
Pub                 0.5189847  0.1832741   2.832 0.004657 **
Win                 0.3229017  0.1774970   1.819 0.068971 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.63 on 3349 degrees of freedom
Multiple R-squared:  0.04889, Adjusted R-squared:  0.04065
F-statistic: 5.936 on 29 and 3349 DF, p-value: < 2.2e-16
```

With the model summary above, we can actually see the significance of variable which R tested with t-test and f-test, the model itself had significant with

F-statistic at 5.9 on 29 and 3349 df at 0.001 confidence level

R-squared at 0.04 which takes as a bad fit

Also, with the test of coefficients above we can see that not many of coefficients and dummy variables are significant, notably there were

- AnnualTax > the researcher suggested that whether riders paid annual tax or not, it certainly does not contribute much to the severity index that we are regressing because the model suggests that the more people paid tax, the more severity index it increase
- NoTraining > this dummy variable surely deserves the place and surely describe the accident and erratic behavior of riders because the model suggests that if people have no training at all, the more severity index it increase.
- Mod_Eq > this variable also have correlation with the accident statistics and if they had more of modification which may decrease the motorcycle safety, the more severity index it increase.
- Exp_App > this variable also have correlation with the accident statistics and if they had more exposure (experience), it may contribute to more severity index.
- Ext_Eq > this variable have negative correlation with the accident statistics, that is if they had more extra safety equipment, the less severity index they would have.
- Pub, Win > this set of dummy variables suggest that working environments and conditions differ the severity index with general riders as highest severity index
- Inner, Middle > this set of dummy variables also suggest that working zone of operations differ the severity index with outer zone as highest severity index.
- Vol_insurance > this dummy variable suggest that if rider have voluntary insurance, they may suffer less severity index.
- TrafficViolation > this variable suggest that if rider had more encounter with traffic violation, they would suffer more severity index.
- Education > this variable has negative correlation that is if rider had more education, they would suffer less severity index

With these variables and more unsignificant variables, we would need more input to decide and determine what data is redundant, or cause multicollinearity.

- Second model (II : val2)

```
Call:
lm(formula = si ~ . - psc - NoNurture - MaritalStatus - Self_Practice,
    data = sq_all)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6296 -0.9381 -0.6918  0.1992  8.2922 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.542e+00  2.682e-01   5.747 9.87e-09 ***
Age              -2.923e-03  3.419e-03  -0.855 0.392570    
Total_Ridehour    3.651e-05  2.087e-03   0.017 0.986043    
TrafficViolation  5.165e-02  1.830e-02   2.822 0.004806 ** 
Gender           -2.751e-02  1.120e-01  -0.246 0.806005    
Education        -8.422e-02  3.657e-02  -2.303 0.021328 *  
PersonalIncome   -7.056e-02  4.683e-02  -1.507 0.131960    
AnnualTax         2.931e-01  8.027e-02   3.651 0.000265 ***
NoTraining        1.890e-01  5.595e-02   3.379 0.000737 ***
Licence_Temp      1.446e-01  1.606e-01   0.900 0.368114    
Licence_Personal -9.887e-02  1.006e-01  -0.983 0.325798    
Licence_Public   -7.545e-02  1.290e-01  -0.585 0.558526    
CCSize           -1.987e-02  5.851e-02  -0.340 0.734185    
Ext_Eq           -9.079e-02  3.530e-02  -2.572 0.010150 *  
Mod_Eq            3.450e-01  9.558e-02   3.609 0.000312 ***
Inner            -1.874e-01  7.511e-02  -2.494 0.012662 *  
Middle           -1.241e-01  7.317e-02  -1.696 0.090009 .  
Pub               7.951e-02  1.642e-01   0.484 0.628194    
Win              -5.850e-02  1.436e-01  -0.407 0.683786    
Tot_ins          -5.412e-02  3.301e-02  -1.640 0.101182    
Tot_exp           2.415e-04  3.162e-03   0.076 0.939115    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.645 on 3358 degrees of freedom
Multiple R-squared:  0.02904, Adjusted R-squared:  0.02326 
F-statistic: 5.022 on 20 and 3358 DF,  p-value: 1.731e-12
```

With the model summary above, we can actually see the significance of variable which R tested with t-test and f-test, the model itself had significant with

F-statistic at 5.0 on 20 and 3358 df at 0.001 confidence level

R-squared at 0.023 which takes as a bad fit worse than Model I

Also, with the test of coefficients above we can see that not many of coefficients and dummy variables are significant, notably there were mostly the same with model I, therefore, let us take a look at non-significant variables as listed

- Licenses group > mostly anticipated that the group had little to no impact on the correlation and severity because most of rider have license anyway and most of them had random effect on the severity index
- Ride hours > it is surprising that the variable that corresponds exposures to accident had very little effect on the severity index, it may

be because we need to factor date and time of ride hours and be more specific to specify the impact of this variable.

- Rider type > when remove experience variable group, it seems that rider type also hold no significance over the correlation and severity index, mostly it correlates with experience variable group which would likely create multicollinearity problem.
- Total insurance and experiences > this is no surprise because the researcher suspects that this compounded variable would hold no significance over severity index because of model I results.

We also test ANOVA (f-test) to test if model II is better than model I which the test suggests that it is better, resulting below

```
> anova(val, val2)
Analysis of Variance Table
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3349 8901.4
2   3358 9087.2 -9   -185.76 7.7654 2.144e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Third model (III : val3)

```
Call:
lm(formula = si ~ AnnualTax + NoTraining + Mod_Eq + Ext_Eq +
    Inner + Middle + Tot_ins + Tot_exp + TrafficViolation + Education,
    data = sq_all)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8268 -0.9435 -0.6998  0.1808  8.3177

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.302450   0.148750   8.756  < 2e-16 ***
AnnualTax         0.283273   0.078598   3.604 0.000318 ***
NoTraining        0.179636   0.055729   3.223 0.001279 **
Mod_Eq            0.364219   0.093393   3.900 9.81e-05 ***
Ext_Eq           -0.108428   0.025422  -4.265 2.05e-05 ***
Inner            -0.211598   0.074445  -2.842 0.004505 **
Middle           -0.140799   0.072717  -1.936 0.052920 .
Tot_ins          -0.064468   0.032536  -1.981 0.047620 *
Tot_exp          -0.004180   0.002022  -2.067 0.038788 *
TrafficViolation  0.053339   0.018223   2.927 0.003445 **
Education        -0.084955   0.033781  -2.515 0.011954 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.645 on 3368 degrees of freedom
Multiple R-squared:  0.02582, Adjusted R-squared:  0.02293
F-statistic: 8.926 on 10 and 3368 DF,  p-value: 1.214e-14
```

With the model summary above, we can actually see the significance of variable which R tested with t-test and f-test, the model itself had significant with

F-statistic at 8.9 on 10 and 3368 df at 0.001 confidence level

R-squared at 0.023 which takes as a bad fit worse than Model I

Also, with the test of coefficients above we can see that most coefficients and dummy variables are significant, notably there were mostly the same with model I and II, therefore, let us take a look some changes when we take out all of non-significant group

- Total insurance > this is a surprise because the researcher had suspected that this compounded variable would hold no significance over severity index because of model I results. Although the results had shown that it had only 0.05 significance, it correlates with severity index in expected way which reduce severity index.
- Total experience > this is not a surprise because the researcher had suspected that this compounded variable would hold some significance over severity index because of more experience would mean that less likely to have accident.

  Although the results had shown that it had only 0.05 significance, it correlates with severity index in expected way which reduce severity index.

We also test ANOVA (f-test) to test if model III is better than model II which the test suggests that it is worse, resulting below

```
> anova(val2, val3)
Analysis of Variance Table
  Res.Df    RSS  Df Sum of Sq      F Pr(>F)
1   3358 9087.2
2   3368 9117.3 -10   -30.148 1.1141 0.3471
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore, we reject model III and we can use model II as the baseline model for usage in the next part. We also test ANOVA (f-test) to test if model III is better than model I, which the test suggests that it is better

```
> anova(val, val3)
Analysis of Variance Table
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1   3349 8901.4
2   3368 9117.3 -19   -215.91 4.2753 1.546e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Forth model (IV : val4)

Taken account of the literature review, we can see the summary of the model below

```
Call:
lm(formula = si ~ Age + Tot_exp + Gender + Ext_Eq + Tot_ins +
    Education + TrafficViolation + Inner + Middle + AnnualTax +
    NoTraining + Mod_Eq, data = sq_all)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7924 -0.9450 -0.7015  0.1824  8.3102

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.412231   0.184916   7.637 2.88e-14 ***
Age              -0.003270   0.003220  -1.015 0.309952
Tot_exp          -0.002670   0.002512  -1.063 0.287857
Gender            0.011443   0.110432   0.104 0.917477
Ext_Eq           -0.113378   0.026055  -4.352 1.39e-05 ***
Tot_ins          -0.062801   0.032587  -1.927 0.054043 .
Education        -0.088026   0.033973  -2.591 0.009610 **
TrafficViolation  0.052466   0.018251   2.875 0.004069 **
Inner            -0.209196   0.074518  -2.807 0.005024 **
Middle           -0.140604   0.072735  -1.933 0.053307 .
AnnualTax         0.277369   0.078854   3.518 0.000441 ***
NoTraining        0.178496   0.055750   3.202 0.001379 **
Mod_Eq            0.355364   0.093949   3.783 0.000158 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.646 on 3366 degrees of freedom
Multiple R-squared:  0.02612, Adjusted R-squared:  0.02265
F-statistic: 7.523 on 12 and 3366 DF,  p-value: 6.997e-14
```

With the model summary above, we can actually see the significance of variable which R tested with t-test and f-test, the model itself had significant with

F-statistic at 7.5 on 12 and 3366 df at 0.001 confidence level

R-squared at 0.022 which takes as a bad fit worse than Model I

Also, with the test of coefficients above we can see that most of coefficients and dummy variables are significant, notably the variables from the literature review did not have that much impact to severity index. Therefore, let us take a look for some interested variables

- Age > Age is a very interesting variable because most of the researches pointed out that the older riders are, more erratic they become, but in this model, it does not seem so, in fact it went opposite way, the researcher suggest it may because of experience, when riders get older, the more proficient they become
- Total experience > It was discussed for many models, but it was insignificant here, it may be because of multicollinearity with other variables such as total insurance

- Gender > It was insignificant here; it may be because of unbalanced data which gives out insignificant coefficients for gender, moreover, it may be interpret as women have less decisive ability than men, but it is insignificant
- Training > Training is discussed as training is crucial and very much significant at 0.001 level of confidence. Therefore, the level of no training has more weight than others, and it is as expected
- Education > Education also crucial as more education level the riders have, the less severity index should be. Since the results is very significant and expected, it holds no more discussion here.

We also test ANOVA (f-test) to test if model IV is better than model II which the test suggests that it is worse, resulting below

```
> anova(val2, val4)
Analysis of Variance Table
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1   3358 9087.2
2   3366 9114.5 -8   -27.343 1.263 0.2582
```

- Fifth Model (V : val5)

```
Call:
lm(formula = si ~ Age + Tot_exp + Gender + Ext_Eq + Education +
    TrafficViolation + NoTraining, data = sq_all)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8613 -0.9471 -0.7400  0.1088  8.2200

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.546679   0.159283   9.710  < 2e-16 ***
Age              -0.005576   0.003208  -1.738 0.082328 .
Tot_exp          -0.002409   0.002515  -0.958 0.338262
Gender            0.019577   0.110663   0.177 0.859595
Ext_Eq           -0.116011   0.026101  -4.445 9.09e-06 ***
Education        -0.097946   0.033478  -2.926 0.003460 **
TrafficViolation  0.063541   0.018167   3.498 0.000475 ***
NoTraining        0.159766   0.052412   3.048 0.002319 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.653 on 3371 degrees of freedom
Multiple R-squared:  0.01585, Adjusted R-squared:  0.0138
F-statistic: 7.754 on 7 and 3371 DF,  p-value: 2.478e-09
```

With the model summary above, we can actually see the significance of variable which R tested with t-test and f-test, the model itself had significant with

F-statistic at 7.7 on 7 and 3371 df at 0.001 confidence level
R-squared at 0.014 which takes as a bad fit worse than Model II

Also, with the test of coefficients above we can see that most of coefficients and dummy variables are significant, notably some variables from the literature review did not have that much impact to severity index. Therefore, let us take a look for some changes in interested variables

- Age > Age is a very interesting variable because most of the researches pointed out that the older riders are, more erratic they become, also in this model, it does seem that it went opposite way, the researcher suggest it may because of experience, when riders get older, the more proficient they become. In other models, age is derived and compounded to other variables, therefore it is significant when alone, but with other variables, it is insignificant.
- Total experience > It was discussed for many models, but it was insignificant here, when tested isolated from other variables, it still hold no significant, therefore, it holds no correlation here.
- Gender > It was insignificant here; it may be because of unbalanced data which gives out insignificant coefficients for gender, moreover, when tested isolated from other variables, it still hold no significant, therefore, it holds no correlation here.

We also test ANOVA (f-test) to test if model V is better than model II which the test suggests that it is better, resulting below

```
> anova(val2, val5)
Analysis of Variance Table
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1   3358 9087.2
2   3371 9210.7 -13   -123.49 3.5102 1.795e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We could use model V for other tests, but it is seemingly worse fitted than others, therefore, we can continue use model II for now.

2. Prediction of behavior

- First Model (I : pre)

```
Call:
glm(formula = psc ~ . - si, family = binomial(link = "logit"),
    data = sq_all)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3367  -0.6667  -0.5511  -0.3498   2.5121

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.4635481  0.4818460  -3.037 0.002386 **
Age              -0.0031258  0.0060539  -0.516 0.605621
Exp_Gen           0.0004191  0.0067159   0.062 0.950239
Exp_Win          -0.0106078  0.0117345  -0.904 0.366004
Exp_App           0.2141570  0.0599111   3.575 0.000351 ***
Total_Ridehour    0.0014026  0.0034767   0.403 0.686642
TrafficViolation  0.0733449  0.0293088   2.502 0.012332 *
Gender           -0.0744463  0.1821182  -0.409 0.682701
MaritalStatus    -0.0584682  0.0946336  -0.618 0.536682
NoNurture        -0.0306166  0.0430606  -0.711 0.477077
Education        -0.1586834  0.0635058  -2.499 0.012464 *
PersonalIncome   -0.1913421  0.0770524  -2.483 0.013018 *
AnnualTax         1.3428898  0.1764077   7.612 2.69e-14 ***
Self_Practice    -0.1187315  0.1135486  -1.046 0.295725
NoTraining        0.2351707  0.0879129   2.675 0.007472 **
Licence_Temp      0.0761522  0.2444127   0.312 0.755366
Licence_Personal -0.1255518  0.1644845  -0.763 0.445282
Licence_Public   -0.3126637  0.2278551  -1.372 0.170000
CCSize            0.0528103  0.0963848   0.548 0.583753
Ext_Eq           -0.1767533  0.0605687  -2.918 0.003520 **
Mod_Eq            0.2821433  0.1391032   2.028 0.042529 *
Inner            -0.3368764  0.1209586  -2.785 0.005352 **
Middle           -0.2411988  0.1156982  -2.085 0.037094 *
Pub               0.5958511  0.3005858   1.982 0.047446 *
Win               0.5939720  0.3058518   1.942 0.052134 .
Tot_ins          -0.0865141  0.0543217  -1.593 0.111244
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3126.6  on 3378  degrees of freedom
Residual deviance: 2956.9  on 3353  degrees of freedom
AIC: 3008.9

Number of Fisher Scoring iterations: 5
```

With the model summary above, we can actually see the significance of variable which R tested with z-test and likelihood ratio test, the model itself had significant with

LR-test at 169.7 on 25 df at 0.001 confidence level

AIC at 3009 which takes as a bad fit

Also, with the test of coefficients above we can see that most of coefficients and dummy variables are insignificant, mostly the variables are like model I in validation. Therefore, it will not be discussed that much here.

That being said, we can now predict and see how predictions are below

Sample code

```
prob<-predict(pre,type="response")
pred<-ifelse(prob>0.5,1,0)
confusionMatrix(data=factor(pred,levels=c(0,1),labels=c("Not
severe","Severe")),reference=factor(sq_all$psc,levels=c(0,1),labels=c
("Not severe","Severe")))
confusion_matrix <- as.data.frame(table(pred, sq_all$psc))
colnames(confusion_matrix) <- c('Prediction','Actual','Freq')
ggplot(data = confusion_matrix, mapping = aes(x = Actual, y =
  Prediction)) + geom_tile(aes(fill = Freq)) +  geom_text(aes(label =
  sprintf("%1.0f", Freq)), vjust = 1) + scale_fill_gradient(low =
  "yellow", high = "red",trans = "log")
```

The results being below,

```
Confusion Matrix and Statistics

            Reference
Prediction   Not severe Severe
  Not severe       2784    583
  Severe              6      6

               Accuracy : 0.8257
                 95% CI : (0.8125, 0.8383)
    No Information Rate : 0.8257
    P-Value [Acc > NIR] : 0.511
                  Kappa : 0.0131

 Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.99785          Specificity : 0.01019
Pos Pred Value : 0.82685       Neg Pred Value : 0.50000
Prevalence : 0.82569           Detection Rate : 0.82391
Detection Prevalence : 0.99645 Balanced Accuracy : 0.50402

             'Positive' Class : Not severe
```

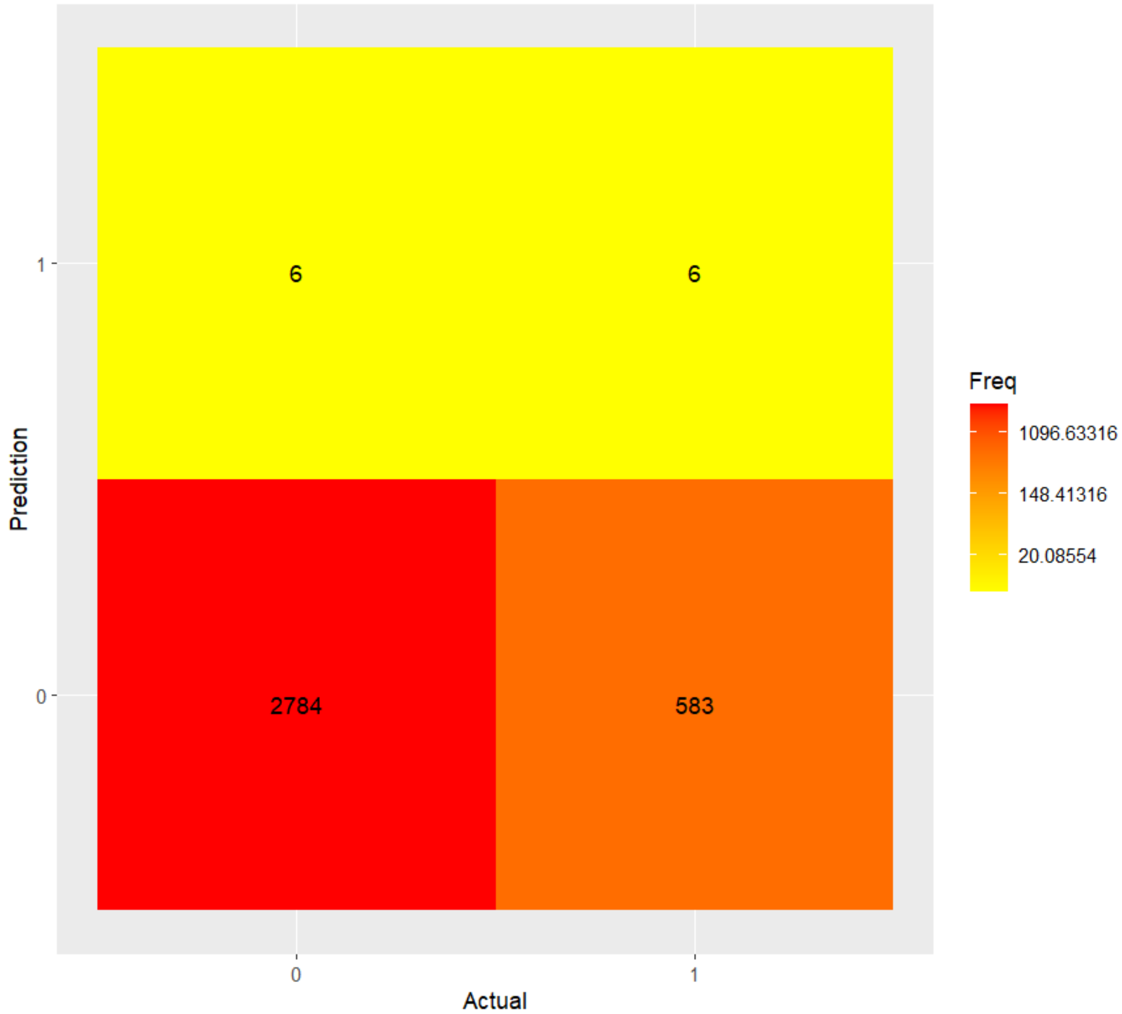

Figure 8 : Predictive Crashes Confusion Matrix

With this figure, we can see that although the model did not fit perfectly and AIC score was bad, it still predict with accuracy 82.57 %. Therefore, we can safely say that this model can predict predictive crashes at good accuracy.

- Second Model (II : pre2)

```
Call:
glm(formula = psc ~ . - si - NoNurture - MaritalStatus - Self_Practice,
    family = binomial(link = "logit"), data = sq_all)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2522  -0.6679  -0.5592  -0.3557   2.4977

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.1226415  0.4513727  -2.487  0.01288 *
Age              -0.0047680  0.0054833  -0.870  0.38454
Total_Ridehour    0.0022045  0.0034539   0.638  0.52330
TrafficViolation  0.0794498  0.0282428   2.813  0.00491 **
Gender           -0.0913203  0.1806001  -0.506  0.61310
Education        -0.1508700  0.0630317  -2.394  0.01669 *
PersonalIncome   -0.2053878  0.0761000  -2.699  0.00696 **
```

```
AnnualTax         1.3115076  0.1742821   7.525 5.26e-14 ***
NoTraining        0.2083833  0.0849675   2.453  0.01419 *
Licence_Temp      0.0138759  0.2433297   0.057  0.95453
Licence_Personal -0.1384838  0.1642508  -0.843  0.39916
Licence_Public   -0.3099387  0.2306479  -1.344  0.17902
CCSize            0.0611078  0.0958297   0.638  0.52369
Ext_Eq           -0.1899984  0.0600115  -3.166  0.00155 **
Mod_Eq            0.2804980  0.1385078   2.025  0.04285 *
Inner            -0.3293406  0.1205649  -2.732  0.00630 **
Middle           -0.2427019  0.1153203  -2.105  0.03533 *
Pub               0.1124194  0.2694378   0.417  0.67651
Win              -0.0548727  0.2525678  -0.217  0.82801
Tot_ins          -0.0776107  0.0533328  -1.455  0.14561
Tot_exp          -0.0005331  0.0051539  -0.103  0.91762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3126.6  on 3378  degrees of freedom
Residual deviance: 2972.0  on 3358  degrees of freedom
AIC: 3014

Number of Fisher Scoring iterations: 5
```

With the model summary above, we used model II for the model prediction, we can actually see the significance of variable which R tested with z-test and likelihood ratio test, the model itself had significant with

LR-test at 154.6 on 20 df at 0.001 confidence level

AIC at 3014 which takes as a bad fit and worse than model I

Also, with the test of coefficients above we can see that most of coefficients and dummy variables are insignificant, mostly the variables are like model I in validation. Therefore, it will not be discussed that much here.

That being said, we can now predict and see how predictions are below

```
Confusion Matrix and Statistics

            Reference
Prediction   Not severe Severe
  Not severe       2784    583
  Severe              6      6

               Accuracy : 0.8257
                 95% CI : (0.8125, 0.8383)
    No Information Rate : 0.8257
    P-Value [Acc > NIR] : 0.511
                  Kappa : 0.0131

 Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.99785          Specificity : 0.01019
Pos Pred Value : 0.82685       Neg Pred Value : 0.50000
Prevalence : 0.82569           Detection Rate : 0.82391
Detection Prevalence : 0.99645 Balanced Accuracy : 0.50402

              'Positive' Class : Not severe
```

With the same results, we can take a look briefly at **Figure 8** and see that although the model did not fit perfectly and AIC score was worse than model I, it still predict with accuracy 82.57 %. Therefore, we can safely say that this model can predict predictive crashes at good accuracy, but with no improvement from model I.

- Model III (I : qpre1)

```
Call:
polr(formula = as.factor(qpsc) ~ . - psc - si, data = sq_all,
    Hess = TRUE, method = c("logistic"))

Coefficients:
                      Value Std. Error   t value
Age               -0.0022566   0.004624 -0.488038
Total_Ridehour    -0.0016265   0.002594 -0.627154
TrafficViolation   0.0739443   0.025733  2.873498
Gender            -0.1716884   0.146617 -1.170998
MaritalStatus     -0.0351259   0.073387 -0.478639
NoNurture          0.1104939   0.031083  3.554852
Education         -0.0902199   0.046614 -1.935459
PersonalIncome     0.0004351   0.059720  0.007285
AnnualTax         -0.3335096   0.096218 -3.466194
Self_Practice     -0.0271336   0.084544 -0.320939
NoTraining         0.3536449   0.071574  4.940987
Licence_Temp       0.3109352   0.192875  1.612109
Licence_Personal -0.1351086   0.123614 -1.092988
Licence_Public    -0.0092703   0.159140 -0.058253
CCSize             0.0001144   0.073894  0.001549
Ext_Eq            -0.1157129   0.044908 -2.576693
Mod_Eq             0.3176538   0.114506  2.774127
Inner             -0.2334777   0.093136 -2.506853
Middle            -0.1810499   0.090397 -2.002820
Pub               -0.1512578   0.203816 -0.742128
Win               -0.1439461   0.178797 -0.805082
Tot_ins           -0.1301331   0.043193 -3.012848
Tot_exp           -0.0024502   0.003958 -0.619129

Intercepts:
    Value   Std. Error t value
0|1 -0.7325  0.3453    -2.1214
1|2  1.1245  0.3472     3.2391
2|3  3.8736  0.4031     9.6087

Residual Deviance: 5894.716
AIC: 5946.716
```

With the model summary above, we can actually see the significance of variable which R tested with t-test and likelihood ratio test, the model itself had significant with

LR-test at 176.9 on 25 df at 0.001 confidence level

AIC at 5947 which takes as a bad fit

With the model using regression same as the model I in both two variations, therefore, it will not be discuss here except cut points between classes which it seems that class 2 and 3 has unusually high cut points which it may affect the results of prediction

That being said, we can now predict and see how predictions are using same code type before, we have here confusion matrix

```
              Reference
Prediction     Least Likely Less Likely More Likely Most Likely
  Least Likely         2696         568           0           0
  Less Likely            91          19           0           0
  More Likely             3           2           0           0
  Most Likely             0           0           0           0

Overall Statistics

               Accuracy : 0.8035
                 95% CI : (0.7897, 0.8168)
    No Information Rate : 0.8257
    P-Value [Acc > NIR] : 0.9996

                  Kappa : 0.0012

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Least Likely    Less Likely   More Likely  Most Likely
Sensitivity            0.96631        0.032258            NA           NA
Specificity            0.03565        0.967384       0.99852            1
Pos Pred Value         0.82598        0.172727            NA           NA
Neg Pred Value         0.18261        0.825635            NA           NA
Prevalence             0.82569        0.174312       0.00000            0
Detection Rate         0.79787        0.005623       0.00000            0
Detection Prevalence   0.96597        0.032554       0.00148            0
Balanced Accuracy      0.50098        0.499821            NA           NA

Likelihood ratio test

  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -3035.8
2  26 -2947.4 23 176.94  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With this results, we can see that although the model did not fit perfectly and AIC score was bad, it still predict with accuracy 80.35 %. Therefore, we can safely say that this model can predict likelihood of severe crashes at good accuracy.

Although this model was intended to be expanded and more thorough investigation of the previous model, it seemed that the model itself confuses and predicts erroneous more than expected.

The results is visualized as below

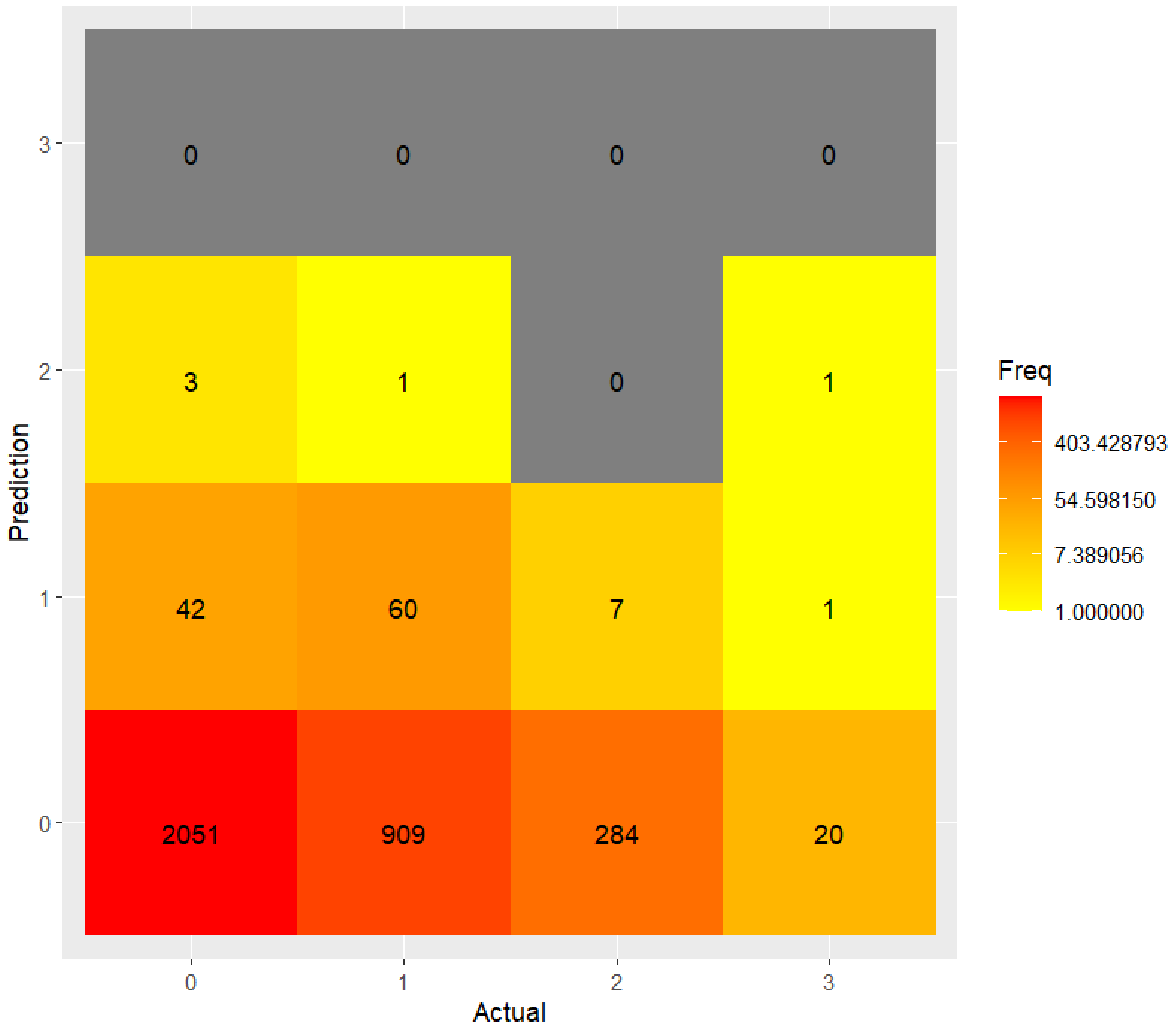

Figure 9 : Quaternary Predictive Severe Crashes Confusion Matrix

- Model IV (II : qpre2)

```
Call:
polr(formula = as.factor(qpsc) ~ Age + Tot_exp + Gender + Ext_Eq +
    Education + TrafficViolation + NoTraining, data = sq_all,
    Hess = TRUE, method = c("logistic"))

Coefficients:
                     Value Std. Error t value
Age              -0.002013   0.003949 -0.5097
Tot_exp          -0.005054   0.003139 -1.6098
Gender           -0.174500   0.143205 -1.2185
Ext_Eq           -0.088010   0.032347 -2.7208
Education        -0.122457   0.042050 -2.9121
TrafficViolation  0.069878   0.024824  2.8149
NoTraining        0.548626   0.064655  8.4855

Intercepts:
    Value   Std. Error t value
0|1  0.0920  0.1957     0.4703
1|2  1.9224  0.2008     9.5717
2|3  4.6717  0.2870    16.2781

Residual Deviance: 5965.75
AIC: 5985.75
```

```
Likelihood ratio test #Model I and II

  #Df  LogLik  Df  Chisq Pr(>Chisq)
1  26 -2947.4
2  10 -2982.9 -16 71.034  6.572e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the model summary above, we can actually see the significance of variable which R tested with t-test and likelihood ratio test, the model itself had significant with

LR-test at 105.9 on 7 df at 0.001 confidence level

AIC at 5985 which takes as a bad fit and worse than model I

With the model using regression same as the model I in both two variations, therefore, it will not be discussed here except cut points between classes which it seems that class 2 and 3 has unusually high cut points and for cut points class 0 and 1 differs from the model I which the researcher suspects that it may affect the results of prediction.

That being said, we can now predict and see how predictions are using same code type before, we have here confusion matrix

```
Reference
Prediction     Least Likely Less Likely More Likely Most Likely
  Least Likely         2756         574           0           0
  Less Likely            30          15           0           0
  More Likely             4           0           0           0
  Most Likely             0           0           0           0

Overall Statistics

               Accuracy : 0.8201
                 95% CI : (0.8067, 0.8329)
    No Information Rate : 0.8257
    P-Value [Acc > NIR] : 0.812

                  Kappa : 0.0219

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Least Likely   Less Likely   More Likely   Most Likely
Sensitivity               0.98781      0.025467            NA          NA
Specificity               0.02547      0.989247      0.998816           1
Pos Pred Value            0.82763      0.333333            NA          NA
Neg Pred Value            0.30612      0.827834            NA          NA
Prevalence                0.82569      0.174312      0.000000           0
Detection Rate            0.81563      0.004439      0.000000           0
Detection Prevalence      0.98550      0.013318      0.001184           0
Balanced Accuracy         0.50664      0.507357            NA          NA
```

With this results, we can see that although the model did not fit perfectly and AIC score was worse than other models, it still predict with accuracy 82.01% and increase with more restricting model. Thus, we can say that there is an optimum point and set of variables that can achieve highest accuracy.

Therefore, we can safely say that this model can predict likelihood of severe crashes at good accuracy.

The results is visualized as below

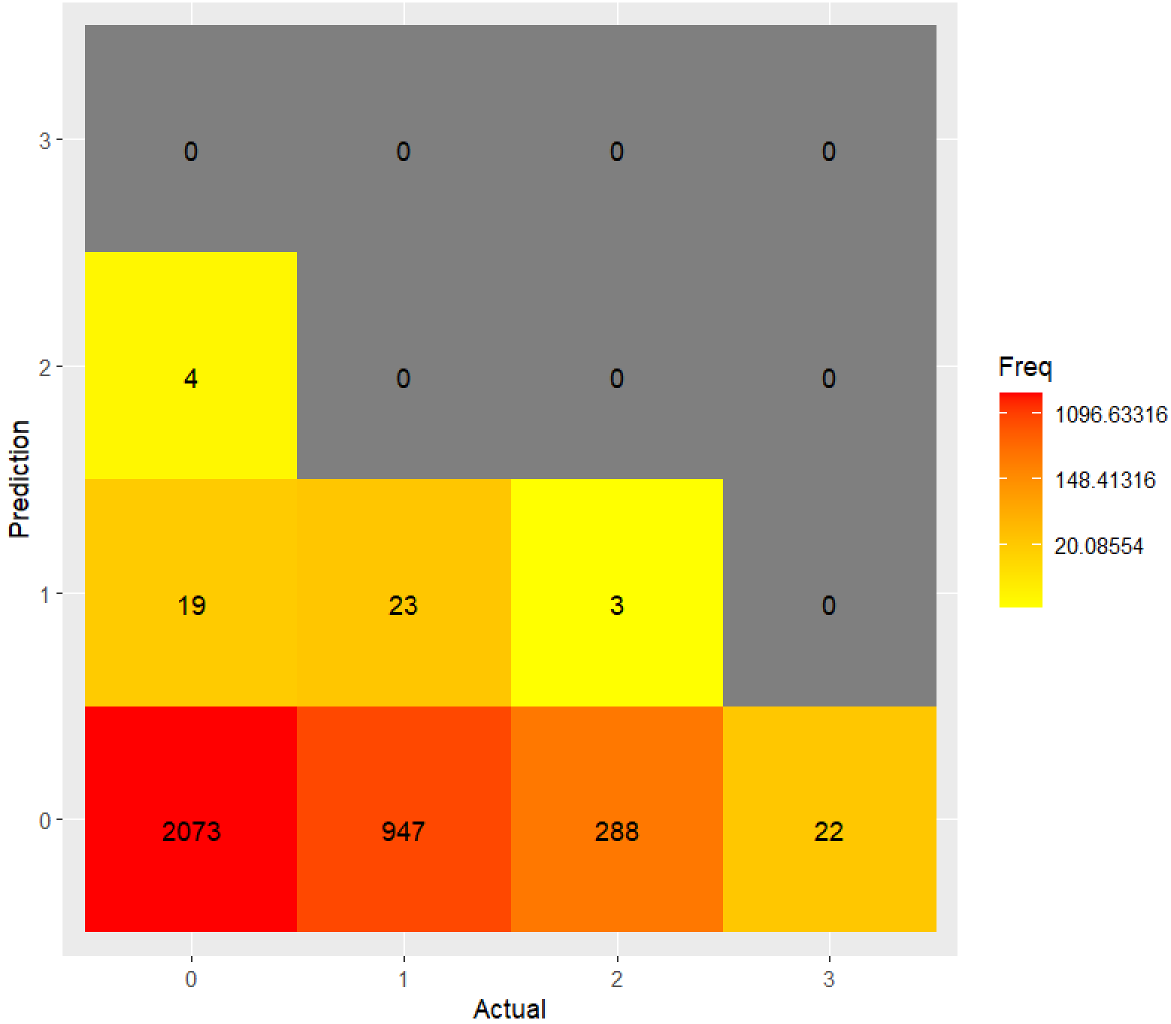

Figure 10 : Quaternary Predictive Severe Crashes Confusion Matrix 2

Summary of results

1. Hypothesis I : Socioeconomic variables such as age, education have significant effects on severity index

Table 4 : Summary table of hypothesis I

Variables	Results	Confidence level
Age	Conditionally significance	0.1
RiderType	No significance	-
Zone	No significance	-
Total_Ridehour	No significance	-
Gender	No significance	-
MaritalStatus	No significance	-
NoNurture	No significance	-
Education	Significant	0.05 – 0.01
PersonalIncome	No significance	-
AnnualTax	Significant	0.001
Compul_Insurance	No significance	-
Vol_Insurance	No significance	-
HealthInsurance	No significance	-
AccidentInsurance	No significance	-
LifeInsurance	Significant	0.001

2. Hypothesis II : Motorcycles related variables such as training, modification have significant effects on severity index

Table 5 : Summary table of hypothesis II

Variables	Results	Confidence level
Exp_Gen	No significance	-
Exp_Win	No significance	-
Exp_App	Significant	0.001
Total_Ridehour	No significance	-
SelfPractice	No significance	-
NoTraining	Significant	0.01 – 0.001
License Personal	No significance	-
License Public	No significance	-
License Temp	No significance	-
NoneLicense	No significance	-
CCSize	No significance	-
Mod_Eq	Significant	0.01 – 0.001
Ext_Eq	Significant	0.05 – 0.001

3. Hypothesis III : the more restricted model is, the more accuracy and distinction it will hold

Table 6 : Summary table of hypothesis III with restriction ranking from most restricted to no restriction (for SI)

Model	R^2	Median Residual	Maximum Residual
Model V	0.0138	-0.7400	8.2200
Model IV	0.02265	-0.7015	8.3102
Model III	0.02293	-0.6998	8.3177
Model II	0.02326	-0.6918	8.2922
Model I	0.04065	-0.5983	8.6935

Table 7 : Summary table of hypothesis III with restriction ranking from most restricted to no restriction (for prediction)

Model	AIC	Accuracy	Residual deviance
Model II	3014	82.57 %	2972
Model I	3008.9	82.57 %	2956.9
Model IV	5985.75	82.01 %	5965.75
Model III	5946.716	80.35 %	5894.716

With that we can summarize our hypotheses verification that tested with statistics as a table below

Table 8 : Summary table for all hypotheses

Hypothesis	Verification
I	Unable to reject for education, annual tax, and life insurance
II	Unable to reject for win experience, no training, extra equipment, and modification equipment
III	Unable to reject

Discussion

The results from the models are quite clear that the severity index that we use is not the best type or method of severity index calculation, it may seem that the type of calculation are mostly the cause of poor performance by all models.

This can be tested by using the feature/ parameter directly to test and regress for the coefficients. We tested for `SumFatality` which yielded and resulted in the model that $R^2 = 0.025$ which is no improvement compared to all models.

There are some variations between the variables in different models, and because of size of variables and model, it seems to be that the cause of variations are from multicollinearity and correlation within the dataset itself which variate through different set of variables used to regress.

If we take a look at the model that use to regress for severity index (SI), we can see that most of variables that regressed are not significant, if they are significant, they are not significant to that much confidence. Therefore, we regress them with simple linear regression, the result that they still have no significant at any level.

This maybe because of the method of regression that we use linear regression, but severity index grows with exponential rate, therefore the model that used may not be suitable for this type of calculation. Thus, the researcher suggest using the package and function `nls` (non-linear regression) to regress on, to better explain the data.

Moreover, the data that we tested later show signs of heteroscedasticity which the results below show rejection of null hypothesis of homoscedasticity, therefore, this may be another reason that the model fitted very poorly

```
> gqtest(val, order.by = ~., data = sq_all, fraction = 6)

      Goldfeld-Quandt test

data:  val

GQ = 17.732, df1 = 1657, df2 = 1656, p-value < 2.2e-16

alternative hypothesis: variance increases from segment 1 to 2
```

The variables that we regressed on and significant, it seems some of them have some irregularities which cause the coefficient turn up in unexpected way, notably, `AnnualTax (Annual tax payment)` which means that if riders paid tax, the more severity index it becomes.

This is not that absurd, but can be explained using correlation, it may be mean that riders that paying tax mean that they still riding motorcycles and working as motorcyclist. If that holds, it means that the more exposure they have and more severity index it will be.

For other relevant variables that we regressed, they mostly acted the way the researcher expected them to be. Therefore, they had no value to discuss further than expected outcome.

If we take a look at the model that use to for prediction for riders' behavior, we use logistic regression and ordered logistic regression, with logistic regression, it is pretty straightforward and restricted to binary outcome, but for ordered logistic, it is more unbounded.

With verified hypothesis III, we may see the reason why restricted/ control model is better, it is because the more thorough we investigate, the model will have more error dividing between the cut points and made erroneous choice that create poor accuracy in the model. Therefore, with less cut points , the model have fewer trouble dividing the data between cut points. Thus, fewer mistakes, more accuracy.

For relevant variables that we regressed, they mostly acted the way the researcher expected them to be. Therefore, they had no value to discuss further than expected outcome.

Lastly, with Python 3.10 using same method of penalty and using machine learning method with library Categorical Boosting (CatBoost), and Extreme Gradient Boosting (XGBoost) we have prediction and confusion matrix accuracy around 64 – 67 % with sample confusion matrix below

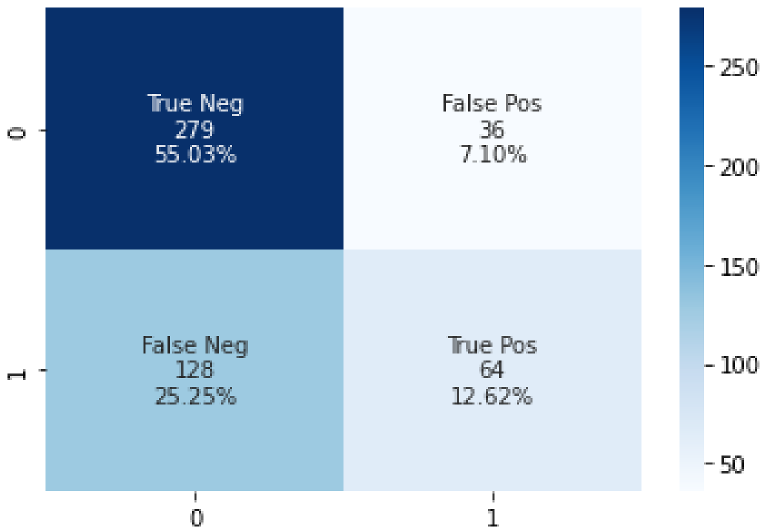

Figure 11 : Confusion matrix from XGBoost with accuracy 67.65%

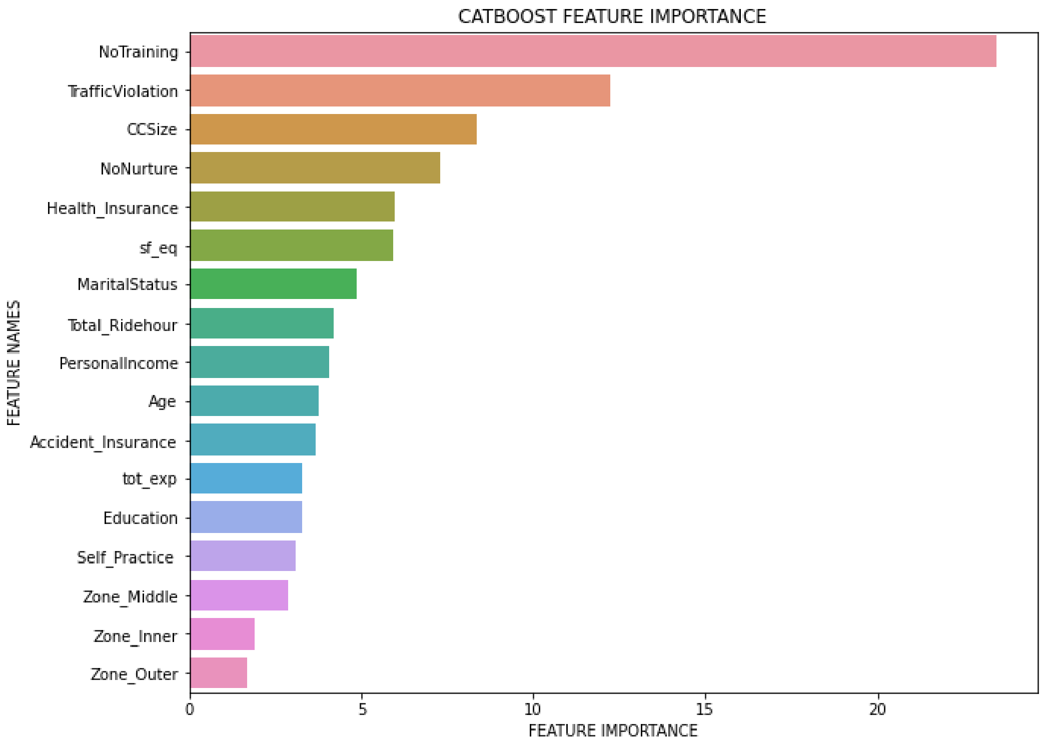

Figure 12 : Feature importance from CatBoost with accuracy 64.02%

It seems that the model suggests nearly the same feature importance, or what we called weights, with some socioeconomic variables that changed and have some significant in the model. It can also be noted that these parameters/ variables also affect the model with some confidences and have importance enough to be filled in the model.

Summary

From the results of analysis, we can summarize and verify hypotheses as follows, hypothesis I is partly rejected for most of variables except education, annual tax, and life insurance, hypothesis II is partly rejected for most of variables except for win experience, no training, extra equipment, and modification equipment and hypothesis III is unable to reject. The severity index model have badness of fit at R^2 near 0.03 and we may choose optimized model II to regress on severity index. The predictive model have badness of fit at AIC around 3000 and accuracy at 82.5 %. The likelihood model have badness of fit at AIC around 6000 and accuracy around 81 %.

Limitations

From the results and summary, we can mostly list the limitation as in 2 ways, first being data-based limitation which come from same group that made data unbalanced and disrupt the models' results. Second being technique-based limitation which discussed in previous sessions that some techniques are not suitable for regress and pre-process on this data.

Conclusions and suggestions

From the results of analysis, we can conclude socioeconomic factors that significantly affect the severity index are education level, annual tax payment, and life insurance with confidence level around 0.1 – 0.001 depending on variable. We may see that policy implication are greatly revolve around education and social policy that regulate and standardize quality of work environments and quality of life

For motorcycles related factors II that significantly affect the severity index are win experience, no training, extra equipment, and modification equipment with confidence level around 0.05 – 0.001. We may see that policy implication can be implement about modification and extra equipment for the motorcycle, moreover, standardizing training, quality of work environments and experience-based work can produce positive effect for the accident prevention.

Lastly the predictive model have badness of fit at AIC around 3000 and accuracy at 82.5 %. The likelihood model have badness of fit at AIC around 6000 and accuracy around 81 %. Most of riders have no encounter or little to none with accident, or severe accident.

References

Baral S., Kanitpong K., (2015). **Factors affecting the severity of motorcycles accidents and casualties in thailand using probit and logit model.** Journal of the Eastern Asia Society for Transportation studies. Vol 11

Oltaye. Z et. Al. (2021). **Prevalence of Motorcycle Accidents and Its Associated Factors Among Road Traffic Accident Patients in Hawassa University Comprehensive Specialized Hospital, 2019.** Open Access Emergency Medicine.

Song, X. et. Al. (2021). **Determinants and Prediction of Injury Severities in Multi-Vehicle-Involved Crashes.** Int. J. Environ. Res. Public Health 2021, 18, 5271. https://doi.org/10.3390/ijerph18105271

Cao, Yi. et. Al. (2020) **An Assessment Method of Urban Traffic Crash Severity Considering Traveling Delay and Non-Essential Fuel Consumption of Third Parties.** Sustainability 2020, 12, 6806; doi:10.3390/su12176806

Champahom, et. Al. (2022). **Factors affecting severity of motorcycle accidents on Thailand's arterial roads: Multiple correspondence analysis and ordered logistics regression approaches.** IATSS Research 46 (2022) 101–111.

Chumpawadee, U. et. Al. (2015). **factors related to motorcycle accident risk behavior among university students in northeastern thailand.** Southeast Asian J Trop Med Public Health. Vol 46 No. 4 July 2015.

NCDOT. (2013). **Chapter 14 : Severity of accident.**

European Commission. (2015). **Road accident statistics.**

Appendix

R Code for usage

```
library(corrplot)
library(readxl)
library(xlsx)
library(haven)
library(dplyr)
library(Hmisc)
library(car)
library(pscl)
library(caret)
library(ggplot2)
library(lmtest)
library(MASS)

#Data Cleaning
sq_all <- read_excel("C:/Users/north/Desktop/CU/Y3T2/Stats Trans
Eng/Term Paper/SRQ.xlsx",sheet = 1)
sq_all$Ext_Eq <- sq_all$Equip_SideBox + sq_all$Equip_RearBox +
sq_all$Equip_SideBag + sq_all$Equip_FrontStorage +
sq_all$Equip_PhoneEar + sq_all$Equip_Windshield +
sq_all$Equip_PhoneGrabber
sq_all$Equip_SideBox <- sq_all$Equip_RearBox <- sq_all$Equip_SideBag
<- sq_all$Equip_FrontStorage <- sq_all$Equip_PhoneEar <-
sq_all$Equip_Windshield <- sq_all$Equip_PhoneGrabber <- NULL
sq_all$Mod_Eq <- sq_all$Modify_Engine + sq_all$Modify_intake +
sq_all$Modify_Wheel + sq_all$Modify_ColorBody
sq_all$Modify_Engine <- sq_all$Modify_intake <- sq_all$Modify_Wheel
<- sq_all$Modify_ColorBody <- sq_all$Modify_None<- NULL
summary(sq_all)

sq_all$Inner<-ifelse(sq_all$Zone=="Inner",1,0)
sq_all$Middle<-ifelse(sq_all$Zone=="Middle",1,0) #outer is base
sq_all$Pub <- ifelse(sq_all$RiderType==1,1,0)
sq_all$Win <- ifelse(sq_all$RiderType==2,1,0) #app is base
hist(sq_all$RiderType, main = 'Histogram of Rider Type' ,xlab="Rider
Type", ylab="No. of rider")
sq_all$RiderType <- sq_all$Zone <- NULL
sq_all$NoneLicence <- NULL #No License = base

#Prelim
nprel <- lm(SumFatality ~ 1, data =sq_all)
nprel2 <- lm(SumInjured ~ 1, data =sq_all)
nprel3 <- lm(SumNear ~ 1, data =sq_all)

prel <- lm(SumFatality ~ . -SumInjured - SumNear, data =sq_all)
summary(prel)
anova(nprel, prel)
prel2 <- lm(SumInjured ~ . - SumNear -SumFatality, data =sq_all)
summary(prel2)
anova(nprel2, prel2)
prel3 <- lm(SumNear ~ .-SumInjured -SumFatality, data =sq_all)
summary(prel3)
```

```
anova(nprel3, prel3)

#Adjusted Rate
nprel <- lm(SumFatality/Total_Ridehour ~ 1, data =sq_all)
nprel2 <- lm(SumInjured/Total_Ridehour ~ 1, data =sq_all)
nprel3 <- lm(SumNear/Total_Ridehour ~ 1, data =sq_all)

prel <- lm(SumFatality/Total_Ridehour ~ . -SumInjured - SumNear,
data =sq_all)
summary(prel)
anova(nprel, prel)
prel2 <- lm(SumInjured/Total_Ridehour ~ . - SumNear -SumFatality,
data =sq_all)
summary(prel2)
anova(nprel2, prel2)
prel3 <- lm(SumNear/Total_Ridehour ~ .-SumInjured -SumFatality, data
=sq_all)
summary(prel3)
anova(nprel3, prel3)

sum(is.na(sq_all))
sq_all[is.na(sq_all)] <- 0

sq_all$si <- (sq_all$SumFatality*9 + sq_all$SumInjured*5 +
sq_all$SumNear*1)/(sq_all$SumFatality+sq_all$SumInjured+sq_all$SumNe
ar)
sq_all$psc <- ifelse(sq_all$si > 1,1,0)

summary(sq_all$si)
sq_all$si[is.na(sq_all$si)] <- 0
hist(sq_all$psc, main = 'Histogram of Predictive Crashes'
,xlab="Predictive Crashes", ylab="No. of rider")
summary(sq_all$psc)
sq_all$psc[is.na(sq_all$psc)] <- 0
hist(sq_all$si, main = 'Histogram of Adjusted Severity Index'
,xlab="Adjusted Severity Index", ylab="No. of rider")
hist(sq_all$SumFatality)
hist(sq_all$SumInjured)
hist(sq_all$SumNear)
hist(sq_all$Age,main = "Histogram of Riders' Age" ,xlab="Riders'
Age", ylab="No. of rider")
hist(sq_all$Exp_Win)
hist(sq_all$Gender,main = "Histogram of Riders' Gender"
,xlab="Riders' Gender", ylab="No. of rider")
hist(sq_all$Total_Ridehour)
hist(sq_all$Mod_Eq, main = 'Histogram of Modification Equipment'
,xlab="Modification Equipment", ylab="No. of rider")

names(sq_all)
#Validation of behavior
sq_all$SumFatality <- sq_all$SumInjured <- sq_all$SumNear <- NULL
val <- lm(si ~ . -psc , data = sq_all )
summary(val)
```

```
#Decrease of data
sq_all.cor = cor(sq_all)
corrplot(sq_all.cor)

sq_all$Tot_ins <- sq_all$Life_Insurance + sq_all$Accident_Insurance
+ sq_all$Health_Insurance + sq_all$Compul_Insurance +
sq_all$Vol_Insurance
sq_all$Life_Insurance <- sq_all$Accident_Insurance <-
sq_all$Health_Insurance <- sq_all$Compul_Insurance <-
sq_all$Vol_Insurance <- NULL
sq_all$Tot_ins[is.na(sq_all$Tot_ins)] <- 0

sq_all$Tot_exp <- sq_all$Exp_Win + sq_all$Exp_App + sq_all$Exp_Gen
sq_all$Exp_Win <- sq_all$Exp_App <- sq_all$Exp_Gen <- NULL
sq_all$Tot_ins[is.na(sq_all$Tot_ins)] <- 0

val2 <- lm(si ~ . -psc -NoNurture - MaritalStatus -Self_Practice ,
data = sq_all)
summary(val2)
anova(val, val2)

#No Lit Review Suggestion
#Choose only significant variable group

val3 <- lm(si ~ AnnualTax + NoTraining + Mod_Eq + Ext_Eq + Inner +
Middle
           +Tot_ins +Tot_exp + TrafficViolation + Education, data =
sq_all)
summary(val3)
anova(val, val3)
anova(val2, val3)

#Take Lit Review Suggestion
val4 <- lm(si ~ Age + Tot_exp + Gender + Ext_Eq + Tot_ins +
Education +
             TrafficViolation + Inner + Middle + AnnualTax +
NoTraining + Mod_Eq, data = sq_all)
summary(val4)
anova(val2, val4)

val5 <- lm(si ~ Age + Tot_exp + Gender + Ext_Eq +  Education +
             TrafficViolation + NoTraining, data = sq_all)
summary(val5)
anova(val2, val5)
anova(val3, val5)
anova(val4,val5)


#Prediction of behavior
npre <- glm(psc ~ 1, family = binomial(link = "logit"), data =
sq_all )
summary(npre)

pre <- glm(psc ~ .-si, family = binomial(link = "logit"), data =
sq_all )
summary(pre)
```

```
lrtest(npre, pre)
prob<-predict(pre,type="response")
pred<-ifelse(prob>0.5,1,0)
confusionMatrix(data=factor(pred,levels=c(0,1),labels=c("Not
severe","Severe")),reference=factor(sq_all$psc,levels=c(0,1),labels=
c("Not severe","Severe")))
confusion_matrix <- as.data.frame(table(pred, sq_all$psc))
colnames(confusion_matrix) <- c('Prediction','Actual','Freq')
ggplot(data = confusion_matrix, mapping = aes(x = Actual, y =
Prediction)) +
  geom_tile(aes(fill = Freq)) +
  geom_text(aes(label = sprintf("%1.0f", Freq)), vjust = 1) +
  scale_fill_gradient(low = "yellow", high = "red",trans = "log")

#Improvement
pre2 <- glm(psc ~ . -si -NoNurture - MaritalStatus -Self_Practice,
family = binomial(link = "logit"), data = sq_all )
summary(pre2)
lrtest(pre, pre2)
prob2<-predict(pre,type="response")
pred2<-ifelse(prob2>0.5,1,0)
confusionMatrix(data=factor(pred2,levels=c(0,1),labels=c("Not
severe","Severe")),reference=factor(sq_all$psc,levels=c(0,1),labels=
c("Not severe","Severe")))
confusion_matrix2 <- as.data.frame(table(pred2, sq_all$psc))
colnames(confusion_matrix2) <- c('Prediction','Actual','Freq')
ggplot(data = confusion_matrix2, mapping = aes(x = Actual, y =
Prediction)) +
  geom_tile(aes(fill = Freq)) +
  geom_text(aes(label = sprintf("%1.0f", Freq)), vjust = 1) +
  scale_fill_gradient(low = "yellow", high = "red",trans = "log")

#No improvement

sq_all$qpsc <- ifelse(sq_all$si >= 9,3,ifelse(sq_all$si >= 5
,2,ifelse(sq_all$si >= 1,1,0)))
summary(sq_all$qpsc)
hist(sq_all$qpsc, main = 'Histogram of Quaternary Predictive Severe
Crashes' ,xlab="Quaternary Predictive Severe Crashes", ylab="No. of
rider")
nqpre <- polr(as.factor(qpsc) ~ 1, data = sq_all, Hess=TRUE, method
= c("logistic"))
summary(nqpre)
lrtest(nqpre,qpre1)
qpre1 <- polr(as.factor(qpsc) ~ . - psc - si, data = sq_all,
Hess=TRUE, method = c("logistic"))
summary(qpre1)
pred3<-predict(qpre1)
confusionMatrix(data=factor(pred3,levels=c(0,1,2,3),labels=c("Least
Likely","Less Likely","More Likely","Most
Likely")),reference=factor(sq_all$psc,levels=c(0,1,2,3),labels=c("Le
ast Likely","Less Likely","More Likely","Most Likely")))
confusion_matrix3 <- as.data.frame(table(pred3, sq_all$qpsc))
colnames(confusion_matrix3) <- c('Prediction','Actual','Freq')
ggplot(data = confusion_matrix3, mapping = aes(x = Actual, y =
Prediction)) +
```

```
  geom_tile(aes(fill = Freq)) +
  geom_text(aes(label = sprintf("%1.0f", Freq)), vjust = 1) +
  scale_fill_gradient(low = "yellow", high = "red",trans = "log")

qpre2 <- polr(as.factor(qpsc) ~ Age + Tot_exp + Gender + Ext_Eq +
Education + TrafficViolation + NoTraining , data = sq_all,
Hess=TRUE, method = c("logistic"))
summary(qpre2)
lrtest(qpre1,qpre2)
lrtest(nqpre,qpre2)
pred4<-predict(qpre2)
confusionMatrix(data=factor(pred4,levels=c(0,1,2,3),labels=c("Least
Likely","Less Likely","More Likely","Most
Likely")),reference=factor(sq_all$psc,levels=c(0,1,2,3),labels=c("Le
ast Likely","Less Likely","More Likely","Most Likely")))
confusion_matrix4 <- as.data.frame(table(pred4, sq_all$qpsc))
colnames(confusion_matrix4) <- c('Prediction','Actual','Freq')
ggplot(data = confusion_matrix4, mapping = aes(x = Actual, y =
Prediction)) +
  geom_tile(aes(fill = Freq)) +
  geom_text(aes(label = sprintf("%1.0f", Freq)), vjust = 1) +
  scale_fill_gradient(low = "yellow", high = "red",trans = "log")
```