Leading Causes of Accidents in United Kingdom

Yashitha Jeet

Background

Accidents has been a leading cause of death across the world for many years. Today it has reached a point where 1.3 million people die in road crashes every year and this averages to about 3,200 deaths a day. The causes of accidents range could from road condition, pedestrians, vehicle drivers, environmental factors like weather and other unforseen circumstances. It is imperative to understand and analyse these underylying causes as this will allow traffic engineers to come up with preventive measures to reduce accidents rates and thereby reduce the loss of life and properties. A statistical analysis of existing accident data will allow us to identify specific road conditions or junctions that are accident prone i.e., predict the probablitiy of accident severity in those areas and obtain valuable insights on taking steps for the future. Studying accident data will not only save lives but will also save money and time. It will alllow countries to understand the flaws in their road system and develop a more streamlined road system.
Additionally, a major advantage of analysing accident data is the information that it can provide to companies that are focused on developing self-driving cars. With Pittsburgh being one of the biggest hubs for this technology, I gained a lot of exposure to self-driving cars and also got the opportunity to be a passenger a few times. I believe that gaining more insight on accident prone conditions and understanding the causes could benefit the companies in training the autonomous vehicles.
This project focuses on data that is a subset of the traffic data collected by the UK government from 2000 to 2016 that consisted of over 1.6 million accidents obtain from police reports. The raw dataset at large can be obtained from the UK Department of Transport. For the purpose of this project, we will be looking at the accidents reported from 2012 to 2014 and analyzing specific variables.

Variables

This data-set was taken from an open-source database known as kaggle and consisted of 464,697 observations (n = 464,697) with 33 variables of different factors related to accidents. This project only looked at 9 variables (p = 9) which were both qualitative and quantitative. The observations of each variables were already reported in the data-base and no calculations or research had to be done to obtain the values. The variables of interest include the following:

Response Variables:

Qualitative: - Accident Severity: Three levels ranging from 1 (least severe) to 3 (most severe)

Predictor Variables

Qualitative:

- Junction Control: The type of junction control at the accident scene. (Levels: Authorised person, Automatic traffic signal, Giveway or uncontrolled, Stop Sign)
- Urban/Rural: Urban = 1, Rural = 2

- Weather Condition: The type of weather condition on the day of the accidents. (Levels: Fine with high winds, Fine without high winds, Fog or mist, Raining with high winds, Raining without high winds, Other)
- Road Type: The type of road (Levels: dual carriageway, one way street, roundabout, single carraigeway, slip road and unknown)
- Road Surface Condition: The condition of the surface of the road (Levels: Dry, Wet/Damp, Flood (Over 3cm of water), Frost/Ice, Snow)
- Day of the week: Ranging from 1 (Monday) to 7 (Sunday)

Quantitative:

- Number of Vehicles: The number of vehicles involved in the accident
- Speed Limit: Speed limit of the vehicle causing the accident

The other variables provided information about the presence of police force, number of casualities, road number, road class etc. The reason for choosing the aformentioned variables were because these variables seem to be potential causes for accidents and analysing them would provide interesting insights.

Hypothesis

We hypothesize that the severity of accidents is greater on weekends, urban areas and near traffic signals. We also hypothesize that speed limit, road type, and weather conditions have a significant effect on accident severity.

Since our variable of interest is a categorical ordinal variable with three levels - 1, 2, 3 we will be treating these as a count variable with a poisson distribution. Therefore, we will use a generalized logistic regression for our analysis with for Poisson family of distribution. The model is as follows:

Data Organization and Data Cleaning

The data-set was in the form of a well-formatted csv file. Since the dataset was from an open-source and already formatted, there was not much data cleaning that needed to be done. The cleaning only involed the removal of variables as mentioned above. Apart from that, the only anomalies was the missing values in the observations for the Junction Control variable. Due to the large number of missing values, we coded these as NA so that they would not effect our analysis. No other variable had any missing values. Additionally, the qualitative variables needed to be converted to factors in order to be treated as categorical variables in the analysis.

```
# Data Cleaning
accidents = read.csv("accidents_2012_to_2014.csv")
# Only keep variables of interest
accidents = accidents[-c(1:6,9, 10, 12:16, 19, 21:25, 28:29, 31:33)]
attach(accidents)
dat = accidents
```

```r
# Data Cleaning -- removing empty cells and converting categorical variables to factors
accidents$Day_of_Week = factor(accidents$Day_of_Week,
                               levels = sort(unique(accidents$Day_of_Week)))
accidents$Road_Type = factor(accidents$Road_Type,
                             levels = unique(accidents$Road_Type))
accidents$Junction_Control = factor(accidents$Junction_Control,
                                    levels = unique(accidents$Junction_Control))
accidents$Weather_Conditions = factor(accidents$Weather_Conditions,
                                      levels = unique(accidents$Weather_Conditions))
accidents$Road_Surface_Conditions = factor(accidents$Road_Surface_Conditions,
                                           levels = unique(accidents$Road_Surface_Conditions))
accidents$Urban_or_Rural_Area = factor(accidents$Urban_or_Rural_Area,
                                       levels = unique(accidents$Urban_or_Rural_Area))

head(accidents,5 )
```

```
##   Accident_Severity Number_of_Vehicles Day_of_Week          Road_Type
## 1                 3                  2           5 Single carriageway
## 2                 3                  2           4 Single carriageway
## 3                 3                  2           3     One way street
## 4                 3                  1           4 Single carriageway
## 5                 3                  1           3 Single carriageway
##   Speed_limit         Junction_Control      Weather_Conditions
## 1          30 Automatic traffic signal Fine without high winds
## 2          30  Giveway or uncontrolled Fine without high winds
## 3          30  Giveway or uncontrolled Fine without high winds
## 4          30  Giveway or uncontrolled Fine without high winds
## 5          30  Giveway or uncontrolled Fine without high winds
##   Road_Surface_Conditions Urban_or_Rural_Area
## 1                     Dry                  1
## 2                     Dry                  1
## 3                     Dry                  1
## 4                     Dry                  1
## 5                     Dry                  1
```

Analysis

We first do an explanatory data analysis on all our variables of interest. From the summary output and the tables we get an idea of the distribution of each categorical variable in the different levels. As for the two quantitative variables, we displayed a histogram for Speed Limit and noticed majority of the vehicles involved in the accident had a speed limit between 20 to 30 mph.

```r
# Explanatory Data Analysis
summary(accidents)[,-1]
```

```
##  Number_of_Vehicles Day_of_Week              Road_Type
##  Min.   : 1.000     1:50304     Single carriageway:351268
##  1st Qu.: 1.000     2:66725     One way street    :  9074
##  Median : 2.000     3:70316     Roundabout        : 31852
##  Mean   : 1.828     4:69835     Dual carriageway  : 65998
##  3rd Qu.: 2.000     5:70644     Slip road         :  4827
##  Max.   :67.000     6:76054     Unknown           :  1678
##                     7:60819
```

```
##   Speed_limit                    Junction_Control
##  Min.   :10.00   Automatic traffic signal: 50208
##  1st Qu.:30.00   Giveway or uncontrolled :232915
##  Median :30.00                           :178610
##  Mean   :38.23   Authorised person       :   677
##  3rd Qu.:40.00   Stop Sign               :  2287
##  Max.   :70.00
##
##                   Weather_Conditions              Road_Surface_Conditions
##  Fine without high winds   :373167   Dry                       :319370
##  Raining without high winds: 57060   Wet/Damp                  :132745
##  Other                     :  8272   Frost/Ice                 :  8140
##  Unknown                   :  8215   Snow                      :  2824
##  Raining with high winds   :  7120   Flood (Over 3cm of water):   863
##  Fine with high winds      :  5011                             :   755
##  (Other)                   :  5852
##  Urban_or_Rural_Area
##  1:307896
##  2:156801
##
##
##
##
##
```

```
table(Day_of_Week)
```

```
## Day_of_Week
##     1     2     3     4     5     6     7
## 50304 66725 70316 69835 70644 76054 60819
```

```
table(Road_Type)
```

```
## Road_Type
##   Dual carriageway     One way street         Roundabout
##              65998               9074              31852
## Single carriageway          Slip road            Unknown
##             351268               4827               1678
```

```
table(Junction_Control)
```

```
## Junction_Control
##                                 Authorised person Automatic traffic signal
##                   178610                       677                     50208
##  Giveway or uncontrolled               Stop Sign
##                   232915                    2287
```

```
table(Weather_Conditions)
```

```
## Weather_Conditions
##       Fine with high winds    Fine without high winds
##                       5011                     373167
##                Fog or mist                      Other
##                       2411                      8272
##    Raining with high winds Raining without high winds
##                       7120                     57060
##    Snowing with high winds Snowing without high winds
```

```
##                          733                         2708
##                      Unknown
##                         8215
```

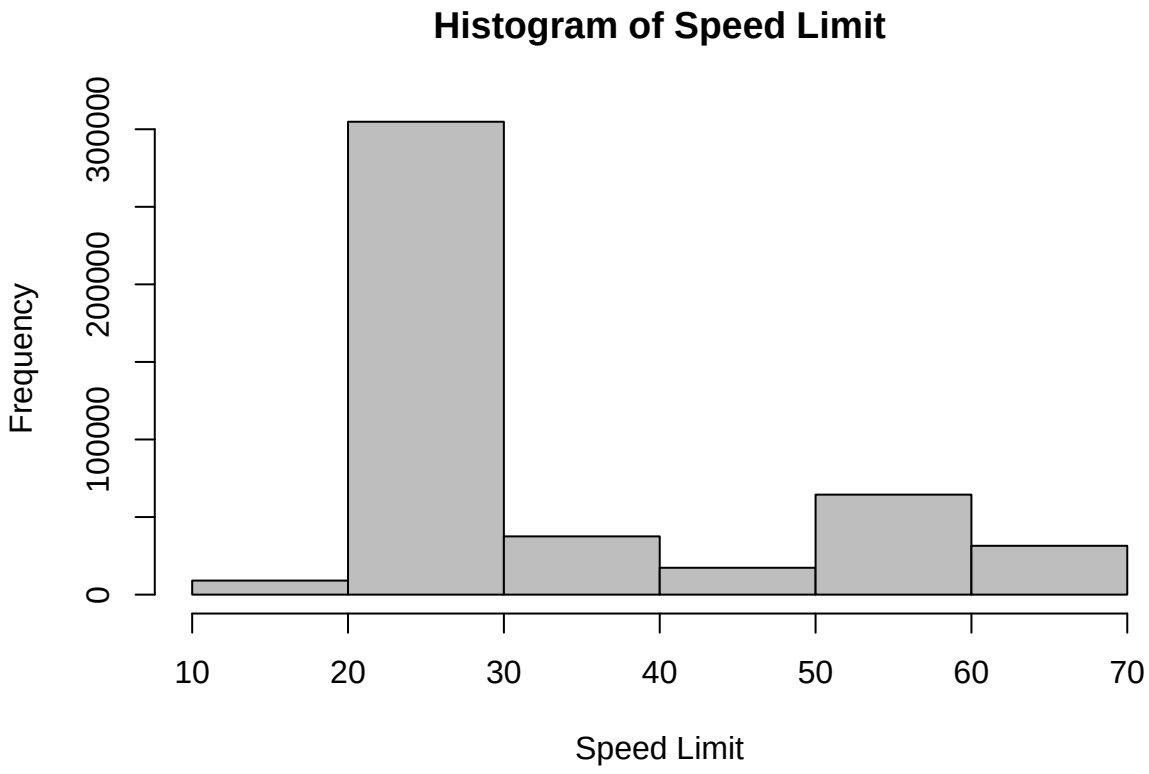
```
table(Road_Surface_Conditions)
```

```
## Road_Surface_Conditions
##                                          Dry
##                          755           319370
## Flood (Over 3cm of water)           Frost/Ice
##                          863             8140
##                         Snow          Wet/Damp
##                         2824           132745
```

```
table(Urban_or_Rural_Area)
```

```
## Urban_or_Rural_Area
##      1      2
## 307896 156801
```

```
hist(Speed_limit, breaks = 8, xlab = "Speed Limit", main = "Histogram of Speed Limit", col = "grey")
```



We will not fit a generalized logistic model to our variables, and model it as follows:

```
library(knitr)
lm.fit = glm(accidents$Accident_Severity ~ accidents$Day_of_Week + accidents$Road_Type + accidents$Speed
               accidents$Junction_Control + accidents$Weather_Conditions  + accidents$Road_Surface_Condi
             accidents$Number_of_Vehicles + accidents$Urban_or_Rural_Area, family = poisson)
#summary(lm.fit)
```

Based on the results from this model, we notice that at least one level of all the predictor variables is significant $p < 0.05$. We will be analyzing each predictor variable to understand its effect on Accident

```
Call:
glm(formula = Accident_Severity ~ ., family = poisson, data = accidents)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.42001   0.05602   0.08255   0.11159   0.19717

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                  1.032e+00  5.052e-03 204.347  < 2e-16 ***
Number_of_Vehicles                           1.544e-02  1.224e-03  12.609  < 2e-16 ***
Day_of_Week2                                 1.253e-02  3.522e-03   3.557 0.000375 ***
Day_of_Week3                                 1.223e-02  3.485e-03   3.510 0.000449 ***
Day_of_Week4                                 1.309e-02  3.490e-03   3.752 0.000175 ***
Day_of_Week5                                 1.191e-02  3.482e-03   3.419 0.000628 ***
Day_of_Week6                                 1.100e-02  3.431e-03   3.205 0.001352 **
Day_of_Week7                                 4.029e-03  3.600e-03   1.119 0.262985
Road_TypeOne way street                      5.802e-03  6.323e-03   0.918 0.358799
Road_TypeRoundabout                          2.228e-02  3.535e-03   6.303 2.91e-10 ***
Road_TypeDual carriageway                    1.842e-02  2.852e-03   6.458 1.06e-10 ***
Road_TypeSlip road                           3.485e-02  8.646e-03   4.031 5.55e-05 ***
Road_TypeUnknown                             1.892e-02  1.447e-02   1.307 0.191131
Speed_limit                                 -6.756e-04  9.802e-05  -6.893 5.46e-12 ***
Junction_ControlGiveway or uncontrolled     -3.160e-03  2.965e-03  -1.066 0.286401
Junction_Control                            -1.268e-02  3.140e-03  -4.039 5.37e-05 ***
Junction_ControlAuthorised person            3.767e-03  2.284e-02   0.165 0.868976
Junction_ControlStop Sign                    6.101e-03  1.264e-02   0.483 0.629298
Weather_ConditionsRaining without high winds 1.069e-02  3.397e-03   3.145 0.001659 **
Weather_ConditionsOther                      9.418e-03  6.835e-03   1.378 0.168218
Weather_ConditionsSnowing without high winds 7.743e-03  1.510e-02   0.513 0.608177
Weather_ConditionsRaining with high winds    1.179e-02  7.445e-03   1.584 0.113239
Weather_ConditionsFine with high winds       2.135e-04  8.499e-03   0.025 0.979954
Weather_ConditionsUnknown                    1.183e-02  6.760e-03   1.750 0.080181 .
Weather_ConditionsFog or mist                2.171e-03  1.231e-02   0.176 0.860040
Weather_ConditionsSnowing with high winds    1.726e-02  2.421e-02   0.713 0.475803
Road_Surface_ConditionsWet/Damp             -7.398e-04  2.555e-03  -0.290 0.772120
Road_Surface_ConditionsFrost/Ice             2.224e-02  6.914e-03   3.217 0.001297 **
Road_Surface_ConditionsSnow                  2.415e-02  1.568e-02   1.541 0.123397
Road_Surface_ConditionsFlood (Over 3cm of water) 1.979e-02  2.052e-02   0.964 0.334833
Road_Surface_Conditions                      7.372e-03  2.203e-02   0.335 0.737867
Urban_or_Rural_Area2                        -1.124e-02  2.607e-03  -4.313 1.61e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 30436  on 464696  degrees of freedom
Residual deviance: 29865  on 464665  degrees of freedom
AIC: 1389717
```

Figure 1: Summary Ouput of GLM

Severity.
- For `Day_of_Week`, only the 7th day i.e., does not seem to be significant with a p-value of 0.262. In particular, we notice that compared to Monday, the other days of week lead to a percentage increase in the mean Accident Severity ranging from 1.1% to 1.6%, holding all other variables fixed. In particular, we notice that Sunday increases the mean Accident Severity count by 0.4%, while Tuesday incraeses it by 1.26%.
- The two types with single carraigeway as a reference group, the two road types that do not seem to be significant are one way streets and unknown type. All the road types have a positive coefficient, indicating a percentage increase in mean Accident Severity when compared to a single carraigeway, holding the other variables constant. In specific, if the type of road is a slip road, there is a 3.5% increase in the mean Accident Severity in comparison to a single carriageway road, holding all other variables constant.
- Speed Limit has a negative significant coefficient with p-value less than 0.05. This negative coefficient indicates that the Speed limit leads to a 0.07% decrease in the mean Accident Severity.
- Junction Control is not a significant predictor of Accident Severity. We should note that in the summary output, the only variable that seems to have a significant p-value is indicative of the values that were coded as NA. The uncontrolled and giveway junctions have a negative coefficient, indicating a decreasing in mean Accident Severity count in comparison to automatic traffic signal junctions, holding all other variables constant.
- The only significant level of the weather condition predictor is raining weather without high winds. It has a positive coefficient indicating a 10.7% increase in the mean count of Accident Severity, compared to fine weather without high winds and holding all other variables constant.
- The only significant level of road surface condition is the condition with frost and ice. This has a positive coefficient, indicating a 22% increase in the mean Accident Severity count, compareed to dry roads and keeping other variables fixed.
- Lastly, whether the location was urban or rural is a significant predictor of Accident Severity. In specific, when you shift from urban to rural, the mean count of Accident Severity decreases by 1.2%, holding all other variables constant.

```r
library(boot)
# Got this function online
#devi = function(y, eta) {
#         deveta = y * log(eta) - eta
#         devy = y * log(y) - y
#         devy[y == 0] = 0
#         mean(2 * (devy - deveta))
#}
#set.seed(1)
#cv.err = cv.glm(data=accidents, lm.fit, K = 5)
#print(paste("Cross-valdiation error", round(cv.err$delta[1],3)))
```

```
> cv.err
$call
cv.glm(data = accidents, glmfit = lm.fit, cost = devi, K = 5)

$K
[1] 5

$delta
[1] 0.06427641 0.06427550
```

Figure 2: Cross Validation

Before we are able to draw any conclusions or make predictions, we need to test the generalized logistic model above and measure model performance. In order to this, we could use cross-validation or bootstrapping.

Since our model has such a large sample size, variance becomes less of an issue and we prefer a technique that is computationally efficient. Thus, we will use k-fold cross validation in order to test model fitness and we will set k as 5 folds. We will use the in-built cross-validation function in R, however, since we are dealing with a poisson family of distributions, we will use the deviance as a basis for the goodness of fit model. Thus, we notice that we get a cross-validation error of 6.43 %.

The next step is to do model selection and in order do this we will use LASSO regression analysis. LASSO will perform variable selectin on our model in order to enhance its prediction accuracy. The given variables in our model could potentially have multi-collinearity so doing a LASSO regression will allow us to gain insight on which variables, if any, to remove from our model and will thereby answer our research question on identifying the major causes.

```
#library(glmnet)
#x = model.matrix(accidents$Accident_Severity ~ accidents$Day_of_Week + accidents$Road_Type + accidents
              #accidents$Junction_Control + accidents$Weather_Conditions  + accidents$Road_Surface_Cond
            #accidents$Number_of_Vehicles + accidents$Urban_or_Rural_Area)
#y = accidents$Accident_Severity
#cv.err2 = cv.glmnet(x, y, family="poisson", alpha=1)
#error = cv.err2$cvm[cv.err2$lambda == 0] * 100
#best.lamda = cv.err2$lambda.min
#best.model = glmnet(x, y, alpha = 1, family="poisson")
#print(paste("Optimal Lambda", best.lamda))
```

```
$lambda.min
[1] 5.210327e-05

$lambda.1se
[1] 0.005990609

attr(,"class")
[1] "cv.glmnet"
```

Figure 3: LASSO Best Lambda

From our LASSO regression we obtain the optimal λ value as 0.00005 and a cross-validated Poisson deviance of 6.43%. The set of variables that LASSO picks are as follows:

```
#predict(best.model, s = best.lamda, type = "coefficients")
```

Based on this variable selection, the only variable that should not be considered in the model is the level of `Weather Condition` that is fine with high winds. This level also seems to have a high p-value of 0.98 in our summary output.

Conclusion

After fitting a generalized logistic regression model for the Poisson family, and performing model fitness as well as selection, we conclude that the significant predictos of Accident Severity are day of the week, type of raod, speed limit, weather condition, and urban/rural. Our results support our hypothesis that speed limit, weather condition and road type have a significant effect on Accident Severity and this can be seen from

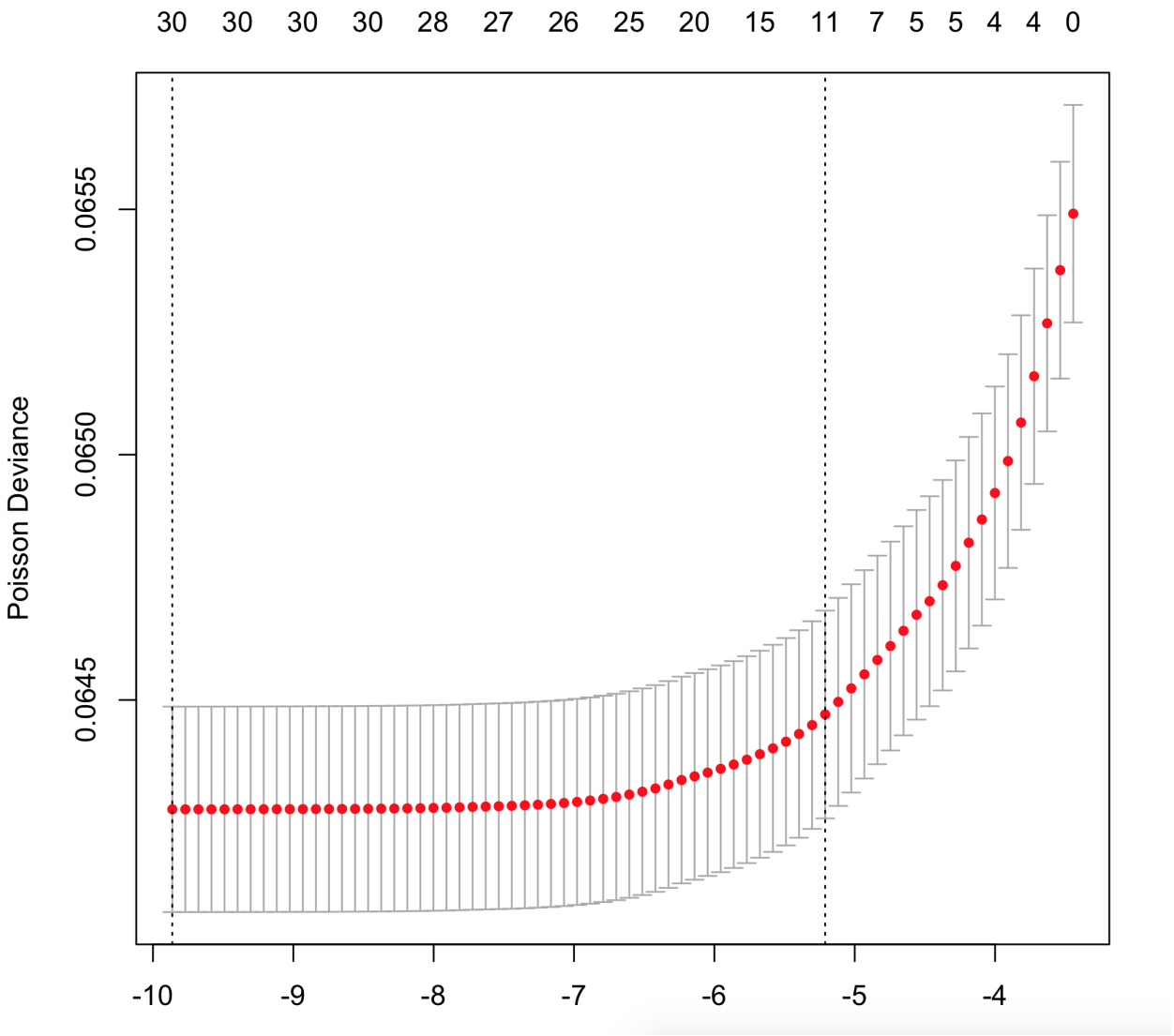


Figure 4: LASSO Best Lambda

```
33 x 1 sparse Matrix of class "dgCMatrix"
                                                                  1
(Intercept)                                          1.0326990309
(Intercept)                                          .
Number_of_Vehicles                                   0.0154163096
Day_of_Week2                                         0.0119149188
Day_of_Week3                                         0.0116231340
Day_of_Week4                                         0.0124959121
Day_of_Week5                                         0.0113156017
Day_of_Week6                                         0.0104181372
Day_of_Week7                                         0.0034518356
Road_TypeOne way street                              0.0056692875
Road_TypeRoundabout                                  0.0221856114
Road_TypeDual carriageway                            0.0183497284
Road_TypeSlip road                                   0.0346204422
Road_TypeUnknown                                     0.0186172922
Speed_limit                                         -0.0006737064
Junction_ControlGiveway or uncontrolled             -0.0029606495
Junction_Control                                    -0.0125073268
Junction_ControlAuthorised person                    0.0034673070
Junction_ControlStop Sign                            0.0060095037
Weather_ConditionsRaining without high winds         0.0104909012
Weather_ConditionsOther                              0.0092053010
Weather_ConditionsSnowing without high winds         0.0074555669
Weather_ConditionsRaining with high winds            0.0115018443
Weather_ConditionsFine with high winds               .
Weather_ConditionsUnknown                            0.0117033998
Weather_ConditionsFog or mist                        0.0018109008
Weather_ConditionsSnowing with high winds            0.0167837646
Road_Surface_ConditionsWet/Damp                     -0.0005954492
Road_Surface_ConditionsFrost/Ice                     0.0221444910
Road_Surface_ConditionsSnow                          0.0241624845
Road_Surface_ConditionsFlood (Over 3cm of water)     0.0195035977
Road_Surface_Conditions                              0.0070233123
Urban_or_Rural_Area2                                -0.0112620544
```

Figure 5: LASSO Variable Selection

the significance we obtained in our analysis. We noticed that accident severity seems to be increased by the accidents that happend on weekdays as compared to weekends, and this result does not seem to support our hypothesis. A potential reason for this could be that there are more cars on the road during weekdays while people commute to work and as a result of this there could be more severe accidents. We notice that the junction control is not a siginficant predictor of accident severity and this is interesting to note because one would assume that the type of junction, i.e, if there is a stop sign or not etc., would be an indicator and plausible cause for accidents. This finding does not support our initial hypothesis of junction control being a signficant predictor. We notice that road conditions of frost and ice have lead to an icnrease in Accident Severity mean count and this could be a useful indicator that during the winter season the roads are not well maintained and this could potentially be one of the reasons of so many road accidents. We also that when we shit from Rural to Urban, the mean count of Accident Severity increases. In other words, accidents in urban areas are more severe than accidents in rural areas, as we had predicted.

Therefoe, in conclusion, there are multiple factors that lead to road accidents and a few significant ones that we can act upon are speed limit, road condition, type of road and the difference between urban and rural areas. Traffic engineers could take into account different road conditions that lead to more severe accidents and work on reforms in order to mitigate this. Additionally, they could look at the road design in Urban and Rural areas in order to obtain a better insight as to why there is such a discrepancy in accident severity in the two. Further analysis using all the accident data that has been collected in the UK, and not only limited to these two years, can allow us to further investigate these identified causes and propose actionable preventative measures.