SDS2 Project : Riding Data Waves - A Bayesian Model for Bicycle Incident Analysis

Marina Zanoni

Contents

1 Abstract

This project aims to classify bicycle accident injuries using a multinomial Bayesian model, leveraging real-world data to conduct an in-depth Bayesian analysis with Markov Chain Monte Carlo (MCMC) simulations. The objective is to apply the techniques studied in our coursework to provide a comprehensive statistical model and analysis.

The dataset includes features that want to consider the relevance of the bikers's characteristics and of the external features. We develop a multinomial Bayesian model to predict the category of injury (BikeInjury) based on these features. The model's parameters are estimated through Bayesian point and interval estimation, and hypothesis testing is conducted to determine the significance of various predictors.

The project also evaluates the ability of Bayesian analysis to recover model parameters using simulated data.

In conclusion, this project demonstrates the effectiveness of a multinomial Bayesian model in classifying bicycle accident injuries, showcasing the application of Bayesian methods to complex real-world data and providing meaningful insights into the factors influencing bicycle accident outcomes.

2 Introduction

Cycling helps reduce traffic congestion and environmental pollution and promote a healthy lifestyle for the general public. However, it could also expose cyclists to dangerous environments, resulting in severe consequences and even death. Transport authorities are noting a rise in urban cycling accidents amidst an increasing cycling population, prompting the need for novel risk-informed cycling safety policies. This project aims to employ Bayesian analysis to assess and predict cycling accidents, considering their severity, focusing on the importance of variables identified in two studies Yang et al. (2021) , Nowakowska (2017). The goal is to analyze these variables and estimate the number of cycling injuries based on severity, highlighting their significant relevance as a primary focus of investigation.

3 Dataset

The dataset used in this study is from Kaggle, containing 57 variables, most of which are categorical discrete variables. The dataset exhibits class imbalance, where severe incidents are underrepresented compared to less severe ones, and unknown incidents have been excluded. Subsampling and overrepresentation techniques were employed to mitigate this issue. In the study conducted in Nowakowska (2017), particular attention was given to balancing the data based on severity levels.

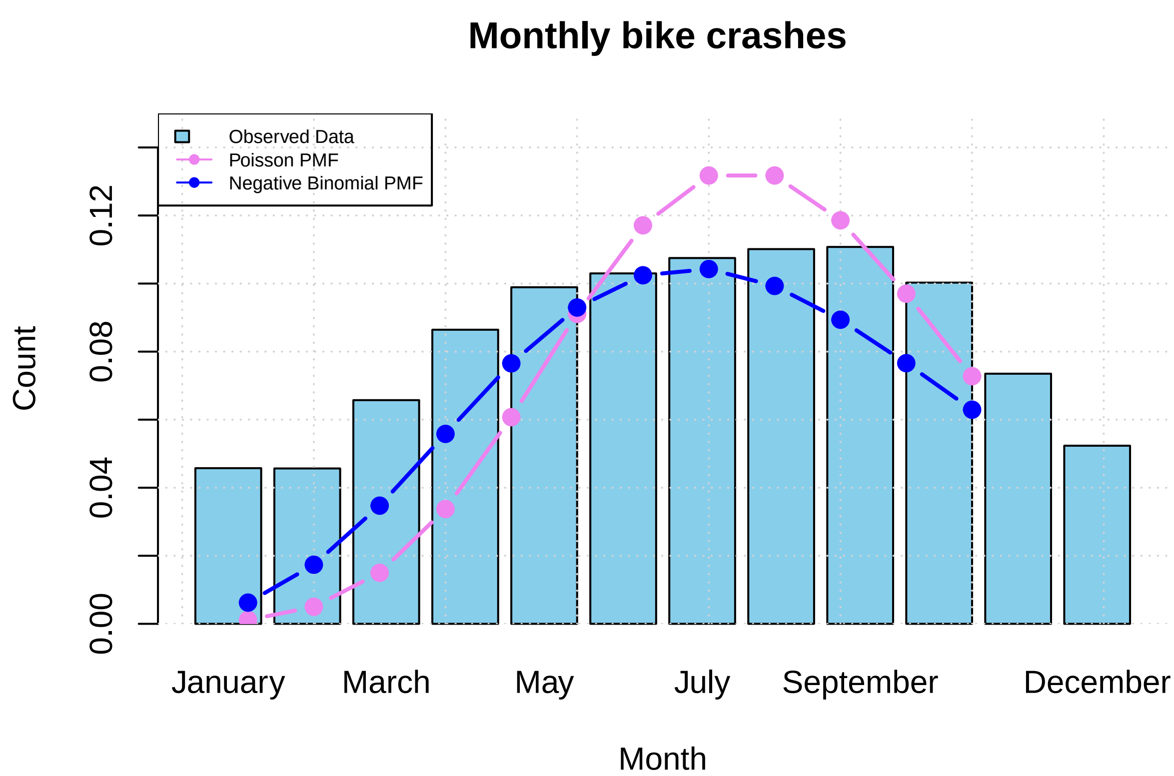
4 Exploratory data analysis

4.1 monthly trend

```r
suppressPackageStartupMessages(library(readr))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(ggplot2))

# 1st IDEA - MONTLY TREND ------------------------------------

# first filtering --------------------
suppressMessages(NCDOT_BikePedCrash <- read_csv("NCDOT_BikePedCrash.csv"))
```

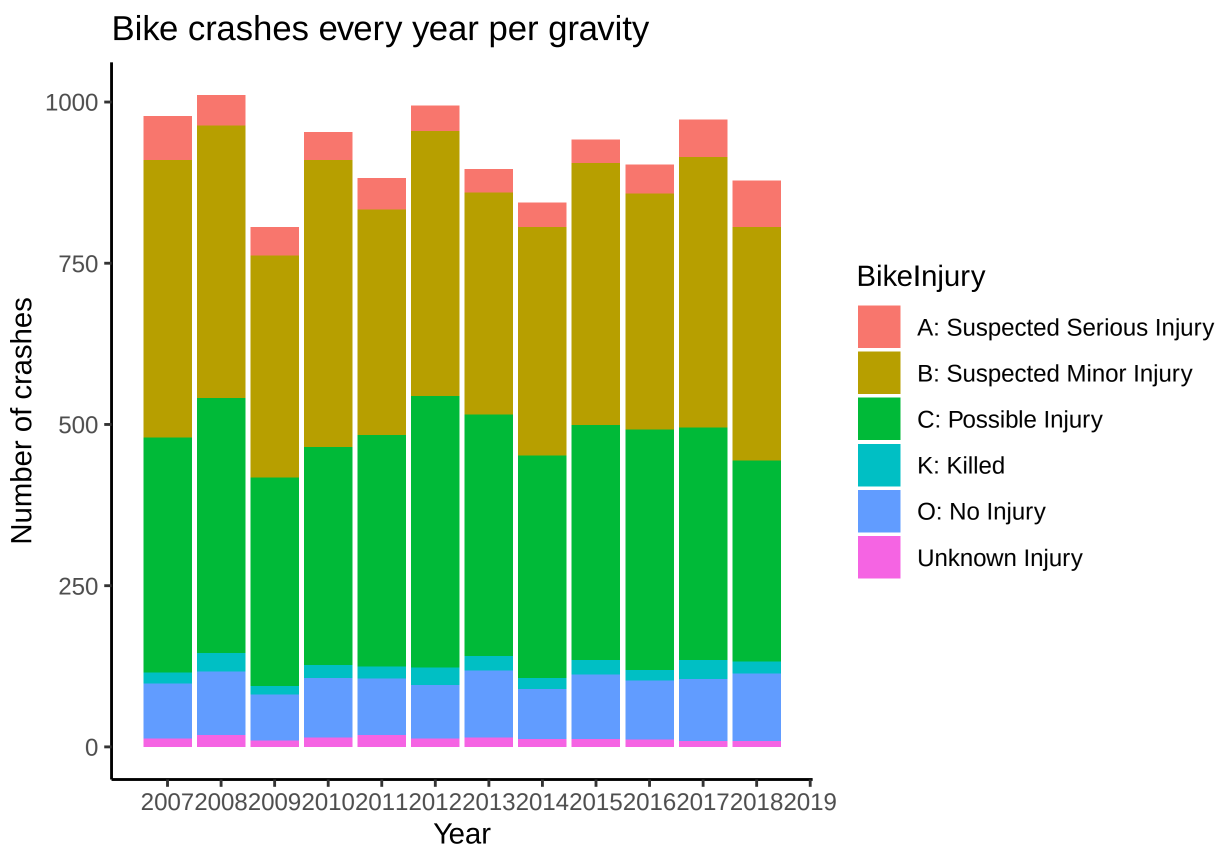
```r
Bike <- NCDOT_BikePedCrash
Bike <- Bike[, !names(Bike) %in% c("X","Y","OBJECTID")]
Bike_sub <- subset(Bike, BikeInjury != "Unknown Injury")
Bike_sub <- subset(Bike, BikeSex != "Unknown")
# Filtering and cleaning the data
clean_bike_df <- Bike_sub %>%
  select(AmbulanceR, CrashYear, CrashDay, CrashMonth, BikeSex,
         CrashHour, CrashAlcoh, CrashSevr, CrashGrp, DrvrInjury,
         BikeAlcFlg, DrvrAlcFlg, BikeAgeGrp, DrvrAgeGrp, BikeInjury,
         LightCond, RdConditio, RdClass, SpeedLimit, Weather,
         TraffCntrl, RdFeature, NumLanes)

# Prior hypothesis -------------
# Ensure the CrashMonth is a factor with the correct order
clean_bike_df$CrashMonth <- factor(clean_bike_df$CrashMonth, levels = month.name)

# Perform group by and count using dplyr
counts <- clean_bike_df %>%
  group_by(CrashMonth) %>%
  summarise(count = n())

# Plot total accident by month
barplot(counts$count/sum(counts$count), names.arg = counts$CrashMonth,
        xlab = 'Month', ylab = 'Count',
        main = 'Monthly bike crashes ',
        col = 'skyblue', border = 'black',
        ylim = c(0, 0.15))
grid()

# Define parameters for prior hypothesis
# Mean parameter of the Poisson distribution
lambda <- 9

# Generate x values for plotting the prior probabiities
x <- 1:12
# Overwrite the Poisson curve
points(x, dpois(x,lambda), type = "b", pch = 19, col = "violet")
lines(x, dpois(x,lambda), type = "b", lwd = 2, col = "violet")


# Plot Negbinomial PMF curve
points(x, dnbinom(x,12,mu=9), type = "b", pch = 19, col = "blue")
lines(x, dnbinom(x,12,mu=9), type = "b", lwd = 2, col = "blue")

# Legend
legend("topleft",
       legend = c("Observed Data", "Poisson PMF","Negative Binomial PMF"),
       fill = c("skyblue", NA, NA),
       border = c("black","white", "white"),
```

```
        pch = c(NA, 19, 19),
        lty = c(NA, 1, 1),
        col = c("skyblue", "violet","blue"),
        cex=0.6)
```



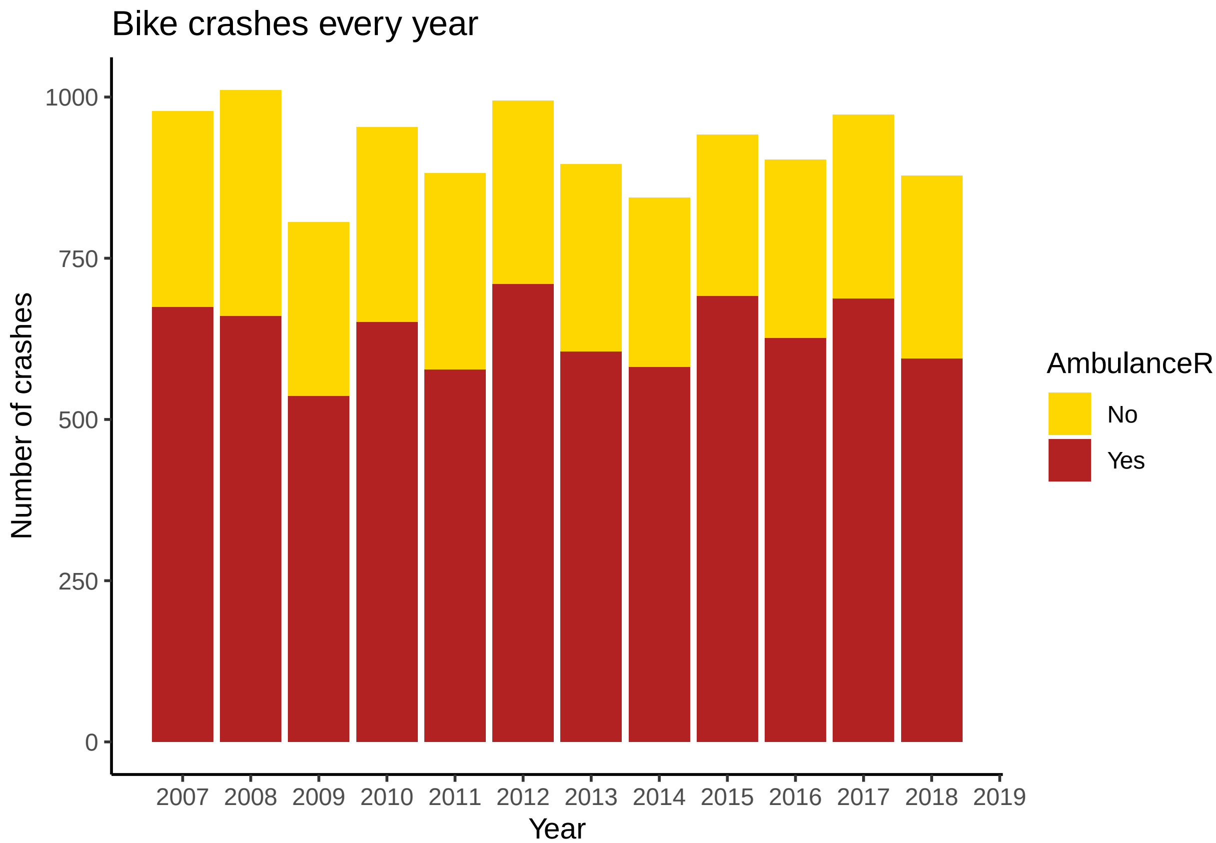
Analyzing whether this trend is consistent across all years is also interesting. The following plot illustrates this constancy.

```
#Yearly dinamics
clean_bike_df%>%
  group_by(CrashYear, BikeInjury)%>%
  count()%>%
  ggplot(aes(x = CrashYear, y = n, fill = BikeInjury)) +
  geom_col() +
  scale_x_continuous(breaks = c(2007:2019)) +
  labs(x = 'Year', y = 'Number of crashes',
       title = 'Bike crashes every year per gravity') +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



Even tough this could be an interesting problem, we rather want to take care of the severity of the accidents among them all. Indeed, as ambulance was necessary in a high number of cases the relevance of the matter is more evident.

```
# Relenvance of the matter
clean_bike_df%>%
  group_by(CrashYear, AmbulanceR)%>%
  count()%>%
  ggplot(aes(x = CrashYear, y = n, fill = AmbulanceR)) +
  geom_col() +
  scale_x_continuous(breaks = c(2007:2019)) +
  labs(x = 'Year',y = 'Number of crashes',
  title = 'Bike crashes every year') +
  scale_fill_manual(values=c("gold", "firebrick")) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black"))
```



5 Feature Selection Process

The **objective** of the experiment is to model the numbers of more or less severe accident involving bikers based on several explanatory variables using a Bayesian multinomial regression approach. The dataset contains factors related to cyclists, drivers, external variables, and road conditions. Explanatory variables include traffic category, crash day, weather conditions, cyclist age, cyclist gender, traffic control, speed limit, light conditions, and the number of lanes. The analysis aims to identify the influence of these variables on the severity of accidents, categorized into no injury, mild injury, and serious/fatal injury.

5.1 Definition of target variable

The classification of accident severity is crucial in understanding the impact and implications of road incidents. We combined the classes to align with the articles and defined a target variable that categorizes accidents into three main classes based on their severity:

- Fatal accidents occur when at least one person dies as a direct result of the crash, either immediately or within 30 days thereafter.
- Serious accidents do not result in fatalities but involve injuries such as life-long disabilities, severe mental or physical impairments, or long-term incapacity to work.
- Light accidents involve minor injuries where individuals suffer temporary harm or health disruptions lasting up to seven days based on medical diagnosis, without severe or fatal consequences.

To implement the first model, it was evident that some variables should be discarded initially to achieve a simple but effective model. Previous studies, particularly Yang et al. (2021), focused on a wide range of hazards influencing the occurrence probability and/or consequence severity of accidents. The variables identified in these studies were categorized into six groups:

1. Cyclist behaviour and personal characteristics
2. Environmental conditions
3. Road infrastructure issue
4. Interaction with other road users
5. Hazardous road conditions
6. Bike-related factors

5.1.1 Variables in the first article

DISTRICT (or analogous): This refers to the region/location where cycling accidents happen.

DAY: The statistics of cycling accidents reveals that the frequency of cycling accidents and their severity varies from days to day. For example, Sunday has the lowest number of accidents in total.

TIME: This variable is classified into two states based on: rush hour or not. According to BBC News and Wikivoyage, rush hour in the UK is typically 7am-10am and 4pm-7pm every weekday; for weekends, there is no specific definition. Following Yang et al. (2021)'s study, '11am-7pm' is set as the rush hour on weekends in Liverpool region because the traffic volume during this period is significantly higher than other time in weekends. Two states are set as rush hour and non-rush hour.

ENCOUNTERING VEHICLE TYPE: Encountering vehicle is the other party colliding with a cyclist on the road. Different encountering vehicle types often result in different accident consequences. For example, it is undoubtedly that colliding with a heavy goods vehicle (HGV) is much more dangerous than the collision with a motorbike for a cyclist if other conditions are kept the same (i.e., speed, environment). Based on the information provided by the accident reports, this variable has six states: Cars, HGV, Public Service Vehicle (PSV), motorcycle, cyclist, and other/unknown.}

WEATHER: This variable refers to weather conditions at the time and location of a cycling accident. As stated in Section 2, bad weather conditions have a major impact on cycling safety and failure to recognize its impact may cause huge loss, injuries and even casualties. The effect of a bad weather condition on cycling safety is mainly because of the reduction in visibility and distraction of cyclists, as well as its impact (e.g. rain) on a hazardous road condition (Joon-Ki Kim et al., 2007). Therefore, this variable needs to be paid much attention, especially in regions where bad weather often occurs. The Department for Transport in the UK defines several states for this variable: fine with high winds, fine without high winds, rain with high winds, rain without high winds, and others.

ROAD SURFACE CONDITION: The road surface condition in this study refers to the surface condition at the time and the place of the cycling accident. According to STATS19 reports, there are five types of a road surface condition in the UK: dry, wet/damp, snow, frost/ice, and flood (where surface water is over 3cm deep). However, in the Liverpool region, the major road surface conditions are dry and wet/damp, as stated by Merseyside Police. Meanwhile, the cycling accident database also tells there are very few accidents occurring on snow/frost/ice/flood road surface and none of them causes fatal consequence. Hence, this variable is classified into three states: dry, wet/damp, and others (i.e. snow, frost, ice and flood).

STREET LIGHTNING:Darkness is the most mentioned environmental hazard for cyclists, according to the literature. Previous researchers have already shown that cycling during late hours, especially at night, is more hazardous than daytime (Juhra et al., 2012). As a solution, the use of sufficient street lighting facilities can effectively tackle the darkness issue. However, in Liverpool, not all the places have the street lighting facilities, or some facilities are broken and not working well, generating an impact on the accident severity accordingly. Based on the STATS19 reports, 'street lighting' is categorized into three states in this study with respect to the darkness types: dark with no street light/unknown, dark with street lights present and lit, and daytime.

SPEED LIMIT: When driving on road, the driver must not drive faster than the speed limit for the type of road and type of vehicle. According to the Highway Code, road safety and vehicle rules of the UK, a speed limit of 30mph the most widely applied compared to other limits. (https://www.gov.uk/speed-limits). 'Speed Limit' is therefore classified into three states: 30mph, above 30mph andbelow 30mph.

ROAD TYPE: The road type associate with a cycling accident refers to the main carriageway on which the accident occurs. STATS19 lists six road types for cycling accidents: roundabout, one way street, dual carriageway, single carriageway, slip road, and unknown road. Nevertheless, among all the collected accident reports, very few occurred on one-way street, slip road and unknown road and none of them caused fatal consequence and hence these three road types are merged into one state - 'others'.

JUNCTION DETAIL: If there are two or more junctions within 20 meters of an accident, the junction that is the closest to the accident is recorded in the report. The UK government classifies junctions into nine categories. Some of them are not relevant in this paper since no relevant accident data are associated with them appropriately. Consequently, the processed states for 'junction detail' are crossroads, roundabout, not at junction, T or staggered junction and 'other' junction.

JUNCTION CONTROL: The existence of control measures at a junction is crucial for reducing the severity of accidents because they are effective in regularizing the behaviour of road users. Different control measures have different efficiency. This variable has four states: automatic traffic signal, give way, 'other' control measures and no control.

AGE OF CYCLIST:Based on the information provided by United Nations Educational, Scientific and Cultural Organisation (UNESCO) and the National Statistics Office of the UK, persons are divided into four groups according to their age bands within the cycling safety context: Child (Under 15), , Youth (15-24), Working adult (24-65) and the elderly (Over 65).

GENDER OF CYCLIST

Class	Variables
Class 1 (1st priority variables)	Age, District, Day, Encountering vessel type, First point of impact
Class 2 (2nd priority variables)	Combined Road Class, Junction control, Junction detail, Manoeuvre of Cyclist, V4, V8, O4, Road Type, V9, Speed limit, Weather, V5, V10, V12, V3, O16, O10, O9
Class 3 (low priority variables)	V1, V7, V11, O14, Cyclist location when accident happens, Street lighting, O5, O3, O11, Time, O8, Road Surface, V13, V2, V6, O7, O15, V14, Skidding, O13, O1, O12, V16, Sex, V15, O6, O2

Table 1: Variable classification

5.1.2 Variables selection in the second article

The resultant data set consists of 1307 records and the following variables:

- Bhv - at-fault driver's behaviour defined by the values:
  - FlCl -following too close (15.65%)

  - DrWrSdRd - driving wrong side of a roadway (3.8%)
  - InSpPrCn - inappropriate speed for the prevailing traffic and weather conditions (36.6%)
  - NGvWy - not giving right of way (23.7%)
  - InTrUTr - incorrect turning or U-turning (3.2%)
  - InOvBp - incorrect overtaking or bypassing (10.9%)
  - PrPsCn - poor psychophysical condition (6,1%)

- AgGrp - at-fault driver's age group: 02 - < 18; 25) (24,8%), 03 <25; 35) (28,4%), 04 - <35; 50) (24,1%), 05 - <50; 65) (17,1%), 06 - at least 65 (5,6%),
- Gndr - at-fault driver's gender: F - female (14,3%), M - male (85,7%),
- Alh - the influence of alcohol or other toxic substances on an at-fault driver: N - no (91,8%), Y - yes (8,2%),
- RdNr - road number: K42 (6,8%), K7 (15,5%), K73 (14,5%), K74 (31,6%), K77 (2,0%), K78 (6,1%), K79 (11,8%), K9 (11,7%)
- AcSvr - accident severity expressed by the status of a road crash according to the highest level of a human casualty harm as follows (Police, 2006; Nowakowska, 2010): LA - light accident (58,8%), SA
- serious accident (29,6%), FA - fatal accident (11,6%).

5.2 modeling and results

The research experiments were carried out in the SAS environment. The LOGISTIC procedure for MLE modelling and the MCMC procedure for Bayesian modelling were the most important. The author's SAS 4GL and macro programs were elaborated, among which the generators of the bootstrap samples and of the balanced training data set played a crucial role. The development of the models with all the input variables included was processed. In the Boot approach, there were the following types of the prior distributions obtained from 95-element sample for each investigated variable: normal for RdNr K73, RdNr K77, RdNr K79, lognormal for Bhv InSpPrCn, Bhv NGvWy, Bhv InTrUTr, Bhv InOvBp, AgGrp 02, AgGrp 04, RdNr K42, RdNr K78, and Weibull for Bhv FlCl, Bhv DrWrSdRd, AgGrp 03, AgGrp 05, Gndr F, Alh N, RdNr K7, RdNr K74.

5.3 data processing

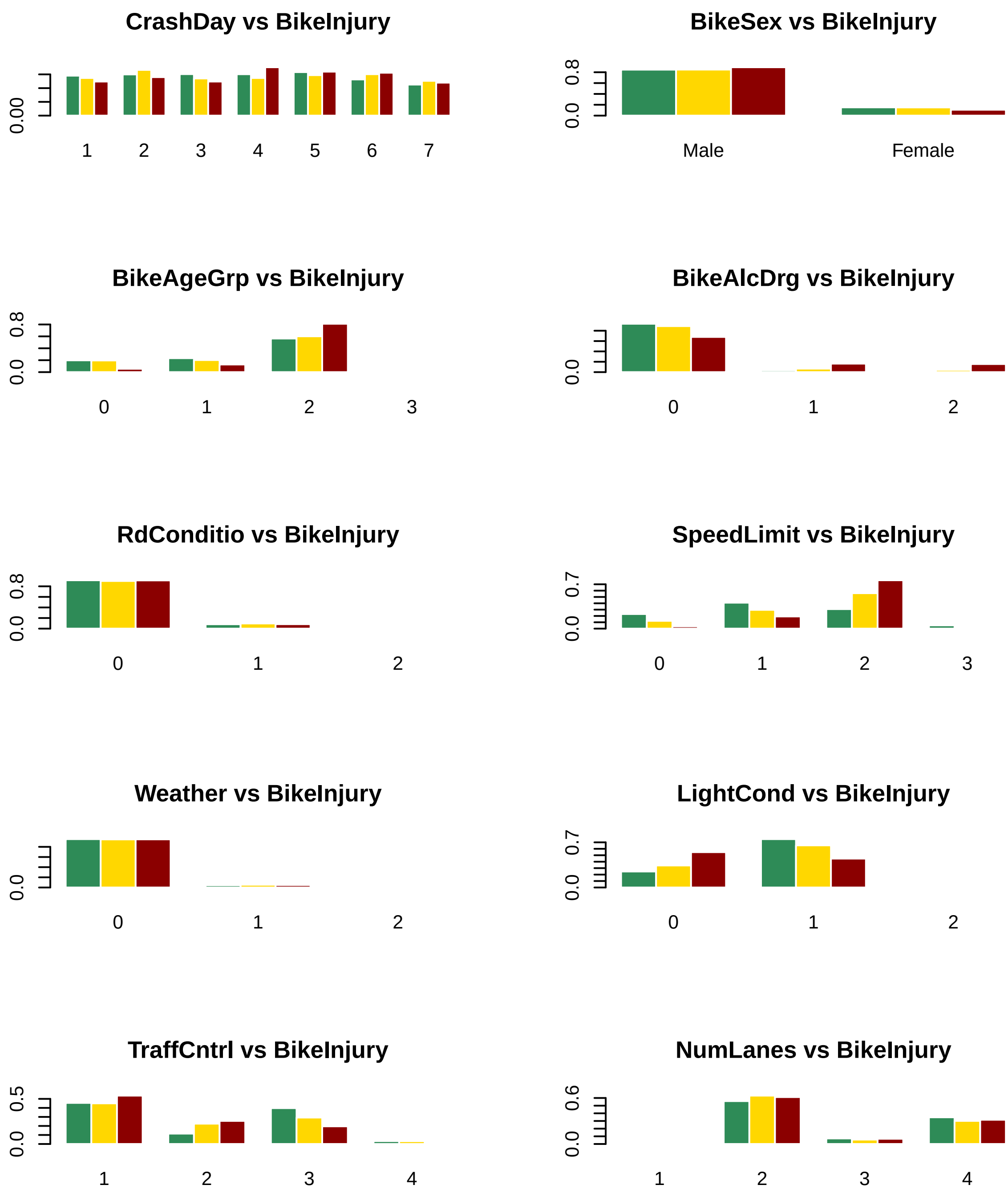
```
Bike_sub<-Bike

# Convert to factor the TARGET VARIABLES
# Eliminating UNKNOWN( "Unknown Injury") ------------------------------------
Bike_sub <- subset(Bike, BikeInjury != "Unknown Injury")
Bike_sub <- subset(Bike, BikeSex != "Unknown")

# Filtering and cleaning the data
clean_bike_df <- Bike_sub %>%
  select(CrashDay, Longitude, Latitude, CrashHour, BikeSex,
  BikeAgeGrp, BikeInjury, BikeAlcDrg, RdConditio,
  SpeedLimit, Weather,  LightCond,  TraffCntrl, NumLanes)
```

```r
# BikeInjury --------------------------
# Define ordered levels
ordered_levels <- c("O: No Injury", "B: Suspected Minor Injury",
                    "C: Possible Injury", "A: Suspected Serious Injury",
                    "K: Killed")

clean_bike_df$BikeInjury <- factor(clean_bike_df$BikeInjury,
                           levels = ordered_levels, ordered = TRUE)

# Map to numeric
clean_bike_df$BikeInjury <- as.numeric(clean_bike_df$BikeInjury) -1

clean_bike_df$BikeInjury<-ifelse(clean_bike_df$BikeInjury==0, 1,
                          ifelse(clean_bike_df$BikeInjury==2, 1,
                          ifelse(clean_bike_df$BikeInjury==3, 2,
                          ifelse(clean_bike_df$BikeInjury==4, 3, 1))))
clean_bike_df$BikeInjury <- as.factor(clean_bike_df$BikeInjury)


# NumLanes ----------------------------
clean_bike_df$NumLanes<-ifelse(clean_bike_df$NumLanes=="1 lane", 1,
                        ifelse(clean_bike_df$NumLanes=="2 lanes", 2,
                        ifelse(clean_bike_df$NumLanes=="3 lanes", 3,
                        ifelse(clean_bike_df$NumLanes=="4 lanes", 4, 4))))

clean_bike_df$NumLanes <- as.factor(clean_bike_df$NumLanes)

# TraffCntrl --------------------------

clean_bike_df$TraffCntrl<-ifelse(clean_bike_df$TraffCntrl=="No Control Present", 1,
              ifelse(clean_bike_df$TraffCntrl=="Double Yellow Line, No Passing Zone", 2,
              ifelse(clean_bike_df$TraffCntrl=="Stop And Go Signal", 3,
              ifelse(clean_bike_df$TraffCntrl=="Stop Sign", 3,
              ifelse(clean_bike_df$TraffCntrl=="Stop Sign", 3, 4)))))
clean_bike_df$TraffCntrl <- as.factor(clean_bike_df$TraffCntrl)

# SpeedLimit -----------------------

clean_bike_df$SpeedLimit<-ifelse(clean_bike_df$SpeedLimit=="5 - 15 MPH", 0,
                         ifelse(clean_bike_df$SpeedLimit=="20 - 25  MPH", 0,
                         ifelse(clean_bike_df$SpeedLimit=="30 - 35  MPH", 1,
                         ifelse(clean_bike_df$SpeedLimit=="Unknown", 3, 2))))

clean_bike_df$SpeedLimit<-as.factor(clean_bike_df$SpeedLimit)

# BikeSex --------------------------------------
clean_bike_df$BikeSex <- factor(clean_bike_df$BikeSex,
       levels = c("Male","Female","Unknown"), ordered = FALSE)
```

```r
clean_bike_df$BikeSex <- factor(clean_bike_df$BikeSex)


# CrashDay ---------------------------------------
weekdays.name<-c("Monday", "Tuesday", "Wednesday", "Thursday",
                 "Friday", "Saturday", "Sunday")
clean_bike_df$CrashDay <- match(clean_bike_df$CrashDay, weekdays.name)

clean_bike_df$CrashDay <- as.factor(clean_bike_df$CrashDay)

# Weather ---------------------------------

clean_bike_df$Weather<-ifelse(clean_bike_df$Weather=="Clear", 0,
            ifelse(clean_bike_df$Weather=="Cloudy", 0,
            ifelse(clean_bike_df$Weather=="Fog, Smog, Smoke", 1,
            ifelse(clean_bike_df$Weather=="Rain", 1,
            ifelse(clean_bike_df$Weather=="Snow, Sleet, Hail,
                   Freezing Rain/Drizzle", 1, 2)))))

clean_bike_df$Weather <- as.factor(clean_bike_df$Weather)

# RdConditio -----------------------------------------------
# 0-1 wet or not wet

clean_bike_df$RdConditio<-ifelse(clean_bike_df$RdConditio == "Dry", 0,
                          ifelse(clean_bike_df$RdConditio == "Unknown", 2, 1))

clean_bike_df$RdConditio <- as.factor(clean_bike_df$RdConditio)

# BikeAgeGrp ----------------------------
clean_bike_df$BikeAgeGrp <- ifelse(clean_bike_df$BikeAgeGrp == "0-5", 0,
                            ifelse(clean_bike_df$BikeAgeGrp == "6-10", 0,
                            ifelse(clean_bike_df$BikeAgeGrp == "11-15",0,
                            ifelse (clean_bike_df$BikeAgeGrp == "16-19",1,
                            ifelse (clean_bike_df$BikeAgeGrp == "20-24",1,
                            ifelse (clean_bike_df$BikeAgeGrp == "Unknown",3,2))))))

clean_bike_df$BikeAgeGrp <- as.factor(clean_bike_df$BikeAgeGrp)

# BikeAlc--------------------------

clean_bike_df$BikeAlcDrg <- as.factor(ifelse(clean_bike_df$BikeAlcDrg == ".", 0,
                 ifelse(clean_bike_df$BikeAlcDrg == "No", 0,
                 ifelse(clean_bike_df$BikeAlcDrg == "Unknown",2,
                 ifelse (clean_bike_df$BikeAlcDrg == "Missing",2,
                 ifelse (clean_bike_df$BikeAlcDrg == ".",2,
                 ifelse (clean_bike_df$BikeAlcDrg == "Unknown", 2, 1)))))))

# LightCond ---------------------------
```

```r
clean_bike_df$LightCond <- as.factor(ifelse(clean_bike_df$LightCond == "Daylight", 1,
                           ifelse(clean_bike_df$LightCond == "Other", 2,
                           ifelse(clean_bike_df$LightCond == "Unknown", 2, 0))))


# Time to Traffic Hour ---------------------
#traffic_hours <- c(7:10, 16:19)
#clean_bike_df$TrafficCategory <- ifelse(clean_bike_df$CrashHour %in% traffic_hours, 1, 0)
```

```r
colors <- c("seagreen", "gold", "darkred")

# Function to create a bar plot for a single variable against the
#target variable 'BikeInjury'
create_plot <- function(data, variable, target, colors) {
  barplot_prop <- function(data, var, target, colors) {
    # Create a table of proportions
    tab <- prop.table(table(data[[var]], data[[target]]), margin = 2)

    # Create the bar plot
    barplot(t(tab), beside = TRUE, col = colors,  main = paste(var, "vs", target),
            ylab = "", xlab = "", border = "white")
  }
  barplot_prop(data, variable, target, colors)
}

par(mfrow = c(3, 2))
for (variable in names(clean_bike_df[,-c(2:4,7)])) {
  create_plot(clean_bike_df, variable, "BikeInjury", colors)
}
```


CrashDay vs BikeInjury
0.00
1 2 3 4 5 6 7
BikeSex vs BikeInjury
0.0 0.8
Male
Female
BikeAgeGrp vs BikeInjury
0.0 0.8
0 1 2 3
BikeAlcDrg vs BikeInjury
0.0
0 1 2
RdConditio vs BikeInjury
0.0 0.8
0 1 2
SpeedLimit vs BikeInjury
0.0 0.7
0 1 2 3
Weather vs BikeInjury
0.0
0 1 2
LightCond vs BikeInjury
0.0 0.7
0 1 2
TraffCntrl vs BikeInjury
0.0 0.5
1 2 3 4
NumLanes vs BikeInjury
0.0 0.6
1 2 3 4

6 Model Formulation

6.1 Multinomial logistic regression

Multinomial logistic regression models are natural extensions of the binomial logistic regression models and are used when the response of interest are multicategorical variables. The Bayesian multinomial regression model aims to predict the severity of cycling accidents (possible/minor, serious, or fatal) based on explanatory variables.

Assuming that the response variable $\mathbf{Y_i} = (\mathbf{X_i}, \ldots, \mathbf{X_K})$ has K levels, where Y_{ik} denotes the frequency of the kth level, the multinomial logistic regression model can be written as:

$$\mathbf{Y_i} \sim \texttt{multinomial}(\pi_\mathbf{i}, \mathbf{N_i})$$

and

$$log\frac{\pi_{ik}}{\pi_{1i}} = \eta_{ik} = \beta_{0k} + \sum_{i=1}^{k} \beta_{jk}\gamma_{jk}x_{ij}$$

for $k = 2, \ldots, K$ where $\pi_{i1}, \ldots, \pi_{ik}^T$ is the vector of the probabilities for each level of variable $\mathbf{Y}$ for individual i with $\pi_i = 1 - \sum_{i=1}^{k} \pi_{ik}$ and γ_{ik} are the usual binary indicators identifying the structure of the model and which variables specify or affect each odds $\pi_{ik}/\pi i1$. It is common practice to use similar structure for all odds; see, for example, in Agresti (2002, chap. 7) for a comprehensive treatment of the subject. Solving is terms of response probabilities results in:

$$\pi_{i1} = \frac{1}{1 + \sum_{k=2}^{K} e^{\eta_{ik}}}$$

$$\pi_{ik} = \frac{e^{\eta_{ik}}}{1 + \sum_{k=2}^{K} e^{\eta_{ik}}} \quad \text{for} \quad k = 2, \ldots, K$$

This can be summarized by:

$$\pi_{ik} = \frac{e^{\eta_{ik}}}{\sum_{k=1}^{K} e^{\eta_{ik}}} \quad \text{con} \quad \eta_{i1} = 0 \quad \text{per} \quad i = 1, 2, \ldots, n$$

6.2 First model

This can be implemented in JAGS using the commands:

```
model {
  # Likelihood
  for (i in 1:N) {
    x[i] ~ dcat(pi[i, 1:3])

    # Definition of categorical variables through a multinomial model
    pi[i, 1] <- 1 / sum(exp(eta2[i]), exp(eta3[i]), 1)
    pi[i, 2] <- exp(eta2[i]) / sum(exp(eta2[i]), exp(eta3[i]), 1)
    pi[i, 3] <- exp(eta3[i]) / sum(exp(eta2[i]), exp(eta3[i]), 1)

    # Definition of linear components for eta1 and eta2
```

```
    eta2[i] <- beta0 + beta1 * CrashDay1[i] + beta2 * CrashDay2[i] +
    beta3 * CrashDay3[i] + beta4 * CrashDay4[i] + beta5 * CrashDay5[i] +
    beta6 * CrashDay6[i] + beta7* CrashDay7[i] + beta8 * CrashHour[i] +
    beta9 * BikeSexFemale[i] + beta10 * BikeAgeGrp1[i] + beta11 * BikeAgeGrp2[i] +
    beta12 * BikeAgeGrp3[i] + beta13 * BikeAlcDrg1[i] + beta14 * BikeAlcDrg2[i] +
    beta15 * RdConditio1[i] + beta16 * RdConditio2[i] + beta17 * SpeedLimit1[i] +
    beta18 * SpeedLimit2[i] + beta19 * SpeedLimit3[i] + beta20 * Weather1[i] +
    beta21 * Weather2[i] + beta22 * LightCond1[i] + beta23 * LightCond2[i] +
    beta24 * TraffCntrl2[i] + beta25 * TraffCntrl3[i] + beta26 * TraffCntrl4[i] +
    beta27 * NumLanes2[i] + beta28 * NumLanes3[i] + beta29 * NumLanes4[i]

    eta3[i] <-gamma0 + gamma1 * CrashDay1[i] + gamma2 * CrashDay2[i] +
    gamma3 * CrashDay3[i] + gamma4 * CrashDay4[i] + gamma5 * CrashDay5[i] +
    gamma6 * CrashDay6[i] + gamma7* CrashDay7[i] + gamma8 * CrashHour[i] +
    gamma9 * BikeSexFemale[i] + gamma10 * BikeAgeGrp1[i] + gamma11 * BikeAgeGrp2[i] +
    gamma12 * BikeAgeGrp3[i] + gamma13 * BikeAlcDrg1[i] + gamma14 * BikeAlcDrg2[i] +
    gamma15 * RdConditio1[i] + gamma16 * RdConditio2[i] + gamma17 * SpeedLimit1[i] +
    gamma18 * SpeedLimit2[i] + gamma19 * SpeedLimit3[i] + gamma20 * Weather1[i] +
    gamma21 * Weather2[i] + gamma22 * LightCond1[i] + gamma23 * LightCond2[i] +
    gamma24 * TraffCntrl2[i] + gamma25 * TraffCntrl3[i] + gamma26 * TraffCntrl4[i] +
    gamma27 * NumLanes2[i] + gamma28 * NumLanes3[i] + gamma29 * NumLanes4[i]
  }

  # Priors for beta parameters
  # Priors parameters for beta
  beta0 ~ dnorm(0, 0.001)
  beta1 ~ dnorm(0, 0.001)
  beta2 ~ dnorm(0, 0.001)
  .
  .
  .

  # Priors parameters for gamma
  gamma0 ~ dnorm(1, 0.001)
  gamma1 ~ dnorm(1, 0.001)
  gamma2 ~ dnorm(1, 0.001)
  .
  .
  .
}
```

Indeed it is advisable to select as baseline the category with the highest number of observations and then transform estimates as desired.

```r
# Subsampling ---------------------------------------
# PROPORTIONAL SAMPLING S
# Numero totale di osservazioni desiderate
n_total <- 400
```

```r
# BALANCED_WAY
# Proportions of the classes in injury already grouped in clean_bike_df
prop_class <- table(clean_bike_df$BikeInjury) / nrow(Bike)

# Calcola il numero di osservazioni per ciascuna classe
n_class <- round(n_total * prop_class)

balanced_sample <- lapply(names(n_class), function(class) {
  # Selecting the indexes of the class
  class_rows <- which(clean_bike_df$BikeInjury == class)
  # Sampling from the indexes
  sampled_indices <- sample(class_rows, n_class[class])
  # returning the datasets for each class, stored in a list
  return(clean_bike_df[sampled_indices, ])
})

# Combine in a single dataset
balanced_dataset <- do.call(rbind, balanced_sample)
dim(balanced_dataset)

# DUMMYVARIABLES------------------------------
categorical_vars <- c("CrashDay", "CrashHour","BikeSex", "BikeAgeGrp", "BikeInjury",
                      "BikeAlcDrg", "RdConditio", "SpeedLimit", "Weather",
                      "LightCond", "TraffCntrl", "NumLanes")
#clean_bike_df$BikeInjury <- as.factor(clean_bike_df$BikeInjury)
is.factor(balanced_dataset$BikeInjury)

# get the dummy variables
dummy_var <- dummyVars(BikeInjury ~ .,
                       data = balanced_dataset[, categorical_vars])
head(dummy_var)
dummy_data <- as.data.frame(model.matrix(~ . - 1,
                                         data = balanced_dataset[categorical_vars]))
colnames(dummy_data)
dummy_data<-dummy_data[,-c(13:14)]    #Removing the Injury variables
colnames(dummy_data)
dim(dummy_data)


# Bayesian MULTILOGISTIC REGRESSION model ------------------------

dd5 <- list(
  "x" = balanced_dataset$BikeInjury,
  "CrashDay1" = dummy_data$CrashDay1,
  "CrashDay2" = dummy_data$CrashDay2,
  "CrashDay3" = dummy_data$CrashDay3,
  "CrashDay4" = dummy_data$CrashDay4,
  "CrashDay5" = dummy_data$CrashDay5,
  "CrashDay6" = dummy_data$CrashDay6,
```

```r
  "CrashDay7" = dummy_data$CrashDay7,
  "CrashHour" = dummy_data$CrashHour,
  "BikeSexFemale" = dummy_data$BikeSexFemale,
  "BikeAgeGrp1" = dummy_data$BikeAgeGrp1,
  "BikeAgeGrp2" = dummy_data$BikeAgeGrp2,
  "BikeAgeGrp3" = dummy_data$BikeAgeGrp3,
  "BikeAlcDrg1" = dummy_data$BikeAlcDrg1,
  "BikeAlcDrg2" = dummy_data$BikeAlcDrg2,
  "RdConditio1" = dummy_data$RdConditio1,
  "RdConditio2" = dummy_data$RdConditio2,
  "SpeedLimit1" = dummy_data$SpeedLimit1,
  "SpeedLimit2" = dummy_data$SpeedLimit2,
  "SpeedLimit3" = dummy_data$SpeedLimit3,
  "Weather1" = dummy_data$Weather1,
  "Weather2" = dummy_data$Weather2,
  "LightCond1" = dummy_data$LightCond1,
  "LightCond2" = dummy_data$LightCond2,
  "TraffCntrl2" = dummy_data$TraffCntrl2,
  "TraffCntrl3" = dummy_data$TraffCntrl3,
  "TraffCntrl4" = dummy_data$TraffCntrl4,
  "NumLanes2" = dummy_data$NumLanes2,
  "NumLanes3" = dummy_data$NumLanes3,
  "NumLanes4" = dummy_data$NumLanes4,
  "N" = nrow(dummy_data)
)

# Parameters to track
params5 <- c("beta0", "beta1", "beta2", "beta3", "beta4", "beta5",
             "beta6", "beta7", "beta8", "beta9", "beta10", "beta11",
             "beta12", "beta13", "beta14", "beta15", "beta16", "beta17",
             "beta18", "beta19", "beta20", "beta21", "beta22", "beta23",
             "beta24", "beta25", "beta26", "beta27", "beta28", "beta29",
             "gamma0", "gamma1", "gamma2", "gamma3", "gamma4", "gamma5",
             "gamma6", "gamma7", "gamma8", "gamma9", "gamma10", "gamma11",
             "gamma12", "gamma13", "gamma14", "gamma15", "gamma16", "gamma17",
             "gamma18", "gamma19", "gamma20", "gamma21", "gamma22", "gamma23",
             "gamma24", "gamma25", "gamma26", "gamma27", "gamma28", "gamma29")

# model
model5 <- jags(data = dd5,
               parameters.to.save = params5,
               model.file = "model5_dummy.txt",
               n.chains = 2,
               n.iter = 10000,
               n.burnin = 2000,
               n.thin = 1)
```

6.3 Diagnostics

```
model5
```

|model5

beta7	4.301	11.062	-20.557	-3.668	5.310	13.344	22.083	1.478	6
beta8	0.067	0.071	-0.062	0.018	0.064	0.113	0.217	1.025	110
beta9	0.330	0.702	-1.114	-0.121	0.352	0.815	1.617	1.001	3500
gamma0	-24.501	11.990	-48.688	-32.666	-24.031	-16.787	-1.455	1.608	5
gamma1	4.293	10.196	-13.933	-3.459	4.629	11.444	24.431	1.402	7
gamma10	10.933	7.682	-0.237	5.004	9.776	15.407	29.257	1.072	41
gamma11	12.719	7.552	2.094	6.750	11.560	17.003	30.809	1.076	39
gamma12	-15.007	22.115	-64.684	-28.966	-11.907	1.717	19.662	1.001	16000
gamma13	1.318	2.082	-3.032	0.011	1.402	2.699	5.224	1.001	16000
gamma14	4.265	1.618	1.175	3.199	4.219	5.297	7.599	1.001	4200
gamma15	-27.783	19.915	-72.982	-40.155	-24.937	-12.047	0.578	1.001	8900
gamma16	-10.360	25.443	-64.612	-26.539	-8.617	7.093	36.084	1.001	16000
gamma17	-0.158	1.807	-3.645	-1.312	-0.216	0.941	3.596	1.001	11000
gamma18	3.132	1.687	0.253	1.965	3.008	4.105	6.826	1.003	940
gamma19	-26.098	19.930	-71.955	-38.345	-23.104	-10.463	2.175	1.001	8300
gamma2	6.985	10.094	-11.173	-0.638	7.436	13.734	27.243	1.398	7
gamma20	-13.356	25.241	-68.191	-28.990	-11.466	3.560	32.594	1.001	16000
gamma21	1.286	31.624	-59.908	-20.035	1.667	22.506	63.888	1.001	16000
gamma22	0.034	1.094	-2.080	-0.694	0.013	0.747	2.194	1.004	480
gamma23	-3.513	29.054	-62.838	-22.668	-2.358	16.631	49.981	1.001	11000
gamma24	-2.248	1.729	-6.035	-3.284	-2.100	-1.064	0.735	1.001	7900
gamma25	0.327	1.144	-1.944	-0.414	0.329	1.077	2.554	1.002	1900
gamma26	-24.184	19.614	-71.821	-35.310	-20.331	-8.788	1.180	1.002	1700
gamma27	-2.650	2.309	-7.656	-4.030	-2.579	-1.152	1.679	1.017	120
gamma28	-27.253	18.439	-70.994	-37.632	-23.338	-12.983	-3.255	1.001	12000
gamma29	-3.895	2.357	-8.960	-5.312	-3.796	-2.333	0.507	1.012	170
gamma3	7.602	10.074	-10.649	-0.092	7.987	14.378	28.193	1.405	7
gamma4	5.836	10.113	-12.482	-1.872	6.274	12.653	26.199	1.392	7
gamma5	-22.669	21.201	-71.289	-35.175	-19.815	-7.594	11.735	1.033	55
gamma6	6.457	10.119	-11.694	-1.310	6.868	13.443	26.788	1.407	7
gamma7	-22.257	21.148	-70.608	-34.781	-19.430	-7.080	11.683	1.031	55
gamma8	0.170	0.111	-0.030	0.092	0.165	0.240	0.400	1.020	110

```
For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 44.9 and DIC = 260.1
DIC is an estimate of expected predictive error (lower deviance is better).
```

$$\eta_{2i} = -32.908 + \cdots + \underset{(11.179)}{13.866} \cdot \texttt{Weather1}_i + \underset{(-1.288)}{31.994} \cdot \texttt{Weather2}_i$$
$$-\underset{(0.788)}{0.503} \cdot \texttt{BikeAgeGrp1}_i - \underset{0.699}{0.788} \cdot \texttt{BikeAgeGrp2}_i - \underset{(18.582)}{23.591} \cdot \texttt{BikeAgeGrp3}_i$$
$$+ \underset{(-0.483)}{0.700} \cdot \texttt{BikeSexFemale}_i - \underset{(19.900)}{27.990} \cdot \texttt{RdConditio1}_i - \underset{(25.005)}{9.942} \cdot \texttt{RdConditio2}_i$$
$$+ \underset{(8.605)}{5.403} \cdot \texttt{NumLanes2}_i + \underset{(8.329)}{5.582} \cdot \texttt{NumLanes3}_i$$

6.4 Coefficients

I selected some coefficients that represent characteristics of the bikers, such as age and sex, as well as some external conditions like the number of lanes and road conditions, which are known to have an impact on the phenomenon. Moreover, these coefficients serve as a preliminary step into a legitimate question: should the government invest in campaigns to raise awareness about bike accidents or rather invest in improving street quality?

The β and γ coefficients represent the effect of each predictor variable on the log-odds of being in a certain category compared to the baseline category (minor injury). For example, β_j represents the effect of the j-th predictor on the log-odds of being in category k relative to the baseline category.

Odds values can range from 0 to infinity and indicate how much more likely it is that an observation is a member of the target group rather than a member of the other group.

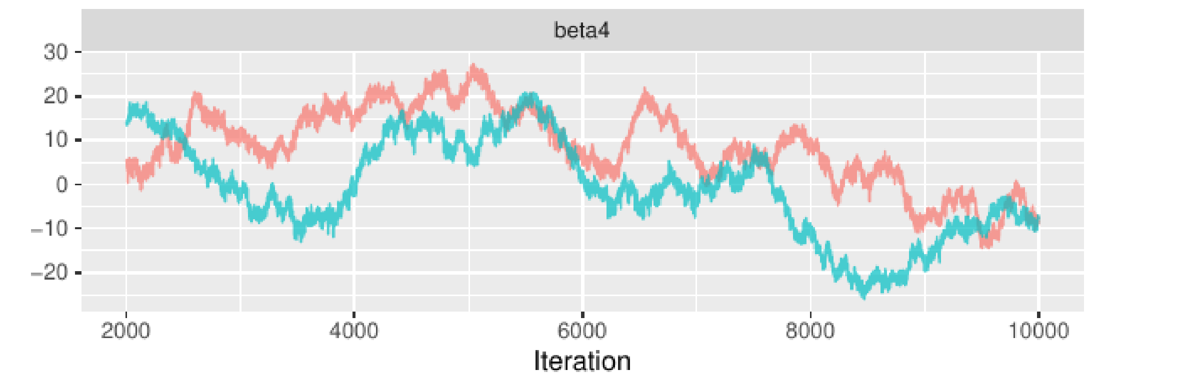
- ($\beta_{10/11/12}$) **and** ($\gamma_{10/11/12}$)**: BikeAgeGrp**: These coefficients have negative values, which is consistent with what was found in Nowakowska (2017). Similarly, for them, the coefficients do not seem significant for either the second or third class. Logically, it is also coherent to expect that as we grow older, we drive more safely, both in attitude and in the type of routes we take.
- (β_9) **and** (γ_9)**: BikeSex**: Here, we expected a negative value as we know that women are involved in less dangerous accidents.
- ($\beta_{15/16}$) **and** ($\gamma_{15/16}$)**: RdCondition**: According to Yang et al. (2021), this coefficient is positive but has a very low impact on the overall analysis. We will discuss this variable further when looking at the convergence of the plots.
- ($\beta_{27/28}$) **and** ($\gamma_{27/28}$)**: NumLanes**: Having a positive value indicates that a greater number of lanes is associated with an increase in the probability of belonging to category 3, thus experiencing a worse accident compared to the reference category, which is also what is expected in the literature.

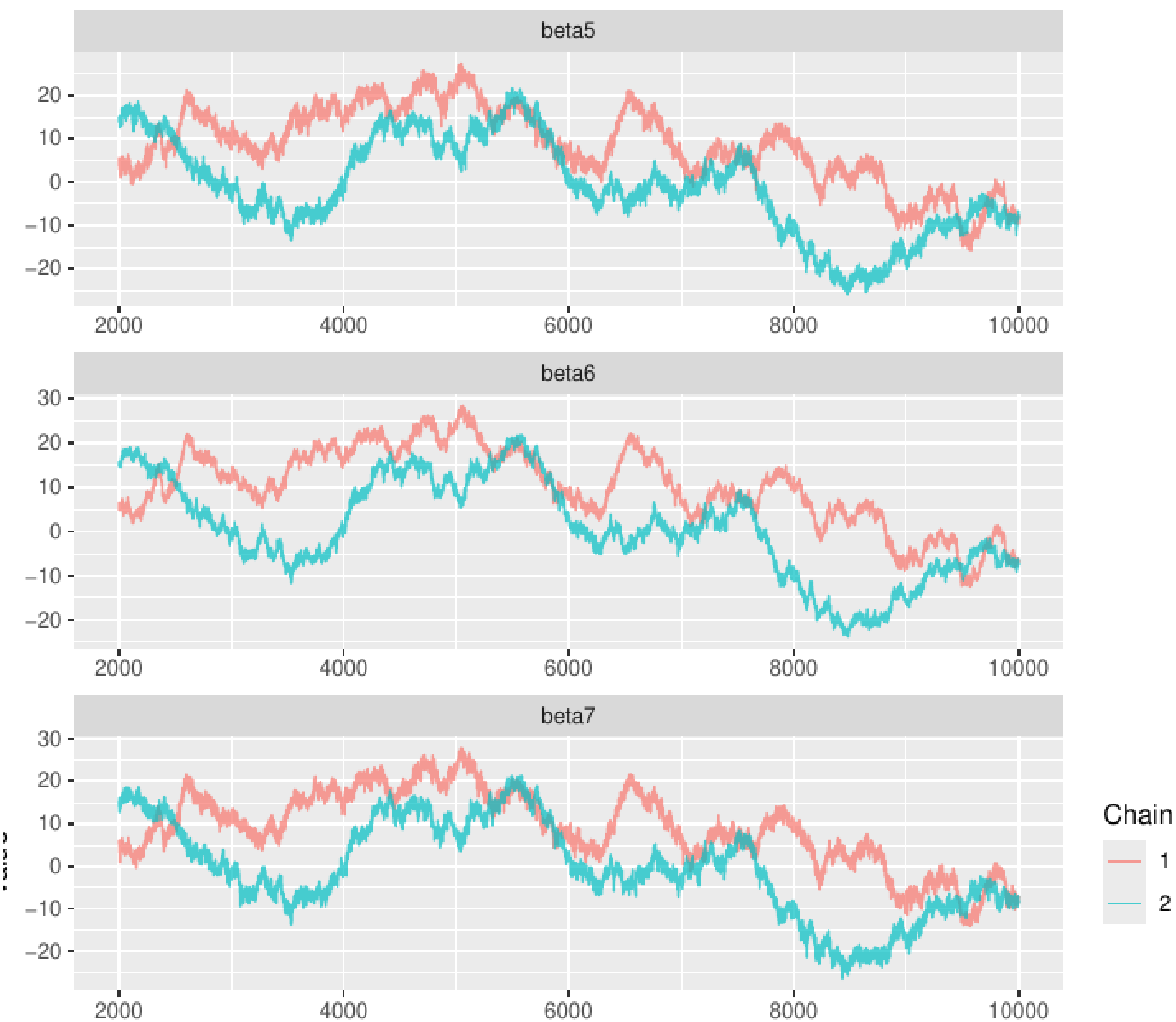
6.5 Second model

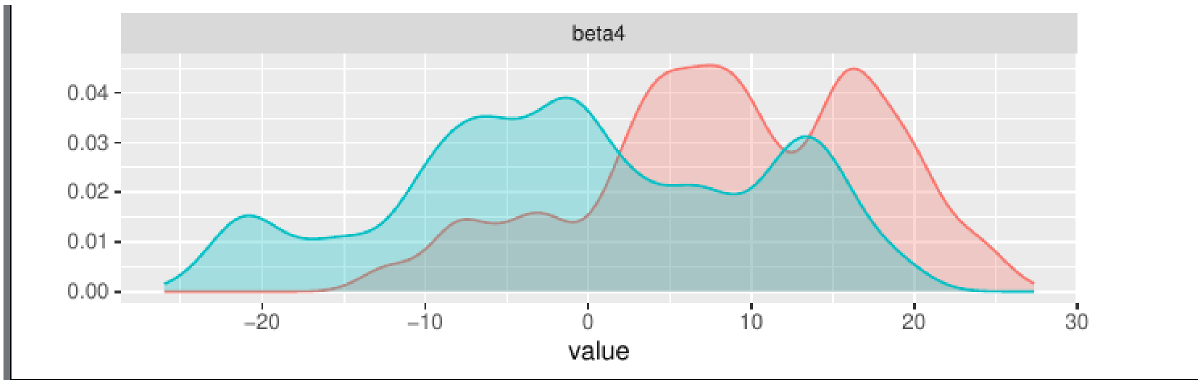
The second model was developed after examining the convergence diagnostics of the first model. Specifically, the coefficients associated with the day of the week and the number of lanes did not exhibit desirable properties and were therefore excluded. The diagnostics considered included the Rhat values, effective sample size (n.eff), trace plots, and density plots, assessed using the ggmcmc library.

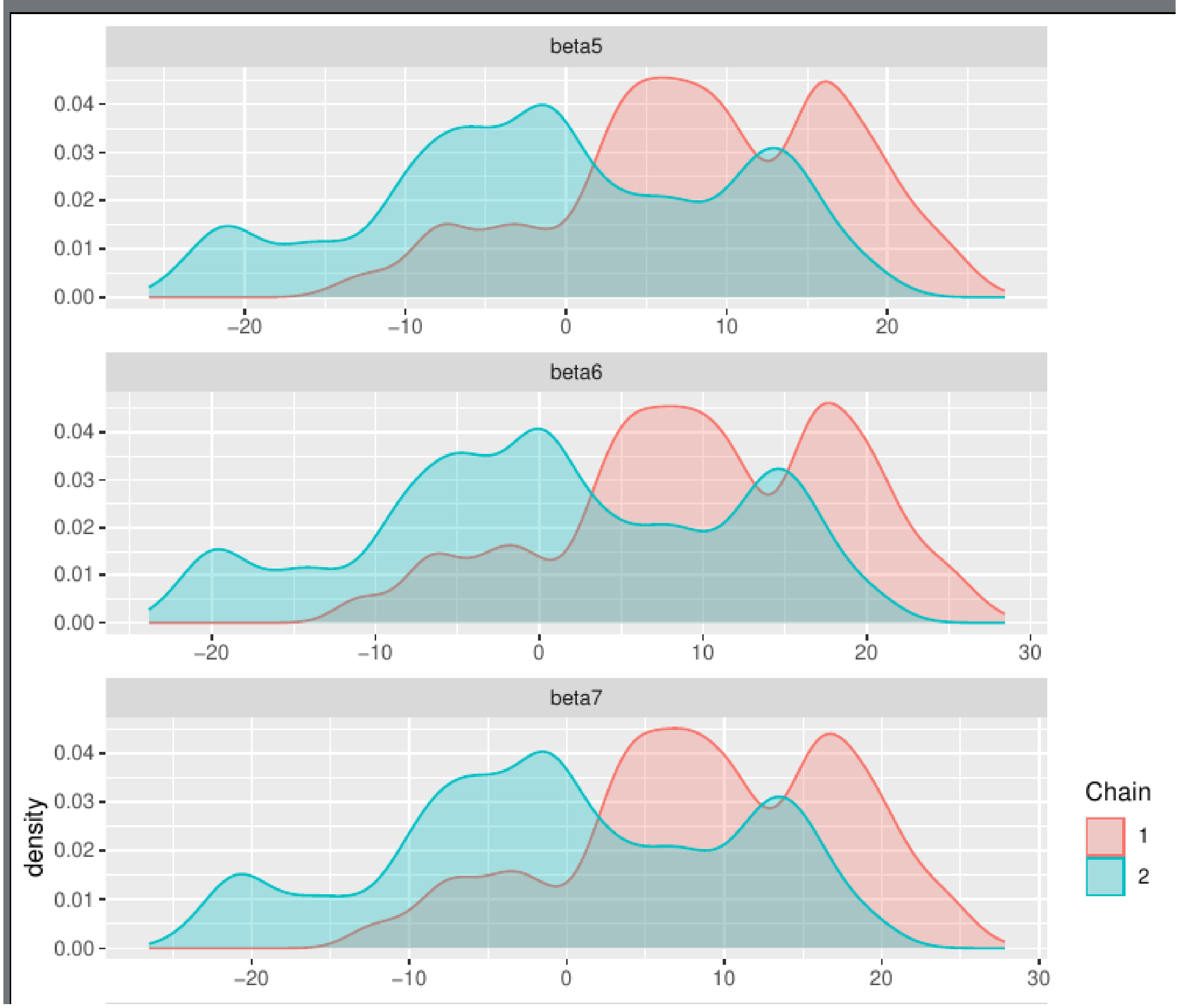
Desirable properties for these diagnostics are:

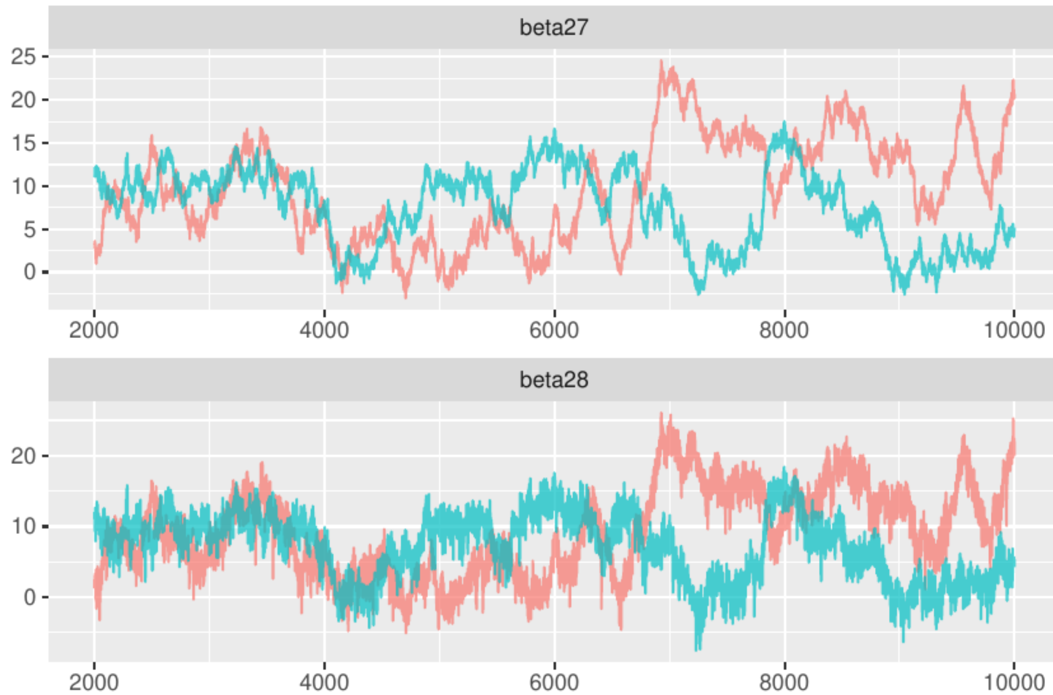
```
Rhat: Values close to 1 indicate convergence.
Effective Sample Size (neff): Higher values indicate better mixing and more reliable estimates.
Trace Plots: These should show good mixing and no apparent patterns or trends.
Density Plots: These should be smooth and unimodal for well-behaved parameters.
```


beta4
30
20
10
0
−10
−20
2000
4000
6000
8000
10000
Iteration

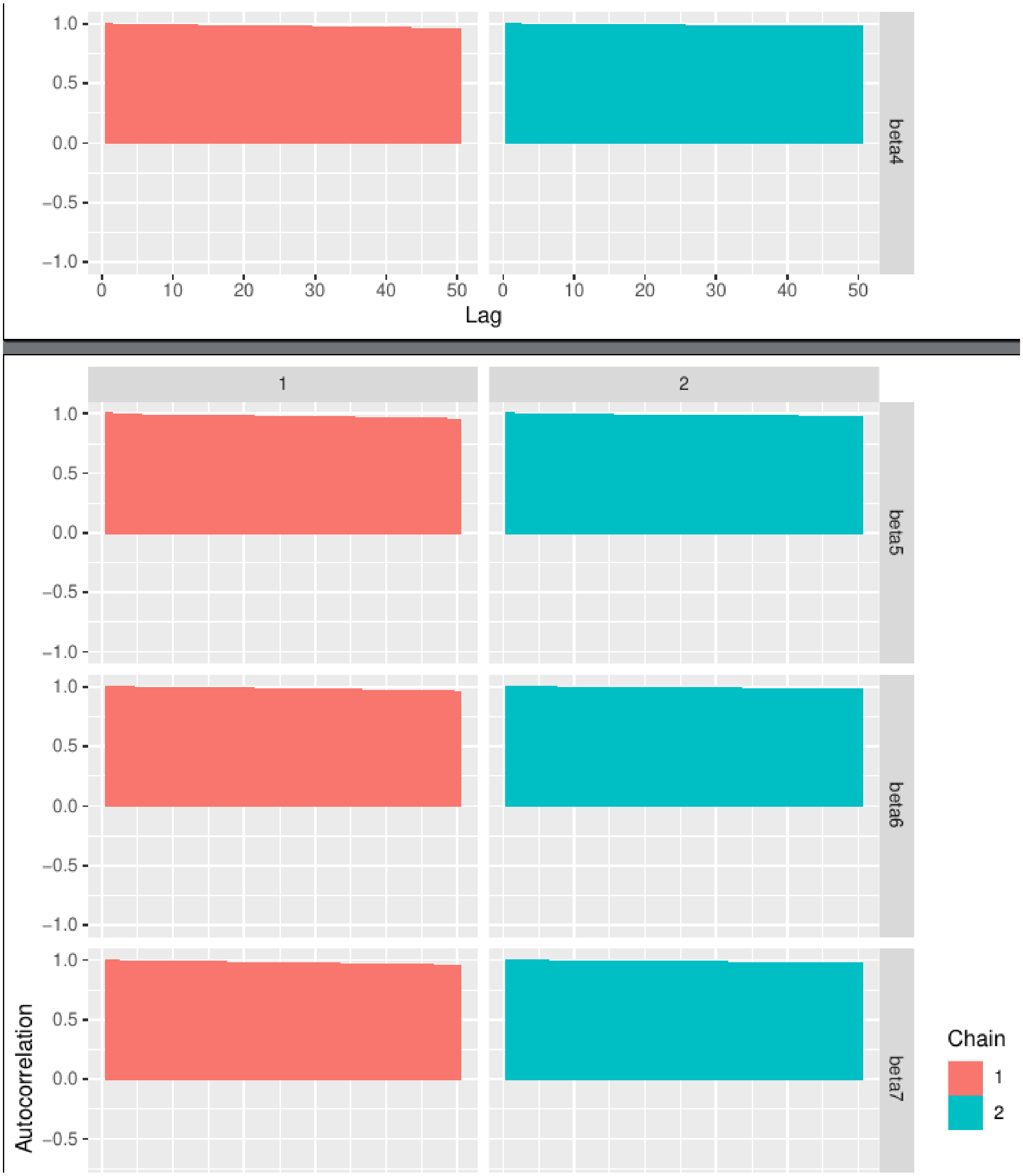

beta5
beta6
beta7
30
20
10
0
−10
−20
2000
4000
6000
8000
10000
Chain
1
2


beta4
0.04
0.03
0.02
0.01
0.00
−20
−10
0
10
20
30
value

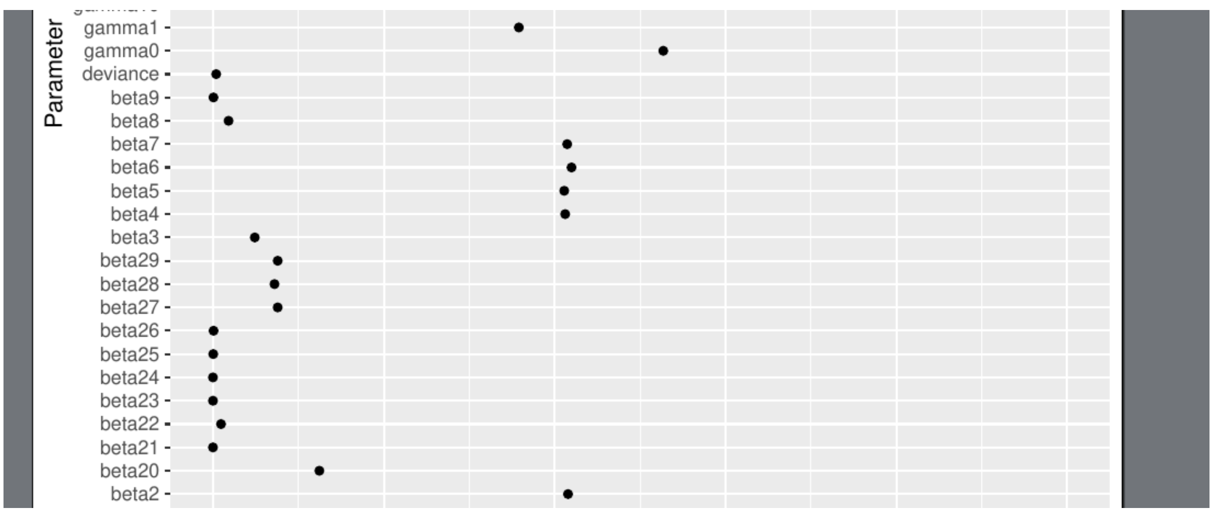

beta5
beta6
beta7
0.04
0.03
0.02
0.01
0.00
−20
−10
0
10
20
30
density
Chain
1
2



By removing the problematic coefficients, the second model aims to achieve better convergence and more reliable parameter estimates. Additionally, we initialize a second set of values in the second chain, taken from one of the two articles, allowing for comparison.



1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50
Lag
beta4
1
2
beta5
beta6
beta7
Autocorrelation
Chain
1
2




```
Inference for Bugs model at "model5_dummy.txt", fit using jags,
 2 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
 n.sims = 2000 iterations saved
           mu.vect sd.vect   2.5%    25%     50%     75%   97.5%  Rhat n.eff
beta0       -3.820   1.608 -7.116  -4.893  -3.726  -2.659  -0.834 1.001  2000
beta1        0.069   0.065 -0.045   0.025   0.064   0.110   0.214 1.001  2000
beta10      -0.529   0.719 -1.927  -1.018  -0.555  -0.043   0.905 1.003  2000
beta11       0.215   0.712 -1.125  -0.284   0.174   0.713   1.648 1.001  2000
beta12     -27.056  19.811 -74.404 -38.473 -23.470 -11.382  -1.362 1.005   650
beta13      18.289  11.040   1.406   9.758  16.844  25.939  42.457 1.068   220
beta14      -0.136  32.900 -63.186 -23.543  -0.366  22.221  64.996 1.004   420
beta15       0.357   0.627 -0.809  -0.059   0.335   0.777   1.639 1.001  2000
beta16      29.429  18.966   3.569  14.741  26.037  40.760  74.134 1.001  2000
beta17      -0.493   0.770 -2.058  -0.983  -0.468   0.016   0.964 1.004   440
beta18      -1.030   0.623 -2.344  -1.409  -1.001  -0.600   0.101 1.000  2000
beta19     -25.484  18.641 -68.769 -36.618 -21.616 -10.997  -1.214 1.000  2000
beta2        0.423   0.643 -0.866   0.001   0.448   0.868   1.628 1.001  2000
beta3       -0.369   0.780 -1.915  -0.886  -0.339   0.170   1.120 1.001  2000
beta4       -0.120   0.631 -1.329  -0.550  -0.102   0.278   1.093 1.004  1400
beta5      -24.235  19.406 -71.059 -34.796 -20.700  -8.974   0.385 1.001  2000
beta6      -25.621  18.861 -71.483 -37.132 -22.102 -10.367  -1.298 1.002   980
beta7      -26.882  19.957 -74.702 -38.152 -23.505 -11.098  -0.813 1.001  2000
beta8      -17.440  11.012 -41.574 -25.326 -15.624  -8.920  -0.754 1.066   230
beta9       -3.897  28.022 -57.235 -22.853  -2.106  15.270  49.559 1.003   620
```

```
gamma0     -13.975   7.310 -35.122 -16.739 -12.309  -9.068  -4.658 1.258    15
gamma1       0.064   0.075  -0.075   0.011   0.061   0.113   0.223 1.002  1100
gamma10     -0.251   1.562  -3.163  -1.250  -0.350   0.662   3.316 1.003  1500
gamma11      1.982   1.400  -0.386   1.000   1.824   2.746   5.294 1.004  1600
gamma12    -27.406  19.969 -72.540 -39.952 -24.409 -11.615   0.625 1.003  1900
gamma13    -11.706  25.117 -62.149 -28.019 -10.130   4.801  34.943 1.001  2000
gamma14      0.322  31.469 -62.948 -21.763   0.840  21.806  60.502 1.001  2000
gamma15     -0.678   0.845  -2.337  -1.249  -0.681  -0.123   1.016 1.002  1900
gamma16     -9.502  26.536 -67.757 -26.115  -7.186   8.446  39.338 1.002  1200
gamma17     -1.887   1.454  -5.273  -2.745  -1.706  -0.900   0.560 1.001  2000
gamma18     -0.315   0.900  -2.130  -0.883  -0.301   0.327   1.351 1.006   380
gamma19    -24.087  18.833 -67.250 -35.257 -20.724  -8.777   0.360 1.002   830
gamma2     -25.181  18.839 -71.251 -36.039 -21.284 -10.307  -1.291 1.001  2000
gamma3       7.154   7.189  -1.779   2.391   5.515   9.853  28.554 1.245    17
gamma4       9.073   7.083   0.769   4.318   7.372  11.781  30.092 1.259    16
gamma5     -16.615  21.202 -65.685 -29.815 -13.387  -0.719  14.758 1.008  2000
gamma6       0.396   1.541  -3.014  -0.514   0.570   1.457   3.015 1.001  2000
gamma7       2.719   1.171   0.373   1.961   2.735   3.510   4.995 1.001  2000
gamma8     -27.826  19.598 -72.584 -39.533 -25.104 -12.344   0.090 1.001  1700
gamma9     -10.345  25.595 -63.470 -26.963  -8.193   7.401  36.275 1.001  2000
```

```
For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 24.6 and DIC = 245.9
DIC is an estimate of expected predictive error (lower deviance is better).
```

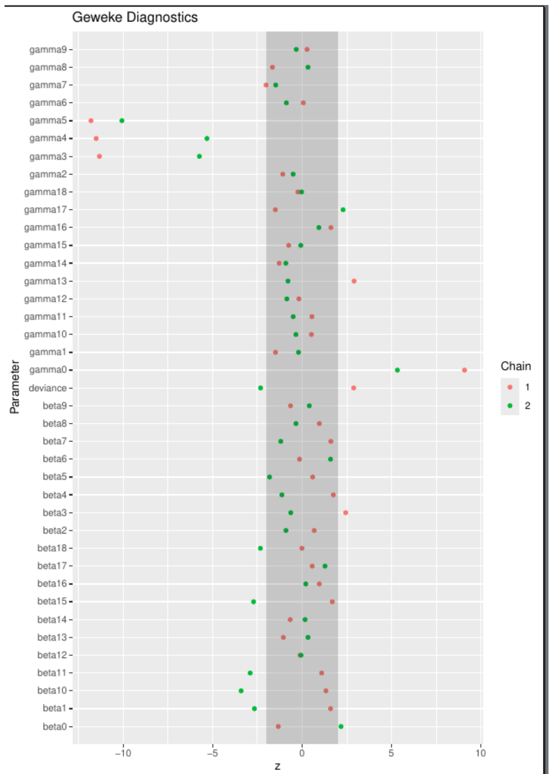
Figure 1: model5 print2

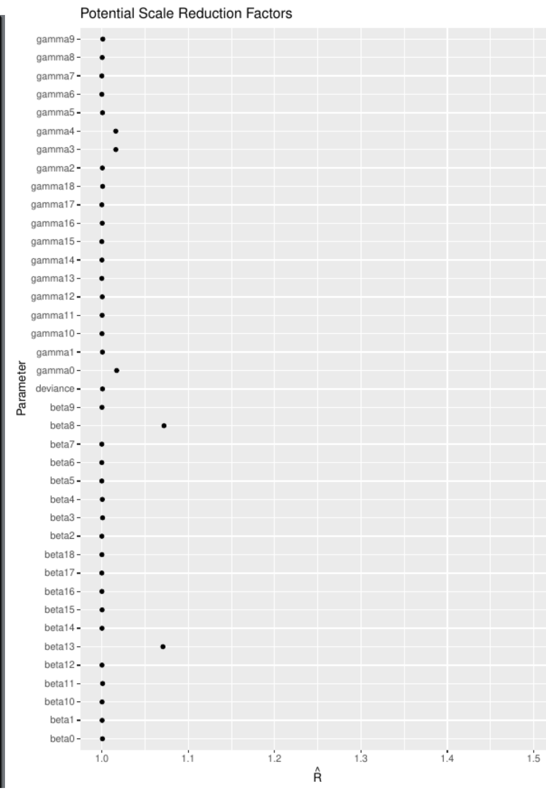
6.6 Convergence diagnostics:

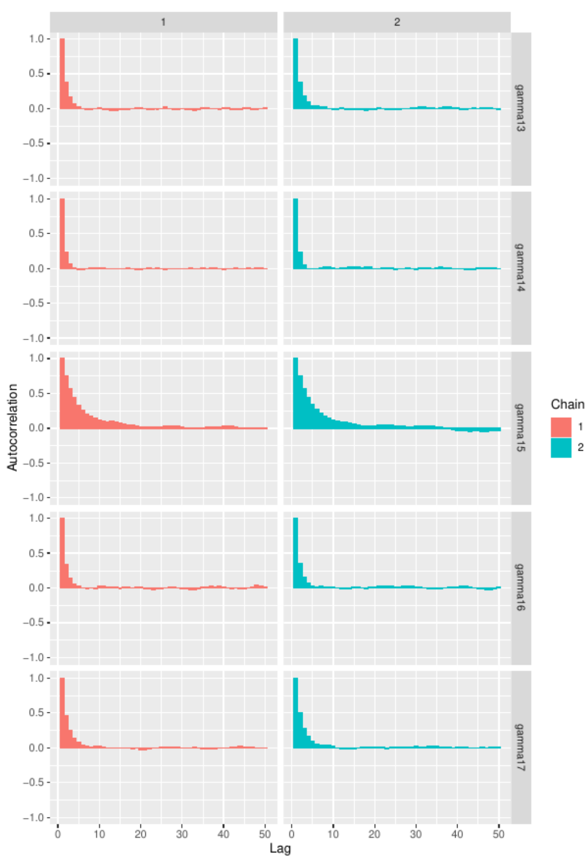
- Convergence diagnostics (such as Gelman-Rubin diagnostics) are used to assess whether the chains have mixed well and converged to a stationary distribution.
  - Parameters with $\hat{R}$ (potential scale reduction factor) close to 1 indicate good convergence.
  - In the given output, most parameters converge well, except for the excluded ones.

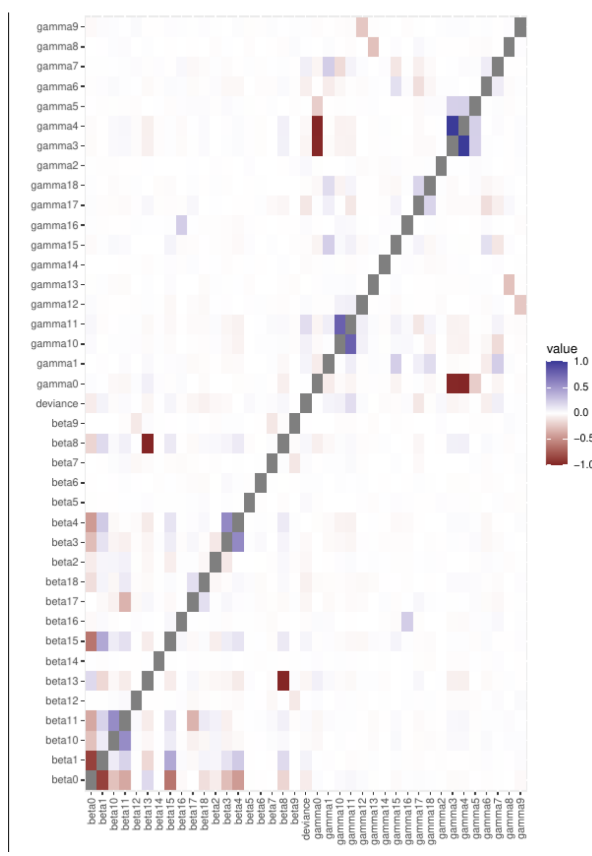
Density Plots: - Density plots for each coefficient provide insights into their posterior distributions. - The shape of these distributions can indicate if the parameters have been estimated accurately. - For instance, if the density plot is highly skewed or has multiple modes, it may suggest issues with model specification or data quality.

Geweke Diagnostics: - Geweke diagnostics compare the means of the first and last parts of the chain. - Significant deviations indicate non-convergence. - For the coefficients included in the model, the Geweke diagnostics mostly show good convergence, except for the excluded ones.

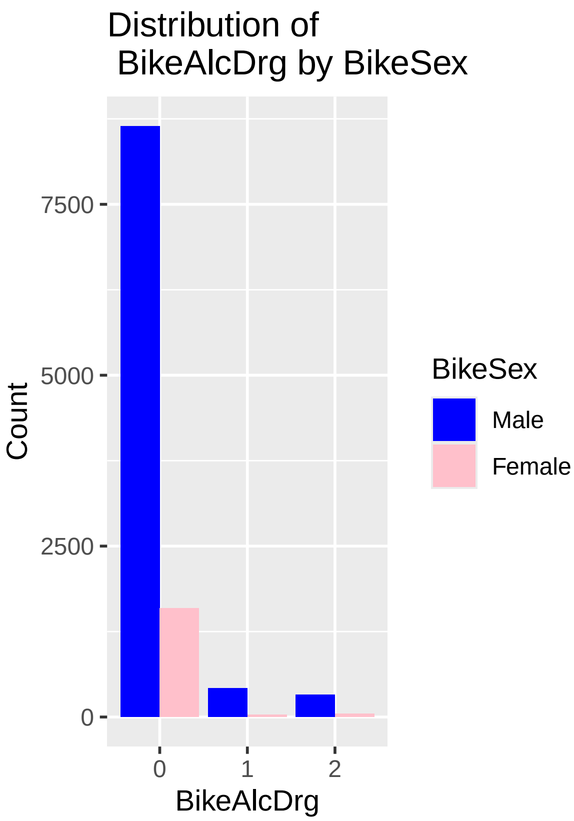

Geweke Diagnostics
Parameter
z
Chain
1
2

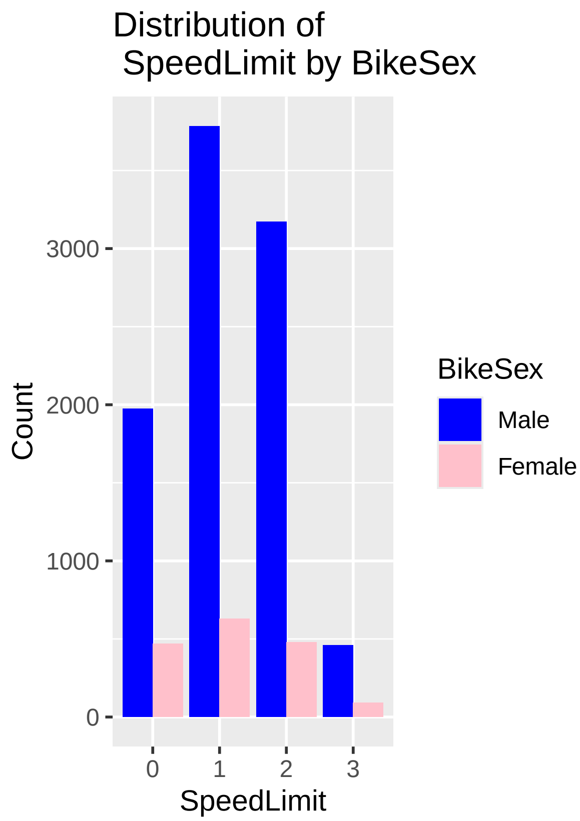

Potential Scale Reduction Factors
Parameter
R̂

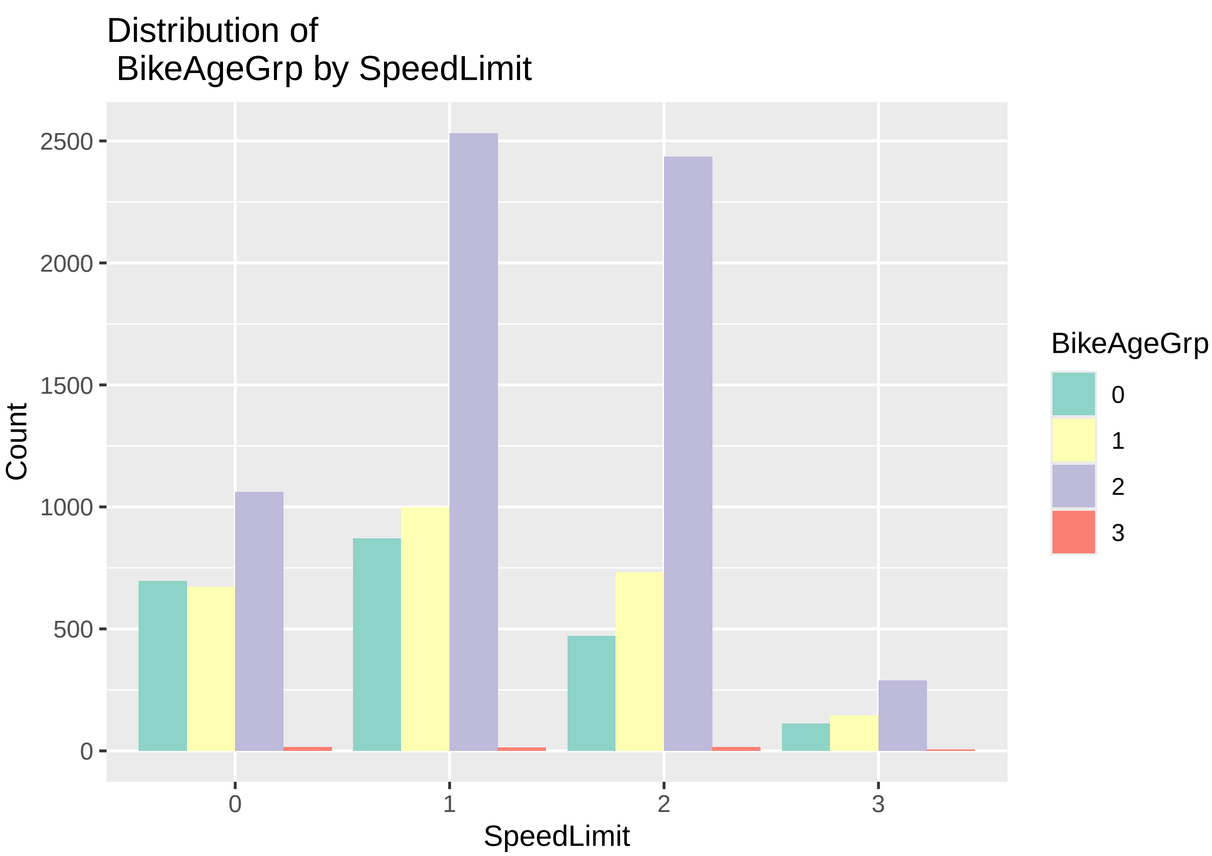

Autocorrelation
Lag
Chain
1
2
gamma13
gamma14
gamma15
gamma16
gamma17

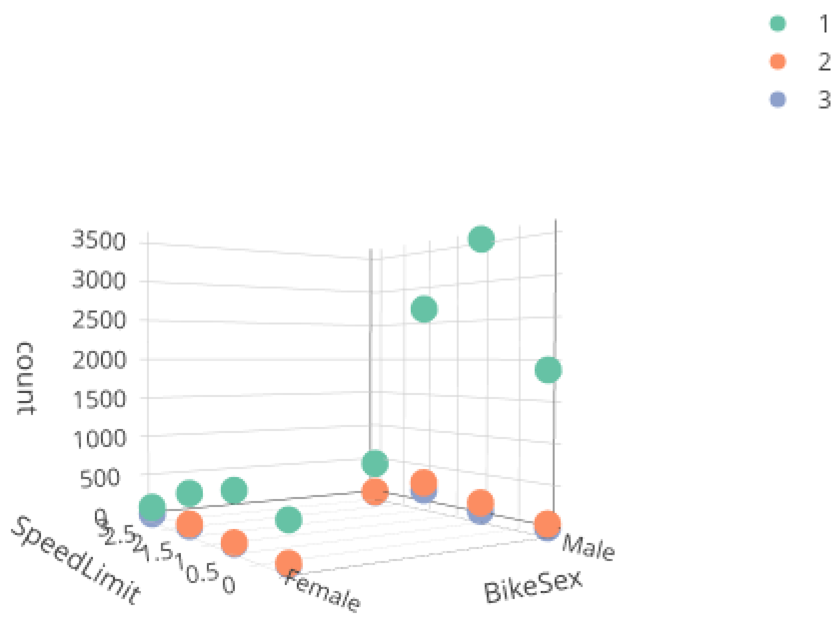

value
1.0
0.5
0.0
−0.5
−1.0

6.7 Third model









In this study, we presented plots illustrating the relationships between several variables, focusing on intuitive associations such as gender (male-female) with alcohol, gender with speed limit, and speed limit with age. These relationships have been well-documented in the literature. We proceeded by testing models where interactions were sequentially introduced. As a result, we observed a slight improvement when including the interaction between sex and speed limit. This enhancement underscores the nuanced impact of these interactions on predicting accident severity, providing valuable insights into the dynamics between demographic factors and environmental conditions.

```
> model5_plus
Inference for Bugs model at "model5_plus.txt", fit using jags,
 2 chains, each with 10000 iterations (first 2000 discarded)
 n.sims = 16000 iterations saved
        mu.vect sd.vect   2.5%     25%     50%     75%   97.5%  Rhat n.eff
beta0    -3.843   1.636 -7.391  -4.872  -3.800  -2.746  -0.824 1.006   460
beta1     0.060   0.065 -0.064   0.017   0.058   0.102   0.192 1.005   620
beta10   -0.026   0.796 -1.559  -0.562  -0.035   0.487   1.602 1.001 11000
beta11    0.426   0.805 -1.074  -0.123   0.411   0.942   2.062 1.001 16000
beta12  -29.656  19.715 -74.888 -41.785 -26.724 -14.341  -1.510 1.001 14000
beta13   19.328  16.593   0.820   7.065  14.418  26.199  63.497 1.621     6
beta14    0.370  31.339 -60.897 -20.984   0.389  21.730  61.556 1.001 16000
beta15    0.262   0.644 -0.952  -0.167   0.244   0.671   1.625 1.003   770
beta16   29.960  19.502   3.812  14.521  26.506  41.707  75.741 1.001  9200
beta17   -0.537   0.788 -2.177  -1.038  -0.498   0.001   0.915 1.001 16000
beta18   -1.009   0.652 -2.382  -1.411  -0.986  -0.569   0.202 1.001 16000
beta19  -25.725  19.229 -72.336 -36.829 -21.559 -10.551  -1.230 1.001  4200
beta2     1.155   1.061 -1.050   0.484   1.194   1.863   3.164 1.001 16000
beta20  -25.816  18.694 -70.484 -36.560 -21.840 -11.135  -1.985 1.001  5300
beta21   -0.579   1.460 -3.442  -1.545  -0.571   0.399   2.267 1.001 16000
beta22  -11.574  25.602 -65.046 -28.286  -9.785   6.067  35.687 1.001 12000
beta3    -0.353   0.807 -1.969  -0.882  -0.339   0.194   1.183 1.002  1600
beta4    -0.180   0.647 -1.411  -0.616  -0.199   0.249   1.123 1.004   560
beta5   -23.639  19.307 -69.198 -35.124 -19.691  -8.312   1.674 1.001  2900
beta6   -25.400  18.673 -70.698 -36.347 -21.692 -10.480  -1.310 1.001 16000
beta7   -26.724  19.611 -73.328 -38.203 -22.819 -11.405  -0.980 1.001 16000
beta8   -18.531  16.580 -62.812 -25.359 -13.624  -6.228  -0.316 1.622     6
beta9    -3.985  28.899 -62.653 -23.445  -2.792  16.129  49.919 1.001  9300
gamma0  -20.177  11.281 -45.296 -28.866 -16.448 -11.097  -4.656 1.143    19
gamma1    0.073   0.083 -0.078   0.015   0.069   0.125   0.244 1.009   190
gamma10  -0.311   1.523 -3.126  -1.323  -0.367   0.612   2.995 1.001  9900
gamma11   1.918   1.360 -0.387   0.983   1.785   2.681   5.082 1.002  3200
```

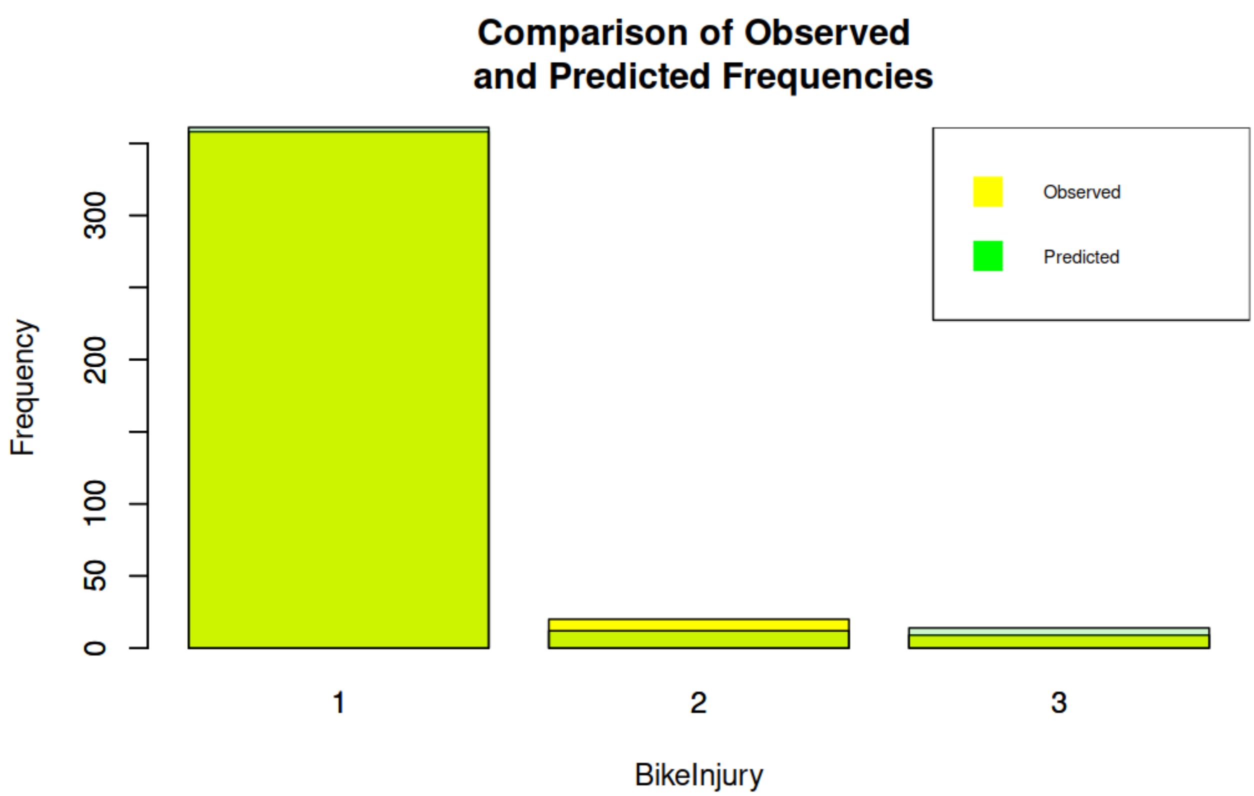
```
gamma12  -28.167  19.899 -73.886 -40.693 -25.318 -12.258   0.274 1.001 16000
gamma13  -11.841  25.188 -65.589 -27.640 -10.187   5.116  34.272 1.002  1400
gamma14    1.244  31.444 -60.028 -19.945   1.191  22.656  63.160 1.001 16000
gamma15   -0.678   0.868  -2.373  -1.249  -0.693  -0.119   1.043 1.001 16000
gamma16   -9.751  25.371 -63.627 -25.907  -8.120   7.570  37.405 1.001 16000
gamma17   -1.842   1.471  -5.180  -2.682  -1.686  -0.822   0.600 1.001  2900
gamma18   -0.274   0.908  -2.119  -0.872  -0.256   0.332   1.473 1.001  4400
gamma19  -23.889  18.865 -68.401 -34.879 -20.216  -8.943   0.683 1.001  7600
gamma2   -31.818  20.265 -77.799 -44.612 -29.430 -16.456  -0.994 1.001 16000
gamma20  -10.082  25.488 -63.992 -26.392  -8.185   7.407  35.756 1.001  6800
gamma21   -9.845  25.774 -63.834 -26.565  -8.054   7.428  38.106 1.001 16000
gamma22   -2.762  29.011 -61.728 -21.955  -1.781  17.287  52.127 1.001  9700
gamma3    13.149  10.981  -1.285   4.108   9.826  21.257  37.854 1.136    20
gamma4    15.173  10.902   1.329   6.051  11.704  23.264  39.862 1.140    20
gamma5   -13.780  23.098 -64.866 -28.376 -11.136   2.764  25.408 1.006   910
gamma6     0.417   1.484  -2.913  -0.461   0.541   1.438   2.940 1.001 16000
gamma7     2.758   1.205   0.317   1.988   2.766   3.555   5.123 1.004   450
gamma8   -28.261  19.864 -73.619 -40.386 -25.294 -12.585  -0.519 1.001  7000
gamma9    -9.884  25.954 -64.653 -26.442  -8.252   7.877  37.050 1.002  1300
deviance 219.759   7.223 207.691 214.625 219.104 224.059 235.823 1.001 16000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 26.1 and DIC = 245.8
DIC is an estimate of expected predictive error (lower deviance is better).
```

7 Predicted values

As we learnt from "Introduction to Bayesian Models: Normal Models" (2009) that: the posterior predictive density $f(y'|\mathbf{y}, m)$ of a model m is frequently used for checking the assumptions of a model and its goodness-of-fit. The main reason is that we can easily generate replicated values y^{rep} from the posterior predictive distribution by adding a single simple step within any MCMC sampler using the likelihood function $f(y^{rep}|\theta(t))$ evaluated at parameter values $\theta(t)$ of the current state of the algorithm; see Section 10.2.

```
model52_balanced <- jags.model("model5_dummy.txt",
                               data = dd5_2,
                               inits = inits5_2,
                               n.chains = 2,
                               n.adapt = 10000)
# Getting the samples
samples <- jags.samples(model52_balanced,
                        variable.names =  params5_2,
                        n.iter = 1)
str(samples)

y_rep <- samples$x_rep[,,1]
```

	1	2	3
Obs	0.93	0.05	0.02
Pred	0.93	0.03	0.04



Further Investigation Using Cross-validation Predictive Densities

In addition to the traditional methods used to evaluate model fitness, an alternative approach that can be employed is the use of cross-validation predictive densities. Although the full predictive distribution ($f(y'|y)$) is beneficial for prediction, it is less suitable for model checking due to the potential issue of double data use. To address this, several authors, including Gelfand, Dey, and Chang (1992); Gelfand (1996); Vehtari and Lampinen (2003); and Draper and Krnjajić (2006), have proposed cross-validatory predictive densities.

This method involves dividing the data (y) into two subsets ((y_1) and (y_2)). The first subset (y_1) is used to fit the model and estimate the posterior distribution of interest. The remaining observations (y_2) are then used for model evaluation and checking by calculating the cross-validatory predictive density:

$f(y_2|y_1) = \int f(y_2|\theta)f(\theta|y_1)d\theta$

A notable challenge with this approach is selecting (y_1) and (y_2) since different splits can yield varying results. To mitigate this, Geisser and Eddy (1979) proposed the leave-one-out cross-validation (CV-1) predictive density, which simplifies the process by considering each observation (y_i) and its complement ($y_{\setminus i}$):

$f(y_i|y_{\setminus i}) = \int f(y_i|\theta)f(\theta|y_{\setminus i})d\theta$

This quantity, also known as the conditional predictive ordinate (CPO), provides a quantitative measure of the effect of observation (i) on the overall prior predictive density ($f(y)$):

$\text{CPO}_i = f(y_i|y_{\setminus i}) = \frac{f(y)}{f(y_{\setminus i})}$

The CPO is equivalent to the posterior predictive ordinate (PPO) and is useful for identifying outliers. Small CPO values indicate observations that are poorly predicted by the model. An overall measure of fit can be constructed by the product of CPOs, referred to as the cross-validation predictive likelihood, which is further elaborated in Chapter 11 “Bayesian Model and Variable Evaluation” (2009) of the referenced book.

8 Frequentist and Bayesian

In the following table, we compare the coefficients obtained from the frequentist model with those obtained from the Bayesian model for the key variables in the study. The parameters β correspond to η_2 and γ correspond to η_3.

Variable	**Frequentist Coefficients**	**Bayesian Coefficients**	**Mean**	**95% CI**
Intercept	-3.318	β_0	-3.841	[-7.270, -0.663]
CrashHour	0.058	β_1	0.071	[-0.059, 0.207]
BikeSexFemale	0.458	β_2	0.404	[-0.953, 1.587]
BikeAgeGrp1	-0.309	β_3	-0.361	[-1.892, 1.095]
BikeAgeGrp2	-0.166	β_4	-0.153	[-1.320, 1.135]
BikeAgeGrp3	-51.534	β_5	-23.767	[-68.688, 0.841]
BikeAlcDrg1	-26.656	β_6	-25.456	[-70.415, -1.298]
BikeAlcDrg2	-30.423	β_7	-26.309	[-71.270, -1.203]
RdConditio1	-41.582	β_8	-14.306	[-37.744, -0.009]
RdConditio2	0.228	β_9	-4.853	[-64.924, 48.851]
SpeedLimit1	-0.498	β_{10}	-0.505	[-1.920, 0.905]
SpeedLimit2	0.185	β_{11}	0.257	[-1.104, 1.650]
SpeedLimit3	-26.296	β_{12}	-26.918	[-73.302, -1.387]
Weather1	42.526	β_{13}	15.102	[0.594, 38.502]
Weather2	0	β_{14}	-0.102	[-61.417, 62.061]
LightCond1	0.233	β_{15}	0.359	[-0.877, 1.690]
LightCond2	57.240	β_{16}	29.720	[3.513, 73.975]
TraffCntrl2	-0.408	β_{17}	-0.546	[-2.206, 0.910]
TraffCntrl3	-0.916	β_{18}	-1.035	[-2.384, 0.169]
TraffCntrl4	-33.567	β_{19}	-25.273	[-69.757, -1.169]
Variable	**Frequentist Coefficients**	**Bayesian Coefficients**	**Mean**	**95% CI**
Intercept	-39.067	γ_0	-27.655	[-48.647, -5.775]
CrashHour	0.052	γ_1	0.076	[-0.077, 0.254]
BikeSexFemale	-31.848	γ_2	-25.351	[-69.868, -1.446]
BikeAgeGrp1	33.864	γ_3	20.335	[-0.751, 42.332]
BikeAgeGrp2	35.165	γ_4	22.336	[1.429, 44.296]
BikeAgeGrp3	-0.251	γ_5	-10.433	[-63.340, 32.257]
BikeAlcDrg1	0.748	γ_6	0.416	[-3.035, 3.014]
BikeAlcDrg2	2.456	γ_7	2.751	[0.350, 5.047]
RdConditio1	-13.817	γ_8	-27.746	[-73.814, -0.396]
RdConditio2	-5.501	γ_9	-10.248	[-64.119, 37.245]
SpeedLimit1	-0.425	γ_{10}	-0.084	[-3.067, 4.345]
SpeedLimit2	1.413	γ_{11}	2.171	[-0.331, 6.524]
SpeedLimit3	-28.123	γ_{12}	-27.192	[-73.137, 0.493]
Weather1	-2.024	γ_{13}	-12.556	[-65.842, 34.862]
Weather2	0	γ_{14}	0.943	[-61.310, 63.410]
LightCond1	-0.668	γ_{15}	-0.624	[-2.284, 1.161]
LightCond2	-0.355	γ_{16}	-9.796	[-62.505, 36.505]
TraffCntrl2	-1.366	γ_{17}	-1.870	[-5.292, 0.580]
TraffCntrl3	-0.224	γ_{18}	-0.266	[-2.190, 1.519]
TraffCntrl4	-40.358	γ_{19}	-24.132	[-70.485, 0.638]

Table 2: Comparison of Frequentist and Bayesian Coefficients

9 Conclusion

The exclusion of non-converging coefficients (CrashDay's and NumLanes's) improved model performance. The analysis of the remaining coefficients provided sometimes logical insights. The model could be improved if we look at the CI, even tough Bayesian and Frequentist approximations of the coeffients were coherent showing converging of the series.

Bibliography

"Bayesian Model and Variable Evaluation." 2009. In *Bayesian Modeling Using WinBUGS*, 389–433. John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9780470434567.ch11.

"Introduction to Bayesian Models: Normal Models." 2009. In *Bayesian Modeling Using WinBUGS*, 151–87. John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/9780470434567.ch5.

Nowakowska, Marzena. 2017. "Selected Aspects of Prior and Likelihood Information for a Bayesian Classifier in a Road Safety Analysis." *Accident Analysis & Prevention* 101: 97–106. https://doi.org/10.1016/j.aap.2017.01.009.

Yang, Zaili, Zhisen Yang, John Smith, and Bostock Adam Peter Robert. 2021. "Risk Analysis of Bicycle Accidents: A Bayesian Approach." *Reliability Engineering and System Safety* 209 (C). https://doi.org/10.1016/j.ress.2021.10746.