STATS 765 Project

Car Crashes and Drinking Driving Regulations in New Zealand

Emma Wang

ewan538@auckladnuni.ac.nz

Outline

1. Background and Goal
2. Data Source
3. Data Processing
4. Data Exploration
5. Analytics and Results
6. Discussion

Section 1

Background and Goal

Drinking driving in New Zealand

Drinking driving was a **long-standing controversial topic**.

"In 2008, driver alcohol/drugs was a contributing factor in 103 fatal crashes, 441 serious injury crashes and 1156 minor crashes."

–Ministry of Health, 2010

"Police estimate that each day in New Zealand, approximately 5923 compulsory breath tests and 2743 mobile breath tests are undertaken and 100 people are charged with drink driving."

–New Zealand Police, 2010

"In 2020, alcohol was a factor in 90 deaths and 262 serious injuries."

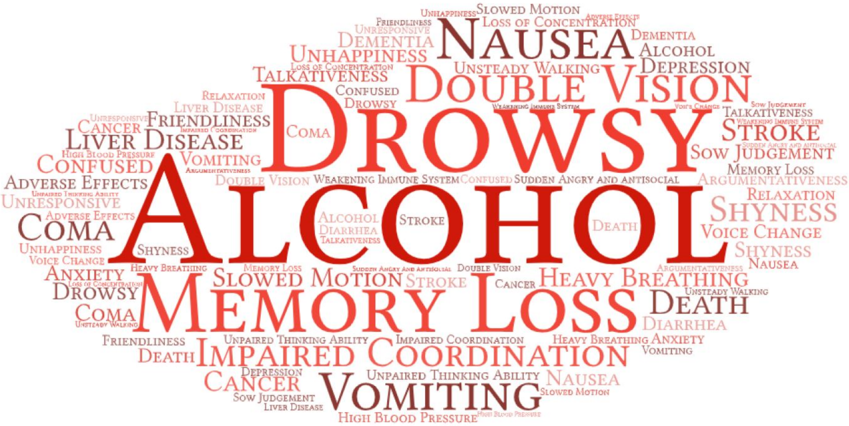
–NZTA, 2022

Alcohol Fact Sheet–Ministry of Health 2010, `https://www.nzta.govt.nz/safety/driving-safely/alcohol/`

Alcohol–NZTA 2022, `https://www.nzta.govt.nz/safety/driving-safely/alcohol/`

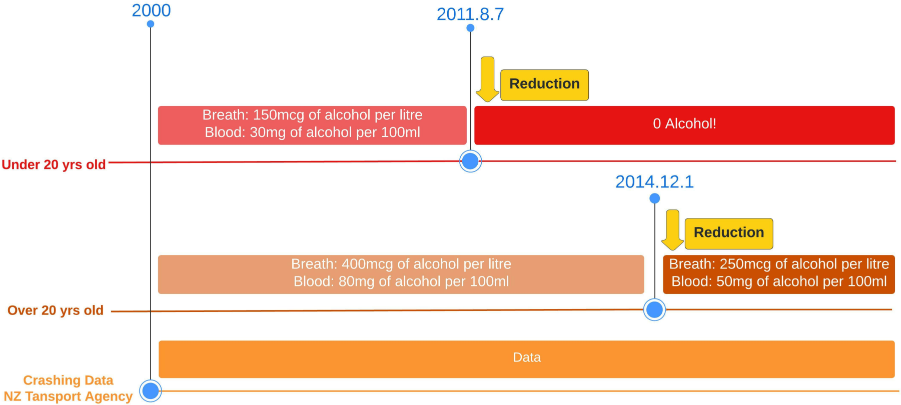
Adverse Effects of Alcohol

Some of the well-known adverse effects of drinking alcohol.



Timeline of Alcohol Limit Changes

The Cabinet has **restricted** the alcohol limit in 2011 and 2014.
The new rules were **documented** in Land Transport Act 1998.


2000
2011.8.7
Reduction
Breath: 150mcg of alcohol per litre
Blood: 30mg of alcohol per 100ml
0 Alcohol!
Under 20 yrs old
2014.12.1
Reduction
Breath: 400mcg of alcohol per litre
Blood: 80mg of alcohol per 100ml
Breath: 250mcg of alcohol per litre
Blood: 50mg of alcohol per 100ml
Over 20 yrs old
Data
Crashing Data
NZ Tansport Agency

Information Available from Ministry of Transport, https://www.transport.govt.nz/about-us/what-we-do/queries/regulation-of-drink-driving-limits/

Goal of the project

The goal of this project is to

- Analyze the crashing data published by NZTA and find out whether the two changes in drink-driving limits (2011/2014) affected the crashing events across New Zealand;
- Gain proficiency in using the techniques to import, clean, impute and build models for large data sets.

Section 2

Data Source

Primary Data Source

Crashing Analysis System

- Publisher: NZ Transport Agency
- Time Period: 2000-2021
- Dimensions: 776878 rows, 72 columns
- Structure: Each row recorded 1 crash

```r
data_raw <- fread("Crash_Analysis_System_(CAS)_data.csv")
dim(data_raw)

## [1] 776878     72
```

Data Source: Crash Analysis System–NZTA, `https://opendata-nzta.opendata.arcgis.com/search?tags=crashes`

Key Response Variables

Key response variables in the data:

- `crashSeverity`: the severity of the most severe case in the data, either `non-injury`, `minor`, `serious` and `fatal`.
- `minorInjuryCount`: the number of minor injuries in the crash.
- `seriousInjuryCount`: the number of serious injuries in the crash.
- `fatalInjuryCount`: the number of deaths in the crash.

Key Covariates

Some of the key covariates:

- `crashYear`: the year when the crash took place, integer between 2000–2021.
- `weatherA`: Weather at the crashing location, either `Fine`, `Mist`, `Light Rain`, `Heavy Rain`, `Snow` or `Unknown`.
- `urban`: urban road or open road, either `Urban` or `Open Road`.

And more. . .

Field Description, `https://opendata-nzta.opendata.arcgis.com/pages/cas-data-field-descriptions`

Section 3

Data Processing

Lack of the alcoholic states of the drivers

Issue: Cannot bulid models for Crash Severity vs. alcoholic measurement.

Solution: Using `crashYear` or derived variables, we can analyze how the Crash Severity changed after 1st and 2nd reductions instead.

Categorization of Time Periods

Issue: The time of each crash was recorded in Year. Inaccurate categories for cases in 2011 and 2014.

Solution: Both changes were near the end of the year (2011.8 and 2014.12).

Cases at the boundary years were classified into the previous periods.

Table: The categorization of the variable `Change`.

Period	**Label**	**Change**
2000-2011	Baseline	0
2012-2014	Post-First Change	1
2015-2021	Post-Second Change	2

High degree of missingness I

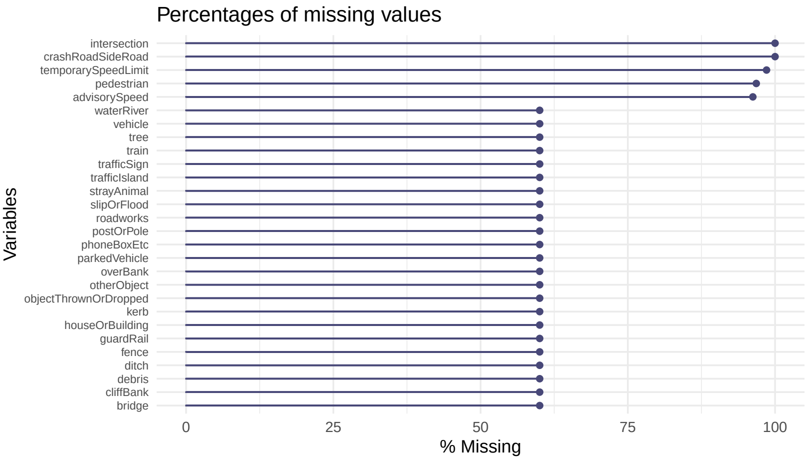
Issue: Lots of missing values in the data.

There were 49 variables with missing values, and there were no complete records.

```
library(naniar)
miss_var_summary(data_raw) %>% filter(pct_miss > 0) %>%
  select(variable) %>% unlist %>% length

## [1] 49

pct_complete_case(data_raw)

## [1] 0
```

High degree of missingness II

Below are all the variables with more than 50% missingness.
5 of those were completely missing.



High degree of missingness III

Solution:

- Delete variables with high missingness ($\geq 90\%$)
- For the other variables (missingness $< 90\%$), determine their importance by:
    - Analyze their correlations with the response
    - Key variables affecting the seveity of crash crashes, using common sense;
- Do Multiple Imputations on the important variables.

Section 4

Data Exploration

Crash Severity and Periods of Change

- There were significant reduction in crashing numbers after first change in 2011.
- Average counts increased after the second reduction in alcohol.

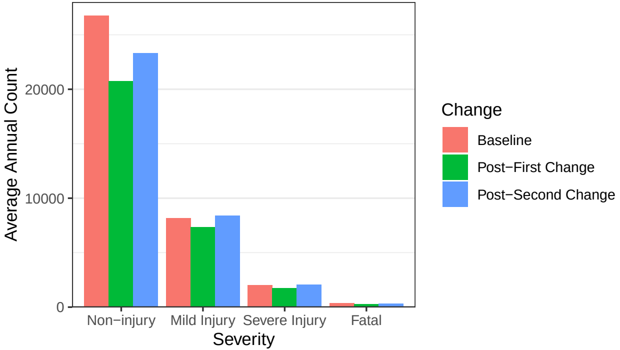


Fig 1. Distribution of the average annual counts of the most severe cases in the crashes across 3 periods of alcohol-limit changes. The average counts during the Post-First-change period were lower than that in the other period.

Injury Counts and Periods of Change

- `Minor Injuries`: Gradient changes at 2012, 2014 and 2016.
- No appearant patterns for `serious` or `fatal` injuries, except for the dramatical increase after 2020.

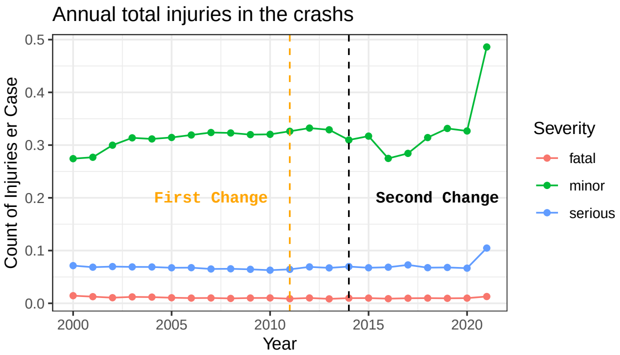


Fig 2. Changes in average injury counts per 10 cases of various severity from 2000 to 2021. The two changes in drinking-driving regulations (2011, 2014) were marked by the dashed vertical lines. Minor injury was the leading type across the time.

Correlation Analysis

No strong correlations between covariates and the response variables.

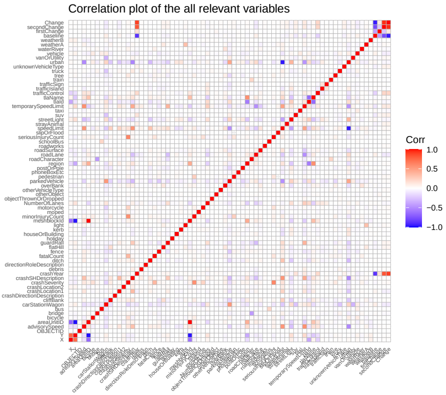


Fig 3. Correlation matrix of all variables in the data set.

Other variables to be imputed

Using common sense, below is a list of variables affecting the severity of the cases.

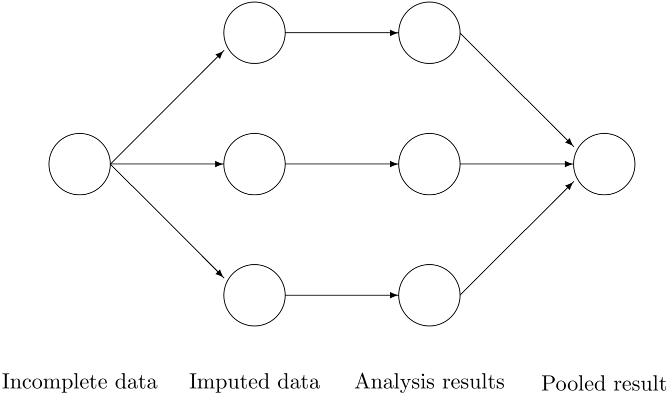
```r
road_conditions <- c("roadSurface", "NumberOfLanes",
                    "speedLimit", "light","roadLane",
                    "streetLight", "trafficControl",
                    "flatHill","roadworks")
location <- c("urban", "region")
other_vehicles <- c("bicycle", "train", "slipOrFlood",
                   "vehicle", "schoolBus","truck",
                   "vanOrUtility", "bus")
time <- c("Change", "crashYear")
weather <- c("weatherA", "weatherB")
other_objects <- c("strayAnimal","bridge",
                  "waterRiver", "cliffBank")
```

Section 5

Analytics and Results

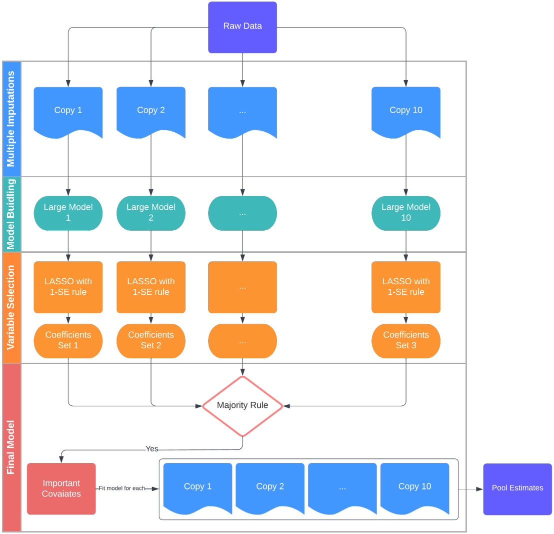
General Work Flow for Multiple Imputation Data I

In Section 5 of Stef's Book on `mice` package [1], he suggested a work flow for imputed data:



After some modifications, below is the work flow for our data.

General Work Flow for Multiple Imputation Data II



Step 1: Multiple Imputations

1. Determined the list of variables to be imputed.
2. `dryrun` to check for initial settings;
3. Modify the settings for `mice()`:
   - `methods`
   - `prediction matrix`
   - `post processing`
4. Run Multiple Imputation and obtain 10 copies of data;
5. Diagnostics.
   - Convergence and Strip plots

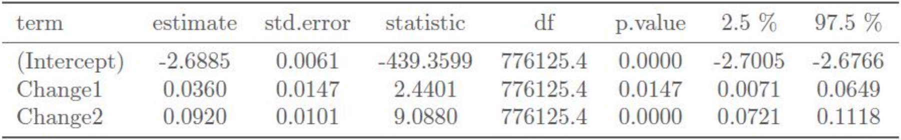
Step 2: Model Buildings–Direction 1 I

Direction 1: The relationship between `crashSeverity` and `Change`.

A logistic regression was built on the probability of fatal or serious crash.

1. Unadjusted model

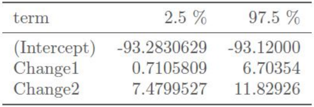
```r
expr1 <- expression(glm(I(crashSeverity %in% c("F", "S")) ~
                          Change, family = "binomial"))
# Apply the expression to each of the 10 copies
fit_unadj <- with(imp_all, expr1)
# Pool the summaries
summary(pool(fit_unadj), conf.int = TRUE)
```



term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
(Intercept)	-2.6885	0.0061	-439.3599	776125.4	0.0000	-2.7005	-2.6766
Change1	0.0360	0.0147	2.4401	776125.4	0.0147	0.0071	0.0649
Change2	0.0920	0.0101	9.0880	776125.4	0.0000	0.0721	0.1118

Step 2: Model Buildings–Direction 1 II

```
# Extract the confidence intervals
confint_unadj <- sum_unadj %>% select(‘2.5 %‘, ‘97.5 %‘)
cbind(term = sum_unadj[,"term"], 100*(exp(confint_unadj)-1))
```



term	2.5 %	97.5 %
(Intercept)	-93.2830629	-93.12000
Change1	0.7105809	6.70354
Change2	7.4799527	11.82926

Although *p*-value for the 2 terms were highly significant, the probabilities of serious or fatal cases increased after the 2 changes. (Dose this make sense?)

Let’s look at the adjusted model instead.

Step 2: Model Buildings–Direction 1 III

Direction 1: The relationship between `crashSeverity` and `Change`.

Build a logistic regression on the probability of fatal or serious crash.

1. Unadjusted model
2. Adjusted model

```
# Function for model selection
select_model <- function(...){
  library(glmnet)
  library(Matrix)
  library(tidyverse)
  i <- (...)
  # Select the data from ith imputation
  data_model <- comp_long %>%
  filter(`.imp` == i) %>%
  select(all_of(var_pred), crashSeverity)}
```

Step 2: Model Buildings–Direction 1 IV

```
# X matrix
X <- data_model %>%
select(all_of(var_pred)) %>%
sapply(as.numeric) %>%
Matrix()
# response
y <- data_model %>%
with(if_else(crashSeverity == c("F", "S"), 1, 0))
#  Fit model
fit <- glmnet(X, y, family = "binomial")
# Cross Validation to choose for lambda
xval <- cv.glmnet(X, y)
# Extract coefficients using 1-SE rule
return(coef(fit, s = xval$lambda.1se))
```

Step 2: Model Buildings–Direction 1 V

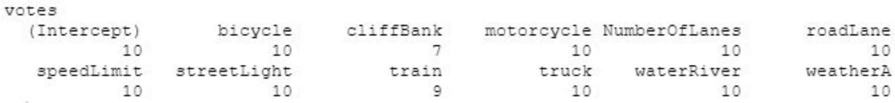
Using parallel computation to loop over all the copies of data:

```
# Make new cluster
cl <- makeCluster(5)
# Export the objects to each cluster
clusterExport(cl, c("select_model", "comp_long", "var_pred"))
# Do the computation
par_output <- parLapply(cl, 1:m, fun = select_model)
# Stop the cluster
stopCluster(cl)
```

Step 2: Model Buildings–Direction 1 VI

Use the majority rule:

```
fit_all <- par_output %>% as.mira
# Find the names of variables where the coefficients were not
terms <- lapply(fit_all$analyses,
                function(x) row.names(x)[which(x != 0)])
# Use majority rule
votes <- unlist(terms)
table(votes)
```



```
votes
  (Intercept)       bicycle      cliffBank    motorcycle NumberOfLanes      roadLane
           10            10              7            10            10            10
   speedLimit   streetLight          train         truck    waterRiver      weatherA
           10            10              9            10            10            10
```

Those variables would be included in the final adjusted model.

Step 2: Model Buildings–Direction 1 VII

The final adjusted model is:

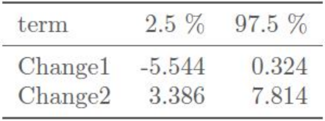
```
expr2 <- expression(glm(I(crashSeverity %in% c("F", "S")) ~
             Change + bicycle + cliffBank
            + motorcycle + NumberOfLanes
            + roadLane + speedLimit + streetLight
            + train + truck + waterRiver
            + weatherA, family = "binomial"))
fit_adj <- with(imp_all, expr2)
summary(pool(fit_adj), conf.int = TRUE)
```

Step 2: Model Buildings–Direction 1 VIII

term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
(Intercept)	-4.833	0.041	-117.664	23829.546	0.000	-4.914	-4.753
Change1	-0.027	0.015	-1.750	739118.446	0.080	-0.057	0.003
Change2	0.054	0.011	5.073	40334.152	0.000	0.033	0.075
bicycle	1.589	0.019	82.858	48157.041	0.000	1.552	1.627
cliffBank	0.109	0.021	5.079	35.982	0.000	0.066	0.153
motorcycle	2.003	0.015	135.223	227283.707	0.000	1.974	2.032
NumberOfLanes	-0.130	0.006	-20.098	10997.001	0.000	-0.143	-0.117
roadLane2-way	1.081	0.024	44.228	18422.447	0.000	1.033	1.129
roadLaneNull	0.826	0.248	3.326	629656.885	0.001	0.339	1.312
roadLaneOff road	1.678	0.046	36.724	67986.188	0.000	1.588	1.767
speedLimit	0.020	0.000	77.246	1840.222	0.000	0.019	0.020
streetLightNull	-0.234	0.013	-17.544	163149.772	0.000	-0.260	-0.208
streetLightOff	-0.468	0.018	-26.731	35861.969	0.000	-0.502	-0.433
streetLightOn	0.015	0.017	0.887	223074.756	0.375	-0.018	0.049
train	1.123	0.151	7.438	23.084	0.000	0.811	1.435
truck	0.337	0.015	22.495	421056.271	0.000	0.308	0.366
waterRiver	0.626	0.082	7.616	15.506	0.000	0.451	0.801
weatherAHail or Sleet	0.174	0.328	0.532	771964.408	0.595	-0.468	0.817
weatherAHeavy rain	-0.218	0.025	-8.685	746897.086	0.000	-0.267	-0.169
weatherALight rain	-0.224	0.014	-15.671	607242.994	0.000	-0.252	-0.196
weatherAMist or Fog	-0.148	0.037	-3.962	745445.467	0.000	-0.220	-0.075
weatherANull	-1.054	0.064	-16.334	773904.260	0.000	-1.181	-0.928
weatherASnow	-0.282	0.103	-2.742	687901.572	0.006	-0.483	-0.080

Step 2: Model Buildings–Direction 1 IX

```
# Extract the confidence intervals
confint_adj <- sum_adj %>%
filter(term %in% c("Change1", "Change2")) %>%
  select(‘2.5 %‘, ‘97.5 %‘)
cbind(term=c("Change1", "Change2"),100*(exp(confint_adj)-1))
```



term	2.5 %	97.5 %
Change1	-5.544	0.324
Change2	3.386	7.814

In the adjusted model, the coefficient of `Change2` was significant, but not for `Change1`.

- Comparing the the baseline period, there was 95% confidence that the odds of fatal or serious cases after the second reduction of alcohol limit in 2014 increased by somewhere between 3.39% to 7.81%.
- No significant changes seen after 1st reduction.

Step 2: Model Buildings–Direction 2 I

Direction 2: The relationship between `minorInjuryCount` and `crashYear`.

Build a piece-wise linear regression on the counts of minor injuries per case.

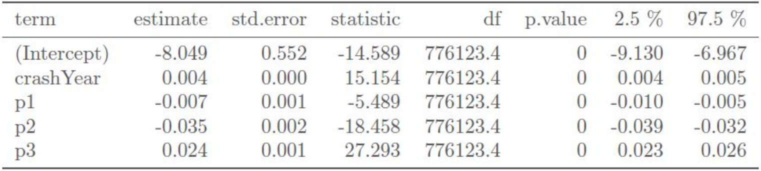
Table: The segments of `crashYear` in the piece-wise linear regression.

Period	**Code in the Model**
2011-2014	`pmin(pmax(crashYear - 2011, 0), 3)`
2014-2016	`pmin(pmax(crashYear - 2014, 0), 2)`
2016-2021	`pmax(crashYear - 2016, 0)`

Step 2: Model Buildings–Direction 2 II

1. Unadjusted model

```
fit_lm_unadj <- list(10)

for(i in 1:10){
comp <- comp_long %>% filter(‘.imp‘ == i)
fit_lm_unadj[[i]] <- lm(minorInjuryCount ~ crashYear
                        + p1 + p2 + p3, data = comp)}
mira_lm_unadj <- fit_lm_unadj %>% as.mira
# Pool the estimates for confidence intervals
summary(pool(mira_lm_unadj), conf.int = TRUE)
```



term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
(Intercept)	-8.049	0.552	-14.589	776123.4	0	-9.130	-6.967
crashYear	0.004	0.000	15.154	776123.4	0	0.004	0.005
p1	-0.007	0.001	-5.489	776123.4	0	-0.010	-0.005
p2	-0.035	0.002	-18.458	776123.4	0	-0.039	-0.032
p3	0.024	0.001	27.293	776123.4	0	0.023	0.026

Step 2: Model Buildings–Direction 2 III

Direction 2: The relationship between `minorInjuryCount` and `crashYear`.

Build a piece-wise linear regression on the counts of minor injuries per case.

1. Unadjusted model
2. Adjusted model

The process was similar to ***Direction 1***, we used a for loop and LASSO.

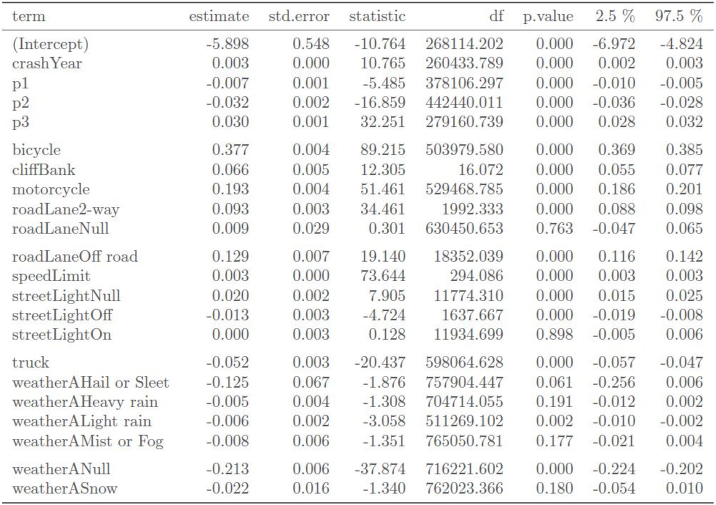
The variables `bicycle`, `cliffBank`, `motorcycle`, `roadLane`, `speedLimit`, `streetLight` and `weatherA` would be included in the final model.

Step 2: Model Buildings–Direction 2 IV

The final adjusted model is:

```r
# Fit model for 10 copies of data
fit_lm_adj <- list(10)
for(i in 1:10){
  comp <- comp_long %>% filter(‘.imp‘ == i) %>%
    select(all_of(var_pred_lm), minorInjuryCount)
  fit_lm_adj[[i]] <- lm(minorInjuryCount ~ crashYear
                        + p1 + p2 + p3 + bicycle
                        + cliffBank + motorcycle + roadLane
                        + speedLimit + streetLight + truck
                        + weatherA, data = comp)}
mira_lm_adj <- fit_lm_adj %>% as.mira
# Pool the estimates and show confidence intervals
summary(pool(mira_lm_adj), conf.int = TRUE)
```

Step 2: Model Buildings–Direction 2 V



term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
(Intercept)	-5.898	0.548	-10.764	268114.202	0.000	-6.972	-4.824
crashYear	0.003	0.000	10.765	260433.789	0.000	0.002	0.003
p1	-0.007	0.001	-5.485	378106.297	0.000	-0.010	-0.005
p2	-0.032	0.002	-16.859	442440.011	0.000	-0.036	-0.028
p3	0.030	0.001	32.251	279160.739	0.000	0.028	0.032
bicycle	0.377	0.004	89.215	503979.580	0.000	0.369	0.385
cliffBank	0.066	0.005	12.305	16.072	0.000	0.055	0.077
motorcycle	0.193	0.004	51.461	529468.785	0.000	0.186	0.201
roadLane2-way	0.093	0.003	34.461	1992.333	0.000	0.088	0.098
roadLaneNull	0.009	0.029	0.301	630450.653	0.763	-0.047	0.065
roadLaneOff road	0.129	0.007	19.140	18352.039	0.000	0.116	0.142
speedLimit	0.003	0.000	73.644	294.086	0.000	0.003	0.003
streetLightNull	0.020	0.002	7.905	11774.310	0.000	0.015	0.025
streetLightOff	-0.013	0.003	-4.724	1637.667	0.000	-0.019	-0.008
streetLightOn	0.000	0.003	0.128	11934.699	0.898	-0.005	0.006
truck	-0.052	0.003	-20.437	598064.628	0.000	-0.057	-0.047
weatherAHail or Sleet	-0.125	0.067	-1.876	757904.447	0.061	-0.256	0.006
weatherAHeavy rain	-0.005	0.004	-1.308	704714.055	0.191	-0.012	0.002
weatherALight rain	-0.006	0.002	-3.058	511269.102	0.002	-0.010	-0.002
weatherAMist or Fog	-0.008	0.006	-1.351	765050.781	0.177	-0.021	0.004
weatherANull	-0.213	0.006	-37.874	716221.602	0.000	-0.224	-0.202
weatherASnow	-0.022	0.016	-1.340	762023.366	0.180	-0.054	0.010

Step 2: Model Buildings–Direction 2 VI

After adjustment, the three terms p1, p2 and p3 were still highly significant. The coefficients were slightly smaller than the unadjusted model.

- The term p1 showed that, on top of the general trend before 2011, there was a significant gradient change since then. With 95% confidence, there was a further reduction of somewhere between 0.5 to 1 minor injuries per 100 cases.
- Similarly, the coefficient of p2 showed that there was greater reduction in the gradient between 2014–2016, comparing the Post-first change period. Compared to the general trend before 2014, we have 95% confidence that there was a further reduction of somewhere between 2.8 to 3.2 minor injuries per 100 cases.

The unadjusted and adjusted models gave similar results in terms of estimates and p-values.

Section 6

Discussion

Conclusion

Conclusion drawn from the analysis:

- The probability of fatal or serious cases increased after 2st change. No significant changes seen after the 1nd change of alcohol limit.
- The counts of minor injury per case also dropped after the 2 changes. The reduction was more significant after 2nd change (2014–2016).

Those, however, did not prove the effectiveness of the restricted alcohol limit. We can just tell there were reduced injuries in certain time periods.

Limitations in the Analysis

The followings have may have caused bias in the analysis:

- The Multiple Imputation method was an approximation of the complete data;
- The model should be adjusted by other variables such as the number of cars on road each year;
- The piece-wise linear models exhibited low `R-squared`. Probably LASSO with *1-SE rule* was too harsh on variable selection, or extra variables are needed out of the data set.

Some thoughts of drinking driving. . .

- From the analysis, we are unsure if the changed regulations in 2011/2014 did work;
- Drinking driving is still a serious issue. Restricted laws in alcohol limits did not completely eliminate it.
- Together with campaigns, at least they have raised the self-awareness of the public on the issue. (The key!)

Marketing campaigns of NZTA

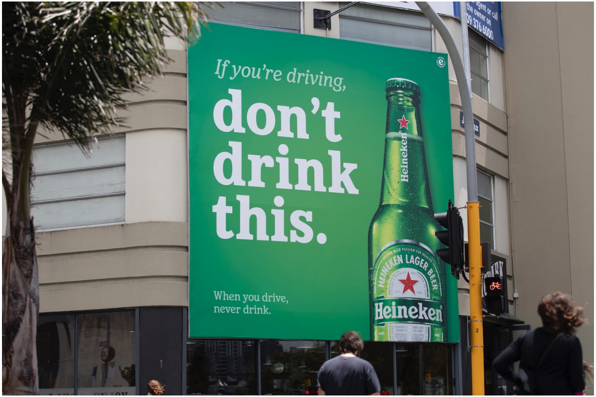
There are 2 pieces of current drinking driving campaigns.

- *Drink-driving campaign: that's a fail*, `www.nzta.govt.nz/safety/what-waka-kotahi-is-doing/marketing-campaigns/current-marketing-campaigns/thats-a-fail/`
- *Drink-driving campaign: doors*, `www.nzta.govt.nz/safety/what-waka-kotahi-is-doing/marketing-campaigns/current-marketing-campaigns/doors/`
- *Standard Drink*, `https://www.newzealandnow.govt.nz/resources/alcohol-and-driving-not-worth-the-risks`

Both of them target male 25–40 years old.

AD campaigns

By Heineken in Dec, 2020.



Source of photo: CampaignBrief `https://campaignbrief.co.nz/2020/12/04/heineken-asks-kiwis-not-to-drink-heineken-in-anti-drink-driving-campaign-via-saatchi-saatchi/`

Extra Information on Drinking Driving...

- Alcohol - Ministry of Health, https://www.health.govt.nz/your-health/healthy-living/addictions/alcohol-and-drug-abuse/alcohol
- Alcohol and drugs limits - NZ Transport Agency, https://www.nzta.govt.nz/roadcode/general-road-code/road-code/about-limits/alcohol-and-drugs-limits/

Reference I

Stef van Buuren.
Flexivle inputation of missing data.
Chapman & Hall/CRC, 2021.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
Regularization paths for generalized linear models via coordinate descent.
Journal of Statistical Software, 33(1), 2010.

R Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2022.

Reference II

Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liskiewicz, and George TH Ellison.
Robust causal inference using directed acyclic graphs: the r package 'dagitty'.
International Journal of Epidemiology, 45(6):1887–1894, 2016.

Stef van Buuren and Karin Groothuis-Oudshoorn.
mice: Multivariate imputation by chained equations in r.
Journal of Statistical Software, 45(3):1–67, 2011.