Crash Data Analysis Report

Sam Oliszewski

Table of Contents

Executive Summary

This analysis studies data that is a compilation of car crash data from New Zealand in 2009. It contains information about the number of crashes involving trucks, bicycles, and motorcycles by hour of day and day of week for the year 2009. Three different datasets were combined for this study and restructured to present the number of crashes by vehicle type for each hour of day and day of week. This analysis was performed on the data compilation, which contains fields for Time, Day, Vehicle, and Crashes. Time is a categorical description of the time of day, with possible values being: early morning, morning, afternoon, and evening. Day reflects either weekday (corresponding to Monday, Tuesday, Wednesday, and Thursday), Friday, Saturday, or Sunday. Vehicle describes the type of vehicle involved in the crash with values truck, bicycle, or motorcycle. Crashes acts as the response variable for this dataset and reflect the number of crashes that occurred.

Three questions were addressed in this study:

1. What day of the week should we expect to observe the most number of crashes?
2. How does the number of crashes change according to vehicle type?
3. How many bicycle crashes should we expect to observe on a Weekday afternoon?

The results from the data analysis showed:

1. The 95% confidence interval of weekday crashes is between 2.16 and 4.56 more crashes than weekend days.
2. Motorcycle involvement increases the number of crashes more than truck or bicycle involvement and the 95% confidence interval is between 10.78 and 13.26 more crashes.
3. The 95% confidence interval for the number of weekday afternoon bicycle crashes is between 162.41 and 238.80.

In summary, non-Friday weekdays and motorcycle involvement are observed to increase the number of crashes.

Analysis

Methods

A negative binomial model was fit with Crashes as the response variable and Day, Hour and Vehicle as the predictors. The necessity of an interaction term was tested and determined to be unnecessary. Using the fitted model, the number of crashes that would occur under various conditions was predicted. Since we are looking to determine the number of crashes as a result of the predictors, and the response variable is a count, Poisson regression is appropriate. There is no evidence of zero-inflation because there are no zero counts in the data, so the fitted model does not need to account for this. However, there is evidence of over dispersion in the data, particularly by time of day, and therefore a negative binomial model would be the most appropriate model to fit. The response variable is a count and since there is evidence of over dispersion standard Poisson regression is not appropriate, and a negative binomial model is preferred. The residuals for the fitted model look like they come from a normal distribution.

This model will be used to answer each of the following questions.

What day of the week should we expect to observe the most number of crashes?

Results

The number of crashes on a day of week in the weekday category (non-Friday weekdays) is with 95% confidence an estimate of 3.36 and between 2.16 and 4.56 crashes higher than the number of crashes on other days (Friday-Sunday). Weekday (non-Friday) crashes have the most impact on the number of crashes when compared to weekend (Friday-Sunday) crashes. It is worthwhile to note that the weekday category is aggregated counts for Monday-Thursday data and may provide somewhat different results than a disaggregated study of the same data. Further, this aggregation means that we expect more crashes to occur on weekdays because it is a four-day total of crashes, rather than a one-day as is analyzed for the remaining days of the week. This means that when we observe slightly larger crash counts for the weekday, we actually can interpret these differences to suggest that weekdays (on a per-day basis) should observe less crashes than weekend days.

How does the number of crashes change according to vehicle type?

Results

With 95% confidence, the number of crashes involving bicycles is an estimate of 5.34 and between 4.09 and 6.59 crashes more than crashes with other transportation types. The number of crashes involving motorcycles is an estimate of 12.02 and between 10.78 and 13.26 crashes more than crashes with other transportation types. The number of crashes involving trucks is an estimate of 5.61 and between 4.36 and 6.86 crashes more than crashes with other transportation types.

How many bicycle crashes should we expect to observe on a Weekday afternoon?

Results

With 95% confidence, we estimate the number of bicycle crashes on weekday afternoon to be 200.61 and between 162.41 and 238.80.

Conclusions

This analysis aimed to answer the following questions:

1. What day of the week should we expect to observe the most number of crashes?
2. How does the number of crashes change according to vehicle type?
3. How many bicycle crashes should we expect to observe on a Weekday afternoon?

The following conclusions were drawn for each respective question:

1. The 95% confidence interval for weekday (non-Friday) crashes is between 2.16 and 4.56 higher than the number of crashes on weekend (Friday-Sunday) days. However, since the weekday category is aggregated total crashes for Monday-Thursday, we would expect to see more crashes on a weekday and therefore, since there are only a few more crashes observed on a weekday, we actually could conclude that weekday driving is safer from these results.
2. Motorcycle involvement increases the number of crashes more than truck or bicycle involvement and the 95% confidence interval is between 10.78 and 13.26 more crashes.
3. The 95% confidence interval for the number of weekday afternoon bicycle crashes is between 162.41 and 238.80.

A more granular analysis could be performed next to determine how these results change if weekdays are disaggregated. Further, the time of day could also be disaggregated and studied as separate hours to determine the time of day at which most crashes occur. Data also could be collected outside of New Zealand to try to expand the inference to a larger population.

Overall, the aggregated day and time model reasonably fit the data and allowed for inferences to be drawn, despite some limitations that exist in terms of who these conclusions can be applied to as a result of aggregation and the narrow scope which the data was collected from (New Zealand drivers only). Further, the aggregation of the non-Friday weekdays is somewhat confusing for interpreting the results of this study because the aggregation makes it look like more crashes occur on weekdays, but if this number were a per-day figure, it would be more clear that the non-Friday weekdays actually observe less crashes than weekend (Friday-Sunday) days. Additional caveats to the conclusions drawn in this study relate to the structure of the data. Since the data only provides crash information by vehicle type and is missing information about all instances of transportation involving the vehicles types studied, we cannot as easily understand the context of the results we obtain; thus making the results difficult to interpret. Additional data collection would help make the results more meaningful.

Appendix

The following code performed the data analysis.

First, the required libraries were loaded.

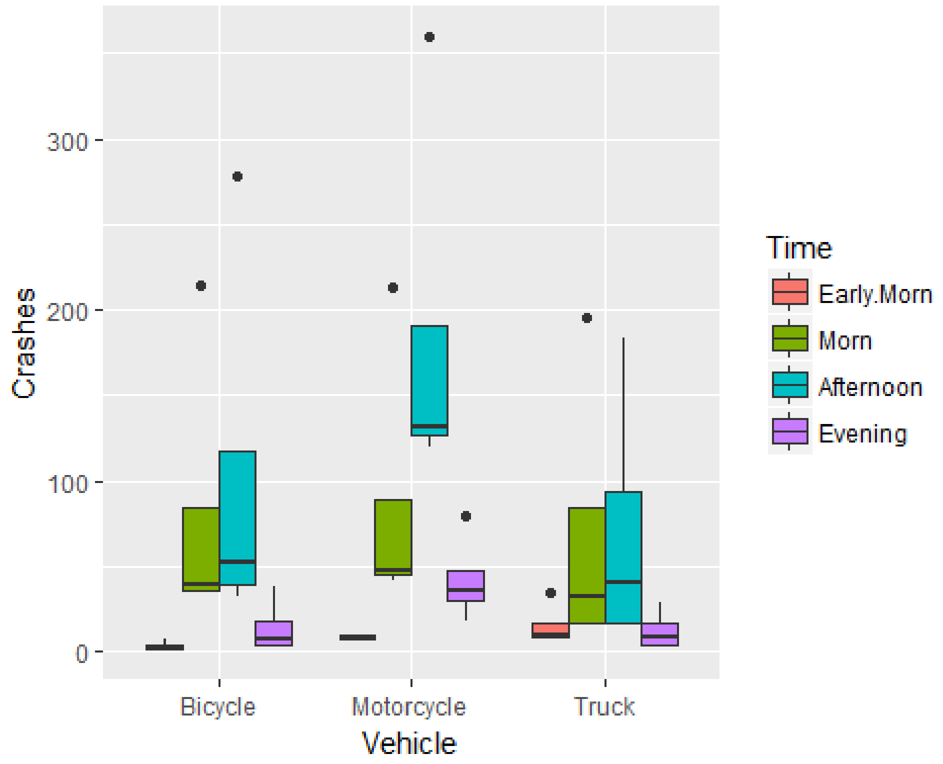
```
library(VGAM)

## Warning: package 'VGAM' was built under R version 3.4.4

## Loading required package: stats4

## Loading required package: splines

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.4.3

## -- Attaching packages ---------------------------------- tidyverse 1.2.1 -
-

## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0

## Warning: package 'ggplot2' was built under R version 3.4.2

## Warning: package 'tibble' was built under R version 3.4.3

## Warning: package 'tidyr' was built under R version 3.4.3
## Warning: package 'readr' was built under R version 3.4.3

## Warning: package 'purrr' was built under R version 3.4.3

## Warning: package 'dplyr' was built under R version 3.4.2

## Warning: package 'stringr' was built under R version 3.4.2

## Warning: package 'forcats' was built under R version 3.4.3

## -- Conflicts ------------------------------------ tidyverse_conflicts() -
-
## x tidyr::fill()   masks VGAM::fill()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(MASS)

## Warning: package 'MASS' was built under R version 3.4.3

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

Next, the data was compiled and tidied.

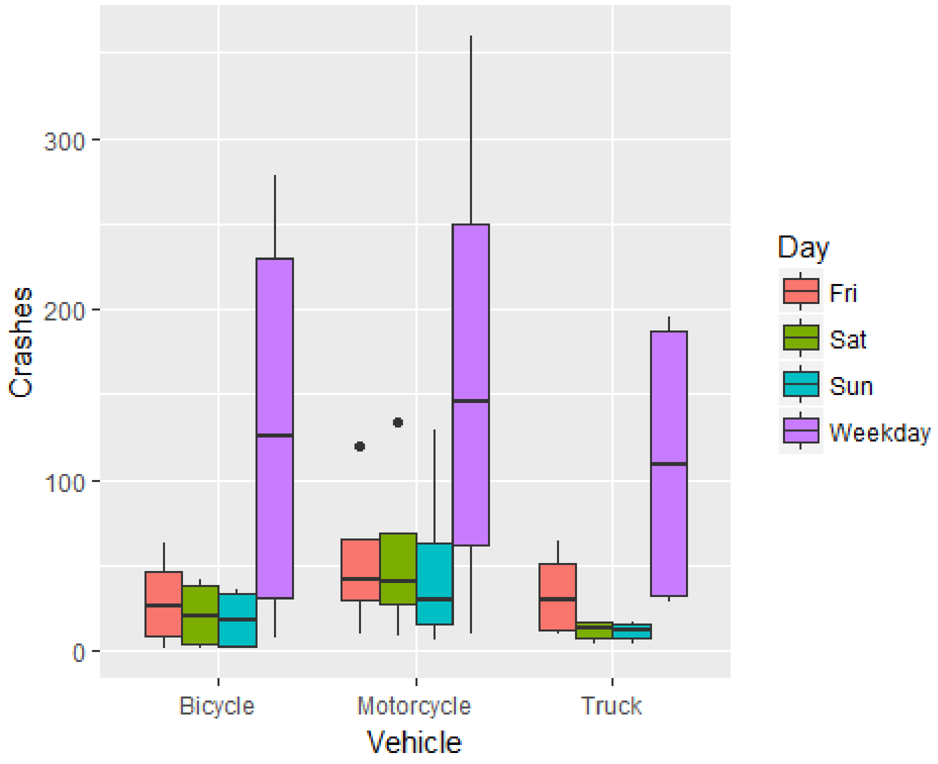
```
data1 <- crashtr %>% mutate(Hour = row_number()-1) %>% gather(-Hour,
key="Day",value="Truck")

## Warning: package 'bindrcpp' was built under R version 3.4.2

data2 <- crashmc %>% mutate(Hour = row_number()-1) %>% gather(-Hour,
key="Day",value="Motorcycle")
data3 <- crashbc %>% mutate(Hour = row_number()-1) %>% gather(-Hour,
key="Day",value="Bicycle")
crashes <- left_join(data1, data2, by=c("Hour","Day")) %>% left_join(data3,
by=c("Hour","Day"))
crashes_time_class_summary <- mutate(crashes,
                                     Time = cut(Hour,
                                                breaks = c(-1, 5.5, 11.5,
18.5, 25),
                                                labels = c("Early.Morn",
"Morn", "Afternoon", "Evening")),
                                     Day = ifelse((Day != "Fri" & Day !=
"Sat" & Day != "Sun"),"Weekday",
                                                  as.character(Day))) %>%
                              group_by(Day, Time) %>%
                              summarize(Truck = sum(Truck), Bicycle =
sum(Bicycle),
                                        Motorcycle = sum(Motorcycle)) %>%
                              gather(-Day, -Time, key="Vehicle",
value="Crashes")

filter(crashes_time_class_summary, Crashes == 0)

## # A tibble: 0 x 4
## # Groups:   Day [0]
## # ... with 4 variables: Day <chr>, Time <fct>, Vehicle <chr>,
## #   Crashes <int>
```

Then, some exploratory plots were created to gather some initial data on the dataset.

```
ggplot(crashes_time_class_summary, aes(Vehicle, Crashes, fill=Time)) +
  geom_boxplot()
```



```
ggplot(crashes_time_class_summary, aes(Vehicle, Crashes, fill=Day)) +
  geom_boxplot()
```



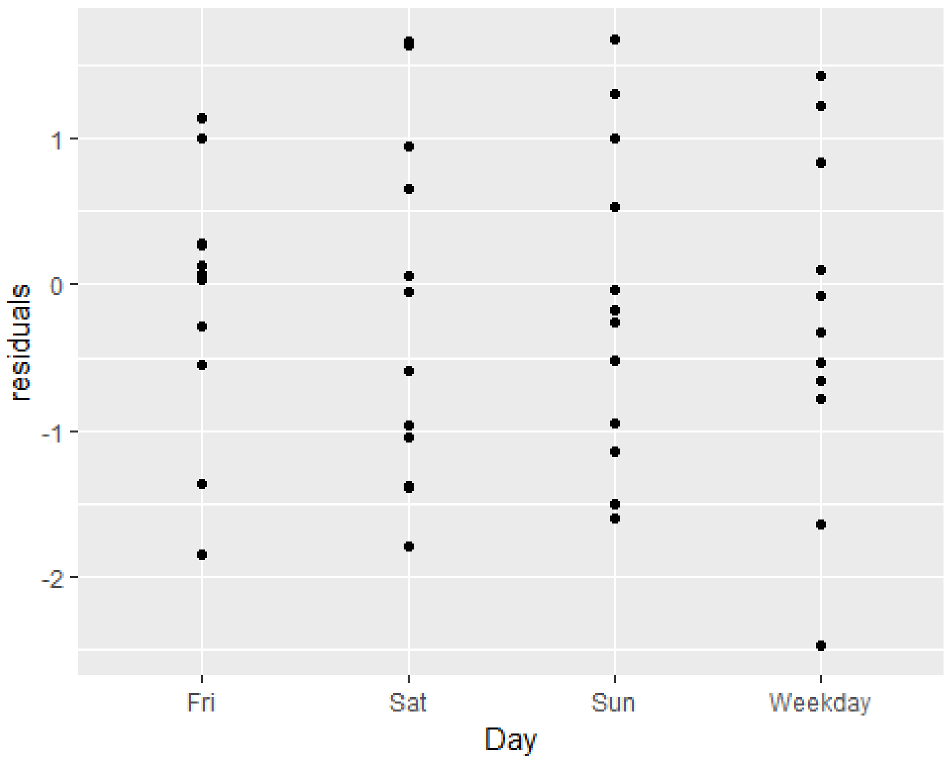
There is evidence of variation between day and time.

A negative binomial model was fit to the data.

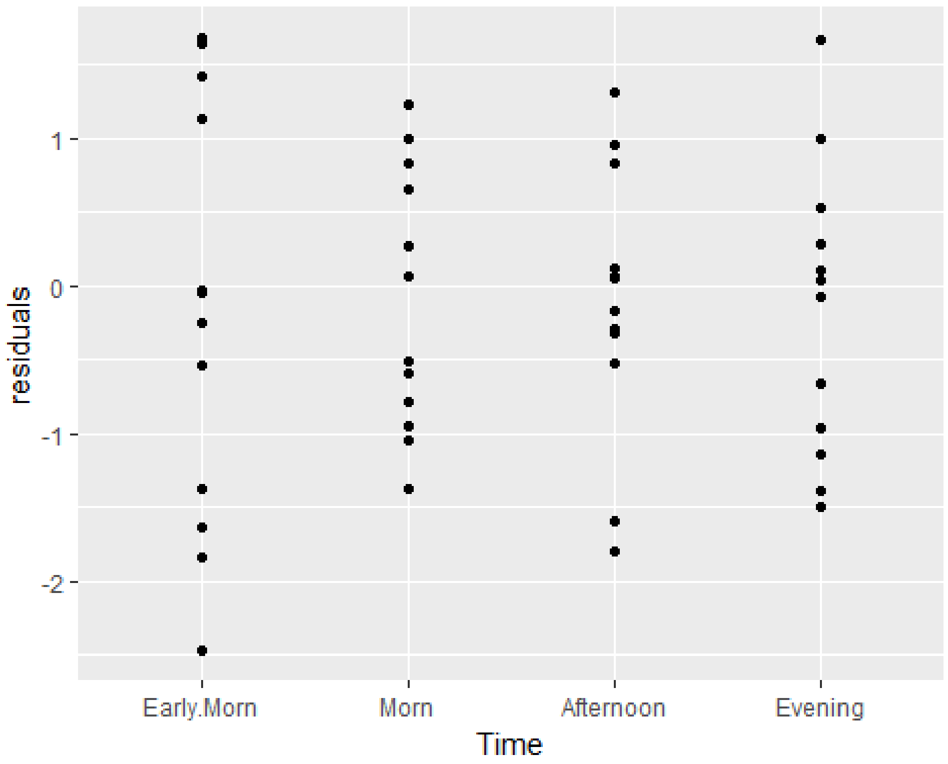
```r
mod <- glm.nb(Crashes ~ Vehicle + Time + Day -1, data =
crashes_time_class_summary)
```

Then, the residuals from the fitted model were assessed.

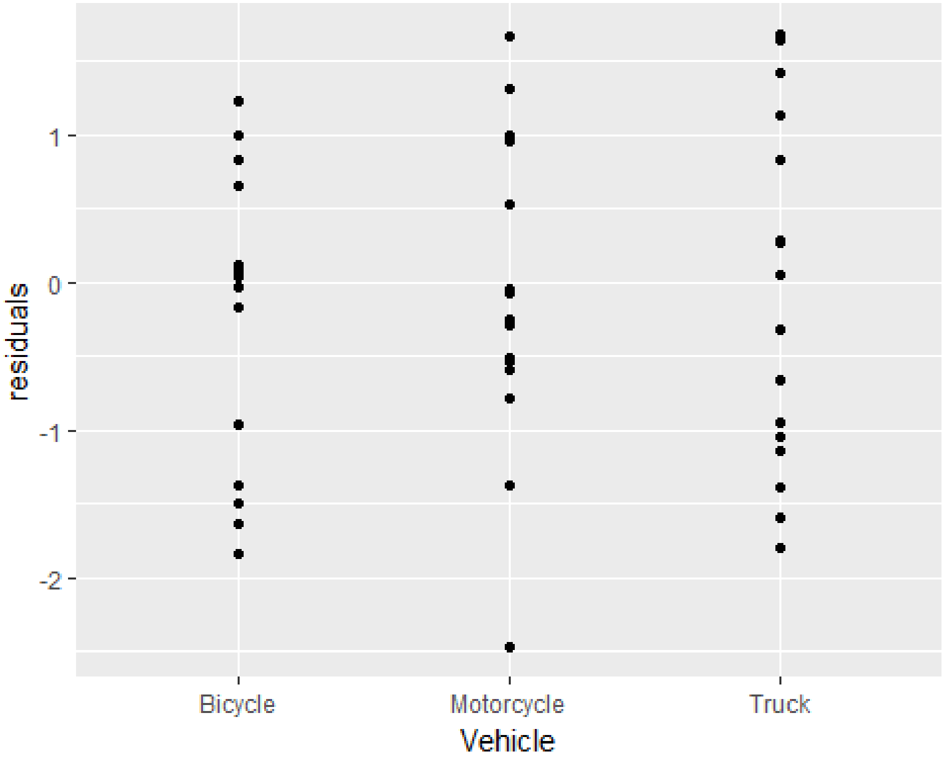
```r
crashes_time_class_summary$residuals <- residuals(mod)
crashes_time_class_summary$fitted.vals <- mod$fitted.values
ggplot(data = crashes_time_class_summary, aes(x = Day, y = residuals)) +
  geom_point()
```



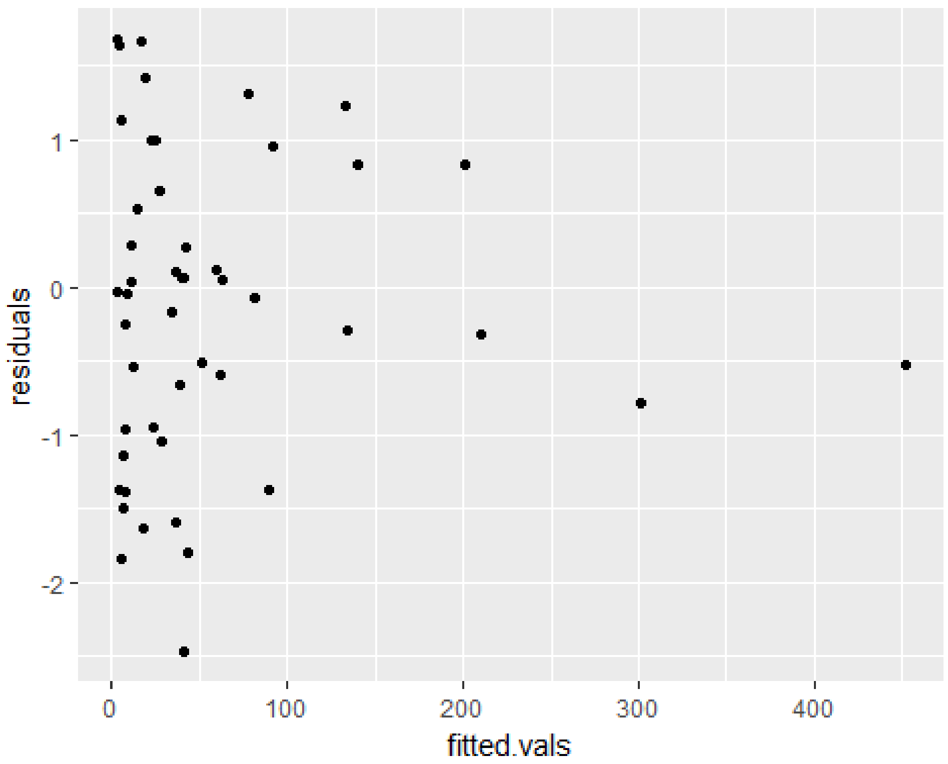
```r
ggplot(data = crashes_time_class_summary, aes(x = Time, y = residuals)) +
  geom_point()
```



```
ggplot(data = crashes_time_class_summary, aes(x = Vehicle, y = residuals)) +
  geom_point()
```



```
ggplot(data = crashes_time_class_summary, aes(x = fitted.vals, y =
residuals)) +
  geom_point()
```



The residuals for the fitted model look like they come from a normal distribution.

Finally, a prediction was made to answer one of the questions of interest.

```
predict.data <- data.frame(c("Weekday"),
                           c("Afternoon"),
                           c("Bicycle"))
colnames(predict.data) <- c("Day","Time","Vehicle")

predictions <- predict(mod, predict.data, type="response", se.fit=TRUE)

to_predict <- mutate(predict.data, Row = row_number())
(prediction_summary <- data.frame(predictions) %>%
    group_by(Row = row_number()) %>%
    summarize(Estimate = round(fit,2), Lower.Bound = round(fit - se.fit,2),
Upper.Bound = round(fit + se.fit,2),
              SE = round(se.fit,2)) %>% left_join(to_predict) %>%
    as.data.frame())

## Joining, by = "Row"

##   Row Estimate Lower.Bound Upper.Bound   SE     Day      Time Vehicle
## 1   1   200.61      162.41       238.8 38.2 Weekday Afternoon Bicycle
```