Analysis of the Crash Data for New Zealand Drivers

Sam Oliszewski

Oregon State University

ABSTRACT

This analysis studies car crash data from New Zealand that contains information about the number of recorded crashes involving trucks, bicycles, and motorcycles grouped by the hour of the day and day of the week of the incident during the year 2009. Three different datasets were examined for this study and restructured to present the number of crashes by vehicle type during different time-of-day intervals for each day of the week. This analysis used the variable Time as a categorical description of the time of day, with the following possible values: early morning, morning, afternoon, and evening. The variable Crashes was used as the response variable for each of these datasets. Three questions were addressed in this study: 1. What day of the week should we expect to observe the greatest 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 Wednesday afternoon? The results from the data analysis showed: 1. The day of the week that one would expect to observe the greatest number of crashes varies by vehicle type and was determined to be Thursday, Friday, and Saturday for bicycles, trucks, and motorcycles, respectively. 2. Since the time of the day that most crashes are expected to occur is the same for all three of the vehicle types, the number of crashes changes according to vehicle type in that each vehicle has different days where one would expect to observe the greatest number of crashes. 3. It was determined that between 60 and 72.52 bicycle crashes are expected to be observed on a Wednesday afternoon.

Keywords: log-linear modeling, negative binomial modeling, New Zealand car crashes

ANALYSIS OF THE CRASH DATA FOR NEW ZEALAND DRIVERS

Three sets of car crash data from New Zealand drivers were examined in this data analysis. Each dataset contained information about the number of crashes involving a given vehicle type (trucks, bicycles, or motorcycles) grouped by the hour of the day and day of the week that the incident occurred during the year 2009. The original datasets were restructured to present the number of crashes for a given vehicle type during different time-of-day intervals for each day of the week. Two explanatory questions and one predictive question were addressed in this study: 1. What day of the week should we expect to observe the greatest 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 Wednesday afternoon? This paper will begin by discussing the exploratory analysis that was performed to provide insight on the data being studied. Next, the approach for model determination is described, in addition to discussion about the limitations of the selected models. Finally, conclusions drawn from the analysis are reported, and caveats of the analysis are offered to aid in the overall interpretation of the study results.

Exploratory Analysis

This section discusses the exploratory analysis performed to investigate the datasets.

Examining the Data

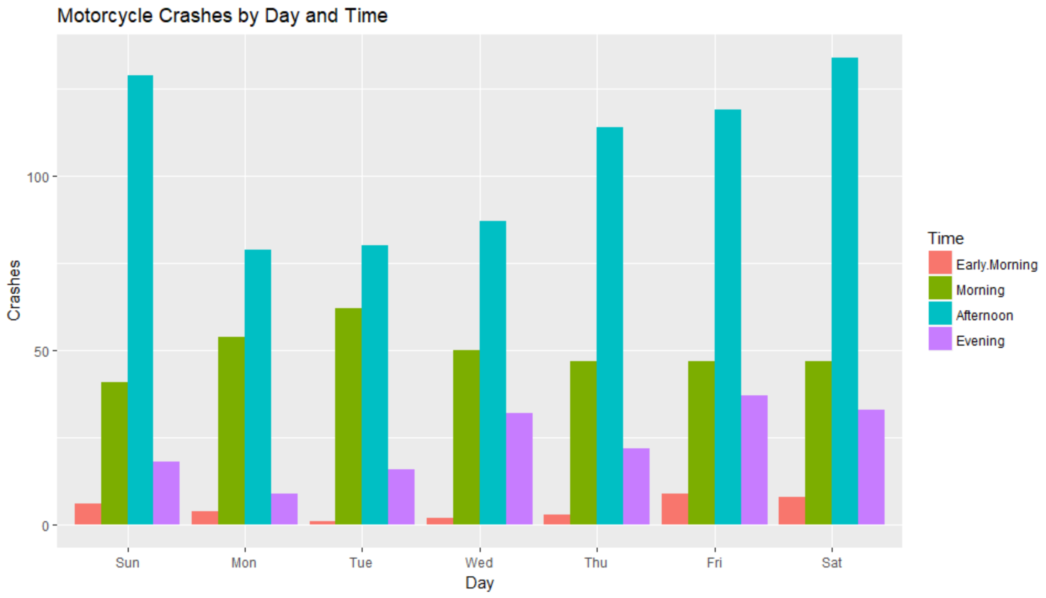
Original Variables. The raw data for this analysis came from three separate datasets in the R library VGAM: crashtr, crashmc, and crashbc. Each dataset was of the same structure and contained variables for each of the seven days of the week and row numbers corresponding to the time of day that crashes were observed. A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices A, B, and C.

Restructuring the Data. To perform this analysis, each of the original datasets was restructured. Each of the restructured datasets contained the following variables: Day, Time, and Crashes. Also of note is that the Time variable was converted into four general time-of-day categories, rather than representing the exact hour of the day. These categories were as follows: Early Morning, Morning, Afternoon, and Evening. The associated time frames for each category are: 12:00 AM-4:59 AM (Early Morning), 5:00 AM-11:59 AM (Morning), 12:00 PM-5:59 PM (Afternoon), and 6:00 PM-11:59 PM (Evening). A sample of the data, as well as descriptions of possible values for each variable, can be referenced in Appendices D, E, and F.

Visualizing Relationships in the Data

Figures of Predictors Against Response. To begin this analysis, it is beneficial to examine plots for each vehicle type depicting how the day of the week and time of the day are related to the number of crashes. These plots are shown below:

` Figure 1. Bicycle Crashes by Day and Time



Day

Figure 2. Motorcycle Crashes by Day and Time

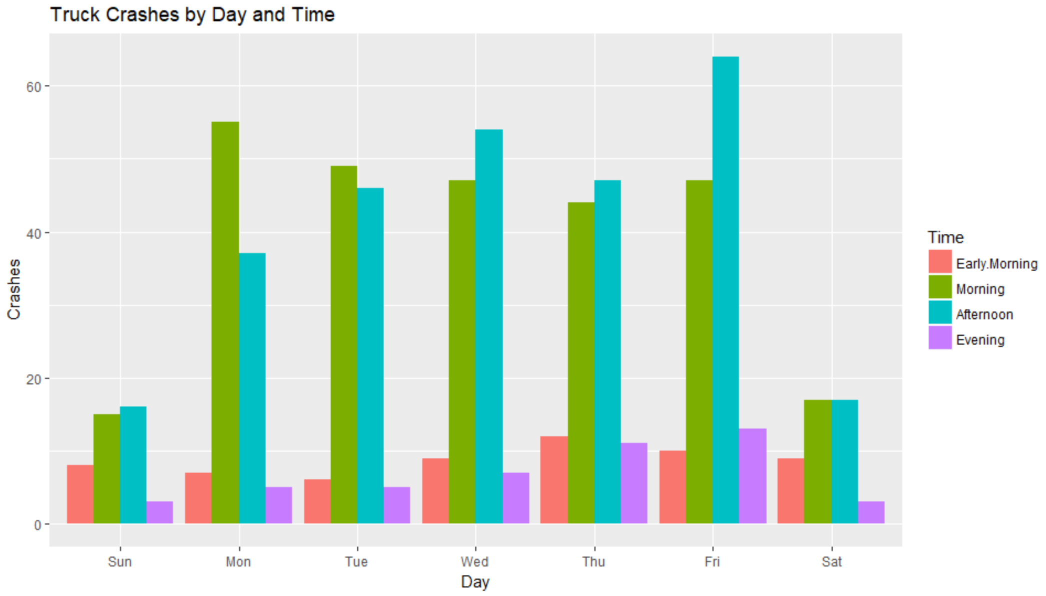


Figure 3. Truck Crashes by Day and Time

Conclusions. After examining these plots, there is some evidence of influence on the number of crashes according to the day of the week and time of the day for each of the vehicle types. For all vehicle types, the most crashes appear to occur in the morning and afternoons. For trucks and bicycles, the most crashes appear to occur during weekdays, whereas for motorcycles the most crashes appear to occur on weekends. This is likely related to the popular types of vehicles used for transportation to and from work as well as vehicles used for commercial use. Additionally, there is no evidence of zero inflation in the data because there is only one occurrence of a zero count; therefore, there is no need to consider zero inflated models when determining the appropriate model to fit the data.

Explanatory Problems

Overview

There are two goals in this section: 1. To determine which day of the week we expect to observe the greatest number of crashes. 2. To determine how the number of crashes changes according to vehicle type. To accomplish these goals, first a log-linear regression model must be constructed for each of the three datasets. These models will help determine the dispersion parameter for the data. If over-dispersion is observed in any of the datasets, a negative binomial model will then be fit to the data. Once the best model has been identified for each of the datasets, a summary of the model will be reported that will identify the difference in the number crashes with respect to the day of the week. This will indicate which day of the week we would expect to observe the greatest number of crashes. The models will also indicate during which time of the day the greatest number of crashes are expected to occur as well. To determine how the number of crashes changes according to vehicle type, a comparison of the day of the week and time of the day where the greatest number of crashes are expected for each vehicle type will be performed.

Methods

Fitting a Log-Linear Model. A log-linear regression model for each vehicle dataset including the predictors Day and Time of the following form was fit:

$$log(Crashes_i) = \beta_0 + \beta_1 DayMon_i + \beta_2 DayTue_i + \beta_3 DayWed_i + \beta_4 DayThu_i + \beta_5 DayFri_i + \beta_6 DaySat_i + \beta_7 TimeMorning_i + \beta_8 TimeAfternoon_i + \beta_9 TimeEvening_i$$

First, the calculation of the dispersion parameter for each model must be calculated. The dispersion parameter for the truck, motorcycle, and bicycle datasets are 1.33, 2.95, and 1.12, respectively. This suggests there is evidence of over dispersion in the motorcycle dataset and a negative binomial model should be fit to this data. For the truck and bicycle datasets, there is no strong evidence of over dispersion according to the dispersion parameter alone. For these datasets, evaluating the residuals for the fitted models would give more insight as to whether the log-linear model is appropriate to fit.

These residual plots are shown below:

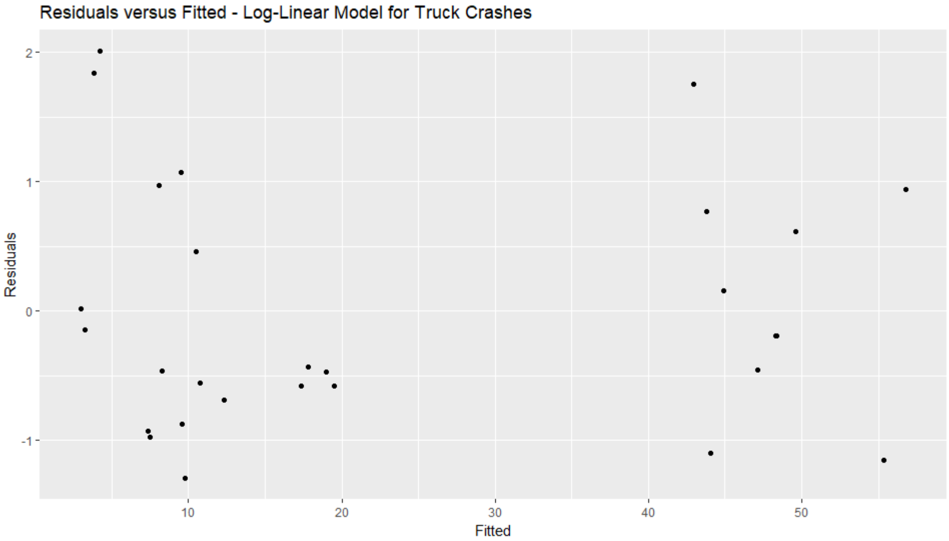


Figure 4. Residuals versus Fitted Truck Data

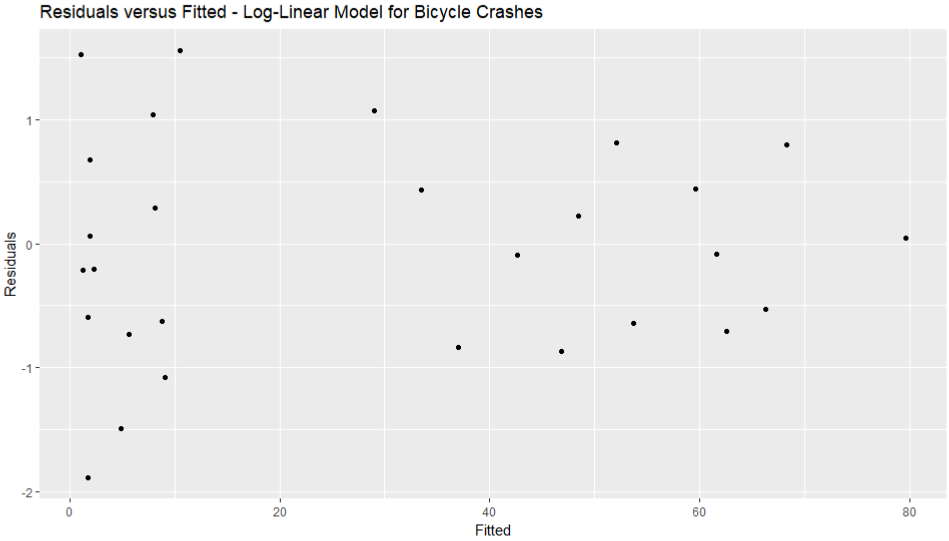


Figure 5. Residuals versus Fitted Bicycle Data

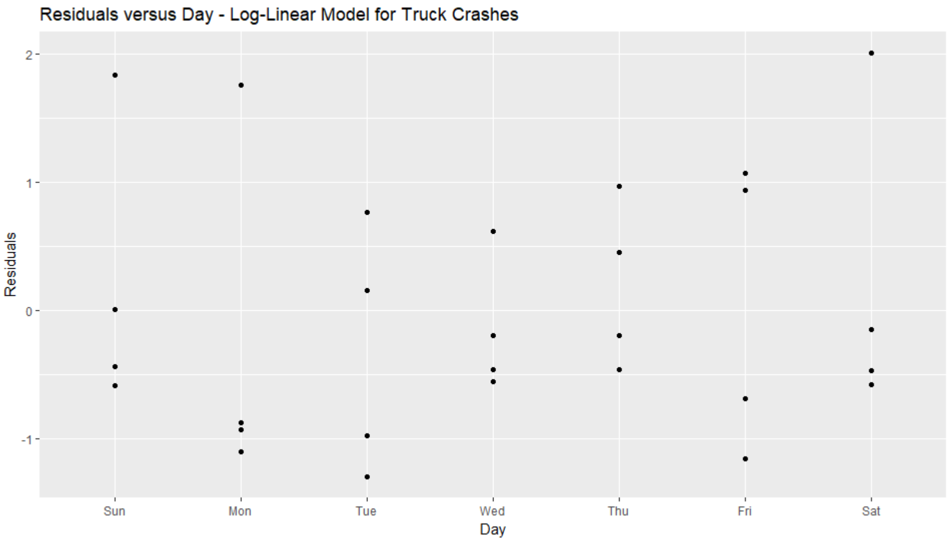


Figure 6. Residuals versus Day Truck Data

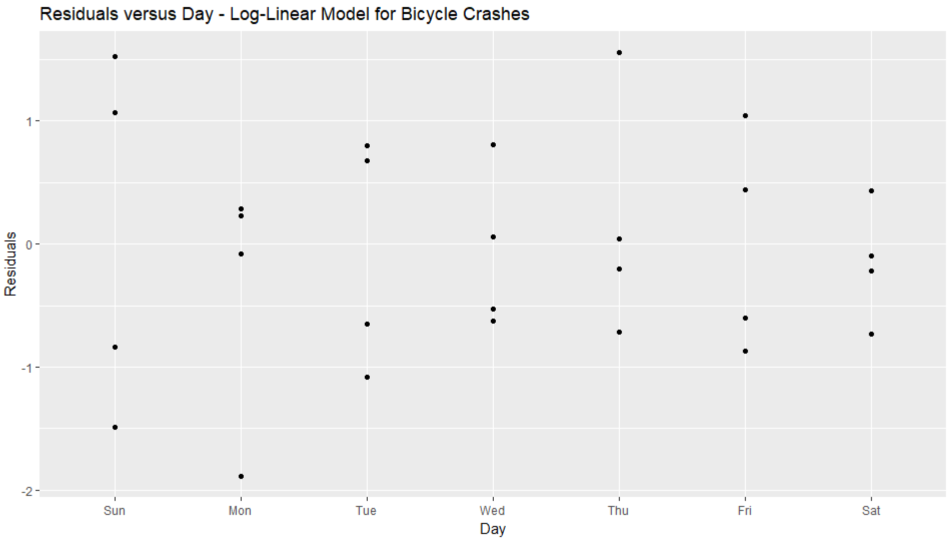


Figure 7. Residuals versus Day Bicycle Data

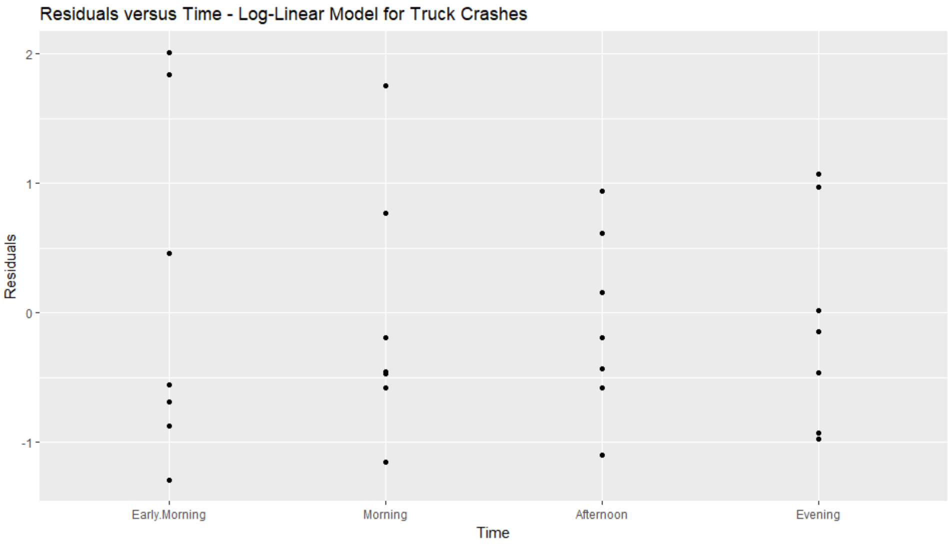


Figure 8. Residuals versus Time Truck Data

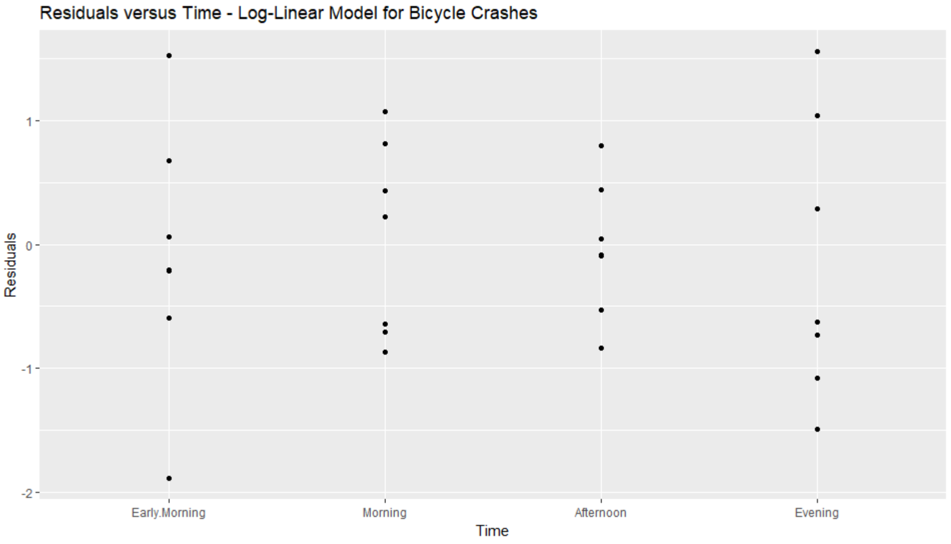


Figure 9. Residuals versus Time Bicycle Data

Based on these residual plots, there is no evidence that the log-linear model is inappropriate for the truck and bicycle datasets. The residuals are all relatively close to zero, ranging mostly between -2 and 2, and do not show any patterns that would suggest they are not normally distributed. Therefore, log-linear models are appropriate to apply to the truck and bicycle datasets.

Fitting a Negative Binomial Model. A negative binomial model for the motorcycle dataset including the predictors Day and Time of the following form was fit:

$$log(Crashes_i) = \beta_0 + \beta_1 DayMon_i + \beta_2 DayTue_i + \beta_3 DayWed_i + \beta_4 DayThu_i + \beta_5 DayFri_i + \beta_6 DaySat_i \ + \ \beta_7 TimeMorning_i + \beta_8 TimeAfternoon_i + \ \beta_9 TimeEvening_i$$

To assess the fit of this model, first a residuals versus fitted plot was generated. This plot is shown below:

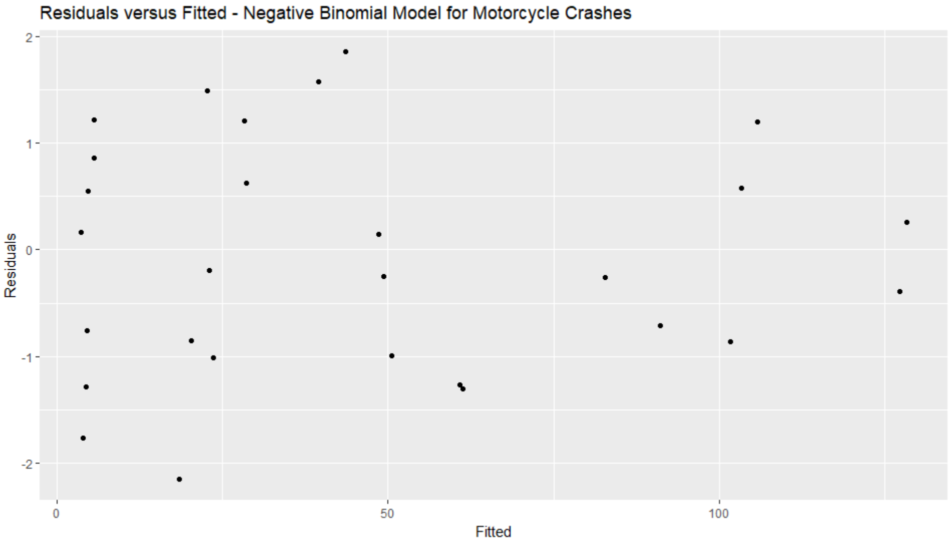


Figure 10. Residuals versus Fitted Motorcycle Data

To further assess whether the negative binomial model is appropriate to use, residuals versus explanatory variable plots for both of the explanatory variables were created. These plots are shown below:

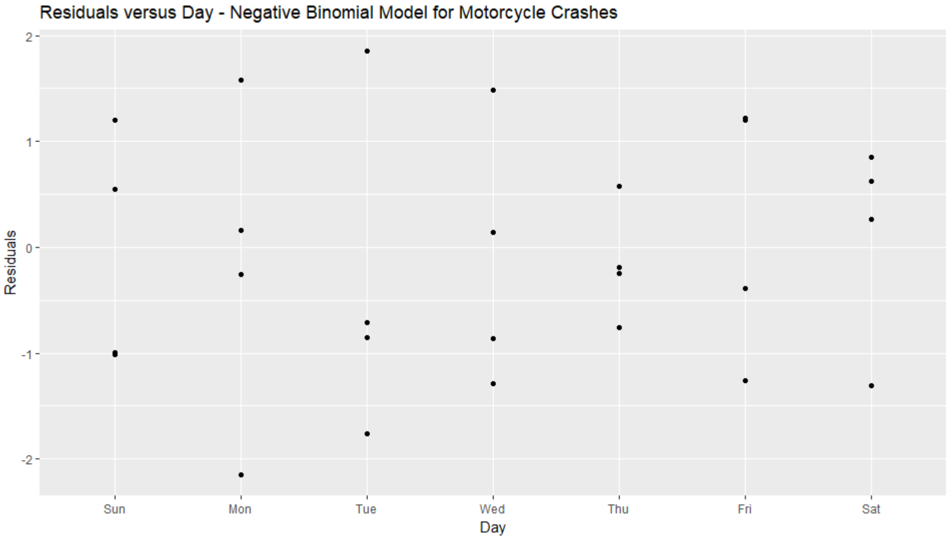


Figure 11. Residuals versus Day Motorcycle Data

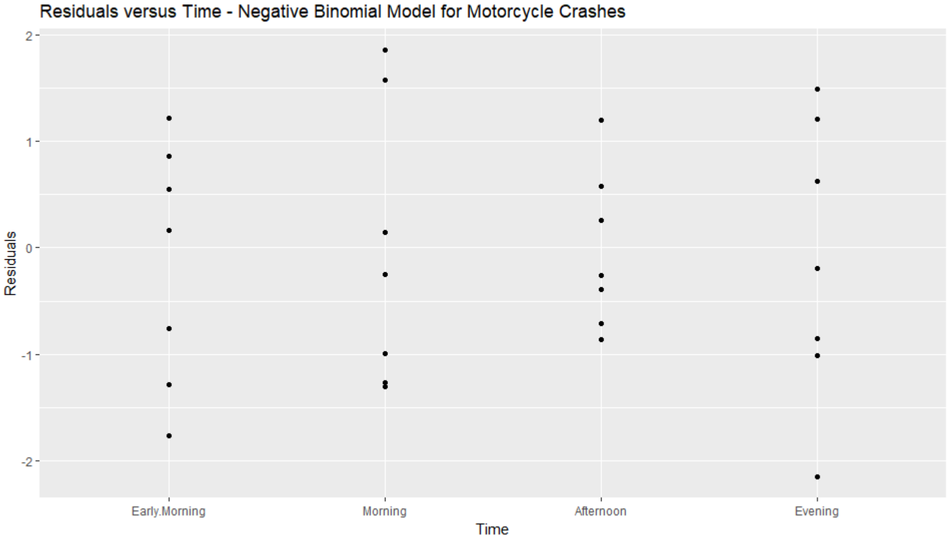


Figure 12. Residuals versus Time Motorcycle Data

Based on these residual plots, there is no evidence the negative binomial model is inappropriate for this data. The residuals are all relatively close to zero, mostly ranging between -2 and 2, and do not exhibit any pattern that would suggest they are not normally distributed. Therefore, the negative binomial model is appropriate to apply to the motorcycle data.

Summarizing the Preferred Models. With the best model for each dataset selected, a summary was generated to determine the following: 1. Which day of the week we expect to observe the greatest number of crashes. 2. How the number of crashes changes according to vehicle type.

The following table summarizes the fitted log-linear model for the truck data:

Term	**Point Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	1.35	0.20	6.87	**6.33e-12**
DayMon	0.91	0.18	4.96	**7.07e-07**
DayTue	0.93	0.18	5.08	**3.82e-07**
DayWed	1.02	0.18	5.70	**1.23e-08**
DayThu	1.00	0.18	5.53	**3.17e-08**
DayFri	1.16	0.18	6.56	**5.36e-11**
DaySat	0.09	0.21	0.43	0.67
TimeMorning	1.50	0.14	10.61	**2e-16**
TimeAfternoon	1.53	0.14	10.81	**2e-16**
TimeEvening	-0.26	0.19	-1.34	0.18

The following table summarizes the fitted log-linear model for the bicycle data:

Term	**Point Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	0.07	0.31	0.22	0.83
DayMon	0.51	0.15	3.43	**0.00**
DayTue	0.61	0.15	4.19	**2.74e-05**
DayWed	0.58	0.15	3.96	**7.37e-05**

DayThu	0.77	0.14	5.38	**7.61e-08**
DayFri	0.48	0.15	3.18	**0.00**
DaySat	0.14	0.16	0.88	0.38
TimeMorning	3.30	0.29	11.23	**2e-16**
TimeAfternoon	3.54	0.29	12.10	**2e-16**
TimeEvening	1.52	0.32	4.78	**1.77e-06**

The following table summarizes the fitted negative binomial model for the motorcycle data:

Term	**Point Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	1.55	0.21	7.35	**1.97e-13**
DayMon	-0.25	0.16	-1.49	0.14
DayTue	-0.15	0.16	-0.92	0.36
DayWed	-0.04	0.16	-0.25	0.80
DayThu	-0.02	0.16	-0.15	0.88
DayFri	0.18	0.16	1.18	0.24
DaySat	0.19	0.16	1.24	0.22
TimeMorning	2.38	0.20	12.01	**2e-16**
TimeAfternoon	3.11	0.19	16.04	**2e-16**
TimeEvening	1.62	0.21	7.85	**4.21e-15**

Using these models, the day of the week that we would expect to observe the greatest number of crashes can be determined by identifying the largest positive coefficient among the day of week terms. Sunday is incorporated into the intercept term and therefore any positive coefficient for other days of the week would indicate that it has a greater impact on the number of crashes than Sunday. For trucks, the day of the week we expect to observe the greatest number of crashes on is Friday, with a point estimate of 1.16, a standard error of 0.18, a z-statistic of 6.56, and a p-value of 5.36e-11. For bicycles, the day of the week we expect to observe the greatest number of crashes on is Thursday, with a point estimate of 0.77, a standard error of 0.14, a z-statistic of 5.38, and a p-value of 7.61e-08. For motorcycles, the day of the week we expect to observe the greatest number of crashes on is Saturday, with a point estimate of 0.19, a standard error of 0.16, a z-statistic of 1.24, and a p-value of 0.22. For all three vehicle types, the time of day when we expect to observe the greatest number of crashes is the afternoon. For trucks, the afternoon time indicator term had a point estimate of 1.53, a standard error of 0.14, a z-statistic of 10.81, and a p-value of 2e-16. For bicycles, the afternoon time indicator term had a point estimate of 3.54, a standard error of 0.29, a z-statistic of 12.10, and a p-value of 2e-16. For motorcycles, the afternoon time indicator term had a point estimate of 3.11, a standard error of 0.19, a z-statistic of 16.04, and a p-value of 2e-16. Based on these results, one can compare how the number of crashes changes according to vehicle type by comparing how day and time influence the number of crashes for each vehicle type.

Conclusion

Based on the raw output discussed above, several conclusions can be made and the two explanatory questions this analysis studied can be answered.

Identifying the Day When the Most Crashes Are Expected to Occur. Based on the summary of the fitted models for the truck, bicycle, and motorcycle data, the day that we expect the greatest number of crashes to occur differs for each vehicle type. For trucks, Friday is when we expect to observe the greatest number of crashes and with 95% confidence this day of the week is expected to increase the number of crashes by between 2.66 and 3.82 crashes. For bicycles, Thursday is when we expect to observe the greatest number of crashes and with 95% confidence this day of the week is expected to increase the number of crashes by between 1.88 and 2.48 crashes. For motorcycles, Saturday is when we expect to observe the greatest number of crashes and with 95% confidence this day of the week is expected to increase the number of crashes by between 1.03 and 1.42 crashes.

Identifying How the Number of Crashes Changes by Vehicle Type. Based on the summary of the fitted models for the truck, bicycle, and motorcycle data, in addition to the discussion above regarding the influence on the number of crashes according to the day of the week, the crash behavior for each vehicle type can be compared. As noted previously, the day of the week where the greatest number of crashes is expected to occur varies for each vehicle type. Therefore, this is the first notable observation of how the number of crashes changes by vehicle type. Secondly, for all three vehicle types, the time of day when the greatest number of crashes is expected to occur is the same— the afternoon. This indicates that crash behavior for each vehicle type is similar based on the time of day, but each vehicle type has a different day of the week where more crashes are expected and may indicate that the odds of a crash are higher.

Limitations. Of note on the conclusions from this study is that there is no information in the original data that indicates the number of vehicles on the road for each vehicle type. Therefore, it is difficult to interpret the results of this analysis to determine whether the odds of a crash are more likely at a certain time of day or day of the week. All that can be determined is how the number of crashes is influenced by these predictors. Without context for these values, the results of this study have questionable practical significance.

Prediction Problem

Overview

The goal in this section is to use a model to predict the number of crashes that we expect to observe on a given day of the week and time of day. The objective is to use the predictive model to determine how many bicycle crashes would occur on a Wednesday afternoon. A summary of the prediction results will be reported.

Methods

Fitting a Log-Linear Model. In the previous section, a log-linear regression model for the bicycle dataset including the predictors Day and Time of the following form was fit:

$$log(Crashes_i) = \beta_0 + \beta_1 DayMon_i + \beta_2 DayTue_i + \beta_3 DayWed_i + \beta_4 DayThu_i +$$

$$\beta_5 DayFri_i + \beta_6 DaySat_i \; + \; \beta_7 TimeMorning_i +$$

$$\beta_8 TimeAfternoon_i + \; \beta_9 TimeEvening_i$$

Recall, that the following table summarizes the fitted log-linear model for the bicycle data:

Term	**Point Estimate**	**Standard Error**	**z-value**	**p-value**
Intercept	0.07	0.31	0.22	0.83
DayMon	0.51	0.15	3.43	**0.00**
DayTue	0.61	0.15	4.19	**2.74e-05**
DayWed	0.58	0.15	3.96	**7.37e-05**
DayThu	0.77	0.14	5.38	**7.61e-08**
DayFri	0.48	0.15	3.18	**0.00**
DaySat	0.14	0.16	0.88	0.38
TimeMorning	3.30	0.29	11.23	**2e-16**
TimeAfternoon	3.54	0.29	12.10	**2e-16**
TimeEvening	1.52	0.32	4.78	**1.77e-06**

Using this log-linear model, we can predict the number of crashes we would expect to occur on a given day of the week and time of day. For the prediction interval, we simply increase the level of error in the confidence interval. Using this model, the number of crashes on a Wednesday afternoon has a point estimate of 66.26, with a standard error of 6.26.

Conclusion

With 95% confidence, the number of bicycle crashes we expect to observe on a Wednesday afternoon is between 60 and 72.52. Referring to the initial plot of the number of crashes according to day and time from the exploratory analysis, this value is reasonable for the given day and time for bicycles.

Limitations. This model was able to determine the number of bicycle crashes one would expect to observe on a given day and time, but there are caveats to this model's predictive power. For example, all of the predictors in this model are indicator terms. Therefore, there are only a finite number of combinations possible and thus this model isn't particularly applicable to other settings. If other predictors were introduced into the model, there would be more room for honest prediction to occur, but this model simply can only evaluate what the expected number of crashes would be for a given day and time for bicycles in New Zealand in 2009. This does not give a whole lot of insight into what future values may be and thus has limited practical significance.

Conclusion

This analysis aimed to answer three questions of interest: 1. What day of the week should we expect to observe the greatest 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 Wednesday afternoon? Ultimately, it was determined that the day of the week that one would expect to observe the greatest number of crashes varies by vehicle type and was determined to be Thursday, Friday, and Saturday for bicycles, trucks, and motorcycles, respectively. Additionally, it was determined that the time of day that most crashes are expected to occur is the same for all three of the vehicle types— the afternoon. Therefore, the number of crashes changes according to vehicle type in that each vehicle has different days where one would expect to observe the greatest number of crashes. Finally, it was determined that between 60 and 72.52 bicycle crashes are expected to be observed on a Wednesday afternoon. To improve the value of this analysis, further research should be performed to determine the number of vehicles on the road according to the time of day and day of the week so that the number of crashes can be put into proper context. Additionally, data should be collected over multiple years to determine if there is trend, cycles, or seasonality in the data. Lastly, it would be beneficial to study how the number of crashes by day and time changes at different points of the year. This could be studied by recording the number of crashes by time for each day of the year and indicating what day of the week the date fell on. This additional research would provide a lot of meaningful insight into the crash behavior in New Zealand.

APPENDIX A

Dataset: crashtr (VGAM)
Description: This dataset from the VGAM library in R contains the number of crashes by day of the week and time of the day involving trucks in New Zealand during 2009.

Sample Data:

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	0	1	1	1	0	2	1
1	1	1	1	0	2	0	2
2	0	3	2	1	1	3	2
3	0	0	2	1	1	0	3
4	3	0	3	4	2	3	0

Description of Variables:

Variable	Description	Possible Values
{Row Num}	Corresponds to the time of day the crash occurred (0 = 12:00 AM).	Integer from 0-23
Mon	The number of crashes involving trucks on Monday.	Any integer
Tue	The number of crashes involving trucks on Tuesday.	Any integer
Wed	The number of crashes involving trucks on Wednesday.	Any integer
Thu	The number of crashes involving trucks on Thursday.	Any integer
Fri	The number of crashes involving trucks on Friday.	Any integer
Sat	The number of crashes involving trucks on Saturday.	Any integer
Sun	The number of crashes involving trucks on Sunday.	Any integer

APPENDIX B

Dataset: crashmc (VGAM)
Description: This dataset from the VGAM library in R contains the number of crashes by day of the week and time of the day involving motorcycles in New Zealand during 2009.

Sample Data:

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	1	0	0	0	5	1	1
1	0	1	0	1	1	3	1
2	0	0	1	2	1	3	1
3	2	0	0	0	0	0	1
4	0	0	0	0	0	0	0

Description of Variables:

Variable	Description	Possible Values
{Row Num}	Corresponds to the time of day the crash occurred (0 = 12:00 AM).	Integer from 0-23
Mon	The number of crashes involving motorcycles on Monday.	Any integer
Tue	The number of crashes involving motorcycles on Tuesday.	Any integer
Wed	The number of crashes involving motorcycles on Wednesday.	Any integer
Thu	The number of crashes involving motorcycles on Thursday.	Any integer
Fri	The number of crashes involving motorcycles on Friday.	Any integer
Sat	The number of crashes involving motorcycles on Saturday.	Any integer
Sun	The number of crashes involving motorcycles on Sunday.	Any integer

APPENDIX C

Dataset: crashbc (VGAM)

Description: This dataset from the VGAM library in R contains the number of crashes by day of the week and time of the day involving bicycles in New Zealand during 2009.

Sample Data:

	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	0	0	0	0	0	1	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	2
3	0	0	0	0	0	0	1
4	0	0	0	1	0	0	0

Description of Variables:

Variable	Description	Possible Values
{Row Num}	Corresponds to the time of day the crash occurred (0 = 12:00 AM).	Integer from 0-23
Mon	The number of crashes involving bicycles on Monday.	Any integer
Tue	The number of crashes involving bicycles on Tuesday.	Any integer
Wed	The number of crashes involving bicycles on Wednesday.	Any integer
Thu	The number of crashes involving bicycles on Thursday.	Any integer
Fri	The number of crashes involving bicycles on Friday.	Any integer
Sat	The number of crashes involving bicycles on Saturday.	Any integer
Sun	The number of crashes involving bicycles on Sunday.	Any integer

APPENDIX D

Description: This dataset restructures the crashtr dataset from the VGAM library in R and contains the number of crashes by day of the week and time of the day involving trucks in New Zealand during 2009.

Sample Data:

Day	Time	Crashes
Fri	Early.Morning	10
Fri	Morning	47
Fri	Afternoon	64
Fri	Evening	13
Mon	Early.Morning	7

Description of Variables:

Day: Corresponds to the day of the week the crashes occurred.
Possible Values: Text

- Mon → Monday
- Tue → Tuesday
- Wed → Wednesday
- Thu → Thursday
- Fri → Friday
- Sat → Saturday
- Sun → Sunday

Time: Corresponds to the time of day the crashes occurred.
Possible Values: Text

- Early.Morning— Time between 12:00 AM and 4:59 AM
- Morning— Time between 5:00 AM and 11:59 AM
- Afternoon— Time between 12:00 PM and 5:59 PM
- Evening— Time between 6:00 PM and 11:59 PM

Crashes: The number of crashes that occurred.
Possible Values: Any integer greater than or equal to 0.

APPENDIX E

Description: This dataset restructures the crashmc dataset from the VGAM library in R and contains the number of crashes by day of the week and time of the day involving motorcycles in New Zealand during 2009.

Sample Data:

Day	Time	Crashes
Fri	Early.Morning	9
Fri	Morning	47
Fri	Afternoon	119
Fri	Evening	37
Mon	Early.Morning	4

Description of Variables:

Day: Corresponds to the day of the week the crashes occurred.
Possible Values: Text

- Mon → Monday
- Tue → Tuesday
- Wed → Wednesday
- Thu → Thursday
- Fri → Friday
- Sat → Saturday
- Sun → Sunday

Time: Corresponds to the time of day the crashes occurred.
Possible Values: Text

- Early.Morning— Time between 12:00 AM and 4:59 AM
- Morning— Time between 5:00 AM and 11:59 AM
- Afternoon— Time between 12:00 PM and 5:59 PM
- Evening— Time between 6:00 PM and 11:59 PM

Crashes: The number of crashes that occurred.
Possible Values: Any integer greater than or equal to 0.

APPENDIX F

Description: This dataset restructures the crashtr dataset from the VGAM library in R and contains the number of crashes by day of the week and time of the day involving bicycles in New Zealand during 2009.

Sample Data:

Day	Time	Crashes
Fri	Early.Morning	1
Fri	Morning	41
Fri	Afternoon	63
Fri	Evening	11
Mon	Early.Morning	0

Description of Variables:

Day: Corresponds to the day of the week the crashes occurred.
Possible Values: Text

- Mon → Monday
- Tue → Tuesday
- Wed → Wednesday
- Thu → Thursday
- Fri → Friday
- Sat → Saturday
- Sun → Sunday

Time: Corresponds to the time of day the crashes occurred.
Possible Values: Text

- Early.Morning— Time between 12:00 AM and 4:59 AM
- Morning— Time between 5:00 AM and 11:59 AM
- Afternoon— Time between 12:00 PM and 5:59 PM
- Evening— Time between 6:00 PM and 11:59 PM

Crashes: The number of crashes that occurred.
Possible Values: Any integer greater than or equal to 0.