11/3/2019

Abhisek Gautam [13679042]

Analysis of Road Accidents’ Severity using Multilevel Modeling

# Introduction

Road accidents may cause deaths and it may leave a lot of people seriously injured. In 2016, 5.34 road deaths occurred in Australia per 100,000 people (Bureau of Infrastructure, Transport and Regional Economics 2018). The number of vehicles registered each year in Australia is increasing with more than 1.5% per annum, inline with a similar increase in the population (Australian Bureau of Statistics 2019, Australian Bureau of Statistics 2018). There is thus an increasing call for road safety by the government, which can be noticed with the Towards Zero campaign setup by the Victorian government. The dataset that was analysed was collected also collected from a government website; from the Department of Planning, Transport and Infrastructure. Previously, our team had done a binomial logistic regression on the dataset to find out what were the causes of high severity and fatalities in road accidents of Victorian Roads. The team had concluded that type of accident, light condition, type of road and type of vehicle during the time of the accident had the most effect on the severity of road accidents.

Low p (hence, statistically significant) values were observed on the vehicle’s part, such as the age of the vehicle and the type of vehicle (like buses, vans, taxis and cars). This study dives deeper into the vehicle aspect of road accidents to answer the question of whether vehicle’s make and model have an impact on the severity of road accidents or not.

# Details on the Previous Study

The previous study included a detailed exploratory data analysis (EDA) for the road accident data, which was used to choose the independent variables. The final result of the analysis is shown in Appendix 1. It shows that the type of vehicle is significant in determining the severity of the accident. However, there may be an underlying effect within, where the make and the model of the vehicles would be causing an impact on the accidents’ severity. Such a structure is not considered by the previous model and without looking into more details, it cannot be assumed that Vehicle Type, in general, is significant. A simple multilevel model would be able to factor the underlying variables.

A screenshot of a cell phone

Description automatically generated

*Figure 1: A snapshot of previous analysis showing Vehicle Type variable and its significance*

It is also worthwhile to find out how predictors on the different levels account in increasing or decreasing the severity of accidents. The group predictors (such as the age of the vehicle, damage done to a vehicle, its speed) can be conjugate with different hierarchies, and their effects can be analysed separately using a multilevel model. Using multilevel fixed effects model, inferences can be derived beyond a group. Using multilevel random effects model, the effect of both types of group level predictor variables and group level dummy variables can be estimated (Magnusson, 2015).

# Structure of the dataset

The total dataset consists of 28 Vehicle types, from light such as bicycles to heavy ones. With the aim of making this study more specific, it will factor out two wheelers and heavy wheelers and will only take everyday vehicles from the small category of vehicles only, viz. car, station wagon, utility and taxi only.

A close up of a device

Description automatically generated

*Figure 2: A bar graph showing the count of individual vehicle types in the dataset*

The data-frame used in the analysis is as follows:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| ACCIDENT\_NO | The Unique ID of the accident |
| ACCIDENTDATE | The date of the accident |
| VEHICLE\_MAKE | The Make of the vehicle |
| VEHICLE\_MODEL | The Model of the vehicle |
| Vehicle\_Type\_Desc | Description of the Type of the Vehicle (Car, Station Wagon) |
| VEHICLE\_YEAR\_MANUF | The year of Manufacture of the Vehicle |
| LEVEL\_OF\_DAMAGE | Damage Level: 1-6, 1 includes severely damaged |
| SPEED\_ZONE | The speed zone of the road which the accident occurred (This is assumed to be the speed of vehicle during crash and is taken as continuous- as speed of the vehicle is not always equal to speed zone’s speed) |
| ACCIDENT\_YEAR | Year of the accident, derived from ACCIDENTDATE (continuous) |
| NUM\_OF\_YEARS | Age of Car, obtained from ACCIDENT\_YEAR – VEHICLE\_YEAR\_MANUF |
| SEVERE | Is the accident severe? Binary: Yes and No |

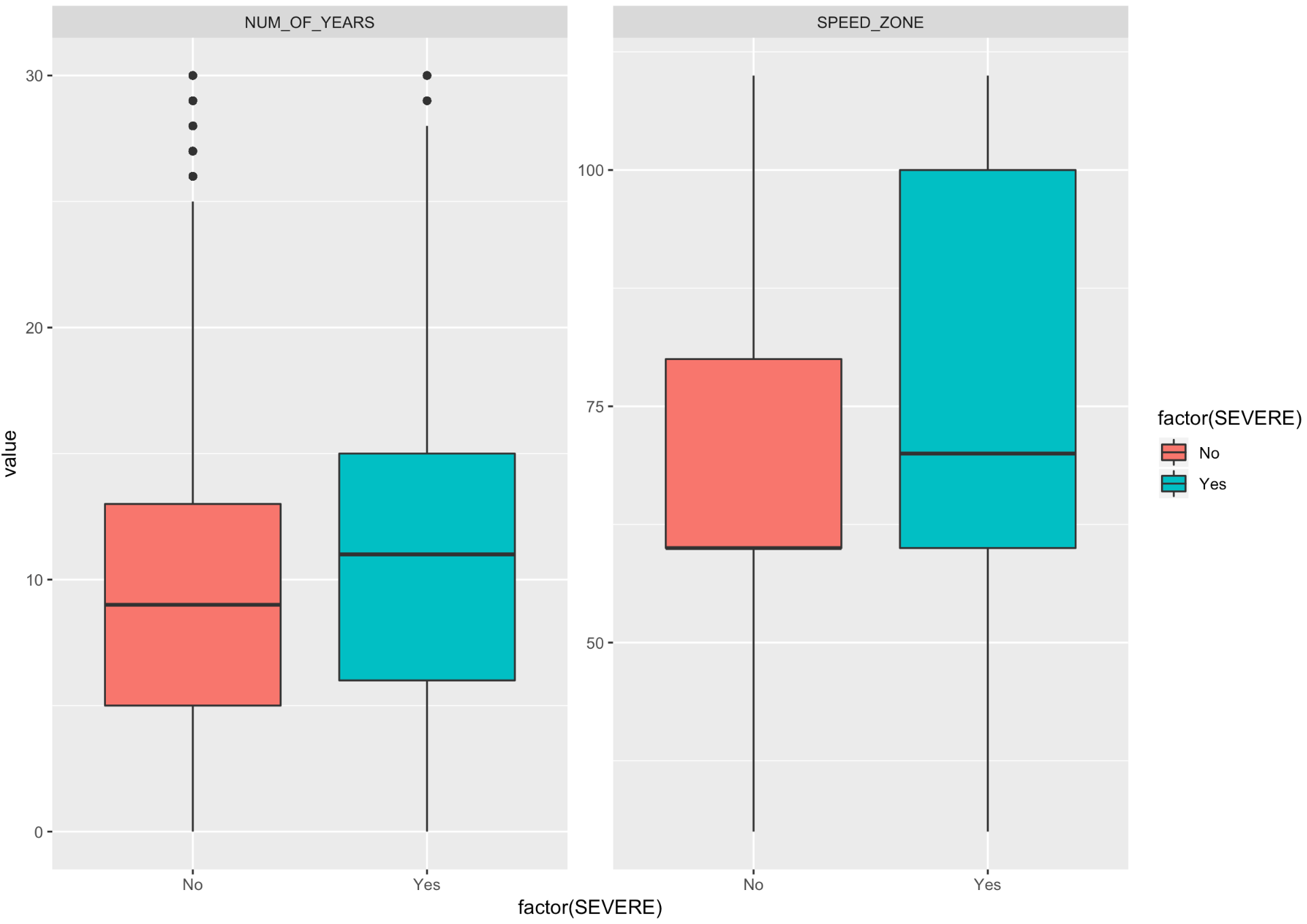
Now, checking for collinearity between the continuous variables NUM\_OF\_YEARS and SPEED\_ZONE, we see that there it is minimal. Hence, both of them can be used in our model.

A close up of text on a white background

Description automatically generated

*Figure 3: Checking collinearity between continuous residual variables*

Looking at the distribution of severity on continuous variables (Figure 4), Number of Years (i.e. the age of car) and Speed Zone, we find that there is not much impact of the age of car on the severity of accidents, whereas with regards to the speed zone, accidents are more severe at higher speed zone areas.



*Figure 4: A box plot showing the distribution of continuous variables*

Second test is to check for the ratios of the response variable, SEVERE. From Figure 5, it can be seen that there are no zeros, and also some of the ratios are large which may be due to the dataset being from the year 2006 to 2018. We will need to be careful with this during modeling our data.

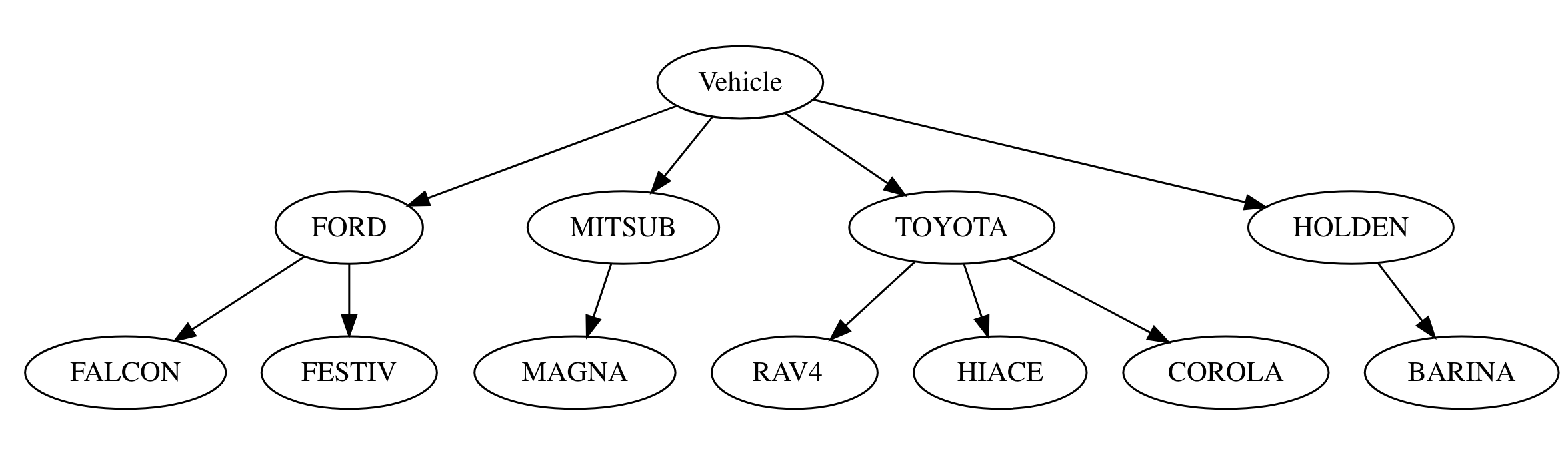
A close up of text on a white background

Description automatically generated

*Figure 5: A table showing the number and percentage of Response variable*

Now, regarding the levels in the data, a vehicle’s make and Vehicle’s model is a sample of wider population of model and make respectively. Hence, they can be used as levels for multilevel analysis. It is useful to compare between higher level units; hence we will be comparing the vehicle’s makes.

The data has a 2-level nested structure hierarchy. The lowest level (Level 2) is vehicle’s model, and highest level (Level 1) is the vehicle’s make, as shown in Figure 6.



Vehicle Model

Vehicle Make

*Figure 6: A plot and a relationship diagram showing 2-level hierarchy in the dataset*

Thus, we will be taking the VEHICLE\_MAKE and VEHICLE\_MODEL as our two levels for multilevel modeling. There are no predictor variables for VEHICLE\_MAKE in this dataset. For VEHICLE\_MODEL level, will take LEVEL\_OF\_DAMAGE, SPEED\_ZONE and NUM\_OF\_YEARS as predictor variables.

The response variable, SEVERE is binary, having values ‘No’ and ‘Yes’. Hence, from the nature of the data and the response variable, the type of regression that we will be using is a logistic multilevel regression to access whether the vehicle’s attributes play a significant impact on the severity of accidents or not.

# Multilevel Model Design and Interpretation

Multilevel model treats units of analysis as independent observations and helps in drawing conclusions about the effects in those observations individually. The motivation to take up multilevel modeling can be due to the wish to make amendments to previously made inferences, or an interest in group effects of a research, or even to estimate the group effects using the group level predictors together. Whatever the motivation might be, there are a series of steps that are performed while doing multilevel regression. We will be performing each of those steps, taking inferences and moving onto the next until the best combination of variables is found.

We will be using `glmer` function of `lme4` package in R to compute the multi-level models with the optimizer `bobyqa` and value of nAGQ as 0, for optimizing the speed of the modeling. To compare the models, we will use the function AIC which stands for ‘An Information Criterion’ with log likelihood of the model. Lower value of AIC means a better model.

Now, the first step of doing multilevel modeling is to create a null model, with no effects, just the intercepts. A null model can be taken as the base model for analysis, and it will help to determine whether controlling the factor is useful or not.

From the data, we have more ‘No’ and less ‘Yes’. The lower coded group is ‘No’ in this analysis.

A screenshot of a cell phone

Description automatically generated

*Figure 7: A null model with only the response variable SEVERE*

After the null model, a random intercept model is made using `glmer` using random intercept of VEHICLE\_MAKE. From this model, we can see that the intercept’s estimate has changed towards the ‘No’ category. The variance due to random effect VEHICLE\_MAKE is very small, so it is not significant.

A screenshot of text

Description automatically generated

*Figure 8: A random intercept model with VEHICLE\_MAKE independent variable*

Now, again adding a second random intercept VEHICLE\_MODEL to the above model to note the changes in the variance of the intercept (Figure 9), we can see that the estimate again changes more towards ‘No’. The variance of VEHICLE\_MODEL is a little more than that of VEHICLE\_MAKE, hence it can be said that VEHICLE\_MODEL plays a bigger role.

A screenshot of text

Description automatically generated

*Figure 9: A random intercept model with VEHICLE\_MAKE and VEHICLE\_MODEL independent variables*

Adding fixed variables NUM\_OF\_YEARS and SPEED\_ZONE to the model, we have slight change in the intercepts. The positive value of Estimate indicates that adding these variables has increases the chance of getting a ‘Yes’, though they are very small. This means that there is a slight increase in the chances of having a higher severity when the vehicle gets older, or when the vehicle is in the higher speed zone (essentially meaning that the vehicle is in higher speed). This is shown in Figure 10.

A screenshot of a cell phone

Description automatically generated

*Figure 10: A random intercept model with added Level 2 fixed variables*

Now, the last thing to do is to add a random slope in the model. A subset of the dataset where the accident year is 2018 was sued to see the effect as the computation did not complete for the entire dataset. It can be seen that the NUM\_OF\_YEARS has had a little impact on the Variance of the Random Effects model, so it is also not that significant. Adding other intercept values also does not have considerable effect on the data. Hence, we will be excluding it from our analysis.

A screenshot of text

Description automatically generated

*Figure 11: A random slope random intercept model applied on a subset of dataset to analyse the effects*

We will now be taking the AIC to compare between our models and finding out the best model among them. Figure 12 shows the AIC values for all the models discussed above. It shows that the model `m3`, which takes NUM\_OF\_YEARS and SPEED\_ZONE as the fixed variables and the random intercepts of our multilevel variables VEHICLE\_MAKE and VEHICLE\_MODEL gives the most efficient model.

A screenshot of a cell phone

Description automatically generated

*Figure 12: Performance of models, compared using AIC*

Though we have compared our models and found that our model m3 is the best fitting model, it has variance of intercepts only 0.081 for VEHICLE\_MODEL and 0.0617 for VEHICLE\_MAKE. Since the model is a same slope and random intercept model, the parallel regression lines which are drawn are separated by a very small distance. This implies that the regression lines drawn by using our independent variables will be very close to each other, which means that there is no effect of either VEHICLE\_TYPE or VEHICLE\_MODEL on our response variable individually. So, the answer to whether the vehicle’s make or model impacts severity of accidents is negative.

# Conclusion

The original study conducted by the group found that type of vehicle played a significant role in the severity of accidents. It had given an impression that except two wheelers, all the vehicle types were statistically significant. However, this assumption is contradicted by analysing deeper and performing multilevel modeling by taking a subset and using the vehicle’s make and model. The reason that the previous study had given those to be statistically significant may be due to overfitting, since the data points were in the order of 180,000.

The main thing to take from this study is that the simple regression models, though would give a good approximation, may be incorrect. It is necessary to go deeper into details if one were to set a firm assumption on the model. Lastly, this study has come up with the conclusion that the vehicle’s attributes are not statistically significant when it comes to the severity of the accidents.

# References

Taimur, U., Liping, F., & F. M. M., Luis. (2016). *Injury severity analysis: comparison of multilevel logistic regression models and effects of collision data aggregation*. Retrieved from https://link.springer.com/article/10.1007/s40534-016-0096-4

Gudmundur, F., & Fred, L. M. (2004). *Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents*. Retrieved from https://www.sciencedirect.com/science/article/pii/S0001457502001355

Bureau of Infrastructure, Transport and Regional Economics 2018, *International road safety comparisons 2016*, BITRE, Canberra.

Magnusson, K. (2015). *Training for multilevel modeling*. Retrieved from http://www.bristol.ac.uk/cmm/learning/multilevel-models/samples.html

Taylor, J. J. (2011). *Confusing Stats Terms Explained: Residual*. Retrieved from http://www.statsmakemecry.com/smmctheblog/confusing-stats-terms-explained-residual.html

Blissett, R. (2017). *Multilevel Modeling in R*. Retrieved from https://rpubs.com/rslbliss/r\_mlm\_ws

Dr. Buchanan, E. M. (2019). *R - Binary Logistic Multilevel Models*. Viewed from https://www.youtube.com/watch?v=6MexZiX-2W8

Jovanovic, M. (2015). *R Playbook: Introduction to Multilevel/Hierarchical Models*. Retrieved from https://complementarytraining.net/r-playbook-introduction-to-multilevelhierarchical-models/

Borowski, T. G. *Chapter 16: Multilevel Modeling*. Retrieved from https://ademos.people.uic.edu/Chapter16.html

# Appendices

## Appendix I

A screenshot of a computer

Description automatically generated

*Appendix I: The final model of our group’s study on Road Accidents*