MAS61004 - Assessed Project

Allan Cousins, Matthew Knowles, Oliver Hewitt

2/23/2023

1 Executive Summary

Motivation for the investigation: to understand the relationship between speed limit (plus other factors) and accident severity.

Our focus is to assess the risk of a fatal or severe injury, given that one has already occurred - using road safety data exclusively from the year 2020. Analysis in the report will be carried out in R using Generalised Linear Models (GLMs), as it would be incorrect to assume the data is normally distributed.

(Outline conclusions here)

2 Introduction

The main purpose of this investigation is to describe the relationship primarily between accident severity and speed limit, and other factors included in the data provided by the UK Government's Department for Transport. The data provided are public domain, and include: "Road Safety Data - Accidents 2020.csv", "Road Safety Data - Vehicles 2020.csv" and "Road Safety Data - Casualties 2020.csv", merged by an accident index present in all three data sets.

Accident severity is categorised as either "fatal", "serious" or "slight", and the aim is to conduct the analysis treating it as a binary variable - i.e. grouping severity into one of two categories, "fatal or serious" and "slight". With this, it effectively eliminates a number of ways to analyse the data, but has potential to give more accurate results given the correct statistical application.

Some primary objectives of this investigation:

- Out of a select few variables in all three data sets (including speed limit), which have the most impact on accident severity when they change slightly?
- Assess the accuracy of the model used to fit the data, and how appropriate it is.

The explanation as to why the model was used will be outlined in the Method, along the corresponding R code or summary outputs found in the Appendix. The Results section will be an interpretation of the outputs from the model, and some relevant visualizations of the data. Ultimately, both of the above points will be reviewed (with the help of the Method and the Results).

3 Method

Our primary analysis, as outlined in the summary, will be GLMs fitted to the data in R. We firstly aim to build a model that predicts accident severity based on speed limit.

Initially cleaning and preparing the data was important, and we remove any values with -1 as to not impede on the results of the GLM. The Vehicles and Casualties data each contain three variables including the accident index, which will be merged onto the Accidents data (see Appendix, A1).

As mentioned previously, we wouldn't assume the data are normally distributed nor would we expect the outcome of our residuals to be normally distributed. We assume that a large demographic of the UK were not involved in a traffic accident in 2020, and an even smaller demographic were involved once or multiple times. In addition, not all response variables considered in the analysis are continuous - hence GLM would be an appropriate method to use under this non-normality and non-continuous response assumption.

We determined that Logistic Regression with a logit link function was the most appropriate method, given the use of dichotomous data (as we need accident severity to have two possible outcomes in our case). In R, we altered the data frame to have a binary response variable with two classes, as opposed to three.

With the final data frame created in A1, we then need assess to how well the model works on new data. For the purposes of this investigation, we will subset this data. We can split the data, and train the model using about 80% of the data – and call it our training set. The remaining 20% we use as our independent test set. Ultimately, this is to train the model on just the training set, and then see how well this predicts the other data (see Appendix, A2).

We fit the GLM to the training data and specify that the response variable is binomial, using a logit link function (see Appendix, A3). We train the model on all predictor variables, but initially observe the model with exclusively one predictor variable being the speed limit, where we consider odds ratio and it's exponent - to measure the association between speed limit ad accident severity alone. In Logistic Regression, the exponential function of the regression coefficient is the odds ratio associated with a one-unit increase in the exposure. [3]

4 Results

4.1 Descriptive plots:

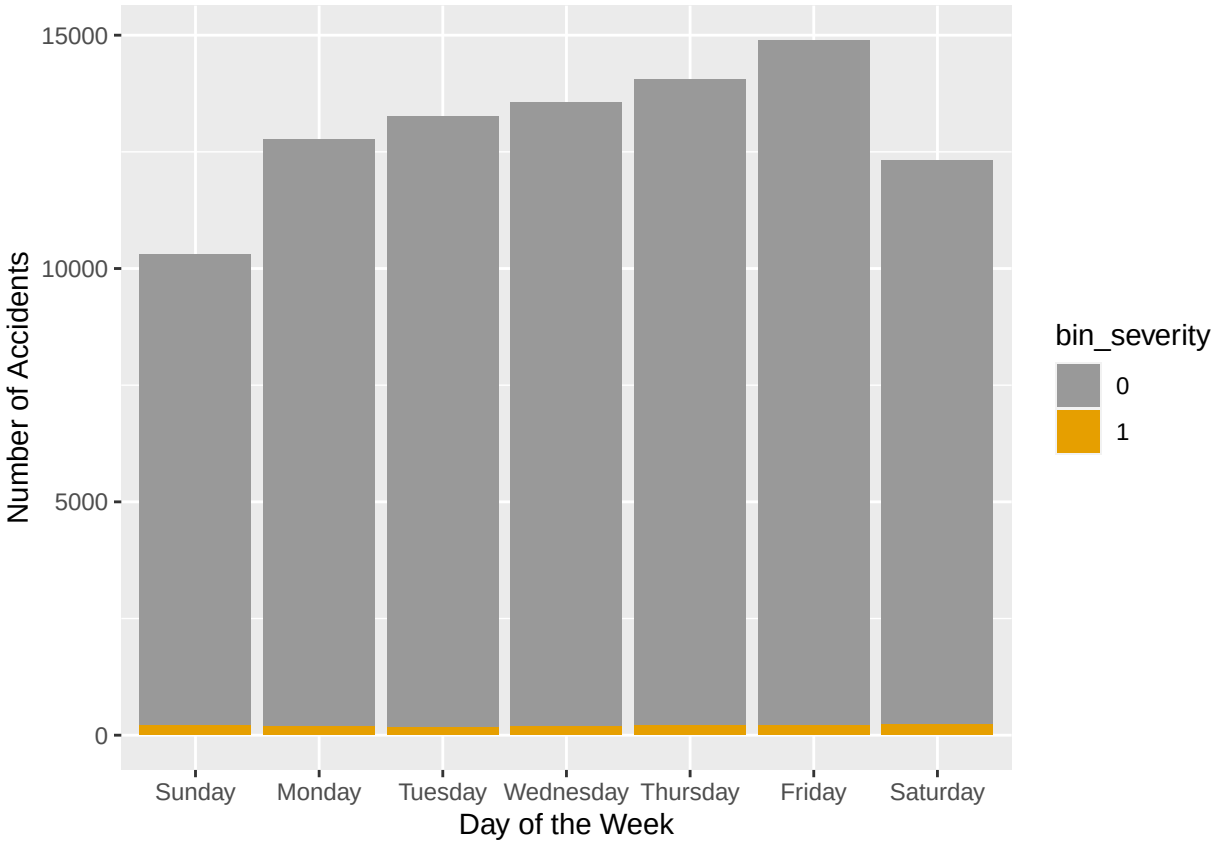

Figure 1: Bar chart presenting Number of Accident by Day

4.2 GLM:

We use the standard function *summary()* to produce result summaries of our model which uses the *glm()* function, considering all predictor variables (see Appendix, A3).

A4 is our summary of the *glm()* function fitting the Binomial model with all predictor variables (A3). We use this output to see which predictor variables have an effect on the response variable by the p value seen in the final column *Pr(>|z|)* - the predictor variable has an effect on the response when $p < 0.05$.

Denote the significance of these variables by the number of stars on each row: we see that the sex of the driver involved in the accident, light conditions and speed limit bear the most significance on the outcome. However, these results can be misleading, as they may be neither statistically significant nor practically important. [1]

For now, we will assess in a little bit more depth the GLM function using just *speed_limit* as a predictor variable (see Appendix, A5). We then observe the summary output and focus on the coefficients (see Appendix, A6): the coefficient of the variable *speed_limit* is 0.043836, and as it's positive it can be deduced that the chance of observing a "fatal or serious" accident increases with a higher speed limit on the road. From here we can the exponent of this value to obtain the odds ratio value 1.044747 (see Appendix, A7).

In essence, for every unit increase in the speed limit of the road, the odd ratio increases on average a constant factor of roughly 4.47%. Then, we construct a 95% confidence interval for the estimated model coefficient and take the exponent of this result. [2]

This implies that with a 95% confidence interval, for every unit increase in speed limit the observation of a "fatal or serious" accident becomes between roughly 4.24% and 4.71% more likely (see Appendix, A8).

5 Appendix

A1: Cleaning and compilation of the main data frame

```
# We begin by converting to a binary variable by severity.
# Accidents with a severity of 1 stay as such, and others are converted to a 0.
accidents <- accidents %>%
  mutate(bin_severity = ifelse(.data$accident_severity == 1, 1, 0))

sub_accidents <- accidents %>%
    select(
        "ï..accident_index",
        "bin_severity",
        "day_of_week",
        "road_type",
        "speed_limit",
        "light_conditions",
        "weather_conditions",
        "road_surface_conditions"
    )

sub_vehicles <- vehicles %>%
    select(
        "ï..accident_index",
        "sex_of_driver",
        "age_band_of_driver"
    )

sub_casualties <- casualty %>%
    select(
        "ï..accident_index",
        "sex_of_casualty",
        "age_band_of_casualty"
    )

df_sub <- merge.data.frame(sub_vehicles, sub_casualties, by = "ï..accident_index")
df_sub_unique <- unique(df_sub)
df <- merge.data.frame(sub_accidents, df_sub_unique, by = "ï..accident_index")


# We want all variables, except speed limit, to be factors.

df <- df %>%
    mutate(across(c(where(is.numeric), -speed_limit), as.factor))

df$bin_severity <- as.factor(df$bin_severity)
```

```r
# We need to remove any rows in which there is a -1 in a column of the data.

has.neg <- apply(df, 1, function(row) any(row == -1))
df <- df[-which(has.neg), ] %>%
    select(-ï..accident_index)
```

A2: Training and test data

```r
df$id <- 1:nrow(df)
df_train <- df %>% sample_frac(0.8)
df_test <- anti_join(df, df_train, by = "id")
```

A3: *glm* function in R using the *binomial* argument with the training data

```r
fit <- glm(
    data = df_train,
    formula = bin_severity ~ .,
    family = binomial(link = "logit")
)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

A4: Summary of A3

```r
summary(fit)
```

```
##
## Call:
## glm(formula = bin_severity ~ ., family = binomial(link = "logit"),
##     data = df_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9924  -0.2114  -0.1493  -0.1120   3.6980
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.999e+01  7.548e+02  -0.026  0.97888
## day_of_week2               -3.581e-01  7.772e-02  -4.607 4.08e-06 ***
## day_of_week3               -2.457e-01  7.532e-02  -3.263  0.00110 **
## day_of_week4               -1.772e-01  7.340e-02  -2.414  0.01579 *
## day_of_week5               -8.951e-02  7.141e-02  -1.253  0.21006
## day_of_week6               -3.107e-01  7.370e-02  -4.216 2.48e-05 ***
## day_of_week7                1.867e-02  7.154e-02   0.261  0.79409
## road_type2                  1.193e-01  4.416e-01   0.270  0.78714
```

```
## road_type3                 7.722e-01  1.708e-01   4.521 6.14e-06 ***
## road_type6                 1.435e+00  1.651e-01   8.691  < 2e-16 ***
## road_type7                 2.684e-01  2.480e-01   1.082  0.27906
## road_type9                 3.912e-01  4.429e-01   0.883  0.37711
## speed_limit                4.570e-02  1.557e-03  29.347  < 2e-16 ***
## light_conditions4          3.880e-01  5.848e-02   6.634 3.26e-11 ***
## light_conditions5          4.952e-01  2.037e-01   2.431  0.01507 *
## light_conditions6          8.363e-01  5.594e-02  14.949  < 2e-16 ***
## light_conditions7          1.403e-02  1.978e-01   0.071  0.94343
## weather_conditions2       -4.257e-01  7.495e-02  -5.680 1.35e-08 ***
## weather_conditions3        4.474e-01  4.004e-01   1.117  0.26392
## weather_conditions4        3.703e-01  1.220e-01   3.035  0.00240 **
## weather_conditions5       -2.397e-01  1.384e-01  -1.732  0.08334 .
## weather_conditions6       -1.305e+01  3.249e+02  -0.040  0.96795
## weather_conditions7        2.140e-01  1.733e-01   1.235  0.21677
## weather_conditions8        2.972e-01  1.170e-01   2.539  0.01112 *
## weather_conditions9       -4.963e-01  2.280e-01  -2.177  0.02950 *
## road_surface_conditions2   1.074e-01  5.057e-02   2.124  0.03364 *
## road_surface_conditions3 -1.410e+01  2.321e+02  -0.061  0.95157
## road_surface_conditions4 -6.755e-01  2.310e-01  -2.924  0.00346 **
## road_surface_conditions5 -2.764e-01  3.929e-01  -0.703  0.48186
## road_surface_conditions9 -1.205e+01  1.544e+02  -0.078  0.93780
## sex_of_driver2            -3.442e-01  5.149e-02  -6.684 2.33e-11 ***
## sex_of_driver3            -9.105e-01  2.924e-01  -3.114  0.00184 **
## age_band_of_driver2        1.228e+01  7.548e+02   0.016  0.98702
## age_band_of_driver3        1.176e+01  7.548e+02   0.016  0.98757
## age_band_of_driver4        1.272e+01  7.548e+02   0.017  0.98656
## age_band_of_driver5        1.298e+01  7.548e+02   0.017  0.98628
## age_band_of_driver6        1.296e+01  7.548e+02   0.017  0.98630
## age_band_of_driver7        1.289e+01  7.548e+02   0.017  0.98637
## age_band_of_driver8        1.296e+01  7.548e+02   0.017  0.98630
## age_band_of_driver9        1.313e+01  7.548e+02   0.017  0.98613
## age_band_of_driver10       1.285e+01  7.548e+02   0.017  0.98642
## age_band_of_driver11       1.291e+01  7.548e+02   0.017  0.98636
## sex_of_casualty2          -3.563e-01  4.609e-02  -7.731 1.07e-14 ***
## sex_of_casualty9          -1.255e+01  2.797e+03  -0.004  0.99642
## age_band_of_casualty2     -4.465e-01  2.528e-01  -1.766  0.07733 .
## age_band_of_casualty3     -3.208e-01  2.272e-01  -1.412  0.15799
## age_band_of_casualty4     -2.849e-01  1.858e-01  -1.534  0.12511
## age_band_of_casualty5     -1.959e-01  1.792e-01  -1.093  0.27420
## age_band_of_casualty6     -1.999e-01  1.731e-01  -1.155  0.24812
## age_band_of_casualty7     -5.706e-02  1.748e-01  -0.326  0.74415
## age_band_of_casualty8     -5.211e-02  1.754e-01  -0.297  0.76645
## age_band_of_casualty9      1.369e-01  1.770e-01   0.774  0.43907
## age_band_of_casualty10     5.479e-01  1.829e-01   2.996  0.00273 **
## age_band_of_casualty11     1.102e+00  1.824e-01   6.044 1.50e-09 ***
## id                         1.262e-06  4.511e-07   2.799  0.00513 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27049  on 138712  degrees of freedom
## Residual deviance: 24326  on 138658  degrees of freedom
## AIC: 24436
##
## Number of Fisher Scoring iterations: 16
```

A5: *glm* function in R using the *binomial* argument with the training data, using exclusively the predictor variable *speed_limit*

```
fit_spd_lim <- glm(
    data = df_train,
    formula = bin_severity ~ speed_limit,
    family = binomial(link = "logit")
)
```

A6: Summary of A5

```
summary(fit_spd_lim)
```

```
##
## Call:
## glm(formula = bin_severity ~ speed_limit, family = binomial(link = "logit"),
##     data = df_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3521  -0.2293  -0.1486  -0.1486   3.1450
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.813945   0.060907  -95.46   <2e-16 ***
## speed_limit  0.043775   0.001171   37.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27049  on 138712  degrees of freedom
## Residual deviance: 25662  on 138711  degrees of freedom
## AIC: 25666
##
## Number of Fisher Scoring iterations: 7
```

A7: Odds ratio, using the coefficent of speed limit from A6

```
exp(coefficients(fit_spd_lim)[2])
```

```
## speed_limit
##    1.044747
```

A8: Construction of a 95% confidence interval for the speed limit coefficient (using A5 and A6)

```
confint.default(fit_spd_lim)[2,]
```

```
##      2.5 %     97.5 %
## 0.04148008 0.04606948
```

```
exp(confint.default(fit_spd_lim)[2,])
```

```
##    2.5 %   97.5 %
## 1.042352 1.047147
```

6 References

[1] Gelman, A., Stern, H. (2006): "The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant"

[2] Hartmann, K., Krois, J., Waske, B. (2018): E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin.

[3] Szumilas M. (2010). Explaining odds ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, 19(3), 227–229.