Preliminary Analysis of RTC Data

Matthew Knowles

2023-03-08

1: Data Cleaning

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)) %>%
    filter(light_conditions != -1)
```

We now reduce the size of the dataset to contain only variables we care about

```
accidents <- accidents %>%
    select(
        "bin_severity",
        "first_road_class",
        "day_of_week",
        "road_type",
        "speed_limit",
        "junction_detail",
        "second_road_class",
        "light_conditions",
        "weather_conditions",
        "road_surface_conditions"
    )
```

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

```
accidents <- accidents %>%
    mutate(across(where(is.integer), as.factor))
accidents$speed_limit <- as.integer(accidents$speed_limit)

head(tibble(accidents))
```

```
## # A tibble: 6 x 10
##   bin_severity first_r~1 day_o~2 road_~3 speed~4 junct~5 secon~6 light~7 weath~8
##          <dbl> <fct>     <fct>   <fct>     <int> <fct>   <fct>   <fct>   <fct>
## 1            0 6         3       6             2 0       0       1       9
## 2            0 3         2       6             2 9       6       1       1
## 3            0 5         4       6             3 3       6       4       1
## 4            0 3         4       6             3 0       0       4       1
## 5            0 3         4       6             3 3       5       4       1
## 6            0 3         4       2             2 3       6       4       1
## # ... with 1 more variable: road_surface_conditions <fct>, and abbreviated
## #   variable names 1: first_road_class, 2: day_of_week, 3: road_type,
```

```
## #   4: speed_limit, 5: junction_detail, 6: second_road_class,
## #   7: light_conditions, 8: weather_conditions
```

We can now also create a train and test set for later on.

```
accidents$id <- 1:nrow(accidents)
accidents_train <- accidents %>% sample_frac(0.7)
accidents_test <- anti_join(accidents, accidents_train, by = "id")
```

2: Fitting

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

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

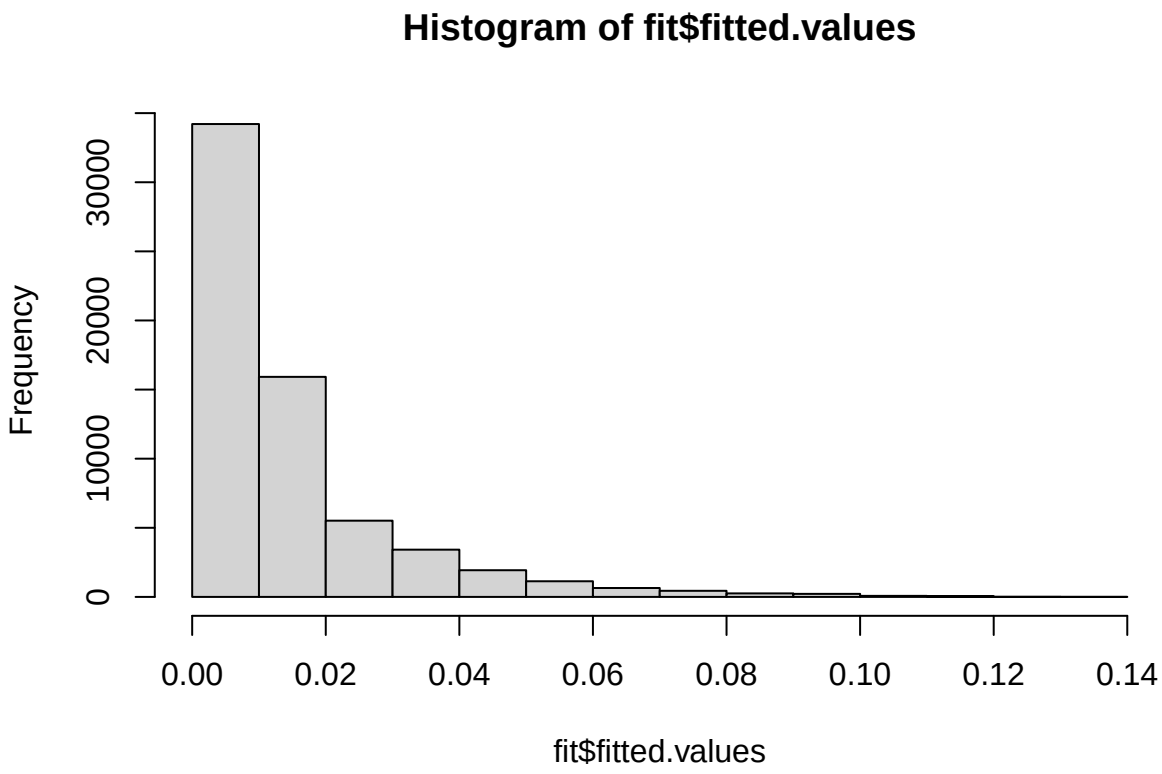
Let us take a look at the summary of this model.

```
summary(fit)
```

```
##
## Call:
## glm(formula = bin_severity ~ ., family = binomial(link = "logit"),
##     data = accidents_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5447  -0.1844  -0.1355  -0.1033   3.7138
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.338e+01  4.279e+03  -0.008  0.99378
## first_road_class2         1.595e-02  5.337e-01   0.030  0.97615
## first_road_class3         2.395e-01  1.689e-01   1.418  0.15620
## first_road_class4         6.649e-02  1.961e-01   0.339  0.73461
## first_road_class5        -1.874e-01  2.517e-01  -0.745  0.45649
## first_road_class6        -2.168e-01  1.899e-01  -1.141  0.25368
## day_of_week2             -1.998e-01  1.249e-01  -1.600  0.10954
## day_of_week3             -3.288e-01  1.278e-01  -2.573  0.01008 *
## day_of_week4             -2.211e-01  1.245e-01  -1.776  0.07576 .
## day_of_week5             -2.437e-01  1.233e-01  -1.977  0.04805 *
## day_of_week6             -2.991e-01  1.222e-01  -2.447  0.01439 *
## day_of_week7             -6.531e-03  1.183e-01  -0.055  0.95598
## road_type2                1.330e-01  5.133e-01   0.259  0.79553
## road_type3                2.118e-01  3.157e-01   0.671  0.50233
## road_type6                7.297e-01  3.081e-01   2.368  0.01787 *
## road_type7                1.726e-01  4.081e-01   0.423  0.67236
## road_type9               -6.093e-01  7.728e-01  -0.788  0.43046
## speed_limit               3.717e-01  2.791e-02  13.317  < 2e-16 ***
## junction_detail0          1.326e+01  4.217e+03   0.003  0.99749
## junction_detail1          1.206e+01  4.217e+03   0.003  0.99772
## junction_detail2          1.249e+01  4.217e+03   0.003  0.99764
```

```
## junction_detail3          1.261e+01  4.217e+03   0.003  0.99761
## junction_detail5          1.212e+01  4.217e+03   0.003  0.99771
## junction_detail6          1.264e+01  4.217e+03   0.003  0.99761
## junction_detail7          1.168e+01  4.217e+03   0.003  0.99779
## junction_detail8          1.265e+01  4.217e+03   0.003  0.99761
## junction_detail9          1.272e+01  4.217e+03   0.003  0.99759
## junction_detail99        -2.534e-01  4.221e+03   0.000  0.99995
## second_road_class0        1.386e+01  7.223e+02   0.019  0.98469
## second_road_class1        1.394e+01  7.223e+02   0.019  0.98460
## second_road_class2        1.546e+01  7.223e+02   0.021  0.98292
## second_road_class3        1.367e+01  7.223e+02   0.019  0.98490
## second_road_class4        1.423e+01  7.223e+02   0.020  0.98428
## second_road_class5        1.431e+01  7.223e+02   0.020  0.98419
## second_road_class6        1.406e+01  7.223e+02   0.019  0.98447
## light_conditions4         5.988e-01  8.897e-02   6.730 1.70e-11 ***
## light_conditions5         7.800e-01  2.882e-01   2.706  0.00681 **
## light_conditions6         6.888e-01  9.797e-02   7.031 2.05e-12 ***
## light_conditions7         1.571e-01  2.754e-01   0.570  0.56849
## weather_conditions2      -5.176e-01  1.309e-01  -3.954 7.68e-05 ***
## weather_conditions3      -2.316e-01  1.025e+00  -0.226  0.82118
## weather_conditions4       1.691e-01  2.224e-01   0.760  0.44712
## weather_conditions5       1.109e-01  1.989e-01   0.558  0.57705
## weather_conditions6       3.188e-01  1.042e+00   0.306  0.75960
## weather_conditions7      -3.940e-02  3.317e-01  -0.119  0.90545
## weather_conditions8      -6.611e-01  2.869e-01  -2.304  0.02120 *
## weather_conditions9      -6.566e-01  3.445e-01  -1.906  0.05666 .
## road_surface_conditions1  1.210e-02  7.275e-01   0.017  0.98673
## road_surface_conditions2  1.612e-02  7.299e-01   0.022  0.98238
## road_surface_conditions3 -1.498e+01  6.210e+02  -0.024  0.98076
## road_surface_conditions4 -5.942e-01  8.157e-01  -0.728  0.46635
## road_surface_conditions5 -1.602e+00  1.248e+00  -1.284  0.19930
## road_surface_conditions9 -1.247e+01  2.453e+02  -0.051  0.95946
## id                        3.923e-06  1.487e-06   2.637  0.00835 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9821.6  on 63838  degrees of freedom
## Residual deviance: 9003.3  on 63785  degrees of freedom
## AIC: 9111.3
##
## Number of Fisher Scoring iterations: 17
```

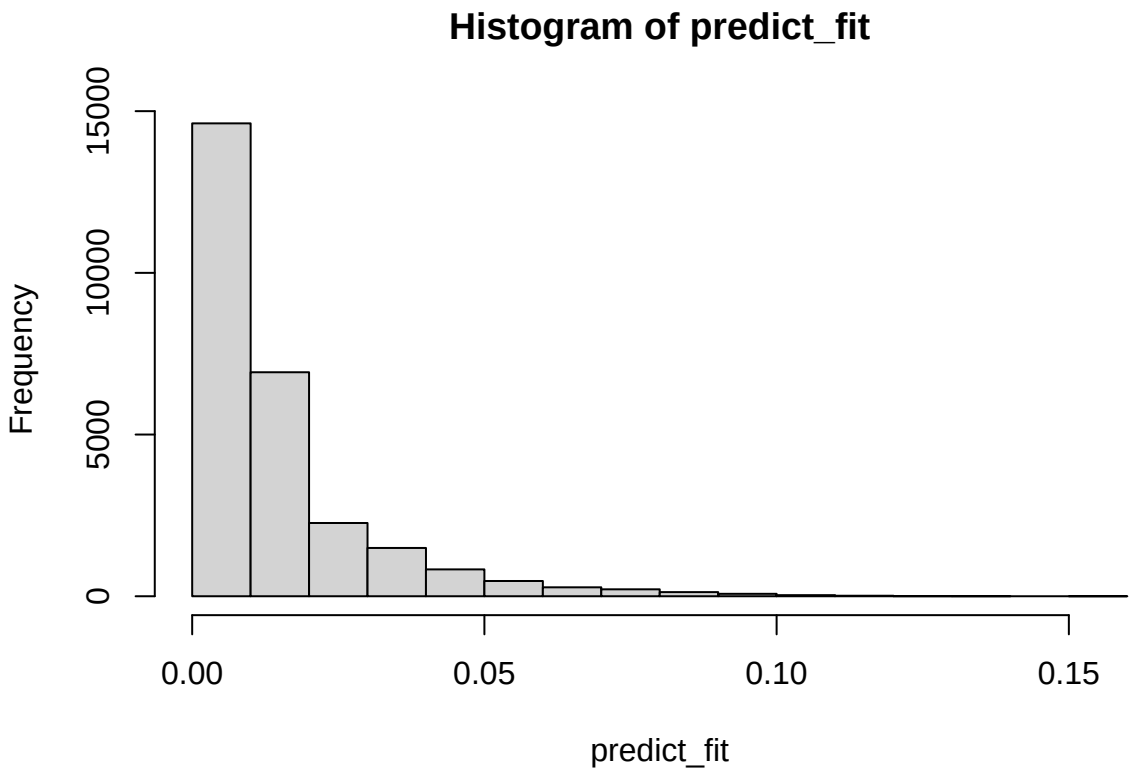
```
hist(fit$fitted.values)
```



3: Predicting Unseen Values

```
predict_fit <- predict(
    fit,
    newdata = accidents_test,
    type = "response"
)
```

```
hist(predict_fit)
```


Histogram of predict_fit
Frequency
0
5000
10000
15000
0.00
0.05
0.10
0.15
predict_fit