Analysis Report of the U.S. Department of Transportation's Fatality Reporting System (FARS) in 2016

By
Abdoulaye Kaloga

Introduction to Dataset:

These are the datasets taken from the U.S Department of Transportation's Fatality Analysis Reporting Systems (FARS) in 2016. It has 51 rows and 23 columns containing the following information in the respective columns:

- State: Factor w/ 51 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 9
- Population : Int 4858979 738432 6828065 2978204 39144818
- Vehicle miles travelled millions: Int 67257 5045 65045….
- Fatal crashes : Int 783 60 810 472 2925 506 253 122 23 2699 ...
- Deaths : Int 849 65 893 531 3176 546 266 126 23 2939 ...
- Deaths.per.100.000.population : Int 18 9 13 18 8 10 7 13 3 15 ...
- Deaths.per.100.million.vehicle.miles.traveled : Int 1 1 1 2 1 1..
- Car occupants : Int 356 13 260 191 1161 175 108 49 6 924 ...
- Pickup and SUV occupants: Int 279 23 217 181 581 165 46 15..
- Larges trucks occupants : Int 16 0 14 17 23 13 7 2 0 31 ...
- Motorcyclists : Int 64 10 131 73 456 103 49 18 2 577 ...
- Pedestrians : Int 98 12 153 43 742 59 45 35 13 628 ...
- Bicyclists Int 9 0 29 3 129 13 3 3 1 150 ...
- Unknown mode of transport: Int 27 7 89 23 84 18 8 4 1 118 ...
- Single vehicle : Int 471 41 491 266 1806 302 157 67 16 1579 ...
- Multiple vehicle : Int 378 24 402 265 1370 244 109 59 7 1360 ...
- Unrestrained fatally injured occupants: Int 355 15 251 190 549..
- Restrained fatally injured occupants: Int 251 15 180 151 1065..
- Unknown restraint status of fatally injured occupants: Int 41 8 6..
- Urban: Int 249 31 554 155 1802 284 217 60 23 1238 ...
- Rural: Int 480 33 335 376 1366 260 46 66 0 453 ...
- X: logi NA NA NA NA NA NA ...
- X.1: logi NA NA NA NA NA NA ...

Multiple Linear Regression is the type of model I will be performing in analysing the data and concluding my findings.

Descriptive data Analysis:

Describing my findings from the descriptive view.

Numerical Summary:

- After reading the csv file on the R studio , I called on the function head() to give me the 5–number summaries, it returned the first parts of the dataset which appears to be quantitative variables and from the introduction section we can just tell.

- To find the correlation and variance between variables, I called on these functions: cor(), var() and corrplot(). These function returned figures and graphical visualisation from what I could report that there is strong correlation and variance amongst variables.

Deaths per 100.000 population appears to be moderately to strongly related:

"Unrestrained fatally injured occupants" (positively)
"Fatal Crashes" (moderately)
"Population"(negatively)
"Single vehicle" (positively)
"Rural"(positively)
"Car occupants"(positively)

The relationship between Deaths per 100.000 population appear to be linear:

"Unrestrained fatally injured occupants (linear)
"Fatal Crashes" (linear)
"Population" (linear)
"Single vehicle" (linear)
"Rural"(linear)
"Car occupants"(linear)

There is multicollinearity in:

"Unrestrained fatally injured occupants" is strongly connected to "Fatal Crashes", "Population", "Single vehicle".
"Population" is strongly connected to "Fatal crashes", "Single vehicle".

"Car occupants" is strongly connect to "Population", "Fatal crashes", "Single vehicle", "Unrestrained fatally injured occupants" and "Fatal Crashes".

Please Referred to the diagram.

Further descriptions were done on the dataset, and it turns out that the dataset has some entries errors (missing values in columns X, X.1). There are some unusual (asymmetric) distribution and some outliers(these outliers can be seen on the graphical summaries of the variables).

Graphical Summary:

By introducing boxplot, histograms, scatterplots we could see the graphical summaries of the variable. Please referred to the diagrams attached in the folder.

Histogram:

All SEVEN variables are right skewed (asymmetric) except "Deaths per 100.000 population".

Scatterplot:

Most variables showed a positive relationship.

boxplot:

All the Eight variables has outliers.

Determining the Best Predictive Model:

68% of the variation in Deaths per 100.000 population is explained by all the selected predictor variables.

From the t-tests, we can make educational estimation if each variable is significantly different from zero or not.

It appears that higher "Deaths per 100 million vehicle miles travelled" are to increased "Deaths per 100.000 population". Higher "Population" is related to increase in "Deaths per 100.000 population". Higher "Unrestrained fatally injured occupants" is connected to increase in "Deaths per 100.000 population".

F-test statistic is 16.21. To compare it to F-critical value and find the rejection region, we would need to look it up in statistical tables or compute it. Instead, a straightforward way is to look at the p-value.

Therefore, the p-value is smaller than the significance level 0.05. We reject the null hypothesis (H0 : The independent don't have significant effect is no the dependent). The insurance model is not looking so good in this respect.

The summary anova does not reveal a lot of significant parameters.

We have the Adjusted R-squared(68%) looking moderately good let's see if we can make it better. By use the technique such as the forward stepwise regression to sequentially test each of our predictor variables.
Our Adjusted R-squared has reduced strangely. How can we optimise it? We had noticed before some of the variables are not normalised, therefore, we will use log10() to improve our model and if necessary reduce the number of predictor variables.

Preform analysis of the best regression model:

Coefficients and model fit:

It indicates that the intercept and slope variables are less than 0.05 that means in million:

- One unit change in "Deaths per 100 million vehicles miles travelled", the fatality rate increases by 6.382e+00.
- One unit change of the "Population" ,the fatality rate decreases per - -9.325e-07 .
- One unit change in log10 of the "Fatal crashed" ,the fatality rate increases per 2.186e+01 .
- One unit change in "Single vehicle", the fatality rate increases per 5.286e-02.
- One unit change in "Fatal crashed", the fatality rate decreases per -2.361e-02 .

Model Diagnostics:

After altering the model using Cook's distance, we finally get these plots as it explained here :

1. Residuals vs Fitted
   - The graph indicates a spread out residuals around the line without distinct patterns, that is a good indication we don't have non-linear relationships.
2. Normal Q-Q plot
   - This plot indicates whether our residuals are normally distributed. It appears that our residuals followed more or less a straight line and we conclude our fitter looks better.
3. Scale-location
   - This plot shows an equal (randomly) spread out points. The assumption of equal variance (homoscedasticity) is also present.

Outliers / Influential values:

An outlier is an observation where the response does not correspond to the model fitted to the bulk of the data. There might be several reasons for that:

Our data is really atypical.
Misprint in the data.
4. Residuals vs Leverage

- As this plot help us locate the influential cases that have an effect on the regression. Our plot has no influential point as it appears all the data point to be inside the cook's distance.

Report of the original vs new model:

AIC (Akaike information criterion) is a measure of how well the model fits the data less a penalty parameter for how the complex model is. We want to minimize the AIC value so the lower AIC the better.

- AIC of original model: 254.698
- AIC of new model: 149.005
- Adjusted R-squared original model: 0.6728
- Adjusted R-squared new model: 0.9588
- Residual Standard Error original model: 2.728
- Residual Standard Error new model: 0.9678

Possible Remedial Measures Section:

Nonlinearity was found amongst the data, we transformed the independent using log10() which gave us a good fitted model with the predictor variables against the dependent variable.

Interpretation of the results of the final model:

- The AIC of the final model is 147.6523 from 149.005
- The fatality rate on average is 12.05882.
- Adjusted R-squared for the final model: 0.9592
- Residual Standard Error for the final model: 0.9633
- One unit change in log10 of the "Population" the fatality rate decreases per -2.346e+01.
- One unit change in log10 of the "Fatal crashed" the fatality rate increase per 2.186e+01.

Conclusion and recommendation:

We can assume in the global spectrum, there is significant decreased in population as more accidents occurs. We do not know what age or gender are more involved in this tragic accident in order to segment the population and

charge insurance fees based on age or gender, we can only assumed based on the data we have.

My recommendation will be to collect more data this time by taking in to consideration the age and gender of the population.