PA Model Solution June 14, 2019

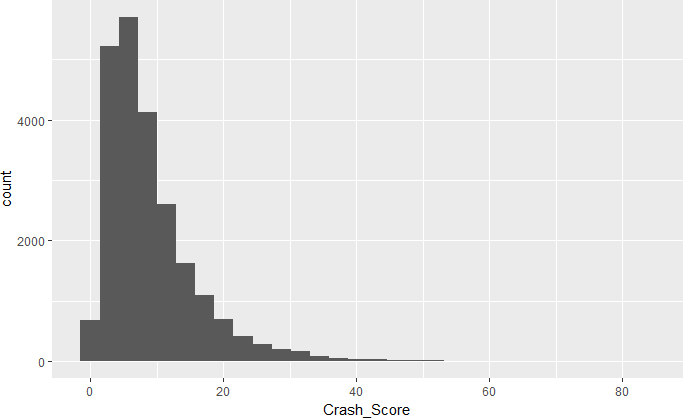
Exam PA June 2019 Project Report Template

**Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.**

As indicated in the instructions, work on each task should be presented in the designated section for that task.

# Task 1 – Explore the relationship of each variable to *Crash\_Score* (5 points)

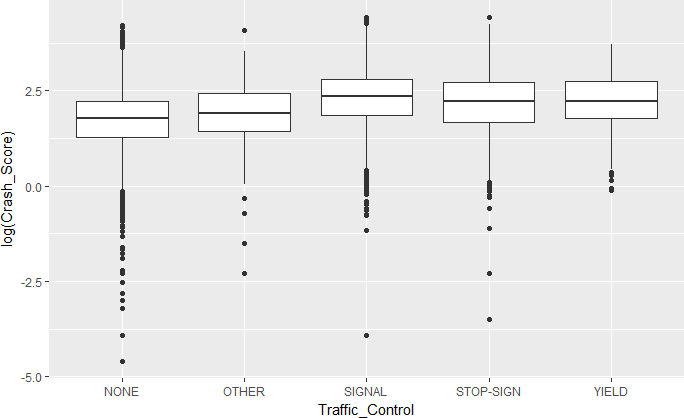
For the target variable Crash Score, the median is 7.16, the mean is 9.11, and the maximum is 83.41. This indicates that the distribution is skewed to the right. A histogram confirms this:



As a result, I explored boxplots of the log of the target variable split among the factors of each variable. Differences were observed for the following variables:

* Time\_of\_Day: Low Crash Score for period 1 (midnight to 4am)
* Rd\_Feature: High for INTERSECTION and RAMP
* Rd\_Character: High for OTHER
* Rd\_Configuration: High for TWO‐WAY‐UNDPROTECTED‐MEDIAN
* Rd\_Surface: Low for OTHER
* Rd\_Conditions: High for WET, low for OTHER
* Light: Low for DARK‐NOT‐LIT and OTHER
* Weather: Low for OTHER
* Traffic\_Control: Low for NONE and OTHER
* Work\_Area: Low for NO.

The plot for Traffic\_Control is provided below. The others can be obtained from the R code.



Looking at means and medians for the logarithms of crash scores reveals some other possible relationships beyond those already mentioned:

* Month: Higher in months October (10) through March (3)
* Rd\_Class: Higher for US HWY
* Rd\_Surface: Also, higher for the two ASPHALT levels relative to the two CONCRETE levels An example using Rd\_Surface appears below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Rd\_Surface** | **mean** | **median** | **n** |
| <fctr> | <dbl> | <dbl> | <int> |
| SMOOTH ASPHALT | 1.946158 | 1.986504 | 20007 |
| COARSE ASPHALT | 1.913299 | 1.931521 | 1997 |
| CONCRETE | 1.688533 | 1.704746 | 692 |
| GROOVED CONCRETE | 1.708268 | 1.796747 | 371 |
| OTHER | 1.450069 | 1.611677 | 70 |

It appears there are several variables that may predict the target variable, but it should also be noted that none stand out as making large differences.

# Task 2 – Reduce the number of factor levels where appropriate (5 points)

The following combinations are made:

* Time\_of\_Day: Time 1 = OVERNIGHT, Times 2 and 6 = LATE‐EARLY, Times 3‐5 = DAYTIME. They have different means and medians and make sense with regard to accident severity.
* Rd\_Feature: Combine INTERSECTION and RAMP into one level, INTERSECTION‐RAMP, and combine the others into OTHER. Intersection and ramp accidents are more likely to involve multiple vehicles and hence more damage.
* Rd\_Character: Based on differing mean scores, combine the three with OTHER into OTHER and combine the remaining levels as STRAIGHT or CURVED.
* Traffic\_Control: Combine NONE and OTHER into OTHER and the others into CONTROLLED to reflect some sort of control.

The other predictor variables either show little difference between the factor levels or enough differences throughout so that no obvious groupings exist.

# Task 3 – Use observations from principal components analysis (PCA) to generate a new feature (9 points)

Running the PCA on these three variables shows that only 22% of variation is explained by the first principal component (PC) and 35% by the first two PCs. However, the loadings may highlight interesting relationships among these variables.

The largest loadings on the first PC are:

* Rd\_ConditionsDRY: ‐0.51
* Rd\_ConditionsWET: 0.50
* WeatherCLEAR: ‐0.46
* Weather RAIN: 0.43

Applying these weights creates a variable that is strongly positive for rain/wet conditions and strongly negative for dry/clear conditions. It makes sense to pair up each of these as they would typically appear together, e.g. rain leads to wet roads.

Based on these results, I created a new feature, WETorDRY, based on the Rd\_Conditions and Weather variables, deleting these two but retaining the Light variable as is.

# Task 4 –Select an interaction (7 points)

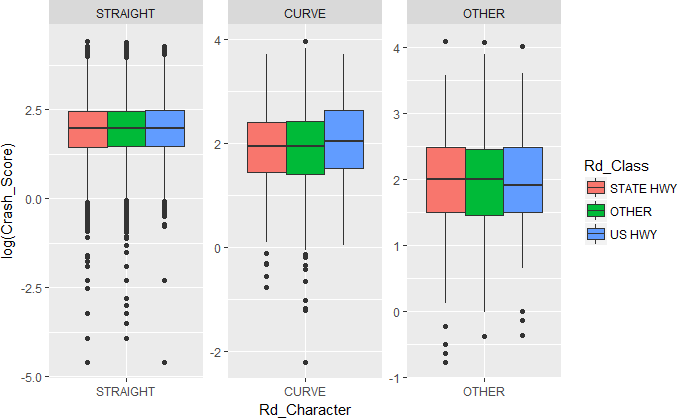
*The best candidates began by explaining what they were looking for when searching for an interaction. They next proposed an interaction based on an understanding of traffic behavior. They then used a graph to either confirm or dispel their interaction. In some cases, it was then necessary to try a new pair of variables.*

*Candidates were not required to create the interaction variable at this task. However, points were lost if the selected interaction was not used in subsequent tasks.*

*While the plots presented here used the logarithm of Crash\_Score, it was not necessary to do so to earn full credit.*

An interaction is indicated when changing the level of one variable alters how levels of the other variables affect the target.

A first thought is Rd\_Character and Rd\_Class. Changing Rd\_Character from STRAIGHT to CURVE may have a different effect depending on the Rd\_Class. U.S. highways may have gentler curves than state highways and hence a different effect. As seen below, there does seem to be an interaction. For the CURVE cases the crash score is higher for US HWY. While this is contrary to my intuition, the interaction appears to be worth considering.



I’ll use this one for future work.

# Task 5 – Select a distribution and link function (10 points)

Before building a GLM, I split the data into training (75%) and testing (25%) sets. The average target value was 9.108 for the training set and 9.109 for the testing set, so the built‐in stratification of the target variable worked well.

I have retained Year as a numeric variable in case there is a trend effect that needs to be accounted for. I converted Month to a factor variable so that any seasonal effect can be determined.

Before investigating various GLMs, an OLS regression was run using the variables previously created but not the interaction from the previous task. Key values were an AIC of 113,388 and an RMSE of 6.2481 on the testing set. This provides a benchmark for further model development.

The only link function I will consider is the log link. This link ensures that all predictions are positive values, which is a characteristic of the target variable. The log link is also easy to interpret.

Because the target variable is highly right skewed, a skewed distribution such as the gamma seems appropriate. For the gamma distribution with the log link and the interaction term, the AIC is 102,336 and the RMSE is 6.2278. Both are improvements over the OLS model in that smaller values are preferred.

Another skewed distribution is the inverse Gaussian. Running it with the log link did not converge, so this combination could not be evaluated.

A Gaussian distribution with the log link, while not right skewed, will ensure positive predictions. For the normal distribution with the log link and the interaction term, the AIC is 113,208 and the RMSE is 6.2390.

Based on these numerical results (the gamma distribution has a lower AIC and essentially the same RMSE), the gamma distribution with the log link will be used from here on. Given that the two models have the same number of parameters, the lower AIC indicates that the loglikelihood is larger and thus the model fits better to the training data.

# Task 6 – Select features using AIC or BIC (12 points)

When a regression model is constructed using a large set of predictor variables there is a risk of overfitting. While additional variables can only improve the fit to the training data, they may actually decrease the fit against unseen (testing) data. One of the methods for handling this is the use of penalized likelihood, also known as information criteria. When fitting models by maximum likelihood, additional variables never decrease the loglikelihood value. An information criterion demands that for an additional variable to be included it must not just increase the loglikelihood, it must do so by at least a specific amount. Two popular criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). For AIC, adding a variable requires an increase in the loglikelihood of two per parameter added. For BIC, the required per parameter increase is the logarithm of the number of observations. For the training dataset it is log(17,354) = 9.76 per parameter.

For this problem, BIC is a more conservative approach as there is a greater penalty for each parameter added, requiring more evidence to support additional variables. Our goal in this project is identify the key variables that relate to the target variable. As such, it makes sense to take a conservative approach and work with as few variables as necessary. Thus, BIC makes the most sense for this analysis.

Similarly, forward selection is more likely to end up with fewer variables. With forward selection, you start with no variables and then add variables until there is no improvement by the selected criterion. Backward selection starts with all the variables and sequentially removes them until no improvement results. It seems more likely that forward selection will result in a simpler model and hence that approach will be used.

When employing BIC and forward selection, the final model uses the following four features:

* Rd\_Feature
* Rd\_Configuration
* Time\_of\_Day
* Traffic\_Control

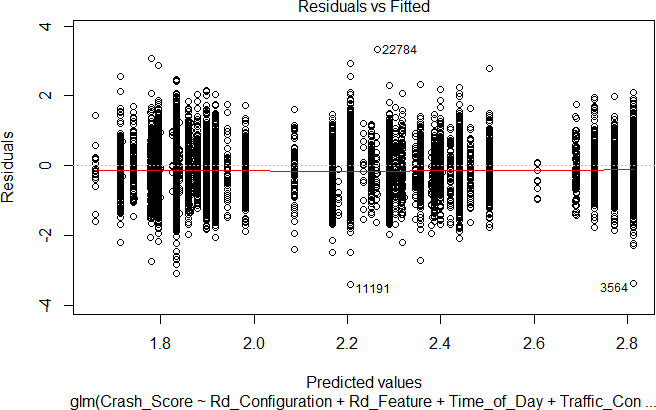
When running the model with these four variables I noticed that Rd\_Configuration = UNKNOWN was not significant compared to the base class. Forward selection with factor variables does not consider

individual factor levels, so this is a possible outcome of this method. Options available include combining this level with the base level or binarizing this variable and rerunning the analysis. Although this level is clearly not significant (p = 0.96), I’m not sure what to do with it as folding it into the base class may be hard to explain. There are only 57 records with this value, so leaving it in should not affect other conclusions.

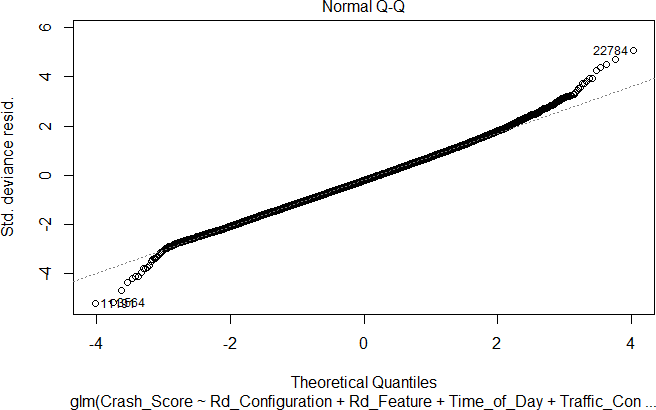
# Task 7 – Validate the model (9 points)

Running the model with the four features identified in Task 6 produced an RMSE of 6.2301 on the test data. Training and testing the assistant’s OLS model on the same data produced an RMSE of 6.2481. The RMSE on unseen data for the GLM is slightly lower despite having far fewer features, suggesting that the GLM is the better model.

The following plot of residuals versus fitted values shows that the model performs well. Because all the predictors are factor variables there are only 2x5x3x2 = 60 possible predicted values, which explains the vertical array. All the vertical bars are centered near zero and spread symmetrically in each direction, indicating constant variance and near zero mean for residuals.



The q‐q plot shows that the normal distribution assumption (for the residuals) is maintained for most values, but not near the extremes. It appears a fatter‐tailed model may do better.



# Task 8 – Interpret the model (9 points)

The gamma model with a log link and four features was fit to the full dataset. The results are:

Coefficients:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | 1.838598 | 0.006812 | 269.913 | < 2e-16 | \*\*\* |
| Rd\_FeatureINTERSECTION-RAMP | 0.533185 | 0.012282 | 43.411 | < 2e-16 | \*\*\* |
| Rd\_ConfigurationONE-WAY | -0.059694 | 0.017914 | -3.332 | 0.000863 | \*\*\* |
| Rd\_ConfigurationTWO-WAY-PROTECTED-MEDIAN | 0.055913 | 0.014045 | 3.981 | 6.89e-05 | \*\*\* |
| Rd\_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN | 0.366560 | 0.010399 | 35.251 | < 2e-16 | \*\*\* |
| Rd\_ConfigurationUNKNOWN | 0.016293 | 0.087165 | 0.187 | 0.851719 |  |
| Time\_of\_DayOVERNIGHT | -0.112799 | 0.023406 | -4.819 | 1.45e-06 | \*\*\* |
| Time\_of\_DayLATE-EARLY | -0.042275 | 0.011383 | -3.714 | 0.000205 | \*\*\* |
| Traffic\_ControlCONTROLLED | 0.072497 | 0.011890 | 6.097 | 1.10e-09 | \*\*\* |

Due to the use of the log link, an appropriate way to interpret coefficients is to exponentiate them and subtract 1. The following table provides that interpretation:

|  |  |  |
| --- | --- | --- |
| Feature | Coefficient | Interpretation |
| Road Feature = INTERSECTION‐RAMP | 0.533 | 70% increase in Crash Score compared to non‐ intersection‐ramp. Crashes at intersections and ramps are likely to involve multiple cars, resulting in higher crash scores. |
| Rd\_Configuration = ONE‐  WAY | ‐0.060 | 6% decrease in Crash Score compared to TWO‐WAY‐  NO‐MEDIAN. Less chance of a head‐on collision. |
| Rd\_Configuration = TWO‐ WAY‐PROTECTED‐MEDIAN | 0.056 | 6% increase in Crash Score compared to TWO‐WAY‐NO‐ MEDIAN. This seems odd as protection should minimize head‐on collisions, but perhaps overall speeds are  higher. |
| Rd\_Configuration = TWO‐ WAY‐UNPROTECTED‐ MEDIAN | 0.367 | 44% increase in Crash Score compared to TWO‐WAY‐ NO‐MEDIAN. This seems odd as a median should reduce head‐on collisions, but perhaps overall speeds are higher. |
| Rd\_Configuration =  UNKNOWN | 0.016 | 2% increase in Crash Score compared to TWO‐WAY‐NO‐  MEDIAN. |
| Time of Day = OVERNIGHT | ‐0.113 | 11% decrease in Crash Score versus daytime (8am to 8pm). It is possible that drivers might be more cautious  at night and there are fewer cars on the road. |
| Time of Day = LATE‐EARLY | ‐0.042 | 4% decrease in Crash Score versus daytime (8am to 8pm). LATE‐EARLY includes times from 4AM to 8AM and 8PM to 12AM. During these times, there are likely fewer cars on the road compared to daytime, leading to lower crash scores. |
| Traffic Control = CONTROLLED | 0.072 | 7% increase in Crash Score versus no control. Traffic controls tend to be used in areas where there is a lot of traffic. Crashes in these areas are likely to involve  multiple vehicles. |

One would expect higher crash scores for crashes associated with multiple vehicles and higher speeds. The model output generally aligns with that intuition, however, the results for Rd\_Configuration are not intuitive without more subject matter expertise.

# Task 9 – Investigate ridge and LASSO regressions (12 points)

There are a variety of methods to reduce overfitting. I previously used an information criterion, BIC, to reduce model complexity. An alternative to reducing the number of variables used is to reduce the coefficients of each variable. This is done by adding a penalty to the loglikelihood that relates to the size of the coefficients. This diminishes the effect, particularly for features that have limited predictive power. There are two approaches to doing this, which come under the general term regularization. (A third approach, a combination of the two, will not be discussed here.) In both cases there is a hyperparameter to estimate that controls the extent of the reduction. This is normally selected through cross‐validation. The specific methods explored are:

* Ridge regression: The penalty is proportional to the sum of the squares of the estimated coefficients. All of the coefficients are reduced but none are reduced to zero. Hence, all variables are retained.
* LASSO regression: The penalty is proportional to the sum of the absolute values of the estimated coefficients. All of the coefficients are reduced and some may be reduced to zero, effectively removing that variable.

Ridge regression is not recommended for this problem. Our goal is to identify variables that best predict Crash\_Score and with all variables retained this approach will not be useful.

LASSO provides an alternative to forward and backward selection for variable selection. One advantage is that through cross‐validation it selects the best hyperparameter using the same criterion (RMSE) that will ultimately be used to judge the model against unseen data.

Regularization methods requires binarization of categorical variables, so unlike the stepAIC performed earlier, which treated all factor levels of one variable as a single object to remove or retain in the model, the LASSO removes individual factor levels if they are not significant with respect to the base level.

In running the two regressions, ridge produced an RMSE of 6.2516 and LASSO produced an RMSE of 6.2463. The LASSO removed 20 factor levels. Note that 10 of them were related to the Month variable.

Based on the above considerations and the fact that neither regularization approach improved the RMSE, I recommend that the original regression be used.

# Task 10 – Consider a decision tree (5 points)

A regression decision tree is an alternative method of linking predictors to a target variable. A tree divides the feature space into a finite, non‐overlapping set of buckets. All observations in a given bucket have the same predicted value. As with GLMs, there are methods to control overfitting such as cost‐ complexity pruning. The advantages of trees relative to GLMs are:

* They can be easier to interpret, provided there aren’t too many buckets. As the name implies, a tree‐like diagram can be constructed that indicates which observations get into the various buckets. Depending on the link function, the coefficients in a GLM may be difficult to explain and interpret.
* Categorical data is automatically handled. There is no need to binarize or determine a base class.

*The following additional advantages also received credit. This list is not exhaustive.*

* + *Interactions are automatically handled. There is no need to identify potential interactions prior to fitting the tree.*
  + *Variables are automatically selected. Some variables simply do not appear in the tree.*
  + *They can produce non‐linear relationships between the predictor variables and the response. (This is a weaker response because for this problem most all variables are categorical and hence linearity is not an issue.)*

Disadvantages relative to GLMs are:

* Even with pruning, there can be considerable overfitting to the training set.
* When underlying data changes, break points for decision trees can change significantly, leading to low user confidence in the model

*The following additional disadvantages also received credit. This list is not exhaustive.*

* + *When fitting a single tree, the locally greedy algorithm is unlikely to find a globally optimal tree.*
  + *With continuous predictors, the bucketing of features means that some small changes can lead to a large change in the prediction while other small changes can lead to no change in the prediction. (This is a weaker response because for this problem most all variables are categorical and hence are already bucketed.)*

# Task 11 – Executive summary (20 points)

To: North Carolina Department of Transportation

From: Actuarial Consulting Firm

Subject: Exploration of factors relating to crash severity