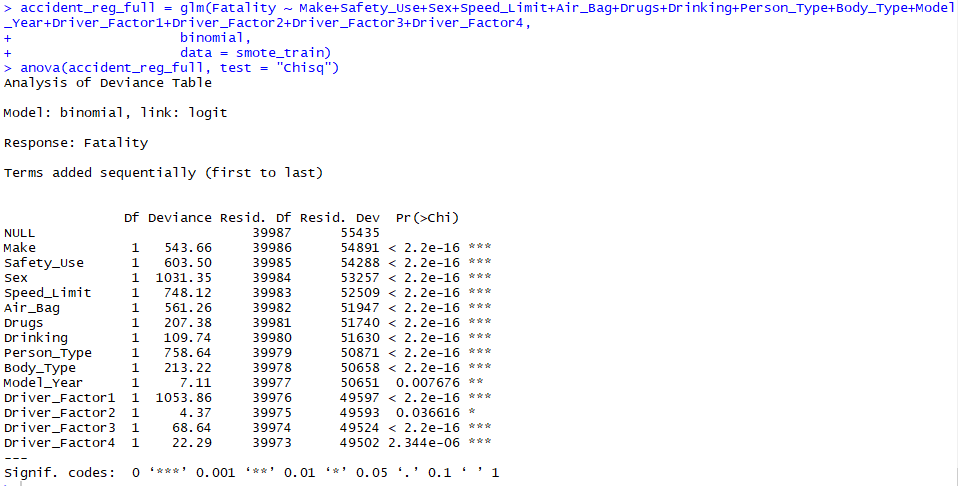
Data Modeling

Logistic Regression Since our objective here is to predict fatalities, which is a binary number

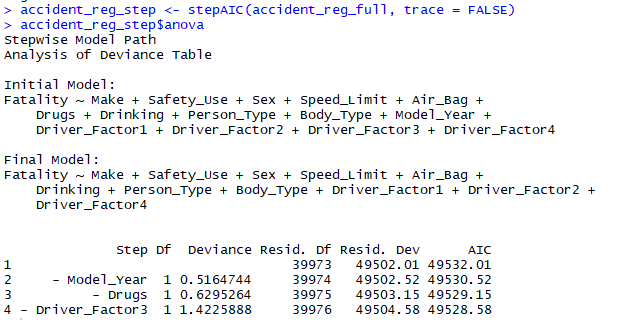
with 0 being the passenger survived and 1 being the passenger did not survive an accident, I have used Logistic Regression.

1. Coefficients and Interpretations

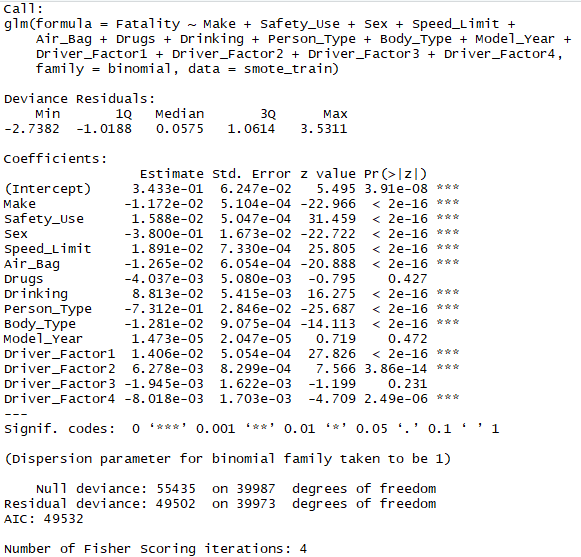
Initial Model Run:



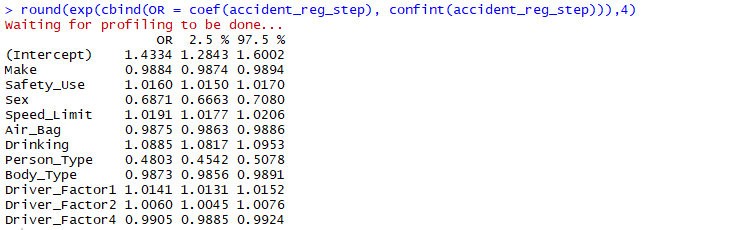
Variables selected by Stepwise Regression:



From the results above, we can see in the anova of the logistical regression that all of the variables except for Driver\_Factor2 and Model\_Year, are significant at the .001 level of significance. However, when we choose a model using stepwise regression, we got a different result. The final model suggested by the stepwise algorithm leaves out Driver\_Factor3, Model\_Year, and Drugs.



Above is the summary of the original logistical regression. We see from the p values (the Pr(>|z|) column), that the logistical regression is similar to our stepwise regression. The main difference is that we can see that State is only significant at the .05 level. Given this result, it makes sense to find the 95% confidence interval for each of the coefficients.

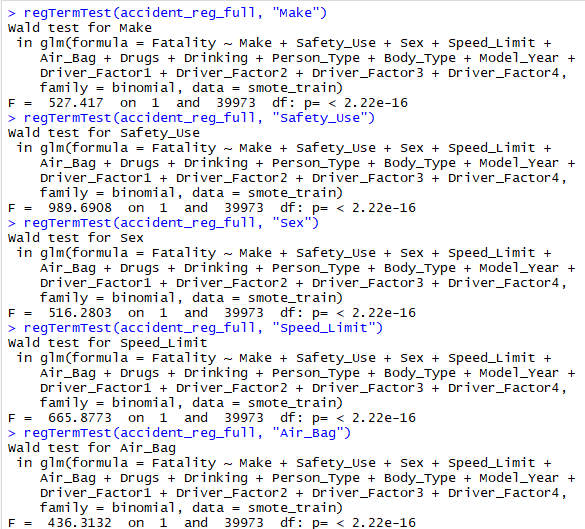


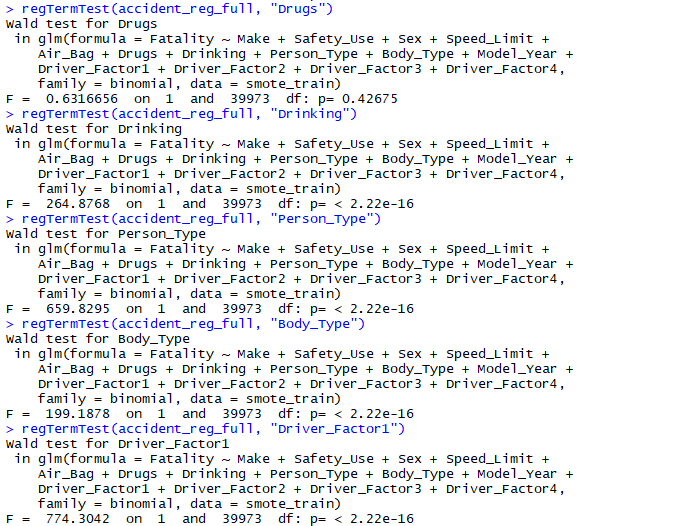
Interpreting the Results

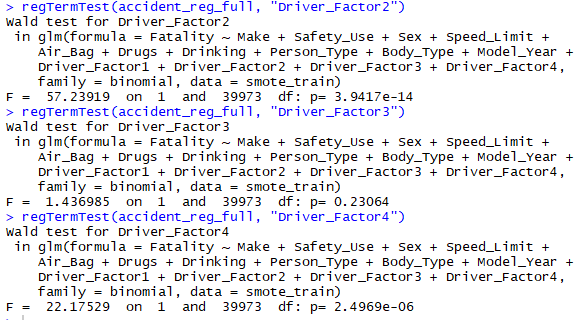
We can see from these results that Drinking has the highest value of 1.0885. Next we have Driver\_Factor1, Speed\_Limit , Safety\_Use, and Driver\_Factor2 as the most significant factors that contribute to a fatality.

Statistical Tests for Individual Predictors

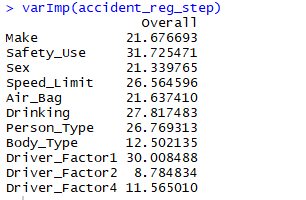
We have used a Wald test to evaluate the statistical significance of each coefficient in the model. As we can see from each of the test, our results mostly line up with earlier tests. This shows our most significant factors are Safety\_Use, Speed\_Limit, and Driver\_Factor1 as still being significant. This also shows that Person\_Type, Driver\_Factor4 and Body\_Type is more significant than in earlier tests as well.







Using another test, we can look at the t-statistic for each parameter. This is done by using the varImp function in the caret package:



As we can see, our most important parameters are Safety\_Use, Driver\_Factor1, Drinking, Person\_Type, and Speed\_Limit.

Model Assessment

In order to assess our model, we will do a goodness of fit test.

Our original equation after stepwise regression was:

glm(formula = Fatality ~ Make + Safety\_Use + Sex + Speed\_Limit +

Air\_Bag + Drugs + Drinking + Person\_Type + Body\_Type + Model\_Year +

Driver\_Factor1 + Driver\_Factor2 + Driver\_Factor3 + Driver\_Factor4,

family = binomial, data = a\_reduc)

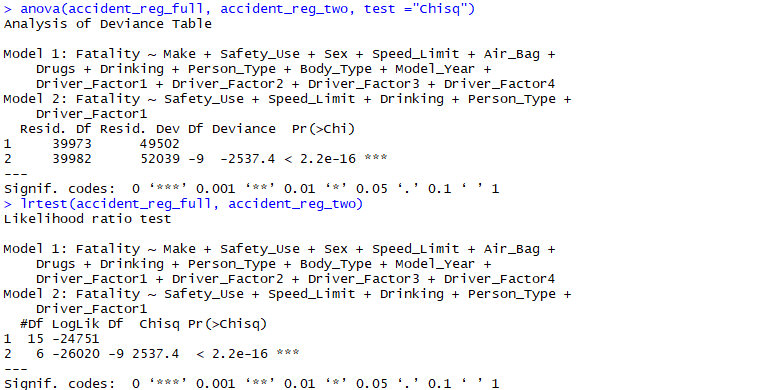
Next, we will take only the factors indicated to perform the best:

glm(formula = Fatality ~ Safety\_Use + Speed\_Limit +

Drinking + Person\_Type +

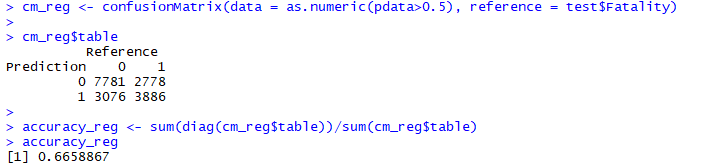
Driver\_Factor1,

family = binomial, data = a\_reduc)



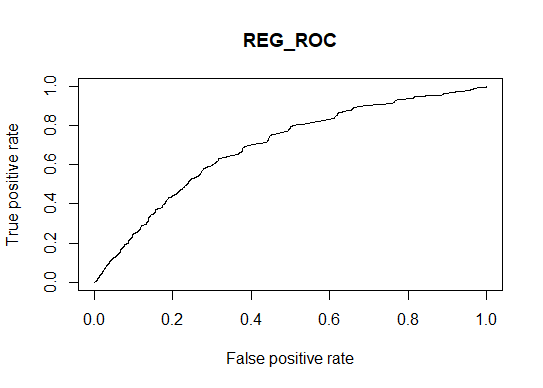
We are using anova and lrtest to compare our two models. As we can see, our new model performs much better with a very significant Pr(>Chi) score.

Moving ahead with our new model, we will test the prediction accuracy of this model:



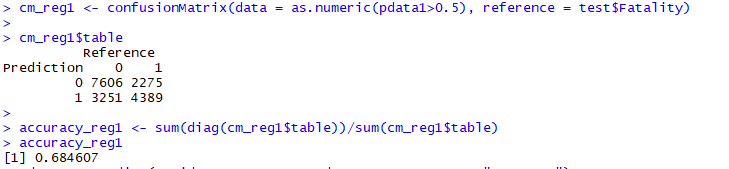
This shows our model is successful 66.6% of the time when attempting to predict fatalities.

ROC Plot



Strengths and Weaknesses

Looking at our two models, we can see that the original model has a slightly better predication rate:



So, we have 68.5% success rate, and our modified model has a 66.6% success rate. The strengths of the original model is we have better predictions. However, the weakness is that we still have large amounts of unnecessary data that isn’t a good predictor of the outcome. In our modified model, all the predictor variables are significant. Yet, this still doesn’t prove to have a good enough prediction rate.

Choice of Model

Given the choice of the two models, we would reluctantly choose the original model. While there is unnecessary data, in the end it has a better prediction rate. This better prediction value would lead us to the conclusion it is the best.