The Application of Logistic Regression to Examine the Predictors of Car Crashes Caused by Alcohol

1 Introduction

Based on the report of the US Department of Transportation, there are approximately 30 people die per day due to the motor vehicle crashes involving an alcohol-impaired driver, and at the same time lots of people get injured during the issues. According to the study of National Highway Traffic Safety Administration in recent years, the alcohol-involved crashes are with an economic impact of more than 59 billion dollars per year, which is relatively significant. This project will use the dataset from the Pennsylvania Department of Transportation, including data of 53,260 car crashes in the City of Philadelphia from 2008 to 2012. The aim of this project is to identify the predictors of accidents related to drunk driving.

In this project, the relationships between the driving behavior predictor variables, neighborhood features, and the involvement of alcohol in crashes of Philadelphia in block group level. More specifically, we want to find out the relationship between the binary dependent variable (DRINKING_D) with the following binary predictors:

- **FATAL_OR_M:** Crash resulted in fatality or major injury (1 = Yes, 0 = No)
- **OVERTURNED:** Crash involved an overturned vehicle (1 = Yes, 0 = No)
- **CELL_PHONE:** Driver was using cell phone (1= Yes, 0 = No)
- **SPEEDING:** Crash involved speeding car (1 = Yes, 0 = No)
- **AGGRESSIVE:** Crash involved aggressive driving (1 = Yes, 0 = No)
- **DRIVER1617:** Crash involved at least one driver who was 16 or 17 years old (1 = Yes, 0 = No)
- DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old

(1 = Yes, 0 = No)

We speculate that all the factors above can be related to the drunk-driving behaviors. Drunk-driving could lead to more significant crash accidents in general, also may be caused by factors including speeding of vehicles, aggressive driving behaviors, and abuse of mobile phones of drivers while driving. As a result, an overturn of vehicle might occur. Furthermore, it is relatively easy and frequent for younger drivers to be affected by the alcohol in some party or dining activities, and it might be difficult for them to figure out the importance of drinking in driving behaviors, and for the old people, it may be hard for them to make the right identification in telling the importance of drinking in driving behaviors. We also take the following continuous predictors into consideration:

- PCTBACHMOR: % of individuals 25 years of age or older who have at least a bachelor's degree in the Census Block Group where the crash took place
- MEDHHINC: Median household income in the Census Block Group where the crash took place

In the analysis of this project, R studio is used to run the logistic regression.

2 Methods

Instead of the OLS regression, the logistic regression is used in this project to assess the relationships between dependent variables and the predictor variables. It is because of the characteristic of the OLS regression that the relationship is based on how much the dependent variable alters if one of the predictor variables is changed by one unit while all other variables are held constant. In this analysis, the dependent variable is binary and coded as 0 and 1, and it can only change from 0 to 1 or from 1 to 0. It is therefore illogical to expect a one-unit change in the predictor variable to result in a beta change in the dependent variable.

2.1 Concept of Logistic Regression

The logistic regression could solve the problem by predicting the log odds of a specific outcome in the dependent variable using a combination of binary and continuous predictor variables. Odds are defined as the probability of an event occurring, and the odds ratio is the ratio of two odds, which can be represented through dividing the probability of an event occurring by 1 minus the probability of that event occurring. The logit model is as follows:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 FATAL_{OR_M} + \beta_2 OVERTURNED \; + \; \beta_3 CELL_{PHONE} \\ + \beta_4 SPEEDING \; + \; \beta_5 AGGRESSIVE \; + \; \beta_6 DRIVER1617 \\ + \; \beta_7 DRIVER65PLUS \; + \; \beta_8 PCTBACHMOR \; + \; \beta_9 MEDHHINC \; + \varepsilon$$

Logistic function:

$$P(DRINKING_D = 1) =$$

$$\frac{1}{1 + e^{-\beta_0 - \beta_1 FATAL_{OR_M} - \beta_2 OVERTURNED - \beta_3 CELL_{PHONE} - \beta_4 SPEEDING - \beta_5 AGGRESSIVE - \beta_6 DRIVER1617 - \beta_7 DRIVER65PLUS - \beta_8 PCTBACHMOR - \beta_9 MEDHHINC}}$$

In the equation above, $p = P(DRINKING\_D = 1)$ is the probability when the dependent variable DRINKING_D is equal to 1, β_0 is the intercept, from β_1 to β_9 are the beta coefficients of the 9 predictors and ε is the residual. Quantity $\frac{p}{1-p}$ is called the odds, $ln\left(\frac{p}{1-p}\right)$ is called the log odds.

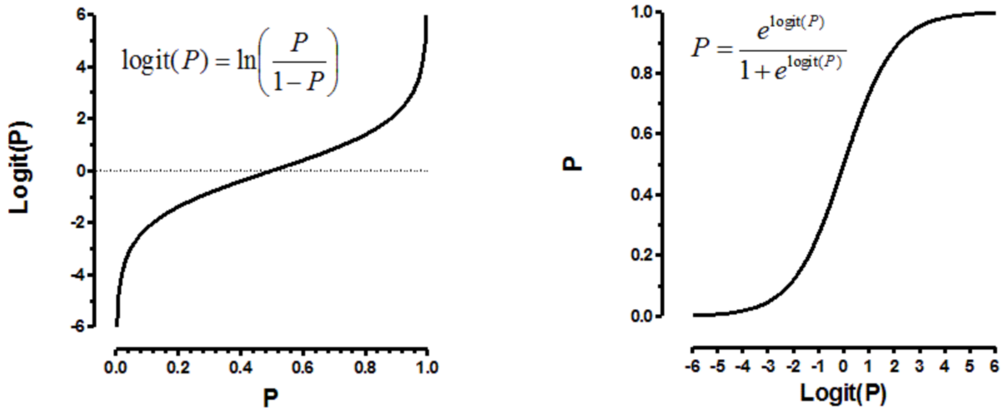
In a linear regression model, the predicted probabilities ($\hat{y}$) can range from -∞ to +∞, while all the probability values must be fall into between 0 and 1. Thus, the logistic function serves as a transformation like:

- The closer the $\hat{y}$ value from our linear regression model is to -∞, the closer our predicted probability is to 0.
- The closer the $\hat{y}$ value from our linear regression model is to +∞, the closer our predicted probability is to 1.

- No predicted probabilities are less than 0 or greater than 1.

The logistic function is the result of solve for $p = P(DRINKING\_D = 1)$ from the logit function. Figure 1 shows both of the two functions, the left one is the logit function and it’s the natural logarithm of the odds. The logit of 0.5 is 0. The logit of any P between 0 and 0.5 is negative, and the logit of any P between 0.5 and 1.0 is positive. The logistic function on the right works well for models where the dependent variable is binary, since the outcome is the probability of an event occurring. It’s more appropriate to explain instead of just a value of 0 or 1.

Figure 1. Graphs of the logit function and the logistic function



2.2 Hypothesis Tested for Each Predictor

For each predictor x_i, we will test the following two hypothesis:

The **Null hypothesis** will be:

$$H_0: \beta_i = 0$$

The **Alternative hypothesis** will be:

$$H_a: \beta_i \neq 0$$

The hypothesis can be translated into the context of the drunk driving analysis, for example, the null hypothesis is that including the predictor FATAL_OR_M into our model doesn't help to increase the odds ratio of drivers being identified as drunk. The alternative hypothesis states that there is a relationship between drinking driver indicator and crash resulted fatality or major injury, and the odds ratio of drivers being identified as drunk is different if there is a brutal crash. The same analogy can be applied to the rest of the predictors in the dataset.

To decide whether we should reject the null hypothesis, the Wald statistic, known as the z-value can be useful because the quantity $\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}$ follows a standard normal distribution with mean equals to 0, and standard deviation equals to 1. As a result, we can take advantage of the Wald statistic quantity by using it to derive p-values for each predictors using the standard normal distribution tables. Since logistic regression is different from OLS regression, the coefficients of each predictor are no longer meaningful without exponentiating the coefficients because this step helps us to derive odds ratios. Odds ratios are greatly adopted by statisticians.

2.3 Assessment of Model's Quality

Once we develop a logistic regression, the process of assessing the quality of the model is different than OLS regression because we are calculating the log odds of whether the driver has consumed alcohol while driving in logistic regression instead of predicting numerical values. In this case, we can no longer relying on the R-squared value to find out the percent of variance that our model can explain; on the other hand, there are other techniques available for us to check the quality of our model such as Akaike Information Criterion, specificity, sensitivity, and misclassification.

Akaike Information Criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. It estimates the relative amount of information that is lost by a given model: the less information a model loses, the higher the quality of that model. Thus, a lower AIC means less information loses, and is indicative of a better fit.

The next critical aspect of the assessing the quality of a model is checking the sensitivity, specificity, and misclassification rates. The definition of sensitivity, also called the true positive rate, measures the proportion of actual positives which are correctly identified. In the context of this assignment is the proportion of actual drivers who are drunk that is correctly identified by our model. On the other hand, specificity, also called true negative rate, tells us the proportion of actual drivers who are not drunk that is correctly identified by the model. Finally, misclassification rate is the proportion of incorrect predictions, including falsely identified drunk drivers who didn't consume alcohol actually and falsely identified not drunk drivers who did consume alcohol. An ideal model we are looking for is a model with higher sensitivity and specificity and lower misclassification rate. To get the input of sensitivity, specificity, and misclassification, we need to get the predicted value of our dependent variable, DRINKING_D. Unlike OLS regression, the predicted value is no longer a discrete value, we are predicting the probability of each driver being drunk or not instead. The equation used to calculate the probability for one of the models is:

$$P(\text{DRINKING\_D} = 1) = \hat{y}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_7 x_{7i}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_7 x_{7i}}}$$

The probability $\hat{y}_i$ is how likely for each observation to be identified as drunk, and we want to observe high probability for drivers who are drunk and low probability who didn't consume any alcohol at all. Determining whether a probability we predicted is considered as drunk driver or the opposite depends on the cut off threshold we choose. If we set the cut-off too high, the number of observations predicted as drunk will be too small, and if it being set too low, we might make a lot of false predictions. It is important to test different cut-offs and compare the results to find an optimal cut-off threshold to achieve the best result in specificity, sensitivity, and misclassification rate.

A useful technique we can use to find an optimal balance between specificity and sensitivity is by examining the ROC curve. ROC curve shows the relationship between sensitivity and false positive rate (1 - specificity), namely sensitivity on the y-axis and false positive rate on the x-axis. The minimum requirement for the predictive power of a model should be producing a curve that is at least to the left of the 45-degree line. Whereas, a perfect model should have the highest sensitivity and lowest false positive rate, which means the curve is extreme close to the top left corner of the plot. There are two ways to find the optimal cut-off based on ROC curves, Youden Index and minimum distance from the upper left corner of the graph. The idea of Youden index is maximizing the sensitivity and specificity of a model, and minimum distance from the upper left corner is similar to Youden index because the values sensitivity and specificity will automatically get large as the curve gets closer to the top left corner. Therefore, we will use the minimum distance from the upper left corner method to find the optimal probability cut-off point in this assignment. Moreover, as the curve get close to the upper left corner, the area under the curve (AUC) will also increase. As AUC increases, the predictive power of our model will also increase because higher sensitivity and specificity make the model become better at predicting whether a driver has consumed alcohol or not. The qualification of a model with excellent accuracy is expect to have an AUC value around 0.9 - 1, an AUC value equals to 0.8 - 0.9 indicates good accuracy, 0.7-0.8 means fair accuracy, 0.6 - 0.7 means poor accuracy, and 0.5 – 0.6 means we fail to develop an adequate model.

2.4 Assumptions of Logistic Regression

Just like some assumptions of OLS, the logistic regression still assume that the observations should be independent and there is no multicollinearity. But unlike OLS regression, a logistic regression’s dependent variable must be binary and doesn’t need the linear relationship between the dependent variable and each predictor. In addition, the residuals’ distribution is no need to be normal and there is no assumption of homoscedasticity.

2.5 Exploratory Analyses

Before running a logistic regression on a dataset, most statisticians may want to do some exploratory analyses. To find out whether there is an association between the dependent variable and binary predictors, we will run the cross-tabulations. Then, we will conduct the Chi-Squared test to examine whether there is an association between each binary predictor and the binary dependent variable, which is whether alcohol was involved in a crash. The two types of data are both categorical variables. The null hypothesis states that the proportion of fatalities for crashes that involve drunk drivers is the same as the proportion of fatalities for crashes that don't involve drunk drivers. The alternative hypothesis states that the proportion of fatalities for crashes that involve drunk drivers is different from the proportion of fatalities for crashes that don't involve drunk drivers.

For the continuous predictors, we can compare the means of each predictor for both values of the binary dependent variable. Then we use the independent samples t-tests to examine whether there were significant differences in mean values of **PCTBACHMOR** and **MEDHHINC** for crashes that involved alcohol and those that didn't. The null hypothesis is that there is no difference in the mean values of each continuous predictor whether or not alcohol is involved. The alternative hypothesis is that there is a difference in the mean values of each continuous predictor whether or not alcohol is involved.

3 Results

3.1 Exploratory Analyses Results

Table 1. Tabulation of the number and proportion of crashes that is involved and not involved drunk driving

	No Alcohol	**Alcohol Involved**
Number of Crashes	40879	2485

Percentage	94.3%	5.7%

Table 1. above shows the number and proportion of crashes that involves drunk driving and not involves. The number of crashes with no alcohol is nearly 16 times greater than those with alcohol.

Table 2. Cross tabulation between alcohol involvement and factors that may be related to a crash with percentage differences and their significance

	No Alcohol (DRINKING_D = 0)		**Alcohol Involved (DRINKING_D = 1)**			
	Number	Percentage	Number	Percentage	Total χ^2	P-value
FATAL_OR_M	1181	2.9	188	7.6	1369	<0.0001
OVERTURNED	612	1.5	110	4.4	722	<0.0001
CELL_PHONE	426	1.0	28	1.1	454	0.69
SPEEDING	1261	3.1	260	1.05	1521	<0.0001
AGGRESSIVE	18522	45.3	916	36.9	19438	<0.0001
DRIVER1617	674	1.6	12	0.5	686	<0.0001
DRIVER65PLUS	4237	10.4	119	4.8	4356	<0.0001

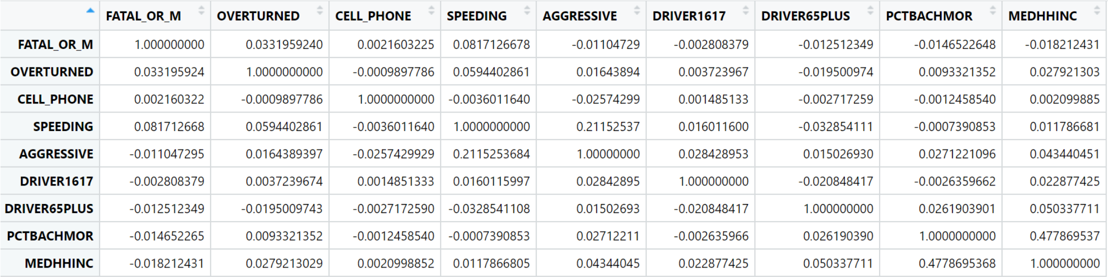
Table 2. above shows that all the predictors except the CELL_PHONE usage have χ^2 p-values lower than the 0.05 cutoffs, which means there is a significant association between the alcohol involvement and all the predictors except for cell phone usage. We can reject the Null Hypothesis and in favor of the alternative hypothesis.

Table 3. Table examining whether the means of the two continuous predictors seem to differ for the different levels of the dependent variable

	No Alcohol (DRINKING_D = 0)		**Alcohol Involved (DRINKING_D = 1)**		**t-test**
	Mean	**SD**	**Mean**	**SD**	**P-value**
PCTBACHMOR	16.57	18.21	16.61	18.72	0.9137
MEDHHINC	31483.05	16930.1	31998.75	17810.5	0.16

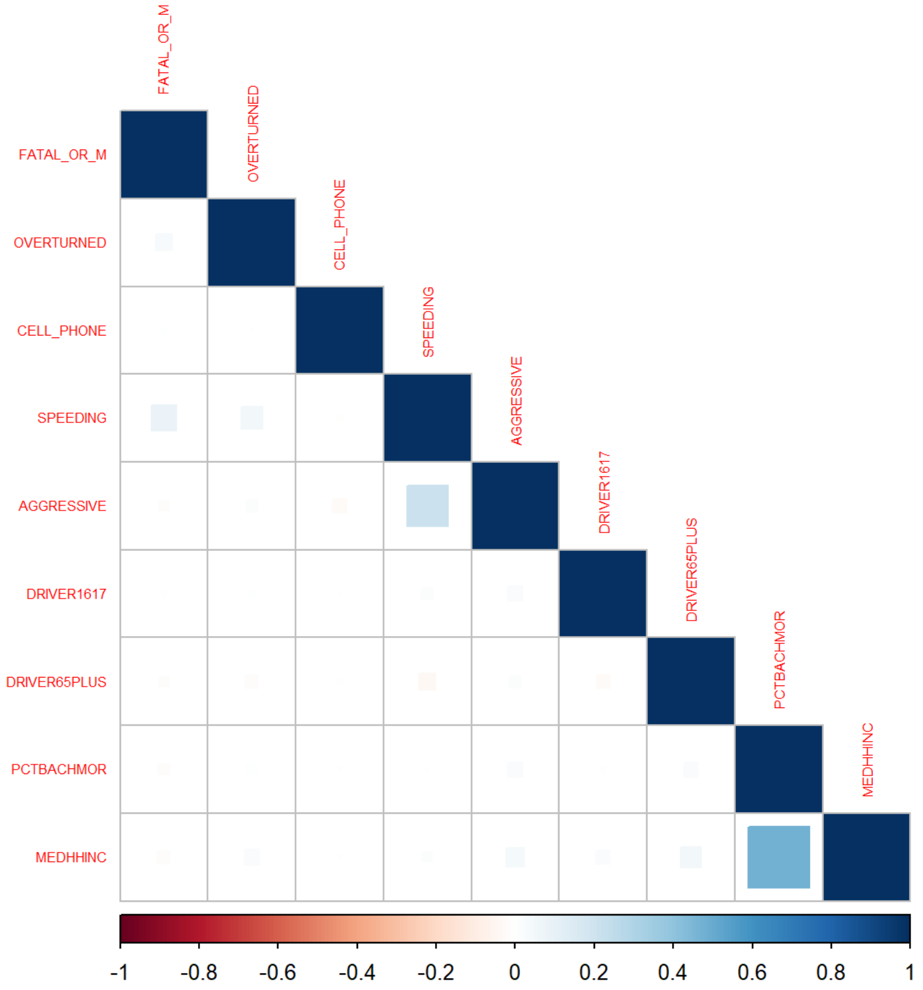
From the table 3, we can know that both of the two continuous predictors' P-value is larger than 0.05 from the comparison of the mean and standard deviation values, which means we can’t reject the Null Hypothesis that there is no significant association between the dependent variable and percentage of people with bachelor’s degrees or median household income.

Table 4. Pearson correlations matrix between all the variables



	FATAL_OR_M	OVERTURNED	CELL_PHONE	SPEEDING	AGGRESSIVE	DRIVER1617	DRIVER65PLUS	PCTBACHMOR	MEDHHINC
FATAL_OR_M	1.000000000	0.0331959240	0.0021603225	0.0817126678	-0.01104729	-0.002808379	-0.012512349	-0.0146522648	-0.018212431
OVERTURNED	0.033195924	1.0000000000	-0.0009897786	0.0594402861	0.01643894	0.003723967	-0.019500974	0.0093321352	0.027921303
CELL_PHONE	0.002160322	-0.0009897786	1.0000000000	-0.0036011640	-0.02574299	0.001485133	-0.002717259	-0.0012458540	0.002099885
SPEEDING	0.081712668	0.0594402861	-0.0036011640	1.0000000000	0.21152537	0.016011600	-0.032854111	-0.0007390853	0.011786681
AGGRESSIVE	-0.011047295	0.0164389397	-0.0257429929	0.2115253684	1.00000000	0.028428953	0.015026930	0.0271221096	0.043440451
DRIVER1617	-0.002808379	0.0037239674	0.0014851333	0.0160115997	0.02842895	1.000000000	-0.020848417	-0.0026359662	0.022877425
DRIVER65PLUS	-0.012512349	-0.0195009743	-0.0027172590	-0.0328541108	0.01502693	-0.020848417	1.000000000	0.0261903901	0.050337711
PCTBACHMOR	-0.014652265	0.0093321352	-0.0012458540	-0.0007390853	0.02712211	-0.002635966	0.026190390	1.0000000000	0.477869537
MEDHHINC	-0.018212431	0.0279213029	0.0020998852	0.0117866805	0.04344045	0.022877425	0.050337711	0.4778695368	1.000000000

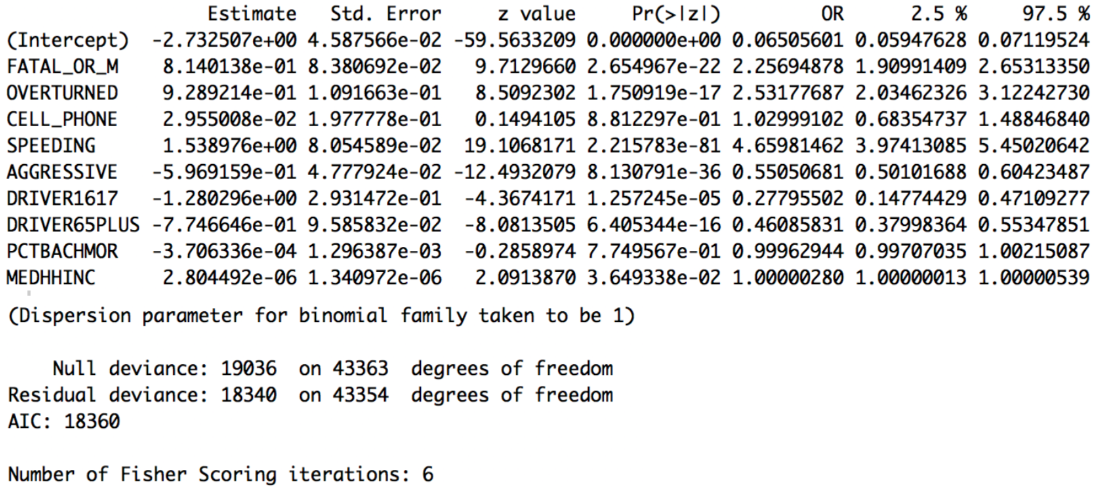
Figure 1. Pearson correlations matrix between all the variables



From the table and figure, we can find that only a few predictors have multicollinearity between. Multicollinearity is defined as several independent variables in a model are correlated or tend to vary with other variables. The Pearson correlation is a measure of linear correlation between two sets of data. The value ranges from -1 to 1. A value of 0 means there is no correlation, 1 means total positive correlation, -1 means total negative correlation. The Pearson correlation used between two binary is not appropriate, it's based on the assumption that variables are continuous. For the binary data, Spearman's correlation may be a better choice. The strongest correlation is between median household income and percent with bachelor's degrees, with a value of 0.48. The second one is between SPEEDING and AGGRESSIVE driving, with a correlation of 0.21. The other correlations are all around 0, which means no strong relationship.

3.2 Logistic Regression Results

Table 5. Logistic regression output results including all the predictors



```
                 Estimate   Std. Error     z value     Pr(>|z|)         OR      2.5 %     97.5 %
(Intercept)  -2.732507e+00 4.587566e-02 -59.5633209 0.000000e+00 0.06505601 0.05947628 0.07119524
FATAL_OR_M    8.140138e-01 8.380692e-02   9.7129660 2.654967e-22 2.25694878 1.90991409 2.65313350
OVERTURNED    9.289214e-01 1.091663e-01   8.5092302 1.750919e-17 2.53177687 2.03462326 3.12242730
CELL_PHONE    2.955008e-02 1.977778e-01   0.1494105 8.812297e-01 1.02999102 0.68354737 1.48846840
SPEEDING      1.538976e+00 8.054589e-02  19.1068171 2.215783e-81 4.65981462 3.97413085 5.45020642
AGGRESSIVE   -5.969159e-01 4.777924e-02 -12.4932079 8.130791e-36 0.55050681 0.50101688 0.60423487
DRIVER1617   -1.280296e+00 2.931472e-01  -4.3674171 1.257245e-05 0.27795502 0.14774429 0.47109277
DRIVER65PLUS -7.746646e-01 9.585832e-02  -8.0813505 6.405344e-16 0.46085831 0.37998364 0.55347851
PCTBACHMOR   -3.706336e-04 1.296387e-03  -0.2858974 7.749567e-01 0.99962944 0.99707035 1.00215087
MEDHHINC      2.804492e-06 1.340972e-06   2.0913870 3.649338e-02 1.00000280 1.00000013 1.00000539

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19036  on 43363  degrees of freedom
Residual deviance: 18340  on 43354  degrees of freedom
AIC: 18360

Number of Fisher Scoring iterations: 6
```

As shown in the regression output above, all the predictors appeared to be significant based on the extremely small p-values except these three predictors: CELL_PHONE, PCTBACHMOR, and MEHHINC. The three predictors had a high p-value each, which prevented us from rejecting the null hypothesis. Median household income seems to be significant, but only at the 0.01 level. Consequently, we can say that for each predictor changing from 1 category to the other or increasing by 1 unit will affect the odds of drunken driving. For each predictor:

If there are fatalities involved in an accident while holding other predictors in the model unchanged, the odds of drunken driving is going to be $e^{\beta_1} = e^{0.8140138} = 2.26$. Alternatively, we can say that if there is a fatality, the odds of drunk driving go up by $(e^{\beta_1} - 1) * 100\% = 126\%$ than that of accidents without fatalities.

If the accident involves overturned vehicles without changing other predictors, the odds of drunken driving is going to be $e^{\beta_2} = e^{0.9289214} = 2.53$ times the odds than drunk driving without overturned vehicles. Alternatively, the odds of drunk driving go up by $(e^{\beta_2} - 1) * 100\% = 153\%$ than that of accidents without overturned vehicles.

If the driver involved in an accident because using cellphone, the odds of drunken driving is $e^{\beta_3} = e^{0.02955} = 1.029$ times the odds of drunken driving without using cell phone. There is an $(e^{\beta_3} - 1) * 100\% = 2.99\%$ higher odds than without using cell phone while other predictors stay constant.

If the driver was speeding and involved in an accident, the odds is $e^{\beta_4} = e^{1.5389} = 4.659$ times the odds that accident related to drunken driving. The odds of drunken driving will go up by 365.9% if speeding involved when other predictors are unchanged.

If the driver was driving aggressively and involved in an accident, the odds is $e^{\beta_5} = e^{-0.5969} = 0.5505$ times the odds that accident related to drunken driving. The odds of drunken driving will decrease by 44.96% if aggressive driving behavior involved when other predictors are unchanged.

If the crash involved at least one driver who was 16 or 17 years old while other predictors unchanged, the odds of drunken driving will be $e^{\beta_6} = e^{-1.2802} = 0.2779$ times the odds that 16 - 17 years old drivers involved accidents caused by drunken driving. Said differently, the odds will decrease by 72.20%.

If the crash involved at least one driver who was 65 years old and above while other predictors unchanged, the odds of drunken driving will be $e^{\beta_7} = e^{-7.7466} = 0.00043$ the odds that 65 years old and above drivers involved accidents caused by drunken driving. Said differently, the odds will decrease by 99.956%.

For 1% increase in the bachelor's degree population and other predictors being constant, the odds of accident caused by drunken driving is $e^{\beta_8} = e^{-0.0037} = 0.99963$ times the odds that no increase in bachelor's degree in population. Said alternatively, the odds will decrease by 0.0369% for 1 % increase in bachelor's degree population.

For 1% increase in the median household income and other predictors being constant, the odds of accident caused by drunken driving is almost the same odds as before.

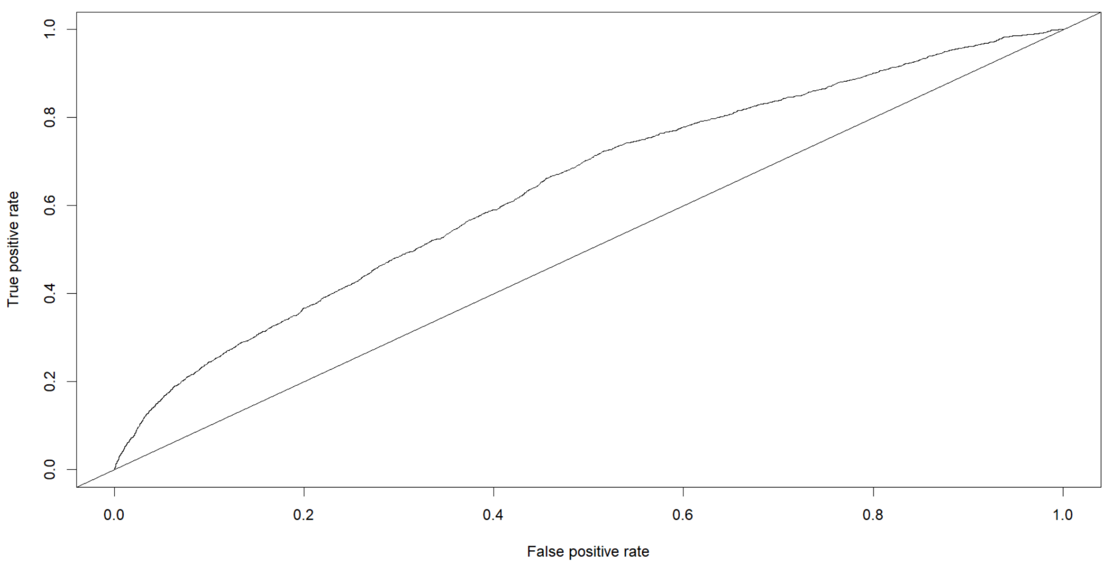
Table 6. Specificity, Sensitivity, and Misclassification rates

Cut-off Value	**Sensitivity**	**Specificity**	**Misclassification Rate**
0.02	0.984	0.058	0.889
0.03	0.981	0.064	0.884
0.05	0.735	0.469	0.516

0.07	0.221	0.914	0.126
0.08	0.185	0.939	0.105
0.09	0.168	0.946	0.097
0.1	0.164	0.948	0.097
0.15	0.104	0.972	0.078
0.2	0.023	0.995	0.060
0.5	0.002	0.999	0.057

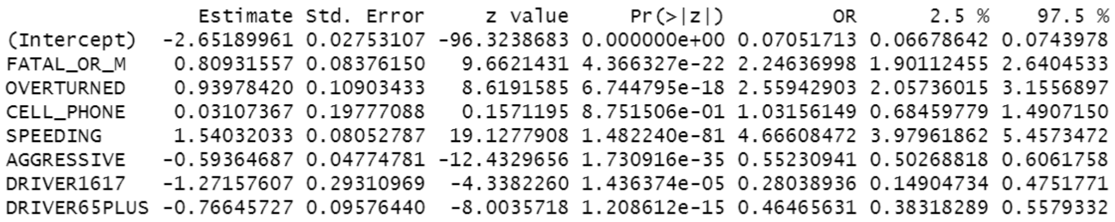
The sensitivity, specificity, misclassification rate calculated from different cut-off values shows that the highest misclassification rate is generated by cut-off value equals to 0.02 (misclassification rate = 0.889), and the lowest misclassification rate is generated by cut-off value equals to 0.5 (misclassification rate = 0.057).

Figure 2. ROC Curve



The ROC curve above shows that the optimal cut-off rate is about 0.06365, which is that is selected by minimizing the distance from the upper left corner of the ROC curve. And it leads to model sensitivity equal to 0.6607 and specificity equal to 0.5452. Moreover, the area under the curve is 0.6398, and the model did do a poor job in terms of accuracy referring to the AUC values guide.

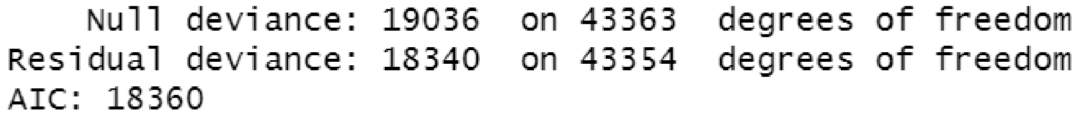
Table 7. Results of the logistic regression with only binary predictors



```
                Estimate Std. Error     z value     Pr(>|z|)         OR      2.5 %    97.5 %
(Intercept)  -2.65189961 0.02753107 -96.3238683 0.000000e+00 0.07051713 0.06678642 0.0743978
FATAL_OR_M    0.80931557 0.08376150   9.6621431 4.366327e-22 2.24636998 1.90112455 2.6404533
OVERTURNED    0.93978420 0.10903433   8.6191585 6.744795e-18 2.55942903 2.05736015 3.1556897
CELL_PHONE    0.03107367 0.19777088   0.1571195 8.751506e-01 1.03156149 0.68459779 1.4907150
SPEEDING      1.54032033 0.08052787  19.1277908 1.482240e-81 4.66608472 3.97961862 5.4573472
AGGRESSIVE   -0.59364687 0.04774781 -12.4329656 1.730916e-35 0.55230941 0.50268818 0.6061758
DRIVER1617   -1.27157607 0.29310969  -4.3382260 1.436374e-05 0.28038936 0.14904734 0.4751771
DRIVER65PLUS -0.76645727 0.09576440  -8.0035718 1.208612e-15 0.46465631 0.38318289 0.5579332
```

The results of the reduced model shown here are same as the first model we developed, predictor CELL_PHONE had a p-value of 0.875, which prevented us from rejecting the null hypothesis that predictor doesn't help to increase the odds ratio of drivers being identified as drunk and involved in an accident.

It is time to choose which model we are going to keep; however, both models showed exactly same AIC value, 18360, which is hard to tell whether the full model or the reduced model is better without more advanced techniques to verify that. But dropping the continuous predictors didn't result in any change in AIC, so we might use the simple model with just binary predictors.



```
    Null deviance: 19036  on 43363  degrees of freedom
Residual deviance: 18340  on 43354  degrees of freedom
AIC: 18360
```

4 Discussion

The drunk driving could cause great loss of economy and could threaten the safety of individuals. In this project, we have explored the relationship between certain neighborhood features, the driving behavior and the occurrence of alcohol in motor crashes of Philadelphia in block group level. We use the logistic regression model to predict the odds of the involvement of alcohol in a crash, and take fatalities, cellphone usage, speeding, aggressive driving and some other characteristics into consideration.

According to the derived result of this project, the factors including the speeding driving, aggressive driving, overturned vehicles, fatality, 16/17 years old drivers, drivers older than 65 years old, and median household values are relatively strong predictors of drunken driving. Also, the abuse of cellphones by drivers and the percentage of individuals with a bachelor's degree are factors less related to the occurrence of the alcohol crashes.

The results are not surprising, for the reason that the input of alcohol could significantly affect the driving behaviors, and as a result the speeding and aggressive driving could occur, and even the overturned cars in serious crashes. The model works appropriately in this project.

Paul Allison, a leading expert on logistic regression pointed our that substantial bias may occur under the circumstance that the possibility of a small number of cases on the rarer of the two outcomes for the dependent variable. In our dataset, the number of drunk driving is 2485 out of 43364 total motor crashes. The sample is good enough for us to conduct logistic regression.

The limitation of this model is probably the blank of other types of data that relevant to the dependent variable, which are highly statistically associated with the dependent variables that are absence from this project. It can be concluded that our model's performance is poor due to the low AUC score of 0.66.