Examine the Predictors of Car Crashes Caused by Alcohol

Introduction

Drunk driving is a dangerous criminal behavior and remains a prominent cause of traffic crashes, injuries and death (Hedlund and McCartt 2021). Everyday about 30 people die in traffic accidents caused by a alcohol-impaired drivers in the United States. This happens every 50 minutes and the property loss is more than $44 billion annually (CDC Injury Center, 2021). Although accidents caused by drunk drivers have decreased in recent years due to serious regulation and laws, it is still important to understand the indicator measures of drunk driving to prevent such controllable accidents from happening.

Under this premise, this study explores the predictors of accidents related to drunk driving in the City of Philadelphia for the years 2008 – 2012. There are several potential predictors in terms of drivers' behavior, census block group information and the features of crash. Drunk drivers become more unconscious and easily distracted, so they might use cell phones when driving, drive cars at a higher speed or drive aggressively, and the crash tends to involve overturned vehicles, fatality, or major injury. The age of the driver is also considered as one predictor, since young drivers have less experience and ignore the laws easily, while old drivers are more easily affected by alcohol. The economic and educational condition of a census block group might also predict the crashes, as we assume that accidents are more likely to happen in poorer and less educated regions. This study will utilize logistic regression to examine these predictors of car crashes caused by alcohol in R.

Methods

In our previous analysis, continuous dependent variables are included in OLS regression. The equation of OLS regression model is as follows:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon \tag{1}$$

Where β_1 is interpreted as the amount by which the dependent variable Y changes when predictor x_1 increases by 1 unit.

However, when Y is a binary variable, which means a categorical variable that only includes one of two values ("Binary Variable – Learndatasci" 2021), there will be some problems with using OLS regression. For example, if Y, binary variable that represents the presence of hospital, is either 0 or 1, where 1 represents there's a hospital while 0 doesn't, it is informative to say that a 1 unit increase in x1 results in a β_1 increase in Y makes no sense, since Y can only change from 0 to 1 or 1 to 0.

Under this circumstance, logistic regression is considered to fit a binary response model. Log odds play an important role in logistic regression, where odds are defined as the ratio between the probability of desirable outcomes and the probability of undesirable outcomes, or between the undesirable outcomes and the probability of desirable outcomes. The odds ratio is defined as the ratio of odds of event A given that event B happens and the odds of event A given that event B doesn't happen. It represents how many times A happens given that event B does not happen. The regression equation for the logit model with multiple predictors is as follows:

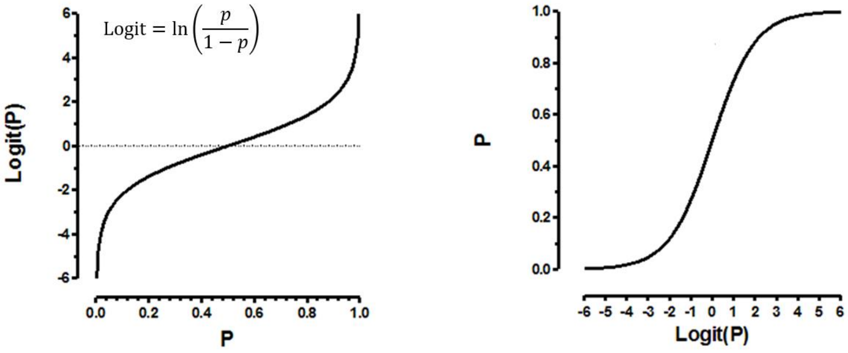
$$Logit(p) = Logit\big(P(Y=1)\big) = l\big(Odds(Y=1)\big) \; = ln \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots \beta_9 x_9 + \varepsilon \qquad (2)$$

Where:

- Y=1 represents that the crash involves a drinking driver, as Y is a binary variable that symbolizes the drinking driver indicator (1 = Yes, 0 = No).
- $p = P(Y = 1)$ represents the probability that the crash involves a drinking driver.
- $Odds(Y = 1) = \frac{p}{1-p}$ represents the odds of the crash which involves a drinking driver.
- X1, …, x9 represent a series of independent variables, including FATAL_OR_M (whether crash resulted in fatality or major injury), OVERTURNED (whether crash involved an overturned vehicle), CELL_PHONE (whether driver was using cell phone), SPEEDING (whether crash involved speeding car), MEDHHINC (Median household income in the Census Block Group where the crash took place), AGGRESSIVE (whether crash involved aggressive driving) ,DRIVER1617 (whether crash involved at least one driver who was 16 or 17 years old), DRIVER65PLUS (Crash involved at least one driver who was at least 65 years old). The DRIVER1617 and DRIVER65PLUS are binary predictors while others are continuous predictors.
- βi is the coefficient of the independent variable X_i. β_0 denotes the intercept and ε denotes the residual.

Following the equation above, the logit function is defined as $Logit(p) = ln \left(\frac{p}{1-p}\right)$, as figure 1 (a) shows, the logarithm of odds ratio. The inverse of logit function is called logistic function, a S-shaped curve as figure 1 (b) shows. Its equation is as follows:

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7 + \varepsilon}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7 + \varepsilon}} = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_7 x_7 - \varepsilon}} \qquad (3)$$



(a) (b)

Figure 1. logit and logistic function

Logistic function is a function that transforms the log-odds to the probability. It is often used to estimate the parameters of a logistic model, where the dependent variables are binary and the predictors can be continuous or binary. The value of p, which ranges from 0 to 1, is depended by the value of $\beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7 + \varepsilon$. According to the equation (3), when $\beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7 + \varepsilon$ gets really big, p is close to 1, otherwise it approaches 0 when $\beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7 + \varepsilon$ become small. Clearly logistic function solves the problems of the OLS regression model, which is likely to give predicted values that are beyond range (0,1), and works well for models where the dependent variable is binary.

When assessing contributions of each of the predictors, people use the Wald Statistics, a quality denoted as $\frac{\widehat{\beta}_i}{\sigma_{\widehat{\beta}_i}}$, to determine predictors' significance. In Wald test, we assume that the statistic we are looking at is in fact an estimator $\widehat{\theta}$ of a parameter θ, that is unbiased and asymptotically normal. Under the null hypothesis, the outcome (T_{obs}) is less than 1.96 with the probability of 0.95, thus we could reject the null hypothesis when T_{obs} is larger than 1.96 (Ramesh).

Odds ratio measures the change-per-one-unit based on the logistic regression we calculated by exponentiating the coefficients (Szumilas, 2010). In practice, most statisticians favor the odds ratio rather than the β coefficients.

For this study, the following hypotheses are proposed:

- H_0: $\beta_i = 0$ ($OR_i = 1$)
- H_a: $\beta_i \neq 0$ ($OR_i \neq 1$)

The R-Squared value is a common way to assess the generalizability and accuracy of the model. However, for the logistic regression model where the dependent variable is binary. Thus for this study, we are going to use Akaike Information Criterion (AIC) to compare models. It is calculated based on 1) the number of independent variables used to build the model and 2) the maximum likelihood estimate of the model. Usually, the lower AIC value is, the better the model fit.

Specificity, Sensitivity and Misclassification Rate are the three indicators to assess the model. Sensitivity, also called the true postive rate, meausres the proportion of actual postives which are correctly identified as such, while specificity, known as true negative rate, measures the proportion of negatives which are correctly identified as such. Specificity and Sensitivity are two complementary indicators that measure the proportion of the actual false negative rate and the actual true positive rate respectively. While the Misclassification Rate measures the total fraction of the predictions that were wrong. The formula for calculating the Specificity, Sensitivity and Misclassification Rates are listed below:

$$Sensitivity\ = \frac{Number\ of\ True\ Positives}{Number\ of\ True\ Positives\ +\ Number\ of\ False\ Negatives}$$

$$Specificity\ = \frac{Number\ of\ True\ Negatives}{Number\ of\ True\ Negatives\ +\ Number\ of\ False\ Positives}$$

$$Misclassification\ Rate = \frac{Number\ of\ False\ Positive\ +\ Number\ of\ True\ Negatives}{Number\ of\ True\ Positives\ +\ TrueNegatives\ +\ False\ Positives\ +\ False\ Negatives}$$

We use residuals to assess the model fit in linear regression, the same logic applies to logistic regression as well based on the formula $\varepsilon = y_1 - \hat{y}_i$. However, the $\widehat{y_i}$ here indicates the probability that Y =1. Thus, we use the following formula:

$$P_{(y=1)} = \hat{y}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1\cdot x_{1i}+\ldots+\hat{\beta}_\varepsilon\cdot x_{\varepsilon i}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1\cdot x_{1i}+\ldots+\hat{\beta}_\varepsilon\cdot x_{\varepsilon i}}}$$

Where we want our model can predict a high probability of Y=1 if Y is actually 1 and low probability of Y-1 when Y is 0.

Specificity (false positive rate) in relation to the sensitivity (true positive rate) at different cut-off rates from 0 to 1. There are a few different ways to identify the optimal cut-off rate based on the ROC curves: the first is to find a point where the sum value of sensitivity and specificity is maximized, this is also called as Youden Index; another way is to find points where has the minimum distance to the from the upper left corner of the graph, in which specificity = 1 and sensitivity = 1. In this report, we are going to use the second way to identify the optimal cut-off value.

By calculating the Area under the ROC Curve, which is known as AUC, we can measure the accuracy of the model. The AUC usually ranges between 0.5 and 1. Basically, the higher AUC indicates a better model, where both sensitivity and specificity is relatively high. The following numbers are the common guide to assess the accuracy:

- .90-1 = excellent
- .80-.90 = good
- .70-.80 = fair
- .60-.70 = poor
- .50-.60 = fail

In fact, some statisticians would argue that they would accept the models with the AUC > 0.7. The assumptions of OLS regression and Logistic Regression share a lot in common as below: independence of observations and no multicollinearity. OLS regression's assumptions also include linear relationship between dependent variable (DV) and each predictor, homoscedasticity and no multicollinearity. The unique assumptions of Logistic Regression require no multicollinearity, binary dependent variables as well as Larger samples.
Pearson correlations are used to look at associations between two continuous (ideally normal) variables and Chi-squares are used to test associations between two categorical variables). We examined the Pearson correlations between all the predictors (both binary and continuous) for assessing multicollinearity. Cross-tabulation is computed along with the Chi-square analysis, which helps identify if the variables of the study are independent or related to each other. The null hypothesis is that there is no association between the dependent and binary predictors, while alternative hypothesis states that there is association between these two variables.
Since we want to look at an association between continuous variables (PCTBACHMOR and MEDHHINC) and binary variables (DRINKING_D), we use independent samples t-tests here. Looking at Table 2, mean values of PCTBACHMOR and MEDHHINC are not significantly associated with each other since their p-values are larger than 0.01. In that case, we failed to reject the null hypothesis for the alternative hypothesis.

Results

There are 40879 crashes involving non-drinking drivers, which accounts for 94.27% of the sum, and the remaining number of crashes is 2485, which involves drinking drivers, occupying 5.73%.

	No Alcohol Involved		**Alcohol Involved**		**Total**	**χ^2 p-value**
	(DRINKING_D = 0)		**(DRINKING_D = 1)**			
	N	**%**	**N**	**%**	**N**	
FATAL_OR_M: Crash resulted in fatality or major injury	1181	2.90%	188	7.60%	1369	2.522202e-38
OVERTURNED: Crash involved an overturned vehicle	612	1.50%	110	4.40%	722	1.551762e-28
CELL_PHONE: Driver was using cell phone	426	1.0%	28	1.10%	454	0.6872569
SPEEDING: Crash involved speeding car	1261	3.10%	260	10.5%	1521	6.249562e-84
AGGRESSIVE: Crash involved aggressive driving	18522	45.3%	916	36.9%	19438	2.00079e-16
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old	674	1.6%	12	0.5%	686	6.115619e-06
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old	4237	10.4%	119	4.8%	4356	2.75703e-19

Table 1. the cross-tabulation of the dependent variable with each of the binary predictors

Table 1 shows the cross-tabulations of the dependent variable with each of the binary predictors. Beside CELL_PHONE, all the binary predictors' p-values are smaller than 0.05, then we reject the null hypothesis for alternative hypothesis: the dependent variable is associated with the predictors.

	No Alcohol Involved		**Alcohol Involved**		**t-test p-value**
	(DRINKING_D = 0)		**(DRINKING_D = 1)**		
	Mean	**SD**	**Mean**	**SD**	
PCTBACHMOR: % with bachelor's degree or more	16.56986	18.21426	16.61173	18.72091	0.9137
MEDHHINC: Median household income	31483.05	16930.1	31998.75	17810.5	0.16

Table 2. the means of the continuous predictors for both values of the dependent variable

Table 2 shows whether the means of the two continuous predictors seem to differ for the different levels of the dependent variable. Both the P-values of these continuous predictor, the PCTBACHMOR and MEDHHINC, are greater than 0.05, suggesting that we fail to reject null hypothesis: there is no significant association between the dependent variable and each of the continuous predictors.

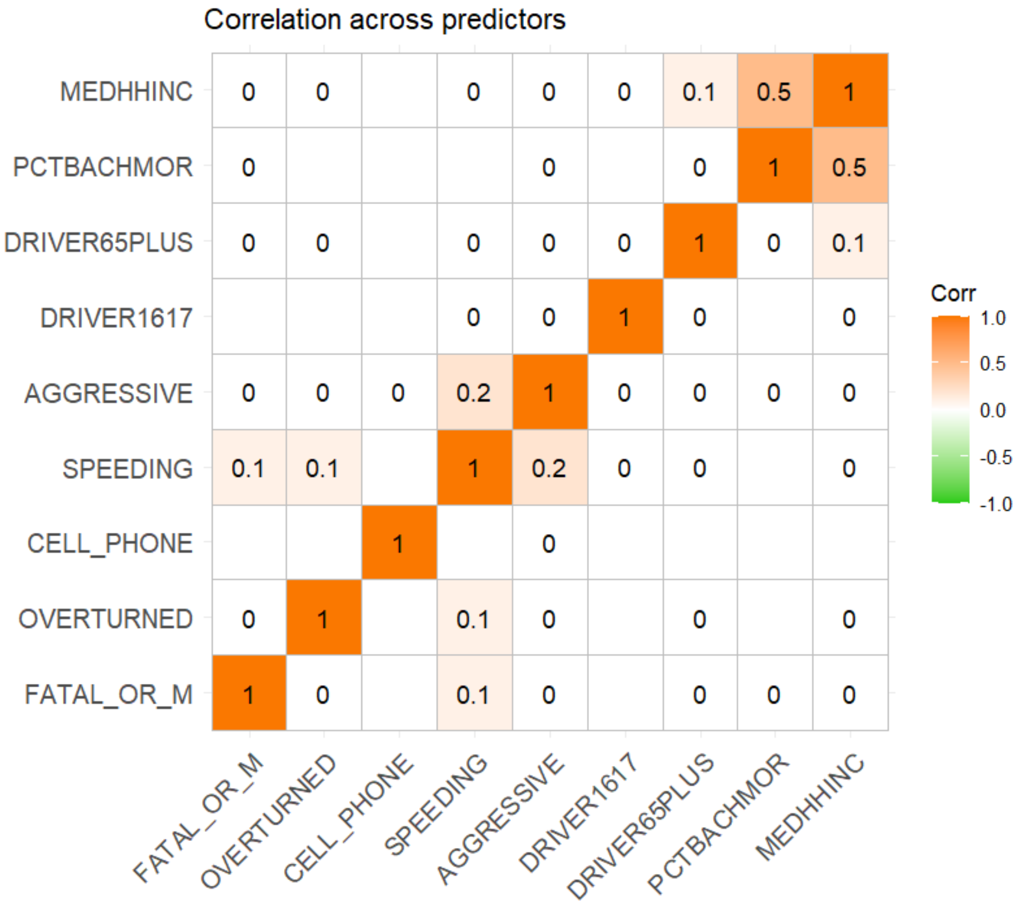


Figure 2. the pairwise Pearson correlations for all the binary and continuous predictors

One potential limitation of using Pearson correlations is that it assumes a linear relationship between DV and each predictor and no multicollinearity exists between predictors. Pearson's correlations are most ideally suited when you have multivariate normality. Within the context of regression analysis, there is assumption that no multicollinearity should exist and very strong linear association between two predictors should not exist. So Pearson’s correlations can still be used to access the presence of multicollinearity even if it is noy ideal. Chi Square test is usually chosen to examine the association between binary variables.

To examine the multicollinearity between each predictors, the correlation matrix in Figure 2 is plotted. Since the correlation coefficients between any two predictors are nearly smaller than 0.5, and most areas in this matrix is white (the deeper orange means the higher coefficient correlation between two predictors), there is no multicollinearity between these predictors.

Table 3 shows the logistic regression results. The p-values for all predictors except cellphone and PCTBACHMORE are smaller than 0.05, suggesting that all the predictors, except cell phone (p-value=8.812297e-01) and percentage of bachelors (p=7.749567e-01), are significant. The odds ratios and their 95% confidence interval are calculated. For a one-unit increase in FATAL_OR_M, the odds of there being a drinking driver go up by a factor of 2.25694878, holding that all other predictors are constant. The ratio between the odds of there being a drinking diver in a crash resulting in fatality or major injury and the odds of there being a drinking diver in a crash not resulting in fatality or major injury is 2.25694878. For a one-unit increase in OVERTURNED, the odds of there being a drinking driver go up by a factor of 2.53177687, holding that all other predictors are constant. The ratio between the odds of there being a drinking diver in a crash involving an overturned vehicle and the odds of there being a drinking diver in a crash not involving an overturned vehicle is 2.53177687. For a one-unit increase in CELL_PHONE, the odds of there being a drinking driver go up by a factor of 1.02999102, holding that all other predictors are constant. The ratio between the odds of there being a drinking diver in a crash when using a cell phone and the odds of there being a drinking diver in a crash not using a cell phone is 1.02999102. But this variable is not significant. For a one-unit increase in SPEEDING, the odds of there being a drinking driver go up by a factor of 4.65981462, holding that all other predictors are constant. The ratio between the odds of there being a drinking diver in a crash involving a speeding car and the odds of there being a drinking diver in a crash not involved speeding car is 4.65981462. The OR for AGGRESSIVE, DRIVER1617, DRIVER65PLUS, PCTBACHMOR are all smaller than 1, suggesting there were negative association between these predictors and independent variables. Imagine further that in drinking driver 1, MEDHHINC=a, and in drinking driver 2, MEDHHINC=a+1. Then, if we were to divide the odds of there being a drinking driver 2 by the odds of there being a drinking driver 1, the ratio of those odds would be 1.00000280.

```
                    OR          2.5 %         97.5 %
(Intercept)     0.06505601  0.05947628  0.07119524
FATAL_OR_M      2.25694878  1.90991409  2.65313350
OVERTURNED      2.53177687  2.03462326  3.12242730
CELL_PHONE      1.02999102  0.68354737  1.48846840
SPEEDING        4.65981462  3.97413085  5.45020642
AGGRESSIVE      0.55050681  0.50101688  0.60423487
DRIVER1617      0.27795502  0.14774429  0.47109277
DRIVER65PLUS    0.46085831  0.37998364  0.55347851
PCTBACHMOR      0.99962944  0.99707035  1.00215087
MEDHHINC        1.00000280  1.00000013  1.00000539
```

```
                    Estimate      Std. Error      z value          Pr(>|z|)
(Intercept)       -2.732507e+00   4.587566e-02  -59.5633209   0.000000e+00
FATAL_OR_M         8.140138e-01   8.380692e-02   9.7129660     2.654967e-22
OVERTURNED        9.289214e-01   1.091663e-01   8.5092302     1.750919e-17
CELL_PHONE         2.955008e-02   1.977778e-01   0.1494105     8.812297e-01
SPEEDING           1.538976e+00   8.054589e-02  19.1068171     2.215783e-81
AGGRESSIVE        -5.969159e-01   4.777924e-02  -12.4932079    8.130791e-36
DRIVER1617        -1.280296e+00   2.931472e-01  -4.3674171     1.257245e-05
DRIVER65PLUS      -7.746646e-01   9.585832e-02  -8.0813505    6.405344e-16
PCTBACHMOR        -3.706336e-04   1.296387e-03  -0.2858974     7.749567e-01
MEDHHINC           2.804492e-06   1.340972e-06   2.0913870    3.649338e-02
```

Table 3. the logistic regression with all predictors

Table 4 demonstrates the sensitivity, specificity, and misclassification rates for our models with different cut-off rates. We can observe that the sensitivity value and misclassification value decreases when we increase the cut-off value, while the specificity value gradually gets closer to 1. Among them, the lowest and the highest misclassification rates were identified when cut-off rate equals to 0.02, and the highest misclassification rates were identified when cut-off rate is 0.5.

Cut-off	Sensitivity	Specificity	Misclassification Rates
0.02	0.983501006	0.05807382764	**0.8888940135**
0.03	0.9799582463	0.0639203503	0.8853815224
0.05	0.7348088531	0.4690917097	0.5156812102
0.07	0.2213279678	0.9138188312	0.1258647726
0.08	0.1847082495	0.9386237432	0.1045798358
0.09	0.1682092555	0.9459624746	0.09860713956
0.1	0.1641851107	0.9482130189	0.0967161701
0.15	0.1042253521	0.9722106705	0.07752974818
0.2	0.02293762575	0.9953765992	0.06034959875
0.5	0.001609657948	0.9999021502	**0.05730559911**

Table 4. Assessment of the Cut-off Values (See This google sheet)

Figure 3 is the ROC Curve of the model. We can use the cut-off rate for the ROC curve to find the optimal cut-off point maximizing specificity and sensitivity as a balance. The point of the line where having the minimum sum distance to x-axis and y-axis has the cut-off value has the optimal rate, in which is 0.06365151, with the sensitivity value of 0.66076459 and the specificity value of 0.54524328. Area under ROC curve (AUC, which stands for Area Under Curve) is a measure of prediction accuracy of the model (how well a model predicts 1 response as 1's and 0 responses as 0's). Higher AUCs mean that we can find a cut-off value for which both sensitivity and specificity of the model are relatively high. In our model, AUC value equals 0.6399, which means the probability that the model can correctly rank two randomly selected observations, where one is a drinking driver and other one is not, 63.99% of the time the probability of being a drinking driver in the first observation will be higher than it in the second observation . Since it is less than 0.7, indicating this is a poor model.

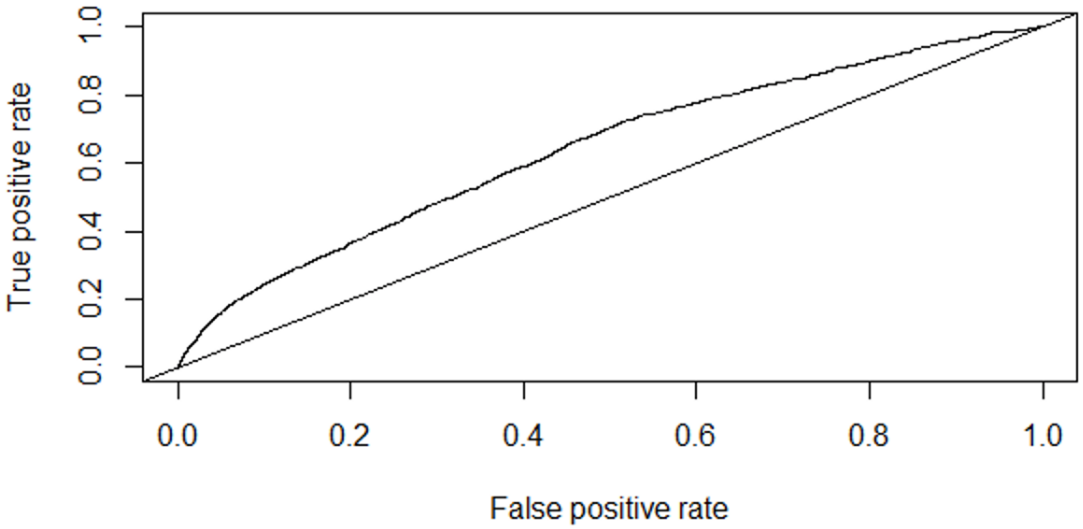


Figure 3. ROC Curve

```
            Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1961  -0.3692  -0.3153  -0.2764   3.0093
```

```
                         Coefficients:
Estimate          Std. Error     z value       Pr(>|z|)
(Intercept)      -2.65190     0.02753    -96.324  < 2e-16 ***
FATAL_OR_M       0.80932      0.08376      9.662  < 2e-16 ***
OVERTURNED      0.93978      0.10903      8.619  < 2e-16 ***
CELL_PHONE      0.03107      0.19777      0.157    0.875
SPEEDING          1.54032      0.08053     19.128  < 2e-16 ***
AGGRESSIVE       -0.59365   0.04775    -12.433  < 2e-16 ***
DRIVER1617       -1.27158    0.29311   -4.338 1.44e-05 ***
DRIVER65PLUS   -0.76646    0.09576   -8.004 1.21e-15 ***
                                  ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 19036  on 43363  degrees of freedom
Residual deviance: 18344  on 43356  degrees of freedom
                           AIC: 18360
```

Table 5. The logistic regression with the binary predictors

In the second regression model, we removed two continuous variables from the model, only kept binary variables. According to p-values in both outcomes, we decide all predictors' significance stays the same, as cellphone is an insignificant variable with p-value larger than 0.05, and the rest are significant ones with p-value less than 0.01.

We usually use Akaike Information Criterion (AIC) to compare models. The lower AIC value is, the more accurate and generalizable the model is. If a model is more than 3 AIC units lower than another, then it is considered significantly better than that model. Looking at outcomes from two models, the AIC of logistic regression with the binary predictors only is 18360.47, while the AIC of the logistic regression with all predictors is 18359.63. The second model is better because it’s simpler and has fewer variables, and its quality is no worse than that of the first model.

Discussion

With p-values less than 0.01, FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS are strong predictors of crashes that involve drinking. CELL_PHONE and PCTBACHMOR aren’t associated with the dependent variable since their p values are larger than 0.05. I was expecting that cell phone use might be a significant predictor of drunk driving because people who are under the influence might engage in other unsafe behaviors when they're behind the wheel, but this doesn't seem to be the case with our data.

According to Paul Allison’s writing, if we have a rarity of events in a large number of samples, we can use the modeling rare events methods, known as penalized likelihood, to reduce small-sample bias . In our model, the final data set contains the 43,364 crashes, with 5.73% of values of ‘1’ for the dependent variable, that is, 2485 cases involving drinking drivers. The logistic regression is appropriate here, since this is not the rarity of events.

When it comes to limitations of logistic regression, the AUC of the model is relatively low and we need to include additional predictors in the model that might improve the fit. One of the limitations is using Pearson correlations to assess multicollinearity, since the predictors don’t have a strong linear relationship within the context of regression analysis. It can be easily misinterpreted as a high degree of correlation from large values of the correlation coefficient does not necessarily mean a high linear relationship between the two variables.

Drunk driving is already one of the biggest causes of accidents in the United States, thus the predictors may only be contributing to the car crash accidents itself rather than drunk driving. In this case, drunk driving should be the predictor for ruthless driving activities like cell phone usage, speeding, aggressive driving etc, instead of vice versa: drivers who use cell phones when driving may not be the result of drunk driving, but rather he/she has always been a ruthless driver who uses cell phones, drives aggressively etc and ignores the laws and regulation, then he/she may also be easier to conduct drunk driving and cause accidents.

Also, it is not appropriate to incorporate predictors about the census block information of the place where the crash took place. Where the accident took place does not necessarily indicate the characteristics of the driver, they may come from a completely different place and had an accident in the middle of the trip. Thus, it is more possible to look at the census block where the driver comes from rather than the block where the accident took place.

References

[1] "Binary Variable – Learndatasci". 2021. *Learndatasci.Com*. https://www.learndatasci.com/glossary/binary-variable/.

[2] Hedlund, J H, and A T McCartt. “Drunk Driving: Seeking Additional Solutions.” TRID, April 30, 2002. https://trid.trb.org/view/724480.

[3] "Impaired Driving: Get The Facts | Motor Vehicle Safety | CDC Injury Center". 2021. *Cdc.Gov*. https://www.cdc.gov/transportationsafety/impaired_driving/impaired-drv_factsheet.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fmotorvehiclesafety%2Fimpaired_driving%2Fimpaired-drv_factsheet.html.

[4] Ramesh, Johari. “MS&E 226: “Small” Data Lecture 15: Examples of hypothesis tests (v3).” *Stanford University*.

[5] Szumilas, Magdalena. “Explaining odds ratios.” Journal of the Canadian Academy of Child and Adolescent Psychiatry. 19,3 (2010): 227-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/