### MUSA 500 HW 3

# **Using Logistic Regression to Examine Predictors of Car Crashes Caused by Alcohol**

Anna Duan, Jingyi Li, Jamie Song

## 1. Introduction

According to the US Department of Transportation, traffic accidents involving alcohol-impaired drivers represent a significant source of morbidity and mortality in the United States, causing nearly 30 deaths and numerous injuries each day. Moreover, the National Highway Traffic Safety Administration estimated the economic cost of alcohol-related crashes to exceed $59 billion annually. These figures underscore the importance of understanding the factors contributing to accidents caused by alcohol and developing effective strategies to mitigate the issue.

In this assignment, our goal is to use logistic regression to identify predictors associated with car crashes resulting from alcohol-impaired driving. The dataset is retrieved from OpenDataPhilly.org and comprises a total of 53,260 car crashes. Our analysis will concentrate on a subset of 43,364 incidents that occurred within Philadelphia's residential block groups (Brusilovskiy), for we excluded 9,896 crashes from non-residential block groups where both the median household income and vacancy rates were zero. The selection of predictors is based on the hypothesis that these variables may be associated with and have an impact on the response variable. The continuous and binary predictors are listed below:

* **FATAL\_OR\_M**: Indicates whether the crash resulted in fatality or major injury (1 = Yes, 0 = No). Drunk driving is associated with decreased reflexes, elevating the risk of fatal or serious injury accidents.
* **OVERTURNED**: Flags crashes involving an overturned vehicle (1 = Yes, 0 = No). An overturned vehicle often signals a serious accident, and impaired driving can lead to loss of vehicle control.
* **CELL\_PHONE**: Indicates if the driver was using a cell phone (1= Yes, 0 = No). Distraction while driving, a common cause of accidents, may be linked to high-risk behaviors associated with drunk driving.
* **SPEEDING**: Denotes whether the crash involved a speeding car (1 = Yes, 0 = No). Speeding may suggest a propensity for risk-taking, and such behavior could be connected to drunk driving.
* **AGGRESSIVE**: Marks crashes involving aggressive driving (1 = Yes, 0 = No). Aggressive driving may manifest as a result of impaired judgment and self-control due to alcohol consumption.
* **DRIVER1617**: Indicates if the crash involved at least one driver aged 16 or 17 (1 = Yes, 0 = No). Young drivers, often inexperienced, may engage in more hazardous driving behaviors under the influence of alcohol.
* **DRIVER65PLUS**: Notes crashes with at least one driver aged at least 65 (1 = Yes, 0 = No). Older drivers, with potentially reduced reflexes, may be more prone to accidents when driving under the influence.
* **PCTBACHMOR**: Represents the percentage of individuals aged 25 or older with at least a bachelor’s degree in the Census Block Group where the crash occurred. Education level may influence driving behavior and knowledge, impacting the incidence of drunk driving.
* **MEDHHINC**: Signifies the median household income in the Census Block Group where the crash occurred. Income levels might reflect residents' socioeconomic status, potentially influencing their safety awareness and behaviors.

We will use Rstudio as our tool to conduct a logistic regression and a rigorous examination of the predictors.

## 2. Methods

### a) Assumptions and Issues in OLS regression

OLS regression is a statistical method used for analyzing the relationship between one or more independent variables and a dependent variable. It seeks to find the optimized line that minimizes the sum of the squared differences (also known as residuals) between the observed and predicted values of the dependent variable. This method is typically used on non-categorical, continuous response variables.

The mathematical formula for simple linear regression (with one independent variable) in OLS is represented as:

Y = β₀ + β₁X + ε

Where Y is the dependent variable, X is the independent variable, β₀ is the y-intercept, β₁ is the slope of the regression line, and ε represents the error term.

For multiple linear regression (with more than one independent variable), the formula generalizes to:

Y = β₀ + β₁X₁ + β₂X₂ +…+ βnXn + ε

Here, X₁, X₂, …, Xn are the independent variables, and β₀, β₁, …, βn are the coefficients associated with each independent variable.

However, when OLS regression is applied to analyze binary dependent variables indicating the occurrence (expressed as 1) or non-occurrence (expressed as 0) of a specific event, several issues may arise. These issues include non-linearity, heteroskedasticity, the distribution of residuals, and problems associated with probability limits. To examine a dataset with binary dependent variables, we need to apply a different regression model — logistic regression.

### b) Logistic Regression

Before diving into the concept of logistic regression, we will first focus on the concept of odds. Probabilities express the likelihood of an event, while odds represent the ratio of the probability of an event happening to the probability of it not occurring. Both probability and odds share a similar calculation logic. Probability is computed as the event occurrence divided by the total possible outcomes, whereas odds are calculated as the ratio of the probability of an event occurring to the probability of it not occurring. To be more specific, if the probability of an event is denoted as p, then the odds are expressed as p/(1-p). 1-p represents the probability of the event not happening.

The odds ratio serves as a metric for assessing the relationship between outcomes in a binary dependent variable. It is calculated by (odds in the first group) / (odds in the second group). An odds ratio below 1 suggests lower odds in the first group, an odds ratio equals to 1 implies no difference in odds between the two groups, and an odds ratio greater than 1 indicates higher odds of the event occurring in the first group.

We will now dive into the equation of logistic regression with multiple predictors. Based on the lecture slides, the logit function equation is:

ln(p/1-p) = β0+ β1X1 + β2X2 + … + βnXn + ε

In the equation above, 𝑝=𝑃(𝑌=1). The quantity 𝑝/(1−𝑝) is called the odds. The quantity ln(𝑝/(1−𝑝)) is called the log odds, or logit. And β₀, β₁​, β₃​, ... , βn are the coefficients associated with the independent variables, X₁, X₂, ... , Xn are the values of the independent variables, *ε* represents the error term.

After doing some algebra, we obtain the logistic function (also known as inverse-logit function) equation:

p = 1 / 1 + e−(*β*0+*β*1*X*1+*β*2*X*2+…+*βnXn*)

In this equation, e is the base of the natural logarithm (approximately 2.71828),p is the probability of Y=1, β₀, β₁​, β₃​, ... , βn are the coefficients associated with the independent variables, X₁, X₂, ... , Xn are the values of the independent variables, ε represents the error term.

The logistic function has an S-shaped curve that ranges between 0 and 1. If β₀ + β₁X₁ + β₂X₂ + … + βnXn = 0, then p = 0.5. As β₀ + β₁X₁ + β₂X₂ + … + βnXn gets really big, p approaches 1 and when β₀ + β₁X₁ + β₂X₂ + … + βnXn gets really small, p approaches 0. As a result, the logistic function avoids predicting extreme probabilities and is more robust to outliers for predicting binary variables.

### c) Hypothesis Test for Each Predictor

In this section, we will do the following hypothesis test for each predictor X𝑖:

H₀: β𝑖 = 0 (𝑂𝑅𝑖 = 1), implying that the ratio of predictor values with a response of 1 for crashes involving drunk drivers is **identical** to the ratio of predictor values with a response of 1 for crashes not involving drunk drivers.

H𝑎: β𝑖 = 1 (𝑂𝑅𝑖 ≠ 1), implying that the ratio of predictor values with a response of 1 for crashes involving drunk drivers **differs from** the ratio of predictor values with a response of 1 for crashes not involving drunk drivers.

In logistic regression, the Wald statistic assesses the significance of individual coefficients. The Wald statistic is given in the formula by quantity of

In this formula, the is the estimated coefficient of the predictor variables, and 𝜎 is the standard error of the estimated coefficient. The p-value for each term may be obtained using the standard normal (z) tables.

Instead of focusing solely on the estimated β coefficients, we emphasize odds ratios as a more interpretable measure. Odds ratios quantify the change in the odds of the event happening (response variable denoted as 1) for a one-unit change in the predictor variable. To obtain the odds ratio, we exponentiate the corresponding coefficient (OR𝑖 = e𝛽𝑖). An exponentiated result greater than 1 indicates an increase in the odds of the event, while a value less than 1 suggests a decrease.

### d) Assess the Quality of Model Fit

R² is commonly used in linear regression to measure the proportion of variance explained by the model. In logistic regression, R² is not the best metric for quality assessment due to differences in the underlying mathematics and possible nonlinear relationship. Thus, its interpretation is not as straightforward as in OLS regression.

Instead of using R², we can use the Akaike Information Criterion (AIC) to assess the quality of the logistic model. AIC is a metric used to assess model performance, particularly valuable for comparing performance of various regression models involving different explanatory variables with the same dependent variable. AIC can balance model fit and complexity and penalizes models for having more complicated parameters. A lower AIC value indicates a better fit.

Sensitivity is the proportion of true positives among all actual positives (True Positives / True Positives + False Negatives). Higher values of sensitivity indicate better performance in correctly identifying events. Specificity measures the proportion of true negatives among all actual negatives (True Negatives / True Negatives + False Positives). Higher values of specificity indicate better performance in correctly identifying non-events. Misclassification rate is the proportion of incorrect predictions, calculated as (False Positives + False Negatives) / total observations. Lower misclassification rate indicates better overall model accuracy.

Fitted (predicted) values (𝑦̂𝑖) are the predicted probabilities of the dependent variable being Y=1 (event occurrence). Depending on different thresholds being chosen, these probabilities are compared to a chosen threshold to classify instances as Y=0 or Y=1. Higher 𝑦̂𝑖 values will correspond to occurrence (Y=1), while lower 𝑦̂𝑖 values will correspond to non-occurrence (Y=0). The thresholds used to split 𝑦̂𝑖 values are referring to cut-offs. Different cutoff values for predicted probabilities can be used to balance specificity and sensitivity, adjusting the trade-off between true positive rates and false positive rates. Cut-off value can be ranging between 0 to 1. A low cut-off may increase sensitivity but might decrease specificity, while a high cut-off may decrease sensitivity but increase specificity. So, the choice of cut-off allows for customization based on the specific needs of the analysis. It helps to find a balance that aligns with the goals and priorities of the model’s purpose.

To figure out the optimal cut-off value, we can plot a ROC which visually represents the true positive rate (sensitivity) against the false positive rate (specificity) at various cut-off values. Several methods can be used to determine the optimal cut-off point on the ROC curve. One common approach is the Youden's J statistic, which identifies the cut-off that maximizes the sum of sensitivity and specificity. Another approach is to identify the point on the ROC curve which minimizes the Euclidean distance to the (0,1) point, known as the nearest point to the perfect classification. In this assignment, we use the method to minimize the distance from the (0,1) point to select the optimal cut-off value.

The Area Under the ROC Curve (AUC) is a metric used to measure the overall quality of a logistic regression model or other classification models. It represents the probability that the model will correctly rank a randomly chosen pair of positive and negative instances. AUC ranges from 0 to 1, where a higher value indicates better discriminatory power. An AUC under 0.5 means that the model is not providing valuable information and the model's predictive performance is worse than random chance, so the model may need reevaluation or adjustment. Therefore, we only analyze AUC values higher than 0.5. A general guide for classification accuracy is:

1. 0.90 - 1 = excellent
2. 0.80 - 0.90 = good
3. 0.70 - 0.80 = fair
4. 0.60 - 0.70 = poor
5. 0..50 - 0.60 = fail

### e) Assumptions of Logistic Regression

To ensure the model's reliability, it is crucial to assess the assumptions of logistic regression, including the following aspects:

1. The dependent variables must be binary
2. Independence of observations
3. No severe multicollinearity
4. Larger samples are needed (at least 50 observations per predictor) than for linear regression because MLE (and not least squares) is used to estimate regression coefficients

We will conduct statistical analysis to make sure the above assumptions are fulfilled. For example, to examine the problem of multicollinearity, we will check Pearson correlation of all the predictors.

In contrast to Ordinary Least Squares (OLS) regression, logistic regression is exempt from the assumptions including:

1. Needs to be a linear relationship between dependent variable and each predictors
2. Assumption of homoscedasticity
3. Residuals need to be normal

### f) Exploratory Analyses

Before conducting logistic regression, statisticians typically perform exploratory analyses to gain insights into the data. In this assignment, we will conduct a cross-tabulations between the dependent variable and binary predictors to see whether there is an association between the two variables. The appropriate approach to statistically assess the association is through the Chi-square (χ2) test. In the case of this assignment, the null and alternative hypotheses for the χ2 test would be as follows:

H₀: the proportion of binary predictors resulted in 1 for crashes that involve drunk drivers is the same as the proportion of binary predictors resulted in 1 for crashes that don’t involve drunk drivers.

H𝑎: the proportion of binary predictors resulted in 1 for crashes that involve drunk drivers is different than the proportion of binary predictors resulted in 1 for crashes that don’t involve drunk drivers.

A high value of the χ2 statistic and a p-value lower than 0.05 suggest that there is evidence to reject the null hypothesis in favor of the alternative, and that there is an association between drunk driving and binary predictors.

For exploratory analyses of the continuous predictors, statisticians usually employ a test known as the independent samples t-test to examine significant differences in mean values. In the case of this assignment, the null and alternative hypothesis for t-test would be as follows:

H₀: There is no significant difference in mean values of continuous predictors for crashes involving and not involving alcohol.

H𝑎: There is a significant difference in mean values of continuous predictors for crashes involving and not involving alcohol.

A high value of the t-statistic, and a p-value lower than 0.05 suggest that there’s evidence to reject the null hypothesis in favor of the alternative, and the drunk driving is statistically significant with the continuous predictors.

## 3. Results

### a) Results of the Exploratory Analyses

Our dataset includes a total of 43,364 incidents. Our dataset is comprised by data of 3 categories:

1. Dependent variable:

**DRINKING\_D**: Indicates whether the driver was involved in drunk driving (1 = Yes, 0 = No).

1. Binary predictors:

**FATAL\_OR\_M**: Indicates whether the crash resulted in fatality or major injury (1 = Yes, 0 = No).

**OVERTURNED**: Flags crashes involving an overturned vehicle (1 = Yes, 0 = No).

**CELL\_PHONE**: Indicates if the driver was using a cell phone (1= Yes, 0 = No).

**SPEEDING**: Denotes whether the crash involved a speeding car (1 = Yes, 0 = No).

**AGGRESSIVE**: Marks crashes involving aggressive driving (1 = Yes, 0 = No).

**DRIVER1617**: Indicates if the crash involved at least one driver aged 16 or 17 (1 = Yes, 0 = No).

**DRIVER65PLUS**: Notes crashes with at least one driver aged at least 65 (1 = Yes, 0 = No).

1. Continuous predictors:

**PCTBACHMOR**: Represents the percentage of individuals aged 25 or older with at least a bachelor’s degree in the Census Block Group where the crash occurred.

**MEDHHINC**: Signifies the median household income in the Census Block Group where the crash occurred.

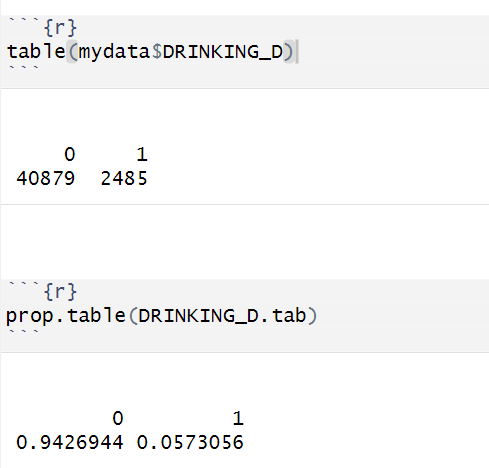


Figure 1: Table of dependent variable DRINKING\_D

From the drinking table processed in R, we can tell that among the 43,364 incidents, 40879 crashes didn’t involve drunk driving, which is about 94.27% of the total incidents. 2485 crashes involved drunk driving, which accounts for 5.73% of incidents in this data set. Therefore, the probability of drunk driving in car crashes in this case is 0.573, and the odds of car crash incidents involving drinking can be computed as 2485 / 40879 = 0.06.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | NO Alcohol involved | (DRINKING\_D = 0) | Alcohol involved | (DRINKING\_D = 1) | Total | Chi^2  p-value |
|  | N | % | N | % | N |  |
| FATAL\_OR\_M | 1181 | 2.90% | 188 | 7.60% | 1369 | <0.001 |
| OVERTURNED | 612 | 1.50% | 110 | 4.40% | 722 | <0.001 |
| CELLPHONE | 426 | 1.00% | 28 | 1.10% | 454 | 0.687 |
| SPEEDING | 1261 | 3.10% | 260 | 10.50% | 1521 | <0.001 |
| AGGRESSIVE | 18522 | 45.30% | 916 | 36.90% | 19438 | <0.001 |
| DRIVER1617 | 674 | 1.60% | 12 | 0.50% | 686 | <0.001 |
| DRIVER65PLUS | 4237 | 10.40% | 119 | 4.80% | 4356 | <0.001 |

Table 1: Cross-tabulation of the binary predictors

The cross-tabulation of the binary predictors with the dependent variable pf DRINKING\_D presents the count and proportion of occurrences is 1 for each predictor category. The table distinguishes between cases where the driver was not under the influence of alcohol and cases where the driver was influenced by alcohol, as well as the total cases of each predictor. Giving the example of the predictor "DRIVER65PLUS," it illustrates that 4237 cases of accident without drunk driving involvement (constituting 10.40% of all non-drunk driving accidents), and 110 accidents involving drunk driving (representing 4.40% of all drunk driving accidents) resulted in an driver who is 65 years old or older. The table reveals that, for predictors such as FATAL\_OR\_M, OVERTURNED, and SPEEDING, the proportion of 1-responses is higher when the driving involves alcohol. Conversely, for predictors like AGGRESSIVE, DRIVER1617, and DRIVER65PLUS, the proportion of 1-responses is lower when the driving involves alcohol.

As stated in class a chi-square test p-value lower than 0.05 suggest that there’s evidence to reject the null hypothesis in favor of the alternative, and that there’s an association between drunk driving and crash fatalities. The chi-square test results for the binary predictors indicate that except for the CELLPHONE predictor, all other binary predictors all have p-value <0.001. Therefore, we can conclude that all binary predictors except for CELLPHONE have significant association with drunk driving.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | NO Alcohol involved | (DRINKING\_D = 0) | Alcohol involved | (DRINKING\_D = 1) | t-test p-value |
|  | Mean | SD | MEAN | SD |  |
| PCTBACHMOR | 16.57 | 18.21 | 16.61 | 18.72 | 0.9137 |
| MEDHHINC | 31483.05 | 16930.10 | 31998.75 | 17810.50 | 0.16 |

Table 2: Cross-tabulation of the continuous predictors

Cross-tabulation of continuous predictors shows the means and standard deviations of PCTBACHMOR and MEDHHINC for crashes that involve alcohol and crashes that didn’t involve alcohol. The means and standard deviations for both PCTBACHMOR and MEDHHINC predictors involving alcohol are both slightly higher than means and standard deviations without involvement of alcohol. However, according to the p-values from t-test, we couldn't reject the null hypothesis. Therefore, we conclude these two continuous predictors do not have significant association with driving involving alcohol.

### b) Hypothesis Test of the Results

We will be discussing which logistic regression assumptions are met and which ones are violated for the problem at hand.

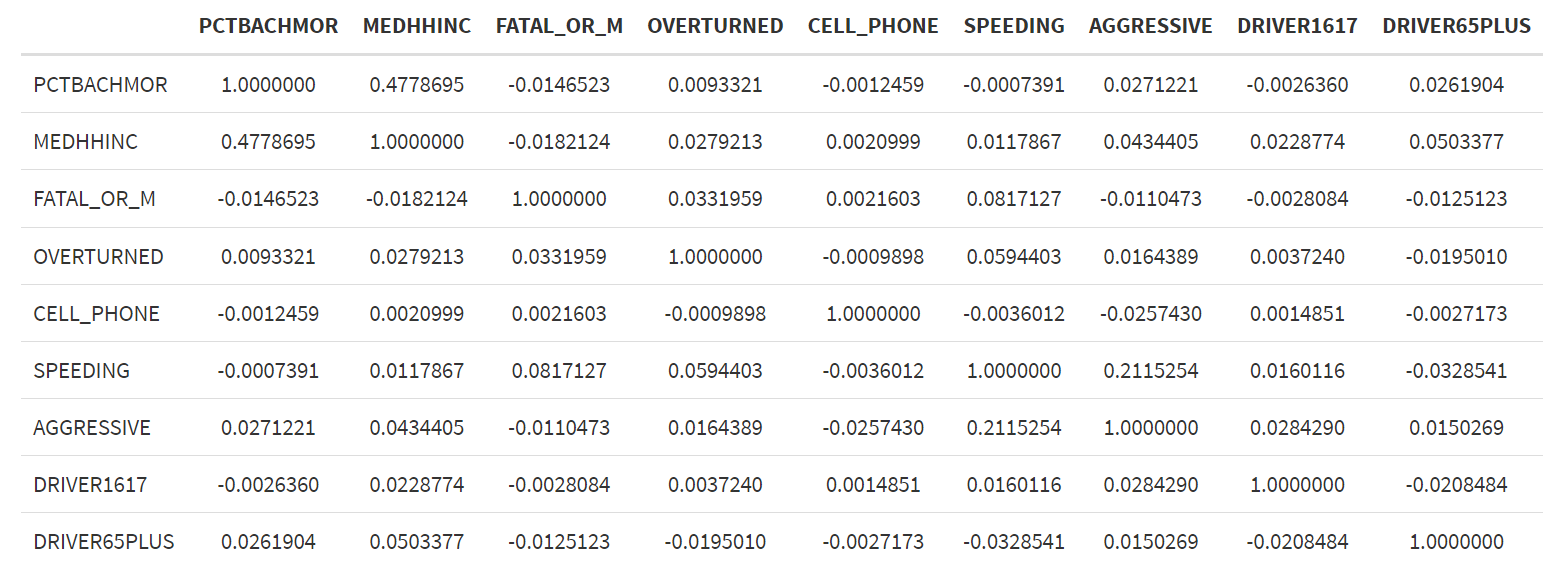


Table 3: Pearson Correlation of binary and continuous predictors

Pearson correlation is used to measure linear relationships between continuous variables. For binary variables with non-linear relationships it may not be able to accurately capture associations.

Multicollinearity occurs when two or more predictors are highly correlated (Daoud, 2017). If a high number appears between two predictors, we need to be aware of the multicollinearity problem. However, by examining the matrix, none of the correlations between two different predictors appear extremely high, so we suggest there is no severe multicollinearity.

### c) Logistic regression results

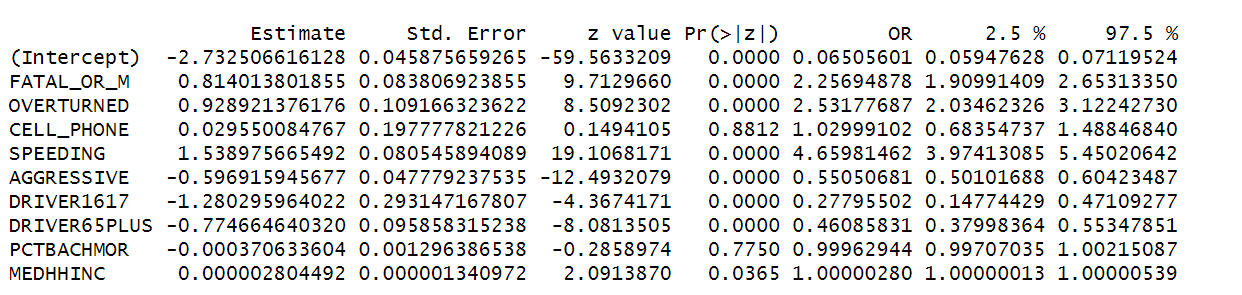


Table 4: Results of the logistic regression with all predictors

Table 4 presents the logistic regression results with both binary and continuous predictors. Looking at the p-value which is shown in the Pr(>|z|) column of the table, we can tell the p-value of CELLPHONE and PCTBACHMOR predictors are greater than 0.05, indicating we have failed to reject the null hypothesis for these two predictors, and they are not statistically significant. The rest of the predictors including FATAL\_OR\_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS, and MEDHHINC all possess p-values less than 0.05, indicating that we have sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis, so they have statistically significant relationships with the dependent variable.

The Estimate column represents the log odds, and the OR column represents the odds ratio for each predictor. For FATAL\_OR\_M, the log odds is 0.814, and the odds ratio is 2.257 with a 95% confidence interval ranging from 1.910 to 2.653. The odds ratio suggests that the odds of alcohol involvement in a crash with fatal and major injuries are higher than in a crash without fatal and major injuries. Given a crash resulted in fatality or major injury, the odds that the incident involved alcohol is increased by (2.257 - 1) \* 100% = 125.7%. The rest of predictors can be interpreted in the same manner – the log odds for OVERTURNED is 0.929, and the odds ratio is 2.531 with a 95% confidence interval ranging from 2.035 to 3.122. Given a crash resulted in an overturned vehicle, the odds of that crash involving alcohol is increased by (2.531 - 1) \* 100% = 153.1%. SPEEDING has the highest log odds and odds ratio. Its log odds is equal to 1.539, and if a crash involved speeding, the odds that it involved alcohol is increased by 366.0%.

AGGRESSIVE has a log odds of -0.597, indicating the odds of alcohol involvement in crashes caused by aggressive driving are lower than in a crash caused by non-aggressive driving. Therefore, if there is a crash with aggressive driving, the odds of the crash involving alcohol is decreased by (1 - 0.551) \* 100% = 44.9%. For DRIVER1617 and DRIVER65PLUS, their log odds are -1.280 and -0.775 respectively, indicating the odds of alcohol involvement in a crash caused by a driver at the ages of 16 and 17 and over 65 are lower than in a crash caused by a driver between 18 and 65. Given a crash involved an underage or elderly driver, the odds of the crash involving alcohol is decreased by 72.2% and 53.9%, respectively.

The odds ratio of MEDHHINC is nearly equal to 1, so its relationship to drunk driving does not have a large magnitude. As the median household income of a block group where a crash took place increases by $1,000, the odds of drunk driving in the crash increases by 0.28%.

|  |  |  |  |
| --- | --- | --- | --- |
| 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 | 0.099 |
| 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 | 0.058 |

Table 5: Specificity, sensitivity and the misclassification rates for the different probability cut-offs.

Table 5 presents the sensitivity, specificity and misclassification rate of 10 different cut-offs. If the predicted probability of DRINKING\_D in the model is equal or larger than the cut-off value, then we consider the driver in the crash to be drunk and assign a value of 1; if the predicted probability of DRINKING\_D in the model is lower than the cut-off value, then we consider the driver in the crash to not be drunk and assign a value of 0.

When the cut-off value is 0.02, the sensitivity is 0.984 and the specificity is 0.058, indicating that we correctly predicted 98.4 percent of drunk drivers, and 5.8 percent of drivers who are not drunk. The misclassification rate calculates the fraction of incorrectly predicted observations. With 0.02 cut-off value, it indicates that 88.9 percent of crashes we have mistakenly identified non-drunk drivers as drunk drivers or vice versa, which is highest among all cut-off values. As cut-off values increase, the misclassification rate drops, so the cut-off value of 0.5 has the lowest misclassification rate of 0.058.

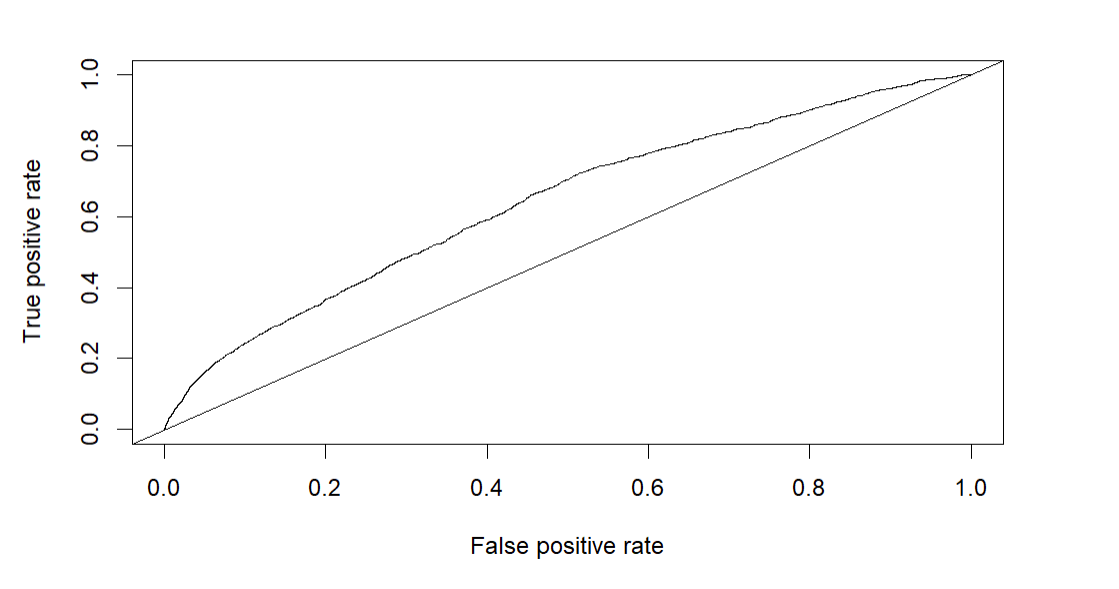


Figure 2: ROC curve

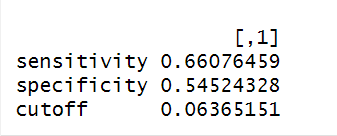


Figure 3: Sensitivity and specificity with optimized cut-off value

The ROC curve visually illustrates the trade-off between sensitivity (true positive rate) and specificity (false positive rate) for different cut-off rates in the logistic regression model. A “perfect” ROC curve should be closely aligned with the top-left border because the upper left corner indicates a true positive rate of 1 and a false positive rate of 0, which indicates the model always correctly predicts whether drunk driving is involved in a crash. The optimal cut-off value is determined by minimizing the distance from the upper-left corner of the graph. In this model, the optimized cut-off value is 0.0637, with 66.08 percent drunk drivers predicted correctly and 54.52 percent of non-drunk drivers predicted correctly.

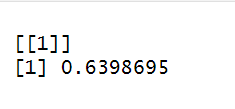


Figure 4: Area under ROC curve

Figure 4 shows the area under the ROC curve. The area under the ROC curve is a measure of the overall performance of accuracy of a model. According to Brusilovskiy, an area of 1 represents a perfect test (prediction); an area of .5 represents a worthless test (prediction). If the area under ROC curve is 0 .9 to 1, the model is performing excellently. Area under the ROC curve from 0.8 to 0.9 is considered as good, 0.7 to 0.8 is fair, 0.6 to 0.7 is poor and 0.5 to 0.6 is fail. In our model, the area under the ROC curve is around 0.64, which falls under the range between 0.6 to 0.7, indicating our model has poor accuracy in predicting alcohol involvement in crashes.

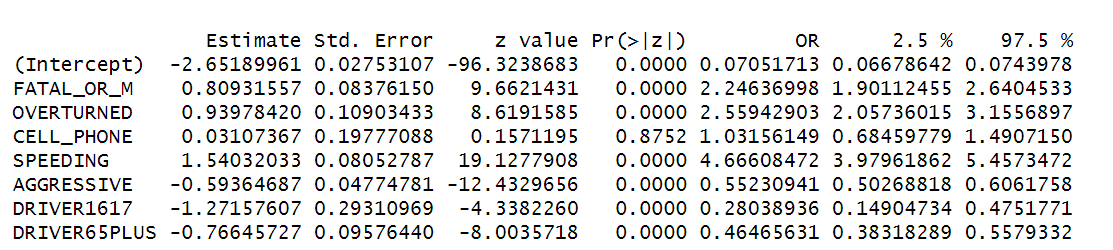


Table 6: Logistic regression with the binary predictors only

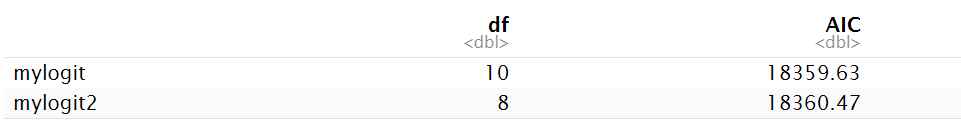


Table 7: AIC for logistic regression with all predictors and with only binary predictors

Comparing the p-value of logistic regression with only binary predictors with logistic regression with all predictors, the results do not change. CELLPHONE remains as not statistically significant with dependent variables, and all rest predictors are significantly associated with the dependent variable. Therefore, we conclude there are no predictors changed from non-significant to significant or vice versa.

The AIC scores compare the overall performance of two models. The logistic regression with all predictors has a score of 18359.63 and the logistic model without continuous predictors has a score of 18360.47. The difference between the AIC scores is less than 1 and can be omitted. Therefore, we cannot really conclude which model has a better performance based on the slight difference.

## 4. Discussion

In this paper, we conducted logistic regression to investigate the predictors of crashes involving drunk driving. We found that variables such as FATAL\_OR\_M, OVERTURNED, SPEEDING, DRIVER1617, DRIVER65PLUS, and AGGRESSIVE were strong predictors of such incidents, while variables like CELL\_PHONE, MEDHHINC, and PCTBACHMOR showed weak associations with the dependent variable.

These results most align with our expectations, for we assume drunk drivers are more likely to cause fatality and major injuries, overturning their vehicles due to impaired driving after they are drunk. We also expect drunk drivers to omit the speed limits. We do not consider income or education level is related to drunk driving. Underage drivers and elderly drivers negatively associated with crashes involving drunk drivers is reasonable as well. They are either limited by law or their own physical condition for alcohol intake. However, we are surprised by the direction of the relationship between the dependent variable and aggressive driving. As our result indicates that if there is a crash resulting in aggressive driving, the odds of an incident involving alcohol will go down, but our expectation is completely opposite.

Logistic regression is appropriate for this analysis due to the binary nature of the variables, and there is no indication that rare events modeling methods by Paul Allison would be more suitable.

Nevertheless, it is crucial to recognize limitations in our analysis. Since we did not result in a well-performed model, there might be potential existence of unobserved biased variables, which often result in decreasing the model fit. Moreover, there exist some challenges in interpreting odds ratios when applied to continuous predictors such as PCTBACHMOR and MEDHHINC.

## 5. Reference

Daoud, J. I. (2017, December). Multicollinearity and regression analysis. In *Journal of Physics: Conference Series* (Vol. 949, No. 1, p. 012009). IOP Publishing.

Brusilovskiy, E.(2023) Lecture 15-17 - Logistic and Multinomial Logistic Regression.pptx. PowerPoint Presentation.

Brusilovskiy, E.(2023) Lecture 16 - Logistic Regression in R. html