

Gabriel Hernandez, Timothy Oliver, Ziyi Tang

Professor Eugene Brusilovskiy

MUSA 5000/CPLN 6710

21 November 2023

## **Assignment 3: The Application of Logistic Regression to Examine the Predictors of Car Crashes Caused by Alcohol**

### **1. Introduction**

According to the US Department of Transportation, almost 30 people a day – or approximately one person every 51 minutes – die in motor vehicle crashes that involve an alcohol-impaired driver. Many more individuals are injured in these crashes. A recent study conducted by the National Highway Traffic Safety Administration has shown that the economic impact of alcohol-related crashes is estimated to be more than \$59 billion annually. There is great merit in identifying predictors of accidents related to drunk driving. Because the relation of alcohol inhibition to a given vehicle crash is a binary conclusion of yes or no, this analysis uses a logistic regression model to achieve the goal of identifying predictors in Philadelphia considering crash data from 2008-2012.

The data used in this analysis was compiled by the Pennsylvania Department of Transportation, and is made available to the public at [OpenDataPhilly.org](https://opendata.philly.org/). The dataset contains 53,260 geocoded car crashes in the City of Philadelphia for the years 2008 – 2012. However, we remove 9,896 crash locations which took place in non-residential block groups, where median household income and vacancy rates are 0, from the data set resulting in a final record count of 43,364 crashes. Because the crash data are geocoded, it is possible to spatially join the data to the 2000 census block group level data set that was used in prior OLS and spatially correlated analyses. After the spatial join, each crash point contains the median

household income and the percent of individuals with at least a bachelor's degree in the block group where the crash took place.

The predictors intended for the logistic regression model and other identification variables are as follows below:

- 1) CRN: Crash Record Number
- 2) DRINKING\_D: Drinking driver indicator (1 = Yes, 0 = No)
- 3) COLLISION: Collision category that defines the crash (0 = Non collision, 1 = Rear-end, 2 = Head-on, 3 = Rear-to-rear (Backing), 4 = Angle, 5 = Sideswipe (same dir.), 6 = Sideswipe (Opposite dir.), 7 = Hit fixed object, 8 = Hit pedestrian, 9 = Other or Unknown)
- 4) FATAL\_OR\_M: Crash resulted in fatality or major injury (1 = Yes, 0 = No)
- 5) OVERTURNED: Crash involved an overturned vehicle (1 = Yes, 0 = No)
- 6) CELL\_PHONE: Driver was using cell phone (1= Yes, 0 = No)
- 7) SPEEDING: Crash involved speeding car (1 = Yes, 0 = No)
- 8) AGGRESSIVE: Crash involved aggressive driving (1 = Yes, 0 = No)
- 9) DRIVER1617: Crash involved at least one driver who was 16 or 17 years old (1 = Yes, 0 = No)
- 10) DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old (1 = Yes, 0 = No)
- 11) AREAKEY: ID of the Census Block Group where the crash took place
- 12) 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

13) MEDHHINC: Median household income in the Census Block Group where the crash took place

Here, we will be regressing the binary dependent variable, DRINKING\_D, on the following binary and continuous predictors: FATAL\_OR\_M, OVERTURNED, CELL\_PHONE, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS, PCTBACHMOR, and MEDHHINC using the R version 4.3.1 and relevant packages to run the logistic regression.

## 2. Methods

### 2.1 Reasoning against use of OLS

The OLS linear regression model is a standard predictive model for similar purposes of identifying and quantifying the impact of predictors on a continuous dependent variable that is represented in the following equation:

$$\begin{aligned} DRINKING\_D = & \\ & \beta_0 + \beta_1(FATAL\_OR\_M) + \beta_2(OVERTURNED) + \beta_3(CELL\_PHONE) \\ & + \beta_4(SPEEDING) + \beta_5(AGGRESSIVE) + \beta_6(DRIVER1617) \\ & + \beta_7(DRIVERS65PLUS) + \beta_8(PCTBACHMOR) + \beta_9(MEDHHINC) + \epsilon \end{aligned}$$

Where

- $\beta_0$  represents the constant or intercept term for the dependent variable when each independent variable is 0 or unobserved in the model.
- $\beta_1, \beta_2, \beta_n$ : represents the coefficients that quantify the relationship between each independent variable and the dependent variable.



- $\varepsilon$  is the error term capturing the variation in the dependent variable,  $Y$ , not explained by the independent variables.

The independent variable coefficients ( $\beta_x$ ) are interpreted as the amount by which the dependent variable  $Y$  changes when said variable increases by 1 unit. However, our current use case has a binary dependent variable that can only change from 0 to 1 or vice versa. Therefore, interpretive statements of an increase in the unit of an independent variable increases the dependent variable by a value of the coefficient does not make sense. Instead, we want to measure the probability of getting its categorical value based on a range of observations. Additionally, the values for our dependent variable must range between 0 and 1 should our  $\hat{y}$  become probabilities. Using OLS to estimate our values can result in invalid estimations outside that range given its own possible range from negative to positive infinity.

## 2.2 Logistic Regression as an Alternative

The logistic regression model seeks to predict the probability that the dependent variable is 1 given the observed values of dependent variables and uses the logit translator function to confine those predictions to the desired range of  $[0, 1]$  as compared to the subtly different probit function. The returned output of the logit function used in this model is referred to as log odds or the natural logarithm of the ratio between a given probability and its alternative. For instance, the equation for the logit model used in this analysis can be written as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(FATAL\_OR\_M) + \beta_2(OVERTURNED) + \beta_3(CELL\_PHONE) + \beta_4(SPEEDING) + \beta_5(AGGRESSIVE) + \beta_6(DRIVER1617) + \beta_7(DRIVERS65PLUS)$$

$$+ \beta_8(PCTBACHMOR) + \beta_9(MEDHHINC) + \varepsilon$$

Where  $p$  is the probability that an outcome or the dependent variable  $Y$  is true,  $1-p$  is the probability that an outcome or the dependent variable  $Y$  is false, and the right-hand symbols have the same meaning as in OLS. For clarity, the ratio between a given probability and its alternative is represented as  $\left(\frac{p}{1-p}\right)$  and is referred to as odds. The logistic function or inverse-logit function involves solving the logit model's equation for the probability—the desired output for our regression model. Therefore, the logistic function can be represented as:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}$$

$$= \frac{1}{1 + e^{\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \dots - \beta_n X_n}}$$

where  $p$  becomes 0.5 for variable coefficients and intercepts totaling zero, approaches 1 with very large coefficient and intercept sums, and approaches 0 with very small coefficient and intercept sums. Due to space, general notation was used, but the representation for this analysis uses the same values for  $X_i$  as in the logit function ranging from  $i = [1,9]$ .

### 2.3 Predictor Hypothesis Tests - Wald Test

Given the logistic regression uses maximum likelihood estimation to estimate parameters, each predictor is also tested using the Wald Hypothesis test with hypotheses

$$H_0: \hat{B}_i = 0 \text{ or that } OR_i = 1$$

$$H_a: \hat{B}_i \neq 0 \text{ or that } OR_i \neq 1$$

where  $\hat{B}_i$  is the maximum-likelihood estimation (MLE) and  $OR_i$  is the odds-ratio for a given coefficient. Given the expected MLE value of 0, the normal distribution of the standardized quantity or the Wald statistic is the output of the hypothesis test represented by

$$\frac{\hat{B}_i - E(\hat{B}_i)}{\sigma \hat{B}_i} = \frac{\hat{B}_i - 0}{\sigma \hat{B}_i} = \frac{\hat{B}_i}{\sigma \hat{B}_i} = Z.$$

Likewise, most statisticians prefer examining odds ratios calculated by exponentiating  $B$  coefficients in this test rather than examining the coefficients themselves.

## 2.4 Assessing Logistic Model Fit

### 2.4.1 General Approaches

The practice of computing an R-squared value for the fitted logistic model is possible to determine goodness of fit with the understanding that a higher value denotes better goodness of fit. However, the R-squared cannot be interpreted as the % of variance explained by the model as it can for an OLS regression. For that reason, it is rarely used to describe a logistic model's goodness of fit. An alternative is comparing the Akaike Information Criterion or AIC of multiple models. The AIC is an estimator of prediction error in a model computed from twice the difference in the number of estimated parameters and the maximum log likelihood of a model with lower criterion values denoting a better model.

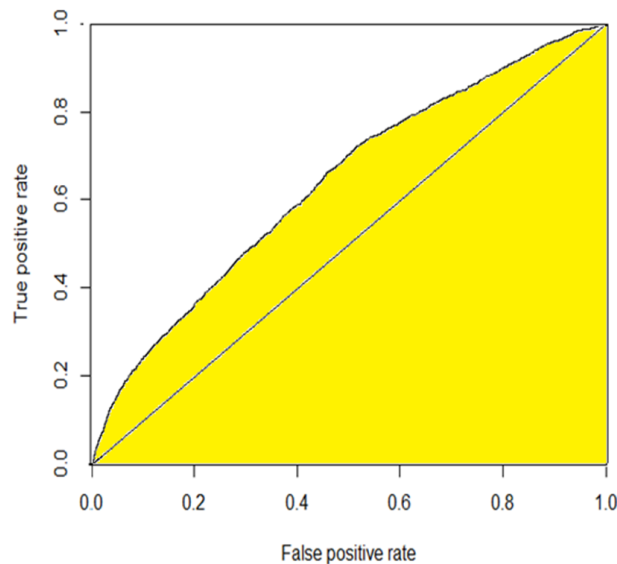
### 2.4.2 Cut-off Rates

Given that a logistic model estimates probabilities for a binary dependent variable, the actual determination of the prediction as 1 or 0 can be influenced by deciding a cut-off threshold. The cut-off is used when predicting or fitting the values of  $y$  by computing the probability of  $y=1$  for an observation with the model's logistic function. If the estimated probability is greater than the cut-off, a fitted value of 1 is assigned while 0 is assigned

otherwise. The decision of cut-offs is mainly arbitrary, so we need a method of comparing the accuracy of the model among these different decision cut-off values. Therein, the accuracy of a model can be broken down into true positives and true negatives where the prediction model accurately predicts an observation to be a 1 or 0, respectively as well as false positives and false negatives where the model predicts a 1 for a false observation or 0 for a true observation, respectively. The number of each of these metrics across the entire dataset can be easily viewed in a cross tabulation of the observed dependent variable with the predictions with the rates of each being reported to describe goodness of fit in a given situation. The rate of true positives is often referred to as sensitivity, the rate of true negatives as specificity, and the ratio of misclassified observations (both false positives and false negatives) from the total observation count is the misclassification rate. A relatively simple way to determine the best cutoff is to use multiple different values and find which one has the lowest misclassification rate for the given dataset. However, a more general criteria for the best cut-off value is to determine which of those tested maximizes both the sensitivity and specificity.

### 2.4.3 ROC Curves

A visualization plotting the sensitivity against the specificity called the Receiver Operating Characteristic (or ROC) Curve is a useful tool originating from British air radar during World War II to determine if blips in radar imaging were targets, friendly crafts, or noise like a bird flock. An example can be seen in [Figure 1](#).



**Figure 1:** *Sample ROC Curve*

It is useful in this analysis by providing two methods of identifying optimal probability cut-offs. First is the Youden Index which is the cut-off that maximizes sensitivity and specificity. Secondly, the value achieving the minimum distance of the ROC Curve from the upper left corner of the graph is an optimal cut-off value due to said corner representing the point where both specificity and sensitivity are 1 (their absolute maximum); this second method can be easily implemented in R with functions to find values maximizing the area under the ROC curve (AUC). Therefore we will use the second method. The AUC is an overall prediction accuracy metric for the model ranging from 0.5 to 1 where some statisticians say an AUC value above 0.7 is generally fine. A more conservative estimate would classify AUC values as follows:



AUC Value Range	Accuracy Descriptor
0.5-0.6	fail (no better than a coin flip for 0.5)
0.6-0.7	poor
0.7-0.8	fair
0.8-0.9	good
0.9-1.0	excellent

## 2.5 Logistic Model Assumptions

Similar to OLS regression, logistic regression has a set of assumptions for its use. Some from OLS regression must hold in logistic regression as well. It is also important to note that interpretation of logistic regression parameters echoes OLS in necessitating the qualifier that all other predictors are held constant. Otherwise, logistic regression assumes...

- 1) Binary Dependent Variable: The dependent variable of the regression must be binary and not continuous or categorical with more than 2 categories.
- 2) Independence of Observations: Each observation in the dataset is independent of other observations such as not being clustered or drawing from the same sampling pools.
- 3) No (Severe) Multicollinearity: Independent variables should not be too highly correlated with each other. Here, a correlation  $r$  value at or above the 0.8 threshold is considered highly correlated.
- 4) Minimum of 50 observations per predictor: MLE is used to estimate regression coefficients rather than least squares as in OLS.

However, logistic regression does not hold OLS's assumptions of presence of a linear relationship between the dependent variable and each independent variable, the assumption of homoscedasticity, nor the normality of residuals.

## 2.6 Common Exploratory Analyses

### 2.6.1 Binary (Categorical) Predictor Association

Considering the importance of probability in predicting a binary dependent variable, gaining an understanding of the relative probabilities among the data that will likely train the model is a common form of exploratory analysis. This is performed by cross-tabulation of the dependent variable (and each of its potential values) and the binary predictors to observe if an association exists between the pairs. This association is tested statistically using the Chi-Square ( $\chi^2$ ) test which determines whether the distribution of one categorical variable varies with respect to the values of another. The null and alternative hypotheses for the Chi-Square test between our dependent variable, `DRIKING_D`, and the `AGGRESSIVE` predictor would be...

$H_0$ : the proportion of aggressive driving in crashes that involve drunk drivers is the same as the proportion of aggressive driving in crashes that do not involve drunk drivers

vs.

$H_a$ : the proportion of aggressive driving in crashes that involve drunk drivers is different than the proportion of aggressive driving in crashes that do not involve drunk drivers

Additionally, a higher test statistic of  $\chi^2$  or p-value lower than or 0.05 significance level suggests enough evidence to reject the null hypothesis in favor of the alternative and an association between the two variables. This test would be repeated between all binary predictors and the dependent variable.

### 2.6.2 Continuous Predictor Association

Comparing the means of continuous predictors for both 1 and 0 values of the dependent variable using an independent samples t-test is also common. The independent samples t-test is appropriate for examining if there are significant differences in the mean values of PCTBACHMOR and MEDHHINC for crashes involving alcohol compared to those that didn't. The test would similarly suggest enough evidence to reject the null hypothesis in favor of the alternative with higher t-values or a p-value below our 0.05 significance level given the null and alternative hypotheses of...

$H_0$ : average values of the (PCTBACHMOR/MEDHHINC) variable are the same for crashes that involve drunk drivers and crashes that don't.

vs.

$H_a$ : average values of the (PCTBACHMOR/MEDHHINC) variable are different for crashes that involve drunk drivers and crashes that don't.

## 3. Results

### 3.1 Exploratory Analysis

#### 3.1.1 $P(Y=1)$

To better understand the dataset for chi-square association tests and beyond, the number and proportion of crashes involving drunk driving was found as shown in [Figure 2](#).

0	1
40879	2485
0	1
0.9426944	0.0573056

Figure 2: Tabulation of DRINKING\_D

Only 2,485 of the total 43,364 crashes or 5.73% involved drunk driving which is a fairly small proportion suggesting that the sensitivity or true positive rate will play little importance in overall model accuracy. So, some measures of the model's goodness of fit will be more skewed for this specific data as opposed to generalizable models.

### 3.1.2 Binary Predictor Associations

<b>Binary Predictors = 1</b>		<b>No Alcohol Involved</b>		<b>Alcohol Involved</b>		<b>Total</b>
		<b>(DRINKING_D = 0)</b>		<b>(DRINKING_D = 1)</b>		
	<b><math>\chi^2</math> p-value</b>	<b>N</b>	<b>%</b>	<b>N</b>	<b>%</b>	<b>N</b>
<b>FATAL_OR_M</b>	2.52E <sup>-38</sup>	1181	2.9%	188	7.6%	1369
<b>OVERTURNED</b>	1.55E <sup>-28</sup>	612	1.5%	110	4.4%	722
<b>CELL_PHONE</b>	0.687	426	1.0%	28	1.1%	454
<b>SPEEDING</b>	6.28E <sup>-84</sup>	1261	3.1%	260	10.5%	1521
<b>AGGRESSIVE</b>	2.00E <sup>-16</sup>	18522	45.3%	916	36.9%	19438
<b>DRIVER1617</b>	6.12E <sup>-6</sup>	674	1.6%	12	0.5%	686
<b>DRIVER65PLUS</b>	2.76E <sup>-19</sup>	4237	10.4%	119	4.8%	4356

**Figure 3:** Binary Predictor Cross-Tabulation and Chi-Squared p-value

For binary predictors, the vast majority appear to have an association to drunk driving. For instance, many percentages of crash types appear to triple when drunk driving is involved such as speeding or having a car overturned. Other predictors like the driver being a young teen, being elderly, or aggressive appear to decrease. Meanwhile, the likelihood of cell phone use changing is not much different across crashes involving alcohol. Similarly, all predictors but CELL\_PHONE have significant p-values with the stand-out predictor having a

p-value of 0.687 for its chi-squared test. Given that all the predictors tested are binary alongside the dependent variable, the degrees of freedom are consistently 1.

### 3.1.3 Continuous Predictor Associations

<i>Continuous Predictors</i>		<b>No Alcohol Involved</b>		<b>Alcohol Involved</b>	
		<b>(DRINKING_D = 0)</b>		<b>(DRINKING_D = 1)</b>	
	<b>t-test p-value</b>	<b>Mean</b>	<b>SD</b>	<b>Mean</b>	<b>SD</b>
<b>PCTBACHMOR</b>	0.9137	16.56986	18.21426	16.61173	18.72091
<b>MEDHHINC</b>	0.16	31483.05	16930.1	31998.75	17810.5

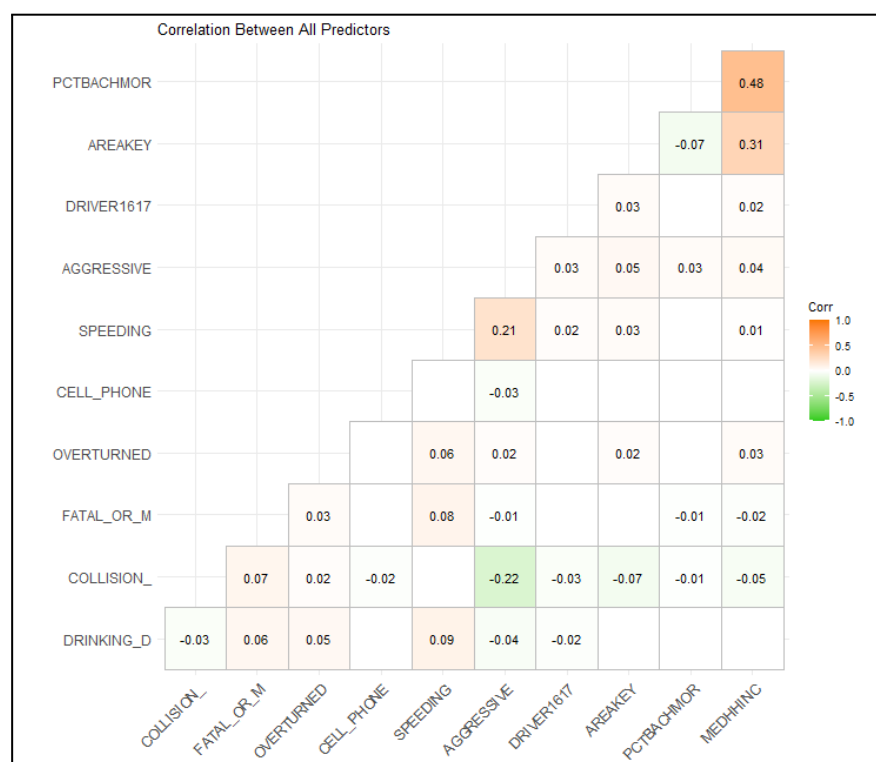
**Figure 4:** *Continuous Predictor Mean Differences and Independent T-test p-value*

For continuous predictors, there appears to be no evident association between the predictors and crashes involving drunk driving. Each of the mean differences fall within a standard deviation of crashes involving the alternative degree of alcohol involvement, and none of the t-test p-values are significant given our 0.05 confidence level. Therefore, we cannot reject the null hypothesis of association, average differences in predictors being the same for crashes involving drunk driving compared to those that do not.

## 3.2 Logistic Regression Assumptions and Predictor Correlations

Considering the four assumptions for logistic regression, three are quickly determined to hold. First, the dependent variable is definitely binary with determination between a crash involving drunk driving or not being possible. Second, there are at least 50 observations per predictor. We have a total of nine predictors and 43,464 observations which well exceed 450


observations. Third, each of the predictors appear independent. None of the binary variables draw from the same population such as a DRIVER16PLUS when we have DRIVER1617—the latter’s probabilities would depend on the former having 1 observations as well. The last assumption to test is that there is no severe multicollinearity. We use Pearson’s  $r$  correlation coefficient and define a correlation  $r$  value at or above the 0.8 threshold as highly correlated. The pairwise correlations for all predictors is shown in [Figure 5](#).



**Figure 5:** *Pairwise Predictor Pearson Correlation Matrix*

No correlation value exceeds 0.48 which is the correlation of the two continuous predictors and below our multicollinearity threshold of 0.8, so we accept that all four of the logistic regression assumptions hold.

There is a consideration to be made that the Pearson correlation aims to quantify the strength of linear relationships and will not have strong values for clear polynomial relationships like quadratic and cubic functions—much less our binary predictor pairs likely


represented by functions of only 4 possibilities. An alternative relationship measure or correlation method like Matthews correlation would likely be more appropriate (Chicco), but Pearson was considered sufficient for this test of multicollinearity. 

### 3.3 Logistic Regression Results

#### 3.3.1 Nine Predictor Model - Including Continuous Variables

	Estimate	Std. Error	z value	Pr(> z )	OR	2.5 %	97.5 %
(Intercept)	-2.732506616128	0.045875659265	-59.5633209	0.000000e+00	0.06505601	0.05947628	0.07119524
FATAL_OR_M	0.814013801855	0.083806923855	9.7129660	2.654967e-22	2.25694878	1.90991409	2.65313350
OVERTURNED	0.928921376176	0.109166323622	8.5092302	1.750919e-17	2.53177687	2.03462326	3.12242730
CELL_PHONE	0.029550084767	0.197777821226	0.1494105	8.812297e-01	1.02999102	0.68354737	1.48846840
SPEEDING	1.538975665492	0.080545894089	19.1068171	2.215783e-81	4.65981462	3.97413085	5.45020642
AGGRESSIVE	-0.596915945677	0.047779237535	-12.4932079	8.130791e-36	0.55050681	0.50101688	0.60423487
DRIVER1617	-1.280295964022	0.293147167807	-4.3674171	1.257245e-05	0.27795502	0.14774429	0.47109277
DRIVER65PLUS	-0.774664640320	0.095858315238	-8.0813505	6.405344e-16	0.46085831	0.37998364	0.55347851
PCTBACHMOR	-0.000370633604	0.001296386538	-0.2858974	7.749567e-01	0.99962944	0.99707035	1.00215087
MEDHHINC	0.000002804492	0.000001340972	2.0913870	3.649338e-02	1.00000280	1.00000013	1.00000539

Figure 6: Nine Predictor Summary

After running multiple logistic regression using the nine predictors specified, 7 of the nine were considered significant with some beta coefficient estimations  resulting in odds ratios communicating drunk driving being a factor of times more or less likely to be involved with the presence of a predictor. Additionally, the intercept ( $B_0$ ) is considered significant with a p-value of 0 and value of -2.73 which corresponds to an odds-ratio of  $e^{-2.73} = 0.0651$ . This is interpreted as a control case where no predictor in the model is present has the odds of a crash involving drunk driving being  $(0.0651-1)*100\%$  or -93.49% which is similar to the proportions presented in section 3.1. We will examine each predictor one at a time describing its significance and interpretations of their odds ratios.

The FATAL\_OR\_M predictor is significant with a very low p-value and an odds-ratio suggesting a car crash involving a fatality or major injury and holding all

other predictors constant having  $(2.26-1)*100\%$  or 126% higher odds of involving drunk driving than crashes without fatalities or major injuries. OVERTURNED is also significant with an odds-ratio suggesting a car crash resulting in an overturned car and holding all other predictors constant having  $(2.53-1)*100\%$  or 153% higher odds of involving drunk driving than crashes where no car was overturned. The CELL\_PHONE predictor is not significant with a p-value of 0.881; therefore there is not enough evidence to suggest an association between drunk driving and crashes involving cell phone use despite the odds-ratio communicating  $(1.03-1)*100\%$  or 3% higher odds of involving drunk driving than crashes without cell phone use holding all other predictors constant.

SPEEDING is significant with an interpretation that crashes involving speeding and holding all other predictors constant have  $(4.66-1)*100\%$  or 366% higher odds of involving drunk driving than those without speeding violations which is the highest degree of impact among all predictors. AGGRESSIVE is significant with an interpretation that crashes involving aggressive behavior and holding all other predictors constant have  $(0.56-1)*100\%$  or 44% lower odds of involving drunk driving than crashes without aggressive behavior. DRIVERS1617 is significant and has odds-ratios communicating crashes with at least one 16 or 17 year old driver and holding all other predictors constant has  $(0.28-1)*100\%$  or 72% lower odds of involving drunk driving than crashes without a 16 or 17 year old driver. Likewise, DRIVERS65PLUS is significant and has odds-ratios communicating crashes with at least one driver over age 64 and holding all other predictors constant has



$(0.46-1)*100\%$  or 54% lower odds of involving drunk driving than crashes without a senior-aged driver.

The second insignificant predictor, PCTBACHMOR, has a p-value of 0.775; therefore there is not enough evidence to suggest an association between drunk driving and the rise of one percent for residents with bachelor's degrees or higher within the census block group of the crash site. The odds-ratio communicates  $(0.9996-1)*100\%$  or 0.04% decrease in odds of involving drunk driving when the percentage of higher-degree holders increases in other census block groups when holding other predictors constant which is notably small as with CELL\_PHONE given the predictor's lack of statistical significance.

Lastly, the significant predictor of MEDHHINC has an odds-ratio communicating a  $(1.000002-1)*100\%$  or 0.0002% increase in odds as the median household income of the crash site's census block increases by one unit holding all other predictors constant.

### *3.3.2 Sensitivity, Specificity, & Optimal Cut-offs*

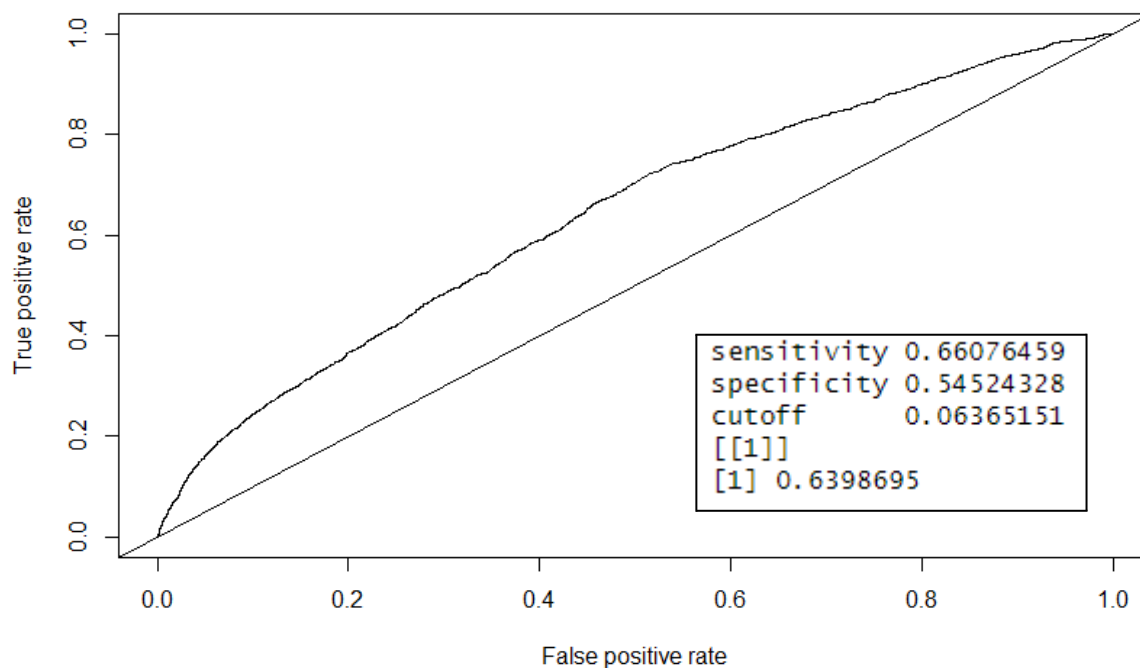
Figure 7 presents the 10 different cut-offs or probability thresholds tested iteratively and their resulting sensitivities, specificities, and misclassification rates.

<u>Cut-off Value</u>	<u>Sensitivity</u>	<u>Specificity</u>	<u>Misclassification Rate</u>
0.02	98.4%	5.8%	88.9%
0.03	98.1%	6.4%	88.4%
0.05	73.5%	46.9%	51.6%
0.07	22.1%	91.4%	12.6%
0.08	18.5%	93.9%	10.5%
0.09	16.8%	94.6%	9.9%
0.1	16.4%	94.8%	9.7%
0.15	10.4%	97.2%	7.8%
0.2	2.3%	99.5%	6.0%
0.5	0.2%	99.99%	5.7%

Figure 7: Metrics of Iterative Cut-offs

Using a minimal misclassification rate as the criteria for the optimal threshold, probabilities above 0.5 as 1 or crashes involving drunk driving were optimal with a misclassification rate and high specificity of 99.99%. Generally, sensitivity and specificity were inversely related. Therefore, the worst performing cut-off (or one with the highest misclassification rate of 88.9%) was 0.02 which was the lowest value tested. Given the small number of samples where the crash involved a drunken driver (2485 records out of 43,364 or 5.7%), the misclassification rate is unproportionally impacted by the rate of true negative classifications or specificity rate in the logistic regression fit of this dataset.

The ROC curve yielded in the analysis can be seen below in [Figure 8](#) alongside its sensitivity, specificity, the value of its optimal cutoff that minimizes the distance to the upper left corner of the plot, and its area under the curve (AUC) from top to bottom. The AUC for the generated ROC curve sits squarely in the poor category at a value of 0.640 although it is not in the failing category. The sensitivity and specificity of 0.661 and 0.545, respectively, are also fairly similar with a higher maximum sum than the maximum from [Figure 7](#) of 0.735 and 0.469 or 1.204 compared to the ROC pair's 1.206. Also, the AUC generally tells us that we are somewhat capable of finding a cut-off value where sensitivity and specificity are both relatively high which is true given both are above 50% with the determined 0.064 cut-off.



**Figure 8:** ROC and Cut-off Metrics Maximizing AUC

The cutoff rate determined with the creation of the ROC curve (0.064) is much smaller than our determined cutoff rate of 0.5. However, it is important to note the criteria

used to determine both. The previous value of 0.5 minimized the misclassification rate which mainly depends on specificity due to the large portion of data not involving drunk drivers. The rate determined here instead maximizes both sensitivity and specificity resulting in 0.064 which is closest to the effect of 0.05 in our [Figure 7](#) in being generally accurate for any data.

### 3.3.3 Seven Predictor Model - Binary Variables Only

	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


[Figure 9](#): 7 Predictor Summary

Use of another logistic regression model where only the binary variables are included as predictors offers some useful comparison to our initial, nine predictor model. All of the same predictors are significant with the p-value for CELL\_PHONE, the only insignificant predictor, remaining similar to the initial model's. Many of the odds ratios are similar with estimates connoting lower odds for the driver and aggressive predictors (as well as the intercept) and the factors of the odds ratios remaining constant. You can use similar equations to find new odds ratio interpretations, but we feel that doing so is unnecessary for comparison in place of other methods like comparing the Akaike Information Criterion (AIC) using [Figure 10](#) below.

=====		
	Dependent variable:	
	-----	
	DRINKING_D	
	(1)	(2)
-----		
FATAL_OR_M	0.814*** (0.084)	0.809*** (0.084)
OVERTURNED	0.929*** (0.109)	0.940*** (0.109)
CELL_PHONE	0.030 (0.198)	0.031 (0.198)
SPEEDING	1.539*** (0.081)	1.540*** (0.081)
AGGRESSIVE	-0.597*** (0.048)	-0.594*** (0.048)
DRIVER1617	-1.280*** (0.293)	-1.272*** (0.293)
DRIVER65PLUS	-0.775*** (0.096)	-0.766*** (0.096)
PCTBACHMOR	-0.0004 (0.001)	
MEDHHINC	0.00000** (0.00000)	
Constant	-2.733*** (0.046)	-2.652*** (0.028)
-----		
Observations	43,364	43,364
Log Likelihood	-9,169.815	-9,172.233
Akaike Inf. Crit.	18,359.630	18,360.470
=====		
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 10: Model Comparison Summary, with AIC

This additional summary table of both logistic models provides quick comparison between them alongside their log likelihoods and Akaike Information Criterion (AIC).

The binary-only model is #2 and has a slightly higher AIC. Similarly, the log likelihood value for model two is further from zero than model one denoting a lower goodness of fit. Therefore, the multiple logistic regression model including continuous variables is suggested to be the better model. 

#### **4. Discussion**

Through this analysis, the association of various behaviors or effects of Philadelphia car crashes and some characteristics of the block groups they occurred in were evaluated in relation to crashes involving drunk driving. Individual tests like chi-squared and independent t-tests were used to determine individual predictor associations while the fitting of two logistic regression models helped to determine the strength of the variables as predictors of drunk driving playing a part in a car crash. Generally, aggressive driving or a driver being ages 16, 17, or beyond 64 had associations of lower odds for drunk driving to be a factor of the crash. Meanwhile, a crash resulting in a fatality, overturning a car, or involving speeding had higher odds of involving drunk driving—especially speeding with nearly 4 times higher odds. The percent of crash block group residents with higher education degrees of at least a bachelor's or cell phone usage in a car crash had no association with the dependent variable nor significance in the logistic models including them. On the other hand, the median household income in the crash's block group was slightly significant despite no individual association to crashes involving drunk driving.

The results are fairly surprising such as the degree to which speeding was associated with drunk driving crashes in comparison to the other predictors and the general lack of association for cell phone usage given media comparison between drunk driving and using the phone behind the wheel. The latter does become less shocking when realizing concurrent

occurrences, however. Otherwise, the scale and direction of the predictor relationships to drunk driving crashes falls within expectations.

While the proportion of drunk driving cases feels small at 5.7%, logistic regression is appropriate for use in this analysis given the rarer 1 case having 2,485 observations out of the total 43,364. The maximum likelihood estimation of the parameters presumably has a sufficient sample of both 0 and 1 cases with over 2,000 observations within an almost 40,000 sample. Otherwise, the model was generally limited by the use of Pearson's correlation to test for multicollinearity. As mentioned earlier, Matthew's correlation would have been a good alternative for binary variables in future logistic regression assumption tests. Furthermore, the choice of continuous variables from the spatially joined data were sufficiently different, but inclusion of another factor like the area's car crash rate or average car throughput would feel more impactful for similar analyses for this dependent variable. However, the logistic models provided consistent interpretations of the behaviors and effects of car crashes that can indicate the involvement of drunk driving.

## Works Cited

Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6 (2020). <https://doi.org/10.1186/s12864-019-6413-7>