Introduction

According to the National Highway Traffic Safety Administration, there were 10,142 people killed in alcohol-impaired-driving crashes in 2019 (NHTSA, 2021). This equates to almost 28 fatalities per day, or approximately one person killed every 52 minutes due to alcohol-impaired-driving. The same NHTSA report estimates that the economic costs of alcohol-impaired-driving related fatalities was $44 billion in 2019. Additionally, traffic related fatalities in urban areas have increased by 34% from 2010 to 2019. Urban governments, such as the City of Philadelphia, must face these alarming statistics and enact policy focused on preventing alcohol-impaired-driving and the resulting fatalities.

From 2008 to 2012, there were 53,260 car crashes in the City of Philadelphia, 1,369 of which resulted in one or more of the drivers dying or being seriously injured, and 188 of those in which alcohol-impaired-driving was involved. In this study, we used the statistical programming language, R, to build a logistic model to identify predictors of alcohol-impaired-driving. Predictors studied included the type of collision (head-on, sideswipe, etc.), whether a cell phone was being used by a driver, and whether a driver was speeding or not. We found that crash types such as rear-ending, crashes at an angle, and hitting a fixed-object were associated with alcohol-impaired driving. We also found that alcohol-impaired-driving was not associated with driver demographics (drivers who are 16-17 or over 65) or driving while distracted or driving aggressively.

Methods

Limitations of OLS

An Ordinary Least Squares (OLS) regression is used when the dependent variable is continuous. OLS produces a slope-intercept form equation that results in a predicted value of the dependent variable. OLS is not appropriate for use when the dependent variable is binary (0/1, True/False, etc.) because the model would contain coefficients and predicted values of the dependent variable for which would result in non-sensical interpretations. For instance, if we are trying to predict whether there is a library in a certain neighborhood (0 = no library present, 1 = present), an OLS equation would indicate that as some predictor variable, perhaps the total population of the neighborhood, increases by one unit, the dependent variable changes by the β-coefficient of the predictor variable. This could result in cases in which a fractional presence or absence of a library was predicted to occur in a particular neighborhood as the total population increased or decreased. However, since we are predicting whether a library will be present or not, and not the number of libraries in each neighborhood, OLS is not useful for these types of situations. For this reason, we use a different type of model known as a logistic regression.

Logistic Regression

The odds is the ratio of the probability a thing will happen over the probability it won't. For example. Odds of event Y=1 can be calculated as:

$$\boldsymbol{Odds(Y=1) = \frac{\#\, event\ Y=1}{\#\, event\ Y \neq 1} = \frac{P(Y=1)}{P(Y \neq 1)} = \frac{p}{1-p}}$$

The odds ratio is the ratio of odds under two different conditions, a way to present the strength of association between risk factors/exposures and outcomes. The Odds Ratio can be calculated as:

$$\boldsymbol{Odds\ Ratio(Y=1) = \frac{\#\, odds\ Y=1, X=a+1}{\#\, odds\ Y=1, X=a} = \frac{Odds(Y=1|X=a+1)}{Odds(Y=1|X=a)}}$$

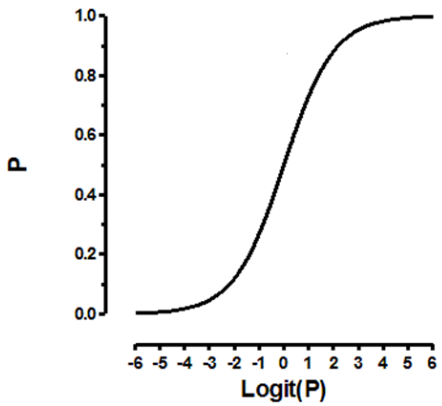
If Odds Ratio>1, that means greater odds of association with the exposure and outcome.
If Odds Ratio=1, that means there is no association between exposure and outcome.
If Odds Ratio<1, that means there is a lower odd of association between the exposure and outcome.

The logit function is the natural log of the odds that Y equals one of the outcomes. The regression equation for the logit model with multiple predictors is listed below.

$$\boldsymbol{Logit\ Form: ln\left(\frac{P_{(DRINKING_D)}}{1-P_{(DRINKING\_D)}}\right)}$$
$$\boldsymbol{= \beta_0 + \beta_1 FATAL_{OR_M} + \beta_2 OVERTURNED + \beta_3 CELL_{PHONE} + \beta_4 SPEEDING + \beta_5 AGGRESSIVE}$$
$$\boldsymbol{+ \beta_6 DRIVER1617 + \beta_7 DRIVER65PLUS + \beta_8 PCTBACHMOR + \beta_9 MEDHHINC + \varepsilon}$$

In the logit form equation, β_0 is the Y intercept (The odds that a driver whose crash didn't result in fatality or major injury, didn't involve an overturned vehicle, didn't involve speeding car, didn't involve aggressive driving, didn't involve at least one driver who was 16 or 17 years old, didn't involve at least one driver who was at least 65 years old and who was not using cell phone will be a drinking driver). $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9$ are coefficients of variables FATAL_OR_M, OVERTURNED, CELL_PHONE, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS, PCTBACHMOR, and MEDHHINC. The beta coefficients are the expected change in the log odds of the dependent variable being on TRUE or 1 or YES as the independent variable increases by one unit (holding other variables constant). ε is the residual (Present in the Logit Form). P is the probability of a car accident driver being a drinking driver.

The logistic function is the inversed logit function. A logistic function is a common S-shaped curve, shown as follows.



For our study with multiple predictors, the logistic function is listed as follows.

$$\boldsymbol{Logistic\ Form}: \boldsymbol{P_{(DRINKING_D)}}$$

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

If $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{1}} + \cdots + \widehat{\boldsymbol{\beta}}_{\boldsymbol{9}}\boldsymbol{x}_{\boldsymbol{9}} = \boldsymbol{0}$, then p=0.5 which means the probability of each outcome of y is equal to 0.5. As $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{1}} + \cdots + \widehat{\boldsymbol{\beta}}_{\boldsymbol{9}}\boldsymbol{x}_{\boldsymbol{9}}$ gets big, p approaches 1, and as $\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{1}} + \cdots + \widehat{\boldsymbol{\beta}}_{\boldsymbol{9}}\boldsymbol{x}_{\boldsymbol{9}}$ gets small, p approaches 0. This is exactly the type of "translator" function which successfully makes the interpretation of a binary dependent variable meaningful.

Hypothesis for each predictor

Then we do the hypotheses test for each predictor x_i with the null hypotheses that the coefficient for x_i (β_i) is equal to 0, and alternative hypotheses that β_i is not equal to 0. The quantity $\frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}-\boldsymbol{E}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}})}{\boldsymbol{\sigma}_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}}} = \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}-\boldsymbol{0}}{\boldsymbol{\sigma}_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}}} = \frac{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}}{\boldsymbol{\sigma}_{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}}} = \boldsymbol{z}$ is called Wald statistic which has a standard normal distribution, and the p-value for each z can be obtained. Odd ratios (OR), which can be calculated by e^{β_i}, are most commonly examined by statisticians rather than estimated coefficients to interpretate the regression results.

Assessment of the model fit

To evaluate the fitness of models, logistic regression is capable of calculating a R-squared, but it is no longer a very useful metric and does not have the same interpretation as OLS. Instead Akaike Information Criterion (AIC) are more frequently used to evaluate the quality of model fit. In general, the Akaike information criterion (AIC) is an estimator of out-of-sample prediction errors and thus of the relative quality of statistical models. Taking a set of models for data, the AIC calculates the quality of each model, relative to the others, and finally provides a means for model selection. Lower AIC values indicate a better-fit model, and a

model with a delta-AIC (the difference between the two AIC values being compared) of more than -2 is considered significantly better than the model it is being compared to.

On top of AIC, specificity, sensitivity, and the misclassification rate can also describe the quality of logistic models. Sensitivity (also known as the true positive rate) measures the proportion of actual positives which are correctly identified as such, and is complementary to the false negative rate. Similarly, specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such, and is complementary to the false positive rate. Higher sensitivity and specificity values indicate better models. In addition, misclassification rate is the incorrect predictions proportion of total predictions, which will be lower for better fitted models. Here are the calculations for these three metrics:

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{Ture\ positive}}{\boldsymbol{Predicted\ positive}} \quad \boldsymbol{Specificity} = \frac{\boldsymbol{Ture\ negative}}{\boldsymbol{Predicted\ negative}} \quad \boldsymbol{Misclassification\ Rate} = \frac{\boldsymbol{False\ prediction}}{\boldsymbol{Total\ prediction}}$$

When calculating the specificity, sensitivity and the misclassification rate, cut-offs are needed to identify what are relatively high probabilities. Considering that a best cut-off value may be determined by optimizing sensitivity, specificity and misclassification rate, we need try using a bunch of different cut-offs to find out the one with best accuracy. Before identifying the cut-off, the possibility of Y = 1 (i.e. the predicted value of y) should be calculated by : $\hat{\boldsymbol{y}} = \boldsymbol{p}(\boldsymbol{Y} = \boldsymbol{1}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i}}$, where x_i is the value of predictors and β_i is the coefficient of x_i.

To calculate the optimal cut-off, the ROC curve method will be applied in this report. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. ROC curve plots two parameters - true positive rate and false positive rate. There are various ways to target the optimal cut-off, including Youden Index (A cut-off for which the sum of Sensitivity and Specificity is maximized, a cut-off for which the ROC curve has the minimum distance from the upper left corner of the graph (i.e., the point at which specificity = 1 and sensitivity = 1). Here, the later method will be applied.

At last, to measure the prediction accuracy of the model, area under ROC Curve (AUC) need to be calculated. And higher AUCs are indicative of cut-off values for which both sensitivity and specificity of the model are relatively high. Here is a rough guide for classifying the accuracy: 0.9 -1 is excellent; 0.8 - 0.9 is good; 0.7 - 0.8 is fair; 0.6 - 0.7 is poor; 0.5 - 0.6 is fail.

The assumptions of logistic regression

- **Dependent Variable must be binary**
- **Independence of observations**
- **No severe multicollinearity**
- **Larger samples**

Because MLE (and not least squares) is used to estimate regression coefficients. At least 50 observations per predictor are needed. In the logistic regression, some assumptions of linear regression are still held such as "Independence of observations" and "No severe multicollinearity". But in Logistic Regression. There's no assumption that there needs to be a linear relationship between dependent variable and each independent variable. But logit of the independent variable should be linear correlated with dependent variable. And there are also no assumptions about homoscedasticity, and residuals don't need to be normal.

Exploratory Analyses

The exploratory analyses used to prepare for a logistic regression differ from those that precede OLS. When both the dependent and predictor variables are categorical, it is useful to identify the proportion of occurrences of a predictor that are associated with the dichotomous outcomes of the dependent variable. These proportions were tabulated using the CrossTable function in R. Disparities in these proportions shed light on the relationship between the predictor variables and the dependent variable. Furthermore, while Pearson's correlation is the statistical test used to measure the linear correlation between two continuous variables, the Chi-Square (χ^2) test is used when the variables are categorical. For instance, if we were to observe the cross-tabulation of the variables **CELL_PHONE** and **OVERTURNED**, the null and alternative hypotheses for the χ^2 test are as follows:

> $\boldsymbol{H_0}$: the proportion of overturned vehicles in crashes in which the driver was using a cell phone is the same as the proportion of overturned vehicles in crashes in which the driver was not using a cell phone.
>
> *vs*.
>
> $\boldsymbol{H_a}$: the proportion of overturned vehicles for crashes in which the driver was using a cell phone is different than the proportion of overturned vehicles in crashes in which the driver was not using a cell phone.

When identifying the correlation between a continuous predictor variable and a dichotomous dependent variable, we calculate the means of the continuous variable associated with the two possible outcomes of the dependent variable. The independent samples t-test is the significance test used for this comparison. For instance, the null and alternative hypothesis for the independent samples t-test of the two continuous predictors in our study, **PCTBACHMOR** and **MEDHHINC**, are as follows:

> $\boldsymbol{H_0}$: the average values of the variable **PCTBACHMOR** (or **MEDHHINC)** are the same for crashes that involve drunk drivers and crashes that do not.
>
> *vs*.
>
> $\boldsymbol{H_a}$: the average values of the variable **PCTBACHMOR** (or **MEDHHINC)** are different for crashes that involve drunk drivers and crashes that do not.

Results

Exploratory Analyses

In this model, we will explore the relationship between **DRINKING_D**, Whether the driver involving the car cash has drunk, and nine independent variables. As the tabulation of the drinking driver indicator (1 = Yes, 0 = No) attached below, 94.3% of crashes, which is made up by 40879 cases, did not involve drunk driving, while around 5.7% of Automobile crashes, or 2485 crashes, was caused by drunk drivers.

	No Alcohol Involved	**Alcohol Involved**
	(DRINKING_D = 0)	**(DRINKING_D = 1)**
Total Number	40879	2485
Proportion	0.9426944	0.0573056

There are seven binary predictors in this regression, which has been listed in a cross-tabulation below. With the p-value of the Chi-Square tests that are less than 0.001, six out of seven binary independent variables, except **CELL_PHONE** that stands for whether driver was using cell phone, are significantly related with the dependent variable. Thus, for these six predictors, we can reject the Null Hypothesis that no relationship exists between the predictor and dependent variable, which means they are not independent. For the insignificant binary predictor, **CELL_PHONE,** we failed to reject the Null Hypothesis, and there is no significant relationship between the cell phone behavior and drunk driving, considering that its χ^2 p-value is 0.687 which is far larger than any significance level.

	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	0.000
OVERTURNED: Crash involved an overturned vehicle	615	0.015	110	0.044	722	0.000
CELL_PHONE: Driver was using cell phone	426	0.01	28	0.011	454	0.687

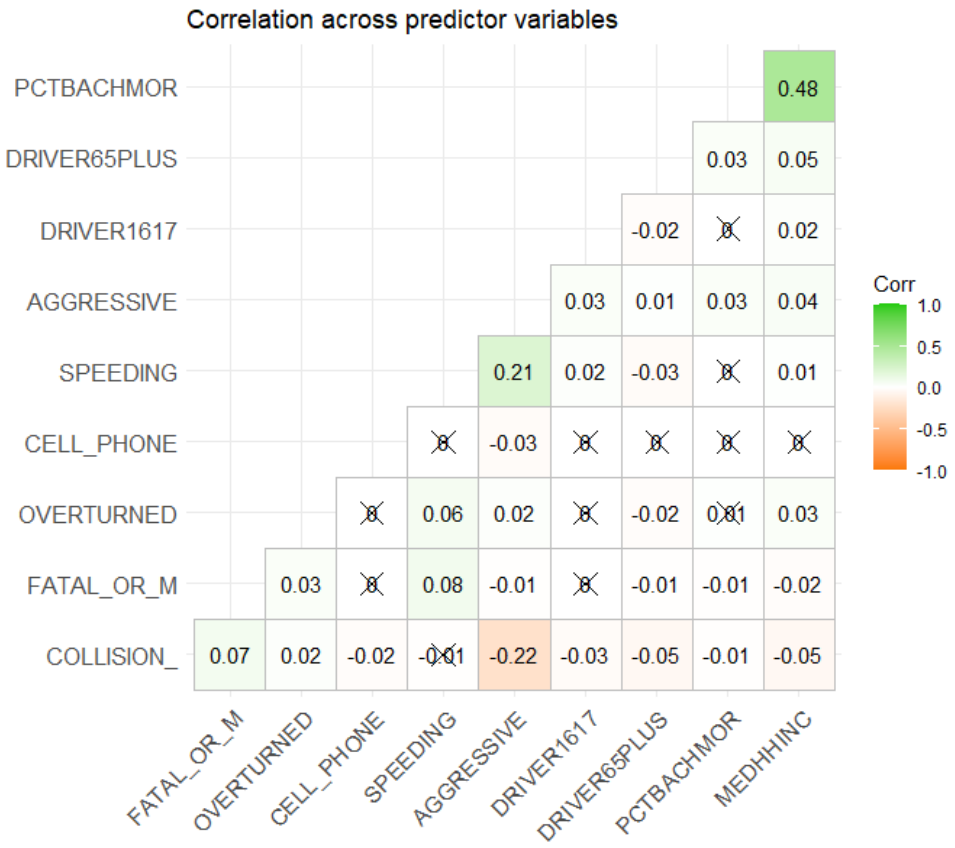
SPEEDING: Crash involved speeding car	1261	0.031	260	0.105	1521	0.000
AGGRESSIVE: Crash involved aggressive driving	18522	0.4533	916	0.369	19438	0.000
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old	674	0.016	12	0.005	686	0.000
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old	4237	0.104	2485	0.057	43364	0.000

For continuous predictors, the independent samples t-test are applied to evaluate their relationships with the dependent variable, and the result of t-tests have been printed below. Given that both p-values of t-tests are larger than 0.05 or 0.01, we have failed to reject the Null Hypothesis that the population means of two different groups, No Alcohol Involved and Alcohol Involved, are equal. So, both **PCTBACHMOR** and **MEDHHINC** are not significantly correlated with the dependent variable, **DRINKING_D.**

	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

Assumptions of Logistic Regression

Recall that the main assumptions of logistic regression are dichotomy of the dependent variable, independence of observations, no severe multicollinearity, and the requirement of at least 50 observations per predictor. The first and third assumptions were met – our dataset contained a dichotomous dependent variable (**DRINKING_B**) and there were 43,364 observations compared to the required 350 observations for seven predictors. Multicollinearity is the severe ($r > 0.8$ or < -0.8) correlation between predictor variables. To



determine if there is multicollinearity between the predictors, the correlation matrix is presented in the figure below.

Since there are no r values that are either greater than 0.8 or less than -0.8, it was determined that there was no multicollinearity between the predictor variables. However, it is important to remember that these Pearson's correlation coefficients are measuring the strength of the linear relationship between predictors. To identify whether the predictors are related to each other, a Chi-square test is used, as explained above.

The logistic regression results

a) The Logistic regression with numeric variables

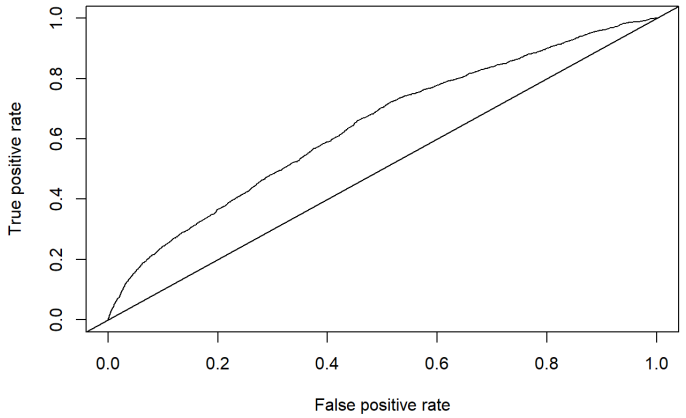
```
glm(formula = DRINKING_D ~ FATAL_OR_M + OVERTURNED + CELL_PHONE +
    SPEEDING + AGGRESSIVE + DRIVER1617 + DRIVER65PLUS + PCTBACHMOR +
    MEDHHINC, family = "binomial", data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1945  -0.3693  -0.3471  -0.2731   3.0099

Coefficients:
               Estimate   Std. Error z value             Pr(>|z|)
(Intercept)  -2.732506616  0.045875659 -59.563 < 0.0000000000000002 ***
FATAL_OR_M    0.814013802  0.083806924   9.713 < 0.0000000000000002 ***
OVERTURNED    0.928921376  0.109166324   8.509 < 0.0000000000000002 ***
CELL_PHONE    0.029550085  0.197777821   0.149               0.8812
SPEEDING      1.538975665  0.080545894  19.107 < 0.0000000000000002 ***
AGGRESSIVE   -0.596915946  0.047779238 -12.493 < 0.0000000000000002 ***
DRIVER1617   -1.280295964  0.293147168  -4.367 0.000012572447127937 ***
DRIVER65PLUS -0.774664640  0.095858315  -8.081 0.000000000000000641 ***
PCTBACHMOR   -0.000370634  0.001296387  -0.286               0.7750
MEDHHINC      0.000002804  0.000001341   2.091               0.0365 *
---
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: 18340  on 43354  degrees of freedom
AIC: 18360

Number of Fisher Scoring iterations: 6
```

	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

The result of logistic regression with all predictors is displayed above. From the result, we can see that except CELL_PHONE and PCTBACHMOR, other variables (FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS, and MEDHHINC) are significant. The OR of FATAL_OR_M is 2.2569, which means when the crash resulted in fatality or major injury (and the values of other independent variables don't change), the odds of accidents related to drunk driving go up by 1.1569%. The OR of OVERTURNED is 2.25317, which means when the crash involved an overturned vehicle (and the values of other independent variables don't change), the odds of accidents related to drunk driving go up by 115.31%. The OR of AGGRESSIVE is 0.5505, which means when the crash involved aggressive driving (and the values of other independent variables don't change), the odds of accidents related to drunk driving go down by 44.95%. The OR of DRIVER1617 is 0.27795, which means when the crash involved at least one driver who was 16 or 17 years old (and the values of other independent variables don't change), the odds of accidents related to drunk driving go down by 72.205%. The OR of DRIVER65PLUS is 0.46085, which means when the crash involved at least one driver who was at least 65 years old (and the values of other independent variables don't change), the odds of accidents related to drunk driving go down by 53.915%. The OR of MEDHHINC is 1.0000028, which means when the median household income increase by one dollar (and the values of other independent variables don't change), the odds of accidents related to drunk driving go down by 0.00028%.

And the specificity, sensitivity and the misclassification rates for the different probability cut-offs are displayed as follows. In the table we can see that the cut-off value 0.5 yields the lowest misclassification rate 0.057, the cut-off value 0.02 yields the highest misclassification rate 0.889.

Cut-off Value	Sensitivity (True Positive)	Specificity (True Negative)	Misclassification Rate
0.02	0.983501006	0.058073828	0.888894013
0.03	0.980684105	0.063926606	0.883543954
0.05	0.734808853	0.46909171	0.51568121
0.07	0.221327968	0.913908257	0.125864773
0.08	0.184708249	0.938715596	0.104579836
0.09	0.168209256	0.945962475	0.09860714
0.1	0.164185111	0.948213019	0.09671617
0.15	0.104225352	0.972210671	0.077529748
0.2	0.022937626	0.995376599	0.060349599
0.5	0.001609658	0.99990215	0.057305599



The ROC curve is displayed above and the optimal cut-off rate, which minimizes the distance from the upper left corner of the ROC curve, is 0.06365151. In this cut-off value, the sensitivity (True Positive) is 0.66076, the specificity (True Negative) is 0.54524. In the above table, the aim is to look for the minimum value of mis-classification rates. Therefore, the optimal cut-off value which yields smallest misclassification rate is 0.5. And in the ROC curve, the aim is to look for the simultaneous maximum values of sensitivity and specificity. Therefore, the optimal cut-off value which yields simultaneously the maximum values of sensitivity and specificity is 0.06.

The area under the ROC curve is 0.6398695. The prediction accuracy of the model depends on how well the model predicts 1 responses as 1's and 0 responses as 0's. Accuracy is measured by the area under the ROC curve. An area of 1 represents a perfect test (prediction); an area of 0.5 represents a worthless test (prediction). 0.6398, the area under the ROC Curve in this model, means that the model represents a poor test (prediction). A rough guide for interpreting area under ROC Curves is displayed as follows.

- 0.90 - 1 = excellent (A)

- 0.80 - 0.90 = good (B)
- 0.70 - 0.80 = fair (C)
- 0.60 - 0.70 = poor (D)
- 0.50 - 0.60 = fail (F)

b) The Logistic regression only with binary variables

```
Call:
glm(formula = DRINKING_D ~ FATAL_OR_M + OVERTURNED + CELL_PHONE +
    SPEEDING + AGGRESSIVE + DRIVER1617 + DRIVER65PLUS, family = "binomial",
    data = mydata)

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 < 0.0000000000000002 ***
FATAL_OR_M    0.80932    0.08376   9.662 < 0.0000000000000002 ***
OVERTURNED    0.93978    0.10903   8.619 < 0.0000000000000002 ***
CELL_PHONE    0.03107    0.19777   0.157                0.875
SPEEDING      1.54032    0.08053  19.128 < 0.0000000000000002 ***
AGGRESSIVE   -0.59365    0.04775 -12.433 < 0.0000000000000002 ***
DRIVER1617   -1.27158    0.29311  -4.338  0.00001436374143293 ***
DRIVER65PLUS -0.76646    0.09576  -8.004  0.00000000000000121 ***
---
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

Number of Fisher Scoring iterations: 6
```

	OR	2.5 %	97.5 %
(Intercept)	0.07051713	0.06678642	0.0743978
FATAL_OR_M	2.24636998	1.90112455	2.6404533
OVERTURNED	2.55942903	2.05736015	3.1556897
CELL_PHONE	1.03156149	0.68459779	1.4907150
SPEEDING	4.66608472	3.97961862	5.4573472
AGGRESSIVE	0.55230941	0.50268818	0.6061758
DRIVER1617	0.28038936	0.14904734	0.4751771
DRIVER65PLUS	0.46465631	0.38318289	0.5579332

The results of the logistic regression with the binary predictors only (without PCTBACHMOR and MEDHHINC) are displayed above. There are not any predictors which are significant in the new model which weren't significant in the original one, or vice versa. From the result, we can see that except CELL_PHONE, other variables (FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, and DRIVER65PLUS) are significant. The AICs for both models(a & b) are the same as 18360, which means these two models have same prediction ability.

Discussion

In this study, we built a logistic model to predict accidents related to drunk driving in the City of Philadelphia, Pennsylvania from 2008 to 2012. Results indicate that use of a cell phone (**CELL_PHONE**) and the demographics of the crash (**PCTBACHMOR** and **MEDHHINC**) were not associated with drunk driving, however, when included in the logistic regression, we found that **MEDHHINC** was a significant predictor of drunk driving. Conversely, **FATAL_OR_M**, **OVERTURNED**, **SPEEDING**, **AGGRESSIVE**, **DRIVER1617**, **DRIVER65PLUS**, and **MEDHHINC** were significant predictors of drunk driving. We were particularly surprised that the use of a cell phone was not associated with drunk driving crashes. Perhaps we have been conditioned to associate vehicle crashes with the use of cell phones, and as such, we thought that there would be a relationship. However, it is possible that a drunk driver who causes a crash is too intoxicated to use a cell phone while driving.

While the proportion of drunk driving to total number of crashes is small, the conventional logistic regression is an appropriate method when compared to the benefits of using Allison's modeling rare events method because the number of drunk driving crashes is greater than 2000 out of over 40,000 total crashes.