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

Fall 2025

MUSA 5000

Jun Luu

Alex Stauffer

Tessa Vu

1. Introduction

Traffic crashes are one of the leading public health challenges in major metropolitan areas throughout the United States (Centers for Disease Control and Prevention, 2024). Within this broader issue, alcohol-impaired driving stands out as a particularly preventable yet deadly factor. The U.S. Department of Transportation reports that approximately 30 people die each day in crashes involving an alcohol-impaired driver, or one every 51 minutes. Beyond the human toll, the National Highway Traffic Safety Administration estimates that alcohol-related crashes impose an economic burden exceeding $59 billion annually. These statistics show the urgency in which we need to identify and understand the factors that contribute to alcohol-involved crashes, particularly at the local level where targeted interventions can be most effective.

Philadelphia presents a compelling context for examining these dynamics. As a densely populated urban center with diverse neighborhoods, varying socioeconomic conditions, and complex traffic patterns, the city experiences thousands of crashes annually. Understanding which characteristics of crashes and their surrounding environments are associated with alcohol involvement can inform targeted prevention strategies by transportation planners, law enforcement agencies, and public health officials. This analysis examines traffic crash data from Philadelphia covering 2008 through 2012 to investigate the relationship between various crash characteristics, driver behaviors, and neighborhood-level factors with the likelihood of alcohol involvement.

This paper 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. Crashes resulting in fatalities or major injuries often signal more severe impacts that may result from the impaired judgment and delayed reaction times characteristic of intoxicated drivers. Behaviors such as speeding and aggressive driving frequently occur with alcohol use, as impairment reduces inhibitions and risk perception. Vehicles overturning during a crash may indicate loss of control consistent with impaired driving. Conversely, cell phone use, while dangerous, represents a distinct form of distraction that may not correlate strongly with alcohol involvement. Driver age also warrants examination: younger drivers aged 16-17 are less experienced and may be more prone to risk-taking behaviors that overlap with underage drinking, while older drivers aged 65 and above typically exhibit lower rates of alcohol involvement due to different lifestyle patterns and health considerations.

Beyond individual crash characteristics, neighborhood socioeconomic factors may influence alcohol-related crash patterns. Areas with higher percentages of residents holding bachelor's degrees or higher may demonstrate greater awareness of drunk driving risks and consequences. Similarly, median household income could relate to alcohol-involved crashes through multiple pathways: higher-income neighborhoods might have better access to alternative transportation

options like rideshare services, while lower-income areas might face different patterns of alcohol availability and consumption. By examining these neighborhood-level variables alongside crash-specific factors, this analysis adopts a multilevel perspective that recognizes both individual behaviors and contextual influences.

This report uses logistic regression analysis conducted in R to examine these relationships. Logistic regression is the appropriate statistical technique when the outcome variable is binary. The analysis will assess which of the selected predictors show statistically significant associations with alcohol involvement and quantify the strength of these relationships through odds ratios.

2. Methods

Limitations of Ordinary Least Squares Regression for Binary Outcomes

When the dependent variable is binary, ordinary least squares (OLS) regression presents several problems. The standard OLS model takes the form:

$$y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

With a binary dependent variable, this approach encounters multiple violations of core regression assumptions. First, the interpretation of coefficients becomes problematic. In OLS, β_1 represents the amount of the dependent variable, y, changes per one-unit increase in x_1, holding other variables constant. However, when y can only take values of 0 or 1, changing y by β_1, no longer makes sense considering the variable can only switch between two states. Also, OLS assumes a linear relationship between predictors and the outcome, an assumption incompatible with binary outcomes. OLS requires normally distributed residuals, yet with binary outcomes, residuals follow a binomial distribution rather than a normal distribution. Finally, OLS assumes homoscedasticity—constant variance of residuals across all predictor values. With binary outcomes, however, residual variance necessarily changes with the predicted probability, creating heteroscedasticity. These violations compromise both the validity of hypothesis tests and the interpretability of results, necessitating an alternative modeling approach.

Logistic Regression

Logistic regression addresses these issues by modeling the probability of the outcome rather than the outcome itself. This approach builds on the concept of odds. The odds of an event represents the ratio of the probability of occurrence to the probability of non-occurrence:

$$Odds = \frac{p}{1-p}$$

where p is the probability that the event occurs. Odds range from 0 (event never occurs) to infinity (event always occurs).

An odds ratio compares the odds of an event occurring under different conditions. Rather than modeling probability directly, logistic regression models the natural logarithm of the odds, called the log-odds or logit. The logit model with multiple predictors takes the form:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

In this model, these variables would be represented:

$$\ln\left(\frac{P(DRINKING_D=1)}{1-P(DRINKING_D=1)}\right) = \beta_0 + \beta_1(\text{FATAL\_OR\_M}) + \ldots + \beta_k(\text{MEDHHINC}) + \epsilon$$

Where P(DRINKING_D=1) represents the probability that a crash involved a drinking driver. The term $\ln\left(\frac{P(DRINKING_D=1)}{1-P(DRINKING_D=1)}\right)$ is the log-odds of alcohol involvement, β_0 is the intercept representing the log-odds when all predictors equal zero, and β_1, β_2, ..., β_9 are coefficients indicating the change in log-odds for a one-unit increase in each predictor, holding others constant.

The predictors in this model include:

- **FATAL_OR_M**: whether the crash resulted in a fatality or major injury
- **OVERTURNED**: whether the crash involved an overturned vehicle
- **CELL_PHONE**: whether the driver was using a cell phone
- **SPEEDING**: whether the crash involved speeding
- **AGGRESSIVE**: whether the crash involved aggressive driving
- **DRIVER1617**: whether at least one driver was aged 16-17
- **DRIVER65PLUS**: whether at least one driver was aged 65 or older
- **PCTBACHMOR**: percentage of residents with a bachelor's degree or higher in the crash location's census block group
- **MEDHHINC**: median household income in the crash location's census block group

The logistic function is the inverse of the logit function. It transforms log-odds back into probabilities, providing the probability form of the model:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}$$

For this analysis:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 FATALORM+\beta_2 OVERTURNED+\cdots+\beta_k MEDHHINC)}}$$

The logistic function produces an S-shaped curve with several properties that make it ideal for binary outcomes. It constrains all predicted probabilities to the interval [0,1]. As the linear predictor approaches negative infinity, the probability approaches 0 and as it approaches positive infinity, the probability approaches 1. This ensures that predicted values always represent valid probabilities. Also, the function is symmetric around 0.5 when the linear predictor equals zero, and the predicted probability equals 0.5, representing equal likelihood of either outcome. These properties make the logistic function well-suited for modeling binary dependent variables, as it restricts predicted probabilities to the appropriate range while allowing each coefficient βi to represent the effect of a predictor on the log-odds of the outcome.

Hypothesis Testing for Predictors

For each predictor in the logistic regression model, we test whether it has a statistically significant effect on the outcome. The null hypothesis states that the predictor has no effect:

$$H_0: \beta_i = 0$$

The alternative hypothesis states that the predictor does have an effect:

$$H_a: \beta_i \neq 0$$

The Wald test assesses the significance of individual coefficients. The Wald statistic is calculated as:

$$W_i = \frac{\hat{\beta_i}}{\sigma(\hat{\beta_i})}$$

$\hat{\beta_i}$ is the estimated coefficient for predictor i and $\sigma(\hat{\beta_i})$ is its standard error. Under the null hypothesis, the Wald statistic follows a standard normal distribution, $N(0,1)$. Large absolute values of the Wald statistic indicate that β_i differs significantly from zero, providing evidence to reject the null hypothesis and conclude that the predictor has a significant effect on the outcome variable.

Rather than interpreting raw β coefficients, most statisticians prefer to examine odds ratios, which are calculated by exponentiating the coefficients: $OR = e^{\beta_i}$. An odds ratio quantifies how the odds of the outcome change for a one-unit increase in the predictor, holding all other predictors constant. For instance, an odds ratio of 2.3 indicates that the odds of the outcome are 2.3 times higher for each one-unit increase in the predictor. Odds ratios greater than 1 indicate that the predictor is associated with increased likelihood of the outcome. Odds ratios less than 1 indicate decreased likelihood, and an odds ratio of exactly 1 indicates no effect.

Assessing Quality of Model Fit

An R-squared value can be calculated for logistic regression; however, it is no longer a very useful metric and does not have the same interpretation as in OLS. In OLS, R-squared represents the proportion of variance in the outcome explained by the model. This interpretation does not apply to logistic regression because the model uses maximum likelihood estimation rather than minimizing squared residuals, and there is no direct concept of variance in a binary outcome variable to be explained in the same way. Therefore, R-squared values in logistic regression are not directly comparable to OLS R-squared and are rarely used as primary indicators of model fit.

The Akaike Information Criterion (AIC) is commonly used to compare models in logistic regression. AIC is an estimator of prediction error and the relative quality of statistical models for a given dataset. It balances the goodness of fit against model complexity by penalizing models with more parameters. The AIC is defined as:

$$AIC = 2k + 2\ln(L^{\wedge})$$

k is the number of estimated parameters in the model and L^ is the maximum value of the likelihood function for the model. A lower AIC value indicates a better fit, as it reflects an appropriate balance between model complexity and goodness of fit.

In logistic regression, fitted (predicted) values $y_i^{\wedge}$ represent the estimated probability that Y=1 for each observation. These probabilities are calculated using the logistic function:

$$y_i^{\wedge} = P(Y = 1) = \frac{1}{1 + \exp(-\beta_0^{\wedge} - \beta_1^{\wedge} x_{1i} - \cdots - \beta_k^{\wedge} x_{ki})}$$

$\beta_0^{\wedge}, \beta_1^{\wedge}, \ldots, \beta_k^{\wedge}$ are the estimated coefficients obtained from the logistic regression model, and $x_{1i}, x_{2i}, \ldots x_{ki}$ are the values of the predictor variables for observation i . To classify observations into binary outcomes, we apply a threshold or cut-off value to these probabilities. For example, observations with $y_i^{\wedge} > 0.5$ might be classified as Y=1, while those with $y_i^{\wedge} \leq 0.5$ are classified as Y= 0. However, it is important to try using different cut-offs for what is considered a "high" probability of Y= 1, as the choice of threshold can significantly impact model performance.

Classification performance is evaluated using several metrics:

1. Sensitivity, also called the true positive rate, measures the proportion of actual positives correctly identified:

$$Sensitivity = \frac{TP}{TP + FN}$$

   TP represents true positives, correctly predicted positives, and FN represents false negatives, actual positives incorrectly predicted as negatives. Higher values of sensitivity are better, indicating that the model successfully identifies a larger proportion of actual positive cases.

2. Specificity, the true negative rate, measures the proportion of actual negatives correctly identified:

$$Specificity = \frac{TN}{TN + FP}$$

TN represents true negatives, correctly predicted negatives, and FP represents false positives, actual negatives incorrectly predicted as positives. Higher values of specificity are better, indicating that the model successfully identifies a larger proportion of actual negative cases.

3. The misclassification rate indicates the proportion of incorrect predictions:

$$Misclassification = \frac{FN + FP}{TN + TP + FN + FP}$$

FN and FP are false negative and false positive counts, respectively. Dividing this by the total number of cases represents the rate at which they are incorrectly labeled. Lower values are better, as they reflect fewer incorrect predictions overall.

Using different cut-off values creates trade-offs between sensitivity and specificity and miscalculation. For example, a lower cut-off value may increase sensitivity but decrease specificity. This means that more actual positives are correctly identified, but more negatives are incorrectly classified as positive. Conversely, a higher cut-off value may increase specificity but decrease sensitivity. Fewer false positives occur, but more actual positives are missed. The cut-off should be chosen based on the specific context and the relative importance of avoiding false positives versus false negatives, the misclassification rate.

Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a graphical representation that plots sensitivity (true positive rate) against 1 – specificity (false positive rate) across all possible threshold values. The ROC curve helps assess the predictive quality of the model by illustrating the trade-off between sensitivity and specificity.

The Youden Index calculates the ideal cut-off threshold:

$$J = Specificity + Sensitivity$$

Another approach involves calculating the minimum distance from the upper-left corner of the ROC plot, where both sensitivity and specificity equal 1. For this report, we will be using the minimum distance approach to determine the optimal cut-off value.

Area Under the Curve

Area under ROC Curve (AUC, which stands for Area Under Curve) is a measure of prediction accuracy of the model. Higher AUCs mean that we can find a cut-off value for which both sensitivity and specificity of the model are relatively high. An AUC of 1.0 indicates perfect separation (the model always ranks positive cases higher), while an AUC of 0.5 suggests the model performs no better than random guessing.

Here is a rough guide for classifying the accuracy:

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

The AUC represents the probability that a model correctly ranks two randomly selected observations, one positive and one negative case. More specifically, if you randomly pick two observations where one is a positive example (outcome = 1) and the other is a negative example (outcome = 0), the AUC measures the likelihood that a model assigns a higher predicted probability to the positive case than to the negative case.

Review of Assumptions of Logistic Regression

Before performing any analyses, we must review the assumptions of OLS regression. The list of assumptions include:

1) independence of observations

2) linear relationship between the dependent variable and each predictor

3) normality of residuals

4) homoscedasticity

5) no multicollinearity

There are key differences between OLS regression assumptions and logistic regression assumptions. First, there is no assumption that the relationship between the dependent variable and each independent variable should be linear. The linearity of log odds of Y = 1 is assumed, but it is not something that is typically tested for in practice. Second, there is no assumption of homoscedasticity. Third, the residuals do not need to be normal. Another difference is that larger samples (~50) are needed for logistic regressions because of the MLE (maximum likelihood estimation), and not least squares. The dependent variable of logistic regressions must be binary.

The OLS and logistic regression share the assumptions of independence of observations and no severe multicollinearity.

Discussion of Exploratory Analysis

Before running a logistic regression, statisticians may want to perform exploratory analyses. Running cross-tabulations between the dependent variable and binary predictors allows

statisticians to determine whether there is an association between the two variables. For our purposes, this means we want to assess the dependent variable DRINKING_D alongside the binary predictor variables: FATAL_OR_M, OVERTURNED, CELL_PHONE, SPEEDING, AGGRESSIVE, DRIVER1617, and DRIVER65PLUS.

The appropriate statistical test for examining the association between two categorical variables is the Chi-Square (χ^2) test. For each categorical binary variable in our equation, we can look at the null and alternative hypotheses for the χ^2 test. Let's look at the null and alternative hypotheses for the cross-tabulations on the dependent variable DRINKING_D and binary variable AGGRESSIVE:

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

vs.

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

A high value of the χ^2 statistic and a p-value lower than 0.05 suggest there is evidence to reject the null hypothesis H_0 in favor of the alternative H_1, and that there is an association between drunk driving and aggressive-driving crashes. This is then repeated for all binary categorial variables in the regression.

We can also compare the means of continuous predictors for both values of the dependent variable. The independent samples' t-tests are the appropriate statistical tests for examining whether there were significant differences in mean values of PCTBACHMORE and MEDHHINC for crashes that involved alcohol and those that did not. With the t-test, we can see whether the average MEDHHINC values are statistically different for crashes that involve drunk drivers and crashes that don't. The null and alternative hypotheses for the independent samples t-test would be as follows:

> H_0: average values of the variable of MEDHHINC **are the same** for crashes that involve drunk drivers and crashes that don't
>
> vs.
>
> H_a: average values of the variable MEDHHINC **are different** for crashes that involve drunk drivers and crashes that don't

A high value t-statistic and a p-value lower than 0.05 suggest that there is evidence to reject the null hypothesis in favor of the alternative. The t-test must be repeated for all continous variables, in our case for PCTBACHMORE.

3. Results

Summary of Alcohol Consumption

Metric	No Alcohol	Alcohol
Count	40,879	2,485
Proportion	0.94	0.06

Table 1

In Table 1, the Summary of Alcohol Consumption, the results indicate that drunk driving crashes are rare events in comparison to ones that do not involve any drinking, this is a mere 6%. This is unsurprising because this dataset is dealing with a very specific outcome in comparison to a larger, more general population of crashes. Clearly, there is an imbalance here, and while it might seem a bit insignificant at first, the Introduction mentions this little 6% is part of a larger issue that costs billions annually in the US.

Cross-Tabulation of Predictors by Alcohol Involvement

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)			
Predictor	N	%	N	%	Total	X2 p-value
FATAL_OR_M	1,181	2.89	188	7.57	1,369	0.00
OVERTURNED	612	1.50	110	4.43	722	0.00
CELL_PHONE	426	1.04	28	1.13	454	0.69
SPEEDING	1,261	3.08	260	10.46	1,521	0.00
AGGRESSIVE	18,522	45.31	916	36.86	19,438	0.00
DRIVER1617	674	1.65	12	0.48	686	0.00
DRIVER65PLUS	4,237	10.36	119	4.79	4,356	0.00

Variable Definitions:

FATAL_OR_M: Crash resulted in fatality or major injury.

OVERTURNED: Crash involved an overturned vehicle.

CELL_PHONE: Driver was using cell phone.

SPEEDING: Crash involved speeding car.

AGGRESSIVE: Crash involved aggressive driving.

DRIVER1617: Crash involved at least one driver who was 16 or 17 years old.

DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old.

Table 2

In Table 2, the Cross-Tabulation of Predictors by Alcohol Involvement, the results show that all predictors *except* CELL_PHONE are statistically significant at $p < 0.05$, meaning to reject the null hypothesis that states there is no association between that specific predictor and the likelihood of a drunk driving crash, in favor of the alternative hypothesis that states the inverse (i.e. these predictors have an association with drunk driving crashes). CELL_PHONE, with its p-value of 0.69, is not statistically significant, meaning fail to reject the null hypothesis that states there is no association between it and drunk driving, so it seems cell phone usage is the only predictor in the list that looks like it is independent from drunk driving incidents.

Unsurprisingly, it looks like fatalities (FATAL_OR_M), OVERTUNED, and SPEEDING have a larger share in the drunk driving bucket than not, massively so for the latter, and cell phone usage also has a very slightly larger proportion, although considered statistically insignificant.

Means of Predictors by Alcohol Involvement

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		
Predictor	**Mean**	**SD**	**Mean**	**SD**	**t-test p-value**
MEDHHINC	31,483.05	16,930.10	31,998.75	17,810.50	0.16
PCTBACHMOR	16.57	18.21	16.61	18.72	0.91

Variable Definitions:
PCTBACHMOR: % with bachelor's degree or more.
MEDHHINC: Median household income.

Table 3

According to Table 3, the Means of Predictors by Alcohol Involvement, and observing MEDHHINC, it looks like $p > 0.05$, meaning fail to reject the null hypothesis that states the average values of the predictor are the same for both groups, so there is no statistically significant difference between MEDHHINC and PCTBACHMOR. Even though the mean of median household income is $515.70 more for drunk driving crashes, the p-value suggests that it is not strong enough determinant for it.

Looking to PCTBACHMOR, $p > 0.05$ at 0.91, which is much larger than that threshold, so this means to fail to reject the null hypothesis as well in favor for the alternative hypothesis, which states that the average values are different. The values for PCTBACHMOR between drunk-driving crash versus drunk driving crash is very marginal. So far, this suggests that maybe individual driver behaviors are stronger than socioeconomic predictors (i.e. speeding).

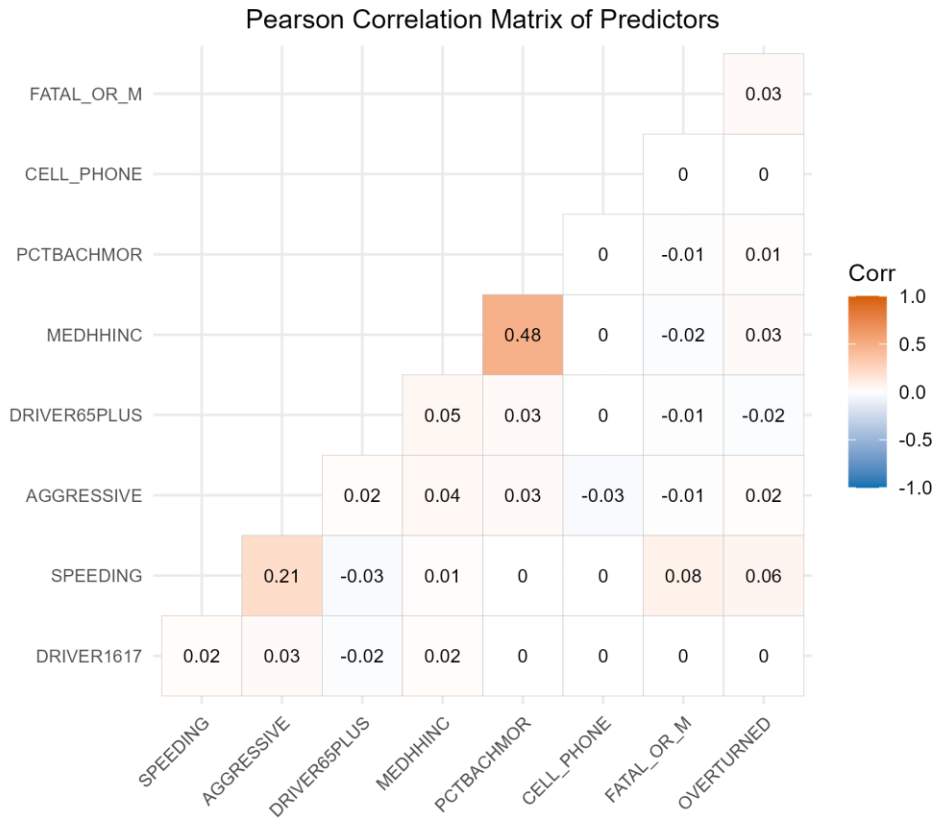

Figure 1

Logistic regression is a bit more relaxed that linear regression, and have four assumptions: first, the dependent variable must be binary; second, the observations must be independent; third, there is no severe multicollinearity; fourth, larger samples are needed in comparison to OLS regression because MLE is used to estimate regression coefficients—at least fifty observations per predictor are needed in comparison to OLS' ten. However, another differentiator is that the relationship between the dependent variable and each of the independent variables does not need to be lin and there is no assumption of homoscedasticity and the residuals do not need to be normal as well.

With the above in mind, the following assumptions in the context of this paper are as follows: DRINKING_D meets the assumption that the dependent variable must be binary; the outcome of one crash may influence another due to the fact that the CDC does aggregate its data geographically, and also the inherent spatial nature of crashes and how they can cause others nearby (i.e. rubbernecking a wreck while driving past, causing another due to immobility in the middle of a main road); there is no severe multicollinearity, but it should be noted that MEDHHINC and PCTBACHMOR have the most collinearity of the others. Each predictor has

more than fifty sample sizes. With the assumption of the log odds of drunk driving crashes being linearly related to the predictors, it was untested.

On multicollinearity, it is when two or more predictors in a multiple regression model are highly correlated with one another. The threshold is –0.90 or 0.90 to indicate severe multicollinearity, although a VIF test could be conducted to make sure, in the context of this paper, the highest Pearson correlation is 0.48, so there is no severe multicollinearity.

Full Logistic Regression Model

	Estimate	Standard Error	Z-Value	P-Value	Odds Ratio	2.5% CI	97.5% CI
(Intercept)	-2.73	0.05	-59.56	0.00	0.07	0.06	0.07
FATAL_OR_M	0.81	0.08	9.71	0.00	2.26	1.91	2.65
OVERTURNED	0.93	0.11	8.51	0.00	2.53	2.03	3.12
CELL_PHONE	0.03	0.20	0.15	0.88	1.03	0.68	1.49
SPEEDING	1.54	0.08	19.11	0.00	4.66	3.97	5.45
AGGRESSIVE	-0.60	0.05	-12.49	0.00	0.55	0.50	0.60
DRIVER1617	-1.28	0.29	-4.37	0.00	0.28	0.15	0.47
DRIVER65PLUS	-0.77	0.10	-8.08	0.00	0.46	0.38	0.55
PCTBACHMOR	0.00	0.00	-0.29	0.77	1.00	1.00	1.00
MEDHHINC	0.00	0.00	2.09	0.04	1.00	1.00	1.00

Variable Definitions:
FATAL_OR_M: Crash resulted in fatality or major injury.
OVERTURNED: Crash involved an overturned vehicle.
CELL_PHONE: Driver was using cell phone.
SPEEDING: Crash involved speeding car.
AGGRESSIVE: Crash involved aggressive driving.
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old.
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old.
PCTBACHMOR: % of individuals 25 years of age or older who have at least a bachelor's degree.
MEDHHINC: Median household income.

Table 4

Table 4, the Full Logistic Regression Model, presents seven predictors as statistically significant and two predictors as statistically insignificant. In other words, SPEEDING, OVERTURNED, FATAL_OR_M, MEDHHINC, AGGRESSIVE, DRIVER65PLUS, and DRIVER1617 are associated with drunk driving crashes controlling for all other predictors. On the note of these significant variables, SPEEDING is the strongest positive predictor with a 4.66 odds ratio that indicates speeding crashes are 4.66 times more likely to be associated with drunk driving compared to crashes without speeding, holding others constant. OVERTURNED and FATAL_OR_M are at 2.53 and 2.26 odds ratios, respectively, so crashes with overturned vehicles are 2.53 times more likely to be associated with drunk driving compared to crashes without speeding, and fatal crashes are 2.26 times more likely to be associated with drunk driving compared to non-fatal crashes. AGGRESSIVE driving with a 0.55 odds ratio, indicates aggressive driving is 45% less likely to be associated with drunk driving crashes, so while aggression is common, it is not a leading predictor with alcohol. MEDHHINC is at 1.00,

meaning no change because for every one-unit increase in median household income it increases by a factor of one, and it is also to be noted that it is considered significant by a margin at p = 0.04, making this negligible. Lastly and unsurprisingly, DRIVER65PLUS and DRIVER1617 are less likely to be associated with drunk driving crashes, with 0.46 and 0.28, respectively—older individuals may not drink as much as those younger than them, and kids aged 16 and 17 cannot legally drink, so that barrier likely attributes to the significant gap between the two's odd ratios. It is likely that individuals in their 20s (specifically undergraduate age bracket) share a proportion that is much higher, but this is an assumption.

The two statistically insignificant predictors are CELL_PHONE and PCTBACHMOR, and getting into the confidence intervals, they include the number 1.00 within their 2.5% and 97.5% bounds, which means they are not statistically significant, and this same pattern can be observed with MEDHHINC, which happened to be barely statistically significant as discussed in the previous paragraph.

Sensitivity, Specificity, and Misclassification Rates

Cutoff	Sensitivity	Specificity	Misclass
0.02	**0.98**	**0.06**	**0.89**
0.03	0.98	0.06	0.88
0.05	0.73	0.47	0.52
0.07	0.22	0.91	0.13
0.08	0.18	0.94	0.10
0.09	0.17	0.95	0.10
0.10	0.16	0.95	0.10
0.15	0.10	0.97	0.08
0.20	0.02	1.00	0.06
0.50	**0.00**	**1.00**	**0.06**

Table 5

Table 5, the Sensitivity, Specificity, and Misclassification Rates, show that the cutoff with the highest error is at 0.02 with a 0.89 misclassification rate. This means that the model is biased toward predicting that drunk driving is involved in a crash, which is associated with the high sensitivity at 0.98. Unfortunately, that is an issue, because it predicts DRINKING_D = 1 for every case, meaning it misclassifies almost all of the non-drunk driving cases (DRINKING_D = 0), associated with the low specificity at 0.06.

The table also shows the lowest error cutoff at 0.20 and 0.50 with a misclassification rate at 0.06. However, it is the 0.50 cutoff that achieves the perfect specificity at 1.00 and perfect sensitivity at 0.00—while this threshold correctly classifies crashes that do not involve alcohol completely, the specificity, it also completely fails to classify on crashes that involve alcohol, the sensitivity. On the other hand, the 0.20 cutoff has a meager 0.02 sensitivity in that regard. Rounding back to

the brief statement of observations in Table 1, the dataset is, again, imbalanced, which explains the dramatic values in Table 5.

While there is no single optimal cutoff rate, it will always depend on the context and the utility of the situation. Are people willing to accept more false negatives, predicting no alcohol was involved in a crash when there was indeed drunk driving? Or false positives, predicting alcohol was involved in a crash when there was no drunk driving? It would need to be questioned, what are the consequences of both and how are those consequences weighed? In this case, the 0.50 cutoff is unacceptable, especially in a public health context, and especially because it costs the US billions annually, but realistically it may be good to consider a 0.07 cutoff rate. This means a balance needs to be set and is discussed in more detail with Figure 2 below.

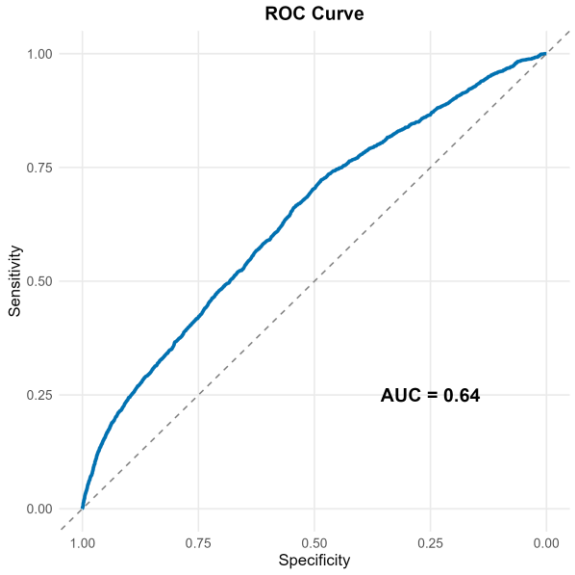

Figure 2

According to Figure 2, the Receiver Operating Characteristic (ROC) Curve, the optimal cutoff rate has to minimize the distance from the upper-left corner at (0, 1). In Table 5, the minimum classification rates were observed, but here Figure 2 provides a visual to simultaneously minimize both sensitivity and specificity, again, finding that statistical balance. From the image, it looks as if the point on the curve that has the least distance to (0, 1) on the graph is indeed (0.53, 0.73), which are specificity and sensitivity, respectively. This indicates that the statistically balanced cutoff is around 0.05, which closely matches the ROC curve values visually.

It can also be observed that this 0.05 cutoff sacrifices specificity compared to the observation in Table 5 declaring a 0.50 cutoff, which sacrifices sensitivity. Meaning 0.05 sacrifices a high number of false positives and 0.50 sacrifices all false negatives with misclassification rates at 0.52 and 0.06, respectively. On the graph, the 0.50 cutoff is in the bottom left corner.

Lastly, the Area Under the Curve (AUC) of the ROC line shows the model's discriminatory ability between the classes across all possible cutoff thresholds; the line represents the closest points outside that AUC to (0, 1). AUC is 0.64 here, and a value of 1 means that the model is perfect. So clearly, this model is not so. However, it does perform better than AUC at 0.50, which essentially means the model does perform any better than a random 50% chance.

Reduced Logistic Regression Model (Binary Predictors)

	Estimate	Standard Error	Z-Value	P-Value	Odds Ratio	2.5% CI	97.5% CI
(Intercept)	-2.65	0.03	-96.32	0.00	0.07	0.07	0.07
FATAL_OR_M	0.81	0.08	9.66	0.00	2.25	1.90	2.64
OVERTURNED	0.94	0.11	8.62	0.00	2.56	2.06	3.16
CELL_PHONE	0.03	0.20	0.16	0.88	1.03	0.68	1.49
SPEEDING	1.54	0.08	19.13	0.00	4.67	3.98	5.46
AGGRESSIVE	-0.59	0.05	-12.43	0.00	0.55	0.50	0.61
DRIVER1617	-1.27	0.29	-4.34	0.00	0.28	0.15	0.48
DRIVER65PLUS	-0.77	0.10	-8.00	0.00	0.46	0.38	0.56

Variable Definitions:
FATAL_OR_M: Crash resulted in fatality or major injury.
OVERTURNED: Crash involved an overturned vehicle.
CELL_PHONE: Driver was using cell phone.
SPEEDING: Crash involved speeding car.
AGGRESSIVE: Crash involved aggressive driving.
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old.
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old.

Table 6

Table 6, the Reduced Logistic Regression Model, shows only the binary predictors, so MEDHHINC and PCTBACHMOR were removed. The most noticeable thing right away is that, despite the differences in including versus excluding the continuous variables, the statistical significance remains the same for the binary predictors, with CELL_PHONE being not significant in both models with $p = 0.88$. It also looks like the continuous variables seem to have not made a substantial difference to the coefficients nor altered the binary conclusions, which suggests again that crashes are most determined by individual action and not very much socioeconomic characteristics that encompass the driver's lived circumstances, so in this case it is the driver's immediate behavior at the time of the crash that poses the most significance.

Presumptively, this does not mean that driver characteristic is the only strong predictor category as physical environment could be at play, but further research would need to be done to look at binary or continuous variables that measure urban density, highway versus inner-city routes, road condition, etc.

Model Comparison: Akaike Information Criterion (AIC)

Model	Degrees of Freedom	AIC
Full Model (All Predictors)	**10**	**18359.63**
Reduced Model (Binary Only)	8	18360.47

Note:
Lower AIC is better.

Looking at the AIC values for both models, despite PCTBACHMOR and MEDHHINC not being as strong of predictors as their binary counterparts with such negligible impact according to their odds ratios, it is the full model that performs slightly better.

4. Discussion

In this analysis, we explored traffic crash data from Philadelphia covering 2008 through 2012 to investigate the relationship between various crash characteristics, driver behaviors, and neighborhood-level factors with the likelihood of alcohol involvement. Our dependent variable, DRINKING_D, was examined against these binary predictors FATAL_OR_M, OVERTURNED, CELL_PHONE, SPEEDING, AGGRESSIVE, DRIVER1617, and DRIVER65PLUS, alongside continuous variables PCTBACHMORE and MEDHHINC.

According to the full logistic regression model, the strongest predictors of crashes that involve drunk driving is SPEEDING with an odds ratio of 4.66 and a p-value of < 0.05, meaning speeding crashes are 4.66 times more likely to be associated with drunk driving compared to crashes without speeding, holding others constant. The next strongest variable is OVERTURNED with an odds ratio of 2.53 and p-value of < 0.05, indicating that crashes with overturned vehicles are 2.53 times more likely to be associated with drunk driving, holding others constant. Finally, FATAL_OR_M follows with an odds ratio of 2.26 and a p-value of < 0.05, meaning that fatal crashes are 2.26 times more likely to be associated with drunk driving compared to non-fatal crashes. The variable that is definitively not associated with drunk driving is CELL_PHONE and PCTBAHCMOR at $p = 0.88$ and $p = 0.77$ respectively.

The results are largely consistent with expectations, though some findings warrant further discussion. SPEEDING being the strongest predictor of alcohol-involved crashes aligns with what we would expect, as alcohol impairs judgment and reaction time, making risky behaviors like speeding more likely among intoxicated drivers. Similarly, OVERTURNED and FATAL_OR_M being positively associated with drunk driving is unsurprising given that impaired motor control and delayed reflexes increase crash severity. The negative associations for DRIVER1617 (OR = 0.28) and DRIVER65PLUS (OR = 0.46) are also intuitive. Minors face legal barriers to alcohol access, and older adults tend to drink less frequently or in smaller quantities.

However, the negative association between AGGRESSIVE driving and drunk driving (OR = 0.55) is a bit surprising. One might expect intoxicated drivers to exhibit more aggressive behaviors, but this finding suggests that aggressive driving is more characteristic of sober crashes, perhaps driven by frustration, road rage, or impatience rather than impairment. Additionally, the lack of significance for CELL_PHONE is interesting but not entirely surprising, as distracted driving and drunk driving may represent distinct risk profiles with different underlying causes. Finally, the negligible effect of MEDHHINC and PCTBACHMOR suggests that socioeconomic context matters less than immediate driver behavior in predicting alcohol involvement—a finding that reinforces the importance of individual-level interventions over neighborhood-level approaches.

While logistic regression is a reasonable starting point for this analysis, the rare events nature of the dependent variable raises concerns about its appropriateness. As observed in Table 1, only 6% of crashes involve alcohol (DRINKING_D = 1), meaning the dataset is substantially imbalanced. Standard maximum likelihood estimation in logistic regression can produce biased coefficient estimates when the outcome is rare, typically underestimating the probability of the event occurring. Paul Allison and others have proposed rare events logistic regression methods—such as Firth's penalized likelihood estimation or the King and Zeng correction—that adjust for this bias by modifying the likelihood function or weighting observations. Given that drunk driving crashes constitute such a small proportion of the sample, these methods would likely produce more accurate and reliable estimates than standard logistic regression. This limitation should be considered when interpreting the model's predictions, particularly the sensitivity and specificity trade-offs observed in Table 5, where the model struggled to correctly classify alcohol-involved crashes without sacrificing accuracy on non-alcohol crashes.

5. Limitations

There were several limitations when trying to use predictors to determine drunk driving as a reason for vehicle crashes. Speeding, overturned vehicles, and fatal or major injuries emerged as the strongest predictors of drunk driving crashes with odds ratios of 4.66, 2.53, and 2.26 respectively, while cell phone use and percentage of residents with bachelor's degrees or higher showed no significant association with alcohol involvement. These results mostly matched our expectations as driver behaviors such as speeding and loss of control align with the impaired judgment characteristic of intoxicated driving, while cell phone use represents a distinct form of distraction independent of alcohol. However, a critical limitation is that only 6% of crashes involve alcohol, suggesting that standard logistic regression may not be appropriate for this rare outcome. Paul Allison's rare events modeling methods, which penalize likelihood estimation, could provide more reliable coefficient estimates and better predictive performance than standard logistic regression when events occur in less than 5-10% of cases. Our model's AUC of 0.64 and the extreme sensitivity-specificity trade-offs observed in Table 5 reflect challenges inherent to modeling such rare events with standard methods.

This analysis has several other limitations. Important unmeasured confounders likely influence the observed relationships. Variables such as continuous driver age, weather conditions, road surface type, speed limits, and neighborhood-level alcohol availability are absent from the dataset and could bias coefficient estimates. While Pearson correlation was used to assess multicollinearity, it is not ideal for binary variables and may overlook complex interactions. Associations can be identified, but it cannot determine whether, for example, speeding causes drunk driving behavior or if alcohol causes speeding. More fundamentally, not all crashes are reported equally. Minor crashes may go unreported, and crashes in certain neighborhoods or times may be policed differently, creating systematic gaps in the data that bias which crashes get recorded as alcohol involved.

Citations

Centers for Disease Control and Prevention. (2024, November 19). *About transportation safety*. U.S. Department of Health & Human Services. https://www.cdc.gov/transportation-safety/about/index.html