Predicting Incidence of Drunk Driving: Applications of Logistic Regression

Ajjit Narayanan

Bill Cohen

CPLN HW3

Introduction

Across the United States, drivers under the influence of alcohol are responsible for an estimated 30 deaths per day according the U.S. Department of Transportation. In addition to the staggering loss of life, drunk driving causes many more non-fatal injuries and significant economic losses, making it a major risk factor when operating a motor vehicle. In this study, we examine motor vehicle accidents in the City of Philadelphia to identify indicators associated with drunk driving in residential neighborhoods.

Out of the 53,260 car crashes that took place in Philadelphia between 2008-2012, we focus on 43,364 accidents that took place in residential block groups where median household income and vacancy rates are higher than 0. The data set was compiled using data from the Pennsylvania Department of Transportation and the U.S. Census Bureau, and includes the following variables: FATAL_OR_M, which indicates whether the crash was fatal or caused a serious injury, OVERTURNED, which indicates if a vehicle was overturned in the crash, CELL_PHONE, which indicates if a cell phone was used, SPEEDING, which indicates if a car was speeding, AGGRESSIVE, which indicates if a driver was behaving aggressively, DRIVER1617, which indicates if a driver was 16 or 17 years of age, DRIVER65PLUS, which indicates if a driver is over the age of 65, PCTBACHMOR, which is the percentage of people with a bachelor's degree or more in the block group where the accident took place, and finally MEDHHINC, which is the median household income for the block group. For this analysis, we run multiple logistic regression using R and Rstudio.

Methods

In this study, we use logistic regression to model the relationship between a binary dependent variable and several categorical and continuous predictor variables. The binary dependent variable, DRINKING_D, takes on values of 0 or 1 indicating whether or not one or more drivers were intoxicated in a given accident.

In a previous paper, we regressed a continuous variable on several predictors using multiple ordinary least squares (OLS) regression. While OLS is equipped to model binary independent variables, it is not appropriate for modeling binary dependent variables, as it yields parameters that define a linear relationship between the dependent variable and each of the predictor variables. In particular, each of the $\beta's$ in OLS is interpreted as the amount by which the dependent variable changes when x_1 increases by one unit. However, if the dependent variable is binary, that means it can only take on values of 0 or 1. So saying that a one unit increase in an independent variable leads to a β increase in the binary variable doesn't make sense. In other words, since the binary variable only takes on one of two values, a linear relationship cannot exist between dependent and independent variables.

One way to get around this problem is instead of predicting the binary variable directly, we predict the probability that the binary variable equals 1. One problem with this approach is that the model may predict values of probability greater than 1 or lower than 0 if the values of the independent variables were extreme enough. Logistic regression addresses this range issue using a translator function that takes the probability values ranging from 0 to 1 and expands the function limits to $-\infty$ to ∞. The function that we use is the logistic function. Before we delve into the specifics of the logistic regression, we first formally introduce the concepts of probability, odds and odds ratios.

The probability of an event occurring is given by the number of observations where the event occurs divided by the total number of observations. Here, the probability that an accident involves a drunk driver (DRINKING_D = 1) is equal to:

$$P(DRINKING\_D = 1) = \frac{\text{\# of accidents where } DRINKING\_D = 1}{\text{Total \# of accidents}}$$

To find the probability that an event does not occur, in this case the probability that a car accident does not involve a drunk driver, we could apply the same approach using DRINKING_D = 0. Or, since there are only two possible outcomes with a binary variable, we can subtract the above probability from 1.

$$P(DRINKING_D = 0) = \frac{\text{\# of accidents where } DRINKING\_D = 0}{\text{Total \# of accidents}} = 1 - P(DRINKING_D = 1)$$

The odds of an event occurring is given by the number of observations where the event occurs divided by the number of observations where the event does not occur. In this study, the odds that an accident involves a drunk driver is given by:

$$Odds(DRINKING\_D = 1) = \frac{\#\text{ of accidents where } DRINKING\_D = 1}{\#\text{ of accidents where } DRINKING\_D = 0}$$

The odds can also be written as only a function of $P(DRINKING\_D = 1)$:

$$Odds(DRINKING\_D = 1) = \frac{P(DRINKING\_D = 1)}{P(DRINKING\_D = 0)} = \frac{P(DRINKING\_D = 1)}{1 - P(DRINKING\_D = 1)} = \frac{p}{1-p}$$

where p = P(DRINKING_D = 1). The natural log of the odds function $\frac{p}{1-p} = \ln(\frac{p}{1-p})$ is known as the log odds, or logit function.

We can write the equation for logistic regression with multiple predictors in terms of the logit function as:

$$ln(\frac{p}{1-p})$$
$$= \beta_0 + \beta_1 FATAL\_OR\_M + \beta_2 OVERTURNED + \beta_3 CELL\_PHONE + \beta_4 SPEEDING + \beta_5 AGGRESSIVE + \beta_6 DRIVER1617 + \beta_7 DRIVER65PLUS + \beta_8 PCTBACHMOR + \beta_9 MEDHHINC$$

where the model parameters are defined as:

β_0 = intercept, or the value of the log odds when all predictors = 0

$\beta_i = E(\hat{\beta}_i)$ = population value of the slope coefficient for predictor i such that the log odds change by a value of β_i when predictor i increases by one unit, with all other predictors held constant.

The full regression equation in logistic form is shown below.

$$p = P(Y = 1) = \frac{e^{\beta_0 + \beta_1 FATAL\_OR\_M + \beta_2 OVERTURNED + \beta_3 CELL\_PHONE + \beta_4 SPEEDING + \beta_5 AGGRESSIVE + \beta_6 DRIVER1617 + \beta_7 DRIVER65PLUS + \beta_8 PCTBACHMOR + \beta_9 MEDHHINC}}{1 + e^{\beta_0 + \beta_1 FATAL\_OR\_M + \beta_2 OVERTURNED + \beta_3 CELL\_PHONE + \beta_4 SPEEDING + \beta_5 AGGRESSIVE + \beta_6 DRIVER1617 + \beta_7 DRIVER65PLUS + \beta_8 PCTBACHMOR + \beta_9 MEDHHINC}}$$

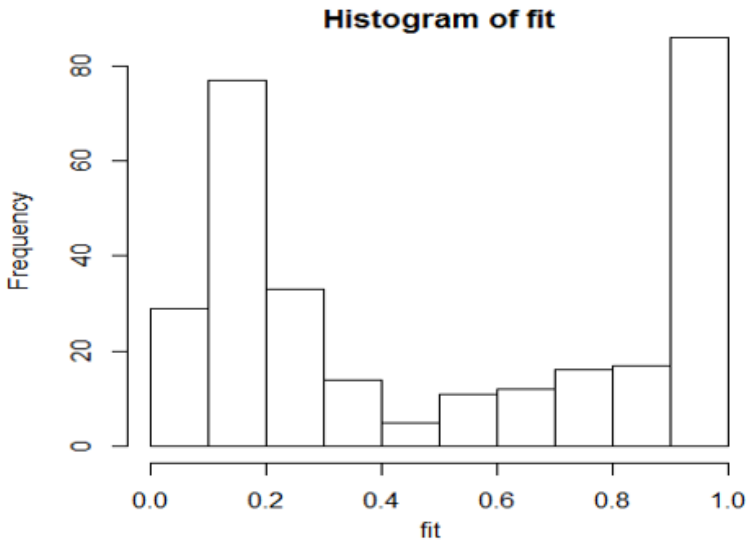
or

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

For each predictor, we run hypothesis tests to see if the effects of that predictors are significant. The test is based on the quantity $\frac{\hat{\beta}_i - E(\hat{\beta}_i)}{\sigma_{\beta_i}} = \frac{\hat{\beta}_i}{\sigma_{\beta_i}}$. This quantity is called a Wald statistic and follows a standard normal distribution. The p-values may be obtained by calculating the Wald score and then consulting a standard normal (z) table. Furthermore, most statisticians prefer not to evaluate the β coefficients directly and instead look at the odds ratios (which are just ratios of odds) which are calculated by exponentiating the coefficients and then taking the ratios.

Assessing Model Fit

In order to assess the quality of model fit under logistic regression, we use different methods than in OLS. An R-squared value may be calculated for logit models but does not have the same interpretation as in OLS and is no longer useful. Instead we look to metrics like the Aikake Information Criterion (AIC) for model selection. The AIC is a measure of model quality relative to other models that is based on the maximum value of the likelihood function. A lower AIC means a better model fit. Some other model fit metrics that are specific to logistic regression are specificity, sensitivity and misclassification rate. In order to explain these, we first have to understand how residuals and fitted values are calculated in logistic regression. Just as in OLS regression, the residuals are equal to $y_i - \hat{y}_i$. However, here the fitted value, $\hat{y}_i$, is the probability that $y = 1$ and is constrained between 0 and 1. If we were to examine the distribution of the fitted values in a logistic regression, it would look something like this:



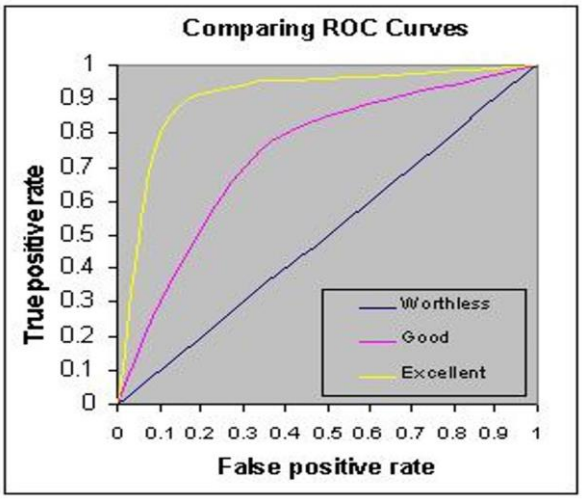
We have to choose a cutoff point so we can classify observations as either low probability or high probability, corresponding to values of 0 and 1 in the binary predicted variable. When we do that, we can generate a table to compare the predicted and actual values:

	Observed 0	Observed 1
Predicted 0	a	b
Predicted 1	c	d

Using this table, we can measure 3 metrics for goodness of fit: sensitivity, specificity, and the misclassification rate. Sensitivity or the true positive rate measures the percentage of actual positives which are correctly identified. In the above table, this is equivalent to $\frac{d}{b+d}$. Specificity or the true negative rate measures the percentage of negatives which are correctly identified as such. In the above table, this is equivalent to $\frac{a}{a+c}$. The misclassification rate is the proportion of all observations that are misidentified. In the case of our table, this is $\frac{b+c}{a+b+C+d}$.

One important thing to keep in mind is that the specific values of a, b, c, and d (and thus the sensitivity, specificity, and misclassification rates) depend on the cut-off value that we choose for determining whether an observation should be considered as high probability. One should try using different cut-off values and see how the goodness of fit measures change.

One tool we can use to help us choose cut-off values is the ROC curve. The ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1-specificity) across different cut-off values. We can use the ROC curve to help us select a good cut-off value by optimizing specificity and sensitivity. The optimal cutoff value is calculated as the point on the ROC curve where the distance from the ROC curve to the upper left of corner of the graph is minimized. Another way to compute a good cut-off value is to maximize the value of Specificity + Sensitivity, and this is called the Youden Index. However, we will not be using this method in our report. Below is an example of what a ROC curve would look like and how the initial minimization method would work.



The ROC curve can also help assess model quality. In particular, the area under the ROC curve (AUC) is a measure of prediction accuracy of the model. It can be interpreted as the probability that any randomly selected observation where the dependent variable is equal to 1 will have a value of $\hat{y}_i$ that is larger than a randomly selected observation with dependent variable equal to 0. Said differently, it is the probability that model's predicted values are correctly ranked. Possible values for the AUC range between 0.5 and 1. A rough guide for classifying the accuracy of AUC is as follows:

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

Assumptions of Logistic Regression

The assumptions of logistic regression are markedly different than the assumptions of OLS. There is still an assumption of no severe multicollinearity. However, in logistic regression, the assumption of strict linear relationships between dependent and independent variables no longer holds. The assumptions of homoscedasticity and normality of residuals is also dropped. Some additional assumptions are that the dependent variable has to be binary (i.e. have values of 0 or 1). There are also larger sample size requirements for logistic regression as compared to OLS regression. There must be at least 50 samples per predictor, as compared to 10 in the OLS case. And finally, the assumption of independence of observations also applies to logistic regression.

Exploratory Analysis

Prior to running logistic regression, statisticians often run some exploratory tests to provide a preliminary understanding of the relationships between the predictor and dependent variables. The tests are slightly different for categorical and continuous predictors, and they are presented in Tables 1 and 2 respectively.

Categorical Predictors

For categorical predictors, we use cross tabulation tables to examine the associations between the variable and the binary predicted variable. Cross tabulations give us the frequency counts of the predicted variable across each value of the predictor variables. They allow us to see whether there is an association between the two variables. We present cross tabulations between the predicted variables and all our categorical predictor variables in the methods section below.

For categorical predictors, a Chi-Square (χ^2) test can also be used to assess whether the distribution of values of the predictor vary significantly across values of the dependent variable. The null hypothesis, H_0, states that the proportion of accidents where the predictor occurs is the same for accidents involving a drunk driver as it is for accidents without a drunk driver, meaning the predictor is not correlated with drunk driving accidents. If we see a high χ^2 value with a p-value below 0.05, we reject H_0 in favor of H_a. The alternative hypothesis, H_a, states that the proportion of instances of the predictor varies significantly with instances of drunk driving. The results of the χ^2 test, including the degrees of freedom and p-values for each categorical variable are presented in Table 1 in the Results section.

Continuous Predictors

For continuous predictors, such as percent of bachelor's degree holders and median household income, we use a t-test to compare the predictor's mean value for each category of the dependent variable (i.e. when DRINKING_D = 0 and when DRINKING_D = 1). The null and alternative hypotheses for the t-test are similar to those for the Chi-Square test. The null hypothesis, H_0, states that the mean is the same for both values of the dependent variable, while the alternative hypothesis, H_a, states that mean varies significantly for different values of the dependent variable. Again, p-values below 0.05 suggest H_0 is rejected in favor of H_a, indicating a association exists between the predictor and dependent variable. These results are presented for the two continuous variables in Table 2 in the Results section.

To test the assumption of multicollinearity between predictors, the Pearson correlation matrix is also presented. Correlation coefficients less than -0.8 or greater than 0.8 indicate a high risk of multicollinearity between our predictor variables.

Results

Here we present and discuss the results of our exploratory analyses and logistic regression.

Exploratory Analysis

Before running our regression, we examine some preliminary characteristics of the data. First, we look at a summary of the count and proportion of accidents involving drunk driving:

```
## DRINKING_D
##     0     1
## 40879  2485

## DRINKING_D
##         0         1
## 0.9426944 0.0573056
```

The first row shows the counts of the DRINKING_D variables and the second row shows the sane variable broken up into percents. In this sample, a large majority of accidents, almost 95%, do NOT involve a drunk driver, while only 5.7% do. This means that in Philadelphia between 2008-2012, the probability of an accident involving a drunk driver is 5.7%. The odds of an accident involving a drunk driver are $\frac{2485}{40897} = 0.061$.

Categorical Variables: Cross Tabulation and Chi-Squared Statistic

Next, we look at the association between the DRINKING_D variable and each of the categorical predictors using cross tabulations and Chi-Square tests. Below are all the pairwise cross tabulations. Table 1 is a summary of all the cross tabulations. It shows us the counts of the independent variables split up by whether or not drivers were drunk. The % is the percentage of drunk driving accidents or non-drunk driving accidents where the independent variable is equal to 1. Table 2 is a summary of the Chi-Square tests for each categorical predictor variable.

Table 1

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		Total
	N	%	N	%	N
FATAL_OR_M: Crash resulted in fatality or major injury	1181	2.89%	188	7.57%	1369
OVERTURNED: Crash involved an overturned vehicle	612	1.50%	110	4.43%	722
CELL_PHONE: Driver was using cell phone	426	1.04%	28	1.13%	454
SPEEDING: Crash involved speeding car	1261	3.08%	260	10.46%	1521
AGGRESSIVE: Crash involved aggressive driving	18522	45.31%	916	36.86%	19438
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old	674	1.65%	12	0.48%	686
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old	4237	10.36%	119	4.79%	4356

Table 2

	Chi-squared test		
	Chi^2	d.f.	p-value
FATAL_OR_M: Crash resulted in fatality or major injury	167.5615	1	0.000000000000000
OVERTURNED: Crash involved an overturned vehicle	122.788	1	0.000000000000000
CELL_PHONE: Driver was using cell phone	0.162071	1	0.687256900000000
SPEEDING: Crash involved speeding car	376.7808	1	0.000000000000000
AGGRESSIVE: Crash involved aggressive driving	67.60186	1	0.000000000000000
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old	20.45167	1	0.000006115619000
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old	80.6047	1	0.000000000000000

Looking at the p-values in Table 2, we can reject H_0 for all of the categorical predictors except for cell phone use (p-value = 0.687), meaning all variables except cell phone use is associated with drunk driving accidents. For accidents involving cell phone use, this makes sense because we see a very similar proportion between those involving a drunk driver (1.13%) and those that do not (1.04%). For accidents involving drunk driving, we see a higher proportion of accidents with fatalities/major injuries, overturned vehicles, and speeding. For accidents that do not involve a drunk driver we see higher rates of aggressive driving, and those involving 16- or 17-year-old drivers, or drivers over the age of 65. For very young and very old drivers, this suggests there may be other factors, perhaps low experience in young drivers or impaired motor skills in older drivers, that might be more likely to cause an accident than alcohol.

Continuous Variable: T-Test

We also examine the relationship between drunk driving accidents and the two continuous variables, percent of bachelor's degree holders and median household income. Table 3 shows the mean and standard deviation (SD) of each predictor for each category of DRINKING_D, as well as the value of the t statistic, the degrees of freedom, and the p-value for the t-test. Table 2 shows the means, SD's and other statistics but the only important rows are PCTBCAHMOR and MEDHINC.

	Estimate	Std. Error	z value	Pr(>|z|)	OR	2.5 %	97.5 %
(Intercept)	-2.732506616128	0.045875659265	-59.5633209	0.0000000000000000000000000000000000000000...	0.06505601	0.05947628	0.07119524
FATAL_OR_M	0.814013801855	0.083806923855	9.7129660	0.0000000000000000000000265496694373523129 4...	2.25694878	1.90991409	2.65313350
OVERTURNED	0.928921376176	0.109166323622	8.5092302	0.0000000000000000175091930989952562102491...	2.53177687	2.03462326	3.12242730
CELL_PHONE	0.029550084767	0.197777821226	0.1494105	0.8812297205088647222837039407750125974416...	1.02999102	0.68354737	1.48846840
SPEEDING	1.538975665492	0.080545894089	19.1068171	0.0000000000000000000000000000000000000000...	4.65981462	3.97413085	5.45020642
AGGRESSIVE	-0.596915945677	0.047779237535	-12.4932079	0.0000000000000000000000000000000000081307...	0.55050681	0.50101688	0.60423487
DRIVER1617	-1.280295964022	0.293147167807	-4.3674171	0.0000125724471279377828539193170942489530...	0.27795502	0.14774429	0.47109277
DRIVER65PLUS	-0.774664640320	0.095858315238	-8.0813505	0.0000000000000006405343748176638523243704...	0.46085831	0.37998364	0.55347851
PCTBACHMOR	-0.000370633604	0.001296386538	-0.2858974	0.7749566677916315260432611466967500746250...	0.99962944	0.99707035	1.00215087
MEDHHINC	0.000002804492	0.000001340972	2.0913870	0.0364933835319284916032245291717117652297...	1.00000280	1.00000013	1.00000539

Right away we see the mean values for both predictors are very similar for accidents with and without alcohol involved, so we'd expect the t-test to show little association with dependent variable. The t-test p-values confirm this. Both p-values are greater than 0.05, so we are unable to reject the null hypothesis that the distribution of the continuous variables is not significantly different drunk driving and non-drunk driving accidents.

Logistic Regression Assumptions

In order to make sure there are no problems with multicollinearity, we look at the Pearson correlation coefficients between all predictors to identify issues of multicollinearity.

	FATAL_OR_M	OVERTURNED	CELL_PHONE	SPEEDING	AGGRESSIVE	DRIVER1617	DRIVER65PLUS	PCTBACHMOR	MEDHHINC
FATAL_OR_M	1.0000000	0.0331959	0.0021603	0.0817127	-0.0110473	-0.0028084	-0.0125123	-0.0146523	-0.0182124
OVERTURNED	0.0331959	1.0000000	-0.0009898	0.0594403	0.0164389	0.0037240	-0.0195010	0.0093321	0.0279213
CELL_PHONE	0.0021603	-0.0009898	1.0000000	-0.0036012	-0.0257430	0.0014851	-0.0027173	-0.0012459	0.0020999
SPEEDING	0.0817127	0.0594403	-0.0036012	1.0000000	0.2115254	0.0160116	-0.0328541	-0.0007391	0.0117867
AGGRESSIVE	-0.0110473	0.0164389	-0.0257430	0.2115254	1.0000000	0.0284290	0.0150269	0.0271221	0.0434405
DRIVER1617	-0.0028084	0.0037240	0.0014851	0.0160116	0.0284290	1.0000000	-0.0208484	-0.0026360	0.0228774
DRIVER65PLUS	-0.0125123	-0.0195010	-0.0027173	-0.0328541	0.0150269	-0.0208484	1.0000000	0.0261904	0.0503377
PCTBACHMOR	-0.0146523	0.0093321	-0.0012459	-0.0007391	0.0271221	-0.0026360	0.0261904	1.0000000	0.4778695
MEDHHINC	-0.0182124	0.0279213	0.0020999	0.0117867	0.0434405	0.0228774	0.0503377	0.4778695	1.0000000

None of the correlation coefficients are greater than 0.8 or less than -0.8, suggesting there are no issues with multicollinearity between any of the predictors. Also to note are that there may be potential problems when using Pearson correlation to measure associations between 2 binary or categorical variables. Especially since our binary variables are sparse (i.e. not a lot of values = 1) this means that the correlation might not be a great measure of similarity. It is just telling us how likely it is that both binary variables are turned on (=1).

Multiple Logistic Regression Analysis

The results of the logistic regression are shown below. We interpret the model parameters, significance tests, and odds ratios (OR) for each predictor and the intercept.

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*	*OR*	*2.5 %*	*97.5 %*
(Intercept)	-2.7325066	0.0458757	-59.5633209	0.0000000	0.0650560	0.0594763	0.0711952
FATAL_OR_M	0.8140138	0.0838069	9.7129660	0.0000000	2.2569488	1.9099141	2.6531335
OVERTURNED	0.9289214	0.1091663	8.5092302	0.0000000	2.5317769	2.0346233	3.1224273
CELL_PHONE	0.0295501	0.1977778	0.1494105	0.8812297	1.0299910	0.6835474	1.4884684
SPEEDING	1.5389757	0.0805459	19.1068171	0.0000000	4.6598146	3.9741309	5.4502064
AGGRESSIVE	-0.5969159	0.0477792	-12.4932079	0.0000000	0.5505068	0.5010169	0.6042349
DRIVER1617	-1.2802960	0.2931472	-4.3674171	0.0000126	0.2779550	0.1477443	0.4710928
DRIVER65PLUS	-0.7746646	0.0958583	-8.0813505	0.0000000	0.4608583	0.3799836	0.5534785
PCTBACHMOR	-0.0003706	0.0012964	-0.2858974	0.7749567	0.9996294	0.9970704	1.0021509
MEDHHINC	0.0000028	0.0000013	2.0913870	0.0364934	1.0000028	1.0000001	1.0000054

All but two variables are significant, with p-values below 0.05. Accidents involving fatalities or major injuries, overturned vehicles, speeding, aggressive driving, 16- or 17-year-old drivers, or those over the age of 65, and the median household income of the census block where the accident took place are significant predictors of car crashes involving an intoxicated driver. Accidents involving a driver on a cell phone and percentage of bachelor's degree holders in the census block where the accident took place are shown not to be significant in the model, both with p-values greater than 0.05, and therefore their results will not be interpreted here.

The model estimate of the coefficient for the intercept, β_0, is -2.733. If all other predictors in the model have values of 0, the log odds of there being a drunk driver involved in a car crash is -2.733. The log odds, -2.733, yields the odds ratio $e^{-2.733} = 0.065$. So, for accidents that did NOT involve a fatality or major injury (FATAL_OR_M = 0), an overturned vehicle (OVERTURNED = 0), a speeding vehicle (SPEEDING = 0), an aggressive driver (AGGRESSIVE = 0), a 16- or 17-year-old driver (DRIVER1617 = 0) or a driver over the age of 65 (DRIVER65PLUS = 0), and accidents in a census block where median household income is 0 (MEDHHINC =0), the odds of their being a drunk driver involved in the crash are 0.065. However, it should be noted that the dataset was initially cleaned to remove observations where MEDHHINC = 0, so this is a purely extrapolated value.

The model estimate of the coefficient for accidents involving a fatality or major injury, β_1, is 0.814. For a one unit increase in FATAL_OR_M, meaning as we go from a crash without a fatality or major injury to a crash with a fatality or major injury, the log odds of there being a drunk driver increases by 0.814, holding all other predictors constant. So, using the odds ratio $e^{0.814} = 2.257$, the odds of there being a drunk driver involved in a crash increase by 2.257, or $(e^{0.814} - 1) * 100\% = 125.7\%$, for accidents with a fatality or major injury compared with those without a fatality or major injury when holding all other predictors constant.

The model estimate of the coefficient for accidents involving an overturned vehicle, β_2, is 0.929. For a one unit increase in OVERTURNED, meaning as we go from a crash without an overturned vehicle to a crash with an overturned vehicle, the log odds of there being a drunk driver increases by 0.929, holding all other predictors constant. So, using the odds ratio $e^{0.929} = 2.232$, the odds of there being a drunk driver involved in a crash increase by 2.232, or $ (e^{0.929} - 1) * 100% = 123.2% $, for accidents with an overturned vehicle compared with those without an overturned vehicle when holding all other predictors constant.

The model estimate of the coefficient for accidents involving at least one vehicle travelling over the speed limit, β_4, is 1.539. For a one unit increase in SPEEDING,

meaning as we go from a crash without a speeding vehicle to a crash with a speeding vehicle, the log odds of there being a drunk driver increases by 1.539, holding all other predictors constant. So, using the odds ratio $e^{1.539} = 4.66$, the odds of there being a drunk driver involved in a crash increase by 4.66, or $(e^{0.1.539} - 1) * 100\% = 366.0\%$, for accidents with a speeding vehicle compared with those without a speeding vehicle when holding all other predictors constant.

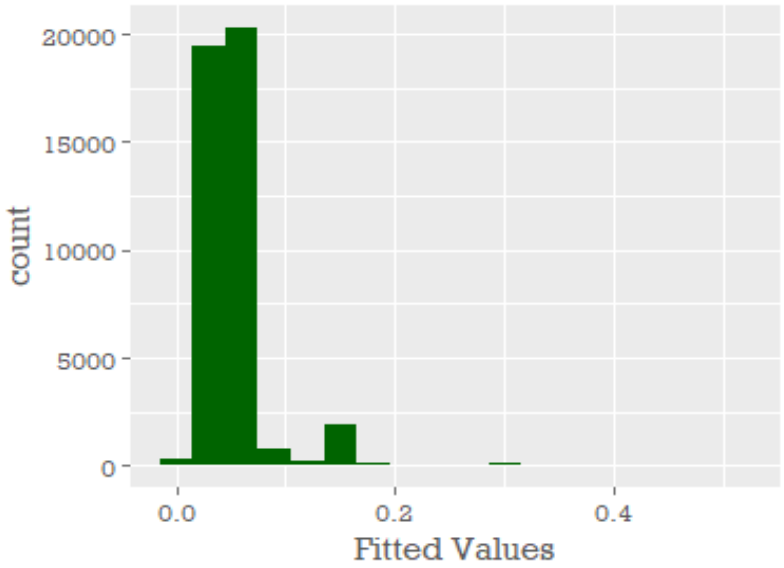
The model estimate of the coefficient for accidents involving an aggressive driver, β_5, is -0.597. For a one unit increase in AGGRESSIVE, meaning as we go from a crash without an aggressive driver to a crash with an aggressive driver, the log odds of there being a drunk driver decreases by 0.597, holding all other predictors constant. So, using the odds ratio $e^{-0.597} = 0.551$, the odds of there being a drunk driver involved in a crash decrease by 0.551, or $(e^{-0.597} - 1) * 100$, for accidents with an aggressive driver compared with those without an aggressive driver when holding all other predictors constant.

The model estimate of the coefficient for accidents involving a 16- or 17-year-old driver, β_6, is -1.280. For a one unit increase in DRIVER1617, meaning as we go from a crash without a 16- or 17-year-old driver to a crash with a 16 or 17 year old driver, the log odds of there being a drunk driver decreases by 1.280, holding all other predictors constant. So, using the odds ratio $e^{-1.280} = 0.278$, the odds of there being a drunk driver involved in a crash decrease by 0.278, or $(e^{-1.280} - 1) * 100\% = 27.8\%$, for accidents with a 16 or 17 year old driver compared with those without a 16 or 17 year old driver when holding all other predictors constant.

The model estimate of the coefficient for accidents involving a driver over the age of 64, β_7, is -0.775. For a one unit increase in DRIVER65PLUS, meaning as we go from a crash without a driver over the age of 64 to a crash with a driver over the age of 64, the log odds of there being a drunk driver decreases by 0.775, holding all other predictors constant. So, using the odds ratio $e^{-0.775} = 0.461$, the odds of there being a drunk driver involved in a crash decrease by 0.461, or $(e^{-0.775} - 1) * 100\% = 46.1\%$, for accidents with a driver over the age of 64 compared with those without a driver over the age of 64 when holding all other predictors constant.

The model estimate of the coefficient for median household income, β_9, is 0.000003. For a one unit ($1) increase in MEDHHINC for the census block where an accident takes place, the log odds of there being a drunk driver increases by 0.000003, holding all other predictors constant. So, using the odds ratio $e^{0.000003} = 1.00$, the odds of there being a drunk driver involved in a crash increase by 1.00, or $(e^{0.000003} - 1) * 100\% = 100\%$, as median household income increases in a given census block when holding all other predictors constant.

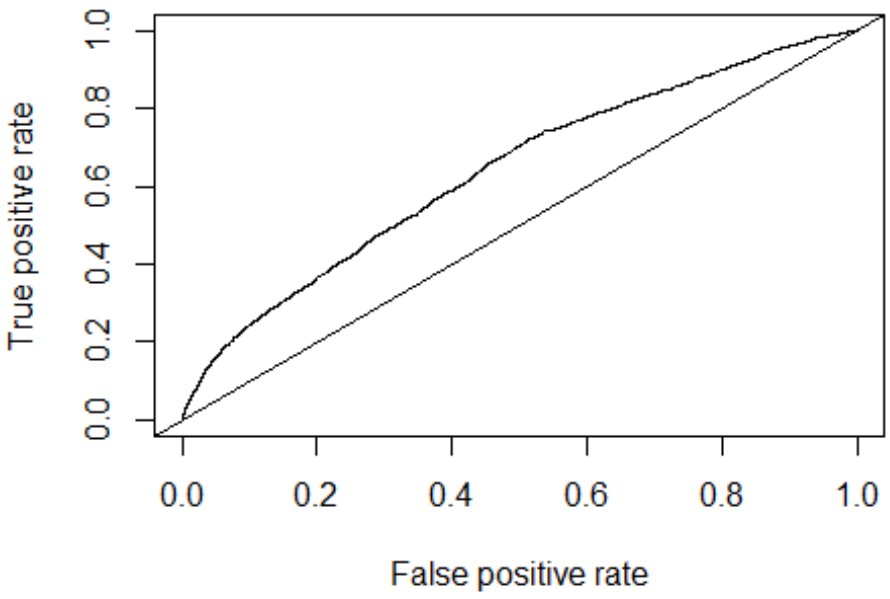
Next, we take a look at the histogram of fitted values.



We need to choose a cutoff value so we can classify points as either high probability of there being a drunk driver or a low probability. Preferably, this cutoff would satisfy the criteria laid out in the Methods section. Below is a summary table of different probability cutoffs and their corresponding Sensitivity, Specificity, and Misclassification rates. The cutoff rate with the lowest Misclassification rate is highlighted.

Cutoff Value	Sensitivity	Specificity	Misclassification Rate
0.02	0.016499	0.0580738	0.888894
0.03	0.0193159	0.0639204	0.883544
0.05	0.2651911	0.4690917	0.5156812
0.07	0.778672	0.9138188	0.1258648
0.08	0.8152918	0.9386237	0.1045798
0.09	0.8317907	0.9459625	0.0986071
0.1	0.8358149	0.948213	0.0967162
0.15	0.8957746	0.9722107	0.0775297
0.2	0.9770624	0.9953766	0.0603496
0.5	0.9983903	0.9999022	0.0573056

The last probability cutoff of 0.5, which is very far to the right in the distribution of fitted probabilities, gives us the minimum misclassification rate of 0.0573056. The cutoff rate that gives us the highest misclassification rate is 0.02, which is the lowest value we tested. In this study cutoff value, the lower the misclassification rate. This suggests the model really doesn't want to label any data points as high probability of being a drunk driving incident (i.e. predicting DRINKING_D = 1). IT seems to want to automatically label all observation, except the very extreme cases, as non-drunk driving incidents. This makes sense given the scarcity of drunk driving incidents in the whole dataset. Next, we calculate the ROC curve and plot it below. We also calculate the Area Under the Curve, and the cutoff value that minimizes the distance between the top left corner of the graph and the curve. All are reported below.



Sensitivity	0.6607646
Specificity	0.5452433
Cutoff	0.0636515
AUC	0.6398695

The cutoff point where the distance from the top left corner of the graph is minimized is 0.065. The accompanying specificity and sensitivity values which are 0.66 and 0.542 respectively. This cut-off is actually above the maximum cutoffs we tested earlier. This makes sense because the largest cutoff there also had the lowest misclassification rates. Raising the cutoff value even more might lead to a better

Sensitivity and Specificity value. The area under the ROC curve is approximately 0.63, which places it in the lower end of the 'poor' rating when it comes to AUC scores.

Finally, we also run a model without our continuous predictor variables and present the results below

	Estimate	Std. Error	z value	Pr(>|z|)	OR	2.5 %	97.5 %
(Intercept)	-2.6518996	0.0275311	-96.3238683	0.0000000	0.0705171	0.0667864	0.0743978
FATAL_OR_M	0.8093156	0.0837615	9.6621431	0.0000000	2.2463700	1.9011246	2.6404533
OVERTURNED	0.9397842	0.1090343	8.6191585	0.0000000	2.5594290	2.0573601	3.1556897
CELL_PHONE	0.0310737	0.1977709	0.1571195	0.8751506	1.0315615	0.6845978	1.4907150
SPEEDING	1.5403203	0.0805279	19.1277908	0.0000000	4.6660847	3.9796186	5.4573472
AGGRESSIVE	-0.5936469	0.0477478	-12.4329656	0.0000000	0.5523094	0.5026882	0.6061758
DRIVER1617	-1.2715761	0.2931097	-4.3382260	0.0000144	0.2803894	0.1490473	0.4751771
DRIVER65PLUS	-0.7664573	0.0957644	-8.0035718	0.0000000	0.4646563	0.3831829	0.5579332

In this model, all coefficients are significant at a 5% Confidence level except for CELL_PHONE, which is the exact same as in the previous regression. The coefficients themselves have also remained relatively stable. We also compare the AIC of the 2 models:

Model	AIC
With Continuous predictors	1018359.6290512
W/o Continuous predictors	818360.4664511

So the model without continuous predictors has significant lower AIC value, indicating that it is a better model. This indicates continuous predictors do not add a lot of predictive power.

Discussion

In this paper, we ran a logistic regression on a dataset of car accidents, with the dependent variable being whether the driver was drunk. The highly significant and positively correlated predictor variables are whether the accident was fatal and/or involved serious injuries, whether the car was overturned, and whether the car was speeding. It's also worth noting that there was a small positive and weakly significant effect of the median house value where the accident occurred. The significant negatively correlated predictors are whether the driver was aggressively driving, whether the driver was 16 or 17, and whether the driver was over the age of 65. For the most part,

these make intuitive sense as we would expect drunk driving incidents to be correlated with fatal, fast, and overturned accidents. The fact that is surprising is that aggressive driving has a negative coefficient. Given that fatal accidents and accidents that involved speeding are positively related, one would assume that aggressive driving is also positively associated. This could indicate that drunk drivers try to be cautious, but because of their inebriated states.

Looking at the results of the regression, it seems that logistic regression may not be the best tool to use here. The reason is the rarity of drunk driving itself. Drunk driving only occurs 5.7% of the time. Our results with the very high cutoff values relative to the histogram of fitted values. The model just seems to want to label every observation as 0. The problem could be with the small sample bias that comes with Maximum Likelihood Estimation of the logistic regression model. A limitation of the analysis is that the direction of causality is not really clear. It doesn't really make sense to say that driving fast increases your probability of being a drunk driver; the causality may go in the opposite direction. From a regression point of view, the independent variables are not necessarily causing the dependent variable. In this case, the modeling rare events methods proposed by Paul Allison could be more appropriate.