Zhenzhao Xu, Hanpu Yao, and Ben Aiken
Statistical and Data Mining Methods - MUSA 500
Professor Eugene Brusilovskiy
Assignment 3

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

Introduction

This investigation examines the relationship between car crashes caused by alcohol in Philadelphia from 2008 to 2012 using logistic regression and predictors including the age of the driver, the driving conditions, the crash outcome, and the crash location. Tragically, alcohol caused car crashes lead to over 10,000 deaths in America every year.[1] If relevant variables with strong correlations are identified and an effective regression model is developed, high risk situations or locations could be addressed with interventions that could prevent future crashes.

Nine predictors were selected due to their potential relationship with crashes and alcohol consumption. Speeding (SPEEDING), driving aggressively (AGGRESSIVE), or using a cell phone (CELL_PHONE) are in the top four most common causes of car accidents, the fourth being alcohol.[2] Additionally, drinking alcohol impairs judgement and could also lead to these three causes. The age of the driver, or more specifically if the driver was either 16-17 (DRIVER1617) or older than 64 (DRIVER65PLUS), are included since younger drivers who have less experience behind the wheel and elderly drivers who have decreased reaction time are often responsible for crashes. The outcome of the crash including if there was a fatal or major injury (FATAL_OR_M) and if the crash involved an overturned vehicle (OVERTURNED) is likely related to the cause of the crash, and in this case could be related to alcohol. Finally, the location characteristics of the crash, specifically the education (PCTBACHMOR) and income levels (MEDHHINC) of the residential block groups are included to determine if there or spatial trends to crash locations.

The regressions and analysis will be completed using R.

Methods

a) OLS Regression Limitations

Ordinary Least Squares, or OLS, regression is a powerful prediction tool when dealing with a dependent variable that is continuous and normal. The model determines how much the

[1] National Highway Traffic Safety Administration. Traffic Safety Facts 2016 data: alcohol-impaired driving. U.S. Department of Transportation, Washington, DC; 2017 Available at: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812450

[2] https://www.huffpost.com/entry/top-15-causes-of-car-accidents_b_11722196

dependent variable increases when a predictor increases by one unit. This regression tool is, however, inadequate with a binary dependent variable. Since the range of possible values are between 0 and 1, OLS regression would be problematic because it may indicate that the dependent variable would change to an amount outside of this range as the predictors change. Since the dependent variable for this investigation is binary: either the driver was drinking or the driver was not drinking, a different regression model must be utilized.

b) Logistic Regression Overview

Instead, logistic regression, which is designed for binary dependent variables, will be used in this report. Logistic regression is derived from the logit function, which determines the log of the odds that a specific outcome will occur. Odds are the probability of an outcome occurring over another outcome. In this instance, the probability of someone crashing due to drinking and driving is:

$$p(drinking) = \frac{\#\ drinking}{\#\ drinking + \#\ not\ drinking} = \frac{\#\ drinking}{\#\ crashes} \qquad (1.)$$

Therefore, the odds of the outcome are:

$$Odds(drinking) = \frac{\frac{\#\ drinking}{\#\ crashes}}{\frac{\#\ not\ drinking}{\#\ crashes}} = \frac{\#\ drinking}{\#\ not\ drinking} \qquad (2.)$$

When the probability of an outcome approaches 1 it is more likely to happen. Conversely, when probability is close to 0 the opposite outcome is likely to happen. Odds are not limited from 0 to 1, but ranges from 0 to ∞. The greater the odds are above 1, the outcome is more likely to occur.

Logit, or the log of odds, models binary dependent variables because it translates a linear function, like OLS for example, into a function that ranges from 0 to 1. In Figure 1, the red line is the linear model and the blue curve has been translated to reflect the binary outcome: 0 or 1.

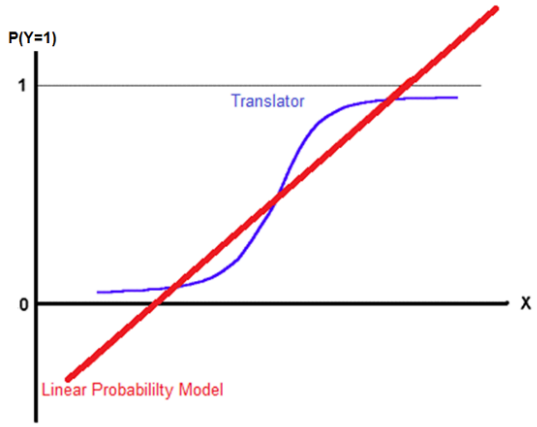

Figure 1 – Linear and Logit Models

The logit function is below:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon \quad (3.)$$

Here, $\frac{p}{1-p}$ is the odds of the examined outcome, therefore $\ln\left(\frac{p}{1-p}\right)$ is the log of the odds. β_0 is the log of the odds when all variables are equal to 0. β_i are the coefficients for the variables x_i. And finally, ε is the error. The logit model with this investigation's independent and dependent variables is below:

$$\ln\left(\frac{p(drinking)}{1-p(drinking)}\right) = \beta_0 + \beta_1 FATALORM + \beta_2 OVERTURNED + \beta_3 CELL\_PHONE + \beta_4 SPEEDING + \beta_5 AGGRESSIVE + \beta_6 DRIVER1617 + \beta_7 DRIVER65PLUS + \beta_8 PCTBACHMOR + \beta_9 MEDHHINC + \varepsilon \quad (4.)$$

Logistic regression is the logit equation rearranged algebraically. This is shown below:

$$p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+x_k+\varepsilon}}{1+e^{\beta_0+\beta_1 x_1+\cdots+x_k+\varepsilon}} \quad (5.)$$

Additionally, the logistic function with this investigation's variables is below. For the sake of space, the variables have been abbreviated.

$$p(drinking) = \frac{e^{\beta_0+\beta_1 F_{OR_M}+\beta_2 OVER+\beta_3 CELL+\beta_4 SPEED+\beta_5 AGG+\beta_6 1617+\beta_7 65PLUS+\beta_8 PCTBACH+\beta_9 MEDHHINC+\varepsilon}}{1+e^{\beta_0+\beta_1 F_{OR_M}+\beta_2 OVER+\beta_3 CELL+\beta_4 SPEED+\beta_5 AGG+\beta_6 1617+\beta_7 65PLUS+\beta_8 PCTBACH+\beta_9 MEDHHINC+\varepsilon}} \quad (6.)$$

6.

c) Logistic Regression Hypothesis Testing

For each predictor a hypothesis test is completed to determine if its coefficient β_i is 0:

$$H_0: \beta_i = 0 \ (OR_i = 1)$$
$$H_a: \beta_i \neq 0 \ (OR_i \neq 1)$$

The quantity below is the difference between the observed ($\hat{\beta}_i$) and estimated $E(\hat{\beta}_i)$ values of the coefficient for the predictor, divided by the standard deviation, which has a standard normal distribution:

$$\frac{\hat{\beta}_i - E(\hat{\beta}_i)}{\sigma_{\hat{\beta}_i}} \quad (7.)$$

According to the H_0, $E(\hat{\beta}_i) = 0$, and the remaining quantity is referred to as z, or the Wald Statistic:

$$z = \frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}} \quad (8)$$

In order to reject the null hypothesis that the predictors have a non-zero coefficient, the z-value must be significant. Instead of comparing estimated coefficients, many statisticians prefer to examine odds ratios. Odds ratios are determined by exponentiating the coefficients.

d) Logistic Regression Assessment

OLS regression is assessed using R^2 values, which indicates what percent of variation is explained by the model. This value can be calculated for logistic regression; however, since logistic regression develops a probability rather than a value, the interpretation is not as straightforward as the percent of variation in OLS.

Another assessment of fit is the Akaike Information Criterion or AIC value. When comparing models, the model with the lower AIC value is a better fit for the data. AIC is calculated with the number of predictors and the maximum likelihood estimate of the model. Since it takes the number of predictors into account, it prevents over-fitting and simple, well fit models have the lowest value.

Commented [EB1]: -1. These are expected values, not estimated values. Beta-hats are the observed, estimated values.

Commented [EB2]: It's the beta coefficient that would be significant. The z-value is either higher than |1.96| (significant) or not (assuming alpha = 0.05).

Commented [EB3]: examining

Commented [EB4]: -1. ?? The reason that R-square isn't really useful in Logistic regression is that logistic regression isn't estimated using OLS, but rather maximum likelihood. And R-square is a measure that makes sense in OLS, where we aim to minimize SSE. This quantity isn't examined in maximum likelihood estimation.

Sensitivity, specificity, and the misclassification rate are also used in the assessment of models. The results of logistic regression, or the fitted/predicted values, y, are probabilities between 0 and 1. The greater the value, or the closer the y is to 1, the more likely that outcome is to occur. A y of 0.1 suggests the outcome only has a 10% chance of occurring. On the other hand, a y value of 0.9 suggest the outcome has a 90% chance of occurring and just a 10% chance of not occurring. When interpreting the results of a logistic regression, creating a "cut off", like 0.5 for example, or rather the level at which a probability is deemed high and the outcome is $y = 1$, is difficult and depends on the model. There is not a specific "cut-off" that is accepted by all statisticians. A histogram of the predicted values can be useful in selecting this cut-off.

Commented [EB5]: Is *predicted* to occur

Commented [EB6]: The issue is that there are different cut-offs that work for different data sets.

The "cut-off" determines the sensitivity and specificity of the model. Sensitivity is the true positive rate: the proportion of positives that the model correctly predicts – in this case the outcome is that the driver had consumed alcohol. Specificity is the true negative rate: the proportion of negatives that the model correctly predicts – in this case that the driver had not consumed alcohol. Alternative to the true negatives and positives is the misclassification rate, which is the rate of positives and negatives that are incorrectly predicted, or the false positives and false negative divided by the total outcomes.

Commented [EB7]: Sensitivity and specificity can be calculated for different cut-offs.

Different "cut-off" probability values will yield different rates, making the decision a complicated trade off. The ideal model optimizes the sensitivity and specificity and limits the misclassification rate. The decision is best informed by testing several cut-offs and comparing the rates, which is easiest with the use of a Receiver Operating Characteristic Curve. The ROC curve plots the true positive rate (sensitivity) compared with the false positive rate (the difference between 1 and the specificity). Figure 2 shows an example of an ROC curve.[3]

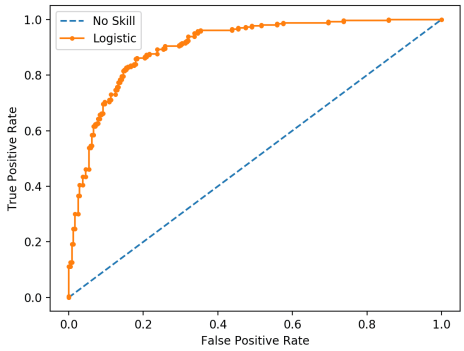

Figure 2 – Example ROC Curve

[3] https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

The blue dotted line in Figure 2 shows a model that is correctly identifying positives half the time, essentially flipping a coin to guess the outcome. The orange curve is a far better fit for the data since it is above the blue guessing line, meaning the false positive predictions of the model (sensitivity) outnumber the false positives (1-specificity).

Commented [EB8]: ??

The further a model's ROC curve is from the blue line and the closer it is to the upper left corner, the better the model is. One way to measure this difference is by examining the area under the curve, or AUC. The AUC of the blue line is .5, since it bisects the plot. Any AUC greater than .5 is a better fit than flipping a coin; however, an AUC between 0.5-0.6 is generally considered a failing model. AUC between 0.6 and 0.7 is poor, 0.7-0.8 is fair, 0.8-0.9 is good, and 0.9-1 is excellent.

In this investigation, the optimal cut-off will be selected by reducing the misclassification rates and examining the ROC curve with the smallest distance between the upper left corner and the curve, or the greatest AUC.

e) Assumptions of Logistic Regression

Logistic regression relies on several assumptions. Just like in OLS there is an assumption of independence of observations and no multicollinearity between predictors. However, unlike in OLS, the dependent variable must be binary and there is no assumption of a linear relationship between the dependent variable and predictors. Additionally, there is not an assumption of homoscedasticity and the residuals do not need to be normal as in OLS. Finally, logistic regression requires at least 50 observations for each predictor, while OLS only needs 10.

f) Exploratory Analysis prior to Modeling

Before running Logistic Regression, many statisticians investigate the variables to better understand the data. Running cross-tabulations between the dependent variable and binary predictors reveals associations and potentially problematic multicollinearity. Cross-tabulation can also identify variables or categories with small frequencies that may be problematic when running the regression.

To further explore the association between variables, the Chi-Square test is used to determine the independence of two categorical variables. The test examines how the distribution of one categorical variable varies with the values of another categorical variable by comparing their proportions using the following hypotheses:

$$H_0: p(x_1) = p(x_2)$$
$$H_a: p(x_1) \neq p(x_2)$$

Commented [EB9]: I don't believe that this is the right mathematical formulation of the hypotheses. You're looking if the distribution of one categorical variable varies across *levels* of another categorical variable. For instance, is the proportion of cell phone users the same (H0) or different (Ha) for drivers who use vs. don't use alcohol?

The result of the Chi-Squared test is a value called χ^2. When there is a high χ^2 value that is significant (the p-value is less than 0.05), then the null hypothesis is rejected. For example, when comparing speeding and overturned cars, if the χ^2 value is high and the p-value indicates

Commented [EB10]: -1. Awkward phrasing. When looking at Chi-square analyses, you're examining associations between two variables, not comparing two variables.

significance there is an association between the variables.

Furthermore, the means and standard deviations of each variable are important to investigate in this exploratory analysis phase. Independent sample t-tests, which measure the mean and standard deviation for predictors for both dependent variable outcomes, drinking or not drinking, are also completed for location measures including PCTBACHMOR and MEDHHINC. The hypotheses associated with the t-test are below:

$$H_0: PCTBACHMORE\ mean\ of\ drinking = PCTBACHMORE\ mean\ of\ not\ drinking$$
$$H_a: PCTBACHMORE\ mean\ of\ drinking \neq PCTBACHMORE\ mean\ of\ not\ drinking$$

The t-test produces the t-statistic, which if high and significant (a p-value less than 0.05), rejects the null hypothesis in favor of the alternative.

Results

a) The results of the exploratory analyses.

It is shown in the tabulation of the dependent variable (Table 1) that among all car crashed in the dataset, 5.73% involved a drinking driver, and the proportion is relatively small compared to the other 94.27% car crashes are not associated with alcohol.

Usually, the rare events in a small dataset may be problematic, because maximum likelihood estimation of the logistic model suffers from small-sample bias. But in this case, the sample size is 43,364 with 2,485 drinking driver samples, so it will not be a problem.

Commented [EB11]: Good!

Table 1 - Tabulation of DRINKING_D

		0: Not Drinking	%	1: Drinking	%
DRINKING_D:	Drinking driver indicator	40879	94.27%	2485	5.73%

Table 2 is the cross-tabulation of the dependent variable with each of the binary predictors. The percentage and number of accidents involving (and not involving) drunk drivers for each predictor are displayed in the table.

These numbers show that there is higher proportion of FATAL_OR_M, OVERTURNED, and SPEEDING crashes with alcohol involved, so there may be a relationship between these predictors and the dependent variable. Lower proportions are found in predictors such as DRIVER1617 and DRIVER65PLUS, which is reasonable considering they are not the main users of alcohol.

Table 2 - Tabulation of Binary Variables

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		Total	Chi-Square	
	N	%	N	%	N	P-value	Chi-Square
FATAL_OR_M: Crash resulted in fatality or major injury	1181	2.90%	188	7.60%	1369	<0.0001	167.56
OVERTURNED: Crash involved an overturned vehicle	612	1.50%	110	4.40%	722	<0.0001	122.79
CELL_PHONE: Driver was using cell phone	426	1.00%	28	1.10%	454	**0.687**	0.16
SPEEDING: Crash involved speeding car	1261	3.10%	260	10.50%	1521	<0.0001	376.78
AGGRESSIVE: Crash involved aggressive driving	18522	45.30%	916	36.90%	19438	<0.0001	67.60
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old	674	1.60%	12	0.50%	686	<0.0001	20.45
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old	4237	10.4	119	4.80%	4356	<0.0001	80.60

In this project, Chi-Square is used to test if there is a difference in each of the predictors when alcohol is involved or not involved.

The results of the Chi-Square test are also presented in Table 2. It shows that all the binary predictors (except CELL_PHONE) are significant with p-value less than 0.0001, so we can reject the null hypnosis and say that there is a significant association between the dependent variable and each of these predictors. The predictor CELL_PHONE has a Chi-Square statistic of 0.16 and its p-value is 0.687, which suggests that the null hypothesis cannot be rejected that CELL_PHONE is independent from the dependent variable.

The means of the continuous predictors for both values of the dependent variable are shown in Table 3. The results suggest that the dependent variable don't have a substantial impact on PCTBACHMOR and MEDHHINC. The difference of mean value of PCTBACHMOR is only 0.1%, and the difference of MEDHHINC is only $500, which can be negligible compared to its standard deviation.

Table 3 - Tabulation of Continuous Variables

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		Independent Samples T-test	
	Mean	SD	Mean	SD	T-statistics	P-value
PCTBACHMOR: % with bachelor's degree or more	16.57	18.21	16.61	18.72	-0.108	0.914
MEDHHINC: Median household income	31483	16930	31998	17810	-1.405	0.160

An Independent Samples T-test is applied to test the hypothesis. Based on the results of Table 3, both continuous variables have a P-value over 0.05 in the Independent Samples T-test, which means that the null hypothesis cannot be rejected that average values of each continuous variable are the same. Thus, there isn't a significant association between the dependent variable and each of the continuous predictors, so PCTBACHMOR and MEDHHINC may not be good predictors in the model.

There are four assumptions in logistic regression. The first assumption is that the dependent variable must be binary. In this project, the dependent variable DRINKING_D is binary, with values of only 0 and 1, so this assumption is met.

The second assumption is that observations should be independent, so there shouldn't be time, space, or other associations between observations. There is no clear evidence that the observations have temporal or spatial relationships, so this assumption is also met.

Commented [EB12]: Good!

The third assumption is that there is not severe multicollinearity between predictors. Pearson correlations are used to measure the associations between predictors. However, one potential limitation of this method is that Pearson correlations assume the use of continuous variables. Here, the binary variables violate the assumption, which may lead to inaccurate results. In this project, multicollinearity is defined as the absolute value of correlation coefficient greater than 0.8. The correlation matrix of predictors in Table 4 shows the correlation between any pair of predictors is < 0.8 (and > - 0.8), so there is no strong multicollinearity. Thus, this assumption of logistic regression is met.

Table 4 - Pearson Correlation of Predictors

	FATAL_OR_M	OVERTURNED	CELL_PHONE	SPEEDING	AGGRESSIVE	DRIVER1617	DRIVER65PLUS	PCTBACHMOR	MEDHHINC
FATAL_OR_M	1.000	0.033	0.002	0.082	-0.011	-0.003	-0.013	-0.015	-0.018
OVERTURNED	0.033	1.000	-0.001	0.059	0.016	0.004	-0.020	0.009	0.028
CELL_PHONE	0.002	-0.001	1.000	-0.004	-0.026	0.001	-0.003	-0.001	0.002
SPEEDING	0.082	0.059	-0.004	1.000	0.212	0.016	-0.033	-0.001	0.012
AGGRESSIVE	-0.011	0.016	-0.026	0.212	1.000	0.028	0.015	0.027	0.043
DRIVER1617	-0.003	0.004	0.001	0.016	0.028	1.000	-0.021	-0.003	0.023
DRIVER65PLUS	-0.013	-0.020	-0.003	-0.033	0.015	-0.021	1.000	0.026	0.050
PCTBACHMOR	-0.015	0.009	-0.001	-0.001	0.027	-0.003	0.026	1.000	0.478
MEDHHINC	-0.018	0.028	0.002	0.012	0.043	0.023	0.050	0.478	1.000

The last assumption is that at least 50 observations per predictor are needed. In this case, the sample size is 43364, which is greater than needed observations, 50 * 9 = 450. Thus, this assumption is also met.

b) The logistic regression results

Table 5 – Logistic Regression Output

```
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)  -2.733e+00  4.588e-02 -59.563  < 2e-16 ***
FATAL_OR_M    8.140e-01  8.381e-02   9.713  < 2e-16 ***
OVERTURNED    9.289e-01  1.092e-01   8.509  < 2e-16 ***
CELL_PHONE    2.955e-02  1.978e-01   0.149   0.8812
SPEEDING      1.539e+00  8.055e-02  19.107  < 2e-16 ***
AGGRESSIVE   -5.969e-01  4.778e-02 -12.493  < 2e-16 ***
DRIVER1617   -1.280e+00  2.931e-01  -4.367 1.26e-05 ***
DRIVER65PLUS -7.747e-01  9.586e-02  -8.081 6.41e-16 ***
PCTBACHMOR   -3.706e-04  1.296e-03  -0.286   0.7750
MEDHHINC      2.804e-06  1.341e-06   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
```

Table 5 displays the results of the logistic regression. Seven out of nine predictors are significant, including FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS, and MEDHHINC. Among them, FATAL_OR_M, OVERTURNED and SPEEDING have positive coefficient estimates, meaning that as each of these predictors go up and other predictors are fixed, the probability of DRINKING_D, a crash related to drunk driving, goes up. On the contrary, AGGRESSIVE, DRIVER1617, DRIVER65PLUS, and MEDHHINC have negative coefficient estimates, meaning that as each of these predictors go up and other predictors are fixed, the probability of DRINKING_D goes down.

Table 6 – Logistic Regression Output

	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

Table 6 shows the odds ratios and confidence intervals of it. The outcome indicates how the odds of an alcohol involved crash is related to different predictors. For example, the odds ratio of SPEEDING is 4.66, meaning an alcohol involved crash is 4.66 times more likely to involve speeding than not involve speeding. In contrast, the odds of an alcohol involved crash occurring

Commented [EB13]: -3. As the speeding goes up by 1 unit (from not speeding), the odds of there being a drunk driver goes up by a factor of 4.66. Said differently, speeding drivers are 4.66 times more likely to be under the influence (not the other way around).

Also, would like to see this for more variables.

with drivers aged 16-17 is lower than the odds of an alcohol involved crash happening with drivers not between the ages of 16 and 17.

Table 7 – Specificity, Sensitivity, and the Misclassification Rates of Different Cut-offs

Cut-off Value	Sensitivity	Specificity	Misclassification Rate
0.02	0.984	0.058	0.889
0.03	0.981	0.064	0.884
0.05	0.735	0.469	0.516
0.07	0.221	0.914	0.126
0.08	0.185	0.939	0.105
0.09	0.168	0.946	0.099
0.1	0.164	0.948	0.097
0.15	0.104	0.972	0.078
0.2	0.057	0.995	0.060
0.5	0.002	1.000	0.057

The cut-off of 0.5 yields the lowest misclassification rate of 0.057. If the dependent variable's odds are greater than 0.5, then 5.7% of the predictions are incorrect. The cut-off of 0.02 yields the highest misclassification 0.889. In this case, if the odds are greater than 0.02 the crash is predicted to involve alcohol and 88.9% of these predictions are incorrect. Therefore, if specificity is prioritized, 0.5 is the most suitable cut-off value.

Commented [EB14]: -2. Cut-off rate is the predicted probability that y=1, not the odds.

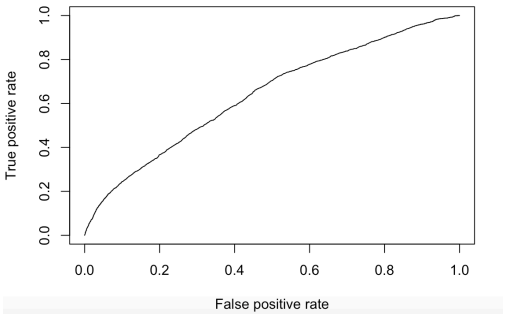

Figure 3 – ROC Curve

Figure 3 shows the ROC curve plot. By minimizing the distance between the upper left corner and the curve, the optimal cut-off selected is:

cutoff	0.06365151
sensitivity	0.66076459
specificity	0.54524328

Compared to the cut-off point found in Table 7 by minimizing the misclassification rates, the ROC curve simultaneously maximizes both sensitivity and specificity. In the case where specificity and sensitivity are weighed equally, this cut-off value is most suitable.

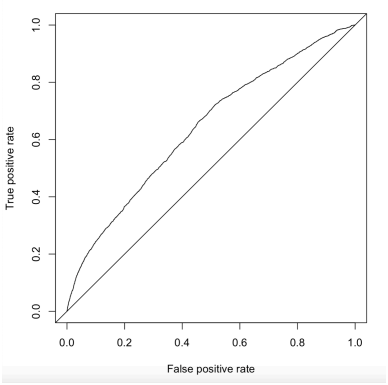

Figure 4 – Area under curve

Figure 4 shows AUC, the area under the ROC curve. AUC of the model is 0.640. As in the Logistic Regression Assessment in the method section above, the performance of this model is poor.

Below is the table of the logistic regression with the binary predictors, excluding continuous variables PCTBACHMOR and MEDHHINC.

Table 8 – Logistic Regression Output of only binary predictors

```
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
```

The result show that all predictors except CELL_PHONE are very significant, just like the previous model. The AIC of the new model remains 18,360, the same as previous model. So, in terms of AIC these models have a very similar performance.

Table 9 – Odds Ratio Output of Only Binary Predictors

	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

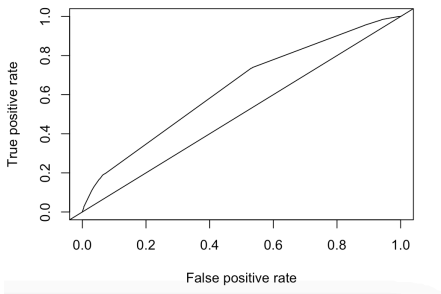

Figure 5 – Area Under Curve of Binary Predictors

In Table 9, the odds ratios don't change much with the previous model. In Figure 5, the AUC is 0.635 slightly lower than the previous model, which means this model is slightly poorer than last one.

Discussion

In the model with all predictors, FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS are of great significance in predicting crashes that involve drunk driving, while CELL_PHONE and PCTBACHMOR are not associated with the depend variable.

The most surprising finding of these results is whether drivers are using cell phones is not relevant to crashes involving alcohol. Common sense suggests that both drunk driving and using a cell phone when driving increase the risk of traffic accidence because one of them decreases a driver's reaction ability and the other distracts the driver. The results do not differ much

between these two dangerous driving behaviors, indicating they increase the risk of traffic crash in different ways.

The unsurprising thing is that speeding is the most related factor to drunk driving. This result corresponds to the fact that alcohol reduces drivers' judgment, including sense of speed and sense of sight. The causality may be that alcohol limits reaction and vision, so drunk drivers may not notice when they are speeding, and then cannot react immediately when a crash happens.

As is mentioned before, rare events in a small dataset may be problematic, because the maximum likelihood estimation of the logistic model suffers from small-sample bias. In this case, although only 5.73% of the dependent variable is '1', the sample size is 43,364 with 2,485 drunk driver samples, so the size of drunk driver samples is big enough and should not be a problem.

There are also several limitations of the analysis. When checking for multicollinearity, the assumption of Pearson correlation is violated. Chi-square Test or similar testing should also be applied to ensure there is no severe multicollinearity in the predictors. Furthermore, the AUC of 0.64 indicates that there is a room for further improvement of the model. Possible improvements may be introducing other predictors like amount of alcohol consumed or time of the day.