Homework 3

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

Yuhao Jia, Zhonghua Yang, Zile Wu

1. Introduction

Alcohol is one of the leading causes of traffic fatalities. According to the National Highway Traffic Safety Administration (NHTSA), drunk driving kills 106 people a day in the U.S. - about one death every 13 minutes. That means more than 38,824 people will be killed by drunk drivers in 2020, accounting for about one-third of all traffic-related deaths. Drunk driving causes more than $44 billion in deaths and damages each year.

Of the 43,464 crashes that occurred in residential block groups of Philadelphia between 2008 and 2012, the percentage of fatal injuries caused by drunk driving was 7.6 percent, compared to 2.9 percent for those who had not been drinking. This again illustrates the seriousness of DUI.

In this study, we will use the statistical programming language R to build a logistic model which is to identify predictors of accidents related to drunk driving. Predictors studied included following variables:

FATAL_OR_M: Crash resulted in fatality or major injury (1 = Yes, 0 = No)
OVERTURNED: Crash involved an overturned vehicle (1 = Yes, 0 = No)
CELL_PHONE: Driver was using cell phone (1= Yes, 0 = No)
SPEEDING: Crash involved speeding car (1 = Yes, 0 = No)
AGGRESSIVE: Crash involved aggressive driving (1 = Yes, 0 = No)
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old (1 = Yes, 0 = No)
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old (1 = Yes, 0 = No)
AREAKEY: ID of the Census Block Group where the crash took place
PCTBACHMOR: % of individuals 25 years of age or older who have at least a bachelor's degree in the Census Block Group where the crash took place
MEDHHINC: Median household income in the Census Block Group where the crash took place.

Some predictors have a more correlated relationship with the dependent variable, such as fatal accidents, speeding and aggressive driving in which DUI rates tend to be higher because alcohol makes drivers more likely to lose control of their minds and make dangerous moves. In the article, we will discuss these issues in detail.

2. Methods

a) Problems with Using OLS Regression for Binary DV.

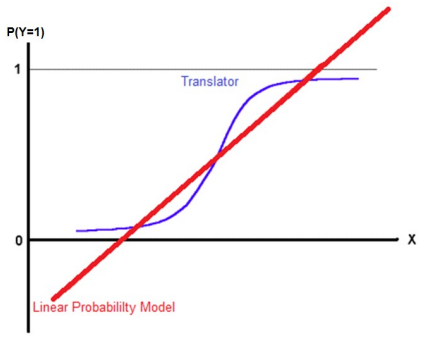
For OLS, the model can be interpreted as the following formula:

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

The relationship between the variables in this formula can be explained by the value of the coefficient, which means that when the predictor variable x changes by one unit, the dependent variable y changes by β1. However, when the dependent variable is binary, the change of y is only binary, which is either 1 or 0. Therefore, it is meaningless to explain the change of y with OLS, because the dependent variable will not increase or decrease the value of β1 due to the change of the prediction variable. Thus, the OLS prediction model describing binary variables is not valid.

b) Logistic Regression

In logistic regression, it is no longer possible to get the regression result by predicting the value of Y, so we can only get the prediction result by other methods. Instead of directly predicted Y value, let's predict the probability of Y occurring, the P(Y=1|X=x). Sometimes our regression models get results in the range of -∞ and +∞, for example, the predicted positive estimate can be 1.3, negative estimate can be -0.2, but for probability values should be 0 to 1, so these calculations have no real value. We need a transformation process to turn the linear regression result into a real and usable probability number using a function that converts the value between -∞ and +∞ to the corresponding value between 0 and 1. The closer the Y result in the linear regression model is to +∞, the closer the probability of return is to 1; The closer the Y result in the linear regression model is to -∞, the closer the probability of return is to 0. Here's a mathematical diagram:



The concept of odds is included in the conversion function. The concept of odds should be understood before understanding the conversion function. Compare odds with probability:

• Probability may be calculated as $\frac{\#\ desirable\ outcomes}{\#\ possible\ outcomes}$

• Odds may be calculated as $\frac{\#\ desirable\ outcomes}{\#\ undesriable\ outcomes}$

For example, set the probability as the probability that there is a hospital in a certain zip code, that is, Y=1 (hospital in zip code), which is expressed by the mathematical formula:

$$P(Y = 1) = \frac{\#\ zip\ codes\ where\ there's\ a\ hospital}{\#\ zip\ codes}$$

According to the above definition of odds and probability, odds can be written as:

$$Odds(Y = 1) = \frac{\#\ zip\ codes\ where\ there's\ a\ hospital}{\#\ zip\ codes\ where\ there's\ no\ hospital} =$$

$$\frac{\frac{\#\ zip\ codes\ where\ there's\ a\ hospital}{\#\ zip\ codes}}{\frac{\#\ zip\ codes\ where\ there's\ no\ hospital}{\#\ zip\ codes}} = \frac{P(Y=1)}{P(Y\neq 1)} = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{p}{1-p}.$$

One type of conversion function is the logit function. A logit model with one predictor is:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \varepsilon$$

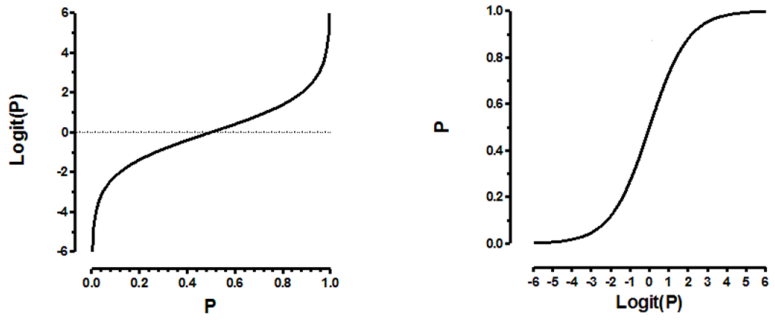
In the interpretation of odds, we can know that $p = P(Y = 1)$ in the above formula is probability. And the quantity on the left $\frac{p}{1-p}$ is called the odds, then to logarithm calculation of odds, $\ln\left(\frac{p}{1-p}\right)$ is called the log odds, or logit.

Another conversion function is the logistic function, which is also called the Inverse-Logit Function. The mathematical formula is:

$$p = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} = \frac{1}{1+e^{-\beta_0-\beta_1 x_1}}$$

There is only one prediction variable in this mathematical formula. Later, I will list the multivariable conversion function with the variables in the prediction process, so as to explain the method of this assignment more intuitively. Logistic function can function in $\hat{\beta}_0 + \hat{\beta}_1 x_1$ value to measure the size of the probability, when $\hat{\beta}_0 + \hat{\beta}_1 x_1$ has a value of 0, the probability is 0.5, when $\hat{\beta}_0 + \hat{\beta}_1 x_1$ is larger, the probability is more close to 1, on the contrary, When the $\hat{\beta}_0 + \hat{\beta}_1 x_1$ is smaller, the probability is close to zero.

The two formulas can be graphed as follows:



The diagram on the left is the logit function, and the diagram on the right is the Inverse-Logit Function.

There is also the concept of odds ratios, which are the exponentiations of prediction variables corresponding to forecast intercepts, and they range from 0 to infinity, meaning they cannot be negative. Odds <1 indicates that there is negative correlation between predictor x and dependent variable, and Odds ratios = 1 indicates that there is no relationship between predictor x and dependent variable, odds > 1 indicates that there is a positive correlation between the predicted factor x and the dependent variable.

The next step is to consider multiple logistic regression, where the formula will bring in multiple predictors and the variable names from this assignment. The case with multiple predictors is not very different from the case with only one predictor. Here are the two formulas:

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

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

In the formula with multivariate variables, we also need to follow the same rule as OLS regression to ensure that there is no multicollinearity. When we interpret and study one variable, we need to keep other predictive variables unchanged.

c) The Hypothesis Tested for Each Predictor

In general, we will conduct two hypothesis tests for each prediction variable x_i, which are:

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

The former is when OR is equal to 1, that is, there is no correlation between the predictor and the dependent variable. This hypothesis test is called null hypotheses. The latter is when OR is not equal to 1, in which case the prediction variable is related to the dependent variable, called alternative hypotheses.

We can know the mean of the $\hat{\beta}_i$ model results E is equal to zero, therefore, there are $\frac{\hat{\beta}_i - E(\hat{\beta}_i)}{\sigma_{\hat{\beta}_i}} = \frac{\hat{\beta}_i - 0}{\sigma_{\hat{\beta}_i}} = \frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}$, And $\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}$conforms to the standard normal distribution, according to the standard definition of score, can get:

$$\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}} = z$$

$\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}$values in the context of logistic regression is sometimes referred to as the wald statistic, this suggests that the wald statistic is accord with a standard normal distribution. P-value can be obtained using the standard normal table of z, which is used to evaluate the contribution value of each predictor variable to the regression results of the dependent variable.

d) The Quality of Model Fit.

We know that in OLS regression, the value of R-squared can be used as the basis for the evaluation of predictive variables, but R-squared cannot be used to explain the value percentage of variables in the model in logistics regression. When conducting logistics regression experiment with tools, R-squared value will not be generated as the evaluation index. In general, we use the stand or fall of residual error to evaluate the results of prediction, residual is defined as the $\varepsilon = y_i - \hat{y}_i$, is to use observed value minus the predicted value, the judgment method is also suitable for logistics regression evaluation period. We also require residual $\varepsilon = y_i - \hat{y}_i$, only here $\hat{y}_i$ is our prediction probability, and more specifically the probability is the result of P (y = 1). If the true value of Y is 1, the probability of predicting P(Y=1) will be higher; if the true value of Y is 0, the

probability of predicting P(Y=1) will be lower.

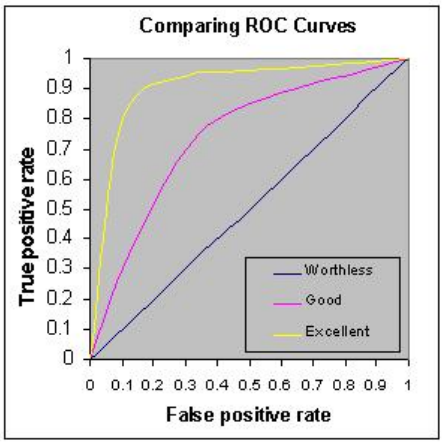
We used an index called AIC in a geographically weighted regression, which is used to measure model performance and help compare different regression models. The smaller AIC is, the better the model can fit the observed data. However, when the amount of data is small, AIC is more likely to select models with too many parameters, and small sample size needs to be corrected, which is called AICc. Neither AIC nor AICc are absolute measures of goodness of fit, but as long as they apply to the same dependent variable, they are useful for comparing models with different explanatory variables.

In determining whether or not we need a threshold to divide all the probability values as the correct result, when the probability of the income is greater than this threshold we can think that the prediction is credible, and when the probability of the income is less than this threshold, we think that the prediction is not acceptable. So we need to find the most appropriate threshold, and we call it the alpha value. This indicator requires different values as the model and the prediction variable need to be chosen, because the distribution of the probability is divided, and in general, the histogram of the predicted value can be observed to determine the object.

There are some definitions associated with predictive values and observations in logical regression. Sensitivity (also called the true positive rate) measures the proportion of actual positives which are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate.

Specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate.

The ROC curve is a method of visualization true positive rate(sensitivity) and false positive rate(specificity). Using the ROC curve to observe the prediction of the model, the relationship between sensitivity and specificity is presented by curved linear form.



It is feasible to use the ROC curve to determine the best cut-off value, and there are many different ways to determine the final judgment value based on the ROC curve. One is Youden Index, which uses the judgment value at the maximum sum of sensitivity and specificity as the threshold value. One is the critical value of the minimum distance

between the ROC curve and the upper left corner of the graph, that is, the point with specificity = 1 and sensitivity = 1. It's really just another way to maximize specificity and sensitivity. In this experiment, we chose the second method.

The Area Under ROC Curve is also meaningful, which is called the AUC (Area Under Curve). AUC is an important index used by models to measure the accuracy of models. The higher AUC value indicates the higher sensitivity and specificity of our final results, which is in line with our expectations. The AUC has a value between 0.5 (the area of the slope below 45 degrees) and 1 (the area of the entire image box). Generally speaking, some rough values can basically determine the quality of the model:

- 0.9-1 = excellent
- 0.8-0.9 = good
- 0.7-0.8 = fair
- 0.6-0.7 = poor
- 0.5-0.6 = fail

In general, the results of the model are acceptable when the AUC is greater than 0.7.

e) Assumptions of Logistic Regression.

Compared to OLS regression, logistic regression has some assumptions that are the same as OLS, but some are not used. In logistic regression, the dependent variable must be binary but doesn't require a linear relationship between the dependent variable and each predictive variable as OLS. The observed values of logistic regression and OLS were independent. For collinearity between variables, OLS requires no multicollinearity, while logistic regression is not so strict, requiring no serious multicollinearity. The residual of logistic regression does not require normality, but the residual of OLS does; Logistic regression does not assume Homoscedasticity, whereas OLS does. Logistic regression has a larger data sample than OLS, because logistic regression uses the maximum likelihood estimation method instead of the least square method, and each prediction variable needs at least 50 observed values, while OLS is usually more than 10 can be carried out.

f) Exploratory Analyses

Unlike the OLS regression model we conducted before. Due to the presence of categorical data in the logistic regression model, we use the Chi-Square (χ^2) test to determine whether the distribution of one categorical variable varies with respect to the values of another categorical variable. If we were to look are a cross-tabulation of the variables DRINKING_D and SPEEDING, the null and alternative hypotheses for the (χ^2) test would be as follows:

H_0: the proportion of speeding that involve drunk drivers is the same as the proportion of speeding that don't involve drunk drivers,

$$vs.$$

H_a: the proportion of speeding that involve drunk drivers is different than the proportion of speeding that don't involve drunk drivers.

A high value of the χ^2 statistic, and a p-value lower than 0.05 suggest that there's evidence to reject the null hypothesis in favor of the alternative, and that there's a

significant association between drunk driving and speeding.

When identify the correlation between a continuous variable and two independent groups, we can employ a test that's called the independent samples t-test. For example, we can see whether the average PCTBACHMOR values are statistically significantly 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 PCTBACHMOR (or MEDHHINC) are the same for crashes that involve drunk drivers and crashes that don't.

$$vs.$$

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

A high value of the t-statistic, and a p-value lower than 0.05 suggest that there's evidence to reject the null hypothesis in favor of the alternative.

3. Results

a) Results of the Exploratory Analyses.

First of all, through the tabulation of the dependent variable, we can see that there was a total of 2,485 accidents involving alcohol, accounting for 5.73% of the total valid data. Specific to the type of accident collision, in the *Hit fixed object* category, the highest percentage of accidents involving alcohol, accounting for 22.7%. In the Head-on; Sideswipe (same dir. & Opposite die.); Hit fixed object and Hit pedestrian categories, the percentage of accidents involving alcohol was more than double the percentage of accidents not involving alcohol.

	No Alcohol Involved (DRINKING_D = 0)	Alcohol Involved (DRINKING_D = 1)
Total Number	40879	2485
Proportion	0.9427	0.0573

Cell Contents

N
N / Col Total

Total Observations in Table: 43364

mydata$COLLISION_		mydata$DRINKING_D 0	1	Row Total
Non collision	0	391 0.010	17 0.007	408
Rear-end	1	8426 0.206	591 0.238	9017
Head-on	2	1747 0.043	190 0.076	1937
Rear-to-rear (Backing)	3	148 0.004	7 0.003	155
Angle	4	14092 0.345	591 0.238	14683
Sideswipe (same dir.)	5	3006 0.074	321 0.129	3327
Sideswipe (Opposite dir.)	6	1159 0.028	111 0.045	1270
Hit fixed object	7	4184 0.102	565 0.227	4749
Hit pedestrian	8	7608 0.186	88 0.035	7696
Other or Unknown	9	118 0.003	4 0.002	122
Column Total		40879 0.943	2485 0.057	43364

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 903.5373 d.f. = 9 p = 1.06981e-188

Secondly, Through the cross-tabulation of the dependent variable with each of the binary predictors. We can learn the following information:

For Predictor FATAL_OR_M, 7.60% of accidents involving drunk drivers and 2.90% of accidents NOT involving drunk drivers. The p-value of the Chi-Square (χ^2) test <0.001 which means that there's a significant association between drunk driving and crash fatalities.

For Predictor OVERTURNED, 4.40% of accidents involving drunk drivers and 1.50% of accidents NOT involving drunk drivers. The p-value of the Chi-Square (χ^2) test <0.001 which means that there's a significant association between drunk driving and crash fatalities.

For Predictor CELL_PHONE, 1.10% of accidents involving drunk drivers and 1.00% of accidents NOT involving drunk drivers. The p-value of the Chi-Square (χ^2) test is 0.687>0.05, which means that there's NOT a significant association between drunk driving and crash fatalities.

For Predictor SPEEDING, 10.50% of accidents involving drunk drivers and 3.10% of accidents NOT involving drunk drivers. The p-value of the Chi-Square (χ^2) test <0.001 which means that there's a significant association between drunk driving and crash fatalities.

For Predictor AGGRESSIVE, 35.90% of accidents involving drunk drivers and 45.30% of accidents NOT involving drunk drivers. The p-value of the Chi-Square (χ^2) test <0.001 which means that there's a significant association between drunk driving and crash fatalities.

For Predictor DRIVER1617, 0.50% of accidents involving drunk drivers and 1.60% of accidents NOT involving drunk drivers. The p-value of the Chi-Square (χ^2) test <0.001 which means that there's a significant association between drunk driving and crash fatalities.

For Predictor DRIVER65PLUS, 4.80% of accidents involving drunk drivers and 10.40% of accidents NOT involving drunk drivers. The p-value of the Chi-Square (χ^2) test <0.001 which means that there's a significant association between drunk driving and crash fatalities.

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		Total	χ^2 p-value
	N	%	N	%	N	
FATAL_OR_M: Crash resulted in fatality or major injury	1181	2.90%	188	7.60%	1369	<0.001
OVERTURNED: Crash involved an overturned vehicle	612	1.50%	110	4.40%	722	<0.001
CELL_PHONE: Driver was using cell phone	426	1.00%	28	1.10%	454	0.687
SPEEDING: Crash involved speeding car	1261	3.10%	260	10.50%	1521	<0.001
AGGRESSIVE: Crash involved aggressive driving	18522	45.30%	916	35.90%	19438	<0.001
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old	674	1.60%	12	0.50%	686	<0.001
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old	4237	10.40%	119	4.80%	4356	<0.001

Thirdly, Through the cross-tabulation of the means of the continuous predictor for both values of the dependent variable. We can learn the following information:

For Predictor PCTBACHMOR: The average PCTBACHMOR values is 16.61 for involving drunk drivers and 16.57 for NOT involving drunk drivers. However, the p-value for t-test is 0.9137>0.05, which means there is NOT a significant association between the average of values of the variable PCTBACHMOR and drunk driving.

For Predictor MEDHHINC: The average MEDHHINC values is 31998.75 for involving drunk drivers and 31483.05 for NOT involving drunk drivers. However, the p-value for t-test is 0.16>0.05, which means there is NOT a significant association between the average of values of the variable MEDHHINC and drunk driving.

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		t-test p-value
	Mean	SD	Mean	SD	
PCTBACHMOR: % with bachelor's degree or more	16.57	18.21	16.61	18.72	0.9137
MEDHHINC: Median household income	31483.05	16930.1	31998.75	17810.5	0.16

b) Logistic Regression Assumptions

The assumptions that the dependent variable is binary and observations are independent are met. We have more than 50 observations per predictor which means the large samples assumption is also met.

The test results of pairwise Pearson correlations for all predictors is shown as follows.

The limitation of using Pearson correlations to measure the associations between binary predictors is that the coefficient is hard to interpret. It's unreasonable to say that with one unit change of a binary predictor is associated with more or less than one unit change of another binary predictor. As shown in the table below, there is no correlation larger than 0.9 or less than -0.9, which means the assumption of no multicollinearity is also met.

	FATAL_OR_M	OVERTURNED	CELL_PHONE	SPEEDING	AGGRESSIVE	DRIVER1617	DRIVER65PLUS	PCTBACHMOR	MEDHHINC
FATAL_OR_M	1.000000000	0.0331959240	0.0021603225	0.0817126678	-0.01104729	-0.002808379	-0.012512349	-0.01465226448	-0.018212431
OVERTURNED	0.033195924	1.0000000000	-0.0009897786	0.0594402861	0.01643894	0.003723967	-0.019500974	0.0093321352	0.027921303
CELL_PHONE	0.002160322	-0.0009897786	1.0000000000	-0.0036011640	-0.02574299	0.001485133	-0.002717259	-0.00124585540	0.002099885
SPEEDING	0.081712668	0.0594402861	-0.0036011640	1.0000000000	0.21152537	0.016011600	-0.032854111	-0.0007390853	0.011786681
AGGRESSIVE	-0.011047295	0.0164389397	-0.0257429929	0.2115253684	1.00000000	0.028428953	0.015026930	0.0271221096	0.043440451
DRIVER1617	-0.002808379	0.0037239674	0.0014851333	0.0160115997	0.02842895	1.000000000	-0.020848417	-0.0026359662	0.022877425
DRIVER65PLUS	-0.012512349	-0.0195009743	-0.0027172590	-0.0328541108	0.01502693	-0.020848417	1.000000000	0.0261903901	0.050337711
PCTBACHMOR	-0.014652265	0.0093321352	-0.0012458540	-0.0007390853	0.02712211	-0.002635966	0.026190390	1.0000000000	0.477869537
MEDHHINC	-0.018212431	0.0279213029	0.0020998852	0.0117866805	0.043440445	0.022877425	0.050337711	0.4778695368	1.000000000

c) Logistic Regression Results.

Below is the result of the logistic regression with all predictors.

The coefficient of FATAL_OR_M is 8.140×10^{-01} and the predictor is significant, indicating that as the variable goes up 1 unit (from 0=No to 1=Yes), the odds of the driver was drinking are about $e^{8.140\times10^{-01}} = 2.2569$ the odds of the driver was not drinking. The odds ratio is with 95% confidence interval.

The coefficient of OVERTURNED is 9.289×10^{-01} and the predictor is significant, indicating that as the variable goes up 1 unit (from 0=No to 1=Yes), the odds of the driver was drinking are about $e^{9.289\times10^{-01}} = 2.5318$ the odds of the driver was not drinking. The odds ratio is with 95% confidence interval.

The coefficient of CELL_PHONE is 2.955×10^{-02} while the predictor is not significant, indicating that we fail to reject the H_0: no correlation between predictor and dependent variable from H_a: as the variable goes up 1 unit (from 0=No to 1=Yes), the odds of the driver was drinking are about $e^{2.955\times10^{-02}} = 1.0300$ the odds of the driver was not drinking. The odds ratio is with 95% confidence interval.

The coefficient of SPEEDING is 1.539×10^{00} and the predictor is significant, indicating that as the variable goes up 1 unit (from 0=No to 1=Yes), the odds of the driver was drinking are about $e^{1.539\times10^{00}} = 4.6598$ the odds of the driver was not drinking. The odds ratio is with 95% confidence interval.

The coefficient of AGGRESSIVE is -5.969×10^{-01} and the predictor is significant, indicating that as the variable goes up 1 unit (from 0=No to 1=Yes), the odds of the driver was drinking are about $e^{-5.969\times10^{-01}} = 0.5505$ the odds of the driver was not drinking. The odds ratio is with 95% confidence interval.

The coefficient of DRIVER1617 is -1.280×10^{00} and the predictor is significant, indicating that as the variable goes up 1 unit (from 0=No to 1=Yes), the odds of the driver was drinking are about $e^{-1.280\times10^{00}} = 0.2780$ the odds of the driver was not drinking. The odds ratio is with 95% confidence interval.

The coefficient of DEIVE65PLUS is -7.747×10^{-01} and the predictor is significant, indicating that as the variable goes up 1 unit (from 0=No to 1=Yes), the odds of the driver was drinking are about $e^{-7.747\times10^{-01}} = 0.4609$ the odds of the driver was not drinking. The odds ratio is with 95% confidence interval.

The coefficient of PCTBACHMOR is -3.706×10^{-04} while the predictor is not significant, indicating that we fail to reject the H_0: no correlation between predictor and dependent variable from H_a: the odds ratio $e^{-3.706\times10^{-04}} = 0.9996$ is the extent to

which the odds of Y=1 (drinking driver) change as the percent of individuals with at least Bachelor's degrees increases by 1 unit. The odds ratio is with 95% confidence interval.

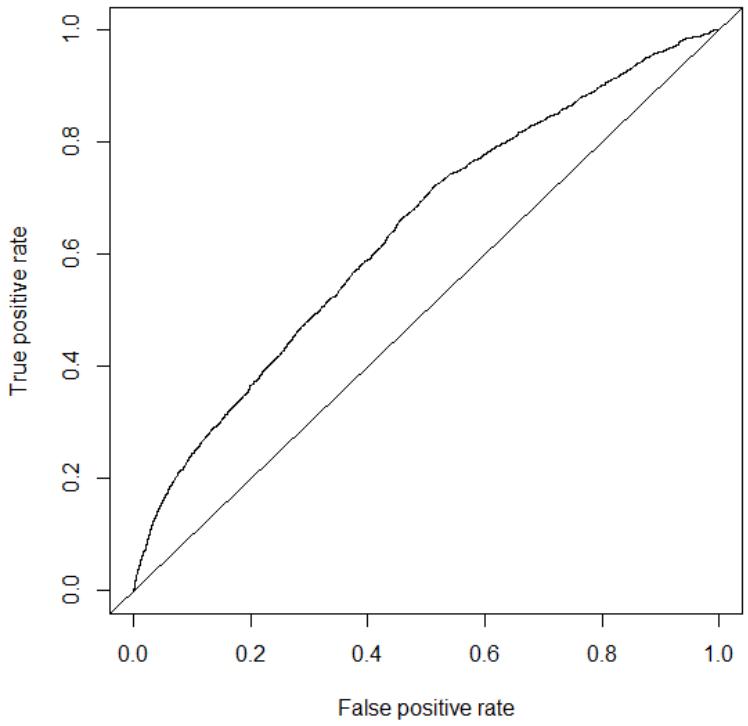
The coefficient of MEDHHINC is -2.804×10^{-06} and the predictor is significant, indicating that the odds ratio $e^{-2.804\times10^{-06}} = 1.0000$ is the extent to which the odds of Y=1 (drinking driver) change as the median household income increases by 1 unit. The odds ratio is approximately 1 with 95% confidence interval, which means this predictor is not related to dependent variable.

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

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

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.733e+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

                Estimate   Std. Error     z value     Pr(>|z|)
(Intercept)  -2.732507e+00 4.587566e-02 -59.5633209 0.000000e+00
FATAL_OR_M    8.140138e-01 8.380692e-02   9.7129660 2.654967e-22
OVERTURNED    9.289214e-01 1.091663e-01   8.5092302 1.750919e-17
CELL_PHONE    2.955008e-02 1.977778e-01   0.1494105 8.812297e-01
SPEEDING      1.538976e+00 8.054589e-02  19.1068171 2.215783e-81
AGGRESSIVE   -5.969159e-01 4.777924e-02 -12.4932079 8.130791e-36
DRIVER1617   -1.280296e+00 2.931472e-01  -4.3674171 1.257245e-05
DRIVER65PLUS -7.746646e-01 9.585832e-02  -8.0813505 6.405344e-16
PCTBACHMOR   -3.706336e-04 1.296387e-03  -0.2858974 7.749567e-01
MEDHHINC      2.804492e-06 1.340972e-06   2.0913870 3.649338e-02
Waiting for profiling to be done ...
                    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
                Estimate   Std. Error     z value     Pr(>|z|)         OR      2.5 %     97.5 %
(Intercept)  -2.732507e+00 4.587566e-02 -59.5633209 0.000000e+00 0.06505601 0.05947628 0.07119524
FATAL_OR_M    8.140138e-01 8.380692e-02   9.7129660 2.654967e-22 2.25694878 1.90991409 2.65313350
OVERTURNED    9.289214e-01 1.091663e-01   8.5092302 1.750919e-17 2.53177687 2.03462326 3.12242730
CELL_PHONE    2.955008e-02 1.977778e-01   0.1494105 8.812297e-01 1.02999102 0.68354737 1.48846840
SPEEDING      1.538976e+00 8.054589e-02  19.1068171 2.215783e-81 4.65981462 3.97413085 5.45020642
AGGRESSIVE   -5.969159e-01 4.777924e-02 -12.4932079 8.130791e-36 0.55050681 0.50101688 0.60423487
DRIVER1617   -1.280296e+00 2.931472e-01  -4.3674171 1.257245e-05 0.27795502 0.14774429 0.47109277
DRIVER65PLUS -7.746646e-01 9.585832e-02  -8.0813505 6.405344e-16 0.46085831 0.37998364 0.55347851
PCTBACHMOR   -3.706336e-04 1.296387e-03  -0.2858974 7.749567e-01 0.99962944 0.99707035 1.00215087
MEDHHINC      2.804492e-06 1.340972e-06   2.0913870 3.649338e-02 1.00000280 1.00000013 1.00000539
```

Below is the table showing the specificity, sensitivity and misclassification rates for the different probability cut-offs. It can be seen that the 0.5 cut-offs yield the lowest misclassification rates while the 0.02 cut-offs yield the highest.

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	0.098
0.15	0.104	0.972	0.078
0.2	0.023	0.995	0.060
0.5	0.002	1	0.057

Below is the ROC curve for the model. The optimal cut-off rate that was selected by minimizing the distance from the upper left corner is 0.064, with 0.66 sensitivity and 0.55 specificity. Compared to the 0.5 cut-offs above, they are similar but not the same. They are different ways of maximizing specificity and sensitivity.



Below shows the area under the ROC Curve calculated in R. The area under ROC Curve (AUC) is a measure of prediction accuracy of the model. In our model, the AUC is 0.6398 indicating that the classifying accuracy is poor.

```
[[1]]
[1] 0.6398695
```

Below is the result of the logistic regression with binary predictors only. Comparing

to the results of the first regression, all predictors which are significant in this regression are significant in the original one. The only not significant predictor CELL_PHONE is also not significant in the original model.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1961  -0.3692  -0.3153  -0.2764   3.0093

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.65190    0.02753 -96.324  < 2e-16 ***
FATAL_OR_M    0.80932    0.08376   9.662  < 2e-16 ***
OVERTURNED    0.93978    0.10903   8.619  < 2e-16 ***
CELL_PHONE    0.03107    0.19777   0.157    0.875
SPEEDING      1.54032    0.08053  19.128  < 2e-16 ***
AGGRESSIVE   -0.59365    0.04775 -12.433  < 2e-16 ***
DRIVER1617   -1.27158    0.29311  -4.338 1.44e-05 ***
DRIVER65PLUS -0.76646    0.09576  -8.004 1.21e-15 ***
---
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

                Estimate Std. Error     z value     Pr(>|z|)
(Intercept)  -2.65189961 0.02753107 -96.3238683 0.000000e+00
FATAL_OR_M    0.80931557 0.08376150   9.6621431 4.366327e-22
OVERTURNED    0.93978420 0.10903433   8.6191585 6.744795e-18
CELL_PHONE    0.03107367 0.19777088   0.1571195 8.751506e-01
SPEEDING      1.54032033 0.08052787  19.1277908 1.482240e-81
AGGRESSIVE   -0.59364687 0.04774781 -12.4329656 1.730916e-35
DRIVER1617   -1.27157607 0.29310969  -4.3382260 1.436374e-05
DRIVER65PLUS -0.76645727 0.09576440  -8.0035718 1.208612e-15
Waiting for profiling to be done ...
                    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
                Estimate Std. Error     z value     Pr(>|z|)         OR      2.5 %    97.5 %
(Intercept)  -2.65189961 0.02753107 -96.3238683 0.000000e+00 0.07051713 0.06678642 0.0743978
FATAL_OR_M    0.80931557 0.08376150   9.6621431 4.366327e-22 2.24636998 1.90112455 2.6404533
OVERTURNED    0.93978420 0.10903433   8.6191585 6.744795e-18 2.55942903 2.05736015 3.1556897
CELL_PHONE    0.03107367 0.19777088   0.1571195 8.751506e-01 1.03156149 0.68459779 1.4907150
SPEEDING      1.54032033 0.08052787  19.1277908 1.482240e-81 4.66608472 3.97961862 5.4573472
AGGRESSIVE   -0.59364687 0.04774781 -12.4329656 1.730916e-35 0.55230941 0.50268818 0.6061758
DRIVER1617   -1.27157607 0.29310969  -4.3382260 1.436374e-05 0.28038936 0.14904734 0.4751771
DRIVER65PLUS -0.76645727 0.09576440  -8.0035718 1.208612e-15 0.46465631 0.38318289 0.5579332
```

Below is a comparison of the AIC for both models. The AIC for the first model is slightly lower than the second one. However, the difference is less than 3, which means we can say that the power of two models is pretty much the same.

	df <dbl>	AIC <dbl>
mylogit	10	18359.63
mylogit1	8	18360.47

4. Discussion

In this research, we explored the cross-tabulation between each predictor and the binary dependent variable DRINKING_D (whether the crash involves a drinking driver). Then we ran two logistic regressions model to examine the relationship between each predictor and the dependent variable. One model includes all predictors and the other only includes binary

predictors.

According to the result of the regression with all predictors, FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617 and DRIVER65PLUS are strong predictors of crashes that involve drunk driving. CELL_PHONE, PCTBACHMOR and MEDHHINC aren't associated with the dependent variable.

CELL_PHONE might be a major cause of crashes, but it's not always related to drunk driving, not surprising that it's not significant. Some of these results are surprising. We expected PCTBACHMOR to be significant while it's not; We expected MEDHHINC to be negative related the dependent variable while it's significantly not associated with drunk driving; We expected AGGRESSIVE to be positive related to the dependent variable while it's significantly negative associated with drunk driving.

Although we have small proportion on one category of the dependent variable, the sample size in our case is pretty large so we believe our model does not suffer much from small-sample bias. The logistic regression is appropriate here and there is no need to use rare events method.

There are still some limitations of this analysis as talked above. It's hard to explain the correlation between binary predictors. The low AUC indicates that we may need more predictors to make the model more powerful.