CPLN 671/MUSA500

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

Yuewen Dai, Xinge Zhang, Yingxue Ou

1) Introduction

In Philadelphia, there's an average of 30 people die every day in a motor vehicle accident caused by a drunk driver, which is a terrible statistic. This has caused individuals injured as well as negative economic impact. It is necessary to analyze the factors related to these accidents in order to predict drunk driving to reduce the accident rate and better protect the safety of the public. So, we acquired the drinking driver data and the possible related predictors, such as the collision reason, the crash result (fatality or major injury), the status of driving before the crash (speeding, overturned, etc.), driver's age (between 16-17, more than 65), and census data. These data belong to the dataset of all 53,260 car crashes in Philadelphia for the year 2008-2012.

Our goal in this research is to identify predictors of accidents related to drunk driving. To predict the drunk driving, we should run the Logistic regression in R to see the extent to which drunk driving might be associated with the predictors. Since the drunk driving data is binary (0&1), and our goal is to predict the probability (ranges 0-1) that the drunk driving might happen, we cannot use the OLS method which we have been using for the past two assignments because the result might exceed 1 or less than 0.

2) Methods

2.1 The problem with using OLS regression when the dependent variable is binary

In OLS regression, the data we are dealing with are continuous variables (with errors normally distributed), which have no upper and lower boundaries for the value. But for binary variables, there are only two values, True=1 and False=0. In the OLS equation, if we're dealing with binary variables, instead of predicting Y, we're predicting P(Y=1|X=x), the probability that Y=1. So, the equation would be:

$$P(Y = 1) = \beta_0 + \beta_1 x_1 + \varepsilon$$

But it's obvious that the probability of Y will increase with x increases and decrease with x decreases, and the value has no upper and lower boundaries. So, if we continue using OLS regression as the predicting method, this will lead to a problem where the predicted probability will be more than 1 or less than 0, which is clearly incorrect. This is why we have to introduce logistic regression method to deal with the binary dependent variable.

2.2 The method of Logistic Regression

The logit model with one predictor looks like the following equations where p equals P(Y=1):

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

The quantity $\frac{p}{1-p}$ is called the odds, and $\ln\left(\frac{p}{1-p}\right)$ is called **log odds**. The odds can be calculated as #desirable outcomes/ #undesirable outcomes, which means the probability that one event will occur divided by the probability that the event will not occur.

The equation of logistic model with one predictor looks like this:

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

The odds ratio e^{β_i} is a statistic that quantifies the strength of the association between two events,[1] for example, the ratio of the odds of A in the presence of B and the odds of A in the absence of B. The odds ratio can range from 0 to infinity, when it is smaller than 1, it indicates that there's a negative association between the predictor i and the dependent variable; when it's equal to 1, it means there's no relationship between them; and when it's larger than 1, meaning that there's a positive association.

In this report, the above equations can be written in the **logit form**:

$$\begin{aligned}\ln\left(\frac{p}{1-p}|The\ driver\ is\ drunk\right) &= \beta_0 + \beta_1 FATAL_{OR_M} + \beta_2 OVERTURNED, + \beta_3 CELL_{PHONE} + \beta_4\, SPEEDING \\ &+ \beta_5\, AGGRESSIVE, + \beta_6 DRIVER1617 + \beta_7\, DRIVER65PLUS \\ &+ \beta_8 PCTBACHMOR + \beta_9 MEDHHINC + \beta_{10} COLLISION + \varepsilon\end{aligned}$$

Or the **logistic form***:

$$p(The\ driver\ is\ drunk) = \frac{e^{\beta_i}}{1 + e^{\beta_i}}$$

***To let the formula looks more simple, I set:**

$$\beta_i = \beta_0 + \beta_1 FATAL_{OR_M} + \beta_2 OVERTURNED, + \beta_3 CELL_{PHONE} + \beta_4 SPEEDING + \beta_5 AGGRESSIVE, + \beta_6 DRIVER1617 + \beta_7 DRIVER65PLUS + \beta_8 PCTBACHMOR + \beta_9 MEDHHINC + \beta 10\ COLLISION$$

Dependent variable:

p: P (The driver is drunk), the probability of the driver is drunk.

Binary variables:

β_1: Crash resulted in fatality or major injury multiplies by e^{β_1} the probability of when the driver is drunk compared to not resulted in fatality of major injury.

β_2: Crash involves an overturned vehicle multiplies by e^{β_2} the probability of when the driver is drunk compared to not involves an overturned vehicle.

β_3: Crash happens when driver is using cell phone multiplies by e^{β_3} the probability of when the driver is drunk compared to the driver is not using cell phone.

β_4: Crash involves speeding car multiplies by e^{β_4} the probability of when the driver is drunk compared to not involves speeding car.

β_5: Crash involves aggressive driving multiplies by e^{β_5} the probability of when the driver is drunk compared to not involves aggressive driving.

β_6: Crash involves at least one driver who was 16 or 17 years old multiplies by e^{β_6} the probability of when the driver is drunk compared to not involves at least one driver who was 16 or 17 years old.

β_7: Crash involves at least one driver who was at least 65 years old multiplies by e^{β_7} the probability of when the driver is drunk compared to not involves at least one driver who was at least 65 years old.

Continuous variables:

β_8: An increase of 1 unit in % 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 multiplies the odds of drunk driver by e^{β_8}

β_9: An increase of 1 unit in Median household income in the Census Block Group where the crash took place multiplies the odds of drunk driver by e^{β_9}

Category variable:

β_{10}: Going up from 1 category of collision to the next multiplies the odds of heart disease by $e^{\beta_{10}}$.

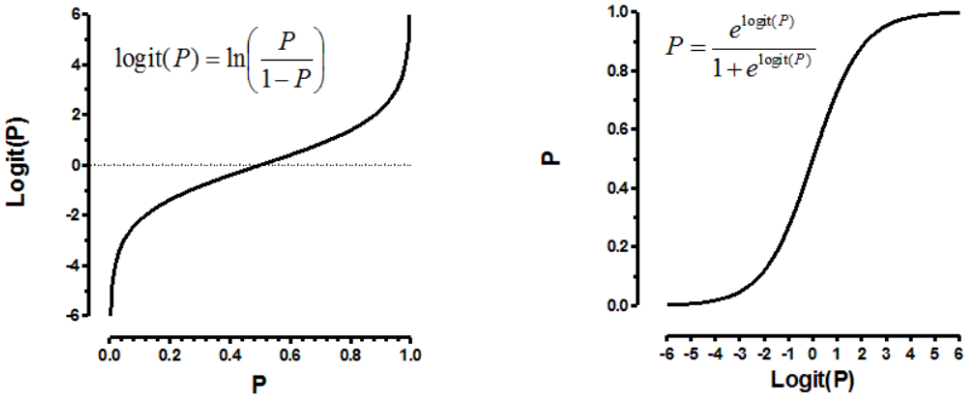

Figure2.1 The logit function and the logistic function

(Source: http://www.graphpad.com/support/faqid/1465/)

The Figure 2.1 shows the comparison between the logit function and the logistic function. The outcomes of the logit function have no upper and lower boundaries, P between 0-0.5 has negative logit values and P between 0.5-1 has positive logit values that can tend to be infinity. The logistic function predicts the logarithm of the odds, it is a model that always predict a value of probability that is between 0-1. Since we want to predict the probability of the binary outcome, and the probability ranges between 0-1, the logistic function is obviously a more appropriate one.

2.3 Hypothesis testing

We are doing the hypothesis test for each of the predictors:

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

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

The **Null Hypothesis** can be interpreted as the odds ratio equals to 1, meaning that there's no relationship between the predictor i and the dependent variable. And the **Alternative Hypothesis** can be interpreted as there's a relationship between the predictor and the dependent variable. The p-value is obtained from the standard normal z tables to test the null hypothesis. The z-value, $\frac{\widehat{\beta}_i}{\sigma_{\widehat{\beta}_i}}$, also called **Wald statistic**, can be used to do the testing because it has a standard normal distribution.

When we are testing the relationship between the dependent variable and each predictor, instead of estimating **the β coefficient**, we look at odds ratios. If the ORi is smaller or larger than 1, then we can reject the null hypothesis that there's no relationship between the predictor and the dependent variable. Estimating the OR can give us a more intuitive result in the test.

2.4 Assess the quality of model fit

In the logistic regression, we also calculate **the R-squared** value for each of the predictor, and same as the OLS model, the higher the R-squared, the better the model. However, the R-squared is no longer as important as in OLS regression, because it cannot be interpreted as the percent of variance explained by the model. Instead, statisticians have come up with a variety of analogues of R squared for multiple logistic regression that they refer to collectively as "pseudo-R squared". These do not have the same interpretation.[2]

When comparing different models with different predictors, one of the ways we can use is to compare the AIC values. **The Akaike Information Criterion (AIC)** is an estimator of prediction error and thereby relative quality of statistical models for a given set of data.[3] AIC estimates the relative amount of information lost by a given model: the lower the AIC, the better the model is.

When we are dealing with the residuals in logistic regression, it can still be defined as $\varepsilon = y_i - \hat{y}_i$, same with the OLS regression, but the interpretation can be different, here, $\hat{y}_i$ is a probability that Y=1 (in our report, the driver is drunk), it ranges from 0 to 1. It can be presented in the equation:

$$P(Y = 1) = \hat{y}_i = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_3 x_{3i}}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_3 x_{3i}}}$$

Our main goal is to predict precisely, to be more specific, we hope the model can predict a high value of Y=1 if the Y is actually 1 (in our report, the driver is actually drunk), and a low probability of Y=1 if the Y is actually 0(in our report, the driver is actually not drunk). In other words, we hope to keep the residuals as small as possible.

To define whether the probability of Y=1(drunk driving) is high or low, we should apply a "cut-off" rate to help us divide the predicted values into two groups. The "cut-off" value is usually chosen by looking at the distribution of values in the histogram. The values whose probabilities are above the threshold will be defined as positives and the values whose probabilities are below the threshold will be defined as negatives. If the cut-off rate is set to be too high, there might be less predicted positives than actual, and if it is set to be too low, there might be less predicted negatives than actual. It's important to find the optimal rate to improve the model's accuracy.

There's a more direct way for us to test the accuracy of the model, that is, to test the **Sensitivity (true positive rate)**, the **Specificity (true negative rate)**, and the misclassification rate of the model. The Sensitive measures the proportion of actual positives which are correctly predicted, the Specificity measures the proportion of negatives which are correctly predicted, and the misclassification rate measures the proportion of predicted values which are falsely identified as such. In our report, the sensitivity is the proportion of the drivers we correctly predicted as drunk. To achieve our goal of increasing the model's accuracy, we should improve the model's sensitivity and specificity, and try to decrease the misclassification rate. It can be done by finding the optimal threshold, as I stated in the previous paragraph.

We can use the ROC curves to examine the predictive ability of our model and find the best cut-off rate. The ROC curve is a plot shows the true positive rate(sensitivity) against the false positive rate(1-specificity).

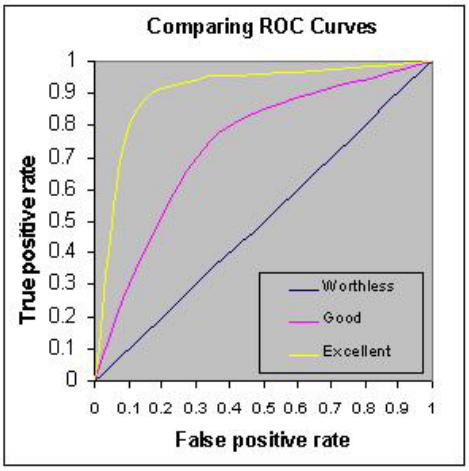

Figure2. 2 The ROC curve

(Source: http://gim.unmc.edu/dxtests/roc3.htm)

There's a 45-degree line in the middle of the plot, where the true positive rate equals to the false positive rate, and it's the worthless ROC. The ROC curve we will have generated should be above this worthless curve, and best one will have the highest true positive related to the lowest false positive rate, it will be close to the left and up boundaries of the plot. We can use the Youden Index or calculating the minimum distance from the upper left corner to identify the optimal cut-off. The Youden Index is a cut-off for which the specificity + sensitivity value is maximized, and similar with minimizing the curve's distance from the upper left corner, where the specificity=sensitivity=1, the closer the curve goes to the corner, the higher the specificity + sensitivity is. In this report, we will use the method of minimizing distance from upper left corner to find the optimal cut-off.

Also, we would like to calculate the Area Under ROC Curves (AUC) to measure the prediction accuracy, a better model has higher AUC value. AUC can be interpreted as the probability that the model correctly ranks two randomly selected observations where one has Y=1 and the other one has Y=0. The higher the AUC, the larger the value sensitivity + specificity is, which can be used to evaluate the cut-offs. The possible AUC values range between 0.5 to 1, 0.9-1 stands for an excellent model, 0.8-0.9 meaning the model is good and 0.7-0.8 represents a fair model. Generally, we can say that a model with an AUC of more than 0.7 is just fine. A model with an AUC which is between 0.6-0.7 is a poor one and if the AUC ranges between 0.6-0.5, the model fails.

2.5 The assumptions of logistic regression

Comparing the OLS regression and the logistic regression, both methods have the **assumptions** that there's no multicollinearity. However, in the logistic regression, the dependent variable must be binary, and there's no assumption that there needs to be linear relationship between the dependent variable and the independent variables, and the residuals don't need to be normally distributed. In OLS regression, we assume that there's homoscedasticity, but in logistic regression, we assume there's not.

2.6 Exploratory analysis

Before running logistic regression, most of the statisticians run cross-tabulations and use the Chi-Square to test whether there's association between the binary dependent variable and one categorical independent variable, said differently, whether the proportion of the results of one categorical variable varies with respect to another categorical variable. For example, in our report, whether the distribution of the drunk drivers varies with respect to the values of the fatalities for crash. The null hypothesis and the alternative hypothesis would be like:

- H_0: the proportion of fatalities for crashes that involve drunk drivers is the same as the proportion of fatalities for crashes that don't involve drunk drivers.

- H_a: the proportion of fatalities for crashes that involve drunk drivers is different than the proportion of fatalities for crashes that don't involve drunk drivers.

For the continuous variable, we can compare the mean values of the continuous predictor for the different levels of the dependent variable. We employ a test called the independent samples t-test to do it. For the variable PCTBACHMOR, the null hypothesis and the alternative hypothesis for the independent samples t-test would be like:

- H_0: the average values of the variable PCTBACHMOR are the same for crashes that involve drunk drivers and crashes that don't.

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

3) Results.

3.1 Exploratory results

Table3.1 Summary statistics

	DRINKING_D		
	0	1	Total
N	40,879	2,485	43,364
%	94.27%	5.73%	100.00%

Firstly, we summarized and looked at the distributions of the dependent variable. In this case, it is **DRINKING_D** (Drinking driver indicator 1 = Yes, 0 = No). From **table 3.1**, we can see that the portion of drunk drivers in our dataset is quite small, which means there might be a tendency for the model to predict better in the no-drinking driver cases.

Table3.2 Try cross-tabulation of the dependent variable and independent variables:

		No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		Total	χ2 p-value
		N	%	N	%	N	
FATAL_OR_M	Crash resulted in fatality or major injury	1,181	2.89%	188	7.57%	1,369	2.5222E-38
OVERTURNED	Crash involved an overturned vehicle	612	1.50%	110	4.43%	722	1.55176E-28
CELL_PHONE	Driver was using cell phone	426	1.04%	28	1.13%	454	0.6872569
SPEEDING	Crash involved speeding car	1,261	3.08%	260	10.46%	1,521	6.24956E-84
AGGRESSIVE	Crash involved aggressive driving	18,522	45.31%	916	36.86%	19,438	2.00079E-16
DRIVER1617	Crash involved at least one driver who was 16 or 17 years old	674	1.65%	12	0.48%	686	6.11562E-06
DRIVER65PLUS	Crash involved at least one driver who was at least 65 years old	4,237	10.36%	119	4.79%	4,356	2.75703E-19

The null hypothesis of the Chi-Square tests between all the independent variables (FATAL_OR_M, OVERTURNED, CELL_PHONE, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS) and the dependent variable (DRINKING_D) is that there is no relationship between the independent and the dependent variable. As is customary in the social sciences, we'll set our alpha level at 0.05. From **table 3.2,** we can see that only the p-value, **0.687**, of CELL-PHONE is higher than 0.05, which means we cannot reject the null hypothesis. Said differently, whether the driver was using a cell phone is not significantly associated with whether the driver is drinking alcohol.

Table3.3 Try a t-test between the continuous predictors and the dependent variable

		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	0.91
MEDHHINC	Median household income	31,483.05	16,930.10	31,998.75	17,810.50	0.16

The null hypothesis of the t-test between **the continuous predictors (PCTBACHMOR, MEDHHINC) and the dependent variable (DRINKING_D)** is that there is no relationship between the independent and the dependent variable. As is customary in the social sciences, we'll set our alpha level at 0.05. From table 3.3, we can see both the p-value of PCTBACHMOR (0.91), MEDHHINC (0.16) are higher than the alpha level. Thus, we cannot reject the null hypothesis of the t-test. Said differently, there is no significant association between the dependent variable and each of the continuous predictors.

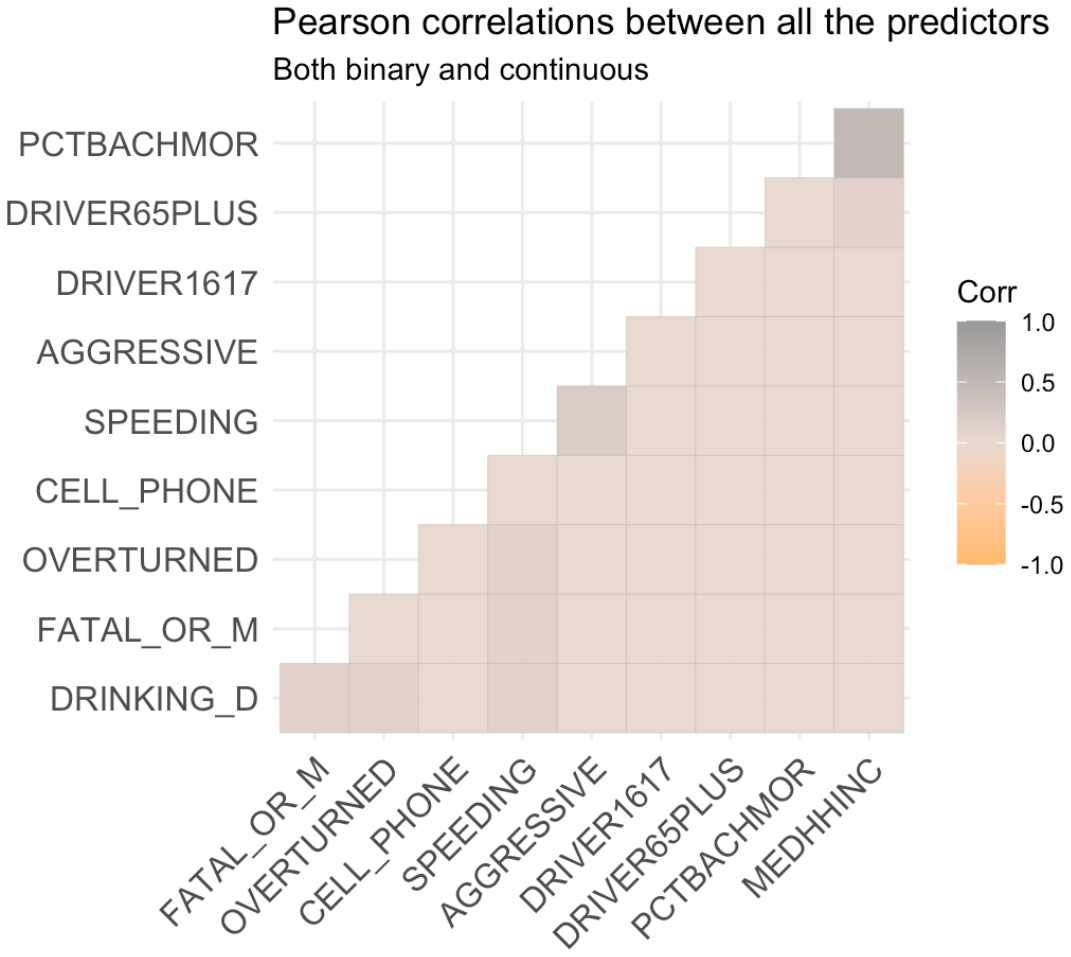

Figure3.1 Pearson correlation matrix between all the predictors

Table3.4 Pearson correlation between all the predictors

	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	-0.0146522648	-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	-0.0012458540	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	0.04344045	0.022877425	0.050337711	0.4778695368	1.000000000

Within the context of regression analysis, we deal with the assumption of no multicollinearity -- or, in other words, the assumption of no strong linear correlation between predictors. We use the pairwise **Pearson correlation matrix** to evaluate it this time, from the correlation matrix, we can notice that all the absolute values of r lie between 0.01 to 0.48 (<0.8), and there isn't severe multi-collinearity between the predictors, that is the predictors are not very strongly correlated with each other.

However, since we are using **Pearson correlations** to measure the associations between binary predictors and continuous predictors, it is important that all correlations computed are comparable. The correlation between the binary variables might not be linear. And the distributions of the binary predictors are likely to be skewed. Thus, using Pearson correlation here has a severe limitation.

3.2 Regression results

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

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

Coefficients:
                Estimate   Std. Error z value             Pr(>|z|)
(Intercept)  -2.732506616  0.045875659 -59.563 < 0.0000000000000002 ***
FATAL_OR_M    0.814013802  0.083806924   9.713 < 0.0000000000000002 ***
OVERTURNED    0.928921376  0.109166324   8.509 < 0.0000000000000002 ***
CELL_PHONE    0.029550085  0.197777821   0.149               0.8812
SPEEDING      1.538975665  0.080545894  19.107 < 0.0000000000000002 ***
AGGRESSIVE   -0.596915946  0.047779238 -12.493 < 0.0000000000000002 ***
DRIVER1617   -1.280295964  0.293147168  -4.367 0.000012572447127938 ***
DRIVER65PLUS -0.774664640  0.095858315  -8.081 0.000000000000000641 ***
PCTBACHMOR   -0.000370634  0.001296387  -0.286               0.7750
MEDHHINC      0.000002804  0.000001341   2.091               0.0365 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19036  on 43363  degrees of freedom
Residual deviance: 18340  on 43354  degrees of freedom
AIC: 18360

Number of Fisher Scoring iterations: 6
```

Figure3.2 The output of regression results of all predictors

Table3.5 The output of summary table

	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

As we can see, **FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS** are significant. **CELL_PHONE, PCTBACHMOR, MEDHHINC** are not significant. We can interpret each one of them as follows:

INTERCEPT:

The estimate is -2.732. In this case, the estimated coefficient for the intercept (-2.732) is the log odds of the driver being drunk where all predictors are 0.

Said differently, -2.732 is the log odds of the driver being drunk where

- **FATAL_OR_M** = 0 (i.e., Crash did not result in fatality or major injury), AND
- **OVERTURNED** = 0 (i.e., Crash did not involve an overturned vehicle), AND
- **SPEEDING** = 0 (i.e., Crash did not involve speeding car), AND
- **AGGRESSIVE** = 0 (i.e., Crash did not involve aggressive driving), AND
- **DRIVER1617** = 0 (i.e., Crash did not involve at least one driver who was 16 or 17 years old), AND
- **DRIVER65PLUS** = 0 (i.e., Crash did not involve at least one driver who was at least 65 years old), AND
- **CELL_PHONE** = 0 (i.e., Driver was not using cell phone), AND
- **PCTBACHMOR** = 0 (i.e., 0% 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** = 0

FATAL_OR_M:

The estimate is 0.8140. This means that for a 1 unit increase in **FATAL_OR_M**: (i.e., as we go from crash did not result in fatality or major injury to crash resulted in fatality or major injury), the log odds of the driver being drunk go up by 0.814, holding other predictors constant.

Said differently, a 1 unit increase in **FATAL_OR_M:** (i.e., as we go from crash did not result in fatality or major injury to crash resulted in a fatality or major injury), he odds of the driver was drunk change by $(e^{\beta_1} - 1) * 100\% = (2.257 - 1) * 100\% = 125.7\%$, holding other predictors constant.

OVERTURNED:

The estimate is 0.9289. This means that for a 1 unit increase in OVERTURNED (i.e., as we go from crash did not involve an overturned vehicle to involve an overturned vehicle), the log odds of there being a drunk driver go up by .9289, holding other predictors constant.

Said differently, for a 1 unit increase in **OVERTURNED** (i.e., as we go from crash did not involve an overturned vehicle to involve an overturned vehicle), the odds of a crash did not involve a drunk driver change by $(e^{\beta_1} - 1) * 100\% = (2.532 - 1) * 100\% = 153.2\%$, holding other predictors constant.

SPEEDING:

The estimate is 1.538. This means that for a 1 unit increase in **SPEEDING**, the log odds of there being a drunk driver increase by 1.538, holding other predictors constant.

Said differently, for a 1 unit increase in **SPEEDING**, the odds of there being a drunk driver change by $(e^{\beta_1} - 1) * 100\% = (4.65527 - 1) * 100\% = 365.5\%$ -- that is, they go up by 365.5%, holding other predictors constant.

AGGRESSIVE:

The estimate is -0.59691. This means that for a 1 unit increase in **AGGRESSIVE** (i.e., as we go from crash-involved aggressive driving to crash did not involve aggressive driving), the log odds of there being a drunk driver go down by 0.597, holding other predictors constant.

Said differently, for a 1 unit increase in **AGGRESSIVE** (i.e., as we go from crash-involved aggressive driving to crash did not involve aggressive driving), the odd of there being a drunk

driver change by $(e^{\beta_1} - 1) * 100\% = (0.5505 - 1) * 100\% = -44.94\%$, holding other predictors constant.

DRIVER1617:

The estimate is -1.2803. This means that for a 1 unit increase in **DRIVER1617** (i.e., as we go from a crash that did not involve at least one driver who was 16 or 17 years old to a crash did involve at least one driver who was 16 or 17 years old), the log odds of there being a drunk driver go down by 1.2803, holding other predictors constant.

Said differently, for a 1 unit increase in **DRIVER1617** (i.e., as we go from a crash that did not involve at least one driver who was 16 or 17 years old to a crash that did involve at least one driver who was 16 or 17 years old), the odds of there being a drunk driver changed by $(e^{\beta_1} - 1) * 100\% = (0.2779539 - 1) * 100\% = -72.20\%$, holding other predictors constant.

DRIVER65PLUS：

The estimate is -0.7747. This means that for a 1 unit increase in DRIVER65PLUS (i.e., as we go from a crash did not involve at least one driver who was at least 65 years old to a crash did involve at least one driver who was at least 65 years old), the log odds of there being a drunk driver go down by 0.7747, holding other predictors constant.

Said differently, for a 1 unit decrease in DRIVER65PLUS (i.e., as we go from a crash did not involve at least one driver who was at least 65 years old to a crash did involve at least one driver who was at least 65 years old), the odds of there being a drunk driver changed by $(e^{\beta_1} - 1) * 100\% = (0.460842 - 1) * 100\% = -53.92\%$, holding other predictors constant.

3.3 Model evaluation

From the **table 3.6**, we can see the misclassification rates are negatively correlated with the **cut-off values**. The lowest cut-off values yielded the highest misclassification rates. And the highest cut-off values yielded the lowest misclassification rates. This might relate to there being more false cases in our dataset (mentioned in part 3.1.1).

Table3.6 Specificity, sensitivity, and misclassification rates for the 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
0.10	0.164	0.948	0.097
0.15	0.104	0.972	0.078
0.20	0.023	0.995	0.060
0.50	0.002	1.000	0.057

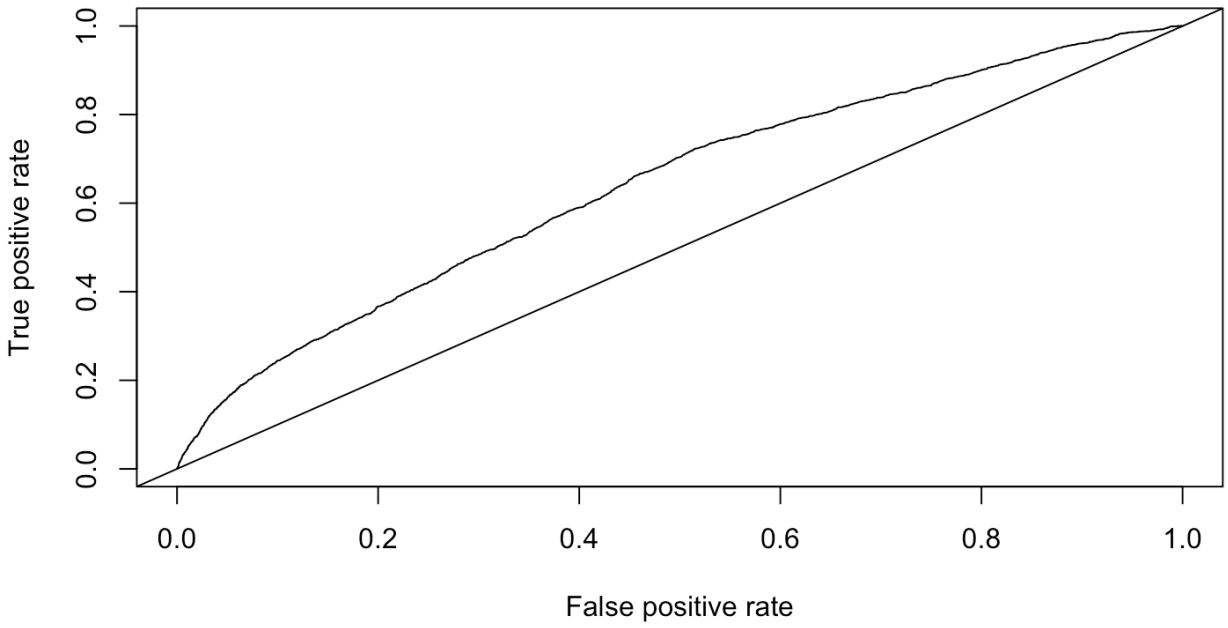

Figure3.3 The ROC curve selected by the optimal cut-off rate

The ROC curve is a way to plot the true positive rate (sensitivity) against the false positive rate (1 - specificity). A cut-off for which the ROC curve has the minimum distance from the upper left corner of the graph – i.e., the point at which specificity = 1 and sensitivity = 1. This is just a different way of maximizing specificity and sensitivity. We can get the optimal cut-off point and corresponding sensitivity and specificity.

The area under the ROC curve = 0.6398695 (AUC, which stands for Area Under Curve) is a measure of the prediction accuracy of the model (how well a model predicts 1 response as 1's and 0 responses as 0's). Higher AUCs means that we can find a cut-off value for which both sensitivity and specificity of the model are relatively high. In this case, AUC is between .60-.70. This means the model might be poor in prediction.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1962  -0.3692  -0.3153  -0.2765   3.0092

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.65149    0.02741 -96.749  < 2e-16 ***
FATAL_OR_M    0.80933    0.08376   9.662  < 2e-16 ***
OVERTURNED    0.93974    0.10904   8.619  < 2e-16 ***
SPEEDING      1.54033    0.08053  19.128  < 2e-16 ***
AGGRESSIVE   -0.59381    0.04774 -12.440  < 2e-16 ***
DRIVER1617   -1.27149    0.29311  -4.338 1.44e-05 ***
DRIVER65PLUS -0.76648    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 43357  degrees of freedom
AIC: 18358

Number of Fisher Scoring iterations: 6
```

Figure3.4 The results of the logistic regression with the binary predictors only

Table 3.7 The output of summary table

	Estimate	Std. Error	z value	Pr(>|z|)	OR	2.5 %	97.5 %
(Intercept)	-2.6514929	0.02740579	-96.749387	0.000000e+00	0.07054582	0.06683011	0.07440992
FATAL_OR_M	0.8093334	0.08376190	9.662310	4.359234e-22	2.24641011	1.90115697	2.64050252
OVERTURNED	0.9397382	0.10903525	8.618664	6.773965e-18	2.55931141	2.05726201	3.15555055
SPEEDING	1.5403331	0.08052818	19.127877	1.479787e-81	4.66614447	3.97966709	5.45742023
AGGRESSIVE	-0.5938127	0.04773588	-12.439547	1.594046e-35	0.55221782	0.50261639	0.60606091
DRIVER1617	-1.2714863	0.29310870	-4.337934	1.438281e-05	0.28041452	0.14906095	0.47521864
DRIVER65PLUS	-0.7664837	0.09576430	-8.003856	1.205820e-15	0.46464404	0.38317285	0.55791837

From the **table 3.7** above, we can see there are no predictors which are significant in the new model but weren't significant in the original one. Also, we can see that the **Akaike Information Criterion (AIC)** for both models is nearly the same (different less than 3), which means including the continuous variable in the regression does not affect the quality of the regression model in the end.

4) Discussion and Limitations

FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, and DRIVER65PLUS are strong predictors of crashes that involve drunk driving. CELL_PHONE, PCTBACHMOR, and MEDHHINC are not associated with the dependent variable. The results are not surprising. Because from the previous parts, the chi-square test result shows that CELL_PHONE, PCTBACHMOR, and MEDHHINC are not significantly related to the dependent variable.

In this case, using logistic regression might have a problem, because the total size of the dataset is not large enough and the % of cases with values of '1' for the dependent variable is too small, which means that the model might perform better in predicting '0' cases. Also, since we are using Pearson correlations in evaluating the correlations between binary predictors, there might be a risk that predictors have nonlinear relation with each other.

5) Reference

[1] Definition of Odds Ratio, Wikipedia: https://en.wikipedia.org/wiki/Odds_ratio

[2] Pseudo R squared values for multiple logistic regression, Principles of Regression, GraphPad: https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_mult_logistic_gof_pseudo_r_squared.htm

[3] Stoica, P.; Selen, Y. (2004), "Model-order selection: a review of information criterion rules", IEEE Signal Processing Magazine (July): 36–47, doi:10.1109/MSP.2004.1311138, S2CID 17338979