

# **Homework 3: The Application of Logistic Regression to Examine the Predictors of Car Crashes Caused by Alcohol**

Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-11-20

## **1 Introduction**

## **2 Methods**

### **2.1 Logistic Regression Motivation and Foundation**

The dependent variable in this report (DRINKING\_D) is a binary indicator: it takes the value 1 (True) or 0 (False). This presents a major limitation for Ordinary Least Squares (OLS) regression. OLS assumes a continuous dependent variable and estimates coefficients that can take any value from  $-\infty$  to  $+\infty$ , interpreting each coefficient as the expected change in the dependent variable for a one-unit increase in a predictor. When the outcome is binary, this interpretation breaks down. A predicted value of 0.65, for example, is not meaningful when the outcome can only be 0 or 1.

Because of these issues, logistic regression is more appropriate for modeling a binary outcome. Logistic regression works by transforming the probability of the event into a metric that can take any real value: the log-odds (also called the logit). To understand this transformation, it helps to introduce the concept of odds.

While probability is defined as  $\text{Pr}(\text{event}) = \frac{\#\text{desirable outcomes}}{\#\text{possible outcomes}}$ , the odds of an event are defined as  $\frac{\#\text{desirable outcomes}}{\#\text{undesirable outcomes}}$ . In the context of this report, the odds of drink-driving are  $\frac{\#\text{with drink driving}}{\#\text{without drink driving}}$ .

Logistic regression models the log of the odds, or the logit, as a linear function of the predictors. Because the log-odds range from  $-\infty$  to  $+\infty$ , the model avoids the limitations of OLS. Exponentiating a logistic regression coefficient produces an odds ratio (OR), which describes

how the odds of the outcome change for a one-unit increase in a predictor. An  $OR > 1$  indicates increased odds of the event, while an  $OR < 1$  indicates decreased odds.

To model the probability that a crash involved a drinking driver, we use a logistic regression model with DRINKING\_D as the dependent variable and a set of binary and continuous predictors. The binary predictors indicate whether specific conditions applied to the crash: whether the crash resulted in a fatality or major injury (FATAL\_OR\_M), whether the vehicle was overturned (OVERTURNED), whether the driver was using a cell phone at the time of the crash (CELL\_PHONE), whether the crash involved speeding (SPEEDING), whether aggressive driving was involved (AGGRESSIVE), whether at least one driver was 16 or 17 years old (DRIVER1617), or whether at least one driver was 65 or older (DRIVER65PLUS). In addition, the model includes continuous block-group-level predictors, specifically the percent of adults with at least a bachelor's degree (PCTBACHMOR) and the median household income (MEDHHINC) for the location where the crash occurred. For this report, the logit model expresses the log-odds of a drinking-driver crash as a linear function of the predictors. The model is:

$$\begin{aligned} \ln\left(\frac{p}{1-p}\right) = & \beta_0 + \beta_1(\text{FATAL OR M}) + \beta_2(\text{OVERTURNED}) + \beta_3(\text{CELL PHONE}) \\ & + \beta_4(\text{SPEEDING}) + \beta_5(\text{AGGRESSIVE}) + \beta_6(\text{DRIVER1617}) \\ & + \beta_7(\text{DRIVER65PLUS}) + \beta_8(\text{PCTBACHMOR}) + \beta_9(\text{MEDHHINC}). \end{aligned}$$

Here,  $p$  is the probability that DRINKING\_D = 1, meaning the crash involved a drinking driver. The term  $\ln\left(\frac{p}{1-p}\right)$  is the logit, or the natural log of the odds of a drinking-driver crash. Each  $\beta_k$  represents the change in the log-odds associated with a one-unit increase in the corresponding predictor, holding the others constant.

Binary predictors such as FATAL\_OR\_M or SPEEDING shift the log-odds by  $\beta_k$  when the indicator changes from 0 to 1. Continuous predictors such as PCTBACHMOR and MEDHHINC shift the log-odds proportionally to their values.

We can rewrite the model by solving for  $p = P(\text{DRINKING\_D} = 1)$  (note: to make the formula fit on the report, we used  $z$  to notate the  $\beta$  coefficients and predictors) :

$$p = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}$$

$$\begin{aligned} \text{where } z = & \beta_0 + \beta_1(\text{FATAL OR M}) + \beta_2(\text{OVERTURNED}) + \beta_3(\text{CELL PHONE}) \\ & + \beta_4(\text{SPEEDING}) + \beta_5(\text{AGGRESSIVE}) + \beta_6(\text{DRIVER1617}) \\ & + \beta_7(\text{DRIVER65PLUS}) + \beta_8(\text{PCTBACHMOR}) + \beta_9(\text{MEDHHINC}) \end{aligned}$$

This expression uses the logistic function, which transforms any real-valued input into a valid probability between 0 and 1. The denominator contains the exponential function  $\exp(\cdot)$ , which

ensures that  $p$  is always bounded between 0 and 1, regardless of the values of the predictors or coefficients.

## 2.2 Hypothesis Testing Overview

In logistic regression, each predictor  $x_i$  is tested for the null hypothesis,  $H_0$ , that the beta coefficient,  $\beta_i$ , is 0 against the alternative hypothesis  $H_a$  that  $\beta_i$  is not 0:

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

The z-value, also known as the Wald statistic in logistic regression, is the test statistic that we calculate under the null hypothesis. We calculate this statistic by dividing the estimated beta coefficient,  $\hat{\beta}_i$ , by its standard error or  $\sigma_{\hat{\beta}_i}$ :

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

Under the null hypothesis, the Wald statistic follows an approximately standard normal distribution,  $N(0,1)$ . This property allows us to compute the two-tailed p-value as the probability of observing a statistic as extreme, or more extreme, than the calculated statistic if the null hypothesis were true. If the p-value is  $< 0.05$ , we can reject the null hypothesis in favor of the alternative hypothesis that  $\beta_i$  is not 0. Rather than interpreting the raw beta coefficients, statisticians prefer use the odds ratio,  $OR_i$ , which can be calculated by exponentiating  $\hat{\beta}_i$ :

$$OR_i = e^{\hat{\beta}_i}$$

The odds ratio expresses the effect of a predictor on the dependent variable in multiplicative terms. Specifically, it represents how the odds of the event change for a one-unit increase in the predictor, holding other variables constant. The null and alternative hypothesis can be adapted for the odds ratio, where the null hypothesis is the predictor has no effect on the odds ( $OR = 1$ ) and the alternative hypothesis is that the predictor increases or decrease the odds of the event ( $OR \neq 1$ ):

$$H_0 : OR = 1$$

$$H_a : OR \neq 1$$

Conceptually, the odds ratio is the ratio of the odds with the predictor present to the odds with the predictor absent. Thus, if the odds ratio equals 1, it indicates that the odds are the same: the predictor did not change the odds of the outcome. Alternatively, if the odds ratio is significantly above or below 1, the predictor increased or decreased the odds. The confidence intervals for the odds ratios can be calculated by exponentiating the coefficient confidence

intervals. These intervals provide a range of plausible values for the true odds ratio, reflecting the uncertainty of the estimate. In the context of logistic regression, the presence of a 1 in the confidence interval indicates the predictor's effect is not statistically significant while a confidence interval entirely above or below 1, indicates that the predictor increased or decreased the odds.

All coefficient estimates, z-values, and p-values were extracted in R from the fitted logistic regression model's summary. Odds ratios and their confidence intervals were calculated by exponentiating the original coefficient estimates and confidence intervals, then merged with the extracted coefficients for interpretation.

### 2.3 Assessing Model Quality of Fit

In our analysis, goodness of the model's fit was evaluated in various ways. In Ordinary Least Squares (OLS) regression,  $R^2$  is used to evaluate model fit as it is a statistic that returns the proportion of total variance in the dependent variable explained by the independent variable. Unlike in OLS regression, logistic regression doesn't model a continuous outcome. In logistic regression the dependent variable,  $Y$  is binary, taking a value of 1 to indicate the occurrence of an event or 0 to indicate its absence. Therefore, since there is no longer a meaningful attribution of unexplained and explained variance in the dependent variable,  $R^2$  can no longer be interpreted as the percent of variance explained by the model. Similarly to linear regression, residuals,  $\varepsilon_i$ , are calculated as the difference between the observed values of the dependent variable,  $y_i$ , and the predicted values of the dependent variable,  $\hat{y}_i$ :

$$\varepsilon_i = y_i - \hat{y}_i$$

In logistic regression, however, the predicted values,  $\hat{y}_i$ , represent the probability that  $Y = 1$ , while  $y_i$  represent the binary outcome ( $Y = 1$  or  $Y = 0$ ). Thus, residuals represent the difference between the observed binary outcome and the model's predicted probabilities. Theoretically a model of good fit predicts high probabilities of  $Y = 1$  if  $y_i$  actually equals 1 and a low probability of  $Y = 1$  if  $y_i$  is actually 0. In order to determine what is considered high probability and low probability, a cut-off value is imposed on the  $\hat{y}_i$  values. Cut-off values are then evaluated based on their specificity, sensitivity, and misclassification rates. Sensitivity, also called the true positive rate, is the proportion of actual positives that are correctly identified:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

In this analysis, the sensitivity rate is the proportion of observed  $y_i = 1$  values correctly predicted as 1. Specificity, also called the true negative rate, is the proportion of actual negatives that are correctly identified as negatives:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

The specificity rate in this analysis is the proportion of observed  $y_i = 0$  values correctly predicted as 0. The misclassification rate is the proportion of incorrectly identified positive and negative  $y_i$  values based on the total number of predictions:

$$\text{Misclassification} = \frac{\text{False Negatives} + \text{False Positives}}{\text{True Negatives} + \text{True Positives} + \text{False Positives} + \text{False Negatives}}$$

In R, we called upon `fit.binary` and set the fit parameter to various different values to simulate how various cut off values would impact the sensitivity, specification, and misclassification rate. In other words, we use multiple cut-off values to compare the trade-offs of each cut-off threshold. Ideally, the chosen threshold will achieve higher sensitivity and specificity while minimizing the misclassification rate.

Receiver Operating Characteristics (ROC) curves are another tool for evaluating cut-off values. The ROC curve plots sensitivity against the false positive rate ( $1 - \text{specificity}$ ) across all possible cut-off values of  $\hat{y}_i$ . The baseline for evaluating ROC curves called the “worthless” ROC is a 45 degree line where sensitivity and the false positive rate are equal across all cut-off values, meaning the predictions are no better than a random guess. Effective models produce ROC curves that lie above this diagonal baseline. ROC curves can be used to determine the cut-off value that balances the sensitivity and specificity rate, characteristics that indicate a good model. One common way to determine the optimal cut-off value is to use the Youden Index, which identifies the cut-off that maximizes the sum of sensitivity and specificity is maximized:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

This corresponds to the point on the ROC curve farthest above the diagonal line, or equivalently, the point closest to the top-left corner of the graph where sensitivity and specificity both equal 1. To identify the optimal cut-off value, we implemented a function in R that is conceptually similar to the Youden Index as it attempts to find the point that minimizes the distance to this ideal point.

In addition to identifying an optimal cut-off, we can also calculate Area Under Curve (AUC) for our ROC curve as a measure of the model’s overall predictive accuracy. The AUC quantifies the model’s ability to discriminate between positive and negative outcomes across all possible cut-offs. An AUC of 1 (area of the entire graph) indicates perfect classification or discrimination while a value of 0.5 (area under the 45 degree line) indicates no better than random guessing. AUC can be interpreted as the probability that the model assigns a higher predicted probability to a randomly chosen positive case than to a randomly chosen negative case. Higher AUC values therefore reflect stronger overall discriminative ability across all possible cut-off values, implying that at least one threshold exists where both sensitivity and specificity are relatively high. In this analysis, the AUC was computed in R using the `performance` function from the `ROCR` package. We relied on commonly established thresholds for evaluating model accuracy based on AUC values where 0.90–1.00 indicates excellent accuracy, 0.80–0.90 good, 0.70–0.80 fair, 0.60–0.70 poor, and 0.50–0.60 indicates the model failed.

Another measure used to evaluate logistic regression model fit is the Akaike Information Criterion (AIC). Although the absolute value of the AIC is not interpretable on its own, it provides a basis for comparing two or more models. Specifically, AIC combines the log-likelihood of the predicted probabilities with a penalty for the number of estimated parameters. Lower AIC values indicate a more favorable balance between model complexity and goodness of fit.

## 2.4 Assumptions of Logistic Regression

Logistic regression models the relationship between a set of predictors and a binary dependent variable by expressing the log-odds of the outcome as a linear function of the predictors:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

The predicted probability of the outcome is obtained by applying the logistic transformation:

$$p = \frac{e^\eta}{1 + e^\eta}, \quad \eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

Here,  $p$  represents the predicted probability of the event occurring,  $X_k$  are the predictor variables, and  $\beta_k$  are the corresponding regression coefficients. This formulation ensures that predicted probabilities remain between 0 and 1, while allowing the model to use a linear combination of predictors on the log-odds scale.

### 2.4.1 Logistic regression assumptions

Logistic regression shares several assumptions with Ordinary Least Squares (OLS) regression, while relaxing others. As in OLS, logistic regression assumes that observations are independent of one another. Independence ensures that the estimated coefficients and their standard errors are valid. The model also assumes that the predictors are not perfectly collinear. Severe multicollinearity inflates standard errors and reduces the reliability of coefficient estimates.

However, several OLS assumptions do not apply to logistic regression. Logistic regression does not assume homoscedasticity of residuals, because the variance of a binary dependent variable is a function of its mean. The model also does not require that residuals follow a normal distribution. In addition, the model does not assume a linear relationship between the predictors and the outcome on the original probability scale. Instead, it assumes linearity only in the log-odds, as shown in the equations above.

## 2.5 Exploratory Analyses Prior to Logistic Regression

Before fitting a logistic regression model, statisticians often conduct preliminary analysis to assess the potential associations between the dependent variable and the predictors.

To assess whether multicollinearity was a concern, we calculated Pearson correlation coefficients between each predictor and the dependent variable. The correlations were all small in magnitude, indicating no severe multicollinearity, as shown below.

```
[1] 0.0005334492
```

```
[1] 0.007058232
```

```
[1] 0.06216164
```

```
[1] 0.05321245
```

```
[1] 0.00193325
```

```
[1] 0.09321369
```

```
[1] -0.03948341
```

```
[1] -0.02171699
```

```
[1] -0.04311372
```

### 2.5.1 Cross-tabulations for Binary Predictors

When both the dependent variable and a predictor are categorical, a cross-tabulation provides a simple way to examine the distribution of outcomes across different categories of the predictor. To formally test whether the distribution of the dependent variable varies across levels of a binary predictor, the appropriate statistical method is the Chi-Square ( $\chi^2$ ) test of independence.

For the  $\chi^2$  test, the null hypothesis states that the two categorical variables are independent; that is, the proportion of positive and negative outcomes is the same for both levels of the predictor. The alternative hypothesis states that the variables are not independent, meaning that the distribution of the dependent variable differs across categories of the predictor. A large  $\chi^2$  statistic and a p-value below the conventional significance threshold (e.g., 0.05) provide

evidence against the null hypothesis and suggest that an association exists between the two categorical variables.

We then examined the association between the dependent variable and each binary predictor using cross-tabulations and Chi-Square tests. The results are shown below.

Cell Contents

	N	
	N / Col Total	

Total Observations in Table: 43364

	mydata\$FATAL_OR_M		
mydata\$DRINKING_D	0	1	Row Total
0	39698	1181	40879
	0.945	0.863	
1	2297	188	2485
	0.055	0.137	
Column Total	41995	1369	43364
	0.968	0.032	

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 167.5615 d.f. = 1 p = 2.522202e-38

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 166.0354 d.f. = 1 p = 5.434077e-38

Cell Contents

	N	
Chi-square contribution		
N / Col Total		

Total Observations in Table: 43364

		mydata\$OVERTURNED		Row Total
		0	1	
mydata\$DRINKING_D	0	40267	612	40879
		0.117	6.919	
		0.944	0.848	
mydata\$DRINKING_D	1	2375	110	2485
		1.927	113.824	
		0.056	0.152	
Column Total		42642	722	43364
		0.983	0.017	

Cell Contents

	N	
	N / Col Total	

Total Observations in Table: 43364

		mydata\$CELL_PHONE		Row Total
		0	1	
mydata\$DRINKING_D	0	40453	426	40879
		0.943	0.938	
	1	2457	28	2485
		0.057	0.062	
Column Total		42910	454	43364
		0.990	0.010	

#### Statistics for All Table Factors

##### Pearson's Chi-squared test

---

Chi^2 = 0.162071      d.f. = 1      p = 0.6872569

##### Pearson's Chi-squared test with Yates' continuity correction

---

Chi^2 = 0.09065262      d.f. = 1      p = 0.7633491

#### Cell Contents

-----		
	N	
	N / Col Total	
-----		

Total Observations in Table: 43364

| mydata\$SPEEDING

mydata\$DRINKING_D	0	1	Row Total	
0	39618	1261	40879	
	0.947	0.829		
1	2225	260	2485	
	0.053	0.171		
Column Total	41843	1521	43364	
	0.965	0.035		

Statistics for All Table Factors

Pearson's Chi-squared test

---

Chi^2 = 376.7808 d.f. = 1 p = 6.249562e-84

Pearson's Chi-squared test with Yates' continuity correction

---

Chi^2 = 374.6039 d.f. = 1 p = 1.861184e-83

Cell Contents

	N
	N / Col Total

Total Observations in Table: 43364

	mydata\$AGGRESSIVE	0	1	Row Total	
mydata\$DRINKING_D					
0	22357	18522	40879		

	0.934	0.953	
1	1569	916	2485
	0.066	0.047	
Column Total	23926	19438	43364
	0.552	0.448	

#### Statistics for All Table Factors

##### Pearson's Chi-squared test

Chi^2 = 67.60186 d.f. = 1 p = 2.00079e-16

##### Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 67.2607 d.f. = 1 p = 2.378758e-16

#### Cell Contents

	N	
	N / Col Total	

Total Observations in Table: 43364

	mydata\$DRIVER1617		Row Total
mydata\$DRINKING_D	0	1	
0	40205	674	40879
	0.942	0.983	
1	2473	12	2485

	0.058	0.017	
-----	-----	-----	-----
Column Total	42678	686	43364
	0.984	0.016	
-----	-----	-----	-----

Statistics for All Table Factors

Pearson's Chi-squared test

Chi^2 = 20.45167 d.f. = 1 p = 6.115619e-06

Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 19.7097 d.f. = 1 p = 9.014275e-06

Cell Contents

	N
	N / Col Total

Total Observations in Table: 43364

	mydata\$DRIVER65PLUS		Row Total
mydata\$DRINKING_D	0	1	
0	36642	4237	40879
	0.939	0.973	
1	2366	119	2485
	0.061	0.027	
Column Total	39008	4356	43364

	0.900	0.100	
----- ----- ----- -----			

#### Statistics for All Table Factors

##### Pearson's Chi-squared test

Chi^2 = 80.6047 d.f. = 1 p = 2.75703e-19

##### Pearson's Chi-squared test with Yates' continuity correction

Chi^2 = 79.9888 d.f. = 1 p = 3.765375e-19

### 2.5.2 Comparing Means of Continuous Predictors

For continuous predictors, it is often useful to compare their mean values across the two categories of the binary dependent variable. The appropriate statistical test for comparing the means of a continuous variable between two independent groups is the independent samples t-test.

For the t-test, the null hypothesis states that the mean value of the continuous predictor is the same across both groups of the dependent variable. The alternative hypothesis states that the means differ between the two groups. A large absolute value of the t-statistic and a p-value below the specified significance level (e.g., 0.05) provide evidence to reject the null hypothesis and conclude that there are significant differences in mean values between the groups.

Finally, we compared the mean values of the continuous predictors across the two categories of the dependent variable using independent samples t-tests. The t-test results for PCTBACHMOR and MEDHHINC are shown below.

0	1
31483.05	31998.75

0	1
16930.1	17810.5

##### Welch Two Sample t-test

```

data: mydata$PCTBACHMOR by mydata$DRINKING_D
t = -0.10842, df = 2777.5, p-value = 0.9137
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
-0.7991398  0.7153982
sample estimates:
mean in group 0 mean in group 1
16.56986      16.61173

```

#### Welch Two Sample t-test

```

data: mydata$MEDHHINC by mydata$DRINKING_D
t = -1.4053, df = 2763.9, p-value = 0.16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
-1235.2508  203.8544
sample estimates:
mean in group 0 mean in group 1
31483.05      31998.75

```

## 3 Results

### 3.1 Exploratory Analysis

Before beginning the logistic regression, we must do some exploratory analysis of the data and check assumptions of the regression model. Below is a summary table of the dependent variable, displaying both the count and proportion of crashes that involved a drinking driver versus those that did not.

Table 1: Distribution of DRINKING\_D (Drunk Driving Indicator)

DRINKING_D	Count	Proportion
0	40879	0.943
1	2485	0.057

The distribution of the dependent variable shows that the vast majority of crashes did not involve a drinking driver: 40,879 crashes (94.3%). Only 2,485 crashes, or about 5.7%, involved a drinking driver.

It is also useful to examine the relationships between the dependent variable, DRINKING\_D, and each of the binary predictors. Table 1 presents the cross-tabulations of DRINKING\_D with each predictor, along with the proportion of crashes in each category. For each predictor, the table also includes the Chi-Square p-value to indicate whether the distribution of drinking-driver crashes differs significantly across its categories.

Table 2: Cross-Tabulation of DRINKING D with Binary Predictors

Predictor	No Alcohol Involved (DRINKING D = 0)		Alcohol Involved (DRINKING D = 1)		Total	Chi-square p-value
	Num	Pct.	Num	Pct.		
FATAL_OR_M	1181	2.89%	188	7.57%	1369	<0.001
OVERTURNED	612	1.50%	110	4.43%	722	<0.001
CELL_PHONE	426	1.04%	28	1.13%	454	0.763
SPEEDING	1261	3.08%	260	10.46%	1521	<0.001
AGGRESSIVE	18522	45.31%	916	36.86%	19438	<0.001
DRIVER1617	674	1.65%	12	0.48%	686	<0.001
DRIVER65PLUS	4237	10.36%	119	4.79%	4356	<0.001

The Chi-Square tests indicate whether there is a significant association between DRINKING\_D and each binary predictor. For most predictors (FATAL\_OR\_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, and DRIVER65PLUS) the p-values are less than 0.001, which is far below our significance threshold of 0.05. This allows us to reject the null hypothesis of independence for these variables, suggesting that the occurrence of a drinking-driver crash is significantly associated with these factors.

In contrast, the p-value for CELL\_PHONE is 0.763, well above 0.05, indicating that we fail to reject the null hypothesis. There is no statistically significant association between drinking-driver crashes and whether the driver was using a cell phone at the time of the crash.

Overall, these results suggest that most of the binary predictors are significantly related to the likelihood of a crash involving a drinking driver, except for CELL\_PHONE.

To further explore factors associated with drinking-driver crashes, we next examine the continuous predictors, PCTBACHMOR and MEDHHINC, comparing their means and standard deviation across crashes with and without alcohol involvement and conducting independent samples t-tests,

Table 3: Summary of Continuous Predictors by DRINKING D

No Alcohol Involved (DRINKING D = 0)	Alcohol Involved (DRINKING D = 1)
---	--------------------------------------

Predictor	Mean	SD	Mean	SD	t-test	p-value
PCTBACHMOR	16.56986	18.21426	16.61173	18.72091		0.914
MEDHHINC	31483.05472	16930.10159	31998.75292	17810.49735		0.160

The summary statistics for the continuous predictors show that the mean percentage of individuals with a bachelor's degree or higher (PCTBACHMOR) is very similar between crashes with no alcohol involvement (16.57%) and those with alcohol involvement (16.61%). The independent samples t-test yields a p-value of 0.914, which is far above our significance threshold of 0.05. This indicates that we fail to reject the null hypothesis, suggesting no significant difference in PCTBACHMOR between the two groups.

Similarly, the mean median household income (MEDHHINC) is slightly higher for alcohol-involved crashes (\$31,998) compared to non-alcohol-involved crashes (\$31,483), but the t-test p-value of 0.160 indicates that this difference is not statistically significant. Again, we fail to reject the null hypothesis, implying that MEDHHINC is not significantly associated with the likelihood of a crash involving a drinking driver.

Overall, the t-test results suggest that neither of the continuous predictors shows a significant association with DRINKING\_D in this dataset.

### 3.2 Logistic Regression Assumption Checks

As previously mentioned, a key assumption of logistic regression is that there is no severe multicollinearity between predictors. We attempted to test whether our data violated this assumption by creating a pairwise Pearson coefficient matrix that included all our predictors.

Table 4: Pearson correlation matrix among predictors

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

Table (#) presents the Pearson correlation coefficients between all binary and continuous predictors. Pearson coefficients,  $r$ , range from 1 to -1 and can be interpreted as 1 indicating strong positive linear correlation, -1 indicating strong negative linear correlation, and 0 indicating no linear correlation. Because correlation coefficients are rarely perfectly negative or positive, the threshold considered to indicate moderate correlation is an absolute value of  $r$  ( $|r|$ ) between 0.5 and 0.8 while  $|r| < 0.5$  indicates weak correlation and  $|r| > 0.8$  indicates strong correlation.

The coefficients that include binary variables are all uniformly small, near-zero values. Because the Pearson correlation coefficient is designed to measure the strength of a linear relationship between continuous variables, it is not an ideal measure of association when applied to binary predictors. As a result, Pearson correlation is not as accurate for assessing relationships between binary predictors (or between binary and continuous predictors) and may misrepresent the true association. This means that in our analysis, while the Pearson coefficients are uniformly small, we must be cautious or explore an alternative method in interpreting them.

The correlation coefficient between the two continuous variables, PCTBACHMOR and MEDHHINC, is, as expected, the highest as  $r=0.478$ . Multicollinearity is considered to occur when two or more predictors are very strongly correlated ( $|r| > 0.9$ ). The correlation between PCTBACHMOR and MEDHHINC is far below this threshold, suggesting no severe multicollinearity.

### 3.3 Logistic Regression Results

#### 3.3.1 Logistic Regression with All Predictors

To identify the predictors associated with alcohol-related crashes, we estimated a logistic regression model that included all binary and continuous predictors. The estimated coefficients, p-values, and odds ratios are presented as shown below.

Waiting for profiling to be done...

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

```

DRIVER1617  0.14774429  0.47109277
DRIVER65PLUS 0.37998364  0.55347851
PCTBACHMOR   0.99707035  1.00215087
MEDHHINC     1.00000013  1.00000539

```

The logistic regression model includes all binary and continuous predictors. This model helps identify which factors are associated with crashes that involve alcohol. Table X presents the estimated coefficients, standard errors, p-values, odds ratios, and 95% confidence intervals for each predictor.

Several predictors are statistically significant. Crashes that resulted in a fatality or major injury show higher odds of involving alcohol, with an odds ratio of approximately 2.26. Crashes involving overturned vehicles are also more likely to involve alcohol, with an odds ratio of about 2.53. Speeding has the strongest association among all predictors, with an odds ratio of approximately 4.66, indicating substantially higher odds of alcohol involvement. Aggressive driving is negatively associated with alcohol involvement, with an odds ratio of 0.55. Both age-related indicators are significant: crashes involving 16–17-year-old drivers or drivers aged 65 or older have lower odds of involving alcohol, with odds ratios of 0.28 and 0.46, respectively.

Two predictors are not statistically significant in this model. Cell phone use does not show a meaningful association with alcohol involvement. The percentage of residents with at least a bachelor's degree (PCTBACHMOR) is also non-significant, with an odds ratio close to 1. Median household income is statistically significant but has an odds ratio of 1.0000028, indicating a negligible substantive effect.

Overall, the model indicates that specific crash characteristics (fatality, overturning, and speeding) and driver demographics (teenage or senior drivers) are important predictors of alcohol involvement, while cell phone use and neighborhood-level sociodemographic variables contribute little additional explanatory power.

### **3.3.2 Sensitivity, Specificity, Misclassification**

To evaluate model performance across different probability thresholds, we computed sensitivity, specificity, and the overall misclassification rate for each cut-off value. The results are summarized as shown below.

Cell Contents	
-----	
	N
	N / Col Total
-----	

Total Observations in Table: 43364

		mydata\$DRINKING_D		Row Total
fit.binary	0	1		
FALSE	2374	41	2415	
	0.058	0.016		
TRUE	38505	2444	40949	
	0.942	0.984		
Column Total	40879	2485	43364	
	0.943	0.057		

Cell Contents

-----
N
N  /  Col  Total
-----

Total Observations in Table: 43364

		mydata\$DRINKING_D		Row Total
fit.binary	0	1		
FALSE	2613	48	2661	
	0.064	0.019		
TRUE	38266	2437	40703	
	0.936	0.981		
Column Total	40879	2485	43364	

	0.943	0.057	
-----	-----	-----	-----

Cell Contents			
	N		
	N / Col Total		

Total Observations in Table: 43364

	mydata\$DRINKING_D		
fit.binary	0	1	Row Total
FALSE	19176	659	19835
	0.469	0.265	
TRUE	21703	1826	23529
	0.531	0.735	
Column Total	40879	2485	43364
	0.943	0.057	

Cell Contents			
	N		
	N / Col Total		

Total Observations in Table: 43364

	mydata\$DRINKING_D		
fit.binary	0	1	Row Total
FALSE	37356	1935	39291
	0.914	0.779	
TRUE	3523	550	4073
	0.086	0.221	
Column Total	40879	2485	43364
	0.943	0.057	

Cell Contents
-----
N
N / Col Total
-----

Total Observations in Table: 43364

	mydata\$DRINKING_D		
fit.binary	0	1	Row Total
FALSE	38370	2026	40396
	0.939	0.815	
TRUE	2509	459	2968
	0.061	0.185	
Column Total	40879	2485	43364
	0.943	0.057	

Cell Contents

		N	
		N / Col Total	

Total Observations in Table: 43364

	mydata\$DRINKING_D		
fit.binary	0	1	Row Total
FALSE	38670	2067	40737
	0.946	0.832	
TRUE	2209	418	2627
	0.054	0.168	
Column Total	40879	2485	43364
	0.943	0.057	

Cell Contents

		N	
		N / Col Total	

Total Observations in Table: 43364

	mydata\$DRINKING_D		
fit.binary	0	1	Row Total

FALSE	38762	2077	40839	
	0.948	0.836		
-----	-----	-----	-----	-----
TRUE	2117	408	2525	
	0.052	0.164		
-----	-----	-----	-----	-----
Column Total	40879	2485	43364	
	0.943	0.057		
-----	-----	-----	-----	-----

Cell Contents

-----				
	N			
	N / Col Total			
-----				-----

Total Observations in Table: 43364

	mydata\$DRINKING_D		
fit.binary	0	1	Row Total
FALSE	39743	2226	41969
	0.972	0.896	
-----	-----	-----	-----
TRUE	1136	259	1395
	0.028	0.104	
-----	-----	-----	-----
Column Total	40879	2485	43364
	0.943	0.057	
-----	-----	-----	-----

Cell Contents

	N	
	N / Col Total	

Total Observations in Table: 43364

	mydata\$DRINKING_D		Row Total
fit.binary	0	1	
FALSE	40690	2428	43118
	0.995	0.977	
TRUE	189	57	246
	0.005	0.023	
Column Total	40879	2485	43364
	0.943	0.057	

	N	
	N / Col Total	

Total Observations in Table: 43364

	mydata\$DRINKING_D		Row Total
fit.binary	0	1	
FALSE	40875	2481	43356
	1.000	0.998	

TRUE	4	4	8
	0.000	0.002	
Column Total	40879	2485	43364
	0.943	0.057	

To further evaluate model performance, predicted probabilities were converted to binary classifications using a series of probability cut-offs ranging from 0.02 to 0.50. For each cut-off, a confusion matrix was generated, allowing the calculation of sensitivity, specificity, and the overall misclassification rate.

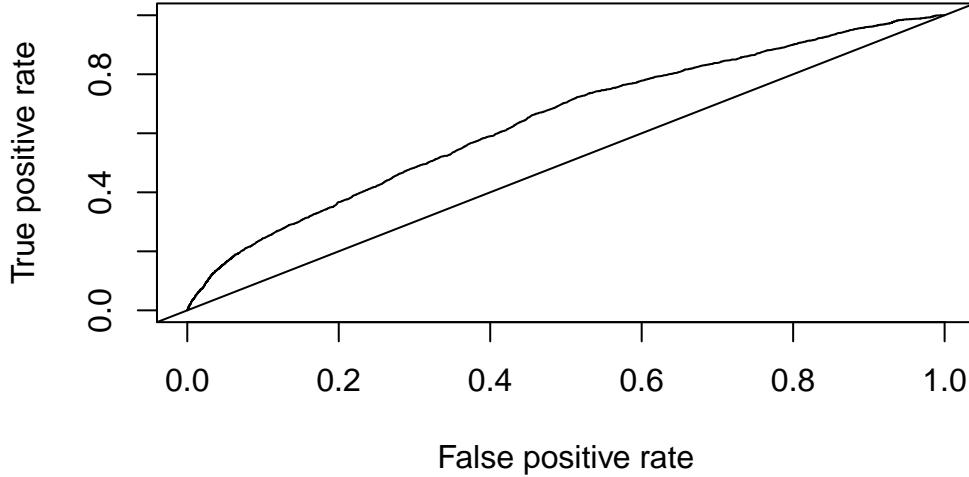
Sensitivity measures the proportion of alcohol-related crashes correctly identified by the model, while specificity measures the proportion of non-alcohol-related crashes correctly classified. The misclassification rate reflects the overall proportion of incorrect predictions. As expected, lower cut-off values result in higher sensitivity and lower specificity, while higher cut-offs reverse this pattern.

Across the tested cut-offs, the lowest misclassification rate occurred at a cut-off of 0.05, indicating that this threshold achieves the best balance between false positives and false negatives for this model. In contrast, extremely low or high cut-offs, such as 0.02 or 0.50, produce substantially higher misclassification rates and therefore perform less effectively as classification rules.

These results illustrate the importance of assessing multiple probability thresholds when evaluating logistic regression models. The optimal choice of cut-off depends on the tradeoff between sensitivity and specificity that is most appropriate for the application.

### 3.3.3 ROC curve & optimal cutoff

To evaluate the model's ability to distinguish between alcohol-related and non-alcohol-related crashes, we generated a receiver operating characteristic (ROC) curve using the predicted probabilities from the full logistic regression model. The ROC curve shows the tradeoff between sensitivity and specificity across all possible probability thresholds, as shown below.



```
[,1]
sensitivity 0.66076459
specificity 0.54524328
cutoff      0.06365151
```

Using the distance-to-(0,1) criterion, we identified the probability cut-off that minimizes the distance to the upper-left corner of the ROC space. The optimal cut-off derived from this approach was approximately 0.06365. This value can be compared with the cut-off of 0.05 identified earlier as the point that yielded the lowest misclassification rate. The difference between these two thresholds reflects the fact that the ROC-based method jointly considers sensitivity and specificity, while the misclassification-based approach evaluates only the proportion of incorrect predictions. Because the two criteria optimize different aspects of model performance, they do not necessarily produce the same probability cut-off.

### 3.3.4 Area Under the Curve (AUC)

The area under the ROC curve (AUC) provides a summary measure of the model's overall discriminative ability. The AUC for this model was 0.6399, as shown below, indicating modest ability to distinguish between alcohol-related and non-alcohol-related crashes. An AUC value of 0.5 suggests no discriminatory power, while values above 0.7 are typically considered acceptable. Thus, while the model performs better than random chance, its ability to accurately classify crashes based on alcohol involvement is limited.

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

### 3.3.5 Reduced model

To assess whether the continuous predictors contributed meaningfully to model performance, we estimated a reduced logistic regression model that included only the binary predictors. The estimated coefficients and odds ratios for this reduced model are presented as shown below.

```
Call:  
glm(formula = DRINKING_D ~ FATAL_OR_M + OVERTURNED + CELL_PHONE +  
    SPEEDING + AGGRESSIVE + DRIVER1617 + DRIVER65PLUS, family = "binomial",  
    data = mydata)  
  
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  
  
Waiting for profiling to be done...  
  
            Estimate Std. Error      z value      Pr(>|z|)      OR  
(Intercept) -2.65189961 0.02753107 -96.3238683 0.0000000e+00 0.07051713  
FATAL_OR_M    0.80931557 0.08376150   9.6621431 4.366327e-22 2.24636998
```

OVERTURNED	0.93978420	0.10903433	8.6191585	6.744795e-18	2.55942903
CELL_PHONE	0.03107367	0.19777088	0.1571195	8.751506e-01	1.03156149
SPEEDING	1.54032033	0.08052787	19.1277908	1.482240e-81	4.66608472
AGGRESSIVE	-0.59364687	0.04774781	-12.4329656	1.730916e-35	0.55230941
DRIVER1617	-1.27157607	0.29310969	-4.3382260	1.436374e-05	0.28038936
DRIVER65PLUS	-0.76645727	0.09576440	-8.0035718	1.208612e-15	0.46465631
	2.5 %	97.5 %			
(Intercept)	0.06678642	0.0743978			
FATAL_OR_M	1.90112455	2.6404533			
OVERTURNED	2.05736015	3.1556897			
CELL_PHONE	0.68459779	1.4907150			
SPEEDING	3.97961862	5.4573472			
AGGRESSIVE	0.50268818	0.6061758			
DRIVER1617	0.14904734	0.4751771			
DRIVER65PLUS	0.38318289	0.5579332			

	df	AIC
full_logit	10	18359.63
binary_logit	8	18360.47

The results of the binary-only model are largely consistent with the full model. Fatal or major injury crashes, overturned vehicles, and speeding remain strong positive predictors of alcohol involvement. Aggressive driving continues to show a negative association. Both age-related predictors—drivers aged 16–17 and drivers aged 65 or older—also remain significant and retain similar effect sizes. As in the full model, cell phone use is not a significant predictor of alcohol involvement.

Comparing the reduced model with the full model shows that removing the continuous predictors does not change the significance of any of the key crash-related or demographic variables. However, the full model has a slightly lower Akaike Information Criterion (AIC = 18359.63) than the reduced model (AIC = 18360.47). Because lower AIC values indicate better model fit, this comparison suggests that the full model provides a marginally better fit, even though the continuous predictors do not substantially alter the significance or magnitude of the main effects.

## 4 Discussion