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

Steven Chang

1. Introduction

According to the US Department of Transportation, nearly 30 people a day - or approximately one person every 51 minutes - die in motor vehicle crashes that involved an alcohol-impaired driver, while many more are injured. A recent study conducted by the National Highway Traffic Safety Administration has shown that the economic impact of alcohol-related crashes is estimated to be more than $59 billion annually.

In this paper, I wish to explore the relationships between certain driving behaviors, select neighborhood characteristics, and the presence of alcohol in motor vehicle crashes at the block group level in Philadelphia. More specifically, I seek to explain the relationships between the alcohol-involvement indicator (**DRINKING_D**, binary dependent variable), and the following binary driving behavior predictor variables:

- Whether the crash resulted in fatality or major injury
- Whether the crash involved an overturned vehicle
- Whether the driver was using the cellphone that resulted in the crash
- Whether the crash involved a speeding vehicle
- Whether the crash involved aggressive driving
- Whether the crash involved at least one driver who is 16 or 17 years old
- Whether the crash involved at least one driver who is 65 years of age or older

I speculate that all of the above factors can be related to drunk driving. Drunk driving is more likely to lead to serious crashes, which may be a result of more aggressive driving, speeding, and distraction from a cell phone. A serious crash may be more likely to result in an overturned vehicle. Younger drivers are also more likely to be in party scenarios in exposure to alcohol, and they may not make the best judgements between drinking and safe driving. I also look at the relationships between the aforementioned dependent variable and the following continuous predictor variables:

- % of individuals 25 years or older who have at least a bachelor's degree
- Median household income of census block group where the crash occurred

In this analysis, I will be using R to run logistic regressions, as it is an effective regression model to handle binary dependent variables.

2. Methods

In this analysis, I move away from OLS regression and instead use logistic regression to explore the relationships between the dependent and predictor variables. This is due to the inherent nature of OLS regression models, stating that the relationship stems from the amount the dependent variable changes from a one-unit change in one of the predictor variables, holding everything else constant. In this analysis, the dependent variable is binary and coded as 0's and 1's, and it can only change from 0 to 1 or from 1 to 0. To say a one-unit change in the predictor variable would result in a Beta change in the dependent variable would therefore make no sense.

Logistic regression can work around this by predicting the log odds of a certain outcome in the dependent variable, from a combination of binary and continuous predictor variables. Odds are defined as the probability of an event occurring, and the odds ratio is the ratio of the two odds. The odds ratio can be computed by dividing the probability of an event occurring by 1 minus the probability of that event occurring.

- Logit form: $\ln\left(\frac{p}{1-p}\right)$ = β_0 + β_1**FATAL_OR_M** + β_2**OVERTURNED** + β_3**CELL_PHONE** + β_4**SPEEDING** + β_5**AGGRESSIVE** + β_6**DRIVER1617** + β_7**DRIVER65PLUS** + β_8**PCTBACHMOR** + β_9**MEDHHINC** + ε
- Logistic form: P(**DRINKING_D** = 1) = $\frac{1}{1+e^{-\beta_0-\beta_1 FATAL\_OR\_M-\beta_2 OVERTURNED-\beta_3 CELL\_PHONE-\beta_4 SPEEDING-\beta_5 AGGRESSIVE-\beta_6 DRIVER1617-\beta_7 DRIVER65PLUS-\beta_8 PCTBACHMOR-\beta_9 MEDHHINC}}$

The equations for the logistic regression model are presented above, where p is the probability that the dependent variable **DRINKING_D** takes on the value of 1, β_0 is the intercept, β_1 through β_9 are the beta coefficients of the 9 predictor variables, and ε is the residual. Collectively, In the logistic equation, $\frac{p}{1-p}$ is the odds of DRINKING_D being 1, and $\ln\left(\frac{p}{1-p}\right)$ is the log odds or logit of DRINKING_D being 1, meaning alcohol was involved in an accident.

In a linear regression function predicting probabilities, issues may arise in that the resulting predicted probabilities ($\hat{y}$) can be anywhere between -∞ and +∞, while all probabilities must fall somewhere between 0 and 1. The logistic function serves as a translator function such that:

- The closer the $\hat{y}$ from the linear regression is to -∞, the closer the predicted probability is to 0
- The closer the $\hat{y}$ from the linear regression is to +∞, the closer the predicted probability is to 1
- No predicted probabilities are less than 0 or greater than 1

The logistic, or inverse-logit function, is written with p on the left-hand side, as the result of solving for p = P(**DRINKING_D** = 1) from the logit function. Figure 1 below shows the graphs of

the logit and logistic functions, with the latter (RIGHT) producing a graph with the probability as the dependent variable on the y-axis within the appropriate confines between 0 and 1. The logistic function works well for models where the dependent variable is binary in that it is predicting for the probability a certain outcome will occur. This makes any interpretations more appropriate and bypasses the meaningless interpretation of one-unit changes in variables that can only be 0 or 1.

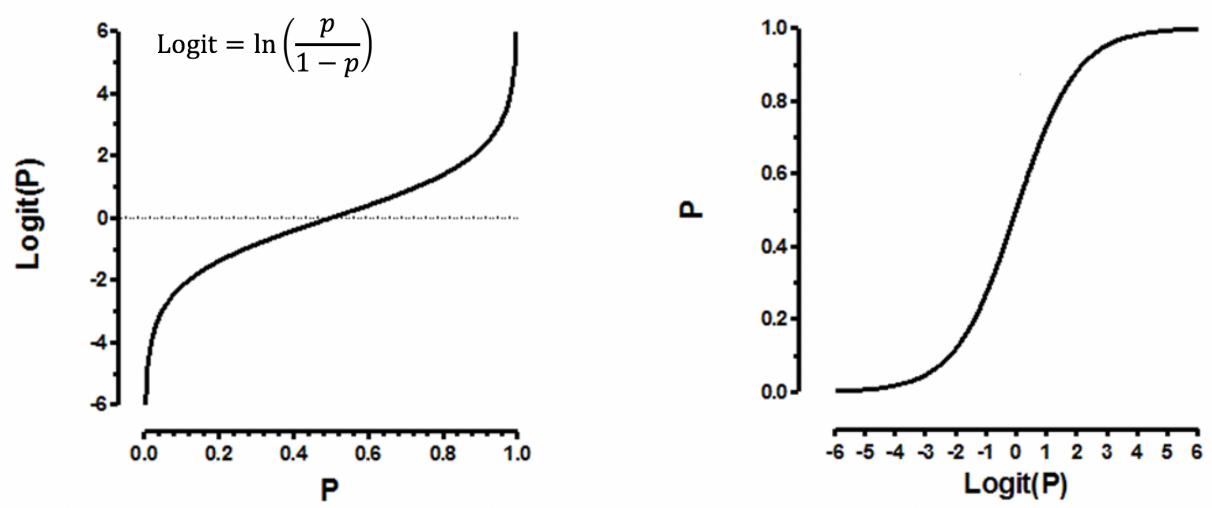


Figure 1: Graphs of the logit function (LEFT) and the logistic function (RIGHT)

For each predictor included in my model, I am testing for the following hypotheses:

- Null hypothesis: beta-coefficient of the predictor variable is 0, or the odds ratio of a certain outcome of the predictor variable is 1
- Alternative hypothesis: beta-coefficient of the predictor variable is NOT 0, or the odds ratio of a certain outcome of the predictor variable is NOT 1

The Wald statistic of each predictor variable, also known as the z-value, is the quantity $\frac{\widehat{\beta}_i}{\sigma_{\widehat{\beta}_i}}$. It has a standard normal distribution. When interpreting the regression results, rather than looking at the beta-coefficients, most statisticians prefer to look at the odds ratios, which are calculated by exponentiating the coefficients.

When assessing the quality of model fit, it is important to note that an R-squared may be calculated for logistic regressions, but it is no longer a very useful metric and doesn't have the same interpretation as in an OLS regression. An R-squared in a logistic regression CANNOT be interpreted as the percentage of variance in the dependent variable that was successfully explained by the model. Instead of relying on the R-squared to assess the quality of model fit, the Akaike Information Criterion (abbreviated as AIC from here on) will be used. The AIC estimates

the amount of information that is lost when a statistical model is used to represent the real-life processes that generated the data. A lower AIC score is indicative of less information lost and a better fit, and therefore a better model will have a lower AIC score than a weaker model.

Other measures of model fit quality include the sensitivity, specificity, and misclassification rate. The sensitivity, also called the true positive rate, measures the proportion of actual positives that are correctly predicted as such, and is complementary to the false negative rate. Specificity, also called the true negative rate, measures the proportion of actual negatives that are correctly predicted as such, and is complementary to the false positive rate. The misclassification rate is, as its name suggests, the proportion of incorrect predictions, and is the sum of actual positives predicted as negatives and actual negatives predicted as positives. Higher values for the sensitivity and specificity measures, and lower values for the misclassification rate are indicative of better model fit. The sensitivity, specificity, and misclassification rate are calculated based on the fitted value of the dependent variable based on some combination of predictor variables. In this case, they can be interpreted as the predicted probabilities that alcohol was involved in a crash based on a series of involvement of other accident-related factors. When determining a suitable cut-off for what is considered a “high” probability of Y=1 when calculating the sensitivity, specificity, and misclassification rate, it is important to adjust the cutoff based on the distribution of probabilities of the dataset at hand. This will ensure that the “high” probabilities are isolated and presented appropriately, rather than containing too few or too many observations.

An ROC curve is a common measure to assess the predictive quality of a logistic regression model by plotting the false positive rate against the true positive rate. The best cutoff value may be determined by optimizing sensitivity and specificity. More specifically, the Youden index selects the cutoff value for which both the sensitivity and specificity are maximized. Another approach entails selecting the cutoff value at which the ROC curve is at its minimum distance to the upper left corner where sensitivity and specificity are both 1. In this analysis, I will use the latter, minimum-distance approach to select the cutoff value.

The area under ROC curve (AUC) is commonly used to measure the predictive accuracy of a model. Higher AUC values indicate that I can find cutoff values for which both sensitivity and specificity are relatively high. AUC values can range between 0.5 (area under 45-degree line) and 1 (area of the whole box), and can be roughly interpreted as follows:

- 0.90 ~ 1.00: Excellent
- 0.80 ~ 0.90: Good
- 0.70 ~ 0.80: Fair
- 0.60 ~ 0.70: Poor

- 0.50 ~ 0.60: Fail

Just like in OLS regression, logistic regression assumes that there is independence of observations and no multicollinearity. Unlike an OLS regression, a logistic regression uses a binary dependent variable, and does not need a linear relationship between the dependent variable and each predictor. Residuals do not need to be normally distributed, and there is also no assumption of homoscedasticity.

Before running a logistic regression on a dataset, most statisticians may wish to do some exploratory analysis. For the binary predictors, I run a cross-tabulation between the dependent variable and each binary predictors to see whether there is an association between the two variables. I then conduct the Chi-Squared test to examine whether there is an association between each binary predictor and the two categories of the binary dependent variable - whether alcohol was involved in a crash. The null hypothesis states that the proportion of 1 and 0 values for each predictor is the same for accidents that involve and don't involve alcohol. The alternative hypothesis states that the proportion of 1 and 0 values for each predictor is NOT the same for accidents that involve and don't involve alcohol.

For the continuous predictors, I compare their respective means of each predictor for both values of the binary dependent variable. I then use the independent samples t-test to examine whether there are significant differences among the means of each predictor for the two categories of the binary dependent variable - whether alcohol was involved. The null hypothesis states that there is no difference in the average values of each continuous predictor (percent with bachelor's degree and income) whether or not alcohol is involved. The alternative hypothesis states that there is in fact a difference in the average values of each continuous predictor whether or not alcohol is involved.

3. Results

3.1: Exploratory Analysis

Figure 2: Tabulation of crashes where alcohol is and isn't involved in the city of Philadelphia

	No Alcohol	Alcohol Involved
Number of Crashes	40879	2485
Percentage of Total	94.27	5.73

Figure 2 above shows the number and proportion of crashes with no alcohol involved still vastly exceeds the respective numbers of crashes involving alcohol, by roughly 16 times.

Figure 3: Cross tabulation between alcohol involvement and factors that may be related to a crash with percentage differences and their significance

	No Alcohol (**DRINKING_D** = 0)		Alcohol Involved (**DRINKING_D** = 1)		Total $\chi 2$	p-value
	N	**%**	**N**	**%**		
FATAL_OR_M: Crash resulted in fatality or major injury	1181	2.9	188	7.6	1369	<0.0001
OVERTURNED: Crash involved an overturned vehicle	612	1.5	110	4.4	722	<0.0001
CELL_PHONE: Driver was using a cellphone	426	1.0	28	1.1	454	0.69
SPEEDING: Crash involved a speeding vehicle	1261	3.1	260	10.5	1521	<0.0001
AGGRESSIVE: Crash involved aggressive driving	18522	45.3	916	36.9	19438	<0.0001
DRIVER1617: Crash involved at least one driver who is 16 or 17	674	1.6	12	0.5	686	<0.0001
DRIVER65PLUS: Crash involved at least 1 driver who is at least 65	4237	10.4	119	4.8	4356	<0.0001

Figure 3 above shows that all the predictor variables with the exception of **CELL_PHONE** usage have $\chi 2$ p-values lower than the 0.05 cutoffs. This suggests that there is reason to reject the null hypothesis in favor of the alternative hypothesis, and that there is an association between alcohol involvement and all the predictor variables except cell phone usage.

Figure 4: Table examining whether the means of the two continuous predictors seem to differ for the different levels of the dependent variable

	No Alcohol (**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

A comparison of the mean values of the two continuous predictor variables across the two levels of the binary dependent variable returned p-values greater than 0.05. I therefore fail to reject the null hypothesis in favor of the alternative hypothesis, and that there is no association between alcohol involvement and median household income or percentage of people with bachelor's degrees. In other words, the average values of the variable **PCTBACHMOR** and **MEDHHINC** are not statistically significantly different for crashes that involved alcohol and crashes that didn't.

Figure 5: Pearson correlations matrix between all predictor variables

```
> correlation <- dataset[c(4:10, 12:13)]
> cor(correlation, method = "pearson")
              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
```

The correlation matrix presented above shows that there is not a lot of multicollinearity between most of the predictor variables. Multicollinearity among predictor variables is defined as two or more predictors are correlated and tend to change in a more joint fashion. A Pearson correlations matrix presents the correlations between all predictor variables (including itself). A value of 1

indicates perfect correlation, while values close to 0 indicate no correlation. The only exceptions to this observation are between SPEEDING and AGGRESSIVE driving, with a correlation of 0.21, and between median household income and percent with bachelor's degrees, with a correlation of 0.48.

3.2: Logistic Regression

Figure 6: Logistic regression output results including binary and continuous predictors

```
> finallogitoutput <- cbind(logitcoeffs, or_ci)
> finallogitoutput
                 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
(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
```

My logistic regression results show that the odds of alcohol involvement is significantly higher in crashes that involved fatalities, overturned vehicles, and speeding. The odds of alcohol involvement is significantly lower in crashes that involved aggressive driving, young drivers 16 or 17 years of age, and older drivers over 65 years of age. The odds that an accident involving fatalities that also involved alcohol are 2.26 the odds that an accident with no fatalities involved alcohol, holding values of all other predictors constant. The odds that an accident involving overturned vehicles that also involved alcohol are 2.53 the odds that an accident with no overturned vehicles involved alcohol. The odds that an accident involving cellphone use that also involved alcohol are 1.03 the odds that an accident with no cellphone usage involved alcohol. The odds that an accident involving speeding that also involved alcohol are 4.66 the odds the odds that an accident with no speeding involved alcohol. The odds that an accident involving aggressive driving are 0.55 the odds that an accident with no aggressive driving involved alcohol. The odds that an accident involving young drivers that also involved alcohol are 0.28 the odds that an accident with no young drivers involved alcohol. The odds that an accident involving old drivers that also involved alcohol are 0.46 the odds that an accident with no old drivers involved

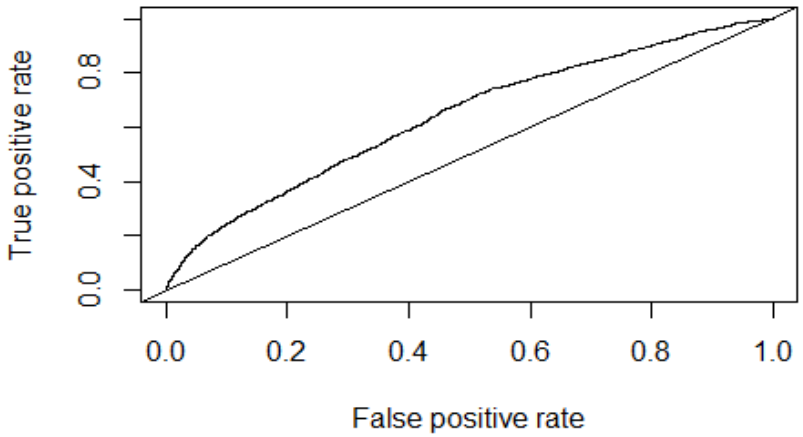
alcohol. A 1% increase in percent residents with bachelor's degrees OR a $1 increase in median household income (all else equal) both result in no changes to the odds that an accident will involve alcohol. Cellphone usage and percent residents with bachelor's degrees do not appear to be statistically significant predictors for alcohol involvement. Median household income seems to be significant, but only at the 0.01 level. All other predictors are significant. The regression returns an AIC score of 18360.

Figure 7: Sensitivity, specificity, and misclassification rate for various cut-off values of logistic regression

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	0.097
0.1	0.164	0.948	0.097
0.15	0.104	0.972	0.078
0.2	0.023	0.995	0.060
0.5	0.002	0.9999	0.057

Figure 7 above shows that a cutoff of 0.5 yields the lowest misclassification rate of roughly 0.057, and a cutoff of 0.02 yields the highest misclassification rate of roughly 0.889.

Figure 8: ROC curve for optimal cutoff selection



The ROC curve presented in Figure 8 above presents an optimal cutoff value of about 0.064, yielding a sensitivity of about 0.66, a specificity of about 0.55, and an area under the curve of about 0.64. This indicates that I have a poor predictive model.

Figure 9: Logistic regression output results including only binary predictors

	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

```
(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
```

My logistic regression results above, including only the binary predictors, show that the odds of alcohol involvement is significantly higher in crashes that involved fatalities, overturned vehicles, and speeding. The odds of alcohol involvement is significantly lower in crashes that involved aggressive driving, young drivers 16 or 17 years of age, and older drivers over 65 years of age.

Cellphone usage does not appear to be a statistically significant predictor for alcohol involvement. The regression returns an AIC score of 18360. The above results are almost identical with the logistic regression that also included the education and income continuous predictors.

Both regression models returned identical AIC scores, so I cannot say that whether including or excluding the continuous predictors necessarily creates a stronger model. However, since dropping the continuous predictors didn't result in a substantial change in AIC, it may be beneficial to just use the simpler model with fewer predictors.

4. Discussion

Drinking and driving is a dangerous combination that is a common causes for great emotional and economic losses. In this paper, I attempt to explore the relationships between certain driving behaviors, select neighborhood characteristics, and the presence of alcohol in motor vehicle crashes at the block group level in Philadelphia. More specifically, I utilized logistic regression models to predict the odds of alcohol involvement in crashes that also involved fatalities, overturned vehicles, cellphone usage, speeding, aggressive driving, young drivers 16 or 17 years old, and old drivers 65 years or older.

My results show that fatality, overturned vehicles, speeding, aggressive driving, young drivers, old drivers, and median household income are strong predictors of drunk driving. Cellphone usage and percent residents with bachelor's degrees in the block group where the accident took place are not statistically significantly associated with the dependent variable.

These results are mostly in line with my expectations. Alcohol consumption can severely impair driving ability, and may cause erratic driving behavior such as speeding and overturning a vehicle, which can then lead to more serious crashes that involve fatalities. My results indicate the odds that alcohol is involved in crashes that also involved fatalities, overturned vehicles, and speeding are much higher than crashes that did not involve these factors. On the other hand, my results also indicate the odds that alcohol is involved in crashes that also involved aggressive driving, young drivers, and old drivers are much lower than crashes that did not involve these factors. This is also consistent with my expectations in that an impaired driver is less likely to be cognitively capable of driving aggressively, drinking laws are in place that can act as a strong deterrent to underage drinking, and older people are less likely to drink in general.

As Paul Allison, a leading expert on logistic regression, points out, substantial bias may be introduced to the model by the presence of a small number of cases on the rarer of the two possible outcomes for the dependent variable. In my dataset, crashes where alcohol is involved

amounts to 2485 out of 43364 total crashes. Since I have more than 2000 observations in the rare category, I should not expect issues using logistic regression. One limitation within the analysis may be the absence of other relevant data that are statistically significantly related to the dependent variable that is not included in my analysis, as seen from the relatively low AUC score of 0.66 from the model indicating poor predictive ability.

5. Appendix (R Code)

```
library(aod)
library(ggplot2)
library(rms)
library(gmodels)
library(ROCR)

dataset <- read.csv("Logistic Regression Data.csv")
head(dataset)
summary(dataset)

#Exploratory analyses

#2a
DRINKING_D.tab <- table(dataset$DRINKING_D)
table(DRINKING_D.tab)
prop.table(DRINKING_D.tab)

#2b
CrossTable(dataset$FATAL_OR_M, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE)
CrossTable(dataset$OVERTURNED, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE)
CrossTable(dataset$CELL_PHONE, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE)
CrossTable(dataset$SPEEDING, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE)
CrossTable(dataset$AGGRESSIVE, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE)
CrossTable(dataset$DRIVER1617, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE)
CrossTable(dataset$DRIVER65PLUS, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE)

#Cross-tabulation like above with Chi-squared included
CrossTable(dataset$FATAL_OR_M, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE, chisq = TRUE)
CrossTable(dataset$OVERTURNED, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE, chisq = TRUE)
CrossTable(dataset$CELL_PHONE, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE, chisq = TRUE)
CrossTable(dataset$SPEEDING, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE, chisq = TRUE)
CrossTable(dataset$AGGRESSIVE, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE, chisq = TRUE)
CrossTable(dataset$DRIVER1617, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE, chisq = TRUE)
```

```
CrossTable(dataset$DRIVER65PLUS, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.chisq =
FALSE, chisq = TRUE)

#2c - mean and standard deviations of 2 continuous variables for alcohol and non-alcohol related
crashes
tapply(dataset$PCTBACHMOR, dataset$DRINKING_D, mean)
tapply(dataset$PCTBACHMOR, dataset$DRINKING_D, sd)

tapply(dataset$MEDHHINC, dataset$DRINKING_D, mean)
tapply(dataset$MEDHHINC, dataset$DRINKING_D, sd)
#tests for significance in above 2 variables across 2 crash categories
t.test(dataset$PCTBACHMOR~dataset$DRINKING_D)
t.test(dataset$MEDHHINC~dataset$DRINKING_D)

#2d - Pearson correlations
correlation <- dataset[c(4:10, 12:13)]
cor(correlation, method = "pearson")

#3 - Logistic regression
logit <- glm(DRINKING_D ~ FATAL_OR_M + OVERTURNED + CELL_PHONE + SPEEDING + AGGRESSIVE +
DRIVER1617 + DRIVER65PLUS
             + PCTBACHMOR + MEDHHINC, data = dataset, family = "binomial")
summary(logit)

exp(cbind(OR = coef(logit), confint(logit)))

logitoutput <- summary(logit)
logitcoeffs <- logitoutput$coefficients
logitcoeffs

or_ci <- exp(cbind(OR = coef(logit), confint(logit)))

finallogitoutput <- cbind(logitcoeffs, or_ci)
finallogitoutput

#Sensitivity, specificity, misclassification
fit <- logit$fitted
fit.binary = (fit>=0.02)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.03)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.05)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.07)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.08)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.09)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.1)
```

```
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.15)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.2)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

fit.binary = (fit>=0.5)
CrossTable(fit.binary, dataset$DRINKING_D, prop.r = FALSE, prop.t = FALSE, prop.c = FALSE,
prop.chisq = FALSE)

#Generate ROC curve
a <- cbind(dataset$DRINKING_D, fit)
colnames(a) <- c("labels", "predictions")
head(a)
roc <- as.data.frame(a)
pred <- prediction(roc$predictions, roc$labels)
roc.perf = performance(pred, measure = "tpr", x.measure = "fpr")
plot(roc.perf)
abline(a = 0, b = 1)

#Calculate sensitivity, specificity, and optimal cutoff
opt.cut = function(perf, pred){
  cut.ind = mapply(FUN=function(x, y, p){
    d = (x - 0)^2 + (y-1)^2
    ind = which(d == min(d))
    c(sensitivity = y[[ind]], specificity = 1-x[[ind]],
      cutoff = p[[ind]])
  }, perf@x.values, perf@y.values, pred@cutoffs)
}

print(opt.cut(roc.perf, pred))

#Calculate area under curve
auc.perf = performance(pred, measure ="auc")
auc.perf@y.values

#Re-runs logistic regression without continuous predictors
logit_2 <- glm(DRINKING_D ~ FATAL_OR_M + OVERTURNED + CELL_PHONE + SPEEDING + AGGRESSIVE +
DRIVER1617 + DRIVER65PLUS,
             data = dataset, family = "binomial")
summary(logit_2)

exp(cbind(OR = coef(logit_2), confint(logit_2)))

logitoutput_2 <- summary(logit_2)
logitcoeffs_2 <- logitoutput_2$coefficients
logitcoeffs_2

or_ci_2 <- exp(cbind(OR = coef(logit_2), confint(logit_2)))

finallogitoutput_2 <- cbind(logitcoeffs_2, or_ci_2)
finallogitoutput_2
```