Examining Predictors of Car Crashes Caused by Alcohol Using Logistic Regression

Akira Di Sandro, Sofia Fasullo, Amy Solano

2024-11-21

Contents

Introduction

In this report, we seek to identify predictors of crashes related to drunk driving in the City of Philadelphia. Drunk driving is a nationwide issue with severe consequences, taking the lives of more than 35 people per day and injuring many more, according to the National Highway Traffic Safety Administration. Alcohol-impaired crashes also present a large financial burden; the NHTSA estimates that these crashes cost the United States more than $68 billion annually.

To explore this issue and the potential predictors in Philadelphia, we will use a dataset that contains information about the crashes that occured in Philadelphia in the years 2008-2012, including their geographic locations. Though 53,260 crashes occured in this time period, we did not include crashes that took place in non-residential block groups. This leaves us with 43,364 crashes. Our data are merged with census block group data to feature PCTBACHMOR and MEDHHINC for a block group, both of which were used as predictors in our previous reports. Predictors from the crash data are FATAL_OR_M, OVERTURNED, CELL_PHONE, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS. We will regress the binary dependent variable DRINKING_D on these predictors using logistic regression in R.

Methods

A binary variable takes on the value of either 1 or 0, with no other options. OLS regression measures the correlation relationships between predictor variables and the dependent variable (Y), with each predictor either causing Y to increase or decrease. However, a binary variable does not increase or decrease continuously, so it cannot be interpreted this way. Predicting the probability of a binary variable taking a value of 1, P(Y=1), makes more sense and creates a dependent variable that does increase continuously, however it is bounded by [0,1] and OLS models do not make sense for bounded variables. In OLS, a predictor that is positively correlated with the dependent variable would always make it increase, and that does not make sense for a probability over 1, say.

Odds are the probabilities that an event occurs over the probability it does not. The formula for the odds is:

$$\text{Odds} = \frac{P}{1-P}$$

This is also called the **Odds Ratio (OR)**. It differs from the probability of an event because that is calculated by the number of times the event occurs over the total number of times it occurs and doesn't occur. Odds are times it occurs over times it doesn't.

OR is easy to interpret. OR = 1 means the event is equally likely to occur or not occur, OR > 1 means it's more likely to occur and OR < 1 means it's more likely not to occur. Taking the log of the OR creates the perfect translator function for a continuous value between 1 and 0 : P(Y=1), in our case. This is the **logistic function**. The formulas for the logistic function can be shown below, with the variables specific to this analysis:

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

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

- $p = P(Y = 1)$ is the probability that the predictor variable is 1, in our case, the probability DRINKING_D = 1, or the accident was due to drunk driving
- β_0 is the baseline log-odds of $Y = 1$ when all other predictors are 0
- β_n demonstrates the relationship between the nth predictor and p such that as the nth predictor increases by 1, p -increases by $(e^{\beta_n} - 1) * 100$

The logistic function is continuous and bounded by 0 and 1, which is perfect to predict p, the probability that Y=1. In addition, the logistic function is symmetrical which displays the inverse relationship between $P(Y = 1)$ and $P(Y \neq 1)$, which is the base of the odds ratio.

For each predictor, the hypothesis test essentially sees whether that predictor has any effect on the probability of Y=1. The null hypothesis is $H_0 : \beta_0 = 0$ which is equivalent to saying OR = 1. This means that changing a certain predictor does not make a difference to the probability of Y=1. The alternative hypothesis is that $H_a : \beta_0 \neq 0$, or that OR $\neq$ 1. This means that the predictor does have an effect on the probability that Y=1.

The equation for the Wald statistic is as follows:

$$W = \frac{\beta^2}{Var(\beta)}$$

This statistic has a χ^2 distribution with one degree of freedom, which can be used to test the hypothesis described above of whether a particular predictor has a result on the odds of the dependent variable.

Rather than looking at the estimated β coefficients, most statisticians prefer to look at odds ratios, which are calculated by exponentiating the coefficients.

In OLS regression, the R-squared value measures the proportion of variance in the dependent variable Y that is explained by the independent variables in the model. However, this measure is not particularly meaningful for logistic regression because logistic regression models the probability that a binary dependent variable takes on a value of, rather than modeling the variance of Y. As a result, the concept of variance

in Y does not translate directly in the context of logistic regression, making the traditional R-squared less useful as an indicator of goodness of fit.

The **Aikaike Information Criterion (AIC)** is a measure of the goodness of fit of a model that takes into account log-likelihood of the model as well as the number of parameters in order to balance overfitting. The formula for AIC is as follows:

$$\text{AIC} = 2k - 2\ln(L)$$

Where k is the number of independent variables in the model, and L is the likelihood of the model, which is the probability of observing the given set of $Y = y_1, y_2, ..., y_n$ variables in the sample. A low AIC value demonstrates a well-fitting model because the log likelihood of the observed Y values is high in comparison to the number of predictor variables used.

A binary variable takes on a value of 1 or of 0. In logistic regression, we predict the probability that Y takes on a value of 1 given a set of predictor variables. This probability does not give any certain answers, but provides a guideline for the prediction of Y, denoted by $\hat{y}$. Using the information provided by the model, we choose a **cutoff value** which specifies at which probability we will start predicting $\hat{y} = 1$. For example, with a cutoff value of 0.5, an observation that yields P(Y=1) = 0.4 with our model will lead us to predict that $\hat{y} = 0$ at that point, and an observation that yields P(Y=1) = 0.7 with our model will lead us to predict that $\hat{y} = 1$ at that point.

Sensitivity is the rate at which true Y=1 points are correctly predicted by the model at a given cutoff rate. A **Type II Error** refers to the rate that a prediction that $\hat{y} = 0$ is incorrect (so Y=1). These are called false negatives. So higher sensitivity relates to lower type II error.

$$\text{Sensitivity} = \frac{\text{true positive predictions}}{\text{true positive predictions} + \text{false negative predictions}}$$

Specificity is the rate at which true Y=0 points are correctly predicted by the model at a cutoff rate. A **Type I Error** refers to the rate that a prediction that $\hat{y} = 1$ is incorrect (so Y=0). These are called false positives. So higher sensitivity relates to lower type I error.

$$\text{Sensitivity} = \frac{\text{true negative predictions}}{\text{true negative predictions} + \text{false positive predictions}}$$

P(Y=1) will have a specific distribution for every model. Depending on what we chose for our cutoff rate to say when $\hat{y} = 1$, we may predict more $\hat{y} = 1$ than $\hat{y} = 0$ or vice versa. If P(Y=1) falls mostly below 0.4 and we choose to call 0.7 a "high" P(Y=1) and thus our cutoff value, it can be assumed that very few observations will be predicted as $\hat{y} = 1$. As a result, the sensitivity and specificity values will change. Sensitivity and specificity are inversely related. If every observation is predicted as $\hat{y} = 1$, there will be a 0% specificity and perfect 100% sensitivity. The inverse is true if every value is predicted as $\hat{y} = 0$. Typically, we want a model that balances sensitivity and specificity. As a result, it is important that we test the sensitivity and specificity rates for multiple cutoff values, and choose an optimal value. The **misclassification rate** is the rate of type I error + type II error. In other words, it is the rate at which any values of Y are predicted incorrectly. For this report, we are evaluating which cutoff rate creates the lowest square of the misclassification rate. Thus, we are balancing sensitivity and specificity.

The ROC curve visualizes the sensitivity rate against the specificity rate at every cutoff value of P(Y=1), from 0 to 1. The higher the curve reaches, the greater sensitivity and specificity the model can reach.

Beyond just visually inspecting the ROC curve graph, we can calculate the area under the ROC curve (abbreviated AUC). The higher the AUC value, the higher the curve and thus sensitivity and specificity values the model can balance. A rough guide for the AUC values and model goodness of fit are as follows:

- 0.90-1 = excellent
- 0.80-0.90 = good
- 0.70-0.80 = fair
- 0.60-0.70 = poor

- 0.50-0.60 = fail

Some assumptions of OLS regression hold for logistic regression:

- Independence of observations
- No severe multicollinearity

Some assumptions of OLS do not hold for logistic regression:

- Linear relationship between dependent variable and predictors: in logistic regression, the dependent variable is binary so this does not hold
- Normality of residuals: because of the bounded and non-linear nature of the logistic function, error terms cannot be normal
- Sample sizes do not have to be as large for OLS as they do for logistic regression, because we need a set of observations where both Y=1 and Y=0.

Before engaging in predictive modeling, we perform exploratory analysis to see whether there are relationships between certain variables we may use as predictors and the dependent variable. Cross-tabulation is the process of comparing all variables of interest with dependent variable and with each other (to assess collinearity). This looks slightly different from correlation cross-tabulations in OLS regression, because the dependent variable in this case is binary.

In one case, both the dependent and predictor variables may be binary. Statisticians frequently employ the **Chi-Square** (χ^2) test to examine whether the distribution of one binary variable depends on another. For example, consider the relationship between the variables DRINKING_D and FATAL_OR_M. The null and alternative hypotheses for the χ^2 test would be:

- H_0 (Null Hypothesis): The proportion of fatalities in crashes involving drunk drivers is the same as the proportion in crashes without drunk drivers.
- H_a (Alternative Hypothesis): The proportion of fatalities in crashes involving drunk drivers differs from the proportion in crashes without drunk drivers.

A large χ^2 statistic, combined with a p-value below 0.05, provides evidence to reject the null hypothesis in favor of the alternative. This would indicate an association between drunk driving and crash fatalities.

We can also observe the relationships between continuous predictor variables and a binary dependent variable by comparing the means of continuous predictors for both values of the dependent variable.

As used in introductory statistics, comparing the mean of a continuous variable across two independent groups typically involves using the independent samples t-test. For instance, we can determine whether the average PCTBACHMOR values differ significantly between crashes involving drunk drivers and those that do not. The null and alternative hypotheses for this test are as follows:

- H_0 (Null Hypothesis): The average PCTBACHMOR values are the same for crashes involving drunk drivers and those that do not.
- H_a (Alternative Hypothesis): The average PCTBACHMOR values differ for crashes involving drunk drivers compared to those that do not.

Results

Exploratory Analysis

Table 1: Proportion of Crashes that involved a Drunk Driver

Drunk Driver Involved?	Proportion
Yes	0.057
No	0.943

According to Table 1, we see that 2,485 out of 43,364, or 5.7% of crashes involved drunk driving.

Table 2: Cross-Tabulation of DV and Binary Predictors

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		Total	
	N	%	N	%	N	χ^2
FATAL_OR_M	1181	2.89%	188	7.57%	1369	0.000
OVERTURNED	612	1.5%	110	4.43%	722	0.000
CELL_PHONE	426	1.04%	28	1.13%	454	0.687
SPEEDING	1261	3.08%	260	10.46%	1521	0.000
AGGRESSIVE	18522	45.31%	916	36.86%	19438	0.000
DRIVER1617	674	1.65%	12	0.48%	686	0.000
DRIVER65PLUS	4237	10.36%	119	4.79%	4356	0.000

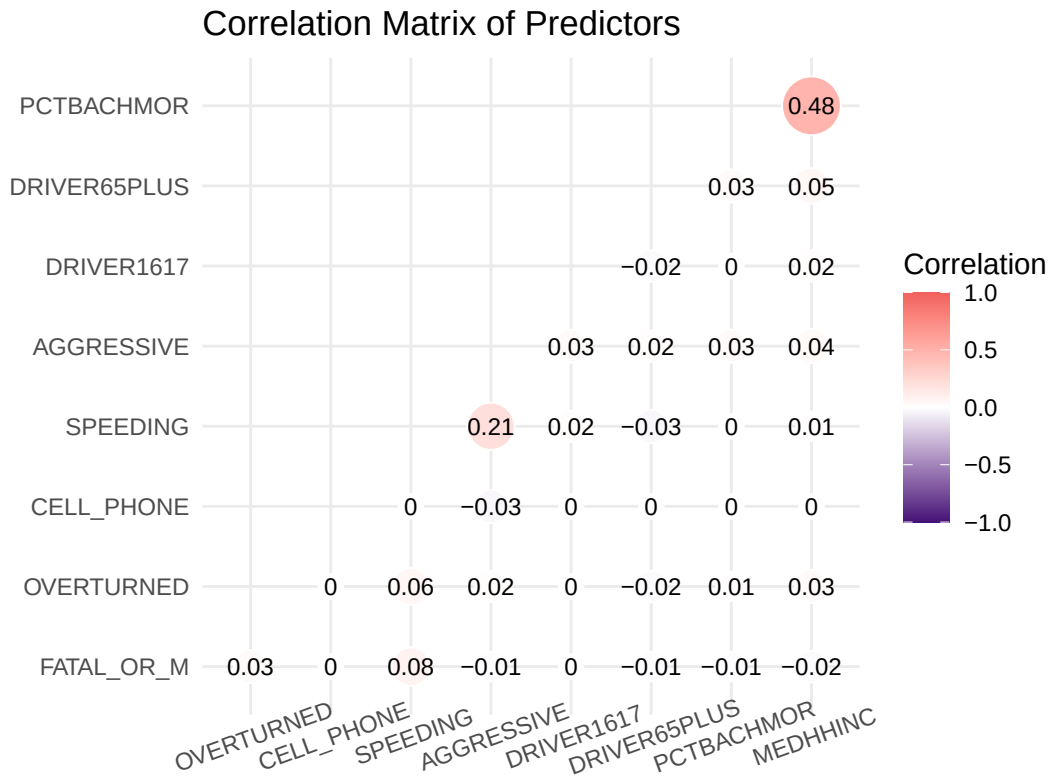
Table 2 shows the cross-tabulation of the dependent variable with each of the binary predictors. We also present the p-values for the Chi-Square test for each binary predictor in this table. We see that most of the binary predictors resulted in a very low p-value, meaning we could reject the null hypothesis and say that there is a significant association between the dependent variable and most of the binary predictors. We see crashes involving a drunk driver being associated with a higher proportion of fatal or major injury crashes, overturned vehicles, and speeding while crashes not involving a drunk driver have a higher proportion of aggressive driving, at least one driver being 16 or 17 years old, and at least one driver being 65 years or older.

The one exception to this is cell-phone usage. The proportion of crashes for which a driver was using a cellphone did not significantly differ between crashes that involved a drunk driver and those that didn't.

Table 3: Table 3. Group Means for Continuous Predictors

	No Alcohol Involved (DRINKING_D = 0)		Alcohol Involved (DRINKING_D = 1)		
	mean	SD	mean	SD	t-test p-value
PCTBACHMOR	16.6%	18.2%	16.6%	18.7%	0.914
MEDHHINC	$31,483.05	$16,930.10	$31,998.75	$17,810.50	0.160

With regards to assumptions for logistic regression, we use a dataset of 43,364 crash observations which is a sufficiently large dataset and we have a binary dependent variable (did the crash involve a drunk driver or not). We also have independence of observations, since each crash is independent of the others.



According to our correlation matrix, we see some multicollinearity between PCTBAHCMOR and MEDHHINC (r = 0.48) and slight multicollinearity between SPEEDING and AGGRESSIVE (r = 0.21). Although an r of 0.48 is quite high, it is still under 0.7 (the threshold we use for high multicollinearity), so we move on with our regression analysis.

A potential limitation of using Pearson correlations to measure associations between binary predictors is that Pearson correlations assume a linear relationship between the variables it's comparing, but the relationship between two binary variables is often not linear.

Logistic Regression

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Nov 21, 2024 - 22:20:15

Table 4:

	Coefficient	p-value	Odds Ratio	2.5 %	97.5 %
(Intercept)	-2.733	0	0.065	0.059	0.071
FATAL_OR_M	0.814	0	2.257	1.910	2.653
OVERTURNED	0.929	0	2.532	2.035	3.122
CELL_PHONE	0.030	0.881	1.030	0.684	1.488
SPEEDING	1.539	0	4.660	3.974	5.450
AGGRESSIVE	-0.597	0	0.551	0.501	0.604
DRIVER1617	-1.280	0	0.278	0.148	0.471
DRIVER65PLUS	-0.775	0	0.461	0.380	0.553
PCTBACHMOR	0	0.775	1	0.997	1.002
MEDHHINC	0	0.036	1	1	1

Here we see the results of the logistic regression with all predictors. We see that FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1617, DRIVER65PLUS, and MEDHHINC are all significant predictors with $p < 0.05$. With the exception of MEDHHINC, all of those predictors in fact have $p < 0.001$. CELL_PHONE and PCTBACHMOR are the two predictors that are not statistically significant. MEDHHINC and PCTBACHMOR basically have coefficients of 0 (when rounding), telling us that they are not great predictors for whether a crash involved a drunk driver.

The interpretations of the coefficients of our statistically significant predictors are as follows. For each predictor, the interpretation assumes that all other variables are held constant.

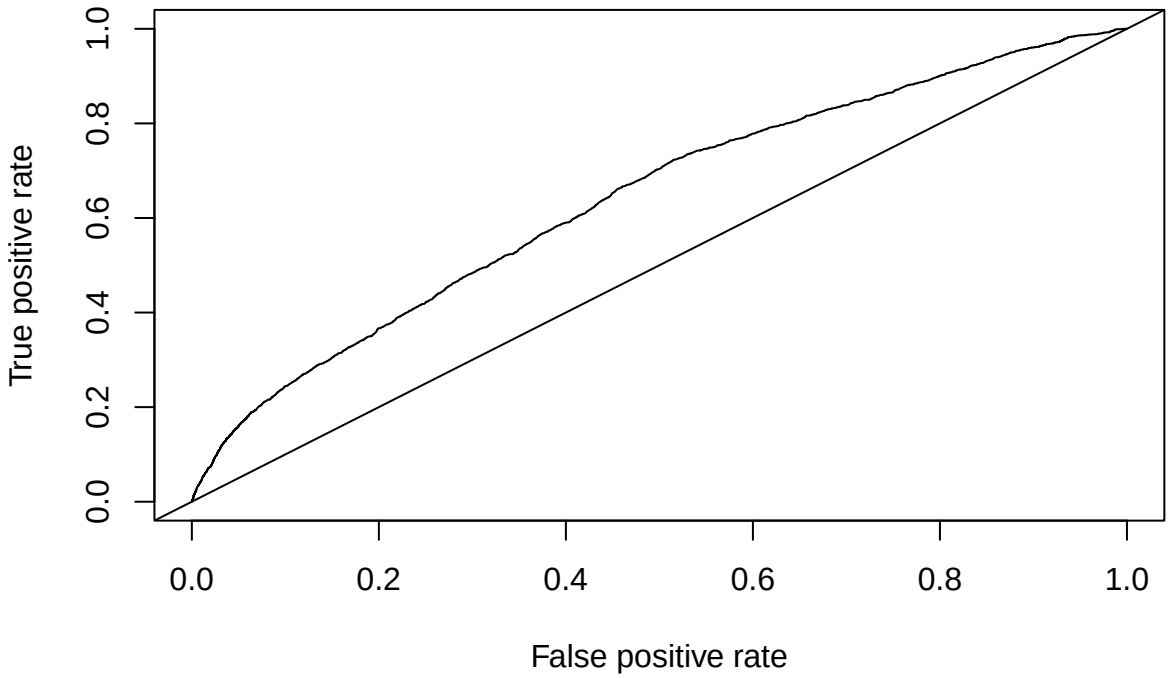
- The odds of a crash involving a drunk driver goes up by $e^{0.814} - 1 = 125.7\%$ when the crash resulted in a fatality or major injury.
- The odds of a crash involving a drunk driver go up by $e^{0.929} - 1 = 153.2\%$ when the crash involves an overturned car.
- The odds of a crash involving a drunk driver go up by $e^{1.539} - 1 = 366.0\%$ when the crash involves speeding.
- The odds of a crash involving a drunk driver are 45% lower ($e^{-0.597} - 1$) when the crash involves an aggressive driver.
- The odds of a crash involving a drunk driver are 72.2% ($e^{-1.280} - 1$) lower when the crash involves a driver who was 16 or 17 years old.
- The odds of a crash involving a drunk driver are 53.9% lower ($e^{-0.775} - 1$) when the crash involves a driver who was 65 years or older.

CELLPHONE, PCTBACHMOR, and MEDHHINC are predictors that have no statistically significant effect on our dependent variable.

Table 5: Sensitivity, Specificity, and Misclassification Rate for First Logistic Model

Cutoff	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

Table 5 presents the specificity, sensitivity, and the misclassification rates for the different probability cut-offs. A cut-off of 0.02 leads to the highest misclassification rate with 88.9% of the observations being misclassified, while a cut-off of 0.50 leads to the lowest misclassification rate with 5.7% of the observations being misclassified. Unfortunately, for a cut-off of 0.50, the sensitivity is very low, at 0.002, or .2% of the crashes involving a drunk driver were correctly predicted by the model at this cut-off (while 100% of the crashes not involving a drunk driver were correctly predicted as such).



The ROC curve tells us that the optimal cut-off rate is 0.06. The sensitivity and specificity are maximized at this cut-off, with a sensitivity of 0.661 and a specificity of 0.545. This cut-off is drastically different from the optimum cut-off rate of 0.50 when minimizing the misclassification rate. The cut-off rate of 0.50 has a misclassification rate of 5.7% despite only being able to accurately predict .2% of the crashes that involved a

drunk driver because of how few crashes involved a drunk driver compared to the ones that didn't (reminder that 5.7% of the total observations of crashes involved drunk drivers).

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

```
auc.perf = performance(pred, measure ="auc")
auc.perf@y.values
```

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

The area under the ROC curve for this model is 0.640, which tells us that our model performs relatively poorly (usually 0.70 is a cut-off for a fair model). There are many nuances to predicting crashes that involve a drunk driver that our set of predictors is not able to capture.

A Second Model

We also looked at a second logistic regression model, this time looking only at the binary predictors. All predictors with the exception of CELLPHONE were statistically significant with very similar coefficients and therefore, odds ratios as the first model.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, Nov 21, 2024 - 22:20:18

Table 6:

	Coefficient	p-value	Odds Ratio	2.5 %	97.5 %
(Intercept)	-2.652	0	0.071	0.067	0.074
FATAL_OR_M	0.809	0	2.246	1.901	2.640
OVERTURNED	0.940	0	2.559	2.057	3.156
CELL_PHONE	0.031	0.875	1.032	0.685	1.491
SPEEDING	1.540	0	4.666	3.980	5.457
AGGRESSIVE	-0.594	0	0.552	0.503	0.606
DRIVER1617	-1.272	0	0.280	0.149	0.475
DRIVER65PLUS	-0.766	0	0.465	0.383	0.558

Table 7: AIC Comparison

	df	AIC
First Logit Model	10	18359.63
Second Logit Model (no continuous predictors)	8	18360.47

Table 7 shows the AIC comparison between the first model with all predictors and the second with only the binary predictors. Although a lower AIC is better, the AIC for the first model is 18,359.63 while the second model has an AIC of 18,369.47, so the difference between the two AICs is less than 1, meaning that the two continuous predictors really don't add much to the first model.

Discussion

The variables FATAL_OR_M, OVERTURNED, SPEEDING, AGGRESSIVE, DRIVER1517, and DRIVER65PLUS were strong predictors of crashes that involve drunk driving, each having a p-value of less than 0.01. The continuous variable MEDHHINC was also a significant predictor of the dependent variable, with a p-value of less than 0.05. The other continuous variable, PCTBACHMOR, was not a significant predictor in the first regression, and both continuous variables were removed from the second regression. The binary predictor CELL_PHONE was not associated with the dependent variable in either regression.

Given that the data are from 2008-2012, CELL_PHONE not being a significant predictor is unsurprising. Though cell phones did exist in those years, cell phone use was much less common, especially smartphone use. This variable may be more influential in a model that uses newer data. Variables like OVERTURNED and SPEEDING are also unsurprising given that drivers under the influence of alcohol have impaired senses that translate into less control while driving.

Paul Allison explains that logistic regression may be problematic for modeling rare events, like when we have 100,000 observations, but only 20 events. In our dataset, 2,485 out of 43,364 crashes included a driver under the influence of alcohol, a total of 5.7%. This number of occurrences of the dependent variable is sufficient and we do not need to use an alternative method for modeling rare events; logistic regression is appropriate. However, this number did present some limitations when choosing an optimal cut-off for our model. A cut-off of 0.50 had a very low sensitivity, only correctly predicting 0.2% of crashes involving a drunk driver. The optimal cut-off of 0.06 determined by the ROC curve maximizes sensitivity and specificity for our model, but the area under the ROC curve is only 0.640, suggesting that our model performs poorly even at the optimal cut-off rate.