MUSA 5000, Homework Number 3, Logistic Regression

Richard Barad, Dave Drennan, Jarred Randall

2023-11-28

Contents

Introduction

Driving while intoxicated and the damage that drunk drivers can cause to individuals, families, and property is a serious issue in the United States that results in fatalities every day - the US Department of Transportation states that almost 30 people a day die as a result of drunk drivers. Understanding what predictors of a crash that are associated with drunk driving can help us better understand the nature of these crashes and the outcomes that can occur when certain behaviors are true in incidents that involve drunk drivers.

For this analysis, we will be using Philadelphia, PA crash data between 2008 and 2012 from the PA Department of Transportation. The data is geocoded and was previously merged with US Census block group data to incorporate socioeconomic features. The original data set includes 53,260 crashes and is filtered to remove nearly 10,000 crashes that occurred in non-residential block groups to contain a total of 43,364 crashes that happened in Philadelphia residential areas.

We will conduct a logistic regression and use a combination of binary and continuous variables, with binary variables represented by a 1 for “Yes” and 0 for “No”. Throughout this report, we will also refer to our logistic model as a logit model. Our dependent variable is DRINKING_D, which indicates if the driver was drunk or not, and we use this binary variable in our logistic regression to regress it on the following predictors:

- FATAL_OR_M: whether the crash resulted in a fatality or major injury. We speculate this may correlate with drunk driving due to the many deaths that are reported to occur as a result of driving while intoxicated.
- OVERTURNED: whether the crash involved an overturned vehicle, which we speculate as a more serious crash incident that is more likely if the driver is drunk.
- CELL_PHONE: whether the crash involved a driver using a cell phone, which we speculate would be associated with drunk drivers since they are more likely to be careless behind the wheel and misjudge their ability to multitask effectively.
- SPEEDING: whether the crash involved speeding, which we speculate would be associated with drunk drivers due to reckless or uninhibited behaviors that are often associated with being drunk.
- AGGRESSIVE: whether the crash involved aggressive driving, which we speculate would be associated with drunk drivers for the same reasons as SPEEDING.
- DRIVER1617: whether the crash involved a driver who is 16 or 17 years old, which we speculate would be associated with drunk drivers due to limited experience or low tolerance for alcohol, as well as the reasons associated with SPEEDING and AGGRESSIVE.

- DRIVER65PLUS: whether the crash involved a driver who is 65+ years old, which we speculate would be associated with drunk drivers due to low tolerance for alcohol, as well as the reasons associated with SPEEDING and AGGRESSIVE.
- PCTBACHMOR: percent of individuals 25+ years old in a block group who have at least a bachelor's degrees in the block group where the crash took place, which we speculate would be associated with lower odds for crashes involving drunk drivers if the area has a higher education level and so may be more aware of the dangers of drunk driving.
- MEDHHINC: median household income in a block group where the crash took place, which we speculate would be associated with lower odds for crashes involving drunk drivers if residents who frequently drive in the area have a higher disposable income to call a cab if intoxicated.

We will use the open source statistical software R to run our exploratory analysis and statistical regression in this report.

Methods

Ordinary least squares (OLS) regression works well when the dependent variable (Y) is continuous, and ideally normally distributed. However, OLS regression does not work well when the dependent (Y) variable is binary (e.g: Yes/No, 1/0, True/False). The beta coefficients in OLS regression represent the amount the dependent variable changes by when a predictor changes by one unit - with a binary variable, Y is either 1 or 0, so determining that a one unit increase in a predictor results in a β change in Y is not useful, when Y can only be 1 or 0. Additionally, OLS regression models for a binary variable could potentially result in Y values which are greater than one or below zero, which is also not possible in terms of the variable's potential values.

When working with a binary variable, it is necessary to apply a translator function to the model. The translator function converts the predicted Y to the probability that Y is equal to 1. The logistic function is a common translator function used for modeling binary variables and is the method we will use in our analysis.

Overview of Logit and Logistic Regression

Before discussing the logistic regression formulas in detail, it is first important to understand how odds are calculated. The formula for odds is: $\frac{Number\ of\ Desirable\ Outcomes}{Number\ of\ Undesirable\ Outcomes}$. For a binary dependent variable, the odds are equal to the probability that $Y = 1$ divided by the probability that $Y = 0$, which can also be expressed as $p/1 - p$ in which p stands for the probability that $Y = 1$. The odds are part of the logit regression formula, which we state for our model as:

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 FATAL\_OR\_M + \beta_2 OVERTURNED + \beta_3 CELL\_PHONE + \beta_4 SPEEDING + \beta_5 AGGRESSIVE +$$

The quantity ln($p/1-p$) is called the log odds or logit and represents the log odds of the predicted dependent variable being a 1. The β_0 coefficient is equal to the log odds of the dependent variable being a one when all independent variables are equal to zero. The β coefficients numbered one through nine represent the change in the log odds of the dependent variable when the indicated independent variable changes by one, while all other independent variables are held constant.

If we solve for p, the logit equation above can also be written as:

$$p = \frac{e^{\beta_0 + \beta_1 FATAL\_OR\_M + \beta_2 OVERTURNED + \beta_3 CELL\_PHONE + \beta_4 SPEEDING + \beta_5 AGGRESSIVE + \beta_6 DRIVER1617 + \beta_7 DRIVER65P}}{1 + e^{\beta_0 + \beta_1 FATAL\_OR\_M + \beta_2 OVERTURNED + \beta_3 CELL\_PHONE + \beta_4 SPEEDING + \beta_5 AGGRESSIVE + \beta_6 DRIVER1617 + \beta_7 DRIVER65}}$$