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

CPLN 671/MUSA 501

According to the US Department of Transportation, almost 30 people a day – or approximately one person every 51 minutes – die in motor vehicle crashes that involve an alcohol-impaired driver. Many more individuals are injured in these crashes. 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. For more information, see the website of the Centers for Disease Control and Prevention.

The goal of the current assignment is to identify predictors of accidents related to drunk driving. The data used in this assignment come from a data set containing all 53,260 car crashes in the City of Philadelphia for the years 2008 – 2012. The data set was compiled by the Pennsylvania Department of Transportation, and is made available to the public at OpenDataPhilly.org. In the past, Azavea, one of Philadelphia's most prominent GIS software development firms, has used these data for a number of interesting analyses, which have been published on the company's website.

Because the crash data are geocoded, it is possible to spatially join the data to the 2000 Census block group level data set that was used for the two previous homework assignments. After the spatial join, each crash point contains the median household income and the percent of individuals with at least a bachelor's degree in the block group where the crash took place.

The data set that you are given for this assignment, ***Logistic Regression Data.csv***, contains the following variables:

1) **CRN:** Crash Record Number
2) **DRINKING_D:** Drinking driver indicator (1 = Yes, 0 = No)
3) **COLLISION:** Collision category that defines the crash (0 = Non collision, 1 = Rear-end, 2 = Head-on, 3 = Rear-to-rear (Backing), 4 = Angle, 5 = Sideswipe (same dir.), 6 = Sideswipe (Opposite dir.), 7 = Hit fixed object, 8 = Hit pedestrian, 9 = Other or Unknown)
4) **FATAL_OR_M:** Crash resulted in fatality or major injury (1 = Yes, 0 = No)
5) **OVERTURNED:** Crash involved an overturned vehicle (1 = Yes, 0 = No)
6) **CELL_PHONE:** Driver was using cell phone (1= Yes, 0 = No)
7) **SPEEDING:** Crash involved speeding car (1 = Yes, 0 = No)
8) **AGGRESSIVE:** Crash involved aggressive driving (1 = Yes, 0 = No)
9) **DRIVER1617:** Crash involved at least one driver who was 16 or 17 years old (1 = Yes, 0 = No)

10) **DRIVER65PLUS:** Crash involved at least one driver who was at least 65 years old (1 = Yes, 0 = No)
11) **AREAKEY:** ID of the Census Block Group where the crash took place
12) **PCTBACHMOR:** % of individuals 25 years of age or older who have at least a bachelor’s degree in the Census Block Group where the crash took place
13) **MEDHHINC:** Median household income in the Census Block Group where the crash took place

Even though the original data set has a total of 53,260 car crashes, for the sake of this assignment, we remove the 9,896 crash locations which took place in non-residential block groups, where median household income and vacancy rates are 0, from the data set. The final data set contains the 43,364 crashes that took place in Philadelphia’s residential block groups.

Here, we will be regressing the binary dependent variable, **DRINKING_D**, on the following binary and continuous predictors: **FATAL_OR_M**, **OVERTURNED**, **CELL_PHONE**, **SPEEDING**, **AGGRESSIVE**, **DRIVER1617**, **DRIVER65PLUS**, **PCTBACHMOR**, and **MEDHHINC**.

INSTRUCTIONS

SUGGESTION: READ THE ENTIRE SET OF INSTRUCTIONS BEFORE STARTING TO WORK ON THE ASSIGNMENT

1) Prior to running any of the analyses, be sure to set the working directory using the setwd command, install the relevant R packages using the install.packages command, and load the relevant packages using the library command.

2) Import the file ***Logistic Regression Data.csv*** into R using the read.csv command. Now, you are ready to do some exploratory analyses.

    a. Using the table and prop.table commands, tabulate the dependent variable, **DRINKING_D**. For example, if you named your R data set ***mydata***, the syntax to generate the tabulation would be

    ```
    DRINKING_D.tab <- table(mydata$DRINKING_D)
    prop.table(DRINKING_D.tab)
    ```

    In your report, in addition to the counts that you can obtain with the table command, you will be asked to report the proportion of crashes that involved a drunk driver using the prop.table command as above.

    b. Using the CrossTable command in the gmodels library, examine the cross-tabulations between the dependent variable, **DRINKING_D**, and the following binary predictor variables: **FATAL_OR_M**, **OVERTURNED**, **CELL_PHONE**, **SPEEDING**, **AGGRESSIVE**, **DRIVER1617**, and **DRIVER65PLUS**.

        i. For the first predictor (**FATAL_OR_M)**, record the number and percentage of 1-responses (i.e., fatalities or major injuries) for both categories of the variable **DRINKING_D**, as well as the total number of 1-responses (i.e., fatalities or major injuries) in the data set. That is, how many fatalities or major injuries were there when the driver wasn't inebriated, when the driver was inebriated, and altogether?

        ii. Repeat step 2.b.i above for the rest of the binary predictors.

        iii. In your report, you will be asked to present the cross-tabulations from 2.b.i and 2.b.ii in a table like the one below:

	No Alcohol Involved		Alcohol Involved		Total
	(DRINKING_D = 0)		(DRINKING_D = 1)		
	N	%	N	%	N
FATAL_OR_M: Crash resulted in fatality or major injury	1181	2.90%	188	7.60%	1369
OVERTURNED: Crash involved an overturned vehicle					
CELL_PHONE: Driver was using cell phone					
SPEEDING: Crash involved speeding car					
AGGRESSIVE: Crash involved aggressive driving					
DRIVER1617: Crash involved at least one driver who was 16 or 17 years old					
DRIVER65PLUS: Crash involved at least one driver who was at least 65 years old					

1. In the table, note that the percentages (e.g., 2.90% and 7.60%) should not add up to 100%. This is not an error. Basically, you're asked to specify what % of accidents involving drunk drivers had fatalities/major injuries, and what % of accidents NOT involving drunk drivers had fatalities/major injuries. If you look at the first row of the table, it says that 1181 accidents not involving drunk driving (which represents 2.90% of all accidents that don't involve drunk driving) had a fatality/major injury. On the other hand, 188 accidents involving drunk driving (which is 7.60% of all accidents involving drunk driving) had a fatality/major injury.

iv. Prior to doing predictive modeling, statisticians often use the Chi-Square (χ^2) test to determine whether the distribution of one categorical variable varies with respect to the values of another categorical variable. If we were to look at a cross-tabulation of the variables **DRINKING_D** and **FATAL_OR_M**, the null and alternative hypotheses for the χ^2 test would be as follows:

> H_0: the proportion of fatalities for crashes that involve drunk drivers is the same as the proportion of fatalities for crashes that don’t involve drunk drivers,
>
> *vs.*
>
> H_a: the proportion of fatalities for crashes that involve drunk drivers is different than the proportion of fatalities for crashes that don’t involve drunk drivers.

As usual, a high value of the χ^2 statistic, and a p-value lower than 0.05 suggest that there’s evidence to reject the null hypothesis in favor of the alternative, and that there's an association between drunk driving and crash fatalities.

1. To carry out the χ^2 test in R to test the hypothesis above, you would use the syntax CrossTable(mydata$FATAL_OR_M, mydata$DRINKING_D, prop.r=FALSE, prop.t=FALSE, prop.chisq=FALSE, chisq=TRUE). The χ^2 results will appear below the cross-tabulation table that you have seen in 2.b.i above. Report the results without the Yates Continuity Correction.

2. Modifying the syntax in 2.b.iv.1 above, run the χ^2 test examining the association between the dependent variable and the remaining binary predictors.

3. In the table 2.b.iii above, add another column called "χ^2 p-value". For each row, present the p-value from the corresponding χ^2 test.

    a. Note, however, that in practice, statisticians generally present not just the p-value, but also the value of the χ^2 statistic and the degrees of freedom, which is the parameter of the χ^2 distribution. The degrees of freedom is calculated as $(R-1)(C-1)$, where R is the number of rows in the cross-tabulation table and C is the number of columns in the cross-tabulation table. Said differently, R is the number of categories of the first variable and C is the number of categories in the second variable. Here, because we are cross-tabulating two binary variables, both R and C are 2, and $df = (R-1)(C-1) = 1$.

c. Now, let's examine whether the means of the two continuous predictors seem to differ for the different levels of the dependent variable. To do this, calculate the group means (and standard deviations) of both predictors (**PCTBACHMOR** and **MEDHHINC**) for crashes that involve drunk drivers and crashes that don't.

    i. In order to do this in R, use the tapply command. For instance, if you want to calculate the average values of the variable **PCTBACHMOR** for crashes that involve drunk drivers and crashes that don't, you would use the following syntax: tapply(mydata$PCTBACHMOR, mydata$DRINKING_D, mean). To calculate the standard deviations of the variable **PCTBACHMOR** for crashes that involve drunk drivers and crashes that don't, you would use the following syntax: tapply(mydata$PCTBACHMOR, mydata$DRINKING_D, sd)

    ii. Present your results in the table below:

	No Alcohol Involved		Alcohol Involved	
	(DRINKING_D = 0)		(DRINKING_D = 1)	
	Mean	SD	Mean	SD
PCTBACHMOR: % with bachelor's degree or more				
MEDHHINC: Median household income				

iii. Recall from introductory statistics classes that in order to compare the mean value of a continuous variable for two independent groups, statisticians usually employ a test that's called the independent samples t-test. For example, we can see whether the average **PCTBACHMOR** values are statistically significantly different for crashes that involve drunk drivers and crashes that don't. The null and alternative hypotheses for the independent samples t-test would be as follows:

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

vs.

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

A high value of the t-statistic, and a p-value lower than 0.05 suggest that there's evidence to reject the null hypothesis in favor of the alternative.

1. To carry out the t-test in R to test the hypothesis above, you would use the syntax t.test(mydata$PCTBACHMOR~mydata$DRINKING_D).

2. Repeat the t-test for the variable **MEDHHINC**.

3. In the table 2.c.ii above, add another column called "t-test p-value". For each row, present the p-value from the corresponding t-test.

    a. Note, however, that in practice, statisticians generally present not just the p-value, but also the value of the t-statistic and the degrees of freedom.

d. Using the instructions from Assignment 1, examine the Pearson correlations between all the predictors (both binary and continuous). Is there evidence of severe multicollinearity here? (Be sure the appropriate R library is loaded).

3) Now that we're done with exploratory analysis, we are ready to proceed to running the logistic regression.

   a. Using the glm command for logistic regression (as shown in the slides), regress the **DRINKING_D** variable on the following predictors: **FATAL_OR_M**, **OVERTURNED**, **CELL_PHONE**, **SPEEDING**, **AGGRESSIVE**, **DRIVER1617**, **DRIVER65PLUS**, **PCTBACHMOR**, and **MEDHHINC**.

      i. Use the summary command to examine the results.

      ii. For each predictor, also compute the odds ratio and the 95% confidence interval (CI) using commands shown in the slides. Use the syntax presented on the slide titled 'Merging Odds Ratios to β Coefficients in R' to merge the odds ratios and CIs to the matrix that contains coefficients and p-values. Present the resulting (merged) matrix in your report.

      iii. Using the syntax presented in the slides, calculate the sensitivity, specificity and misclassification rate for each of the following cut-off values: 0.02; 0.03; 0.05; 0.07; 0.08; 0.09; 0.10; 0.15; 0.20; 0.50 and present them in the table below. In the table, highlight the cut-off value which has the *lowest* misclassification rate.

Cut-off Value	**Sensitivity**	**Specificity**	**Misclassification Rate**
0.02			
0.03			
0.05			
0.07			
0.08			
0.09			
0.1			
0.15			
0.2			
0.5			

      iv. Using the syntax presented in the slides, generate the ROC curve, and identify the optimal cut-off value. Be sure to export the image of the ROC curve (as you will be expected to present it in your report).

      v. Using the syntax presented in the slides, calculate the area under the ROC curve.

b. Re-run the model without the **PCTBACHMOR** and **MEDHHINC** terms. As in 3.a.i and 3.a.ii above, use the summary command to examine results, and calculate the odds ratio and the 95% confidence interval for each predictor. As in 3.a.ii above, merge the odds ratios and confidence intervals to the matrix containing coefficients and p-values. Present the resulting (merged) matrix in your report.

c. Compare the two models using the Akaike Information Criterion (AIC). The results are typically presented at the bottom of the logistic regression output. They may also be obtained using the AIC command. For instance, if the results obtained from the glm command for the first model are saved as *mylogit1* and the results from the second model are saved as *mylogit2*, the R syntax to obtain the AICs from both models would be AIC(mylogit1, mylogit2). Here, recall that lower values of the AIC correspond to a better model.

Now, you are finally ready to start writing your report!

REPORT OUTLINE

A successful report will address *all* the points presented in this outline. You are strongly encouraged to use the outline as a backbone for your report.

The outline below structures your report as a journal article. That is, in the Methods section, only talk about the techniques that you use, present the formulas, etc. Do not present any results in the methods section. In the Results section, actually present the output from R, the tables that you created above, etc., and describe your output.

1) **Introduction (~2 paragraphs)** *Section Title*
   a) State the problem, the importance of the problem, and the setting of the analysis (Philadelphia).

   b) Speculate as to why the predictors we're using might be associated with the response variable.

   c) Indicate that you will be using R to run logistic regression for the analysis.

2) **Methods (~3-4 pages)** *Section Title*
   a) Explain the problems with using OLS regression when the dependent variable is binary

   b) Explain how logistic regression works around these issues.
      i. Prior to diving into logistic regression, introduce the concept of odds. Explain what an odds ratio is, and how it may be interpreted.
      ii. Present the regression equation for the logit model with multiple predictors.
         1. Present the regression equation in both the logit form (i.e., with $\ln\left(\frac{p}{1-p}\right)$ on the left-hand side of the equation) and the logistic form ($p = P(Y = 1)$ on the left-hand side of the equation).
            a. In the formulas in (a) above, instead of writing Y, $x_1 \dots x_k$, etc., write the names of the dependent and independent variables used in our actual model.
            b. Be sure to explain every term in each equation and talk about the logit and logistic functions (feel free to use the graphs in the slides as part of your explanation).
               i. Specifically, talk about the properties of the logistic function and explain why the logistic function works well for models where the dependent variable is binary.

   c) Describe the hypothesis that is tested for each predictor

    i. State the null and alternative hypotheses
    ii. Talk about the Wald statistic and the distribution that it follows
    iii. State that rather than looking at the estimated β coefficients, most statisticians prefer to look at odds ratios, which are calculated by exponentiating the coefficients.

d) Talk about how you assess the quality of model fit.
    i. Mention that an R-squared may be calculated for logistic regression, however it is no longer a very useful metric and doesn't have the same interpretation as in OLS.
    ii. Talk about the Akaike Information Criterion to compare models. In one sentence describe what the AIC is (if the information in the slides isn't sufficient, I suggest Googling it or looking at Wikipedia), and state whether a lower AIC is indicative of a better or worse fit.
    iii. Explain what is meant by specificity, sensitivity and the misclassification rate, and describe how each quantity is calculated. Are higher or lower values of each quantity better or worse?
        1. In your explanation, be sure to mention about how fitted (predicted) values of y, i.e., $\hat{y}$, are calculated and interpreted in logistic regression.
        2. Indicate why you should try using different cut-offs for what is considered a "high" probability of $Y = 1$ when calculating the specificity, sensitivity and the misclassification rate.
        3. Explain what the ROC curve is. Be sure to talk about some methods for calculating the optimal cut-off using the ROC curve, and specify which one you will be using in this report.
        4. Also, explain what we get by calculating the area under the ROC curve, and what might be value ranges for excellent, good, fair, poor and failing models.

e) Talk about the assumptions of logistic regression.
    i. Which assumptions of OLS regression also hold for logistic regression, and which don't?

f) Describe the exploratory analyses that statisticians may want to do before running logistic regression.
    i. Talk about running the cross-tabulations between the dependent variable and binary predictors to see whether there is an association between the two variables.
        1. Say that the appropriate statistical test for examining the association between two categorical variables is the Chi-Square test (described above on pp. 4-5 of this document).
            a. Be sure to mention the null and alternative hypotheses for the test.

ii. Also state that we can compare the means of continuous predictors for both values of the dependent variable.

1. Say that the independent samples t-tests (described above on pp. 5-6 of this document) are the appropriate statistical tests for examining whether there were significant differences in mean values of **PCTBACHMOR** and **MEDHHINC** for crashes that involved alcohol and those that didn't.

a. Mention the null and alternative hypotheses for the t-test.

3) Results (~3-5 pages, including figures & tables)

Section Title

a) Present and discuss the results of the exploratory analyses.

i. Present the tabulation of the dependent variable and comment on the number and proportion of crashes that involves drunk driving.

ii. Present the cross-tabulation of the dependent variable with each of the binary predictors (table on p. 4 above).

1. In the table, add (an) extra column(s) which presents the results of the Chi-Square test (you may present the p-value only, or the Chi-Square statistic, the degrees of freedom *and* the p-value).

2. Discuss whether the Chi-Square test shows that there is a significant association between the dependent variable and each of the binary predictors (i.e., can you reject the Null Hypothesis?).

iii. Present the means of the continuous predictors for both values of the dependent variable (table on p. 5).

1. In the table, add an extra column which presents the results of the independent samples t-test (you may present the p-value only, or the t-statistic, the degrees of freedom *and* the p-value).

2. Discuss whether the t-test shows that there is a significant association between the dependent variable and each of the continuous predictors (i.e., can you reject the Null Hypothesis?).

b) Comment on which logistic regression assumptions are met and which ones are violated for the problem at hand.

i. In particular, be sure to present the matrix showing the pairwise Pearson correlations for all the binary and continuous predictors.

1. Comment on any potential limitations of using Pearson correlations to measure the associations between binary predictors.

2. State whether there is evidence of multicollinearity, and remind the reader how you're defining multicollinearity.

c) Present the logistic regression results

i. First, present the results of the logistic regression with all predictors (**FATAL_OR_M**, **OVERTURNED**, **CELL_PHONE**, **SPEEDING**, **AGGRESSIVE**, **DRIVER1617**, **DRIVER65PLUS**, **PCTBACHMOR**, and **MEDHHINC**).

        1. Interpret the results. Be sure to comment on whether each predictor is significant, and interpret the odds ratio for each predictor.

    ii. Using the table on page 7, present the specificity, sensitivity and the misclassification rates for the different probability cut-offs. Indicate the cut-offs which yield the lowest/highest misclassification rates.

    iii. Present the ROC curve and comment on the optimal cut-off rate that was selected by minimizing the distance from the upper left corner of the ROC curve.

        1. Compare this cut-off rate with the optimal rate in 3.c.ii above

            a. Keep in mind that in 3.c.ii, we're looking at minimum mis-classification rates, and here, we're looking at simultaneously minimizing both sensitivity and specificity.

    iv. Also present and comment on the area under the ROC curve. What does it tell us about our model?

    v. Finally, present the results of the logistic regression with the binary predictors only (i.e., without **PCTBACHMOR** and **MEDHHINC**).

        1. Compare the results of this regression with the results of the first regression: are there any predictors which are significant in the new model which weren't significant in the original one, or vice versa?
        2. Be sure to also present the Akaike Information Criterion (AIC) for both models and indicate which model is better.

4) <u>Discussion (~1-2 paragraphs)</u> *<u>Section Title</u>*

a) In a couple sentences, recap what you did in the paper, and your findings.

    i. Which variables are strong predictors of crashes that involve drunk driving? Which variables aren't associated with the dependent variable?

    ii. Are the results surprising? Discuss, and mention whether the variables you expect to be significant actually are significant, and if so, whether the relationships with the dependent variable are in the direction you would expect.

    iii. Is logistic regression appropriate here? In other words, would the modeling rare events methods proposed by Paul Allison be more appropriate here?

        1. Hint: look at the # and % of cases with values of '1' for the dependent variable.

b) What are some limitations of the analysis? Discuss.