Lab 11 Activity

For this lab activity we will be working with the MplsStops dataset. This is a dataset about nearly all stops made by the Minneapolis Police Department for the year 2017. we will focus on the citationIssued variable, which indicates whether the stop resulted in a traffic violation report. Run the following code to load the data and delete rows with missing data:

```r
library(tidyverse)
dat <- carData::MplsStops %>%
        drop_na()
```

Here is a description of the columns in the data:

Variable	Description
idNum	Character vector of incident identifiers
date	A POSIXlt date variable giving the date and time of the stop
problem	A factor with levels suspicious for suspicious vehicle or person stops and traffic for traffic stops
MDC	A factor with levels mdc for data collected via in-vehicle computer, and other for data submitted by officers not in a vehicle, either on foot, bicycle or horseback. Several of the variables above were recorded only in-vehicle
citationIssued	A factor with levels no yes indicating if a citation was issued
personSearch	A factor with levels no yes indicating if the stopped person was searched
vehicleSearch	A factor with levels no or yes indicating if a vehicle was searched
preRace	A factor with levels white, black, east african, latino, native american, asian, other, unknown for the officer's assessment of race of the person stopped before speaking with the person stopped
race	A factor with levels white, black, east african, latino, native american, asian, other, unknown, officer's determination of race after the incident
gender	A factor with levels female, male, unknown, gender of person stopped
lat	Latitude of the location of the incident, somewhat rounded
long	Longitude of the location of the incident, somewhat rounded
policePrecinct	Minneapolis Police Precinct number
neighborhood	A factor with 84 levels giving the name of the Minneapolis neighborhood of the incident

1. We will try to predict what factors impact whether a traffic violation citation will be issued if individuals get stopped. Out of all the individuals stopped, how many were issued a citation? What is the percentage?

2. Run two logistic regressions, one using `personSearch` as a predictor of `citationIssued` and the other using both `personSearch` and `problem` as predictors of `citationIssued`. Conduct a likelihood ratio test to evaluate whether adding `problem` as a predictor significantly improves the logistic regression. Which regrresion is best according to the likelihood ratio test?

3. Compare the predictions of the logistic regression with just `personSearch` as a predictor and the predictions of the regression with both `personSearch` and `problem` as predictors.

- What is the confusion matrix for the two regressions?
- You may notice something "strange". Look at the predictions that the two regressions make; for both regressions, how many individuals are predicted to be given traffic violation citation?
- Is there any point in comparing accuracy, sensitivity, and specificity of the two regressions?

4. `personSearch` and `problem` are both factor variables with two levels; calculate the **predicted probability** of being issued a citation for a traffic violation given all the possible combinations of `personSearch` and `problem`. Make sure that you specify what combination of `personSearch` and `problem` the prediction is for. See some notes below for help on how to do this:

- Because we have factor variables as predictors, `R` uses dummy coding. See Slide 19 of Lab 9 if you are not sure on how to interpret the slopes of your logistic regression. The `contrasts()` function may help you to figure out how the variable is dummy coded.
- You can pass the logistic regression object to the `coef()` function to get the intercept and regression coefficients (so you don't have to use the `summary()` function and print unnecessary output).
- **Important:** to turn predictions of a logistic regression into probabilities you need to use the logistic function

$$\frac{e^x}{1+e^x}$$

Where x is the prediction of the logistic regression. The logistic function is `plogis()` in `R`. For example, if $x = 0$, then, the predicted probability is

$$\frac{e^0}{1 + e^0} = \frac{1}{1 + 1} = 0.5.$$

Equivalently,

```r
# logistic function in R
plogis(0)
```

```
[1] 0.5
```

5. You should have now calculated all the possible predictions that our logistic regression can make. Does this explain the results from question 3?