# DATA 621 - HW3

Andrew Bowen, Glen Davis, Shoshana Farber, Joshua Forster, Charles Ugiagbe

2023-10-23

## Homework 3 - Logistic Regression

### Overview

In this homework assignment, you will explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).

Your objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. You can only use the variables given to you (or, variables that you derive from the variables provided).

Below is a short description of the variables of interest in the data set:

| Column | Description |
|---|---|
| `zn` | proportion of residential land zoned for large lots (over 25000 square feet) (*predictor variable*) |
| `indus` | proportion of non-retail business acres per suburb (*predictor variable*) |
| `chas` | a dummy var. for whether the suburb borders the Charles River (1) or not (0) (*predictor variable*) |
| `nox` | nitrogen oxides concentration (parts per 10 million) (*predictor variable*) |
| `rm` | average number of rooms per dwelling (*predictor variable*) |
| `age` | proportion of owner-occupied units built prior to 1940 (*predictor variable*) |
| `dis` | weighted mean of distances to five Boston employment centers (*predictor variable*) |
| `rad` | index of accessibility to radial highways (*predictor variable*) |
| `tax` | full-value property-tax rate per $10,000 (*predictor variable*) |
| `ptratio` | pupil-teacher ratio by town (*predictor variable*) |
| `lstat` | lower status of the population (percent) (*predictor variable*) |
| `medv` | median value of owner-occupied homes in $1000s (*predictor variable*) |
| `target` | **whether the crime rate is above the median crime rate (1) or not (0) (*response variable*)** |

### Data Loading

Let's load in the training dataset.

```
train_df <- read.csv('https://raw.githubusercontent.com/ShanaFarber/businessAnalyticsDataMiningDATA621/m

head(train_df)  # preview data
```

```
##   zn indus chas   nox    rm   age    dis rad tax ptratio lstat medv target
## 1  0 19.58    0 0.605 7.929  96.2 2.0459   5 403    14.7  3.70 50.0      1
## 2  0 19.58    1 0.871 5.403 100.0 1.3216   5 403    14.7 26.82 13.4      1
## 3  0 18.10    0 0.740 6.485 100.0 1.9784  24 666    20.2 18.85 15.4      1
## 4 30  4.93    0 0.428 6.393   7.8 7.0355   6 300    16.6  5.19 23.7      0
## 5  0  2.46    0 0.488 7.155  92.2 2.7006   3 193    17.8  4.82 37.9      0
## 6  0  8.56    0 0.520 6.781  71.3 2.8561   5 384    20.9  7.67 26.5      0
```

### Data Exploration

```
dim(train_df)
```

```
## [1] 466  13
```

The dataset consists of 466 observations of 13 variables. There are 12 predictor variables and one response variable (`target`).

All of the columns in the dataset are numeric. Let's take a look at the summary statistics for the variables in the dataset.

```
summary(train_df)
```

```
##        zn              indus             chas              nox        
##  Min.   :  0.00   Min.   : 0.460   Min.   :0.00000   Min.   :0.3890  
##  1st Qu.:  0.00   1st Qu.: 5.145   1st Qu.:0.00000   1st Qu.:0.4480  
##  Median :  0.00   Median : 9.690   Median :0.00000   Median :0.5380  
##  Mean   : 11.58   Mean   :11.105   Mean   :0.07082   Mean   :0.5543  
##  3rd Qu.: 16.25   3rd Qu.:18.100   3rd Qu.:0.00000   3rd Qu.:0.6240  
##  Max.   :100.00   Max.   :27.740   Max.   :1.00000   Max.   :0.8710  
##        rm              age              dis              rad        
##  Min.   :3.863   Min.   :  2.90   Min.   : 1.130   Min.   : 1.00  
##  1st Qu.:5.887   1st Qu.: 43.88   1st Qu.: 2.101   1st Qu.: 4.00  
##  Median :6.210   Median : 77.15   Median : 3.191   Median : 5.00  
##  Mean   :6.291   Mean   : 68.37   Mean   : 3.796   Mean   : 9.53  
##  3rd Qu.:6.630   3rd Qu.: 94.10   3rd Qu.: 5.215   3rd Qu.:24.00  
##  Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.00  
##       tax           ptratio          lstat             medv      
##  Min.   :187.0   Min.   :12.6   Min.   : 1.730   Min.   : 5.00  
##  1st Qu.:281.0   1st Qu.:16.9   1st Qu.: 7.043   1st Qu.:17.02  
##  Median :334.5   Median :18.9   Median :11.350   Median :21.20  
##  Mean   :409.5   Mean   :18.4   Mean   :12.631   Mean   :22.59  
##  3rd Qu.:666.0   3rd Qu.:20.2   3rd Qu.:16.930   3rd Qu.:25.00  
##  Max.   :711.0   Max.   :22.0   Max.   :37.970   Max.   :50.00
```

```
##      target
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4914
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

We can see the mean, median, and interquartile ranges for each of the variables in the dataset.

There does not appear to be any NA values in the dataset. Let's validate this.

```
sum(is.na(train_df))
```

```
## [1] 0
```

There are no null values in the training dataset.

Let's take a look at the distributions for the predictor variables.

```
par(mfrow=c(3,4))
par(mai=c(.3,.3,.3,.3))

variables <- names(train_df)

for (i in 1:(length(variables)-1)) {
  hist(train_df[[variables[i]]], main = variables[i], col = "lightblue")
}
```