Lab 4 Logistic Regression

Problem Statement

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and rank (prestige of the undergraduate institution), affect admission into graduate school. The response variable, admit/do not admit, is a binary variable.

Dataset

The dataset is included in the package `aod`. Install the package and include package using the command `library(aod)`.

Using the following command to load the dataset

```
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
## view the first few rows of the data
head(mydata)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

[More on reading and writing CSV files, see here: https://swcarpentry.github.io/r-novice-inflammation/11-supp-read-write-csv/index.html]

This dataset has a binary response (outcome, dependent) variable called `admit`. There are three predictor variables: `gre`, `gpa` and `rank`. We will treat the variables `gre` and `gpa` as continuous. The variable `rank` takes on the values 1 through 4. Institutions with a `rank` of 1 have the highest prestige, while those with a `rank` of 4 have the lowest.

Questions

1) Get basic descriptives for the entire data set using `summary()`. View the dataset using `View()`.

2) How many observations are there in this dataset?

3) Get the standard deviations for the first three variables (i.e., `admit`, `gre` and `gpa`).

Hint: use `sapply` to apply the `sd` function to each variable in the dataset: `sapply(mydata, sd)`.

Now get the mean `admit`, `gre` and `gpa` in a similar way.

4) Convert `rank` to a factor to indicate that rank should be treated as a categorical variable. (Hint: use `factor()` function)

[More on factors, see the tutorial here: https://swcarpentry.github.io/r-novice-inflammation/12-supp-factors/index.html]

5) Estimate a logistic regression model using the `glm` function, and get the results using the `summary` command.

6) Do you notice variable `rank` is replaced with categorical variables `rank2`, `rank3`, and `rank4` that can only take values of 0 or 1? Recall that the original variable `rank` can take values of 1, 2, 3, or 4. Why isn't a variable `rank1` needed? If rank is 1, what are the values of `rank2`, `rank3` and `rank4`?

7) From the z-statistics and p-values of the variables, report which variables are statistically significant.

8) Use the model to predict the training dataset and store the results to a vector of probabilities `admit.prob`.

9) Create another vector `admit.pred` to show 0 or 1 for `admit.prob`. Let's set the value to be 0 if the probability is less than 0.5, and 1 if the probability is no less than 0.5.

10) Using `table()` function to create a confusion matrix to determines how many observations were correctly or incorrectly classified. Calculate the percentage that the observations were correctly classified.

11) Use the model to predict the average cases in each rank, that is, four new data with mean gre, mean gpa and rank from 1 to 4.