STA 214 Homework 1

Due: Friday, January 20, 12:00pm (noon) on Canvas.

Instructions: There are three parts to this assignment. Part I is a review of some exploratory data analysis from STA 112, Part II is practice with logistic regression models, and Part III is a simulation activity exploring linear vs. logistic regression for binary data. Parts I and II both use a dataset on admission to graduate programs.

Getting started: Begin by downloading the HW1 template from the course website:

`https://sta214-s23.github.io/homework/hw_01_template.Rmd`

Save this template file to your computer, then open it in RStudio. Now read the R Markdown instructions on the course website (`https://sta214-s23.github.io/resources/rmarkdown_instructions/`). As you complete the assignment, you will write down your answers to all questions in the R Markdown file, and include all R code in code chunks. *If a question requires code, you will not receive credit if no code is provided.*

Submission: When you have completed the assignment, knit your homework to HTML and submit on Canvas.

Data

Today we will be working with data on admission to graduate school. We have data on undergraduate students and the outcome of their application to graduate school. The variables are

- `admit`: whether or not the student was admitted (0 = no, 1 = yes)
- `gre`: the student's score on the GRE (graduate record exam)
- `GPA`: the student's grade point average
- `rank`: how prestigious the student's undergraduate institution is (1 = most prestigious, 2, 3, 4 = least prestigious)

Goal: Our goal is to use a student's GPA to predict the probability they are admitted to graduate school.

Loading the data

The data are available from the UCLA, and can be loaded into R with the following command:

```
grad_app <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
```

Here `read.csv` is a function that imports data from a CSV file. We can pass `read.csv` either a local path on our computer, or a URL – in this case, we use the URL where the data is stored online. We have called the data `grad_app` in R.

Copy the command to load the data into the setup chunk of your R Markdown file, and run it.

1 EDA review

In STA 112, you learned tools for *exploratory data analysis* (EDA), in which we explore features of the data such as the available variables and their relationships. Exploratory data analysis is an important step before we do any model fitting or hypothesis testing, because it gets us familiar with the data and any unusual features. It is hard to fit a sensible model when we don't know what the data look like!

A review of exploratory data analysis and R from STA 112 is provided in the codebook on the course website: `https://sta214-s23.github.io/resources/codebook.html`

1. Let's begin by examining the overall size of the data, and looking for any missing values.

   (a) How many rows and columns are in the `grad_app` data?

   (b) What does one row in the data represent?

   (c) Are there any rows with missing values in the `grad_app` data? If so, create a new dataset containing only the rows of `grad_app` with no missing values, and use this new dataset for the remainder of the assignment.

2. Now let's explore the variables of interest. We will start with the response variable: whether the student was admitted to graduate school. Since this is a categorical variable, we can summarize the distribution by the number of observations for each category.

   (a) How many students were admitted to graduate school? How many were not admitted?

   (b) Create a bar chart for `admit` displaying this information. Make sure to properly label the axes of the graph, and provide a title.

3. Next, let's look at `GPA`, our intended explanatory variable. We would like to describe and summarize the distribution of GPA in the observed data. Since GPA is quantitative, we can use a histogram to visualize the distribution, and summary statistics like the mean or median.

   (a) Create a histogram for `GPA`. Make sure to properly label the axes of the graph, and provide a title.

   (b) Calculate appropriate summary statistics for the center (choose between the mean and median) and spread (choose between the standard deviation and the IQR) of the distribution of GPA. Justify your choice of summary statistics.

   (c) Using your summary statistics from (b) and your histogram from (a), write 1–2 sentences summarizing the distribution of GPA. You should discuss the shape (skewness and modality), center, spread, and whether there are any potential outliers.

4. Finally, let's explore the relationship between GPA and whether the student was admitted to graduate school. Since GPA is quantitative and acceptance is binary, one option for visualizing this relationship is with side-by-side boxplots.

   (a) Create side-by-side boxplots with `admit` on the x-axis and `GPA` on the y-axis.

   (b) Calculate the mean and standard deviation of GPA within each group (admitted and not admitted).

   (c) Based on your answers to (a) and (b), does it seem like there is a relationship between GPA and acceptance to graduate school?

2 Logistic regression modeling

In this portion of the assignment, we will fit and use a logistic regression model. The lecture notes, class activities, and course codebook are good references for these questions.

5. Now that we have explored the relationship between GPA and acceptance to graduate school, we want to fit a regression model with GPA as the explanatory variable and acceptance as the response.

   (a) Explain why a linear regression model is *not* appropriate here.

   (b) Write down the population logistic regression model. Make sure to include both parts of the parametric model (the random and the systematic components). Use proper notation and include all subscripts. (See "writing math in R Markdown" in the course codebook, `https://sta214-s23.github.io/resources/codebook.html`)

   (c) Explain why there is no noise term ε_i in your logistic regression model.

   (d) In R, fit your logistic regression model from (b), and report the fitted coefficients.

   (e) Interpret the estimated slope and intercept in terms of the log odds.

   (f) Interpret the estimated slope and intercept in terms of the odds.

6. From question 5, our fitted model is

$$Admit_i \sim Bernoulli(\pi_i)$$
$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -4.36 + 1.05\ GPA_i$$

   (a) What is the predicted probability of acceptance for a student with a 3.8 GPA?

   (b) What is the minimum GPA required to have an estimated probability of acceptance of at least 0.2?

3 Simulating data

(This part of the assignment does not use the grad admissions data)

In class, we learned that a linear regression model is not appropriate for a binary response. In this activity, we will explore some of the issues that arise when we apply linear regression to binary data.

So that we can control everything about the data – including the relationship between the variables – in this section we will create our **own** data. The process of generating our own data is called **simulation**, and we will do it in R.

7. We will simulate a quantitative explanatory variable X_i and a binary response Y_i, with the following relationship:

$$Y_i \sim Bernoulli(\pi_i)$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -3 + 2X_i$$

   (that is, $\beta_0 = -3$ and $\beta_1 = 2$).

(a) First, we need to generate our explanatory variable X_i. We could pick any distribution we want. For simplicity, we will make X_i randomly sampled between 0 and 5. Run the following code in R:

```
x <- runif(200, 0, 5)
```

(This generates 200 observations $X_1, ..., X_{200}$, sampled *uniformly* between 0 and 5.)

(b) Next, we need to calculate π_i for each explanatory variable. Here, transforming the log odds gives us

$$\pi_i = \frac{\exp\{-3 + 2X_i\}}{1 + \exp\{-3 + 2X_i\}}$$

In R, run the following code:

```
pis <- exp(-3 + 2*x)/(1 + exp(-3 + 2*x))
```

(This calculates π_i for each X_i.)

(c) Now that we have π_i for each observation, we can generate Y_i! The function to generate samples from a Bernoulli distribution in R is the `rbinom()` function. Run the following code in R:

```
y <- rbinom(200, 1, pis)
```

(This generates 200 observations $Y_1, ..., Y_{200}$, where each $Y_i \sim Bernoulli(\pi_i)$.)

(d) Finally, let's use our simulated data to fit a linear regression model. Run the following code:

```
m1 <- lm(y ~ x)
summary(m1)
```

Are the estimated coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ close to the true values (-3 and 2)?

(e) For comparison, fit a logistic regression model to the simulated data:

```
m2 <- glm(y ~ x, family = binomial)
summary(m2)
```

Does the logistic regression model better estimate the regression coefficients?

(f) One issue with applying linear regression to a binary response is that we may get predictions outside the range $[0, 1]$. Using your fitted linear regression model from (d), for what values of X do we see $\widehat{Y} > 1$?