Jayoung Kang

Problem Set 2

1. *Chapter 4*
   a. *Question 6: Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, β^0 = −6, β^1 = 0.05, β^2 = 1.*
      i. *Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class*

-6 + 40*0.05 + 1*3.5 = -0.5

$$\frac{\exp(-6 + 40 * 0.05 + 1 * 3.5)}{1 + \exp(-6 + 40 * 0.05 + 1 * 3.5)} = \mathbf{37.75\%}$$

      ii. *How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?*

$$\frac{\exp(-6 + x * 0.05 + 1 * 3.5)}{1 + \exp(-6 + x * 0.05 + 1 * 3.5)} = 50\%$$
$$\exp(-2.5 + 0.05x) = 0.5(1 + \exp(-2.5 + 0.05x))$$
$$0.5(\exp(-2.5 + 0.05x)) = 0.5$$
$$\exp(-2.5 + 0.05x) = 1$$
$$-2.5 + 0.05x = \log(1)$$
$$0.05x = 2.5$$

$$\boldsymbol{x = 50 \text{ hours}}$$

   b. *Question 7: Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was X¯ = 10, while the mean for those that didn't was X¯ = 0. In addition, the variance of X for these two sets of companies was ^σ2 = 36. Finally, 80 % of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year.*

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}.$$

$$\pi_k = probability\ of\ 'Yes'$$

$$p_k(x) = posterior\ probability\ that\ an\ observation\ X = x\ belongs\ to\ class\ 'Yes'$$

$$\frac{80\% * \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu_{yes}\right)^2\right)}{20\% * \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_{no})^2\right) + 80\% * \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\mu_{yes}\right)^2\right)}$$

$$\frac{80\% * \frac{1}{6\sqrt{2\pi}} \exp\left(-\frac{1}{2*36}(4-10)^2\right)}{20\% * \frac{1}{6\sqrt{2\pi}} \exp\left(-\frac{1}{2*36}(4-0)^2\right) + 80\% * \frac{1}{6\sqrt{2\pi}} \exp\left(-\frac{1}{2*36}(4-10)^2\right)} \; 0.7519 = \mathbf{75.19\%}$$

c. *Question 9*

i. *On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?*

$$\frac{\text{probability}}{1-\text{probability}} = 0.37$$
$$prob = 37\%(1-prob)$$
$$137\% prob = 37\%$$
$$prob = \frac{37\%}{137\%} \approx \mathbf{27.01\%}$$

ii. *Suppose that an individual has a 16 % chance of defaulting on her credit card payment. What are the odds that she will default?*

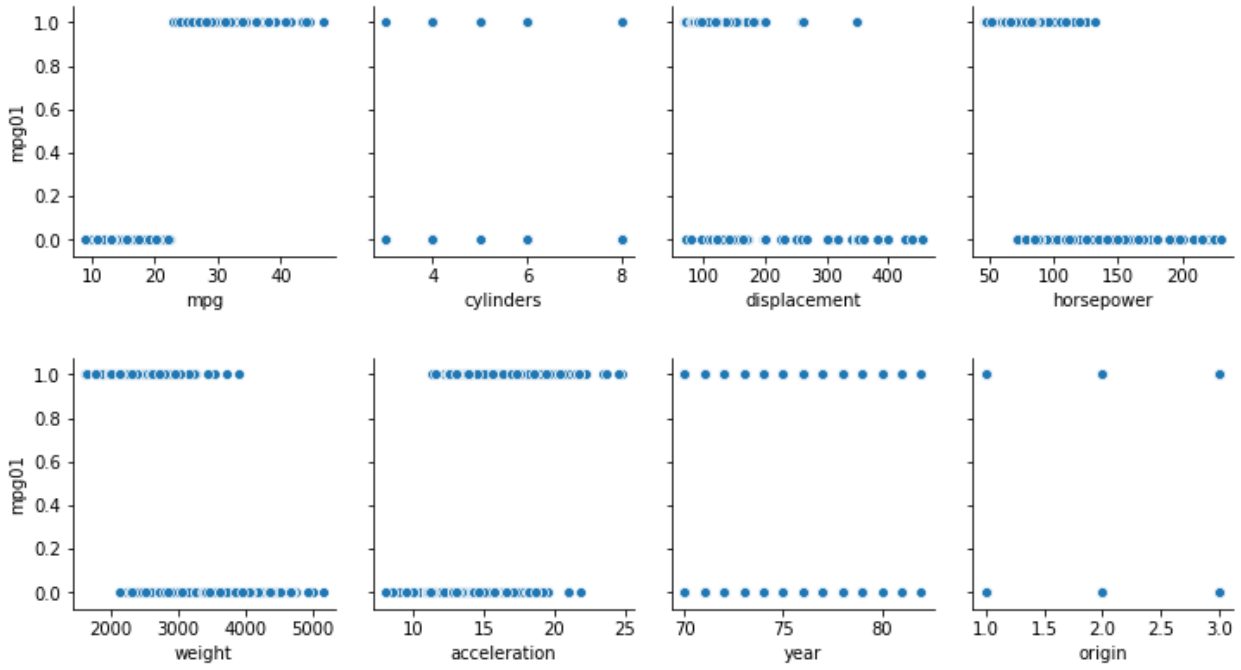
$$\frac{16\%}{1-16\%} \approx \mathbf{0.1905}$$

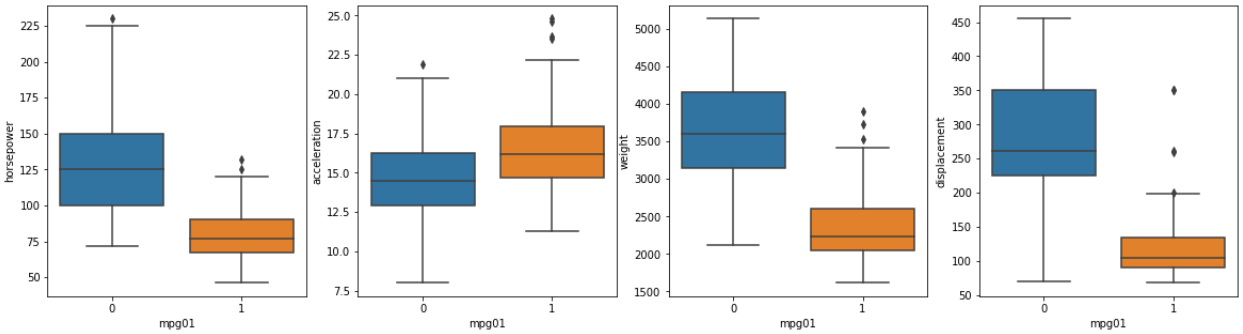
d. *Question 11 a) ~ f) In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.*

i. *Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.*

ii. *Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.*

- Based on the pairwise scatter plot, the variable mpg01 seems to have a positive relationship with mpg and acceleration. However, the positive relationship with mpg is expected and not meaningful because mpg01 is coded from mpg. We can further see the possibility of a positive relationship through the boxplot as we see that the conditional mean of mpg01 increases with acceleration.
- Based on the pairwise scatter plot, mpg01 seems to have a negative relationship with displacement, horsepower and weight. The boxplots also show that the conditional mean of mpg01 decreases with these variables. However the relationship with displacement doesn't seem to be as clearly visible as the other variables.





iii. Split the data into a training set and a test set.

iv. Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

- 8.47%

v. Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

- 11.02%

vi. Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

- 10.17%

2. **Chapter 5**
    a. *Question 5: In Chapter 4, we used logistic regression to predict the probability of default using income and balance on the Default data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.*
        i. *Fit a logistic regression model that uses income and balance to predict default.*
        ii. *Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:*
            1. *Split the sample set into a training set and a validation set.*
            2. *Fit a multiple logistic regression model using only the training observations.*
            3. *Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.*
            4. *Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.*
        iii. *Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.*
            - The 3 error rates are: 3.6%, 2.87%, 2.57%
            - According to the ISLR textbook, the validation set approach can overestimate the test error rate when compared to cross validation because only a subset of the observations are used to fit the model. However, the results we received here seem relatively low and they vary little. The low variation could be because of the relatively large sample size we have to train on.
            - The test errors help validate the logistic model and the low values show that the logistic model shows high accuracy for predicting default.

        iv. *Now consider a logistic regression model that predicts the probability of default using income, balance, and a dummy variable for student. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.*
            - The test error rate obtained: 3.37%
            - These test error rate is very similar to the ones obtained from the previous question. The average of the 3 error rates without the student dummy variable is 3.01% which is not that different from 3.37%.
            - Therefore, it doesn't seem that the addition of the student dummy variable leads to a reduction in the test error rate.