STAT 632, HW 7
Due: Tuesday, May 5

Reading: Section 4.1–4.3 (pp. 127–137) from *An Introduction to Statistical Learning*

The following exercises will use a data set that contains results for US counties from the 2012/16 US presidential elections. Demographic information on counties from the US Census is also provided.[1]

To load the data into R run the following command:

```
county_votes16 <- read.csv("https://ericwfox.github.io/data/county_votes16.csv")
```

Descriptions of relevant variables:

- `trump_win`: indicator variable (1=Trump won, 0=Trump lost)
- `obama_pctvotes`: percent of votes cast for Obama in 2012
- `pct_pop65`: percent over 65 years
- `pct_black`: percent black
- `pct_white`: percent white
- `pct_hispanic`: percent hispanic
- `pct_asian`: percent asian
- `highschool`: percent high school graduate or higher
- `bachelors`: percent with Bachelor's degree or higher
- `income`: per capita income in the past 12 months (in thousands of dollars)

Exercise 1

(a) Fit a simple logistic regression model with `trump_win` as the binary response variable, and `obama_pctvotes` as the predictor. Use `summary()` to print the results, and write down the equation for the estimated logistic regression model. Use this logistic regression model to answer the remaining questions.

(b) Make a scatter plot of the data (i.e., plot the observed zeros and ones on the y-axis and `obama_pctvotes` on the x-axis) and superimpose the fitted logistic curve for the estimated probability of Trump winning. Use `ggplot2` to make the plot.

(c) Use the logistic regression model to estimate the probability of Trump winning in a county with `obama_pctvotes` $= 40, 50$, and 60.

(d) Provide an interpretation of the estimated coefficient $\hat{\beta}_1$ for `obama_pctvotes`.

[1] Source: `https://www.kaggle.com/joelwilson/2012-2016-presidential-elections#county_facts_dictionary.csv`

Exercise 2

(a) Fit a multiple logistic regression model with `trump_win` as the response, and the following 8 demographic variables as predictors: `pct_pop65`, `pct_black`, `pct_white`, `pct_hispanic`, `pct_asian`, `highschool`, `bachelors`, and `income`. Use `summary()` to print the results.

(b) Remove any predictors that are not significant from the model fit in (a).

(c) Provide an interpretation of the signs of the estimated coefficients.

Additional practice on logistic regression (not to be collected).

Practice Problem 1.[2] Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression model and produce estimated coefficients, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, and $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

(b) How many hours would a student with a 3.5 GPA need to study to have a 50% chance of getting an A in the class?

Practice Problem 2. The parameters β_0 and β_1 for the simple logistic regression model can be estimated using the method of maximum likelihood. There are actually no closed form solutions for the parameter estimates, so iterative techniques are used to perform the optimization (e.g., gradient descent). The end of lecture 17 discusses how the likelihood function for logistic regression can be expressed as

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}$$

where $p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$ for $i = 1, \cdots, n$.

(a) Show that the log-likelihood function can be expressed as $l(\beta_0, \beta_1) = \log(L(\beta_0, \beta_1)) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$

(b) Show that $\frac{\partial l(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - p_i)$ and $\frac{\partial l(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} x_i (y_i - p_i)$.

[2] From *An Introduction to Statistical Learning*, Ch. 4, Exercise 6