Homework 4 (Due on November 15th midnight)

Total marks 40

1. Suppose $(y_1, y_2, \dots, y_n)$ is a random sample from a Bernoulli distribution where $Y \sim Bern(p)$ with probability mass function, pmf of Y

$$f(y_i|p) = p^{y_i}(1-p)^{1-y_i};\ y_i \in \{0,1; 1 = success\}$$

   i. Find maximum likelihood estimate of p. **(5 marks)**
   ii. In five independent Bernoulli trials from above Bernoulli process, three successes and two failures were observed. Calculate the maximum likelihood estimates of p in this situation? **(5 marks)**
   iii. Plot the log likelihood function for the data in part ii by performing a grid search over a set of possible values of p parameter. Add a vertical line to the plot at the value of p that maximizes the log-likelihood. **(3 marks)**

2. Consider the case of simple logistic regression where Y is the binary dependent variable and X is the predictor variable with sample data $(x_1, y_1), (x_2, y_2), \dots., (x_n, y_n)$. Binary outcomes are modeled as Bernoulli trials as in Question 1 above. Here

$$f(y_i|p_i) = {p_i}^{y_i}(1-p_i)^{1-y_i};\ y_i \in \{0,1; 1 = Yes\}$$

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

   I. Derive the log-likelihood function, $L(\beta)$ where $\beta = (\beta_0, \beta_1)$. **(3 marks)**
   II. Write a function, ll(), to calculate the log likelihood. **(3 marks)**
   III. Using Default data in the book, ISLR, fit a logistic regression model to predict default given balance as a predictor using glm(). **(3 marks)**
   IV. Use optim() function together with your ll() and initial parameter estimates as zero to calculate maximum likelihood estimates of regression coefficients in part iii. **(3 marks)**
   V. Comment on the maximum likelihood estimates obtained using your work and using glm() in part iii. **(1 mark)**
   VI. Calculate the standard errors of your estimates. Hint: Include the parameter option 'hessian = TRUE' in the function optim() when you call optim() in part iv. **(3 marks)**
   VII. Comment on the standard error estimates obtained using your work and using glm() in part iii. **(1 mark)**

Chapter 5, Q6 on page 199

3. We continue to consider the use of a logistic regression model to predict the probability of default using income and balance on the Default data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed to 100 before beginning your analysis.

    I. Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors. **(2 marks)**
    II. Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model. **(3 marks)**
    III. Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance. **(3 marks)**
    IV. Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function. **(2 marks)**