UMASS Amherst

Biostats 690P: Topics in Health Data Science Spring 2021

Homework 1

Due: Friday, February 19, 2021

(90 Points in total)

1. (15 points) If $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2S}$ are independent random sample of size n_1 and n_2 from normal populations with the means μ_1 and μ_2 and the common variance σ^2, find maximum likelihood estimators for μ_1, μ_2, and σ^2.

2. (20 points) Suppose that $X_1, X_2, \ldots, X_n$ are independent random sample from a Poisson distribution, $X_i \sim \text{Poi}(\lambda)$.

   (a) Find the maximum likelihood estimator, $\widehat{\lambda}$, of λ based on the data $X_1, \ldots, X_n$.

   (b) What is the Fisher information $I(\lambda)$ of the random sample $X_1, X_2, \ldots, X_n$?

3. (25 points) Group testing (pooled testing) has been used to reduce costs in infectious disease screening, since Dorfman's pioneering work to estimate the prevalence of syphilis in World War II soldiers for syphilis in 1943. Suppose that blood samples from nk subjects are analyzed to learn about θ, the proportion of the population infected with SARS-CoV-2. In order to save costs, n pooled sampled are formed, where each pooled sample consists of the blood from k individuals mixed together. The blood analysis results are $Y_1, \ldots, Y_n$, where

$$Y_i = \begin{cases} 1, & \text{if all the } k \text{ individuals are free of disease;} \\ 0, & \text{otherwise.} \end{cases}$$

   (a) Find the probability mass function for Y_i.

   [Hint: the individuals follow Bernoulli distribution with the probability θ to get infected of the disease.]

   (b) Find the maximum likelihood estimator, $\widehat{\theta}$, of θ based on the data $Y_1, \ldots, Y_n$.

   (c) In part (a), Y_i has a probability distribution with parameter λ, where λ is a function of θ, i.e., $\lambda = g(\theta)$. We will use an alternative way to find the maximum likelihood estimator for θ. Find the maximum likelihood estimator, $\widehat{\lambda}$, of λ based on the data $Y_1, \ldots, Y_n$. Next, write out the expression of the function $g(\cdot)$, and then its inverse function $g^{-1}(\cdot)$. Based on the invariance property of maximum likelihood estimator (MLE), what is the MLE for θ?

   [**Hint:** See page 6 of the handout "examples-for-lec1.pdf" for the invariance property of maximum likelihood estimator.]

4. (30 points) **Implementation of Newton's method on a logistic regression model in R software.** Investigators are interested in assessing the association between the baseline cognitive dysfunction and the risk of Alzheimer's disease. In the `AD` data set (AD.csv file), the investigators collected a random sample of $n = 400$ participants of an Alzheimer's disease study. The variables in the data set include standardized baseline cognitive dysfunction measurement (the column named `x`) and the diagnosis of Alzheimer's disease (the column named `y`). As a statistical consultant, you would like to use a logistic regression model to analyze the data, where the response variable Y is *the diagnosis of Alzheimer's disease* (1=diagnosis of Alzheimer's disease, 0=no Alzheimer's disease), and the independent variable X is the *standardized baseline cognitive dysfunction measurement* (the higher, the worse).

(a) Summarize the two variables in the data set based on commonly used summary statistics. What is the proportion of Alzheimer's disease p_0 in this sample?

(b) Write out the logistic regression model for this data set, using α and β to represent the intercept and the regression coefficient of X.

(c) Choose initial estimate for α and β as $\widehat{\alpha}^{(0)} = \log\{p_0/(1-p_0)\}$ and $\widehat{\beta}^{(0)} = 0$, where p_0 is defined in part (a). Use Newton method to find the maximum likelihood estimates $\widehat{\alpha}$ and $\widehat{\beta}$ for α and β. Set the stopping rules for iteration as

$$\max\left(|\widehat{\alpha}^{(\nu+1)} - \widehat{\alpha}^{(\nu)}|, |\widehat{\beta}^{(\nu+1)} - \widehat{\beta}^{(\nu)}|\right) < 10^{-7}.$$

(i) Report the sequence of estimates $\{(\widehat{\alpha}^{(\nu)}, \widehat{\beta}^{(\nu)}), \nu = 0, 1, 2, \ldots\}$ from the Newton method.

(ii) How many iterations did you see for Newton method until the algorithm stop?

(iii) What are the values of maximum likelihood estimates (MLEs) $\widehat{\alpha}$ and$\widehat{\beta}$ you obtain?

(iv) Based on your fitted logistic regression model, estimate the probability of having Alzheimer's disease for a patient who has standardized baseline cognitive dysfunction measurement of 1.0.

[**Hint:** Refer to slides 16-17 of lecture 3. In the R code file, I wrote the R function '`score.fun()`' for the likelihood score function and observed information under logistic regression model with one predictor. You may write a loop using R to update the estimates $(\widehat{\alpha}^{(\nu)}, \widehat{\beta}^{(\nu)})$. You can use `score.fun()` function that I provided. Or you can write your own function to calculate the likelihood score and observed information if you want.]

(d) What is the observed information (matrix) at the MLE $\widehat{\alpha}$ and $\widehat{\beta}$? The inverse of the observed information matrix at the MLE will give you the estimate of variance-covariance between $\widehat{\alpha}$ and $\widehat{\beta}$. What are the standard errors for $\widehat{\alpha}$ and $\widehat{\beta}$?

[**Hint:** use `solve()` function in R to calculate the inverse of a matrix.]

(e) Now, Use the `glm()` function in R to fit a logistic regression model on Y with respect to X in the `AD` data set. Report the output of modeling fit. What are the estimates for α and β you get? What are the corresponding standard errors from the output? How are these estimation results compared to the results you got from your own Newton algorithm in parts (c) and (d)?

Reminder: Please submit your R script for Question 4 as well when you submit your HW1.