STA 711 Homework 3

Due: Friday, February 3, 12:00pm (noon) on Canvas.

Instructions: Submit your work as a single PDF. For this assignment, you may include written work by scanning it and incorporating it into the PDF. Include all R code needed to reproduce your results in your submission.

Maximum likelihood estimation

1. Let $X_1, ..., X_n$ be an iid sample from a distribution with pdf

$$f(x|\lambda, \sigma) = \frac{\sigma^{1/\lambda}}{\lambda} \exp\left\{-\left(1 + \frac{1}{\lambda}\right)\log(x)\right\} \mathbb{1}\{x \geq \sigma\},$$

where $\lambda, \sigma > 0$. Find the maximum likelihood estimates of λ and σ. *(Hint: find $\widehat{\sigma}$ first)*

Score and information

2. Let $X_1, ..., X_n \overset{iid}{\sim} Poisson(\lambda)$. Find the score function $\mathcal{U}(\lambda)$ and the Fisher information $\mathcal{I}(\lambda)$.

3. Consider a clinical trial to compare two treatments. n_1 subjects are given treatment 1, and n_2 subjects are given treatment 2. Let X_1 be the number of people on treatment 1 who respond favorably, and X_2 the number of people on treatment 2 who respond favorably. Assume that $X_1 \sim Binomial(n_1, p_1)$ and $X_2 \sim Binomial(n_2, p_2)$. The quantity of interest is the difference in the two treatments: $\psi = p_1 - p_2$.

   (a) Find the maximum likelihood estimate $\widehat{\psi}$ for ψ.

   (b) Since we have *two* parameters, p_1 and p_2, Fisher information is no longer a scalar. Instead, $\mathcal{I}(p_1, p_2)$ is a 2×2 matrix. By definition, the i, j entry of this Fisher information matrix is

$$[\mathcal{I}(p_1, p_2)]_{ij} = \mathbb{E}\left[\left(\frac{\partial}{\partial p_i}\ell(p_1, p_2|X_1^n)\right)\left(\frac{\partial}{\partial p_j}\ell(p_1, p_2|X_1^n)\right)\right].$$

   Use this definition to find $\mathcal{I}(p_1, p_2)$.

   (c) The definition in part (b) is often a clunky way to calculate Fisher information. Under appropriate regularity conditions, it can be shown that the Fisher information is also

$$[\mathcal{I}(p_1, p_2)]_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial p_i \partial p_j}\ell(p_1, p_2)\right].$$

   Confirm that this second method of calculating $\mathcal{I}(p_1, p_2)$ gives the same answer as in part (b).

   (d) A sufficient condition for the formula in part (c) is given in Lemma 7.3.11 of Casella & Berger, which essentially requires that we can differentiate under the integral sign. Read Section 2.4 of Casella & Berger (particularly Theorem 2.4.2), on rules for differentiating under the integral sign. Then explain why the regularity conditions hold for this problem.

Fisher scoring problems

In class, we learned how to use Fisher scoring to fit a logistic regression model. Recall that the Fisher scoring algorithm estimates the parameters β of a model as follows:

- Start with an initial guess $\beta^{(0)}$
- Update the estimate: $\beta^{(r+1)} = \beta^{(r)} + \mathcal{I}^{-1}(\beta^{(r)})\mathcal{U}(\beta^{(r)})$
- Stop when $\beta^{(r+1)} \approx \beta^{(r)}$

The purpose of these questions is to practice with Fisher scoring.

4. In class, we derived the score $\mathcal{U}(\beta)$ and the information matrix $\mathcal{I}(\beta)$ for logistic regression in the case of a *single* explanatory variable. What happens when we have multiple explanatory variables?

   Suppose that

$$Y_i \sim Bernoulli(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}$$

   We can write the systematic component more concisely as $\log\left(\frac{p_i}{1-p_i}\right) = \beta^T X_i$, where $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ and $X_i = (1, X_{i,1}, ..., X_{i,k})^T$ are $k+1$-dimensional vectors.

   (a) Show that $\mathcal{U}(\beta) = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}\right) X_i$

   (b) Show that $\mathcal{I}(\beta) = \sum_{i=1}^{n} \frac{e^{\beta^T X_i}}{(1+e^{\beta^T X_i})^2} X_i X_i^T$

   Hints: There are a couple different ways to approach this problem. It is probably cleanest to use rules for matrix calculus; that is, what it means to take derivatives when vectors and matrices are involved.

   Remember that $\mathcal{U}(\beta) = \frac{\partial \ell(\beta)}{\partial \beta}$ and $\mathcal{J}(\beta) = -\frac{\partial \mathcal{U}(\beta)}{\partial \beta}$, where $\ell(\beta)$ is the log-likelihood.

   Rules for matrix calculus can be found in the Matrix Cookbook `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf` and in Wikipedia's article on matrix calculus `https://en.wikipedia.org/wiki/Matrix_calculus`. The following rules are particularly helpful:

   - If $\mathbf{x}$ is a vector, $g(\mathbf{x}) \in \mathbb{R}$, and $f : \mathbb{R} \to \mathbb{R}$, then $\frac{\partial f(g(\mathbf{x}))}{\partial \mathbf{x}} = f'(g(\mathbf{x}))\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}$
   - If $\mathbf{x}$ and $\mathbf{a}$ are both vectors, then $\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$
   - If $\mathbf{x}$ and $\mathbf{a}$ are both vectors, and $g(\mathbf{x}) \in \mathbb{R}$, then $\frac{\partial g(\mathbf{x})\mathbf{a}}{\partial \mathbf{x}} = \left(\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}\right)\mathbf{a}^T$

5. In this problem, we will work with the dengue data we discussed in class. A CSV containing the data can be downloaded in R by running

```
dengue <- read.csv("https://sta711-s22.github.io/homework/dengue.csv")
```

For this problem, we are interested in modeling the relationship between platelet count and dengue fever. Let PLT_i denote the platelet count of patient i, and Y_i denote their dengue status (0 = negative, 1 = positive). Our logistic regression model is

$$Y_i \sim Bernoulli(p_i)$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 PLT_i$$

(a) Fit this logistic regression model in R, and report the estimated coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

(b) In R, write a function `U` which calculates $U(\beta)$ using the `dengue` data. For example, if $\beta = (1.8, 0)^T$ then your function should produce the following:

```
U(c(1.8, 0))
[1]   -3211.612 -820195.802
```

(c) In R, write a function `I` which calculates $\mathcal{I}(\beta)$ using the `dengue` data. For example, if $\beta = (1.8, 0)^T$ then your function should produce the following:

```
> I(c(1.8, 0))
            [,1]        [,2]
[1,]    696.2918    161214.3
[2,] 161214.2603 41783775.1
```

(d) Suppose that we use Fisher scoring to estimate β, and our current estimate is $\beta^{(r)} = (1.8, 0)^T$. Calculate the updated estimate $\beta^{(r+1)}$.

(e) Use your code from (b) and (c) to write code which implements Fisher scoring until convergence, beginning with $\beta^{(0)} = (1.8, 0)^T$. For the purpose of this question, stop when

$$\max\{|\beta_0^{(r+1)} - \beta_0^{(r)}|,\ |\beta_1^{(r+1)} - \beta_1^{(r)}|\} < 0.0001$$

Does your final estimate match the estimated coefficients in (a)? How many scoring iterations did it take to converge?

(f) Modify your code from (e) to implement gradient ascent instead of Fisher scoring. Use a learning rate (step size) of $\gamma = 0.0000001$, begin with $\beta^{(0)} = (1.8, 0)^T$, and run for 5000 iterations (do not run until convergence!). Report the estimated coefficients after 5000 steps. Why do you think Fisher scoring performs better here than gradient ascent?