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 $Y_1, ..., Y_n$ be an iid sample from a distribution with pdf

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

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

Score and information

2. Let $Y_1, ..., Y_n \overset{iid}{\sim} Poisson(\lambda)$.
   (a) Find the score function $U(\lambda)$.
   (b) Calculate the Fisher information $\mathcal{I}(\lambda)$ using $Var(U(\lambda))$.
   (c) Calculate the Fisher information $\mathcal{I}(\lambda)$ using $-\mathbb{E}\left[\frac{d^2}{d\lambda^2} \ell(\lambda|\mathbf{Y})\right]$ (the required regularity conditions hold in this example).

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 Y_1 be the number of people on treatment 1 who respond favorably, and Y_2 the number of people on treatment 2 who respond favorably. Assume that $Y_1 \sim Binomial(n_1, p_1)$ and $Y_2 \sim Binomial(n_2, p_2)$. 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

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

   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

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

   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: $\beta^{(r+1)} = \beta^{(r)} + \mathcal{I}^{-1}(\beta^{(r)})U(\beta^{(r)})$
- Stop when $\beta^{(r+1)} \approx \beta^{(r)}$

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

4. 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-s23.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?

5. One alternative to Fisher scoring is *gradient ascent*, variations of which are often used to fit complicated machine learning models for which it is challenging to calculate the Hessian / Fisher information. Rather than the Fisher information, gradient ascent uses a *learning rate* (or *step size*) $\gamma > 0$ to update coefficient estimates.

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

   (a) Modify your code from 4(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?

6. So far, we have applied Fisher scoring to estimate parameters in logistic regression models. How does this relate to estimation for *linear* regression models?

   Consider the model

$$
\begin{aligned}
Y_i &\sim N(\mu_i, \sigma^2) \\
\mu_i &= \beta^T X_i
\end{aligned}
$$

   where $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ and $X_i = (1, X_{i,1}, ..., X_{i,k})^T$. Suppose we observe data $(X_1, Y_1), ..., (X_n, Y_n)$, and we want to estimate β.

   (a) Write down the log likelihood function $\ell(\beta | \mathbf{X}, \mathbf{Y})$.

   (b) Show that the score function, in matrix form, is given by

$$
U(\beta) = \frac{1}{\sigma^2} \mathbf{X}^T (\mathbf{Y} - \mu),
$$

   where $\mu = \mathbf{X}\beta$.

   (c) Set the score equal to 0 and solve for β to get

$$
\widehat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}
$$

   (d) Show that the Hessian of the log likelihood, in matrix form, is given by

$$
H(\beta) = -\frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X}
$$

   (e) As we can see from (c), for *linear* regression we can get a closed form for $\widehat{\beta}$. But for the sake of comparison with logistic regression, let's suppose instead that we use Fisher scoring. Let $\beta^{(0)}$ be *any* initial estimate of β. Show that the result from a single iteration of Fisher scoring is

$$
\beta^{(1)} = \widehat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}
$$

   (in other words, Fisher scoring converges in a single step).