STA 712 Homework 1

Due: Tuesday, September 6, 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.

Probability and inference review questions

The purpose of these questions is to review some key concepts which will be useful in STA 712.

1. The Gamma distribution for random variable Y has probability density function

$$f(y) = \frac{1}{\Gamma(k)\theta^k} y^{k-1} e^{-\frac{y}{\theta}}$$

where $k > 0$ is the shape parameter, $\theta > 0$ is the scale parameter, and

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$$

is the Gamma function evaluated at k.

   (a) In `R`, make a single plot showing the pdf of the Gamma distribution for a few different combinations of k and θ. Be sure to add a legend to your plot, use different line types/colors, and make everything legible.

   (b) Derive $E(Y) = k\theta$.

   (c) Derive $Var(Y) = k\theta^2$.

   (d) Suppose $Y_1, ..., Y_n$ are independent, identically distributed Gamma($k = 1/2, \theta = 2$) random variables. What distribution does $\sum Y_i$ follow? Prove the result using moment generating functions. What is the expected value and variance of this distribution?

2. Suppose $Y_1, ..., Y_n$ are an i.i.d. sample drawn from a Bernoulli(p) distribution.

   (a) Derive the maximum likelihood estimate of p, observed information $\mathcal{J}(p)$, and the Fisher information $\mathcal{I}(p)$.

   (b) Make three separate plots in `R` showing the likelihood function $L(p)$, the log-likelihood function $\ell(p)$, and the score function $U(p)$ for $p \in (0, 1)$. Do this for two cases: $n = 10$ and $\sum y_i = 8$, and $n = 100$ and $\sum y_i = 80$. Compute the MLE, $\mathcal{J}(p)$, and $\mathcal{I}(p)$ for both cases.

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.

3. In class, we derived the score $U(\beta)$ 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 $U(\beta) = \sum_{i=1}^{n} \left(Y_i - \frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}\right) X_i$

   (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 $U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta}$ and $\mathcal{J}(\beta) = -\frac{\partial U(\beta)}{\partial \beta}$, 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 $\frac{\partial g(\mathbf{x})\mathbf{a}}{\partial \mathbf{x}} = \left(\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}}\right)\mathbf{a}^T$

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://sta712-f22.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

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

(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?