STA 712 Homework 2

Due: Tuesday, September 13, 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.

MLE review

1. If $Y \sim Poisson(\lambda)$, then

$$P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

where $\lambda > 0$ and $k = 0, 1, 2, ...$. Suppose we observe $Y_1, ..., Y_n \overset{iid}{\sim} Poisson(\lambda)$.

   (a) Derive the maximum likelihood estimate of λ.

   (b) Derive the observed information $\mathcal{J}(\lambda)$ and the Fisher information $\mathcal{I}(\lambda)$.

   (c) Let $\widehat{\lambda}$ be the maximum likelihood estimate of λ. Show that $Var(\widehat{\lambda}) = \lambda/n$. How does this relate to the Fisher information $\mathcal{I}(\lambda)$?

Sneak peek: Poisson regression

2. So far, we have worked with logistic regression models for a binary response. Later in the course, we will work with other types of response variables, like a Poisson response. This question will give you a preview of Poisson regression, while giving you practice with Fisher scoring.

Suppose that we have the Poisson regression model

$$\begin{aligned} Y_i &\sim Poisson(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}, \end{aligned}$$

and we observe data $(X_1, Y_1), ..., (X_n, Y_n)$, where $X_i \in \mathbb{R}^k$. (Since $\lambda > 0$ for a Poisson variable, $\log(\lambda) \in (-\infty, \infty)$, which makes it reasonable for $\log(\lambda_i)$ to be a linear function of the Xs).

   (a) Let $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)^T$, $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, $\boldsymbol{\lambda} = (\exp\{\boldsymbol{\beta}^T X_1\}, ..., \exp\{\boldsymbol{\beta}^T X_2\})^T$, and $\mathbf{X} \in \mathbb{R}^{n \times (k+1)}$ the design matrix with rows $(1, X_{i,1}, X_{i,2}, ..., X_{i,k})$. Show that

$$U(\boldsymbol{\beta}) = \mathbf{X}^T(\boldsymbol{Y} - \boldsymbol{\lambda}).$$

   (b) Let $\boldsymbol{W} = \text{diag}(\lambda_1, ..., \lambda_n)$, where $\lambda_i = \exp\{\boldsymbol{\beta}^T X_i\}$. Show that

$$\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}.$$

   (c) In R, simulate $n = 500$ observations $(X_1, Y_1), ..., (X_n, Y_n)$. Draw $X_i \overset{iid}{\sim} N(0, 1)$, and $Y_1 \sim Poisson(\lambda_i)$, where $\log(\lambda_i) = -2 + 2X_i$.

(d) Using your simulated data from part (c), fit a Poisson regression model of Y on X, and report the fitted model coefficients. To fit a Poisson regression model in R:

```
glm(y ~ x, family = poisson)
```

(e) Modify your code from HW1, Question 4 to implement Fisher scoring for Poisson regression with the simulated data. Begin with $\beta^{(0)} = (0,0)^T$, and 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 (d)? How many scoring iterations did it take to converge?

(Randomized) quantile residuals

3. In class, we talked about (randomized) quantile residuals as a method of assessing the shape assumption in logistic regression. To formally define quantile residuals, we will follow the textbook (Section 8.3.4.2).

Suppose we have a logistic regression model:

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

We observe data $(X_1, Y_1), ..., (X_n, Y_n)$ and fit the model, producing coefficient estimates $\widehat{\beta}$ which give estimated probabilities $\widehat{p}_i$. The *(randomized) quantile residual* $r_{Q,i}$ for the ith observation is defined by

$$r_{Q,i} = \Phi^{-1}(u), \qquad u \sim \begin{cases} Uniform(1-\widehat{p}_i, 1) & Y_i = 1 \\ Uniform(0, 1-\widehat{p}_i) & Y_i = 0, \end{cases}$$

where Φ is the standard normal CDF.

(a) Show that if $\widehat{p}_i = p_i$ (our estimated probability is correct), then $r_{Q,i} \sim N(0,1)$. *Hint: treat the response Y_i as a random variable, and note that $Y_i \sim Bernoulli(\widehat{p}_i)$ if $p_i = \widehat{p}_i$.*

(b) Show that $\mathbb{E}[r_{Q,i}] > 0$ when $\widehat{p}_i < p_i$, and $\mathbb{E}[r_{Q,i}] < 0$ when $\widehat{p}_i > p_i$.

(c) Write your own function in R to compute randomized quantile residuals for a binary logistic regression model. (Your function may not call the `qresid` function from the `statmod` package).

(d) Using code from the class activity on September 2, generate data for which the logistic regression shape assumption is satisfied. Then create a quantile residual plot using your R function, and show that the residuals $r_{Q,i}$ are randomly scattered around the horizontal line at 0.

(e) Using code from the class activity on September 2, generate data for which the logistic regression shape assumption is *not* satisfied. Then create a quantile residual plot using your R function, and show that the plot shows a violation of the shape assumption.