STA 711 Exam 1

Due: Monday, February 26, 11:00am 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.

Rules: This is an open-book, open-notes exam. You may:

- Use any resources from the course (the textbook, the course website, class notes, previous assignments, etc.)
- Email me, or come to office hours, with specific questions (I may be somewhat less helpful than for regular assignments)
- Use one or two days from your bank of extension days, if you need more time on the exam

You may *not*:

- Use the internet to look up any questions on the exam
- Use any resources outside of the course (other textbooks, notes from other universities, etc.)
- Use WolframAlpha or any generative AI to help solve the problems
- Discuss the exam with anyone else

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.

2. Let $Y_1, ..., Y_n$ be iid random variables with pdf

$$f(y|\mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(\log(y) - \mu)^2\right\}$$

where $\sigma^2 > 0$, $y > 0$, and $\mu \in \mathbb{R}$. Find the maximum likelihood estimators $\widehat{\mu}$ and $\widehat{\sigma}^2$.

3. A random variable X follows a *categorical* distribution with k categories if $X \in \{1, ..., k\}$ and the probability that X is in category j is $P(X = j) = p_j$, with each $p_j \in [0, 1]$ and $\sum_{j=1}^{k} p_j = 1$. We write $X \sim Categorical(p_1, ..., p_k)$. (This is just a generalization of the Bernoulli to more than two categories).

   Suppose that we observe $X_1, ..., X_n \overset{iid}{\sim} Categorical(p_1, ..., p_k)$. Let $n_j = \sum_{i=1}^{n} \mathbb{1}\{X_i = j\}$ (the number of observations in category j), and note that $\sum_j n_j = n$. Find the maximum likelihood estimators $\widehat{p}_j$ of each probability p_j. (*Hint:* You will need to add a constraint that $\sum_j \widehat{p}_j = 1$. Lagrange multipliers may be helpful.)

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

$$f(x|\theta) = \frac{4\theta^4}{x^5} \mathbb{1}\{x \geq \theta\}.$$

Find the maximum likelihood estimator $\widehat{\theta}$.

A Poisson hurdle model

Poisson regression models are commonly used when the response is a *count* variable (i.e., Y_i takes values in $\{0, 1, 2, ...\}$). On HW 3, you showed that the Poisson distribution is an example of an EDM, and so the same general model fitting approach, that we used for logistic regression, can be applied to Poisson data.

However, the Poisson distribution is restrictive about the frequency of 0s: if $Y \sim Poisson(\lambda)$, then $P(Y = 0) = e^{-\lambda}$. If our data have more 0s than a Poisson distribution allows, one option is to fit a *hurdle* model, which models the 0s and the non-zeros separately. In the following questions, we will build towards defining a hurdle model and finding the score and information.

5. Let V be a random variable with support $\{1, 2, 3, ...\}$ (all positive integers). We say that V follows a *positive Poisson* distribution, and write $V \sim PosPoisson(\lambda)$, if

$$P(V = v) = \frac{\lambda^v e^{-\lambda}}{v!(1 - e^{-\lambda})} \qquad \text{for } v = 1, 2, 3, ...$$

   (In other words, V is the conditional distribution of a Poisson restricted to non-zero values).

   Show that the positive Poisson distribution is an example of an EDM by finding ϕ, θ, $\kappa(\theta)$, and $a(v, \phi)$.

6. Suppose that $V \sim PosPoisson(\lambda)$. Using question 5 and properties of EDMs from HW 3, calculate $\mathbb{E}[V]$ and $Var(V)$.

7. Let $Y_i \in \{0, 1, 2, ...\}$ be a count variable of interest, and $X_i = (1, X_{i,1}, ..., X_{i,k})^T \in \mathbb{R}^{k+1}$ be a vector of covariates. Suppose we observe independent samples $(X_1, Y_1), ..., (X_n, Y_n)$ from the following model:

$$\begin{aligned} P(Y_i = 0) &= 1 - p_i \\ Y_i | (Y_i > 0) &\sim PosPoisson(\lambda_i) \\ \log\left(\frac{p_i}{1 - p_i}\right) &= \gamma^T X_i \\ \log(\lambda_i) &= \beta^T X_i, \end{aligned}$$

   where $\gamma, \beta \in \mathbb{R}^{k+1}$ are vectors of coefficients. That is, our model contains two separate pieces: one piece to model the probability that $Y_i = 0$, and another piece to model the non-zero values of Y_i. This is called a *hurdle* model, and is useful for handling count variables with excess 0s (aka *zero inflation*). Our goal is to calculate maximum likelihood estimates $\widehat{\gamma}$ and $\widehat{\beta}$ of the coefficient vectors.

   (a) Write down the log-likelihood $\ell(\gamma, \beta | X, Y)$ for the hurdle model. *Hint:* it may help to

view the pmf of Y_i as a piecewise function:

$$P(Y_i = y) = \begin{cases} 1 - p_i & y = 0 \\ p_i \dfrac{\lambda_i^y e^{-\lambda_i}}{y!(1 - e^{-\lambda_i})} & y > 0 \end{cases}$$

(b) Find the score function $U(\gamma, \beta | X, Y)$. *Hint:*

$$U(\gamma, \beta | X, Y) = \begin{bmatrix} \dfrac{\partial \ell}{\partial \gamma} \\ \dfrac{\partial \ell}{\partial \beta} \end{bmatrix} \in \mathbb{R}^{2(k+1)}$$

(c) Find the Fisher information matrix $\mathcal{I}(\gamma, \beta | X, Y)$. *Hint:* $\mathcal{I}(\gamma, \beta | X, Y) \in \mathbb{R}^{2(k+1) \times 2(k+1)}$