DS-GA 3001: Applied Statistics (Fall 2023-24)
Midterm, Tuesday October 31st
Solutions

Instructions:

- You have **100 minutes**, 4:55PM - 6:35PM
- The exam has 4 problems, totaling 100 points (+5 bonus points).
- Please answer each problem in the space below it.
- You are allowed to carry the textbook, your own notes and other course related material with you. Electronic devices are not allowed.
- Please read the problems carefully.
- We use boldcase letters $\boldsymbol{\theta}, \boldsymbol{x}, \cdots$ to distinguish vectors from scalars.
- Unless otherwise specified, you are required to provide explanations of how you arrived at your answers.
- You can use previous parts of a problem even if you did not solve them.
- The problems may not be arranged in an increasing order of difficulty. If you get stuck, it might be wise to try other problems first.
- Good luck and enjoy!

Full name: _____

N number: _____

1. Binary choice questions. *(40 points)*

For each of the statements, decide if it is "True" or "False". Provide explanations if you think it is "False". Each question is worth 5 points.

(a) In exponential families, the natural parametrization is the unique parametrization such that the log-likelihood in the corresponding GLM is concave.

Solution: False. In the Bernoulli exponential family, the logistic regression and Probit regression give two different parametrizations such that the resulting log-likelihood is concave.

(b) By the delta method, if $X \sim \text{Poi}(\lambda)$, then $\text{Var}(\sqrt{X}) \approx 1/4$ for large λ.

Solution: True. The delta method gives that

$$\text{Var}(\sqrt{X}) \approx \left(\frac{1}{2\sqrt{\lambda}}\right)^2 \text{Var}(X) = \frac{1}{4\lambda} \cdot \lambda = \frac{1}{4}.$$

(c) Among the tests for generalized linear models, in practice the likelihood ratio test is typically preferred to the Wald or score tests because it has the best asymptotic (i.e. sample size $n \to \infty$) performance.

Solution: False. These tests have the same asymptotic performance (i.e. test statistics converge to the same chi-squared distribution). The likelihood ratio test is preferred because it typically enjoys the best *non-asymptotic* performance with finite data.

(d) Given a sample $(y_1, \cdots, y_n) \sim p_\theta$ and an estimator $\widehat{\theta} = f(y_1, \cdots, y_n)$ for θ, Alice believes that bootstrap can also be used to estimate the bias

$$b(\theta) := \mathbb{E}_\theta[\widehat{\theta}] - \theta = \mathbb{E}_\theta[f(y_1, \cdots, y_n)] - \theta.$$

Specifically, Alice draws m bootstrap samples $\mathbf{Y}^{(1)}, \cdots, \mathbf{Y}^{(m)}$, where each sample $\mathbf{Y}^{(j)}$ consists of n i.i.d. draws from $p_{\widehat{\theta}}$. She proceeds to estimate $b(\theta)$ using

$$\widehat{b} = \left(\frac{1}{m} \sum_{j=1}^{m} f(\mathbf{Y}^{(j)}) \right) - \widehat{\theta}.$$

Claim: for large m, the above estimator $\widehat{b}$ is a reasonable estimator of $b(\theta)$.

Solution: True. In fact, LLN gives $\widehat{b} \to b(\widehat{\theta})$ as $m \to \infty$, which is a reasonable plug-in estimator for the true bias $b(\theta)$.

(e) Because the Poissonization step motivated by the profile likelihood is still valid in high dimensions, Lindsey's method remains suitable and practical for estimating high-dimensional densities.

Solution: False. Lindsey's method involves a discretization step at the beginning, which suffers from the curse of dimensionality (i.e. requires exponentially many small cubes in high dimensions).

(f) In a continuous-time hazards model, if two hazard functions satisfy $h_1(t) = 2h_2(t)$ for every t, then $S_1(t) = S_2(t)^2$ holds for the corresponding survival functions.

Solution: True. This is because

$$S_1(t) = \exp\left(-\int_0^t h_1(s)ds\right) = \exp\left(-2\int_0^t h_2(s)ds\right) = S_2(t)^2.$$

(g) In an error-in-variable model $y \sim \mathcal{N}(x\theta, \sigma_y^2)$, Bob knows (y, σ_y) but not (x, θ). In addition, he has access to a pivot $\widehat{x} \in \mathbb{R}^p$ where $x \sim \mathcal{N}(\widehat{x}, \sigma_x^2)$ with a known σ_x. Bob uses the following estimator of θ:

$$\widehat{\theta} = \arg\min_{\theta} \left[\min_{x} \left(\frac{(y - x\theta)^2}{\sigma_y^2} + \frac{(x - \widehat{x})^2}{\sigma_x^2} \right) \right].$$

Claim: this estimator is the maximizer of the profile likelihood for θ.

Solution: True. The joint log-likelihood for (x, θ) is given by

$$\ell(\theta, x) = -\frac{(y - x\theta)^2}{2\sigma_y^2} - \frac{(x - \widehat{x})^2}{2\sigma_x^2} + C,$$

so the profile maximum likelihood estimator $\widehat{\theta}$ is given by

$$\widehat{\theta} = \arg\max_{\theta} \left[\max_{x} \ell(\theta, x) \right] = \arg\min_{\theta} \left[\min_{x} \left(\frac{(y - x\theta)^2}{\sigma_y^2} + \frac{(x - \widehat{x})^2}{\sigma_x^2} \right) \right].$$

(h) For the above problem, Bob proposes the following algorithm to compute $\widehat{\theta}$. Given an initialization x^0, for $t = 0, 1, 2, \cdots$:

i. let θ^{t+1} be the minimizer of $\theta \mapsto \ell(\theta, x^t)$;

ii. let x^{t+1} be the minimizer of $x \mapsto \ell(\theta^{t+1}, x)$.

Here $\ell(\theta, x)$ is the objective function in the above definition of $\widehat{\theta}$. The estimator $\widehat{\theta}$ is then defined to be the final θ^t when the algorithm converges.

Bob argues that this is a reasonable algorithm because: (i) $\ell(\theta, x)$ is not jointly convex in (θ, x); (ii) for a fixed θ (resp. x), $\ell(\theta, x)$ becomes convex in x (resp. θ). Determine if the statements (i) and (ii) are "both true" or "not both true".

Solution: Both true. The objective function ℓ involves the product $x\theta$, so is not jointly convex in (θ, x). However, when x or θ is fixed, the function ℓ is quadratic in the other variable with a positive leading term, and is thus convex.

2. Computation of deviance and Fisher information. *(20 points)*

Let $p_\theta = \mathcal{N}(\theta, 1)$ be a Gaussian location model with mean $\theta \in \mathbb{R}$ and variance 1.

(a) For $\theta_1, \theta_2 \in \mathbb{R}$, compute the deviance $D(\theta_1; \theta_2)$. *(10 points)*

Solution: Writing the Gaussian location model in the exponential family form, the log-partition function is $A(\theta) = \theta^2/2$. Therefore,

$$\begin{aligned} D(\theta_1; \theta_2) &= 2[A(\theta_2) - A(\theta_1) - A'(\theta_1)(\theta_2 - \theta_1)] \\ &= \theta_2^2 - \theta_1^2 - 2\theta_1(\theta_2 - \theta_1) \\ &= (\theta_2 - \theta_1)^2. \end{aligned}$$

(b) For $\theta \in \mathbb{R}$, compute the Fisher information $I(\theta)$. *(10 points)*

Solution:

$$\begin{aligned} &\text{log-likelihood:} && \ell_\theta(y) = -\frac{(y-\theta)^2}{2} - \frac{1}{2}\log(2\pi) \\ &\text{score function:} && s_\theta(y) = \frac{\partial}{\partial\theta}\ell_\theta(y) = y - \theta \\ &\text{Hessian:} && \frac{\partial^2}{\partial\theta^2}\ell_\theta(y) = \frac{\partial}{\partial\theta}s_\theta(y) = -1 \\ &\text{Fisher information:} && I(\theta) = \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}\ell_\theta(y)\right] = 1. \end{aligned}$$

3. Cox model. *(20 points)*

Consider a set of censored survival data from a randomized clinical trial for which we have information on several covariates:

- z_1 is the indicator for treatment, i.e. $z_1 = 1$ represents the treatment group, and $z_1 = 0$ represents the control group;
- $z_2 \in \{1, 2, 3\}$ is the disease stage at the time of randomization, where 1 represents the least severe, and 3 represents the most severe;
- $\boldsymbol{x} \in \mathbb{R}^p$ consists of other features (gender, age, etc.)

Design different Cox models for the following scenarios.

(a) Suppose that the treatment effect does not depend on the disease stage, but the disease stages should be treated as *ordered* categorical variables. Propose a model for the hazard function $h(t \mid z_1, z_2, \boldsymbol{x})$ in this scenario. *(5 points)*

Solution:

$$h(t \mid z_1, z_2, \boldsymbol{x}) = h(t) \exp\big(\boldsymbol{\beta}^\top \boldsymbol{x} + \beta_1 z_1 + \beta_{2,1}\mathbb{1}(z_2 = 1) + \beta_{2,2}\mathbb{1}(z_2 = 2) + \beta_{2,3}\mathbb{1}(z_2 = 3)\big),$$

where $\beta_{2,1} \le \beta_{2,2} \le \beta_{2,3}$ represents that the more severe the disease stage is, the larger the hazard function is.

(b) Modify your model in (a) for $h(t \mid z_1, z_2, \boldsymbol{x})$ to allow the magnitude of the treatment effect to vary with the disease stage. *(5 points)*

Solution:

$$h(t \mid z_1, z_2, \boldsymbol{x}) = h(t) \exp\big(\boldsymbol{\beta}^\top \boldsymbol{x} + z_1(\beta_{1,1}\mathbb{1}(z_2 = 1) + \beta_{1,2}\mathbb{1}(z_2 = 2) + \beta_{1,3}\mathbb{1}(z_2 = 3)) + \beta_{2,1}\mathbb{1}(z_2 = 1) + \beta_{2,2}\mathbb{1}(z_2 = 2) + \beta_{2,3}\mathbb{1}(z_2 = 3)\big),$$

where $\beta_{1,1} \le \beta_{1,2} \le \beta_{1,3}$, $\beta_{2,1} \le \beta_{2,2} \le \beta_{2,3}$.

(c) Given your model in (b), describe a test for the hypothesis that the magnitude of the treatment effect does *not* vary with the disease stage. You only need to write your (H_0, H_1), and need not give details of how to carry out the test. *(5 points)*

Solution:

$$H_0: \quad \beta_{1,1} = \beta_{1,2} = \beta_{1,3}$$
$$H_1: \quad \beta_{1,1}, \beta_{1,2}, \beta_{1,3} \text{ are not all equal}$$

(d) Suppose that before observing (z_1, z_2), we would like to select a subset of features from $\boldsymbol{x}$ via Lasso. Propose a Lasso objective function for the model selection, and explain how you choose the Lasso regularization parameter λ.

You may assume that your dataset is $\{(\boldsymbol{x}_i, t_i, \Delta_i) : i = 1, \cdots, n\}$, where t_i is the death/censored time for individual i, and Δ_i is the indicator of death/censoring. *(5 points)*

Solution:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} - \sum_{i:\Delta_i=1} \log\left(\frac{e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}^\top \boldsymbol{x}_j}}\right) + \lambda\|\boldsymbol{\beta}\|_1,$$

where $R_i = \{j : t_j \geq t_i\}$ is the risk set at time t_i. The parameter λ is chosen by cross validation.

4. Parameter estimation in the choice model. *(20 points + 5 bonus points)*

A choice model attempts to model the decision process of an individual, and is widely used in operations management and behavioral science. A typical dataset in assortment optimization takes the form $\{(S_t, j_t)\}_{t=1}^T$, where $S_t \subseteq \{1, \cdots, N\}$ is an *assortment* (i.e. a subset of products labeled by $\{1, \cdots, N\}$) offered to a customer at time t, and $j_t \in S_t$ is the product purchased by the customer. For simplicity we assume that the customer purchases exactly one product at each time.

To model the decision process of the customer, suppose there is an unknown positive vector $(p_1, \cdots, p_N)$ representing the common preference over the products; here $p_i > 0$. In a simple choice model, when an assortment $S \subseteq \{1, \cdots, N\}$ is offered to the customer, the customer chooses to purchase product $j \in S$ with probability

$$\mathbb{P}(j \mid S) = \frac{p_j}{\sum_{i \in S} p_i}.$$

(a) Write down the overall log-likelihood of $(p_1, \cdots, p_N)$, based on the entire dataset $\{(S_t, j_t)\}_{t=1}^T$. *(5 points)*

Solution:

$$\begin{aligned}
\ell(p_1, \cdots, p_N) &= \sum_{t=1}^T \log \mathbb{P}(j_t \mid S_t) \\
&= \sum_{t=1}^T \log \frac{p_{j_t}}{\sum_{i \in S_t} p_i} \\
&= \sum_{t=1}^T \left(\log p_{j_t} - \log \left(\sum_{i \in S_t} p_i \right) \right).
\end{aligned}$$

(b) For $j \in \{1, \cdots, N\}$ and $\boldsymbol{p} = (p_1, \cdots, p_N)$, let

$$n_j = \sum_{t=1}^{T} \mathbb{1}(j_t = j), \qquad n_j(\boldsymbol{p}) = \sum_{t=1}^{T} \frac{p_j \mathbb{1}(j \in S_t)}{\sum_{i \in S_t} p_i}$$

be the actual and expected number of choosing product j in the data, respectively. Based on your log-likelihood in (a), show that the MLE $\widehat{\boldsymbol{p}}$ should satisfy

$$n_j = n_j(\widehat{\boldsymbol{p}}), \quad \forall j \in \{1, 2, \cdots, N\}.$$

You may assume that the first-order condition holds for $\widehat{\boldsymbol{p}}$. You also do not need to worry about the non-uniqueness of the MLE in this example. *(5 points)*

Solution:

$$\begin{aligned}
\frac{\partial \ell}{\partial p_j} &= \sum_{t=1}^{T} \frac{\partial}{\partial p_j} \left(\log p_{j_t} - \log \left(\sum_{i \in S_t} p_i \right) \right) \\
&= \sum_{t=1}^{T} \left(\frac{\mathbb{1}(j_t = j)}{p_j} - \frac{\mathbb{1}(j \in S_t)}{\sum_{i \in S_t} p_i} \right) \\
&= \frac{n_j}{p_j} - \frac{n_j(\boldsymbol{p})}{p_j}.
\end{aligned}$$

By the first-order condition for $\widehat{\boldsymbol{p}}$, the above derivative should be zero for $\widehat{\boldsymbol{p}}$, i.e. $n_j = n_j(\widehat{\boldsymbol{p}})$.

(c) Find a proper reparameterization of $(p_1, \cdots, p_N)$ by another vector $\boldsymbol{\theta}$, such that the log-likelihood becomes concave in $\boldsymbol{\theta}$. *(5 points)*

Solution: Write $p_j = \exp(\theta_j)$, then the log-likelihood

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{T} \left(\theta_{j_t} - \log \left(\sum_{i \in S_t} e^{\theta_i} \right) \right)$$

becomes concave in $\boldsymbol{\theta}$.

(d) Now suppose that for each product j, the product feature $\boldsymbol{x}_j \in \mathbb{R}^p$ is also available in the data. Propose a reasonable choice model for $\mathbb{P}(j \mid S, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ to include the product features – the overall log-likelihood should be concave in your parametrization. *(5 points)*

Solution: Motivated by the GLM, we may take $\theta_j = \boldsymbol{\beta}^\top \boldsymbol{x}_j$. In other words, the choice model now becomes

$$\mathbb{P}(j \mid S, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x}_j)}{\sum_{i \in S} \exp(\boldsymbol{\beta}^\top \boldsymbol{x}_i)}.$$

(e) To compute the MLE in (a), a natural way is to use the reparametrization in (b) and apply Newton's method to $\boldsymbol{\theta}$. This step requires to compute the Hessian and could be computationally expensive.

An alternative idea is to apply a recursive algorithm similar in spirit to EM. Show that for every $\boldsymbol{p}$, if we define $\boldsymbol{q} = (q_1, \cdots, q_N)$ as

$$q_j = \frac{p_j n_j}{n_j(\boldsymbol{p})}, \quad \forall j \in \{1, \cdots, N\},$$

then the log-likelihood is non-decreasing moving from $\boldsymbol{p}$ to $\boldsymbol{q}$. *(5 bonus points)*

Solution:

$$\begin{aligned}
\ell(\boldsymbol{q}) - \ell(\boldsymbol{p}) &= \sum_{t=1}^{T} \left(\log \frac{q_{j_t}}{\sum_{i \in S_t} q_i} - \log \frac{p_{j_t}}{\sum_{i \in S_t} p_i} \right) \\
&= \sum_{t=1}^{T} \left(\log \frac{n_{j_t}}{n_{j_t}(\boldsymbol{p})} - \log \frac{\sum_{i \in S_t} q_i}{\sum_{i \in S_t} p_i} \right) \\
&\geq \sum_{t=1}^{T} \left(\log \frac{n_{j_t}}{n_{j_t}(\boldsymbol{p})} - \frac{\sum_{i \in S_t} q_i}{\sum_{i \in S_t} p_i} + 1 \right) \quad (\text{because } \log x \leq x - 1) \\
&= \sum_{j=1}^{N} \left(n_j \log \frac{n_j}{n_j(\boldsymbol{p})} - q_j \sum_{t=1}^{T} \frac{\mathbb{1}(j \in S_t)}{\sum_{i \in S_t} p_i} \right) + T \\
&= \sum_{j=1}^{N} \left(n_j \log \frac{n_j}{n_j(\boldsymbol{p})} - \frac{q_j n_j(\boldsymbol{p})}{p_j} \right) + T \\
&= \sum_{j=1}^{N} \left(n_j \log \frac{n_j}{n_j(\boldsymbol{p})} - n_j \right) + T \\
&= \sum_{j=1}^{N} n_j \log \frac{n_j}{n_j(\boldsymbol{p})} \\
&= T \cdot D_{\mathrm{KL}} \left(\frac{(n_1, \cdots, n_N)}{T} \middle\| \frac{(n_1(\boldsymbol{p}), \cdots, n_N(\boldsymbol{p}))}{T} \right) \\
&\geq 0.
\end{aligned}$$