Student ID (NOT your name):

Final Examination: QUESTION BOOKLET

Prof. Will Fithian

Fall 2021

- Do *NOT* open this question booklet until you are told to do so.
- Write your Student ID number (**NOT** your name) at the top of this page.
- Write your solutions in this booklet.
- No electronic devices are allowed during the exam.
- Be neat! If we can't read it, we can't grade it.
- You can treat any results from lecture or homework as "known," and use them in your work without rederiving them, but do make clear what result you're using. You do not need to explicitly check regularity conditions for the theorems from class that required them.
- For a multi-part problem, you may treat the results of previous parts as given (if you don't prove the result for part (a), you can still use it to solve part (b)).
- I have starred some parts which I believe are the most difficult, and which I expect most students won't necessarily be able to solve in the time allotted. They are generally not worth more points than the less difficult parts, so don't waste too much time on them until you're happy with your answers to the latter.
- Be careful to justify your reasoning and answers. We are primarily interested in your understanding of concepts, so show us what you know.
- Good luck!

1. Regression with correlated errors (25 points, 5 points / part).

Some useful facts for this problem:

- For $\mu \in \mathbb{R}^n$ and positive definite $\Sigma \in \mathbb{R}^{n \times n}$, the density for $Z \sim N_n(\mu, \Sigma)$ is
$$p_{\mu,\Sigma}(z) = |2\pi\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right\},$$
where $|\cdot|$ is the determinant (note the exponent of $1/2$ is correct; it should not be $n/2$). The mean is μ and the variance is Σ.
- If $Z \sim N_n(\mu, \Sigma)$, and $A \in \mathbb{R}^{k \times n}$ and $b \in \mathbb{R}^k$ are fixed, then
$$AZ + b \sim N_k(A\mu + b, A\Sigma A').$$

Suppose that for $i = 1, \ldots, n$ we observe fixed covariates $x_i \in \mathbb{R}^d$ and random response $Y_i = x_i'\beta + \varepsilon_i$, for coefficient vector $\beta \in \mathbb{R}^d$ and $\varepsilon_i \in \mathbb{R}$. The errors are multivariate Gaussian with mean zero and positive definite covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. In terms of the full response vector $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n \times d}$ with ith row x_i', we have
$$Y = X\beta + \varepsilon, \quad \text{with } \varepsilon \sim N_n(0, \Sigma).$$
Assume $n \geq d \geq 1$ and X has full column rank. For parts (a) and (b), we will assume Σ is known and we want to estimate β. For (c)-(e) we will assume Σ is unknown.

(a) Show that Y follows a full-rank exponential family model and identify its complete sufficient statistic.

(b) Find the maximum likelihood estimator of β and give its distribution.

(c) Now, for the remainder of the problem, suppose that Σ is unknown so we have to estimate it. To facilitate this, we observe i.i.d. replicates $Y^{(k)}$ for $k = 1, \ldots, m$, with distribution
$$Y^{(k)} = X\beta + \varepsilon^{(k)}, \quad \text{with } \varepsilon^{(k)} \overset{\text{i.i.d.}}{\sim} N_n(0, \Sigma).$$
Note that X and β are the same for $k = 1, \ldots, m$ (they do not depend on k); only the errors change (and the responses change as a result). Define
$$\overline{Y} = \frac{1}{m}\sum_{k=1}^m Y^{(k)}, \quad \text{and } \widehat{\Sigma} = \frac{1}{m-1}\sum_{k=1}^m (Y^{(k)} - \overline{Y})(Y^{(k)} - \overline{Y})'.$$
Show that $\overline{Y}$ and $\widehat{\Sigma}$ are independent of each other.

(d) Show that $\widehat{\Sigma}$ is an unbiased estimator of Σ.

(e) Now assume $n = d$. Is $\widehat{\Sigma}$ UMVU? Why or why not?

Problem 1 answers continued (1):

Problem 1 answers continued (2):

Problem 1 answers continued (3):

2. Contamination model (25 points, 5 points / part).

Suppose that we observe n random variables $X_1, \ldots, X_n \in [0, 1]$. The observations are supposed to come from a uniform distribution, but we suspect that our sample may be contaminated by a small proportion of observations from another, known distribution with Lebesgue density $q(x)$ (q is not necessarily continuous). Assume that for some $C < \infty$, $0 \leq q(x) \leq C$ for all $x \in [0, 1]$. That is, we observe

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x) = 1 - \theta + \theta q(x).$$

Assume $\theta \in [0, b]$ for some $b < 1$.

(a) Show that the maximum likelihood estimator $\hat{\theta}_n$ is consistent for the true value θ_0 as $n \to \infty$.

(b) Give the asymptotic distribution of the maximum likelihood estimator as $n \to \infty$, for $\theta_0 \in (0, b)$. Give an explicit expression for the asymptotic variance in terms of a definite integral (you don't need to check any regularity conditions for this part).

(c) Find a score test for the null hypothesis that there is no contamination, against the alternative that there is some, i.e. test $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. Give an explicit expression for your test statistic and your cutoff, in terms of a definite integral and a quantile of a known distribution.

(d) (*) If we expand the parameter space to $[0, 1)$, is the MLE still consistent?

(e) (*) If $\theta_0 = 0$, give the distribution of the MLE as $n \to \infty$.

Problem 2 answers continued (1):

Problem 2 answers continued (2):

Problem 2 answers continued (3):

3. Two-by-two count table (25 points, 5 points / part).

Some useful facts for this problem:

- For $\theta > 0$, the Poisson density for $X \sim \text{Pois}(\theta)$ is $\frac{\theta^x e^{-\theta}}{x!}$ on $x = 0, 1, \ldots$. The mean and variance are both θ.
- For $\pi_1, \ldots, \pi_d \geq 0$ with $\sum_{i=1}^d \pi_i = 1$, the multinomial density for $X \sim \text{Multinom}(n, \pi)$ is
$$p_{n,\pi}(x) = n! \cdot \prod_{i=1}^d \frac{\pi_i^{x_i}}{x_i!},$$
on $x \in \{0, \ldots, n\}^d$ with $\sum_i x_i = n$.
- Suppose $X_i \sim \text{Pois}(\theta_i)$ with $\theta_i > 0$, independently for $i = 1, \ldots, d$, and let $X_+ = \sum_{i=1}^d X_i$ and $\theta_+ = \sum_{i=1}^d \theta_i$. Then, conditional on $X_+ = n$,
$$(X_1, \ldots, X_d) \sim \text{Multinom}(n, (\theta_1, \ldots, \theta_d)/\theta_+)$$

Assume that $X_{ij} \sim \text{Pois}(\lambda_{ij})$, independently for $i, j \in \{0, 1\}$. We will consider the model with $\lambda_{ij} = \lambda_0 \rho^{i+j}$, for $\lambda_0, \rho > 0$. Except when otherwise specified, assume both parameters are unknown.

(a) Give a complete sufficient statistic for the model and show it is complete.

(b) Assume (for this part **only**) that λ_0 is known, but ρ is unknown. Suggest a UMP test of $H_0 : \ \rho = \rho_0$ vs. $H_1 : \ \rho > \rho_0$. You do not need to give an explicit cutoff for your test but give an explicit formula for the test statistic, explain how you would find the cutoff, and explain why your test is UMP.

(c) Assuming again that both parameters are unknown, suggest a UMPU test of $H_0 : \ \rho = 1$ against $H_1 : \ \rho > 1$. You do not need to give an explicit cutoff for your test but explain how you would calculate it. If the data are $X_{00} = X_{01} = 0$ and $X_{10} = X_{11} = 1$, calculate the (conservative, non-randomized) p-value for your test.

(d) For the same data set, $X_{00} = X_{01} = 0$ and $X_{10} = X_{11} = 1$, find the maximum likelihood estimators for λ_0 and ρ. Give your answers as explicit numbers.

(e) (*) Now suppose we consider a relaxed model $\lambda_{ij} = f(i + j)$, for any strictly positive real-valued function f on $\{0, 1, 2\}$. This includes our previous parametric model as a special case since we could have $f(i + j) = \lambda_0 \rho^{i+j}$. Does

there exist a UMPU test of the null hypothesis that our previous model was correctly specified, against the alternative that it was misspecified but the relaxed model is correct? Explain why or why not. (If you say yes you only need to give enough details to establish that such a test exists).

Problem 3 answers continued (1):

Problem 3 answers continued (2):

Problem 3 answers continued (3):

4. Change point problem (25 points, 5 points / part).

Some useful facts for this problem:

- The Beta distribution Beta(α, β) with parameters $\alpha, \beta > 0$ has density

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad \text{where } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

  with respect to the Lebesgue measure on $(0, 1)$. Its mean and variance are

$$\mathbb{E}_{\alpha,\beta}[X] = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}_{\alpha,\beta}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

- The negative binomial distribution NB(m, θ) with parameters $m \in \{1, 2, \ldots\}, \theta \in (0, 1)$ has probability mass function

$$p_{m,\theta}(x) = \binom{x+m-1}{x}\theta^x(1-\theta)^m, \quad \text{for } x \in \{0, 1, 2, \ldots\}.$$

  Its mean and variance are

$$\mathbb{E}_{m,\theta}[X] = \frac{m\theta}{1-\theta}, \quad \text{Var}_{m\theta}(X) = \frac{m\theta}{(1-\theta)^2}.$$

Assume we observe independent random variables $X_i \sim \text{NB}(m, \theta_i)$ for $i = 1, \ldots, n$. Assume also that the θ_i values are constant except at some integer $k \in \{1, \ldots, n-1\}$ where they change. That is,

$$\theta_i = \begin{cases} \gamma_0 & \text{if } i \le k \\ \gamma_1 & \text{if } i > k \end{cases},$$

for $\gamma_0, \gamma_1 \in (0, 1)$.

Until otherwise specified, assume k is known. Throughout the problem, we will assume m is known.

(a) Calculate the maximum likelihood estimator for γ_0 and find its asymptotic distribution if $k, n \to \infty$. You do not need to check regularity conditions.

(b) Next assume we introduce a prior distribution that $\gamma_0, \gamma_1 \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta)$. Give the posterior distribution for (γ_0, γ_1) given $X_1, \ldots, X_n$, and give the Bayes estimator for squared error loss.

(c) (*) Find the asymptotic distribution of the Bayes estimator for γ_0, holding γ_0 and γ_1 fixed and sending $k, n \to \infty$.

(d) Next, we relax the assumption that k is known. Instead assume $n = 10$ and all we know is that $k \in \{4, 5, 6\}$. Find a minimal sufficient statistic for the three-parameter model with $\gamma_0, \gamma_1 \in (0, 1)$ and $k \in \{4, 5, 6\}$. You do not need to prove it is minimal, as long as you give the right answer.

(e) Continuing with the three-parameter model above, consider a Bayesian approach where we assign priors $k \sim \text{Unif}(\{4, 5, 6\})$ independently of $\gamma_0, \gamma_1 \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta)$. Give a Gibbs sampler algorithm to sample from the posterior distribution of (k, γ_0, γ_1).

Problem 4 answers continued (1):

Problem 4 answers continued (2):

Problem 4 answers continued (3):