Student ID (NOT your name):

Final Examination: QUESTION BOOKLET

Prof. Will Fithian

Fall 2019

- 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. Poisson minimax estimation (24 points, 4 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.
- The Gamma density for $X \sim \text{Gamma}(k, \beta)$, where $\beta > 0$ is the *rate* parameter, is
$$\frac{\beta^k}{\Gamma(k)} x^{k-1} e^{-\beta x}, \qquad \text{on } x > 0,$$
where $\Gamma(k) = \int_0^\infty z^{k-1} e^{-z}\, dz$. The mean and variance are k/β and k/β^2, respectively.
- If $X \sim \text{Gamma}(k, \beta)$ (in the rate parameterization) with $k > 1$, then $\mathbb{E}[X^{-1}] = \beta/(k-1)$.

Consider estimating θ given a single Poisson observation $X \sim \text{Pois}(\theta)$ using the loss function
$$L(d, \theta) = \frac{(d-\theta)^2}{\theta}.$$
Throughout this problem, unless otherwise specified, the risk of a given estimator is always calculated using this loss.

(a) Find the MLE and calculate its risk function.

(b) Show that $\theta \sim \text{Gamma}(k, \beta)$ is a conjugate prior for this problem and give the posterior distribution.

(c) Find the Bayes estimator for the prior from part (b) and the loss L defined above.

(d) (*) Show that the Bayes risk of the Bayes estimator from part (c) is $1/(1+\beta)$

(e) Show that the MLE is minimax relative to the loss L.

(f) Show that the minimax risk for the usual squared error loss — i.e., $L_{\text{SE}}(d, \theta) = (d-\theta)^2$ — is infinite (this motivates changing the loss function to our L, which "adjusts" for the hardness of the problem).

1. Solution.

(a) By inspection the Poisson density $e^{X \log\theta - \theta}/x!$ is an exponential family, so the MLE solves $E_\theta X = X$; the MLE is therefore $\hat{\theta} = X$. Its risk is

$$R(\theta) = \frac{1}{\theta}\mathbb{E}_\theta\left[(X-\theta)^2\right] = 1.$$

(b) The posterior density is

$$\begin{aligned} p(\theta \mid X) &\propto_\theta p(\theta)\cdot p(x \mid \theta) \\ &\propto_\theta \theta^{k-1}e^{-\theta\beta}\cdot\theta^x e^{-\theta} \\ &\propto_\theta \theta^{x+k-1}e^{-\theta(\beta+1)} \\ &\propto \text{Gamma}(x+k, \beta+1). \end{aligned}$$

Hence $\theta \mid X \sim \text{Gamma}(X+k, \beta+1)$. This is true for any setting of the prior parameters so the prior is conjugate.

(c) The Bayes estimator solves

$$\begin{aligned} \delta(X) &= \min_d \mathbb{E}\left[\frac{(d-\theta)^2}{\theta} \mid X\right] \\ &= \min_d d^2\mathbb{E}[\theta^{-1} \mid X] - 2d + \mathbb{E}[\theta \mid X] \\ &= 1/\mathbb{E}[\theta^{-1} \mid X] \\ &= \frac{X+k-1}{\beta+1}. \end{aligned}$$

(d) For the minimization problem in the last part, the minimized value is

$$-\delta(X) + \mathbb{E}[\theta \mid X] = -\frac{X+k-1}{\beta+1} + \frac{X+k}{\beta+1} = \frac{1}{\beta+1}.$$

The Bayes risk, then, is $\mathbb{E}\mathbb{E}\left[\frac{(\delta(X)-\theta)^2}{\theta} \mid X\right] = \frac{1}{\beta+1}$.

(e) Taking arbitrary $k > 1$, the sequence $\Gamma(k, \beta_n)$ has limiting risk equal to 1, which is also the sup-risk of the MLE. Hence it is a least-favorable sequence and the MLE is minimax.

(f) For the usual squared error loss, the Bayes estimator is the posterior mean and the conditional expectation of the loss (given X) is therefore the posterior variance, which is $(X+k+1)/(1+\beta)^2$. The Bayes risk is then

$$\mathbb{E}\frac{X+k+1}{(1+\beta)^2} = \frac{k/\beta+k+1}{(1+\beta)^2},$$

where we use $\mathbb{E}X = \mathbb{E}\theta = k/\beta$. If we fix $k > 1$ and send $\beta \to 0$, the Bayes risk tends to ∞. The minimax risk is larger than any Bayes risk, so it is also infinite.

2. ANOVA with random effects (25 points, 5 points / part).

Some useful facts for this problem:

- For $\sigma > 0$ and $\mu \in \mathbb{R}$, the Gaussian density for $X \sim N(\mu, \sigma^2)$ is $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$. If $\sigma^2 = 0$ then $X = 0$ almost surely.
- For a positive integer k, the χ_k^2 density for $X \sim \chi_k^2$ is
$$\frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}.$$
Its mean and variance are k and $2k$, respectively.

Assume we observe X_{ij} for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, and consider the hierarchical Gaussian model

$$\alpha_i \overset{\text{i.i.d.}}{\sim} N(0, \tau^2)$$
$$X_{ij} \mid \alpha \overset{\text{ind.}}{\sim} N(\mu + \alpha_i, \sigma^2).$$

Define the following quantities for the purposes of this problem:

$$\overline{X}_i = \frac{1}{n} \sum_j X_{ij},$$
$$S_i^2 = \frac{1}{n-1} \sum_j (X_{ij} - \overline{X}_i)^2,$$
$$\overline{X} = \frac{1}{nm} \sum_{i,j} X_{ij}, \quad \text{and}$$
$$S_B^2 = \frac{1}{m-1} \sum_i (\overline{X}_i - \overline{X})^2 \quad \text{(the } B \text{ stands for "between groups").}$$

The parameters $\mu \in \mathbb{R}, \tau^2 \geq 0$, and $\sigma^2 > 0$ are unknown. $\alpha_1, \ldots, \alpha_m$ are unobserved random variables but they are not parameters, and the model could be rewritten without them.

In this problem, unless otherwise stated, you do **NOT** need to show tests and confidence intervals are UMP(U) or UMA(U). Where I ask you to give an explicit formula, it is fine for the formula to be in terms of quantiles of one or more distributions from class.

(a) Show that $S_B^2, S_1^2, \ldots, S_m^2$ are mutually independent and give their distribution.

(b) Find a finite-sample, equal-tailed confidence interval for μ. Give an explicit formula.

(c) Give a finite-sample test of the null hypothesis $H_0 : \tau^2 = 0$ vs $H_1 : \tau^2 > 0$. Give an explicit formula for the test statistic and the critical value.

(d) (*) Find a finite-sample, equal-tailed confidence interval for τ^2/σ^2. Give an explicit formula.

(e) (*) Show that the model (with the additional restriction that $\tau^2 > 0$) is a three-parameter exponential family and $\left(\overline{X},\, S_B^2,\, \sum_i S_i^2\right)$ is a complete sufficient statistic.

2. Solution.

(a) $\overline{X}_i$ and S_i^2 are both functions of $(X_{i1}, \ldots, X_{i,n})$, which are mutually independent across $i = 1, \ldots, m$. Furthermore, $\overline{X}_i$ and S_i^2 are independent of each other within each i, as we have shown by e.g. Basu's theorem. Hence $(\overline{X}_1, \ldots, \overline{X}_m, S_1^2, \ldots, S_m^2)$ are $2m$ independent random variables, with

$$\begin{aligned} \overline{X}_i &\overset{\text{i.i.d.}}{\sim} N(\mu, \tau^2 + \sigma^2/n) \\ S_i^2 &\overset{\text{i.i.d.}}{\sim} \frac{\sigma^2}{n-1}\chi^2_{n-1}. \end{aligned}$$

As a result, $S_B^2 \sim \frac{\tau^2+\sigma^2/n}{m-1}\chi^2_{m-1}$, and it is independent of $(S_1^2, \ldots, S_m^2)$ because it is a function of $(\overline{X}_1, \ldots, \overline{X}_m)$.

(b) Because $\overline{X} = N(\mu, \tau^2/m + \sigma^2/nm)$, we have

$$\sqrt{m}\frac{X - \mu}{\sqrt{S_B^2}} \sim t_{m-1}.$$

As a result, $X \pm \sqrt{\frac{S_B^2}{m}}t_{m-1}(\alpha/2)$ is an exact $1-\alpha$ confidence interval.

(c) (**Common mistake:** Note that we cannot use S_B^2 alone to do this test because its distribution depends on the nuisance parameter τ^2.)

Combining evidence across the within-group sample variances gives us the combined within-group variance

$$S_W^2 = \frac{1}{m}\sum_i S_i^2 \sim \frac{\sigma^2}{m(n-1)}\chi^2_{m(n-1)}.$$

Because S_W^2 is independent of S_B^2, we have

$$\frac{S_B^2}{S_W^2} \sim \frac{\tau^2 + \sigma^2/n}{\sigma^2}F_{m-1,m(n-1)} = (\tau^2/\sigma^2 + n^{-1})F_{m-1,m(n-1)}.$$

As a result,

$$nS_B^2/S_W^2 \sim (n\tau^2/\sigma^2 + 1)F_{m-1,m(n-1)} \overset{H_0}{=} F_{m-1,m(n-1)}$$

so we can reject the null when that statistic is above $F_{m-1,m(n-1)}(\alpha)$.

(d) Using the same logic, for $\tau > 0$ we can get an equal tailed test of $H_0 : \tau^2/\sigma^2 = \rho$ vs. the two-sided alternative $H_1 : \tau^2/\sigma^2 \neq \rho$ by rejecting when $T_\rho = \frac{n}{n\rho+1} S_B^2/S_W^2$ is either above $a = F_{m-1,m(n-1)}(\alpha/2)$ or below $b = F_{m-1,m(n-1)}(1-\alpha/2)$. Hence the acceptance region is

$$\left\{ b \leq \frac{n}{n\rho+1} S_B^2/S_W^2 \leq a \right\}$$

leading to confidence interval

$$C(X) = \left[\frac{S_B^2}{aS_W^2} - \frac{1}{n}, \; \frac{S_B^2}{bS_W^2} - \frac{1}{n} \right].$$

(e) Let $X_i = (X_{i1}, \ldots, X_{in})$ denote the ith group of observations. The X_i are independent multivariate Gaussians with mean $\mu 1 = (\mu, \ldots, \mu)$ and covariance matrix $\Sigma = \sigma^2 I_n + \tau^2 11'$ (that is, the variance of X_{ij} is $\sigma^2 + \tau^2$ and the within-group covariance is τ^2). The inverse covariance matrix has the same form: $\Sigma^{-1} = \theta I_n + \zeta 11'$ for some $\theta(\tau^2, \sigma^2) > 0, \zeta(\tau^2, \sigma^2) < 0$.

As a result, the likelihood is

$$\begin{aligned} p(X) &= (2\pi)^{-nm/2} \prod_{i=1}^m \exp\left\{ -\frac{1}{2}(X_i - \mu 1)'\Sigma^{-1}(X_i - \mu 1) \right\} \\ &= (2\pi)^{-nm/2} \prod_{i=1}^m \exp\left\{ \theta \|X_i\|^2/2 - \zeta(\sum_j X_{ij})^2/2 + (n\mu\zeta + \mu\theta)\sum_j X_{ij} - A(\theta, \zeta, \mu) \right\} \\ &= (2\pi)^{-nm/2} \exp\left\{ \theta \sum_i \|X_i\|^2/2 - \zeta \frac{n^2}{2} \sum_i \overline{X}_i^2 + nm(n\mu\zeta + \mu\theta)\overline{X} - A(\theta, \zeta, \mu) \right\}. \end{aligned}$$

This is a full-rank three-parameter exponential family because the natural parameter space contains an open set, so $T = \left(\sum_i \|X_i\|^2, \sum_i \overline{X}_i^2, \overline{X} \right)$ is a complete sufficient statistic.

We have shown in class that $(m-1)S_B^2 + m\overline{X}^2 = \sum_i \overline{X}_i^2$, and $(n-1)S_i^2 + n\overline{X}_i^2 = \|X_i\|^2$. Therefore, we can reconstruct $\left(\overline{X}, \, S_B^2, \, \sum_i S_i^2\right)$ from T and vice-versa.

3. "And if you ever saw it..." (24 points, 6 points / part).

Some useful facts for this problem:

- For $n \in \{0, 1, \ldots\}$ and $p \in [0,1]^d$ with $\sum p_i = 1$, the multinomial density for $X \sim \text{Multinom}(n, p)$ is

$$p_1^{x_1} \cdots p_d^{x_d} \frac{n!}{x_1! \cdots x_d!}, \qquad \text{on } x \in \{0, \ldots, n\}^d \text{ with } \sum_i x_i = n.$$

An ecologist is interested in estimating the total population of reindeer in a wildlife preserve near the North Pole. She makes two visits to the preserve on two consecutive days and looks for reindeer. Each time she finds a reindeer she marks it with a unique identifying tag, so she can tell if she sees the same reindeer twice (in ecology this type of study is called a *capture-recapture* or *mark-recapture* study).

Assume that the same population of n of reindeer is present in the preserve on both days, and each reindeer on each day has the same probability $\pi \in (0, 1)$ of being seen by her, independently across the reindeer and the days (so the detections / non-detections are like $2n$ i.i.d. "coin flips" each with success probability π). Note that n is the unknown parameter of interest and π is an unknown nuisance parameter.

Let N_{11} denote the number of reindeer she sees both days, N_{10} the number she sees the first day not the second, and N_{01} the number she sees the second day but not the first. (Note that N_{00}, the number of reindeer she sees on neither day, is not observed.)

(a) Write down the likelihood for the model as a function of N_{01}, N_{10}, and N_{11} and show that $T = (N_{01} + N_{10}, N_{11})$ is a sufficient statistic for the model.

You do **NOT** need to show a sufficiency reduction from the Bernoulli model of detected/non-detected "coin flips" for each reindeer-day; after all we do not really get to observe the data for that model because we don't know how many reindeer went undetected on both days. Just start with N_{01}, N_{10}, N_{11} as the data and n and π as the parameters.

(b) (*) Show that T is minimal sufficient (for this part you may assume we already know it is sufficient).

(c) Define the estimator

$$\hat{n} = \frac{(N_{01} + N_{10} + 2N_{11})^2}{4N_{11}}.$$

Show that $\hat{n}$ is consistent in the sense that $\hat{n}/n \xrightarrow{p} 1$ as $n \to \infty$ with π fixed.

(d) Find the asymptotic distribution of $\hat{n}$ from part (c) as $n \to \infty$ with π fixed. You should center and scale appropriately so that it has a non-degenerate limiting distribution (that is, after centering and scaling it shouldn't converge in probability to a constant).

3. Solution.

(a) (**Common mistake:** In applying the factorization theorem we need to treat n as an unknown parameter: $n!/(n - N_{11} - N_{10} - N_{01})!$ is not just a function of the data.)

There are four possible outcomes for each reindeer, labeled 00 (undetected twice), 01 (undetected then detected), 10 (detected then undetected), and 11 (detected twice), which occur respectively with probability $p_{00} = (1-\pi)^2$, $p_{01} = p_{10} = \pi(1-\pi)$, and $p_{11} = \pi^2$. The counts $N_{00}, N_{01}, N_{10}, N_{11}$ record how many times each outcome happens, so

$$\begin{aligned}
&(N_{00}, N_{01}, N_{10}, N_{11}) \sim \text{Multinom}(n, ((1-\pi)^2, \pi(1-\pi), \pi(1-\pi), \pi^2)) \\
&\qquad = (1-\pi)^{2N_{00}} \cdot (\pi(1-\pi))^{N_{10}+N_{01}} \cdot \pi^{2N_{11}} \cdot \frac{n!}{N_{00}!N_{01}!N_{10}!N_{11}!}, \\
&\qquad = \left(\frac{\pi}{1-\pi}\right)^{2N_{11}+N_{10}+N_{01}} \cdot \frac{n!}{(n - N_{11} - N_{01} - N_{10})!} \cdot \frac{1}{N_{01}!N_{10}!N_{11}!},
\end{aligned}$$

where the first two factors are functions of $T = (N_{11}, N_{01} + N_{10})$ and the parameters (n, π) but the last factor is only a function of the data. By the factorization theorem, then, T is sufficient.

(b) If T can be computed from the collection of all likelihood ratios (and if it is also sufficient, as we have just shown it is), then it is minimal sufficient. The likelihood ratio between (n, π) and $(\tilde{n}, \tilde{\pi})$ is

$$\left(\frac{\pi(1-\tilde{\pi})}{\tilde{\pi}(1-\pi)}\right)^{2N_1+N_{10}+N_{01}} \cdot \frac{n!}{\tilde{n}!} \cdot \frac{(\tilde{n} - N_{11} - N_{10} - N_{01})!}{(n - N_{11} - N_{10} - N_{01})!}$$

By taking $n = \tilde{n} = N_{11} - N_{10} - N_{01}$ and varying $\pi/\tilde{\pi}$, we can learn $2N_{11} + N_{10} + N_{01}$; whereas by taking $\pi = \tilde{\pi} = 0.5$ and varying n and $\tilde{n}$, we can learn $N_{11} + N_{11} + N_{10} + N_{01}$; knowing both of these is equivalent to knowing T.

(c) Note that $N_{ij}/n \to p_{ij}$ by LLN, since it is an average of n i.i.d. Bern(p_{ij}) random variables which have finite expectation. Dividing by n^2 in the numerator and n in the denominator and applying the continuous mapping theorem, we get

$$\begin{aligned}\frac{\hat{n}}{n} &= \frac{(N_{01}/n + N_{10}/n + 2N_{11}/n)^2}{4N_{11}/n} \\ &\xrightarrow{p} \frac{(2p_{01} + 2p_{11})^2}{4p_{11}} \quad \text{(continuous since } p_{11} > 0\text{)} \\ &= (2\pi(1-\pi) + 2\pi^2)^2/4\pi^2 = 1.\end{aligned}$$

(d) By grouping together the outcomes 01 and 10 we can get a reduced multinomial

$$(N_{00}, N_{10} + N_{01}, N_{11}) \sim \text{Multinom}(n, (p_{00}, 2p_{01}, p_{11})),$$

which is also a sum of n i.i.d. Multinom$(1, (p_{00}, 2p_{01}, p_{11}))$ random variables which have finite variance. Restricting attention to the two entries we actually observe and then applying the CLT gives

$$\frac{1}{\sqrt{n}}\left(\begin{pmatrix} N_{10} + N_{01} \\ N_{11} \end{pmatrix} - \begin{pmatrix} 2np_{01} \\ np_{11} \end{pmatrix}\right) \Rightarrow N_2\left(0, \Sigma\right)$$

where

$$\Sigma = \begin{pmatrix} 2p_{01}(1-2p_{01}) & -2p_{01}p_{11} \\ -2p_{01}p_{11} & p_{11}(1-p_{11}) \end{pmatrix} = \begin{pmatrix} 2\pi(1-\pi)(1-2\pi(1-\pi)) & -2\pi(1-\pi)\pi^2 \\ -2\pi(1-\pi)\pi^2 & \pi^2(1-\pi^2) \end{pmatrix}$$

We will apply delta method to the function $f(t_1, t_2) = (t_1 + 2t_2)^2/4t_2$:

$$\nabla f(t_1, t_2) = \left(1 + \frac{t_1}{2t_2}, 1 - \frac{t_1^2}{4t_2^2}\right).$$

Applying the delta method to $f\left(\frac{N_{10}+N_{01}}{n}, \frac{N_{11}}{n}\right)$ gives

$$\sqrt{n}\left(\frac{\hat{n}}{n} - 1\right) = \sqrt{n}\left(f\left(\frac{N_{10} + N_{01}}{n}, \frac{N_{11}}{n}\right) - f(2p_{10}, p_{11})\right) \Rightarrow N(0, \sigma^2),$$

where $\sigma^2 = \nabla f(2p_{10}, p_{11})'\Sigma\nabla f(2p_{10}, p_{11})$. After a lot of algebra we can simplify $\sigma^2 = (1-\pi)^2/\pi^2$, but we would award full credit for the unsimplified form as described above.

4. Nonlinear regression (24 points, 6 points / part).

Note the Gaussian density is printed in the preamble of Problem 2.

We are given a sample of n pairs (x_i, Y_i) where $x_1, \ldots, x_n \in \mathbb{R}$ are fixed real numbers and

$$Y_i = g(\alpha + \beta x_i) + \varepsilon_i, \qquad \text{where } \varepsilon_i \overset{\text{ind.}}{\sim} N(0, \sigma^2 h(x_i)).$$

Assume (except where otherwise specified) that:

- $g: \mathbb{R} \to \mathbb{R}$ is a known function which is strictly increasing and infinitely differentiable.
- $h: \mathbb{R} \to (0, \infty)$ is a known continuous function.
- $\alpha, \beta \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown

Finally, let $r_i = Y_i - g(\hat{\alpha} + \hat{\beta} x_i)$ denote the ith residual. Throughout the problem, assume we are estimating the parameter vector $(\alpha, \beta, \sigma^2)$ jointly by maximum likelihood; let $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ denote the joint MLE.

(a) Show that the MLE for α and β is found by setting weighted averages of the residuals to 0:

$$\sum_{i=1}^{n} w_i r_i = \sum_{i=1}^{n} w_i r_i x_i = 0,$$

and give explicit expressions for the weights w_i in terms of the data, the functions g and h, and the maximum likelihood estimators $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$.

(b) Give an explicit expression for the MLE for σ^2, i.e. $\hat{\sigma}^2$, in terms of the data, the functions g and h, and the maximum likelihood estimators $\hat{\alpha}, \hat{\beta}$.

(c) (*) Now assume (for this part **ONLY**) that instead of fixed numbers we observe i.i.d. random variables $X_1, \ldots, X_n$, which are continuous and bounded random variables ($|X_i| \leq B$ almost surely, for some $B > 0$.) Give the asymptotic distribution of the maximum likelihood estimators $(\hat{\alpha}, \hat{\beta})$ in terms of the functions g and h, and expectations of suitable random variables. The limit is taken as $n \to \infty$ with the other parameters fixed.

You may assume without proof that $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ are consistent for the true population values, and that all of the regularity conditions from class for our theorem on the asymptotic distribution of the MLE hold (the log-likelihood and its derivatives are well-behaved in the required sense). You do not need to write down what the conditions are, either.

(**Hint:** it might be easier to do the problem assuming σ^2 is known, and then explain why the answer doesn't change when σ^2 is unknown.)

(d) (*) We now go back to assuming the x_i values are fixed. Now assume $h(z) \equiv 1$ but g is completely unknown (apart from the restrictions described in the preamble: strictly increasing and infinitely differentiable). Give a finite-sample test of $H_0: \ \beta \leq 0$ vs $H_1: \ \beta > 0$. You should provide a test statistic and describe how to calculate the critical value. For full credit you must show your test controls the rejection probability throughout the composite null hypothesis (that is, for all valid choices of g, α, and σ^2.)

Since we are already using the letter α for the intercept, I suggest using a to denote the significance level in your answer.

4. Solution

(a) The log likelihood is

$$\frac{1}{2\sigma^2}\sum_{i=1}^n \left[\frac{(Y_i - g(\alpha+\beta x_i))^2}{h(x_i)} - \frac{1}{2}\log(2\pi\sigma^2 h(x_i))\right]$$

The first derivative with respect to α is

$$\frac{1}{\sigma^2}\sum_{i=1}^n \left[(Y_i - g(\alpha+\beta x_i))\frac{\dot{g}(\alpha+\beta x_i)}{h(x_i)}\right] = \frac{1}{\sigma^2}\sum_i r_i w_i,$$

for $w_i = \dot{g}(\alpha+\beta x_i)/h(x_i)$. Similarly, the gradient with respect to β is

$$\frac{1}{\sigma^2}\sum_{i=1}^n r_i w_i x_i.$$

Any local minimizer of the log-likelihood sets the gradient equal to zero. (**Remark:** we gave full credit for just setting the derivative to zero, but the question statement should have been clearer about the difference between local optimality and global optimality: it is a necessary but not sufficient condition that the gradient should be zero, but it is possible for there to be more than one local minimum. It is also possible, I realized later, to come up with counterexamples where there is no MLE because the likelihood is maximized at infinity.)

Note that our estimate of σ^2 plays no role in determining $\hat{\alpha}$ and $\hat{\beta}$; we would get the same estimators for α and β whether we estimate σ^2 or whether it is known. This will be useful in part (c).

(b) Differentiating with respect to σ^2 gives

$$-\frac{1}{2\sigma^4}\sum_{i=1}^n \left[\frac{(Y_i - g(\alpha+\beta x_i))^2}{h(x_i)}\right] - \frac{n}{2\sigma^2},$$

and setting the derivative equal to 0 while the other parameters are at their MLEs gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left[\frac{(Y_i - g(\hat{\alpha}+\hat{\beta} x_i))^2}{h(x_i)}\right]$$

(c) First assume $\sigma^2 > 0$ is known. Then the only two unknown parameters are α and β, and the score is

$$\frac{1}{\sigma^2}\left(\sum_i r_i w_i, \sum_i r_i w_i X_i\right).$$

The variance of the score conditional on X is

$$\begin{aligned}
&\operatorname{Var}\left(\sum_i (Y_i - g(\alpha+\beta X_i))\frac{\dot{g}(\alpha+\beta X_i)}{\sigma^2 h(X_i)}\begin{pmatrix}1\\ X_i\end{pmatrix} \mid X_i\right)\\
&\quad = \sum_i \frac{\dot{g}(\alpha+\beta X_i)^2}{\sigma^4 h(X_i)^2}\operatorname{Var}(Y_i - g(\alpha+\beta X_i) \mid X_i)\begin{pmatrix}1 & X_i\\ X_i & X_i^2\end{pmatrix}\\
&\quad = \sum_i \frac{\dot{g}(\alpha+\beta X_i)^2}{\sigma^2 h(X_i)}\begin{pmatrix}1 & X_i\\ X_i & X_i^2\end{pmatrix}.
\end{aligned}$$

The expectation of the score given X is zero, so the marginal variance is simply the expectation of the conditional variance:

$$J(\alpha,\beta) = \operatorname{Var}_{\alpha,\beta}(\nabla\ell(\alpha,\beta)) = \frac{n}{\sigma^2}\mathbb{E}\left[\frac{\dot{g}(\alpha+\beta X_i)^2}{h(X_i)}\begin{pmatrix}1 & X_i\\ X_i & X_i^2\end{pmatrix}\right].$$

Note that the exam should have guaranteed $h(X_i)$ didn't have positive density at zero; that could make the expectation infinite. Assuming it is not infinite though, and the conditions hold for our theorem on the asymptotic distribution, then we have

$$\sqrt{n}\left(\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} - \begin{pmatrix}\alpha\\ \beta\end{pmatrix}\right) \Rightarrow N_2(0, J(\alpha,\beta)^{-1}).$$

If instead σ^2 is unknown, nothing actually changes because, as noted in part (a), the MLE $(\hat{\alpha},\hat{\beta})$ is the same regardless of σ^2, which merely scales the log-likelihood up or down. Since it is the same random variable regardless of whether σ^2 is known or estimated, it also has the same limiting distribution regardless.

(d) Let $\mu_i = g(\alpha + \beta x_i)$ and assume without loss of generality that μ_i is non-decreasing in i. If $\beta = 0$ then $\mu_1 = \cdots = \mu_n$ and the Y_i values are i.i.d., but if $\beta > 0$ then the means are increasing too (and if $\beta < 0$ the means are decreasing). We can use a permutation test whose test statistic is meant to pick up correlation between x and μ, for example $T(Y) = x'Y$. We use a Monte Carlo version of the permutation test as usual: take B random permutations and reject if $T(Y)$ is among the $\lfloor a(B+1) \rfloor$ largest of $T(Y), T(\pi_1 Y), \ldots, T(\pi_B Y)$.

If $Y_i = \mu_i + \epsilon_i$ and assume $\beta \leq 0$. Then for a generic permutation π,

$$T(\pi Y) = x'(\pi\mu) + x'(\pi\epsilon).$$

Note that $(x'\epsilon, x'(\pi_1\epsilon), x'(\pi_B\epsilon))$ are exchangeable no matter what, but $x'\mu \leq x'(\pi\mu)$ for all π; therefore $T(Y)$ has a less than a chance of being among the $\lfloor a(B+1) \rfloor$ largest values.