MATH95012

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2020

This paper is also taken for the relevant examination for the
Associateship of the Royal College of Science

Statistical Modelling

Date: 15th May 2020

Time: 13.00pm – 15.00pm (BST)

Time Allowed: 2 Hours

Upload Time Allowed: 30 Minutes

This paper has 4 Questions.

Candidates should start their solutions to each question on a new sheet of paper.

Each sheet of paper should have your CID, Question Number and Page Number on the top.

Only use 1 side of the paper.

Allow margins for marking.

Any required additional material(s) will be provided.

Credit will be given for all questions attempted.

Each question carries equal weight.

SUBMIT YOUR ANSWERS AS SEPARATE PDFs TO THE RELEVANT DROPBOXES ON BLACKBOARD (ONE FOR EACH QUESTION) WITH COMPLETED COVERSHEETS WITH YOUR CID NUMBER, QUESTION NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.

1. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p) random variables where $p \in (0,1)$ is unknown, and let $\Delta_1, \ldots, \Delta_n$ be i.i.d. Bernoulli(π_0) random variables, independent of the X_i, where $\pi_0 \in (0,1)$ is known. Define $Y_i := \Delta_i X_i$. We observe a realisation $(y_1, \delta_1), \ldots, (y_n, \delta_n)$ of the random sample $(Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)$. Hence, the true value of x_i is missing (unobserved) if $\delta_i = 0$.

(a) The joint pmf of (Y_1, Δ_1) has the form

$$f_p(y, \delta) = \begin{cases} \pi_0^{\delta}(1-\pi_0)^{1-\delta}\{p^y(1-p)^{1-y}\}^{\delta}, & (y,\delta) \in \mathcal{S} \\ 0, & (y,\delta) \notin \mathcal{S} \end{cases}$$

where $\mathcal{S}$ is the support of $f_p(y, \delta)$. Find $\mathcal{S}$ and then verify that $f_p(y, \delta)$ is a valid pmf.

(2 marks)

(b) Based on $f_p(y, \delta)$ in part (a):

(i) Compute the Fisher information, $I_n(p)$, for a random sample of size n.

(4 marks)

(ii) Find the Rao-Cramer lower bound for the variance of any unbiased estimator of p.

(2 marks)

(c) Consider the estimator $\tilde{p} := \frac{1}{n\pi_0} \sum_{i=1}^{n} Y_i$.

(i) Show that $\tilde{p}$ is an unbiased estimator of p. (2 marks)

(ii) Does Var$_p(\tilde{p})$ achieve the Rao-Cramer lower bound? (2 marks)

(d) Find the maximum likelihood estimator, $\hat{p}$, of p. (4 marks)

(e) In a particular study, the sample size n is quite large. Considering the large sample properties of each estimator, state whether you would prefer to report the estimate $\hat{p}$ or $\tilde{p}$ and give your reasons for this choice.

(4 marks)

(Total: 20 marks)

2. Let $X \mid \theta \sim$ Binomial(n, θ), and suppose we have specified a prior distribution $\theta \sim$ Beta(α, β) with density

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1, \quad \alpha > 0, \beta > 0.$$

You may use the fact that if $\theta \sim$ Beta(α, β), then $E(\theta) = \frac{\alpha}{\alpha+\beta}$ and Var$(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

(a) Find the posterior distribution $p(\theta \mid x)$. (4 marks)

(b) Compute the posterior mean, $\hat{\theta}_B$. Show that $\hat{\theta}_B$ can be expressed as a weighted average of the prior mean based on $\pi(\theta)$ and the sample proportion x/n. Describe the behavior of $\hat{\theta}_B$ for large and small values of α and β (you may take $\alpha = \beta$ to simplify this discussion). (4 marks)

(c) For parts (c) and (d), we now consider the classical (frequentist) interpretation of probability (i.e., we consider the parameter $\theta \in (0, 1)$ to be fixed).

(i) Define the mean squared error (MSE) of a random variable T as an estimator of θ. (2 marks)

(ii) State and prove the relationship between MSE and the variance and bias of T. (2 marks)

(d) Consider $\hat{\theta}_B$ as an estimator of θ in a frequentist setting.

(i) Find the MSE of $\hat{\theta}_B$. (4 marks)

(ii) Find values of $\alpha > 0$ and $\beta > 0$ such that the MSE of $\hat{\theta}_B$ is constant as a function of θ. What is the MSE of $\hat{\theta}_B$ at these values? On the basis of MSE, would you prefer the estimator $\hat{\theta}_B$ or $\hat{\theta} = X/n$ if $\theta = 1/2$? Would your preference change if $\theta = 0$? (4 marks)

(Total: 20 marks)

3. In this question, we consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for non-random $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, and random $\boldsymbol{\epsilon} \in \mathbb{R}^n$, with $\mathsf{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathsf{Cov}(\boldsymbol{\epsilon}) := \boldsymbol{\Sigma}$.

(a) Assuming that $\mathbf{X}$ has full rank (FR) and that $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ for some $\sigma^2 > 0$, derive the mean and covariance of the least squares estimator

$$\hat{\boldsymbol{\beta}} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

[Note: $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix.] (4 marks)

(b) State the Gauss-Markov theorem, defining any technical terms in the statement. (3 marks)

(c) Derive a level $1-\alpha$ confidence region for $\boldsymbol{\beta}$ under the normal theory assumptions based on the statistic

$$A = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\mathsf{RSS}}\frac{n-p}{p},$$

where RSS is the *residual sum of squares*. Your answer should include justification for the distribution of A. You may use results from lecture without proof, but must verify that the assumptions are met. (4 marks)

(d) For a fixed $k \geq 0$, the *ridge estimator* of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}(k) := (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$. [Note: $\mathbf{I} \in \mathbb{R}^{p\times p}$ is the identity matrix.]

(i) Show that, for $k > 0$, $\hat{\boldsymbol{\beta}}(k)$ exists even if the FR assumption does not hold. (3 marks)

(ii) For $k > 0$, is $\hat{\boldsymbol{\beta}}(k)$ an unbiased estimator of $\boldsymbol{\beta}$? (2 marks)

(iii) Let $\boldsymbol{c} \in \mathbb{R}^n$ be given and suppose that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. Assuming SOA and FR, show that $\mathsf{Var}\{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}(k)\} \leq \mathsf{Var}\{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}\}$. Does this violate the Gauss-Markov theorem? Explain your answer.

(4 marks)

(Total: 20 marks)

4. Suppose that $X_1, \dots, X_m$ are i.i.d. $N(\mu, \sigma^2)$, and that $Y_1, \dots, Y_n$ are i.i.d. $N(\nu, \tau^2)$, independent of the X_js. Let $N := m + n$ denote the total sample size.

(a) Assuming that $\sigma^2 = \tau^2$, express the above setting as a linear model

$$\mathbf{Z} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\beta} := (\mu, \nu)^T$. In your answer, clearly define $\mathbf{Z}$, $\mathbf{W}$, and $\boldsymbol{\epsilon}$ in this model and verify that the normal theory assumptions hold. (4 marks)

(b) Let $\bar{X}_m := \frac{1}{m}\sum_{i=1}^m X_i$ and $\bar{Y}_n := \frac{1}{n}\sum_{j=1}^n Y_i$. Show that $\hat{\boldsymbol{\beta}} = (\bar{X}_m, \bar{Y}_n)^T$ is the least squares estimator of $\boldsymbol{\beta}$. (3 marks)

(c) Let $S_X^2 := \frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X}_m)^2$ and $S_Y^2 := \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. Show that the unbiased estimator $\hat{\sigma}_N^2 :=$ RSS$/(N-2)$ of σ^2 can be expressed as

$$\hat{\sigma}_N^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{N-2}.$$

(3 marks)

(d) Suppose we wish to test the hypothesis $H_0 : \boldsymbol{c}^T\boldsymbol{\beta} = 0$ against $H_1 : \boldsymbol{c}^T\boldsymbol{\beta} \neq 0$ for $\boldsymbol{c} = (1, -1)^T$. Assuming that $\sigma^2 = \tau^2$, show that the t-statistic of this hypothesis can be written as

$$T_N = \frac{\bar{X}_m - \bar{Y}_n}{\hat{\sigma}_N\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

which is the two-sample t-test statistic. What is the exact distribution of T_N under H_0? (3 marks)

(e) Now suppose that σ^2 and τ^2 are arbitrary and $\lambda_N := m/N \to \lambda \in (0,1)$ as $N \to \infty$. Show that under H_0

$$T_N \to_d N(0, V)$$

as $N \to \infty$ and find an explicit expression for V in terms of σ^2, τ^2 and λ. Clearly state any results that you use to show this convergence. (5 marks)

(f) Find two different conditions such that, under H_0, $V = 1$. Describe the impact of $V \neq 1$ on the performance of asymptotic confidence intervals that assume $V = 1$. (2 marks)

(Total: 20 marks)

1. (a) *[Seen Similar]* (2A marks) The support $\mathcal{S} := \{(0,0),(0,1),(1,1)\}$ since $\mathsf{P}(Y_1 = 1, \Delta_1 = 0) = 0$. Moreover, $f_p(y,\delta)$ is clearly non-negative and

$$\sum_{(y,\delta)\in\mathcal{S}} f_p(y,\delta) = f_p(0,0) + f_p(0,1) + f_p(1,1) = (1-\pi_0) + \pi_0(1-p) + \pi_0 p = 1.$$

Students should get 1 mark if they identify the support correctly, but do not verify both that the pmf is non-negative and sums to one.

(b) (i) *[Seen Similar]* (4B marks) The Fisher information for a sample of size $n = 1$ is

$$I_1(p) = E_p\left[\left\{\frac{\partial}{\partial p}\log f_p(Y,\Delta)\right\}^2\right] = -E_p\left\{\frac{\partial^2}{\partial p^2}\log f_p(Y,\Delta)\right\}.$$

We will use the latter. First, we find

$$\begin{aligned}\log f_p(y,\delta) &= \delta\log\pi_0 + (1-\delta)(1-y)\log\pi_0 + \delta\{y\log p + (1-y)\log(1-p)\}\\ \frac{\partial}{\partial p}\log f_p(y,\delta) &= \delta\left[\frac{y}{p} - \frac{1-y}{1-p}\right] = \delta\left[\frac{y-p}{p(1-p)}\right]\\ \frac{\partial^2}{\partial p^2}\log f_p(y,\delta) &= -\delta\left[\frac{y}{p^2} + \frac{1-y}{(1-p)^2}\right]\end{aligned}$$

Hence,

$$\begin{aligned}I_1(p) &= -E_p\left\{-\Delta\left[\frac{Y}{p^2} + \frac{1-Y}{(1-p)^2}\right]\right\}\\ &= E_p\left\{\frac{\Delta Y}{p^2} + \frac{\Delta - \Delta Y}{(1-p)^2}\right\}\\ &= E_p\left\{\frac{\Delta X}{p^2} + \frac{\Delta - \Delta X}{(1-p)^2}\right\}\\ &= \frac{\pi_0 p}{p^2} + \frac{\pi_0 - \pi_0 p}{(1-p)^2} = \frac{\pi_0}{p(1-p)}.\end{aligned}$$

Then, because of the additive property of information, $I_n(p) = nI_1(p) = \frac{n\pi_0}{p(1-p)}$.
Other correct approaches include use of the other definition of information, or computing the information for a sample of size n directly from the likelihood/joint pmf of the sample.

(ii) *[Seen Similar]* (2B marks) Let T be any unbiased estimator of p based on a sample of size n. Then

$$\mathsf{Var}_p(T) \geq \frac{1}{I_n(p)} = \frac{p(1-p)}{n\pi_0}.$$

(c) (i) *[Seen Similar]* (2A marks) Noting that the $Y_i = \Delta_i X_i$ are i.i.d. Bernoulli($\pi_0 p$), we have

$$E_p(\tilde{p}) = \frac{1}{n\pi_0}\sum_{i=1}^{n} E_p(Y_i) = \frac{E_p(Y_1)}{\pi_0} = \frac{\pi_0}{\pi_0}p = p,$$

for each $p \in (0,1)$. Hence, is an unbiased estimator of p.

(ii) *[Unseen]* (2A marks)

$$\mathsf{Var}_p(\tilde{p}) = \frac{1}{n^2\pi_0^2}\sum_{i=1}^{n}\mathsf{Var}_p(Y_i) = \frac{n\pi_0 p(1-\pi_0 p)}{n^2\pi_0^2} = \frac{p(1-\pi_0 p)}{n\pi_0} \geq \frac{p(1-p)}{n\pi_0}$$

with equality holding when $\pi_0 = 1$. Since $\pi_0 \in (0,1)$, $\mathsf{Var}(\tilde{p})$ does not attain the lower bound.

(d) *[Seen Similar]* (4A marks) The maximum likelihood estimator is the solution to

$$\sum_{i=1}^{n} \frac{\partial}{\partial p} \log f_p(Y_i, \Delta_i) = 0.$$

From before, we have

$$\frac{\partial}{\partial p} \log f_p(y, \delta) = \delta \left[\frac{y - p}{p(1-p)} \right].$$

Hence,

$$\sum_{i=1}^{n} \frac{\partial}{\partial p} \log f_p(Y_i, \Delta_i) = \sum_{i=1}^{n} \Delta_i \left[\frac{Y_i - p}{p(1-p)} \right] = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} \Delta_i Y_i = p \sum_{i=1}^{n} \Delta_i.$$

This gives the MLE

$$\hat{p} = \frac{\sum_{i=1}^{n} \Delta_i Y_i}{\sum_{i=1}^{n} \Delta_i} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \Delta_i}.$$

From before, for any $p \in (0, 1)$, we have

$$\frac{\partial^2}{\partial p^2} \log f_p(y, \delta) = -\delta \left[\frac{y}{p^2} + \frac{1-y}{(1-p)^2} \right] \leq 0,$$

which readily implies that $\hat{p}$ maximises the log-likelihood.

(e) *[Unseen]* (4D marks) We would prefer to report the MLE $\hat{p}$. By the central limit theorem,

$$\sqrt{n}(\tilde{p} - p) \to_d N(0, p(1 - \pi_0 p)/\pi_0)$$

and by results for the asymptotic properties of the MLE,

$$\sqrt{n}(\hat{p} - p) \to_d N(0, I_1(p)^{-1}) = N(0, p(1-p)/\pi_0).$$

Hence, both estimators are asymptotically unbiased and the variance of $\hat{p}$ will be lower than that of $\tilde{p}$ in sufficiently large samples.

2. (a) *[Seen Similar]* (4A marks) We will ignore multiplicative constants:

$$\begin{aligned} p(\theta \mid x) &\propto p(x \mid \theta)\pi(\theta) \\ &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}. \end{aligned}$$

In order for $p(\theta \mid x)$ to integrate to one, it must be the density of Beta$(\alpha+x, \beta+n-x)$.

(b) *[Seen Similar]* (4B marks) Based on the posterior Beta$(\alpha+x, \beta+n-x)$ distribution, the posterior mean is

$$\hat{\theta}_B = \frac{\alpha+x}{\alpha+\beta+n} = \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right) + \frac{n}{\alpha+\beta+n}\left(\frac{x}{n}\right).$$

From this expression for $\hat{\theta}_B$, we see that if α and β are small relative to n, $\hat{\theta}_B$ approaches x/n. If, instead, α and β are large relative to n, $\hat{\theta}_B$ approaches $\alpha/(\alpha+\beta)$. In the latter case, if $\alpha = \beta$, then $\hat{\theta}_B$ approaches $1/2$.

(c) *[Seen]*

(i) (2A marks) $MSE_\theta(T) := E_\theta(T-\theta)^2$.

(ii) (2A marks) We have

$$\begin{aligned} MSE_\theta(T) &= E_\theta(T - E_\theta(T) + E_\theta(T) - \theta)^2 \\ &= E_\theta(T - E_\theta(T))^2 + (E_\theta(T)-\theta)^2 - 2E_\theta\{(E_\theta(T)-\theta)(T-E_\theta(T))\} \\ &= E_\theta(T - E_\theta(T))^2 + (E_\theta(T)-\theta)^2 + 0 \\ &= \mathsf{Var}_\theta(T) + \mathsf{bias}_\theta(T)^2. \end{aligned}$$

Partial credit (1 mark) is to be awarded if the relationship is stated but not proven.

(d) *[Unseen]*

(i) (4C marks) From part (b), we have

$$\begin{aligned} \mathsf{bias}_\theta(\hat{\theta}_B) &= E_\theta(\hat{\theta}_B) - \theta \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right) + \frac{n}{\alpha+\beta+n}E_\theta\left(\frac{X}{n}\right) - \theta \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right) + \frac{n}{\alpha+\beta+n}\theta - \theta \\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right) - \frac{\alpha+\beta}{\alpha+\beta+n}\theta \\ &= \frac{\alpha - \theta(\alpha+\beta)}{\alpha+\beta+n} \end{aligned}$$

and

$$\begin{aligned} \mathsf{Var}_\theta(\hat{\theta}_B) &= \mathsf{Var}_\theta\left(\frac{n}{\alpha+\beta+n}\left(\frac{X}{n}\right)\right)^2 \\ &= \left(\frac{n}{\alpha+\beta+n}\right)^2\frac{\theta(1-\theta)}{n} \\ &= \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2}. \end{aligned}$$

We then find that

$$\begin{aligned} MSE_\theta(\hat{\theta}_B) &= \mathsf{Var}_\theta(\hat{\theta}_B) + \mathsf{bias}_\theta(\hat{\theta}_B)^2 \\ &= \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2} + \left(\frac{\alpha-\theta(\alpha+\beta)}{\alpha+\beta+n}\right)^2 \\ &= \frac{n\theta(1-\theta)}{(\alpha+\beta+n)^2} + \frac{(\alpha-\theta(\alpha+\beta))^2}{(\alpha+\beta+n)^2} \\ &= \frac{\alpha^2}{(\alpha+\beta+n)^2} + \frac{n\theta(1-\theta)+\theta^2(\alpha+\beta)^2-2\theta\alpha(\alpha+\beta)}{(\alpha+\beta+n)^2} \\ &= \frac{\alpha^2}{(\alpha+\beta+n)^2} + \frac{\theta^2\{(\alpha+\beta)^2-n\}+\theta\{n-2\alpha(\alpha+\beta)\}}{(\alpha+\beta+n)^2}. \end{aligned}$$

Simplification is not required here, but is helpful for the next part.

(ii) (4D marks) From above $MSE_\theta(\hat{\theta}_B)$ is constant as a function of θ if

$$(\alpha+\beta)^2 = n \quad \text{and} \quad n = 2\alpha(\alpha+\beta).$$

From the second equality, we find that $(\alpha+\beta) = n/(2\alpha)$. Plugging this into the first equality, we have that

$$n^2/(2\alpha)^2 = n \quad \Leftrightarrow \quad \alpha = \frac{\sqrt{n}}{2}.$$

Substituting $\alpha = \frac{\sqrt{n}}{2}$ into either constraint yields $\beta = \alpha = \frac{\sqrt{n}}{2}$. For these values (2 marks)

$$MSE_\theta(\hat{\theta}_B) = \frac{\alpha^2}{(\alpha+\beta+n)^2} = \frac{n}{4(\sqrt{n}+n)^2}.$$

The remaining 2 marks are for the comparison of the two estimators of θ. Since $\hat{\theta}$ is unbiased, $MSE_\theta(\hat{\theta}) = \mathsf{Var}_\theta(\hat{\theta}) = \theta(1-\theta)/n$, which depends on both n and θ. For fixed n and $\theta = 1/2$, $MSE_\theta(\hat{\theta}_B) < MSE_\theta(\hat{\theta}) = (4n)^{-1}$. Hence, if $\theta = 1/2$, we prefer $\hat{\theta}_B$.
For fixed n and $\theta = 0$, $MSE_\theta(\hat{\theta}_B) > MSE_\theta(\hat{\theta}) = 0$. If $\theta = 0$, we instead prefer $\hat{\theta}$.

3. In this question, we consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for non-random $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$, with $n > p$ and random $\boldsymbol{\epsilon} \in \mathbb{R}^n$, with Cov$(\boldsymbol{\epsilon}) := \boldsymbol{\Sigma}$.

(a) *[Seen]* (4A marks) By linearity of expectations,

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

From properties of the covariance,

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

(b) *[Seen]* (3B marks) Assume FR and SOA. For $\boldsymbol{c} \in \mathbb{R}^p$, and any linear unbiased estimator $\mathbf{L}^T\mathbf{Y}$, $E[\mathbf{L}^T\mathbf{Y}] = \boldsymbol{c}^T\boldsymbol{\beta}$, of $\boldsymbol{c}^T\boldsymbol{\beta}$, we have

$$\text{Var}(\mathbf{L}^T\mathbf{Y}) \geq \text{Var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}).$$

That is, $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{c}^T\boldsymbol{\beta}$.

(c) *[Seen]* (4C marks) For $\mathbf{Z} = \frac{1}{\sigma}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ we can write

$$A = \frac{\mathbf{Z}^T\mathbf{P}\mathbf{Z}/p}{\mathbf{Z}^T\mathbf{Q}\mathbf{Z}/(n-p)}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{Q} = \mathbf{I} - \mathbf{P}$ are projection matrices that sum to the identity matrix. It follows from the Fisher-Cochran theorem that $\mathbf{Z}^T\mathbf{P}\mathbf{Z} \sim \chi^2_p$ is independent of $\mathbf{Z}^T\mathbf{Q}\mathbf{Z} \sim \chi^2_{n-p}$. Hence, $A \sim F_{p,n-p}$ and a $1-\alpha$ confidence region for $\boldsymbol{\beta}$ is defined as

$$\{\gamma \in \mathbb{R}^p : A \leq F_{p,n-p,\alpha}\}$$

for the value $F_{p,n-p,\alpha}$ such that $P(X \geq F_{p,n-p,\alpha}) = \alpha$ for $X \sim F_{p,n-p}$.

(d) (i) *[Unseen]* (3D marks) To begin, we note that the ridge estimator exists if $\mathbf{X}^T\mathbf{X} + k\mathbf{I}$ is invertible. Clearly, $\mathbf{X}^T\mathbf{X} + k\mathbf{I}$ is symmetric. Hence, the matrix is invertible if it is positive definite. We can establish this property in multiple ways – students need only provide one correct method.

- For instance, one can note that for any non-zero vector $\boldsymbol{c}$,

$$\boldsymbol{c}^T(\mathbf{X}^T\mathbf{X} + k\mathbf{I})\boldsymbol{c} = \boldsymbol{c}^T\mathbf{X}^T\mathbf{X}\boldsymbol{c} + k\boldsymbol{c}^T\boldsymbol{c} = (\mathbf{X}\boldsymbol{c})^T(\mathbf{X}\boldsymbol{c}) + k\boldsymbol{c}^T\boldsymbol{c} > k\boldsymbol{c}^T\boldsymbol{c} > 0.$$

  Indeed, this shows that $\mathbf{X}^T\mathbf{X} + k\mathbf{I}$ satisfies the definition of a positive definite matrix.

- Another approach is to make use of the fact that if the eigenvalues of $\mathbf{X}^T\mathbf{X} + k\mathbf{I}$ are strictly positive, then the matrix is positive definite (and hence, non-singular). Let λ_j be an eigenvalue of $\mathbf{X}^T\mathbf{X}$ with eigenvector $\boldsymbol{v}_j$. Then

$$(\mathbf{X}^T\mathbf{X} + k\mathbf{I})\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j + k\boldsymbol{v}_j = (\lambda_j + k)\boldsymbol{v}.$$

  Because $\mathbf{X}^T\mathbf{X}$ is positive semi-definite, $\lambda_j \geq 0$. Hence, the eigenvalues of $\mathbf{X}^T\mathbf{X} + k\mathbf{I}$ are all of the form $(\lambda_j + k) > 0$.

Hence, $\hat{\boldsymbol{\beta}}(k)$ exists even if $\mathbf{X}$ does not have full rank.

(ii) *[Seen Method]* (2B marks) Let $k > 0$. By linearity of expectation,

$$E(\hat{\boldsymbol{\beta}}(k)) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \neq \boldsymbol{\beta}.$$

Hence, $\hat{\boldsymbol{\beta}}(k)$ is a biased estimator.

(iii) *[Unseen]* (4C marks) We have that

$$\begin{aligned}\mathsf{Var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}(k)) &= \boldsymbol{c}^T\mathsf{Cov}(\hat{\boldsymbol{\beta}}(k))\boldsymbol{c}\\ &= \boldsymbol{c}^T(\mathbf{X}^T\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^T\mathsf{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X}+k\mathbf{I})^{-1}\boldsymbol{c}\\ &= \boldsymbol{c}^T(\mathbf{I}+k\mathbf{I})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{I}+k\mathbf{I})^{-1}\boldsymbol{c}\\ &= \boldsymbol{c}^T(1+k)^{-1}\mathbf{I}(1+k)^{-1}\boldsymbol{c}\sigma^2\\ &= \sigma^2\boldsymbol{c}^T\boldsymbol{c}(1+k)^{-2}\end{aligned}$$

Since $\mathsf{Var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}(0)) = \mathsf{Var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}})$,

$$\mathsf{Var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}(k)) = \sigma^2\boldsymbol{c}^T\boldsymbol{c}(1+k)^{-2} \le \sigma^2\boldsymbol{c}^T\boldsymbol{c} = \mathsf{Var}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}}).$$

This does not violate the Gauss-Markov theorem, since $\boldsymbol{c}^T\hat{\boldsymbol{\beta}}(k)$ is biased for estimating $\boldsymbol{c}^T\boldsymbol{\beta}$.

4. (a) *[Seen Similar]* (4A marks) We have that

$$\mathbf{Z} := \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \in \mathbb{R}^N, \quad \mathbf{W} := \begin{pmatrix} \mathbf{1}_m & \mathbf{0}_m \\ \mathbf{0}_n & \mathbf{1}_n \end{pmatrix} \in \mathbb{R}^{N\times 2}, \quad \boldsymbol{\epsilon} := \mathbf{Z} - \mathbf{W}\boldsymbol{\beta} \sim N(\mathbf{0}_N, \sigma^2 \mathbf{I}_{N\times N}),$$

where $\mathbf{1}_k \in \mathbb{R}^k$ denotes a vector of ones, $\mathbf{0}_k \in \mathbb{R}^k$ denotes a vector of zeros, and $\mathbf{I}_{k\times k}$ is the identity matrix in $\mathbb{R}^{k\times k}$. Since $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. $N(0, \sigma^2)$, the normal theory assumptions hold.

(b) *[Seen]* (3A marks) We have that

$$(\mathbf{W}^T\mathbf{W})^{-1} = \begin{pmatrix} m & 0 \\ 0 & n \end{pmatrix}^{-1} = \begin{pmatrix} 1/m & 0 \\ 0 & 1/n \end{pmatrix}$$

and

$$\mathbf{W}^T\mathbf{Z} = \begin{pmatrix} m\bar{X}_m \\ n\bar{Y}_n \end{pmatrix}.$$

Hence,

$$\hat{\boldsymbol{\beta}} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z} = \begin{pmatrix} 1/m & 0 \\ 0 & 1/n \end{pmatrix} \begin{pmatrix} m\bar{X}_m \\ n\bar{Y}_n \end{pmatrix} = \begin{pmatrix} \bar{X}_m \\ \bar{Y}_n \end{pmatrix}.$$

(c) *[Seen Similar]* (3A marks) Starting with $\hat{\sigma}^2_N := \mathsf{RSS}/(N-2)$, we have

$$\begin{aligned} \frac{\mathsf{RSS}}{N-2} &= \frac{(\mathbf{Z} - \mathbf{W}\hat{\boldsymbol{\beta}})^T(\mathbf{Z} - \mathbf{W}\hat{\boldsymbol{\beta}})}{N-2} \\ &= \frac{\sum_{i=1}^N (Z_i - (\mathbf{W}\hat{\boldsymbol{\beta}})_i)^2}{N-2} \\ &= \frac{\sum_{i=1}^m (X_i - \bar{X}_m)^2 + \sum_{j=1}^n (Y_j - \bar{Y}_n)^2}{N-2} \\ &= \frac{(m-1)S_X^2 + (n-1)S_Y^2}{N-2}. \end{aligned}$$

(d) *[Seen Method]* (3B marks) The t-statistic has the general form

$$T_N = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2_N \boldsymbol{c}^T(\mathbf{W}^T\mathbf{W})^{-1}\boldsymbol{c}}}.$$

Clearly, when $\mu = \nu$, $\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta} = (\bar{X}_m - \bar{Y}_n) - (\mu - \nu) = \bar{X}_m - \bar{Y}_n$. Also,

$$\boldsymbol{c}^T(\mathbf{W}^T\mathbf{W})^{-1}\boldsymbol{c} = \boldsymbol{c}^T \begin{pmatrix} 1/m & 0 \\ 0 & 1/n \end{pmatrix} \boldsymbol{c} = \frac{1}{m} + \frac{1}{n}.$$

Substituting these quantities into the formula for T_N yields the first part of the question. Moreover, we know that under H_0, $T_N \sim t_{N-2}(0)$.

(e) *[Unseen]* (5D marks) We know that under H_0, for arbitrary σ^2 and τ^2 that

$$\bar{X}_m - \bar{Y}_n \sim N(0, \frac{\sigma^2}{m} + \frac{\tau^2}{n})$$

and hence, that

$$U_N := \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}} \sim N(0, 1).$$

We can write

$$T_N = U_N \frac{\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}}{\sqrt{\frac{\hat{\sigma}_N^2}{m} + \frac{\hat{\sigma}_N^2}{n}}} = U_N \frac{\sqrt{\frac{N\sigma^2}{m} + \frac{N\tau^2}{n}}}{\sqrt{\frac{N\hat{\sigma}_N^2}{m} + \frac{N\hat{\sigma}_N^2}{n}}}.$$

For arbitrary σ^2 and τ^2, by the weak law of large numbers and continuity

$$\hat{\sigma}_N^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{N-2} \to_p \lambda\sigma^2 + (1-\lambda)\tau^2.$$

Moreover, by continuity of the limits

$$\sqrt{\frac{N\hat{\sigma}_N^2}{m} + \frac{N\hat{\sigma}_N^2}{n}} \to_p \sqrt{\frac{\lambda\sigma^2 + (1-\lambda)\tau^2}{\lambda} + \frac{\lambda\sigma^2 + (1-\lambda)\tau^2}{1-\lambda}} = \sqrt{\frac{\lambda\sigma^2 + (1-\lambda)\tau^2}{\lambda(1-\lambda)}}$$

and

$$\sqrt{\frac{N\sigma^2}{m} + \frac{N\tau^2}{n}} \to \sqrt{\frac{\sigma^2}{\lambda} + \frac{\tau^2}{1-\lambda}} = \sqrt{\frac{(1-\lambda)\sigma^2 + \lambda\tau^2}{\lambda(1-\lambda)}}.$$

Hence, by Slutsky's lemma,

$$T_N = U_N \frac{\sqrt{\frac{N\sigma^2}{m} + \frac{N\tau^2}{n}}}{\sqrt{\frac{N\hat{\sigma}_N^2}{m} + \frac{N\hat{\sigma}_N^2}{n}}} \to_d \sqrt{\frac{(1-\lambda)\sigma^2 + \lambda\tau^2}{\lambda\sigma^2 + (1-\lambda)\tau^2}} N(0,1) = N(0,V)$$

with

$$V = \frac{(1-\lambda)\sigma^2 + \lambda\tau^2}{\lambda\sigma^2 + (1-\lambda)\tau^2}.$$

Notes: answers must mention the weak law of large numbers (1 mark) and Slutsky's lemma (1 mark). Other starting points besides U_N are possible.

(f) *[Unseen]* (2B marks)

For given sample sizes $m = 1000$ and $n = 2000$, we can reasonably assume that $\lambda_N = \lambda = 1/3$. From part (e), we have that

$$V = \frac{(2/3)\sigma^2 + (1/3)\tau^2}{(1/3)\sigma^2 + (2/3)\tau^2} = \frac{2\sigma^2 + \tau^2}{\sigma^2 + 2\tau^2}.$$

When $\sigma^2 \neq \tau^2$, there are two cases to consider:

Case 1: Smaller sample has smaller variance. If $\sigma^2 < \tau^2$, then there exists $\delta > 0$ such that $\tau^2 = \sigma^2 + \delta$. We then have

$$V = \frac{2\sigma^2 + \tau^2}{\sigma^2 + 2\tau^2} = \frac{3\sigma^2 + \delta}{3\sigma^2 + 2\delta} < 1.$$

In this case, the limiting distribution of T_N is more concentrated around zero than N(0,1). This implies that the t-test with nominal level α will have Type I-error rate less than α. The corresponding $1-\alpha$ confidence interval will have coverage probability greater than $1-\alpha$.

Case 2: Smaller sample has higher variance. If $\sigma^2 > \tau^2$, then there exists $\delta > 0$ such that $\sigma^2 = \tau^2 + \delta$. We then have

$$V = \frac{2\sigma^2 + \tau^2}{\sigma^2 + 2\tau^2} = \frac{3\tau^2 + 2\delta}{3\tau^2 + \delta} > 1.$$

Here, in contrast to the previous case, the limiting distribution of T_N is *less* concentrated around zero than N(0,1). This implies that the t-test with nominal level α will have Type I-error rate *greater than* α. The corresponding $1-\alpha$ confidence interval will have coverage probability *less than* $1-\alpha$.

One can summarise the two cases as follows: if the smaller group has smaller variance, then inference based on the two-sample t-test statistic is *conservative*. If the smaller group has higher variance, then inference is instead *anti-conservative*.

Alternate solution: The true sampling variance of $\bar{X} - \bar{Y}$ is $\text{Var}(\bar{X} - \bar{Y}) = \sigma^2/m + \tau^2/n$, whether or not $\sigma^2 = \tau^2$.
Squaring the denominator of the two-sample t-test statistic T_N, we can consider

$$\hat{\sigma}_N^2 \left(\frac{1}{m} + \frac{1}{n} \right)$$

as an estimator of $\text{Var}(\bar{X} - \bar{Y})$. From part 4(c), we note that

$$\begin{aligned} E[\hat{\sigma}_N^2] \left(\frac{1}{m} + \frac{1}{n} \right) &= \frac{(m-1)\sigma^2 + (n-1)\tau^2}{N-2} \left(\frac{1}{m} + \frac{1}{n} \right) \\ &\approx \{\lambda_N \sigma^2 + (1-\lambda_N)\tau^2\} \left(\frac{1}{m} + \frac{1}{n} \right) \\ &= \sigma^2/n + \tau^2/m \end{aligned}$$

Hence, if $\sigma^2 < \tau^2$, this expectation is greater than $\text{Var}(\bar{X} - \bar{Y})$. As above, this results in the distribution of T_N being more concentrated around zero than N(0,1), leading to conservative inference. In the case that $\sigma^2 < \tau^2$, this expectation is less than $\text{Var}(\bar{X} - \bar{Y})$. The result is that the distribution of T_N is now less concentrated around zero than N(0,1), such that inference based on the t-test will be anti-conservative.