In all the questions that follow, X is an n by p design matrix with full column rank and P is the orthogonal projection on to the column space of X. Also, let X_0 be the matrix formed from the first $p_0 < p$ columns of X and let P_0 be the orthogonal projection on to the column space of X_0. The vector $Y \in \mathbb{R}^n$ will be a vector of responses and we will define $\hat{\beta} := (X^T X)^{-1} X^T Y$, $\hat{\beta}_0 := (X_0^T X_0)^{-1} X_0^T Y$ and $\tilde{\sigma}^2 := \|(I - P)Y\|^2/(n - p)$.

1. Show that $X^T X$ is invertible and that

$$\operatorname*{argmin}_{b \in \mathbb{R}^p} \|Y - Xb\|^2 = (X^T X)^{-1} X^T Y.$$

2. Consider the linear model

$$Y = X\beta + \varepsilon, \qquad \mathrm{Var}(\varepsilon) = \sigma^2 I. \tag{1}$$

Let $\tilde{\beta}$ be an unbiased linear estimator of β with $X\tilde{\beta} = AY$.

(a) Show that $AP = P$ so we may write $A = P + A(I - P)$.

(b) By considering $\mathrm{Var}(AY) - \mathrm{Var}(PY)$ or otherwise, show that $\hat{\beta}$ is the best linear unbiased estimator for β in linear model above (i.e. show that $\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\hat{\beta})$ is positive semi-definite).

(c) Conclude that for any $x^* \in \mathbb{R}^p$,

$$\mathbb{E}\{\|x^{*T}(\hat{\beta} - \beta)\|^2\} \le \mathbb{E}\{\|x^{*T}(\tilde{\beta} - \beta)\|^2\}.$$

3. Let W be an $n \times n$ diagonal matrix with positive entries. Using the result from question 2 or otherwise, show that the best linear unbiased estimator for β in the model

$$Y = X\beta + \varepsilon, \qquad \mathrm{Var}(\varepsilon) = \sigma^2 W^{-1},$$

is

$$\hat{\beta}^{\mathrm{w}} := (X^T W X)^{-1} X^T W Y.$$

What are the variances of $\hat{\beta}^{\mathrm{w}}$ and $\hat{\beta} := (X^T X)^{-1} X^T Y$? Using your answer to question 1 or otherwise, further show that

$$\hat{\beta}^{\mathrm{w}} = \operatorname*{argmin}_{b \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} W_{ii}(Y_i - x_i^T b)^2 \right\},$$

where x_i^T is the i^{th} row of X.

4. Consider the model $Y = f + \varepsilon$ where $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I$ and $f \in \mathbb{R}^n$ is a non-random vector. Suppose we have have performed linear regression of Y on X so the fitted values are PY. Show that if $Y^* = f + \varepsilon^*$ where $\mathbb{E}(\varepsilon^*) = 0$, $\mathrm{Var}(\varepsilon^*) = \sigma^2 I$ and ε^* is independent of ε, then

$$\frac{1}{n}\mathbb{E}(\|PY - Y^*\|^2) = \frac{\sigma^2 p}{n} + \frac{1}{n}\|(I - P)f\|^2 + \sigma^2.$$

5. (a) Let V and W be subspaces of $\mathbb{R}^n$ with $V \leq W$. Let Π_V and Π_W denote orthogonal projections on to V and W respectively. Show that for all $v \in \mathbb{R}^n$, $\|v\|^2 \geq \|\Pi_W v\|^2 \geq \|\Pi_V v\|^2$.

   (b) Consider the linear model (1) but where only the first p_0 components of β are non-zero. Show that

$$\text{Var}(\hat{\beta}_{0,j}) \leq \text{Var}(\hat{\beta}_j) \qquad \text{for } j = 1, \dots, p_0.$$

   Here $\hat{\beta}_{0,j}$ denotes the j^{th} component of $\hat{\beta}_0$. *Hint: Use the alternative characterisation of $\hat{\beta}_j$ that if $X_j^{\perp}$ is the orthogonal projection of X_j (the j^{th} column of X) on to the orthogonal complement of the column space of X_{-j} (the matrix formed by removing the j^{th} column from X), then $\hat{\beta}_j = \frac{(X_j^{\perp})^T Y}{\|X_j^{\perp}\|^2}$.*

6. Show that the maximum likelihood estimator of σ^2 in the normal linear model ($Y = X\beta + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$) is $\|(I - P)Y\|^2/n$.

7. Let the cuboid C be defined $C := \prod_{j=1}^{p} C_j(\alpha/p)$, where

$$C_j(\alpha) = \left[\hat{\beta}_j - \sqrt{\tilde{\sigma}^2 (X^T X)^{-1}_{jj}}\, t_{n-p}(\alpha/2),\ \hat{\beta}_j + \sqrt{\tilde{\sigma}^2 (X^T X)^{-1}_{jj}}\, t_{n-p}(\alpha/2)\right].$$

   Assuming the normal linear model, show that $\mathbb{P}_{\beta,\sigma^2}(\beta \in C) \geq 1 - \alpha$.

8. Consider the Bayesian linear model $Y = X\beta + \varepsilon$ with $\varepsilon | \omega \sim N_n(0, \omega^{-1} I)$. Let the prior for β conditional on $\omega := \sigma^{-2}$ be $\beta | \omega \sim N_p(0, \omega^{-1}\lambda^{-1} I)$. We leave the prior for ω undefined. Show that the posterior distribution for β satisfies $\beta | Y, \omega \sim N_p(\hat{\beta}_\lambda, \omega^{-1}(X^T X + \lambda I)^{-1})$, where $\hat{\beta}_\lambda := (X^T X + \lambda I)^{-1} X^T Y$. Conclude that $\mathbb{E}(\beta | Y) = \hat{\beta}_\lambda$.

9. Consider the Bayesian linear model in question 8, but now with a prior for (β, ω) as $p(\beta, \omega) = \omega^{-1}$, as in lectures. Let $Y^* = x^{*T}\beta + \varepsilon^*$, where, conditional on ω, $\varepsilon^* \in \mathbb{R}$ is independent of ε and $\varepsilon^* | \omega \sim N(0, \omega^{-1})$. Find the distribution of $Y^* | \omega, Y$. Write down the posterior distribution of the precision $\omega | Y$ and hence find the posterior predictive distribution $Y^* | Y$.

10. This question is about understanding what can happen to the F-test when the expected value of the response is not necessarily linear in β. Consider the model $Y = f + \varepsilon$ where $\varepsilon \sim N_n(0, \sigma^2 I)$ and $f \in \mathbb{R}^n$ is a non-random vector. Define $\beta \in \mathbb{R}^p$ by $X\beta = Pf$, so $Y = X\beta + (I - P)f + \varepsilon$, and partition β as $\beta = (\beta_0^T, \beta_1^T)^T$ where $\beta_0 \in \mathbb{R}^{p_0}$ and $\beta_1 \in \mathbb{R}^{p - p_0}$. Suppose we try to test the hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$ by rejecting the null hypothesis when

$$F := \frac{\frac{1}{p - p_0}\|(P - P_0)Y\|^2}{\frac{1}{n - p}\|(I - P)Y\|^2}$$

    exceeds $F_{p - p_0, n - p}(\alpha)$. We will show that the size of this test (the probability of rejecting the null hypothesis when in fact it is true) is at most α.

    (a) Show that the numerator and denominator of F are independent (no matter which hypothesis is true).

    (b) What is the distribution of $\|(P - P_0)Y\|^2$ under the null hypothesis (i.e. when $Y = X_0\beta_0 + (I - P)f + \varepsilon$)?

(c) By considering the eigendecomposition of $I-P$, show that $\|(I-P)Y\|^2$ has the same distribution as

$$Z_1^2 + \cdots + Z_{n-p}^2,$$

where the Z_i are independent and $Z_i \sim N(\lambda_i, \sigma^2)$ for some λ_i such that

$$\sum_{i=1}^{n-p} \lambda_i^2 = \|(I-P)f\|^2.$$

(d) For any two real-valued random variables A and B, let us write $A \preceq B$ to mean $\mathbb{P}(A > x) \leq \mathbb{P}(B > x)$ for all $x \in (-\infty, \infty)$ (we say A is *stochastically less than* B). Now prove that if $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$ are all independent real-valued random variables and $A_1 \preceq B_1, \ldots, A_m \preceq B_m$, then $A_1 + \cdots + A_m \preceq B_1 + \cdots + B_m$. *Hint: Use induction on m and recall that for real-valued random variables U_1 and U_2 (defined on the same probability space) the tower property of conditional expectation gives us that $\mathbb{P}(U_1 + U_2 > x) = \mathbb{E}\{\mathbb{P}(U_1 > x - U_2 | U_2)\} := \mathbb{E}\{\mathbb{E}(\mathbb{1}_{\{U_1 > x - U_2\}} | U_2)\}$.*

(e) Let $Z \sim \sigma^2 \chi^2_{n-p}$. Show that

$$Z \preceq \|(I-P)Y\|^2.$$

Conclude that the size of the test mentioned at the beginning of this question is at most α.

11. Prove that in the linear model $Y = X\beta + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$, the F-test and t-test for testing the significance of a single predictor are equivalent. That is, prove that, taking $p_0 = p - 1$, so P_0 is the orthogonal projection on to the first $p-1$ columns of X,

$$\frac{\hat{\beta}_p^2}{(X^T X)^{-1}_{pp} \tilde{\sigma}^2} = \frac{\|(P - P_0)Y\|^2}{\frac{1}{n-p}\|(I-P)Y\|^2}.$$

Hint: Look at the hint to question 5 (b).