STATISTICAL MODELLING Part IIC / Michaelmas 2025
Example Sheet 1 (of 4)

In all the questions that follow, X is an n by p design/model 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 $\hat{\sigma}^2 := \|(I-P)Y\|^2/(n-p)$. By normal linear model, we mean the model $Y = X\beta + \varepsilon$, $\varepsilon \mid X \sim N_n(0, \sigma^2 I)$.

1. Show that the maximum likelihood estimator of σ^2 in the normal linear model is

$$\hat{\sigma}^2_{\text{MLE}} = \|(I-P)Y\|^2/n.$$

   Find the distribution of $\hat{\sigma}^2_{\text{MLE}}$ and conclude that $\hat{\sigma}^2_{\text{MLE}}$ is a biased estimator of σ^2 but $\hat{\sigma}^2$ is unbiased. Construct a confidence interval of σ^2 with level $1-\alpha$.

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

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

   Assuming the normal linear model, show that $\mathsf{P}(\beta \in C) \geq 1-\alpha$.

3. Show that $\|(P-P_0)Y\|^2 = \|(I-P_0)Y\|^2 - \|(I-P)Y\|^2 = \|PY\|^2 - \|P_0 Y\|^2$. Use this to show that the size-α generalised likelihood ratio test in the normal linear model for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ rejects H_0 when

$$\frac{\|(P-P_0)Y\|^2/(p-p_0)}{\|(I-P)Y\|^2/(n-p)} > F_{p-p_0,n-p}(\alpha),$$

   where $F_{p-p_0,n-p}(\alpha)$ is the upper-α quantile of $F_{p-p_0,n-p}$.

4. Suppose we observe data (X, Y) generated from a normal linear model. Let $x^* \in \mathbb{R}^p$ be a fixed vector and $\epsilon^* \sim N(0, \sigma^2)$ be independent of (X, Y). Denote $Y^* = (x^*)^T\beta + \epsilon^*$. Construct $(1-\alpha)$-confidence intervals for $(x^*)^T\beta$ and Y^*. Which interval is shorter in length? Can you give an intuitive explanation to your answer?

5. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. and $Y_1 \mid X_1 \sim \mathrm{N}(X_1^T\beta, \sigma^2)$. Denote $\Sigma = \mathsf{E}(X_1 X_1^T)$. Find the asymptotic distribution of the MLE of $\theta = (\beta^T, \sigma^2)^T$ by calculating the Fisher information matrix $I(\theta) = \mathsf{Var}_\theta(\nabla_\theta l(\theta))$ where $l(\theta)$ is the log-likelihood function. Comment on how that is related to the exact distribution of the MLE given X.

6. Consider a random variable $F \sim F_{d_1,d_2}$.

   (a) When $d_1 = 1$, show that the distribution of F is the same as that of T^2 where $T \sim t_{d_2}$.

   (b) What can you say about the distribution of F when $d_2 \to \infty$?

   (c) Check the conclusion in part (a) and your answer to part (b) using R.

7. Data are available on weights of two groups of three rats at the beginning of a fortnight and at its end. During the fortnight, one group was fed normally, and the other was given a growth inhibitor. The weights of the k^{th} rat in the j^{th} group before and after the fortnight are denoted by X_{jk} and Y_{jk}, respectively. It is assumed that $Y_{jk} = \alpha_j + \beta_j X_{jk} + \varepsilon_{jk}$ where ε_{jk}, $j = 1, 2$, $k = 1, 2, 3$ are independent normally distributed noise with mean 0 and unknown variance.

   (a) Let W be the vector of responses, so $W = (Y_{11}, Y_{12}, Y_{13}, Y_{21}, Y_{22}, Y_{23})^T$, and similarly let δ be the vector of random errors. Write down the model above in the form $W = A\theta + \delta$ with $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2)^T$; you should give the design matrix A and noise vector δ explicitly.

(b) The model is to be reparametrised in such a way that it can be specialised to (i) two parallel lines for the two groups, (ii) two lines with the same intercept, (iii) one common line for both groups, just by setting parameters to zero. Give one design matrix that can be made to correspond to (i), (ii) and (iii), just by dropping columns, specifying which columns are to be dropped for which cases.

(c) The data have been imported into R using the code below.

```
X <- c(184, 204, 211, 190, 195, 203)
Y <- c(213, 228, 226, 197, 190, 206)
data <- data.frame(before = X, after = Y,
                   group = gl(2, 3, labels = c("Normal", "Inhibitor")))
```

Find out how R constructs the design matrix by calling `model.matrix(formula, data)` with the following formulas:

- `formula <- after ~ 0 + group + group:before`
- `formula <- after ~ group * before`

Write down the R formulas for constructing the three specialisations in part (b) and check your answers using `plot(data$before, fitted(lm(formula, data)))`.

8. (Tripos 2022/II/13J) Consider the following R code:

```
> n <- 1000000
> sigma_z <- 1; sigma_x1 <- 0.5; sigma_x2 <- 1; sigma_y <- 2; beta <- 2
> Z <- sigma_z * rnorm(n)
> X1 <- Z + sigma_x1 * rnorm(n)
> X2 <- Z + sigma_x2 * rnorm(n)
> Y <- beta * Z + sigma_y * rnorm(n)
> lm(Y ~ Z)
Coefficients:
(Intercept)            Z
  -0.003089     1.999780
> lm(Y ~ X1)
Coefficients:
(Intercept)           X1
  -0.002904     1.600521
> lm(Y ~ X2)
Coefficients:
(Intercept)           X2
  -0.002672     0.997499
```

Describe the phenomenon you observe from the output above, then give a mathematical explanation to this phenomenon. Do you expect the slope coefficient in the second model to be generally smaller than that in the first model? Do you think modifying (for example, doubling) the value of `sigma_y` will substantially alter the slope coefficient in the second model? Justify your answer.

9. In the second practical session, you were asked to write a function called `mylm` in R with arguments `X` ($n \times p$ model matrix) and `Y` (n-vector) that outputs all the numbers reported by `summary.lm` without calling `lm` or using an LLM. The function should also accept a logical vector `S0` of length p and performs the analysis of variance test for the sub-model that only uses the regressors `X0 <- X[, S0]`. If you have not done so already, finish writing your function and test it using an example.

10. In this question, we explore the leverage of the i-th data point (X_i, Y_i), defined as P_{ii} (the i-th diagonal element of P).

(a) In the normal linear model, show that $\mathsf{Var}(Y_i - X_i^T\hat{\beta} \mid X) = \sigma^2(1 - P_{ii})$. [*Hint: Write the residual $Y_i - X_i^T\hat{\beta}$ as a linear transformation of Y.*]

(b) Suppose the design matrix X consists of just a single variable and a column of 1's representing an intercept term (as the first column). Show that the leverage, P_{ii}, of the i^{th} observation satisfies

$$P_{ii} = \frac{1}{n} + \frac{(X_{i2} - \bar{X}_2)^2}{\sum_{k=1}^{n}(X_{k2} - \bar{X}_2)^2},$$

where $\bar{X}_2 := \frac{1}{n}\sum_{k=1}^{n} X_{k2}$. Describe what kind of observations may have a large leverage. [*Hint: Why can we assume that the i^{th} component of the second column is $X_{i2} - \bar{X}_2$ rather than X_{i2}?*]

11. Consider the normal linear model with fixed X. Suppose only the first p_0 components of β are non-zero. Show that

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

Here $\hat{\beta}_j$ and $\hat{\beta}_{0,j}$ denote the j^{th} component of $\hat{\beta}$ and $\hat{\beta}_0$, respectively. [*Hint: Use the partial regression characterisation of $\hat{\beta}_j$.*]

12. This question is about understanding what can happen to the F-test when the linearity assumption does not hold. Consider the model $Y \sim \mathrm{N}(\mu, \sigma^2 I)$ where $\mu \in \mathbb{R}^n$ is non-random. Define $\beta \in \mathbb{R}^p$ by $X\beta = P\mu$, so $Y = X\beta + (I - P)\mu + \varepsilon$, and partition $\beta = (\beta_0^T, \beta_1^T)^T$ as before.

(a) Show that the numerator and denominator of the F-statistic in Question 3 are independent regardless of the value of β_1.

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

(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

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

(d) For any two real-valued random variables A and B, let us write $A \preceq B$ (and say A is *stochastically less than* B) if

$$\mathsf{P}(A > x) \leq \mathsf{P}(B > x), \quad \text{for all } x \in \mathbb{R}.$$

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 the law of total expectation.*]

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

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

Conclude that the size of the F-test in Question 3 is at most α.