UNIVERSITY OF CAMBRIDGE

M. Phil. in STATISTICAL SCIENCE

Tuesday, 1 June, 2010 1:30 pm to 3:30 pm

STATISTICAL THEORY

*Attempt no more than **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Consider the linear model

$$Y = X\beta + \epsilon,$$

where $Y = (Y_1, \ldots, Y_n)^T$ is a vector of responses, where X is a known $n \times p$ design matrix of full rank $p < n$, where $\beta = (\beta_1, \ldots, \beta_p)^T$ is an unknown vector of regression coefficients and where $\epsilon \sim N_n(0, \sigma^2 I)$. Write down expressions for the maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}^2$ of β and σ^2 respectively. State their joint distribution.

Derive an exact $(1-\alpha)$-level confidence set for β. [Standard facts about distributions may be assumed without proof.]

Explain how this confidence set can be used to give an exact size α test of $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$.

In the linear model above with $n = 9$ and $p = 2$, we have

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -4 & -3 & -2 & -1 & 0 & 1 & 2 & 3 & 4 \end{pmatrix}$$

and observe $Y = (-\frac{21}{25}, -\frac{72}{25}, -1, -1,\ 0,\ 1,\ 1, \frac{72}{25}, \frac{21}{25})^T$.

Use your hypothesis test above to test $H_0 : \beta = (0,0)^T$ against $H_1 : \beta \neq (0,0)^T$ at the 5% level.

[*Hint: You may find some of the following information helpful:*

1. *If* $Z \sim \chi^2_2$, *then* $\mathbb{P}(Z \leqslant 5.99) = 0.95$
2. *If* $Z \sim \chi^2_7$, *then* $\mathbb{P}(Z \leqslant 14.1) = 0.95$
3. *If* $Z \sim F_{2,7}$, *then* $\mathbb{P}(Z \leqslant 4.74) = 0.95$
4. *If* $Z \sim F_{2,9}$, *then* $\mathbb{P}(Z \leqslant 4.26) = 0.95$.

]

2 State a result on the consistency of maximum likelihood estimators in general parametric models, specifying carefully any regularity conditions you require.

In each of the two models below, find the maximum likelihood estimator $\hat{\theta}_n$ of θ. Show that in one case $\hat{\theta}_n$ is consistent, and that in the other it is not.

1. $Y_1, \ldots, Y_n$ are independent, each having density
$$f(y;\theta) = e^{-(y-\theta)}, \quad y \geqslant \theta$$
2. $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim \mathrm{Poi}(\theta 2^{-i})$ for some $\theta > 0$.

In the case above where $\hat{\theta}_n$ is consistent, derive an asymptotic distributional result for an appropriately normalised version of $\hat{\theta}_n$.

[In this question, standard distributional results may be assumed without proof.]

3 In the context of high-dimensional linear models, motivate the use of estimators that shrink the least squares estimator towards the origin.

Consider a linear model with an $n \times p$ design matrix of full rank $p < n$. By first describing appropriate standardisations for the response vector and design matrix, write down the penalised least squares problem solved by the ridge regression estimator $\hat{\beta}^R_\lambda$, where $\lambda > 0$. Write down also a closed form expression for $\hat{\beta}^R_\lambda$.

Prove that for sufficiently small $\lambda > 0$, the mean squared error of $\hat{\beta}^R_\lambda$ is smaller than that of the least squares estimator.

4 In the context of multiple testing, define what is meant by the False Discovery Proportion (FDP), and the False Discovery Rate (FDR). Describe the Benjamini–Hochberg (BH) procedure which aims to control the FDR at level α.

Suppose $H_1, \ldots, H_m$ are null hypotheses to be tested, and that the first m_0 of these are the true null hypotheses. Suppose further that the p-values $P_1, \ldots, P_m$ are independent, with $P_1, \ldots, P_{m_0} \sim U(0,1)$. Let R denote the number of rejected hypotheses. Find an alternative way of writing the event

$$\left\{P_1 \leqslant \tfrac{\alpha r}{m}, R = r\right\}$$

in terms of P_1 and the number of rejections in a modified BH procedure applied to $P_2, \ldots, P_m$. Similarly, find an alternative way of writing the event

$$\left\{P_1 \leqslant \tfrac{\alpha r}{m}, P_2 \leqslant \tfrac{\alpha r}{m}, R = r\right\}$$

in terms of P_1, P_2 and the number of rejections in a second modified BH procedure applied to $P_3, \ldots, P_m$. Hence show that

$$\mathbb{E}(\mathrm{FDP}^2) = \frac{\alpha m_0}{m}\mathbb{E}(A) + \frac{\alpha^2 m_0(m_0 - 1)}{m^2},$$

where A is a function of the number of rejections in the first modified BH procedure above which you should specify.

END OF PAPER