b8 STATISTICS **HT 2002**

Problem Sheet 8

1. Percentages of total calories obtained from complex carbohydrates were recorded for 20 male insulin-dependent diabetics who had been on a high carbohydrate diet for six months. Age (x_1), body weight (x_2) and percentage protein in diet (x_3) were tried as explanatory variables and a regression model was fitted to the data. The results were as follows:

	Coefficient	Standard error
Intercept	36.960	13.071
x_1	-0.114	0.109
x_2	-0.228	0.083
x_3	1.958	0.635

Correlations

	Intercept	x_1	x_2
x_1	-0.297		
x_2	-0.606	0.000	
x_3	-0.581	-0.216	-0.132

The residual sum of squares is 567.663 and the total sum of squares is 1092.800.

On the assumption that the normal linear model is appropriate:

(i) Calculate F-tests for (a) the whole regression, (b) the effect of age. Comment on your result for (b).

(ii) Compute 95% confidence intervals for the coefficients of x_2 (body weight), x_3 (protein) and the difference between these coefficients.

(iii) Explain how to compute a 95% confidence interval for the mean carbohydrate response when the values of age, weight and protein are 43, 120 and 16 respectively.

2. The table below gives the number of births per thousand in the states of the USA in three regions for the year 1985.

	North East	South Atlantic	South Central
	14.5	15.5	14.2
	15.5	15.5	14.0
	15.0	15.8	14.9
	14.1	15.1	16.6
	13.5	12.5	14.9
	13.9	14.3	18.2
	14.6	15.6	16.1
	14.0	16.1	18.8
	13.5	14.4	
Totals	128.6	134.8	127.7

A linear model, which allows for different birth rates in the three regions, is fitted. The residual sum of squares is 36.143 and the total sum of squares is 48.063. Test the null hypothesis that the birth rates in the regions are the same. What can you say about the differences between birth rates?

3. Consider the model

$$\mathbf{Y} = X\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where X is an $n \times p$ matrix of rank p $(< n)$, and where the components of $\boldsymbol{\varepsilon}$ are independent with mean 0 and variance σ^2. However we do *not* assume here that the components of $\boldsymbol{\varepsilon}$ are normally distributed. Let $\widehat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \mathbf{Y}$, the usual least squares estimator of $\boldsymbol{\theta}$.

Fix $\mathbf{k} \in \mathbb{R}^p$. Show that $\mathbf{k}^T \widehat{\boldsymbol{\theta}}$ is an unbiased estimator of $\mathbf{k}^T \boldsymbol{\theta}$ (i.e. show $E(\mathbf{k}^T \widehat{\boldsymbol{\theta}}) = \mathbf{k}^T \boldsymbol{\theta}$ for all $\boldsymbol{\theta}$) and find $\mathrm{var}(\mathbf{k}^T \widehat{\boldsymbol{\theta}})$.

Suppose $\boldsymbol{\ell} \in \mathbb{R}^n$ is a fixed vector such that $\boldsymbol{\ell}^T \mathbf{Y}$ is an unbiased estimator of $\mathbf{k}^T \boldsymbol{\theta}$. Use that fact that $\boldsymbol{\ell}^T \mathbf{Y}$ is unbiased to show that $X^T \boldsymbol{\ell} = \mathbf{k}$. Show also that

$$\mathrm{var}(\boldsymbol{\ell}^T \mathbf{Y}) - \mathrm{var}(\mathbf{k}^T \widehat{\boldsymbol{\theta}}) = \sigma^2 \boldsymbol{\ell}^T (I - H)^T (I - H) \boldsymbol{\ell} \geqslant 0$$

where $H = X(X^T X)^{-1} X^T$, and hence that $\mathbf{k}^T \widehat{\boldsymbol{\theta}}$ is a linear unbiased estimator of $\mathbf{k}^T \boldsymbol{\theta}$ with minimum variance.

4. For the one-way analysis of variance model

$$Y_{ij} = \mu + t_i + \varepsilon_{ij}, \qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, p,$$

where the ε_{ij} are independent $N(0, \sigma^2)$ random variables, prove that the function of the treatment effects $\sum_{i=1}^p c_i t_i$ is estimable if and only if $\sum_{i=1}^p c_i = 0$.

An experiment on sugar beets compared times and methods of applying mixed artificial fertilisers. The mean yields of sugar (cwt per acre) were as follows:

$$\begin{aligned}
\text{no artificials,} \quad & \overline{Y}_{1.} &=& \; 38.7, \\
\text{artificials applied in January by ploughing,} \quad & \overline{Y}_{2.} &=& \; 48.7, \\
\text{artificials applied in January by broadcasting,} \quad & \overline{Y}_{3.} &=& \; 48.8, \\
\text{artificials applied in April by broadcasting,} \quad & \overline{Y}_{4.} &=& \; 45.0.
\end{aligned}$$

The average in each case is taken over three replications. The standard error of these averages is 1.22 .

Calculate 95% confidence intervals for the following contrasts:

$$\begin{aligned}
\text{average effect of the artificials,} \quad & \tfrac{1}{3}(t_2 + t_3 + t_4) - t_1, \\
\text{January versus April application,} \quad & \tfrac{1}{2}(t_2 + t_3) - t_4.
\end{aligned}$$