MATH273501

This question paper consists of 8 printed pages, each of which is identified by the reference **MATH2735**.

Statistical tables are provided at the end of this examination paper. Only approved basic scientific calculators may be used.

Examination for the Module MATH2735
(January 2008)

Statistical Modelling

Time allowed: **2 hours**

Attempt not more than FOUR questions.
All questions carry equal marks.

Throughout this examination paper, replacing a subscript by a $\bullet$ denotes that the subscript has been summed over. A bar over a quantity indicates that averaging has taken place. For example, given data y_{ij} for $i = 1, \ldots, t$ and $j = 1, \ldots, n$, use the notation

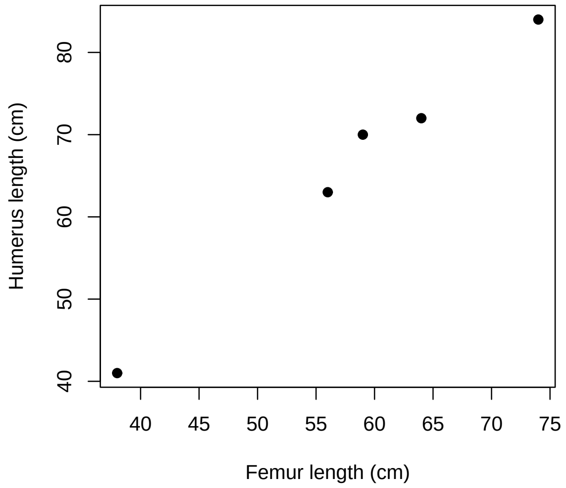
$$y_{i\bullet} = \sum_{j=1}^{n} y_{ij}, \qquad y_{\bullet\bullet} = \sum_{i=1}^{t} \sum_{j=1}^{n} y_{ij},$$

$$\overline{y}_{i\bullet} = \frac{1}{n} \sum_{j=1}^{n} y_{ij}, \quad \text{and} \quad \overline{y}_{\bullet\bullet} = \frac{1}{nt} \sum_{i=1}^{t} \sum_{j=1}^{n} y_{ij}.$$

1. Five complete fossils are available of the extinct species *Archeopteryx*. For each, the lengths of the femur (leg bone) and humerus (wing bone) are shown in the table and figure below.

Archeopteryx bone lengths

Femur length (cm)	38	56	59	64	74
Humerus length (cm)	41	63	70	72	84



(a) Comment on the relationship between femur length and humerus length. Is linear regression a suitable tool to analyse these data?

(b) Given that $S_{xx} = 696.8$ and $S_{xy} = 834$, estimate the regression parameters for these data. Give the regression equation in an uncentred form.

(c) Analysis in R gave the following results.

```
> bone.lm = lm(humerus ~ femur)
> anova(bone.lm)
Analysis of Variance Table
Response: humerus
          Df Sum Sq Mean Sq F value    Pr(>F)
femur      1 998.21  998.21   254.1 0.0005368
Residuals  3  11.79    3.93
```

What null hypothesis has been tested in this ANOVA table? What is the alternative hypothesis? What conclusions can be drawn from the results?

Construct a 95% confidence interval for the regression slope parameter β.

(d) An incomplete *Archeopteryx* skeleton is also available. For this skeleton, the femur length is known to be 50cm.

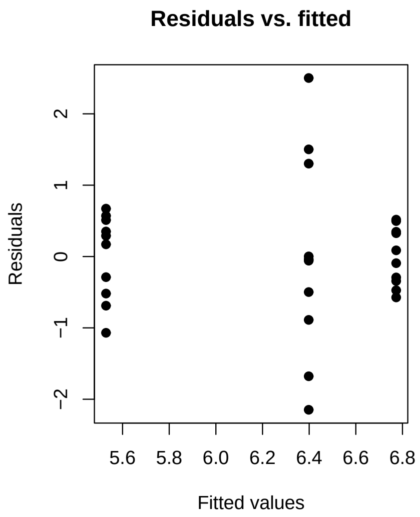
Predict the humerus length for this specimen given the femur length of 50cm. Also construct confidence and prediction intervals for humerus length given a femur length of 50cm. Explain the difference between the two types of interval.

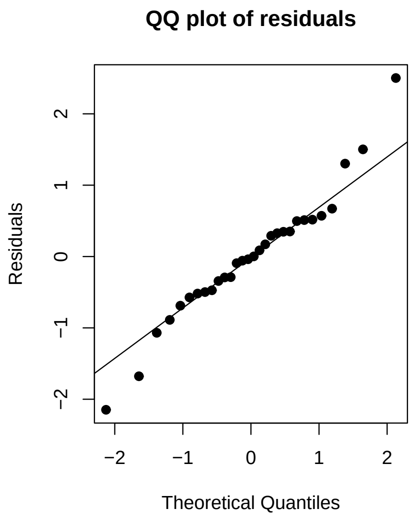
2. In a study of the percentage rubber content of three different varieties of rubber plant, a random sample of 30 plants was selected from a field of one-year old rubber plants. In this sample, there were ten plants each of the three varieties "elongated", "oval", and "transverse". The percentage rubber content of each plant was measured as the response variable and the group totals are shown below.

	Elongated	Oval	Transverse
Totals $y_{i\bullet}$	67.73	63.98	55.29

(a) What model is appropriate to analyse these data? Justify your choice.
Given that $\sum_{ij} y_{ij}^2 = 1197.218$, complete an ANOVA table for these data. What conclusions can you draw?

(b) Explain briefly how plots of residuals against fitted values and normal QQ plots of residuals can be used to assess model adequacy.
Interpret the following plots of residuals for the rubber data. Some R output is given below. Say what hypotheses have been tested and what conclusions can be drawn. What implications do these plots and R code have for your conclusions in part (a)?






```
> bartlett.test(res ~ species)

Bartlett test of homogeneity of variances

data:  res by species
Bartlett's K-squared = 14.1386, df = 2, p-value = 0.0008508

> fligner.test(res ~ species)

Fligner-Killeen test of homogeneity of variances

data:  res by species
Fligner-Killeen:med chi-squared = 4.0618, df = 2, p-value = 0.1312
```

3. An engineer wanted to investigate the forces developed by a circular saw used in cutting a metal plate. He thought that the rate of feed of the metal plate into the saw and the saw speed were the major factors determining the force generated, so he conducted the following experiment. Using four selected feed rates and two representative saw speeds, he cut sixteen test pieces giving the data below.

Saw feed rate data

	Feed rate (cms/second)			
Saw speed	0.2	0.4	0.6	0.8
Low	2.77	2.49	2.60	2.77
(200 r.p.m)	2.69	2.45	2.72	2.88
High	2.86	2.84	2.84	2.90
(500 r.p.m)	2.83	2.79	2.85	2.88

(a) Write down an appropriate ANOVA model to analyse these data. What difference would it make if the engineer had only used one test piece for each saw speed / feed rate combination?

(b) State the ANOVA identity for this model in terms of $Y_{ijk}, \overline{Y}_{i\bullet\bullet}, \overline{Y}_{\bullet j\bullet}, \overline{Y}_{ij\bullet}$, and $\overline{Y}_{\bullet\bullet\bullet}$. Name the individual terms in the identity and say what each one represents.

(c) Explain how the hypothesis H_0: $\alpha_i = 0$ for all i might be tested against H_1: at least one $\alpha_i \neq 0$ by writing down the appropriate test statistic and the null distribution that it would be compared to. How would the test statistic differ from the null distribution when H_0 is false? When would you reject H_0?

(d) Analysis in R gave the following ANOVA table

```
             Df   Sum Sq  Mean Sq F value   Pr(>F)
speed         a        b 0.126025       c  8.836e-05
rate          3 0.096500 0.032167       d  0.001738
speed:rate    3 0.044675 0.014892  6.2049  0.017500
Residuals     8 0.019200 0.002400
```

Given that the total sum of squares is $SS_{TOT} = 0.2864$, find the values `a, b, c,` and `d` missing from the ANOVA table, and state what conclusions can be drawn from the table.

4. Consider the one-way fixed effects ANOVA model

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for $i = 1, \ldots, t;\ j = 1, \ldots, n$ with the ε_{ij} being independently $N(0, \sigma^2)$ distributed. We apply the constraint $\sum_i n_i \alpha_i = 0$.

(a) Use the method of maximum likelihood to find estimates of the parameters μ and α_i.

(b) By comparing the log-likelihood function to the sum of squared errors $S = \sum_{ij} \varepsilon_{ij}^2$, explain why the maximum likelihood estimates will be the same as the least squares estimates in this case.

(c) A factory has three machines producing components. The factory manager selects four components at random from each machine and records the surface finish of each component as shown in the table below.

Surface finish of machined components

Machine		
A	B	C
88	92	79
75	99	62
94	85	53
84	79	56

Analysis of these data in R gave the following results

```
> finish.lm = lm(finish ~ machine)
> anova(finish.lm)
Analysis of Variance Table

Response: finish
          Df  Sum Sq Mean Sq F value   Pr(>F)
machine    2 1625.17  812.58  8.9132 0.007338
Residuals  9  820.50   91.17
```

What conclusions can be drawn from this ANOVA table?

Find the parameter estimates $\widehat{\mu}$ and $\widehat{\alpha}_i$ for these data.

5. (a) Under what circumstances would you use (i) orthogonal contrasts or (ii) multiple comparisons for testing sub-hypotheses in the analysis of variance assuming a one-way layout model?

(b) Briefly describe the method of orthogonal contrasts, giving the hypotheses tested, a formula for calculating the test statistics, and the distribution of the test statistics when the null hypotheses are true.

(c) A firm employs four people to make electronic circuit boards. A time and motion study yields the following data on the times taken by each employee to complete five circuit boards.

Circuit board production data

	Employee			
	A	B	C	D
	18	78	45	93
Time	21	62	23	47
(seconds)	22	72	22	81
	32	57	33	77
	12	84	48	95
Means	21.0	70.6	34.2	78.6

Analysing the data in R gave the result below.

```
> ecb.lm = lm(times ˜ employee)
> anova(ecb.lm)
            Df  Sum Sq Mean Sq F value    Pr(>F)
employee     3 11640.6  3880.2  22.387 5.707e-06
Residuals   16  2773.2   173.3
```

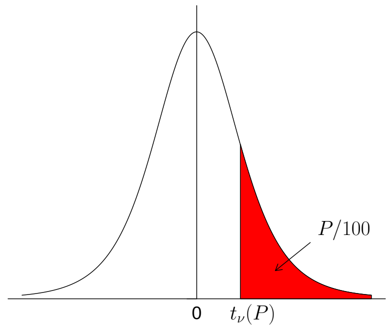
What model has been assumed here and what conclusions can you draw?

(d) The manager points out that employees B and D are fairly new recruits while A and C are experienced workers. He also notes that A and B are women, C and D are men. Use orthogonal contrasts to investigate whether experience or gender are significant factors affecting job time.

What third orthogonal contrast could additionally be examined? (Do not test this contrast.)

Percentage Points of the t-Distribution

This table gives the percentage points $t_\nu(P)$ for various values of P and degrees of freedom ν, as indicated by the figure to the right.



The lower percentage points are given by symmetry as $-t_u(P)$, and the probability that $|t| \geq t_u(P)$ is $2P/100$.

Percentage points P

$\boldsymbol{\nu}$	**10**	**5**	**2.5**	**1**	**0.5**	**0.1**	**0.05**
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
$\boldsymbol{\infty}$	1.282	1.645	1.960	2.326	2.576	3.090	3.291

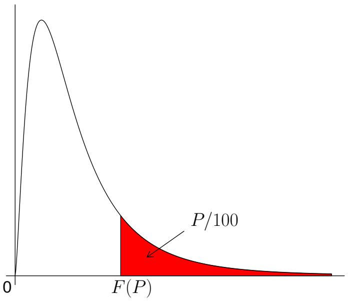
CONTINUED...

5 Percent Points of the F-Distribution

This table gives the percentage points $F_{\nu_1,\nu_2}(P)$ for $P = 0.05$ and degrees of freedom ν_1, ν_2, as indicated by the figure to the right.

The lower percentage points, that is the values $F'_{\nu_1,\nu_2}(P)$ such that the probability that $F \leq F'_{\nu_1,\nu_2}(P)$ is equal to $P/100$, may be found using the formula

$$F'_{\nu_1,\nu_2}(P) = 1/F_{\nu_1,\nu_2}(P)$$



	ν_1								
ν_2	**1**	**2**	**3**	**4**	**5**	**6**	**12**	**24**	∞
2	18.513	19.000	19.164	19.247	19.296	19.330	19.413	19.454	19.496
3	10.128	9.552	9.277	9.117	9.013	8.941	8.745	8.639	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	5.912	5.774	5.628
5	6.608	5.786	5.409	5.192	5.050	4.950	4.678	4.527	4.365
6	5.987	5.143	4.757	4.534	4.387	4.284	4.000	3.841	3.669
7	5.591	4.737	4.347	4.120	3.972	3.866	3.575	3.410	3.230
8	5.318	4.459	4.066	3.838	3.687	3.581	3.284	3.115	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.073	2.900	2.707
10	4.965	4.103	3.708	3.478	3.326	3.217	2.913	2.737	2.538
11	4.844	3.982	3.587	3.357	3.204	3.095	2.788	2.609	2.404
12	4.747	3.885	3.490	3.259	3.106	2.996	2.687	2.505	2.296
13	4.667	3.806	3.411	3.179	3.025	2.915	2.604	2.420	2.206
14	4.600	3.739	3.344	3.112	2.958	2.848	2.534	2.349	2.131
15	4.543	3.682	3.287	3.056	2.901	2.790	2.475	2.288	2.066
16	4.494	3.634	3.239	3.007	2.852	2.741	2.425	2.235	2.010
17	4.451	3.592	3.197	2.965	2.810	2.699	2.381	2.190	1.960
18	4.414	3.555	3.160	2.928	2.773	2.661	2.342	2.150	1.917
19	4.381	3.522	3.127	2.895	2.740	2.628	2.308	2.114	1.878
20	4.351	3.493	3.098	2.866	2.711	2.599	2.278	2.082	1.843
25	4.242	3.385	2.991	2.759	2.603	2.490	2.165	1.964	1.711
30	4.171	3.316	2.922	2.690	2.534	2.421	2.092	1.887	1.622
40	4.085	3.232	2.839	2.606	2.449	2.336	2.003	1.793	1.509
50	4.034	3.183	2.790	2.557	2.400	2.286	1.952	1.737	1.438
100	3.936	3.087	2.696	2.463	2.305	2.191	1.850	1.627	1.283
∞	3.841	2.996	2.605	2.372	2.214	2.099	1.752	1.517	1.002

END