The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2013–2014**

MAS273 Statistical Modelling **2 hours**

Attempt ***ALL FOUR*** *questions. The allocation of marks is shown in brackets. Total marks 85.*

1 This question concerns data on the energy content of municipal waste. The variables are *Energy* (E) - usable energy recovered from waste (kcal/kg), *Plastic* (PL) - % plastic composition by weight, *Paper* (P) - % paper composition by weight, *Garbage* (G) - % garbage composition by weight, *Water* (W) - % moisture by weight.

n heaps of municipal waste were obtained and the regression model to predict energy $E = \beta_0 + \beta_{PL}PL + \beta_P P + \beta_G G + \beta_W W + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ was fitted to the data. The following output was obtained:

```
Coefficients:

             Estimate  Std.  Error   t value  Pr(>|t|)
(Intercept)  2245.09     177.89      12.62    2.4 e-12
 Plastics     28.92       2.82       10.24     2.0e-10
  Paper        7.64       2.31        3.30      0.0029
 Garbage       4.30       1.92        2.24      0.0340
  Water      -37.36       1.83      -20.37    < 2e-16

Residual standard error:  31.5 on 25 degrees of freedom
Multiple R-Squared:  0.964, Adjusted R-squared:  0.958
F-statistic:  168 on 4 and 25 DF,  p-value:  < 2e-16
```

(i) What is the sample size n of the data set? ***(1 mark)***

(ii) What is the value of $\hat{\sigma}$ where $\hat{\sigma}^2$ is the unbiased estimator of σ^2? ***(1 mark)***

(iii) What is the value of the residual sum of squares for the model? ***(2 marks)***

(iv) Suppose heap 1 has 10% more plastic than heap 2 (the rest are the same for both heaps). What is the difference in the usable energy recovered from heap 1 and heap 2? ***(3 marks)***

(v) Compute the 95% confidence interval for β_P. (You may use the fact that $t_{25,0.975} = 2$). ***(3 marks)***

(vi) Test whether *Garbage* can be removed as a predictor. Describe your test clearly and report the findings of your test. ***(4 marks)***

(vii) Is there evidence against $H_0 : \beta_P = \beta_G = \beta_{PL} = \beta_W = 0$? Describe your test clearly and report the findings of your test. ***(4 marks)***

(viii) The model
`Energy ~ Plastics + Paper + Water`
was fitted to the data. The **F** test was computed comparing this model to the original model above. What is the value of the F-statistic for this test? ***(2 marks)***

1 (continued)

(ix) For the model of part (viii), the residual standard error is 33.8 which is larger than in the original model. Does removing a predictor from a model always increase the residual standard error? Explain. ***(3 marks)***

(x) Suppose the percentage of other material (not *Plastic, Paper, Garbage* or *Water*) was computed and added to this model as an additional predictor. Would the R^2 be larger, smaller or the same as in the original model? Explain.
(Note that as the predictors are percentages, the sum of all the four predictors and the percentage of other material add up to 100.) ***(4 marks)***

2 (i) Consider data from the simple linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, \cdots, n$ where x_i's are fixed constants, β_0, β_1 are the unknown coefficients and ϵ_i's are unobserved i.i.d. random variables from $N(0, \sigma^2)$. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the least squares estimators of β_0 and β_1 respectively. We know that

$$\begin{pmatrix} \hat{\beta}_0 \\ \\ \hat{\beta}_1 \end{pmatrix} = N\left(\begin{pmatrix} \beta_0 \\ \\ \beta_1 \end{pmatrix}, \ \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} & -\frac{\bar{x}}{s_{xx}} \\ \\ -\frac{\bar{x}}{s_{xx}} & \frac{1}{s_{xx}} \end{pmatrix} \right)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

(a) Suppose we are given a *new* value of x, say x_h. We want to estimate the *mean* of the dependent variable y_h given x_h. Let's call the mean M.

(α) Find an estimator $\hat{M}$ for M and show that it is unbiased for M. ***(4 marks)***

(β) Find the distribution of $\hat{M}$. ***(4 marks)***

(b) Give a condition on the x_i's under which $\hat{\beta}_0$ is *independent* of $\hat{\beta}_1$. Give reasons. ***(2 marks)***

(ii) A random sound signal has amplitude A at time t given by

$$A = 3\sin(\omega t + \epsilon), \tag{1}$$

where ω is the unknown frequency and ϵ is the random error distributed as $N(0, \sigma^2)$. The signal is observed at times t_1, t_2, t_3 and t_4 to give amplitudes A_1, A_2, A_3 and A_4 respectively. By taking a suitable transformation of (1), find an estimate for ω. ***(6 marks)***

(iii) Suppose we have a data set $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Consider two different linear models $y_i = \alpha + \beta x_i^2 + \epsilon_i$ and $y_i = \alpha + \beta x_i^2 + \gamma e^{x_i} + \epsilon_i$. Compare the residual sum of squares for the two models. Explain. ***(3 marks)***

3 (i) The one-way ANOVA model was used to compare the effectiveness of 3 insecticides on the growth of the plant *Solanum tuberosum*. 9 specimens of the plant were divided randomly into 3 groups, corresponding to the 3 different treatments. The growth of the plant over six months is reported below.

`Trt.1` : 3, 6, 6, `Trt.2` : 5, 6, 7, `Trt.3` : 8, 7, 9.

Denote by μ_1, μ_2, μ_3 the mean growths of the plants from `Trt.1, Trt.2` and `Trt.3` respectively. Is there any evidence against the null hypothesis $H_0 : \mu_1 = \mu_2$? Set up your model and describe the F test to test this hypothesis. Clearly explain all your work and report the P-value in the form $P(F_{?,?} > ?)$. (*You need to fill in the ? marks*). ***(15 marks)***

(ii) A two-way ANOVA *balanced* design was used in a study to determine which drug was most effective in reducing the size of a tumor. The 3 drugs used in the treatments were *Cisplatin, Vinblastine* and *5-fluorouracil*. The number of participants in the study included 24 *males* and 24 *females*. Two models were considered (i: treatment, j: sex of patient)

$$M_1 : y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \epsilon_{i,j,k}, \quad (2)$$

$$M_2 : y_{i,j,k} = \mu + \alpha_i + \beta_j + \epsilon_{i,j,k}, \quad (3)$$

with appropriate restrictions on the parameters. The following test was conducted in `R` to see which model is better.

```
> lm1<-lm(tumor∼trt*sex)
> lm2<-lm(tumor∼ trt+sex)
> anova(lm2, lm1)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	?	4428.1				
2	?	3065.6	?	?	?	0.0004429

(a) Fill in the ?'s in the table. ***(9 marks)***

(b) Which of the models M_1 or M_2 is preferable? ***(2 marks)***

4 An ANCOVA model was considered to describe the relationship between the *cholesterol level* (C) and *age of persons* (A) in the 2 *states* of Iowa ($i = 1$) and Nebraska ($i = 2$). The data set consisted of 11 measurements from Iowa and 19 from Nebraska. The models considered are given below (β_0, β_1, τ_i, $\beta_{1,i}$ are parameters with $\tau_1 = 0$).

$$
\begin{aligned}
M_1: &\quad C_{i,j} = \beta_0 + \tau_i + \beta_{1,i} A_{i,j} + \epsilon_{i,j} \\
M_2: &\quad C_{i,j} = \beta_0 + \epsilon_{i,j} \\
M_3: &\quad C_{i,j} = \beta_0 + \tau_i + \beta_1 A_{i,j} + \epsilon_{i,j} \\
M_4: &\quad C_{i,j} = \beta_0 + \tau_i + \epsilon_{i,j} \\
M_5: &\quad C_{i,j} = \beta_0 + \beta_1 A_{i,j} + \epsilon_{i,j}
\end{aligned}
$$

(i) Draw a diagram to show how the models are nested. ***(5 marks)***

(ii) Using the following output from `R`, find an appropriate model for the data set by reducing as many parameters as possible (**use a size/significance level of 0.05 for your tests**). Explain all your steps. ***(8 marks)***

```
> lm1<-lm(cholesterol~state*age)
> lm2<-lm(cholesterol~1)
> lm3<-lm(cholesterol~state+age)
> lm4<-lm(cholesterol~state)
> lm5<-lm(cholesterol~age)
```

```
> anova(lm2,lm1)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	29	103537				
2	26	48395	3	55142	9.8749	0.00016

```
> anova(lm3,lm1)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	27	49104				
2	26	48395	1	709.05	0.3809	0.5425

```
> anova(lm4,lm1)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	102924				
2	26	48395	2	54529	14.648	5.491e − 05

4 (continued)

```
> anova(lm5,lm1)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	54560				
2	26	48395	2	6165.5	1.6562	0.2104

```
> anova(lm2,lm3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	29	103537				
2	27	49104	2	54433	14.965	4.23e − 05

```
> anova(lm2,lm4)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	29	103537				
2	28	102924	1	612.7	0.1667	0.6862

```
> anova(lm2,lm5)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	29	103537				
2	28	54560	1	48976	25.134	2.673e − 05

```
> anova(lm4,lm3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	102924				
2	27	49104	2	53820	29.593	9.361e − 06

```
> anova(lm5,lm3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	54560				
2	27	49104	1	5456.4	3.0003	0.09466

End of Question Paper