The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2011–2012**

MAS273 Statistical Modelling **2 hours**

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

1 In a survey conducted by the Wall Street Journal, adults who regularly used various products were asked to name a TV commercial they had seen for a specified product category in the past week. Each naming of a commercial is termed a "retained impression". For each product the TV advertising budget (in $ millions) for the same year was recorded. In an R session, the advertising budgets are stored in a vector `budget`, and the corresponding estimated numbers of retained impressions are stored in a vector `imp`. (Data obtained from the Data and Story Library). Below is some edited output from an R session.

```
> lm1<-lm(imp~budget)
> summary(lm1)

Coefficients:
            Estimate Std. Error
(Intercept) 22.16269    7.08948
budget       0.36317    0.09712

Residual standard error: 23.5 on 19 degrees of freedom
Multiple R-squared: 0.424

> qt(0.975,19)
[1] 2.093024
> deviance(lm1)
[1] 10494.11
```

(i) Defining your notation carefully, write down the model that has been fitted to the data and assigned to the variable `lm1`. State the sample size, justifying your answer. ***(4 marks)***

(ii) Give estimates for each parameter in model `lm1`, including the error variance. ***(2 marks)***

(iii) Calculate 95% confidence intervals for each parameter in model `lm1`, excluding the error variance. ***(4 marks)***

(iv) What proportion of variation in the observed retained impressions is described by the regression fit? ***(1 mark)***

(v) Using a suitable hypothesis test, investigate whether there is evidence of a relationship between advertising budget and the number of retained impressions. ***(3 marks)***

(vi) Suppose the standardised residuals are plotted against the fitted values for model `lm1`. What assumption would such a plot be used to check? State two features of the plot that you would inspect to test this assumption. ***(3 marks)***

1 (continued)

(vii) The R session is continued below.

```
> lm2<-lm(imp~budget+I(budget^2))
> deviance(lm2)
[1] 8568.853
> qf(0.95,1,18)
[1] 4.413873
```

Using an F-test, compare model `lm2` against `lm1`. Using suitable notation, state clearly what you are testing, and interpret the result.

(6 marks)

2 (i) Each of the following equations is a mathematical model for the relation between variables y_i and x_i, for $i = 1, \ldots, n$. Each model depends on two unknown parameters γ and δ, and the error term ε_i is normally distributed. State which of these models are linear models, and specify the design matrix X for each linear model.

(a) $y_i = \gamma + \delta \sin(x_i) + \varepsilon_i$.

(b) $y_i = \gamma + \sin(\delta x_i) + \varepsilon_i$.

(c) $y_i = \gamma x_i + \delta x_i^2 + \varepsilon_i$.

(d) $y_i = \gamma x_i + x_i^\delta + \varepsilon_i$. ***(6 marks)***

(ii) In a simplified linear regression model, the intercept and gradient are assumed equal to each other:

$$M_1 : \qquad y_i = \phi(1 + x_i) + \varepsilon_i,$$

for $i = 1, \ldots, n$, with $\varepsilon_i \sim N(0, \sigma^2)$.

(a) Write this model in matrix notation, and find the least squares estimate of ϕ (you may quote, without proof, the general formula for the least squares estimator for a linear model, denoted by $\hat{\boldsymbol{\beta}}$ in your lecture notes). ***(4 marks)***

(b) Given $\sigma^2 = 2$ and $x_i = i$ for $i = 1, 2, 3$, with $n = 3$, calculate the variance of your least squares estimate. Noting that σ^2 is known, and that 1.96 is the 97.5th percentile of the standard normal distribution, give a formula for a 95% confidence interval for ϕ (write your answer in terms of $\hat{\phi}$, the least squares estimator of ϕ). ***(5 marks)***

(c) An alternative model M_2 is to be fitted to the same data:

$$M_2 : \qquad y_i = \theta_0 + \theta_1 x_i + \varepsilon_i.$$

Which model out of M_1 and M_2 would you expect to have the smaller residual sum of squares? Briefly justify your answer. ***(3 marks)***

3 In a car safety experiment, cars containing crash test dummies in the driver's seat were crashed into a wall at 35 miles per hour. The recorded data included a measure of the head injury (known as the head injury criterion) sustained by the dummy, stored in R under the variable name `headinjury` and the type of protection offered to the driver (four different groups: airbag, manual seatbelt, motorised seatbelt or passive seatbelt, stored in R under the variable name `protection`). (Data obtained from the Data and Story Library). Nine cars were tested in each group.

(i) The following model is proposed for the data.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

for $i = 1, \ldots, 36$, where y_i is the recorded head injury criterion and x_i the corresponding type of protection, for the i-th car. Give one criticism of this model, and suggest an alternative, defining your notation carefully.

(4 marks)

(ii) A one way analysis of variance model is fitted to the data, and some R output is given below.

```
            Df  Sum Sq Mean Sq
protection              375777
Residuals               126320

> qf(0.95,3,32)
[1] 2.90112
```

(a) State the missing `Df` and `Sum Sq` values in the above table, briefly justifying your answers. ***(6 marks)***

(b) The 9 measurements for the airbag group are

497, 343, 493, 417, 580, 920, 435, 185, 298.

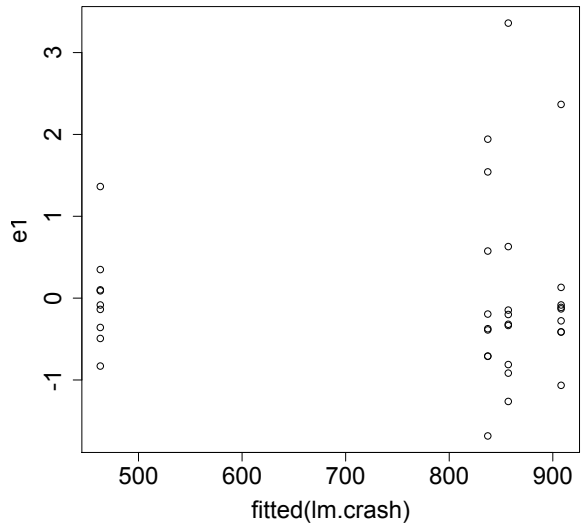
Define θ_1 to be the expected head injury criterion in the airbag group. Calculate the least squares estimate of θ_1. (You may state the form of the least squares estimator in this case without proof). Give a 95% confidence interval for θ_1, assuming a common variance for all four groups, and noting that $t_{32;0.975} = 2.0369$. ***(5 marks)***

(c) Test the hypothesis that there is no difference in mean head injury criterion between the four groups. ***(3 marks)***

3 (continued)

(iii) Below are some further R commands and a resulting plot. Explain the relevance of these commands with regard to your analysis in part (ii). What problem does the plot suggest, and what might be done to rectify it?

```
> lm.crash<-lm(headinjury~protection)
> library(MASS)
> e1<-stdres(lm.crash)
> plot(fitted(lm.crash),e1)
```



(6 marks)

(iv) Suppose another experiment is to be conducted, in which the cars will be driven at different speeds. If the speeds are known, write down a suitable model for analysing the data from this new experiment, defining your notation carefully.

(2 marks)

4 An R dataframe `teengamb` contains data from a study of teenage gambling in Britain (Data obtained from Faraway (2005), original source: Ide-Smith & Lea, 1988). The dependent variable, `gamble` is the gambling expenditure (in pounds per year) for each person, and the three independent variables are `sex` (stored as `female` or `male`); `income`, in pounds per week; `verbal` score, in words out of 12 correctly defined. There are 45 teenagers in the study.

(i) Below is some edited output from an R session.

```
> lm1<-lm(gamble~income+sex+verbal,teengamb)

Coefficients:
            Estimate Std. Error
(Intercept)   1.1788    14.0441
income        4.8981     0.9551
sexmale      22.9602     6.7706
verbal       -2.7468     1.8253

> qt(0.975,43)
[1] 2.016692
> qt(0.995,43)
[1] 2.695102
```

Conduct suitable tests to see whether there is evidence of a relationship between gambling expenditure and each of the three independent variables. ***(6 marks)***

(ii) A more complex model is fitted to the data, with the R command and output shown below.

```
> lm(gamble~income*sex*verbal,teengamb)

(Intercept)    income   sexmale    verbal
     -17.8        4.2      18.1       3.2
income:sexmale    income:verbal   sexmale:verbal
           6.2              -0.6             -3.4
income:sexmale:verbal
                -0.1
```

Defining your notation carefully, write down the fitted regression lines giving the relationships between

(a) gambling expenditure, income and verbal score for females;

(b) gambling expenditure, income and verbal score for males.

(4 marks)

5 A machine used to test the effect of wear on fabrics has three different settings (corresponding to different degrees of abrasion on the fabric). Two different fabrics are to be compared and each type of fabric is used once on each setting for a specified period of time. The loss in weight (in grammes) on each piece of cloth is measured.

		Setting		
		I	II	III
Fabric	A	7	12	10
	B	4	9	13

(i) Defining your notation carefully, write down the most complex model appropriate for these data, stating any necessary parameter constraints.

(4 marks)

(ii) Sketch a plot of the data that allows you to check for the presence of an interaction between the two factors in this model. Comment briefly on your plot. ***(4 marks)***

End of Question Paper