The
University
Of
Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Spring Semester 2012–2013**

MAS273 Statistical Modelling **2 hours**

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

1 Ethanol fuel was burned in a single-cylinder engine. For various settings of the engine compression and equivalence ratio, the emissions of nitrogen oxides were recorded. The variables are:

NOx : Concentration of nitrogen oxides (NO and NO2) in micrograms/J.

C : Compression ratio of the engine.

E : Equivalence ratio is a measure of the richness of the air and ethanol fuel mixture.

A regression model for predicting NOx was fit to the data [$NOx = \beta_0 + \beta_C C + \beta_E E + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$] and the following output was obtained:

```
Call:  lm(formula = NOx~ C + E)
Residuals:
```

Min	1Q	Median	3Q	Max
-1.769	-0.950	-0.254	1.038	2.096

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.559	0.662	3.863	0.000218
C	-0.007	0.031	-0.228	0.820
E	-0.557	0.601	-0.926	0.357

```
Residual standard error:  1.14 on 85 degrees of freedom
Multiple R-squared:  0.01095,   Adjusted R-squared:  -0.01232
F-statistic:  0.4707 on 2 and 85 DF,    p-value:  0.6262
```

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

(ii) Give an estimate of the expected value of NOx when $C = 12$ and $E = 1$. ***(2 marks)***

(iii) Give the numerical value of the unbiased estimator of σ^2.
Give the value of the residual sum of squares for the model. ***(2 marks)***

(iv) Are the p-values for the t-statistics computed for a one or two sided test? ***(1 mark)***

(v) State the mean of the residuals for this model. If there is not sufficient information in the data, say so. Justify your answer. ***(3 marks)***

1 (continued)

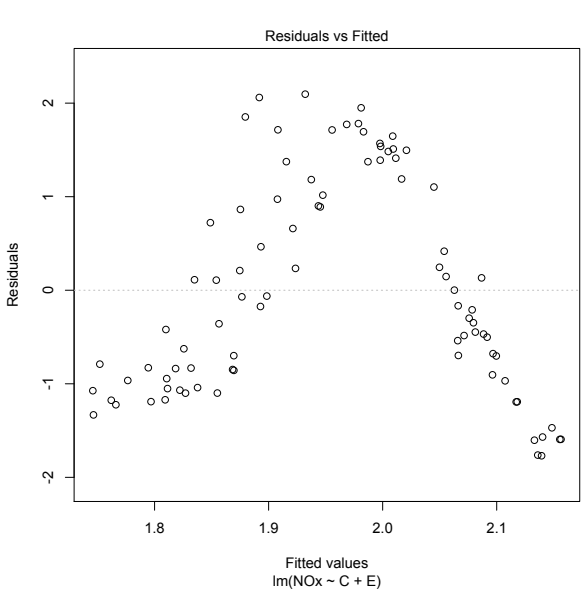
(vi) How would you compute an exact 95% confidence interval for β_E? Give an approximate 95% confidence interval. You will need that $P(N(0,1) \leq 1.96) = 0.975$. ***(5 marks)***

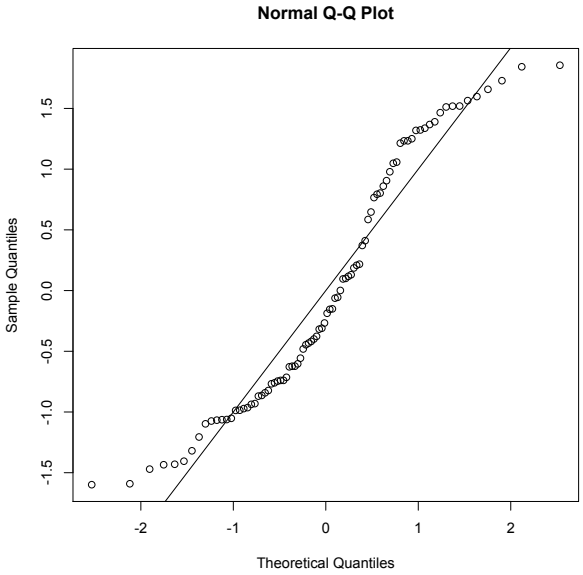
(vii) Is there any evidence against $H_0 : \beta_C = \beta_E = 0$? Describe a test and report your findings. ***(4 marks)***

(viii) Is there any evidence against $\beta_0 = 0$? Describe a test and report your findings. ***(4 marks)***

(ix) The first panel in the figure below shows the residuals versus fitted plot for the model. What can be concluded from the plot? ***(2 marks)***

(x) The second panel in the figure below shows the QQ plot of the residuals for this model. What can be concluded from the plot? ***(2 marks)***





2 (i) Each of the following equations is a mathematical model for the relation between variables y_i and x_i, for $i = 1, \cdots, n$. Each model depends on two unknown parameters α and β, 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. ***(6 marks)***

(a) $y_i = \alpha\beta^{x_i} + \epsilon_i$.

(b) $y_i = \alpha x_i + \beta x_i^2 e^{3x_i} + \epsilon_i$.

(c) $y_i = \beta + \dfrac{\alpha}{1 + x_i} + \epsilon_i$.

(ii) Two compounds A and B having unknown weights α and β respectively, are measured on a spring balance separately. A third compound C, made by mixing *half* of A and *half* of B is then measured on the balance. Each measurement on the spring balance gives a random measurement error. Write a statistical model to describe this data and find the least squares estimates for α and β. ***(6 marks)***

(iii) Consider the linear model $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$, where $\underset{\sim}{\epsilon} \sim N(0, \sigma^2 I)$. Show that the least squares estimatior $\underset{\sim}{\hat{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{y}$ is unbiased for $\underset{\sim}{\beta}$ and the variance of $\underset{\sim}{\hat{\beta}}$ is $\sigma^2 (X^T X)^{-1}$. ***(4 marks)***

(iv) Suppose we have a data set (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n). Consider two different linear models $y_i = \alpha + \beta x_i + \epsilon_i$ and $y_i = \gamma + \delta(1 + x_i) + \epsilon_i$, where α, β, γ, δ are unknown parameters. Explain why the residual sums of squares for the two models are equal. ***(5 marks)***

3 (i) Agriculturists in Sheffield were interested in the effectiveness of 3 insecticides on the growth of a certain plant. Three treatments `trt1`, `trt2` and `trt3`, were used in the study. 30 plants were randomly divided into 3 equal groups for each of the three treatments and the weights of the plant were recorded after 1 year. Below is a *partial* output from `R`.

```
anova(lm(weight~group))

Analysis of Variance Table

Response:  weight
```

	Df	Sum Sq	Mean Sq	F value
group	?	?	?	4.846
Residuals	?	?	0.387	

Complete the ANOVA table by filling in the entries marked by ? .

(6 marks)

(ii) Nine patients are divided at random into three groups. Suppose that each group receives a different treatment for a month and the data below indicate the individuals responses to the treatments.

Group 1: 2, 5, 5, Group 2: 4, 5, 6, Group 3: 8, 7, 9.

Denote by μ_1, μ_2 and μ_3 the mean responses for groups 1, 2 and 3 respectively.

(a) The model is the *One Way Analysis of Variance*. Write down the model, clearly stating any assumptions you make. ***(3 marks)***

(b) Consider the model in part (a). Is there any evidence against the null hypothesis $H_0 : \mu_1 = \mu_2, \mu_3 = 8$? 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*).

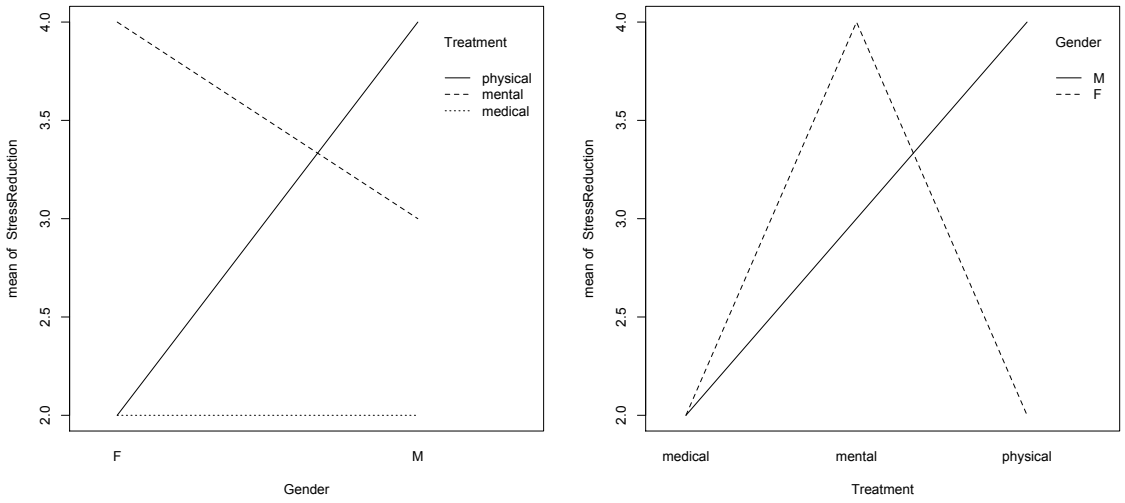
(15 marks)

4 A group of 30 male and 30 female subjects were chosen for a study to understand the effectiveness of 3 treatment options (*mental, medical* and *physical*) on stress reduction. The stress reduction was recorded on a scale from 1 to 5. The subjects were randomly divided so that we have a two-way *balanced* design, with 10 subjects in each (gender, treatment) group. The most general model for the two way ANOVA is of the form

$$SR_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \epsilon_{i,j,k},$$

where $SR_{i,j,k}$ is the stress reduction value for the kth subject in the (i,j)th group. $i = 1, 2$ depending on whether the subject is *male* or *female* and $j = 1, 2, 3$ depending on whether the subject has the treatments *mental, medical* and *physical* respectively.

(i) The interaction plots for the two-way ANOVA are as below



What do the plots suggest? Explain. ***(3 marks)***

 Question 4 continued on next page

4 (continued)

(ii) The following output was obtained from R

```
anova(lm(StressReduction~Treatment*Gender))
Analysis of Variance Table

Response:  StressReduction
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatment	2	23.333	11.667	17.5	1.384e-06 ***
Gender	1	1.667	1.667	2.5	0.1197
Treatment:Gender	2	23.333	11.6667	17.5	1.384e-06 ***
Residuals	54	36.000	0.667		

Each of the first three lines of the ANOVA (`Treatment`, `Gender`, `Treatment:Gender`) corresponds to a hypothesis test.

(a) Write down the null hypothesis for each of these lines, give the degrees of freedom for the F statistic and the P-values for the test. *Note: You don't have to mention the test statistic.* ***(9 marks)***

(b) Do the tests support the plots above? Explain. ***(2 marks)***

End of Question Paper