MATH371301

This question paper consists of 6 printed pages, each of which is identified by the reference **MATH3713**.

Graph paper is provided. Only approved basic scientific calculators may be used.

©UNIVERSITY OF LEEDS

Examination for the Module MATH3713
(January 2005)

Regression and Smoothing

Time allowed: **3 hours**

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

1. (a) Define carefully the multiple linear regression model with n observations and k explanatory variables. Explain why it is intuitively natural to estimate the regression parameter vector, $\boldsymbol{\beta}$, say, by minimizing a certain sum of squares. Show that the resulting estimate of $\boldsymbol{\beta}$ takes the form

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y},$$

where, in this formula you should define the notation and specify the dimensions of the various quantities.

(b) Let $\hat{y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ denote a predicted value of the response variable y_0, say, at a new vector of explanatory variables, $\boldsymbol{x}_0$. Find the form of a 95% prediction interval for y_0.

(c) Find $\hat{y}_0$ if

$$X = \begin{bmatrix} 1 & -1 & 2 \\ 1 & 2 & 0 \\ 1 & -1 & -2 \end{bmatrix}, \quad \boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} -2 \\ 1 \\ 3 \end{bmatrix},$$

and try to evaluate the corresponding prediction interval. Carry out the calculations as far as you can, commenting on any difficulties you encounter.

2. Consider an $(n \times 1)$ vector $\boldsymbol{y}$ and an $(n \times p)$ design matrix X in a multiple linear regression model. Let $H = XCX^T$ denote the hat matrix where $C = (X^T X)^{-1}$.

(a) The raw residuals are defined as the difference between the observed and fitted values of $\boldsymbol{y}$. Show that the vector of raw residuals can be represented in the form

$$\boldsymbol{e} = (I - H)\boldsymbol{y}\,.$$

(b) Define the deletion quantities X_{-i}, C_{-i}, $\boldsymbol{y}_{-i}$, and $\hat{\boldsymbol{\beta}}_{-i}$, and state the range of values that i can take. Starting from the result $C_{-i} = C + C\boldsymbol{x}_i\boldsymbol{x}_i^T C/(1-h_{ii})$ where $\boldsymbol{x}_i^T$ denotes the ith row of X, show that

$$\hat{\boldsymbol{\beta}}_{-i} = \hat{\boldsymbol{\beta}} - \frac{e_i}{1-h_{ii}} C\boldsymbol{x}_i.$$

(c) Cook's distance is defined by

$$D_i = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^T X^T X(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})/(p\hat{\sigma}^2).$$

Explain how this statistic can be used in regression analysis and why its value is often compared to an $F_{p,n-p}$ distribution. In particular, what conclusions would you draw if $n = 36$, $p = 5$, and for some particular value of i, $D_i = 1.5$.

[Hint: the following R output gives various percentiles from the $F_{5,31}$ distribution, rounded to 3 decimal places.]

```
> round(qf(c(.01,.05,.10,.50,.90,.95,.99),5,31),3)
[1] 0.107 0.223 0.315 0.890 2.042 2.523 3.675
```

3. (a) Consider a multiple linear regression with a large number of possible regressor variables. The AIC criterion for a model M containing a subset of the possible regressor variables is given by

$$\text{AIC}(M) = n\log(SS_E(M)/n) + 2p(M).$$

Define the terms in this expression.

Two statisticians have different strategies for model selection.

(i) Statistician A seeks a model M to minimize the criterion AIC(M).

(ii) Statistician B seeks a model M to minimize the criterion $SS_E(M)$.

Explain the intuition behind the strategy of Statistician A. What is the problem with the strategy of Statistician B?

(b) In a dataset involving a response variable `y` and 3 possible regressor variables, `x1, x2, x3`, the following AIC values, rounded to the nearest integer, were obtained.

Model	AIC
-	16
1	-14
2	15
3	-11
12	-45
13	-13
23	-47
123	-46

The entry under "Model" lists the regressor variables in the model. For the top row, only an intercept is present.

Starting with the model 123, describe how the `step` procedure in R moves through various possible models to reach a final choice, the model 23.

(c) For the dataset in part (b), the variance inflation factors took the values 42.5, 1.2, 43.6, and a `summary(lm(y ~ x1+x2+x3))` command in R yielded the following output.

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70657 -0.31900 -0.04266  0.30394  0.93017

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03265    0.09660   0.338    0.738
x1           0.01340    0.01752   0.765    0.451
x2           0.67645    0.08899   7.601 4.55e-08 ***
x3          -0.18521    0.12550  -1.476    0.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.436 on 26 degrees of freedom
Multiple R-Squared: 0.896,      Adjusted R-squared: 0.884
F-statistic: 74.68 on 3 and 26 DF,  p-value: 6.613e-13
```

Explain how it can happen that `x3` is not significant the the table of coefficients, yet it appears in the final model in part (b).

Using the output of the `summary` command, confirm the AIC value in the bottom line of the table in part (b).

4. A solid-fuel rocket propellant loses weight after it is produced. A set of 10 values is available for the weight loss `y` (in kg) and the number of months since production `x`. The R commands given below read in the data, fit 4 models, and produce the plots given in Figure 1. Note that `xr=sqrt(x)` and `xlog=log(x)`.

```
x=(1:10)/4
y=c(0.868, 1.215, 1.498, 1.726, 1.940, 2.130, 2.284, 2.451, 2.595, 2.743)
par(mfrow=c(3,2))
plot(x,y,main="y vs x",sub="(a)")
lm1=lm(y~x)
abline(lm1)
e1=residuals(lm1)
plot(x,e1,main="residuals vs x for model y~x",sub="(b)")
xr=sqrt(x)
lm2=lm(y~xr)
e2=residuals(lm2)
plot(xr,e2,main="residuals vs xr for model y~xr",sub="(c)")
xlog=log(x)
lm3=lm(y~xlog)
e3=residuals(lm3)
```

```
plot(xlog,e3,main="residuals vs xlog for model y~xlog",sub="(d)")
lm4=lm(y~poly(x,degree=2))
e4=residuals(lm4)
plot(x,e4,main="residuals vs x for model y~poly(x,2)",sub="(e)")
par(mfrow=c(1,1))
```

As a result of some further R commands (not provided here) for each fitted model, the following expressions were obtained for the expected response, as a function of `x`, and for the error sum of squares.

lm1 $E(y|x) = 0.84 + 0.80x, \quad SS_E = 0.0722$
lm2 $E(y|x) = 0.01 + 1.74\sqrt{x}, \quad SS_E = 0.0003$
lm3 $E(y|x) = 1.84 + 0.83 \log x, \quad SS_E = 0.0961$
lm4 $E(y|x) = 0.59 + 1.30x - 0.18x^2, \quad SS_E = 0.0053$

Using this output, describe what conclusions can be reached about the relationship between `y` and `x`, focusing on the following points.

(i) Describe in general terms when transformations of `y` and/or `x` are worth considering when modelling the relationship between `y` and `x`.

(ii) Which of these considerations are relevant for the dataset here?

(iii) What information is provided by the residual plots (b) – (e) in Figure 1?

(iv) Comment on the quality of the fits of the different models.

(v) It is desired to predict the amount of weight loss after 5 months. Evaluate the predicted weight loss from each of the 4 models, and comment on the reasons for the differences between the predictors.

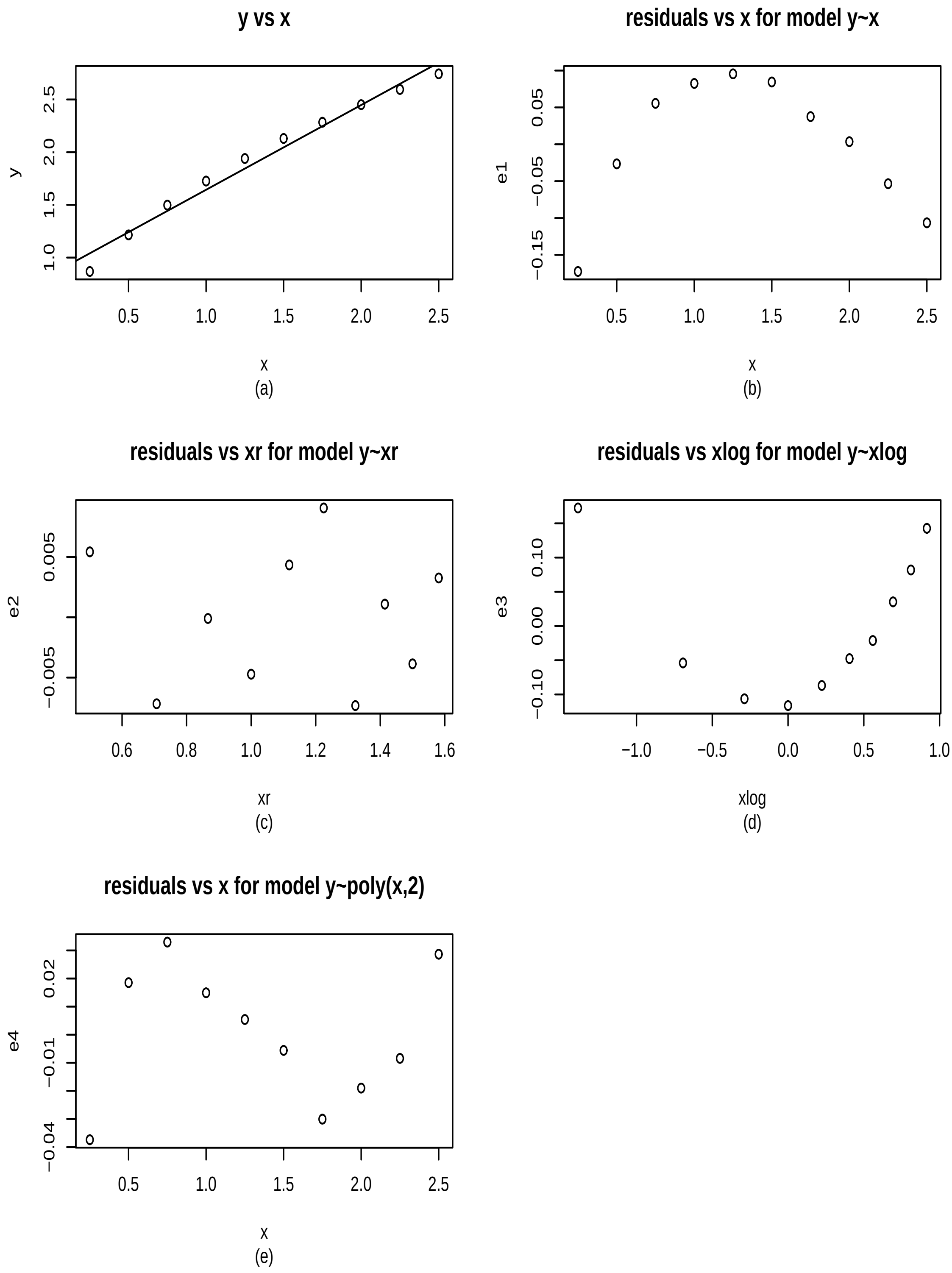

Figure 1: Plots for Question 4.

5. Let $x_1, \ldots, x_n$ be independent observations from a twice continuously differentiable, but otherwise unknown, probability density $f(x),\ x \in \mathbb{R}$.

(a) Define the kernel density estimate $\hat{f}_h(x)$ based on a kernel function $K(x)$ and a bandwidth parameter $h > 0$. State the regularity assumptions usually assumed of $K(x)$.

(b) For fixed x the bias and variance of $\hat{f}_h(x)$ are given by

$$E\{\hat{f}_h(x) - f(x)\} = \frac{h^2}{2} f''(x) \quad \text{and} \quad \text{var}\{\hat{f}_h(x)\} = (nh)^{-1} C_2 f(x),$$

plus smaller order remainder terms. Verify the first of these expressions, give a formula for the constant C_2 and find conditions on n and h under which the bias and variance will tend to 0.

(c) Hence show that if $h = n^{-1/5}$, then $\hat{f}_h(x)$ will converge in mean square to $f(x)$ for any fixed x as $n \to \infty$.

(d) Let $a < b$. Show that as $n \to \infty$ for fixed h,

$$\int_a^b \hat{f}_h(x)\, dx \to h^{-1} \int_a^b \left\{ \int_{-\infty}^{\infty} f(x - y) K(y/h)\, dy \right\} dx.$$

END