MATH371301

This question paper consists of 7 printed pages, each of which is identified by the reference **MATH3713**.

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

Examination for the Module MATH3713
(January 2006)

Regression and Smoothing

Time allowed: **3 hours**

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

1. The multiple linear regression model involving n observations and k explanatory variables can be written in the form

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \epsilon_i \, .$$

(a) Explain each of the terms in this model, the assumptions on the $\{\epsilon_i\}$ and the ranges for the indices i and j.

(b) The maximum likelihood estimator of the regression parameter $\boldsymbol{\beta}$ can be written in matrix form as $\hat{\boldsymbol{\beta}} = CX^T\boldsymbol{y}$. Verify this formula, taking care to define and give the dimensions of C, X and $\boldsymbol{y}$.

(c) A scientist wishes to explore the dependence of January rainfall (in mm) on mean January temperature in Leeds (in °C) by fitting a simple linear regression ($k = 1$) using 20 years of measurements. He obtains estimated regression coefficients $\hat{\beta}_0 = 10.2$ and $\hat{\beta}_1 = 2.4$, with estimated standard deviations 2.2 and 1.3, respectively.

(i) Find the associated t-statistic for $\hat{\beta}_1$. Describe what hypothesis can be tested with this t-statistic and state what conclusions can be reached in this example.

(ii) A second scientist repeats the analysis on the same dataset but in different units, using inches for rainfall and °F for temperature. Give the new values of the estimated regression coefficients, and give the new standard error and t-statistic for $\hat{\beta}_1$. What effect does this change in units have on the conclusion reached in part (i)?

[Hints: (1) Note that 1 inch equals 25.4 mm. (2) If C denotes a temperature on the Celsius scale (°C), then the corresponding value F on the Fahrenheit scale (°F) is given by $F = 32 + (9/5)C$. (3) The following R output gives some upper 2.5% critical values for the t-distribution.]

```
round(qt(.975,c(15,16,17,18,19,20)),3)
[1] 2.131 2.120 2.110 2.101 2.093 2.086
```

2. The following R output gives the result of fitting a multiple linear regression model. Here `y, x1, x2` are all numeric vectors of the same length.

```
> lm1=lm(y~x1+x2)
> summary(lm1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-1.8942 -0.5667  0.1963  0.7334  1.7128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2684     0.9637  -0.279    0.787
x1            3.2793     0.3328   A     4.04e-06 ***
x2            1.8909     0.2319   8.153 1.90e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.263 on 9 degrees of freedom
Multiple R-Squared: 0.9913,     Adjusted R-squared: B
F-statistic: 511.6 on 2 and 9 DF,  p-value: 5.395e-10

> anova(lm1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)

x1         1 1525.50 1525.50 956.803 1.890e-10 ***
x2         1  105.98  105.98  C      1.902e-05 ***
Residuals  D   14.35    1.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

(a) In this R output, 4 entries labelled A – D are missing. Give these values. Also give the sample size of the dataset and the values of SS_T, SS_R, SS_E, the "total", "regression" and "error" sums of squares, respectively.

(b) For the multiple linear regression model, give the definitions for the three types of residuals e_i, r_i, and t_i (raw, standardized and studentized, respectively).

(c) The statistic DFFITS_i is defined by

$$\text{DFFITS}_i = (\hat{y}_i - \hat{y}_{i,-i}) \,/\, \left(\hat{\sigma}_{-i}\sqrt{h_{ii}}\right).$$

Explain this notation and motivate the definition.
It is known that DFFITS_i can also be written in the form

$$\text{DFFITS}_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i.$$

Using the second representation, explain how this statistic combines the effects of leverage and outlyingness.
If $h_{11} = 0.2$ and $t_1 = 6$ for the dataset in (a), what conclusions would you draw?

3. Consider a multiple linear regression model with $k > 1$ regressor variables $x_1, \ldots, x_k$ on n observations.

(a) Why is variable selection an important problem? Explain the method of forward search, describing, in particular, how new variables are added during the algorithm.

(b) Consider a hypothesis test to compare the models

$$M_0 : y \sim x_1 \text{ vs. } M_1 : y \sim x_1 + x_2.$$

Using the error sum of squares for the two models (labelled $SS_E(1)$ and $SS_E(1,2)$, say), show how an F statistic with 1 and $n-3$ degrees of freedom can be constructed to carry out the test.

(c) In a dataset involving a response variable y and three possible regressor variables $x_1,\ x_2,\ x_3$ on $n = 25$ observations, the following table lists the error sums of squares (SS_E) for a collection of models. The entry under "Model" lists the regressor variables in the model. For the top row, only an intercept is present.

Model	SS_E
-	88
1	52
2	48
3	82
12	46
13	29
23	24
123	18

Carry out one step of forwards search and backwards elimination on this dataset, and compare the models chosen in each case.

[Hint: The following R output gives some upper 5% critical values for the F distribution.]

```
> round(qf(.95,1,c(20,21,22,23,24,25)),2)
[1] 4.35 4.32 4.30 4.28 4.26 4.24
```

4. (a) Consider the family of transformations indexed by a real parameter λ, defined by

$$f_\lambda(x) = \frac{x^\lambda - 1}{\lambda}, \quad x > 0, \quad \lambda \neq 0$$

and by

$$f_0(x) = \log(x), \quad x > 0.$$

Show that for $\lambda < 1$, $f_\lambda(x)$ is increasing and concave in $x > 0$ with $f'(1) = 1$. What happens if $\lambda > 1$?

(b) Suppose that in a regression analysis of a real-valued response variable y on a real-valued explanatory variable x, a plot of y vs. x suggests that y is approximately a decreasing convex function of x. Discuss (i) conditions under which applying the transformation in (a) to y or x might be useful, (ii) what values of λ might be appropriate, (iii) why such a transformation might be useful, and (iv) how you might assess the suitability of a particular transformation.

(c) In a study to investigate the effectiveness of slug pellets, various dosages (x = number of pellets per square metre) were sprinkled on a selection of identically-sized English country gardens one evening, and the numbers (y) of dead slugs the next morning were counted. A plot of y vs. x is given in Figure 1. There were three replications at each of 4 doses, so $n = 12$.

A statistician considers three models $M_1 : y \sim x$, $M_2 : y \sim x^{1/3}$ and $M_3 : y \sim x + x^{1/3}$. On the basis of the information in the following R output, investigate which of the models seems to be most appropriate.

```
> x
 [1]  25  25  25  50  50  50 100 100 100 200 200 200
> y
 [1] 332 295 330 337 284 395 545 457 426 560 464 564
> lm1=lm(y~x)
> summary(lm1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-84.609 -26.728   2.326  19.902 121.435 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 298.5217    28.1999   10.59 9.41e-07 ***
x             1.2504     0.2447    5.11 0.000457 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

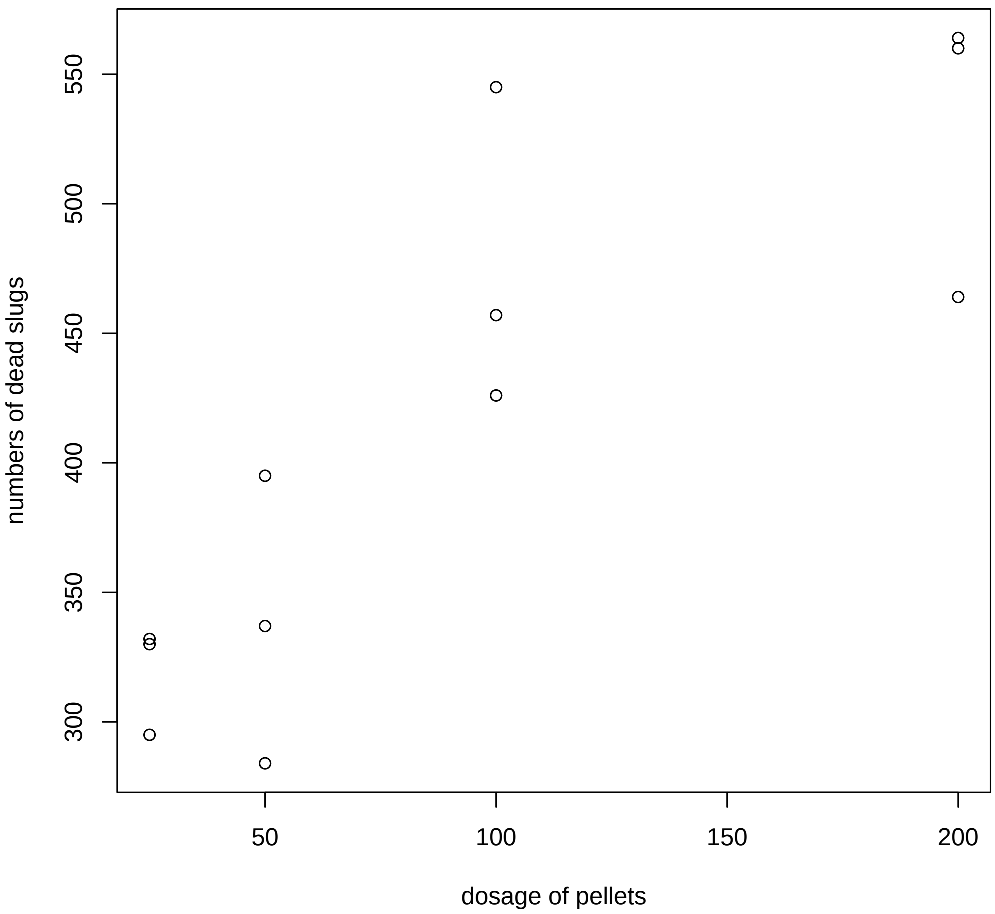

Figure 1: Plot for Question 4: numbers of dead slugs vs. dosage of slug pellets.

```
Residual standard error: 56.81 on 10 degrees of freedom
Multiple R-Squared: 0.7231,     Adjusted R-squared: 0.6954
F-statistic: 26.11 on 1 and 10 DF,  p-value: 0.0004572

> x3=x^(1/3)
> lm2=lm(y~x3)
> summary(lm2)

Call:
lm(formula = y ~ x3)

Residuals:
   Min     1Q Median     3Q    Max
-85.22 -22.07  16.26  23.07 100.31
```

 QUESTION 4 CONTINUED...

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    78.84      60.98   1.293 0.225160
x3             78.82      13.82   5.703 0.000198 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.36 on 10 degrees of freedom
Multiple R-Squared: 0.7648,     Adjusted R-squared: 0.7413
F-statistic: 32.52 on 1 and 10 DF,  p-value: 0.0001976

> lm3=lm(y~x+x3)
> summary(lm3)

Call:
lm(formula = y ~ x + x3)

Residuals:
   Min     1Q Median     3Q    Max
-88.50 -27.67  14.22  26.25  93.94

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.0818   219.8512   0.060    0.954
x            -0.4015     1.2845  -0.313    0.762
x3          103.0109    78.7292   1.308    0.223

Residual standard error: 54.89 on 9 degrees of freedom
Multiple R-Squared: 0.7674,     Adjusted R-squared: 0.7157
F-statistic: 14.84 on 2 and 9 DF,  p-value: 0.001413
```

5. Given data points $x_1 < \cdots < x_n$ and a bandwidth parameter $h > 0$, the kernel density estimate (KDE) is defined by the function

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K((x - x_i)/h).$$

(a) What regularity conditions are usually assumed for the kernel $K(\cdot)$ and what probability model is usually assumed for the data? Make these assumptions for the remainder of the question.

(b) Find the limit of the $\hat{f}_h(x)$ as $n \to \infty$ for fixed h and x.

(c) For fixed data and assuming $x = x_i$ for some $i = 1, \ldots, n$, so that x equals one of the data points, show that

$$\hat{f}_h(x) \to \infty \text{ as } h \to 0, \text{ and}$$
$$\hat{f}_h(x) \to 0 \text{ as } h \to \infty.$$

What is the significance of these results for data analysis?

(d) In the asymptotic theory of kernel density estimation, under the conditions $n \to \infty$, $h \to 0$ and $nh \to \infty$, it can be shown that the bias and variance of $\hat{f}_h(x)$ for fixed x can be approximated by

$$\frac{h^2}{2}A(x) \text{ and } (nh)^{-1}B(x)$$

for certain functions $A(x)$ and $B(x)$, where $B(x) > 0$. Using these results describe the asymptotic behaviour of the mean squared error of $\hat{f}_h(x)$ for fixed x under the following settings:

(i) $h = n^{-4/5}$,
(ii) $h = n^{-1/5}$,
(iii) $h = n^{-1/10}$.

Assuming $A(x) \neq 0$, show that choice (ii) is better than the others. Find a constant c (depending on x) such that $h = ch^{-1/5}$ gives an improvement on choice (ii).

END