MATH371301

This question paper consists of 5 printed pages, each of which is identified by the reference **MATH371301**.

New Cambridge Elementary Statistical Tables are provided. Only approved basic scientific calculators may be used.

Examination for the Module MATH3713

(January 2004)

REGRESSION AND SMOOTHING

Time allowed: **3 hours**

Do not attempt more than four questions.

All questions carry equal marks.

1. Suppose we have data $\{(x_i, y_i), i = 1, \ldots, n\}$ and that we want to obtain a smooth estimate of $m(x) = \mathsf{E}(Y \mid X = x)$. Consider the Nadaraya-Watson estimator given by

$$\hat{m}_h(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)} = \frac{(1/n)\sum K_h(x - x_i)y_i}{(1/n)\sum K_h(x - x_i)},$$

whith K a symmetric kernel function.

(i) Using the more general form

$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n} w_i(x)y_i$$

obtain the form of the weights $w_i(x)$, and discuss the role of the smoothing parameter h.

(ii) What is the limiting behaviour as

(a) $h \to 0$, and

(b) $h \to \infty$?

(iii) Show that $\mathsf{E}[\hat{r}_h(x)]$ can be written as $\int K_h(x - u)r(u)du$, where $r(x) = m(x)f(x)$.

(iv) By making a change of variable and expanding as a Taylor series, show that

$$\mathsf{E}[\hat{r}_h(x)] = r(x) + \frac{h^2}{2}r''(x)\mu_2(K) + o(h^2) \qquad \text{as } h \to 0$$

where $\mu_2(K)$ should be defined.

(v) Discuss how the smoothing parameter h could be chosen in practice.

2. Consider the multiple linear regression model given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$\boldsymbol{y}$	is $n \times 1$ vector of observations
$\boldsymbol{X}$	is $n \times (p + 1)$ full rank matrix of explanatory variables
$\boldsymbol{\beta}$	is $(p + 1) \times 1$ vector of regression coefficients
$\boldsymbol{\varepsilon}$	is $n \times 1$ vector of random errors, with $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$.

Prove that the least squares estimator of $\boldsymbol{\beta}$ is given by

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

Derive the mean and variance of $\hat{\boldsymbol{\beta}}$, taking care to quote any general results that you use. What is the distribution of $\hat{\boldsymbol{\beta}}$, and why?

By considering the residual sum of squares, show that the total (uncorrected) sum of squares can be partitioned into a model sum of squares, $\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}$, and a residual sum of squares $\boldsymbol{y}^T\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\}\boldsymbol{y}$. Quoting any general results that you use, prove that the model sum of squares and the residual sum of squares are independent.

3. (a) Show that the density scaled histogram can be expressed as a step function of the form

$$\hat{f}(x) = \begin{cases} n_i/\{n(t_i - t_{i-1})\} & x \in [t_{i-1}, t_i) \\ 0 & \text{otherwise} \end{cases}$$

where n_i is the number of observations, $\{x_j, j = 1, \ldots, n\}$, in the interval $[t_{i-1}, t_i)$. Show also that the heights in each interval are equivalent to the maximum likelihood estimate of a step function in which the t_i are given.

(b) State up to three distinct criteria for selecting subset regression models. In each case, provide a definition and describe how the criterion would be used in practice.

4. (a) Define the coefficient of multiple determination R^2, and briefly describe its role as a multiple regression diagnostic aid.
Given that $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ where

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

show that R^2 is equal to the square of the correlation between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$.

(b) Suppose $\boldsymbol{X}_{(I)}$ is the design matrix with rows belong to set I removed, and $\boldsymbol{X}_I$ are the rows of $\boldsymbol{X}$ belonging to a set I. Given that

$$(\boldsymbol{A} - \boldsymbol{U}\boldsymbol{V}^T)^{-1} = \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I}_m - \boldsymbol{V}^T\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^T\boldsymbol{A}^{-1}$$

for a square ($(p+1)\times(p+1)$) matrix $\boldsymbol{A}$, and with $\boldsymbol{U}$ and $\boldsymbol{V}$ of dimension $(p+1)\times m$, show that

$$(\boldsymbol{X}_{(I)}^T\boldsymbol{X}_{(I)})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I^T(\boldsymbol{I}_m - \boldsymbol{H}_I)^{-1}\boldsymbol{X}_I(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

where

$$\boldsymbol{H}_I = \boldsymbol{X}_I(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I^T.$$

Hence show that if observation i is removed then the resulting estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x}_i}{(1 - h_{ii})}e_i$$

where $e_i = y_i - \hat{y}_i$ and $h_{ii} = \mathbf{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{x}_i$.

5. The following computer output shows the results of fitting a regression model to the winning time taken (y in minutes) to climb Scottish hills based on distance (x_1 in miles) and height climbed (x_2 in feet) for 35 races.

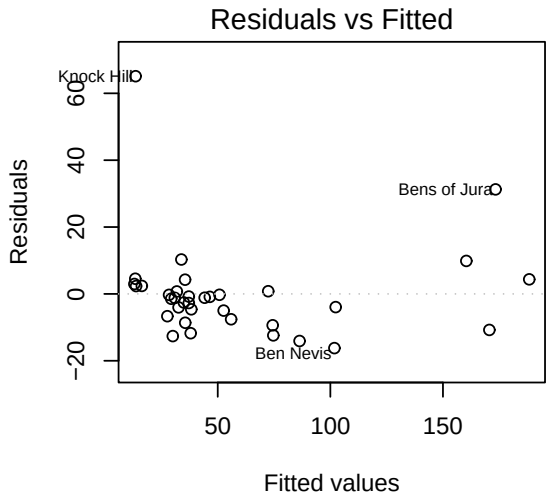
```
> lm1=lm(time ~ ., data=hills)
> summary(aov(lm1))
            Df Sum Sq Mean Sq F value    Pr(>F)
dist         1  71997   71997 334.293 < 2.2e-16 ***
climb        1   6250    6250    A    6.445e-06 ***
Residuals    B   6892      C
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(lm1)

Call:
lm(formula = time ~ ., data = hills)

Residuals:
    Min      1Q  Median      3Q     Max
-16.215  -7.129  -1.186   2.371  65.121

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.992039   4.302734     D       E    *
dist         6.217956   0.601148  10.343 9.86e-12 ***
climb        0.011048   0.002051   5.387 6.45e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:    F   on 32 degrees of freedom
Multiple R-Squared: G,     Adjusted R-squared: 0.914
F-statistic: H on 2 and 32 DF,  p-value: < 2.2e-16
```

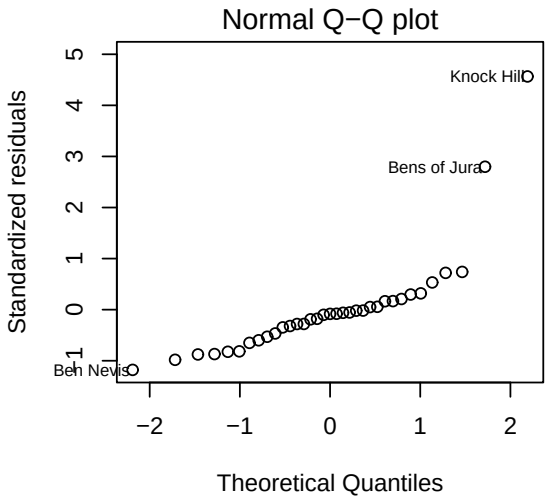
Calculate the missing values at A–H.

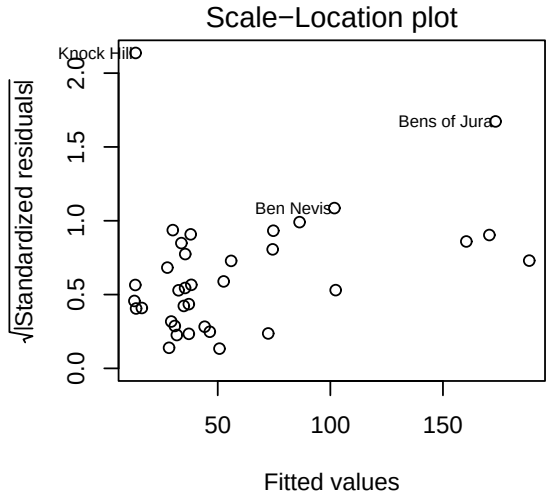
Predict the time taken to climb Ben Venue, which has a distance of 8.1 miles, and height climbed 2700 feet.

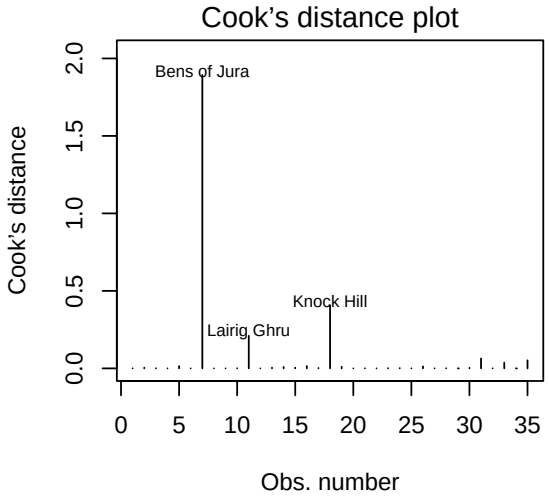
The graphs overleaf show some plots based on the fitted model. Fully describe each of these plots giving mathematical expressions for the quantities in the x and y axes, and interpret the information they provide.

State what further steps you would take in the analysis of these data.


Residuals vs Fitted
Residuals
Fitted values
Knock Hill
Bens of Jura
Ben Nevis


Normal Q−Q plot
Standardized residuals
Theoretical Quantiles
Knock Hill
Bens of Jura
Ben Nevis


Scale−Location plot
√|Standardized residuals|
Fitted values
Knock Hill
Bens of Jura
Ben Nevis


Cook's distance plot
Cook's distance
Obs. number
Bens of Jura
Knock Hill
Lairig Ghru

END