MATH371301

This question paper consists of 6 printed pages, each of which is identified by the reference **MATH371301**.

Required statistical tables are appended to this exam. Only approved basic scientific calculators may be used.

Examination for the Module MATH3713
(January 2007)

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

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

with $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$ a symmetric kernel function.

(i) Discuss the role of the smoothing parameter h. What is the limiting behaviour as

(a) $h \to 0$, and

(b) $h \to \infty$?

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

(iii) 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.

(iv) Given that the expected value of $\hat{f}_h(x)$ can similarly be written as

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

find, in terms of $m(\cdot)$ and $f(\cdot)$ (but not $r(\cdot)$), an asymptotic expression for the expected value of $\hat{m}_h(x)$

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

2. In the multiple linear regression model given by

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

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 $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$.

you may assume that the fitted values of $\boldsymbol{y}$ are given by $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

(a) Show that the residual sum of squares $\sum(y_i - \hat{y}_i)^2$ can be expressed as $\boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y}$ and use this to obtain an estimate of σ^2.

Prove, stating explicitly any results that you make use of, that your estimate is unbiased.

(b) In the above model, suppose that $p = 4$ and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_4)$. Describe in detail how you would test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2, \beta_3 = \beta_4 \\ vs\ H_1 &: \text{at least one of } \beta_1 \neq \beta_2, \beta_3 \neq \beta_4 \end{aligned}$$

3. Consider the multiple linear regression model described in Question 2.

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.

4. (a) Suppose $u_1, \ldots, u_{n_1}$ observations are taken from group 1, and $v_1, \ldots, v_{n_2}$ observations from group 2, with

$$\begin{aligned} u_i &\sim N(\mu_U, \sigma_U^2) \qquad i = 1, \ldots, n_1 \\ v_i &\sim N(\mu_V, \sigma_V^2) \qquad i = 1, \ldots, n_2 \end{aligned}$$

and we want to test $H_0 : \mu_V = \mu_U$.

Write down the usual test statistic under the assumption that $\sigma_U = \sigma_V$, state its distribution, and provide equations so that the test statistic is completely defined in terms of u_i, v_i, and n_1, n_2.

Reformulate the problem as a linear regression model

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

by giving the form of $\boldsymbol{y}, \boldsymbol{X}$ and $\boldsymbol{\beta}$ in terms of μ_U, μ_V and u_i, v_i.

Show that the ordinary least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} * \\ \overline{u} - \overline{v} \end{pmatrix},$$

for some suitable $*$.

Explain how you would use this linear model to test the null hypothesis above.

Show the equivalence of this method (the test statistic, and the distribution) with the usual two-sample t-test.

(b) Define the coefficient of multiple determination R^2, and briefly describe its role and limitations as a multiple regression diagnostic aid.

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.

5. The following computer output concerns data on results of a survey to investigate teenage gambling in Britain. The variables in the data frame `teengamb` are:

sex	0=male, 1=female
status	Socioeconomic status score based on parents' occupation
income	in pounds per week
verbal	verbal score in words out of 12 correctly defined
gamble	expenditure on gambling in pounds per year

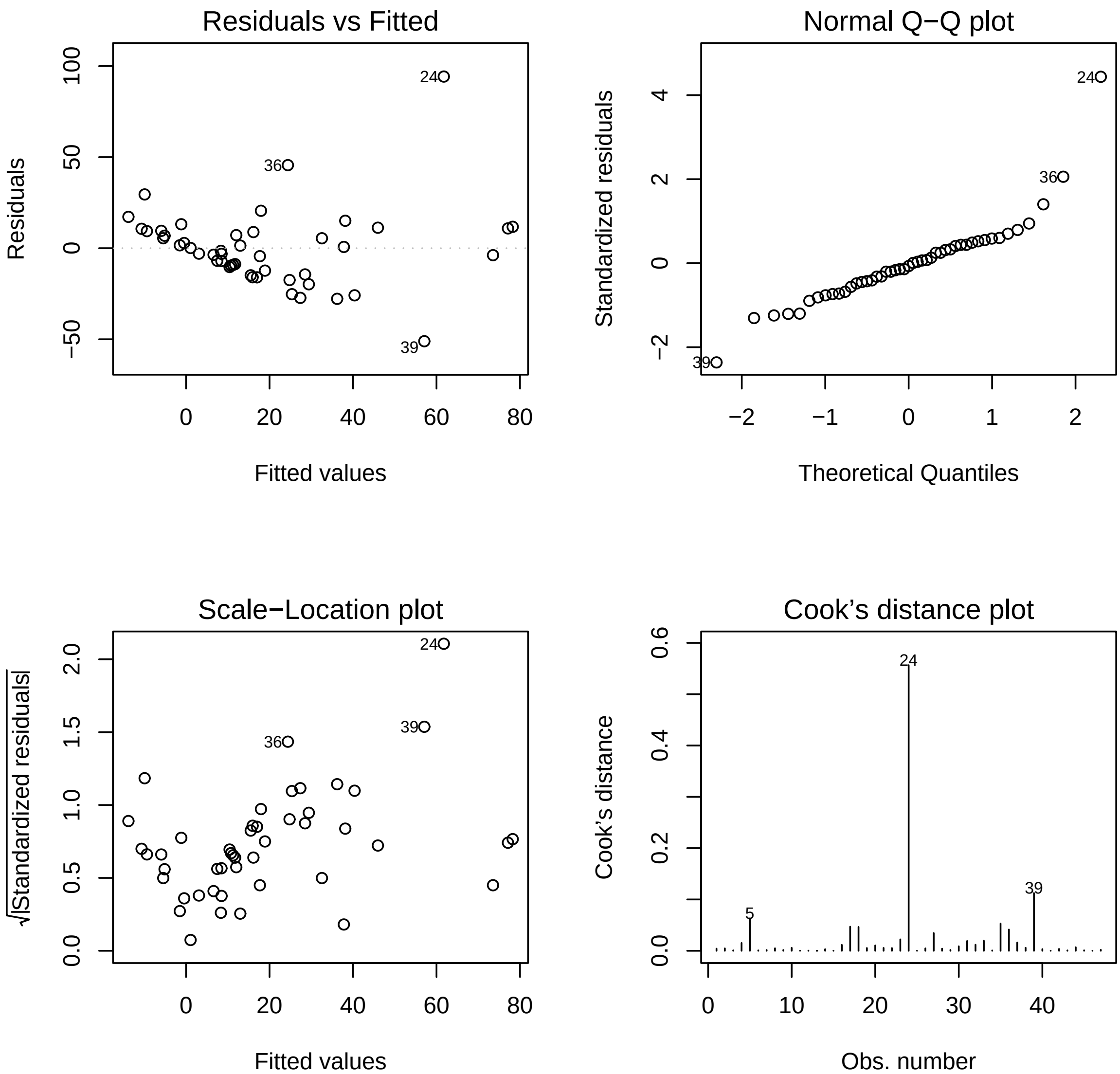
```
> lm1=lm(gamble ~ ., data=teengamb)
> summary(aov(lm1))
            Df  Sum Sq Mean Sq F value    Pr(>F)
sex          1  7598.4  7598.4 14.7584 0.0004066 ***
status       1  3613.0  3613.0  7.0175 0.0113254 *
income       1 11898.6 11898.6 23.1108 1.985e-05 ***
verbal       1   955.7   955.7    A    0.1803109
Residuals    B 21623.8     C
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(lm1)

Call:
lm(formula = gamble ~ ., data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-51.082 -11.320  -1.451   9.452  94.252

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.55565   17.19680     D       E
sex         -22.11833    8.21111  -2.694   0.0101 *
status        0.05223    0.28111   0.186   0.8535
income        4.96198    1.02539   4.839 1.79e-05 ***
verbal       -2.95949    2.17215  -1.362   0.1803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error:   F  on 42 degrees of freedom
Multiple R-Squared: 0.5267,  Adjusted R-squared: G
F-statistic: H on 4 and 42 DF,  p-value: 1.815e-06
```


QUESTION 5 CONTINUED...

Calculate the missing values at A–H.

The graphs below 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. In addition, explain why it is necessary to *standardize* the (raw) residuals.

State what further steps you would take in the analysis of these data, and briefly justify your choice.



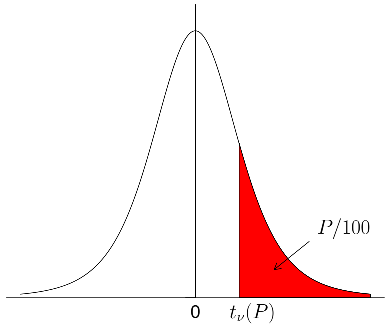
 CONTINUED...

Percentage Points of the t-Distribution

This table gives the percentage points $t_\nu(P)$ for various values of P and degrees of freedom ν, as indicated by the figure to the right.

The lower percentage points are given by symmetry as $-t_\nu(P)$, and the probability that $|t| \geq t_\nu(P)$ is $2P/100$.

The limiting distribution of t as $\nu \to \infty$ is the normal distribution with zero mean and unit variance.



	Percentage points P						
ν	**10**	**5**	**2.5**	**1**	**0.5**	**0.1**	**0.05**
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

END