A15678W1

SECOND PUBLIC EXAMINATION

Honour School of Mathematics and Statistics Part B: Paper SB1

Honour School of Mathematics Part B: Paper SB1

Honour School of Mathematics and Computer Science Part B: Paper SB1

APPLIED AND COMPUTATIONAL STATISTICS

TRINITY TERM 2017

Saturday 03 June, 14:30–17:00

You may submit answers to as many questions as you wish but only the best three will count for the total mark.

You must start a new booklet for each question which you attempt. Indicate on the front sheet the numbers of the questions attempted. A booklet with the front sheet completed must be handed in even if no question has been attempted.

For this paper:

- Permitted calculator series: Casio fx-83, Casio fx-85, Sharp EL-531.
- New Cambridge Statistical Tables are provided.

Do not turn this page until you are told that you may do so

1. (a) [4 marks] Consider the normal linear model

$$y = X\beta + \epsilon$$

where $y = (y_1, \ldots, y_n)^T$, where X is an $n \times p$ design matrix of rank p, where $\beta = (\beta_1, \ldots, \beta_p)$ is vector of unknown parameters, and where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ with $\epsilon_i \sim N(0, \sigma^2)$ and independent for $i = 1, \ldots, n$.

(i) Give the maximum likelihood estimator $\widehat{\beta}$ in terms of X and y, define the *residual sum of squares*, and state their joint distribution.

(ii) Explain how to test the hypothesis $\beta_i = 0$ against the alternative $\beta_i \neq 0$.

(b) [8 marks] Figure (A) shows lean body mass `LBM` (in kilograms) plotted against `Height` (in centimetres) for 44 athletes. Each athlete plays one of three sports (basketball, rowing, swimming), recorded in the categorical variable `Sport` which has three levels, `BBall` (the baseline level), `Row` and `Swim`.

(i) Give a complete mathematical specification of the normal linear model

$$H_1: \quad \texttt{LBM} \sim \texttt{Height} + \texttt{Sport} + \texttt{Height:Sport}$$

(ii) The residual sum of squares for model H_1 is 598.1. The model

$$H_0: \quad \texttt{LBM} \sim \texttt{Height} + \texttt{Sport}$$

has a residual sum of squares of 610.5. Carry out a test, at significance level 5%, of the null hypothesis H_0 against the alternative H_1.

(iii) Figures (B)–(E) are plots for the fit to model H_0. Comment briefly on each of the four plots in relation to goodness of fit. Identify any outliers and recommend what if any follow-up analysis should be done.

(c) [4 marks] Consider the following output for model H_0.

```
Correlation of Coefficients:
       Intercept Height   Row
Height -1.00
Row    -0.27      0.22
Swim   -0.47      0.44    0.74
```

Coefficient	Estimate	Std. Error
`Intercept` $\widehat{\beta}_1$	−51.83	16.41
`Height` $\widehat{\beta}_2$	0.59	0.09
`Row` $\widehat{\beta}_3$	3.79	1.39
`Swim` $\widehat{\beta}_4$	5.25	1.47

For athletes of the same height:

(i) test the hypothesis that rowers have a larger `LBM` than basketball players

(ii) test whether there is a difference in `LBM` between rowers and swimmers.

(d) [6 marks] Recall the general setup in (a) where X is an $n \times p$ design matrix for p variables. Let Z be an $n \times q$ design matrix for q further variables. Suppose the columns of X and Z are orthogonal so that $Z^T X = 0_{q \times p}$. Suppose we fit the model

$$M_0: \quad y = X\beta + \epsilon$$

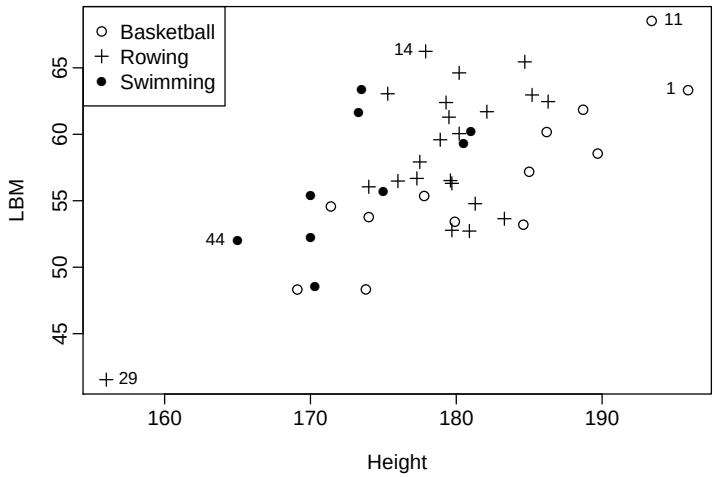
when the true model is

$$M_1: \quad y = X\beta + Z\gamma + \epsilon$$

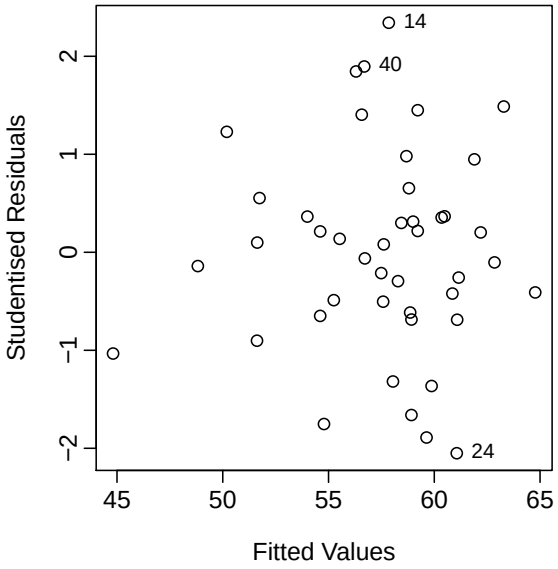
where γ is a q-component parameter vector. Let RSS be the residual sum of squares for the fitted model M_0. Show that

$$E(\text{RSS}) = (n - p)\sigma^2 + \gamma^T Z^T Z \gamma$$

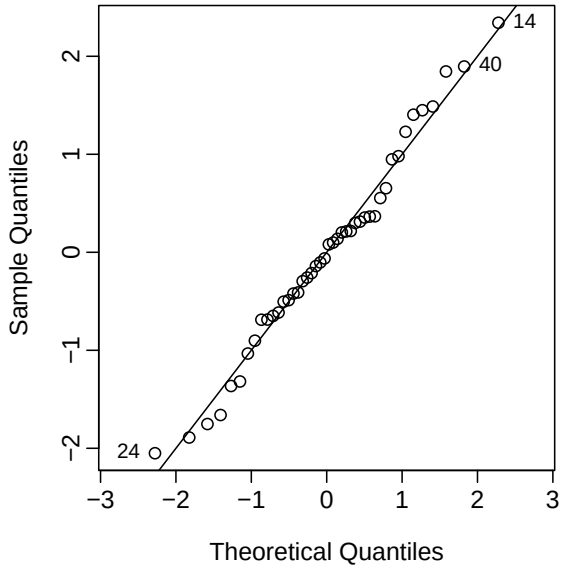
and comment on the implications of this result for model selection.



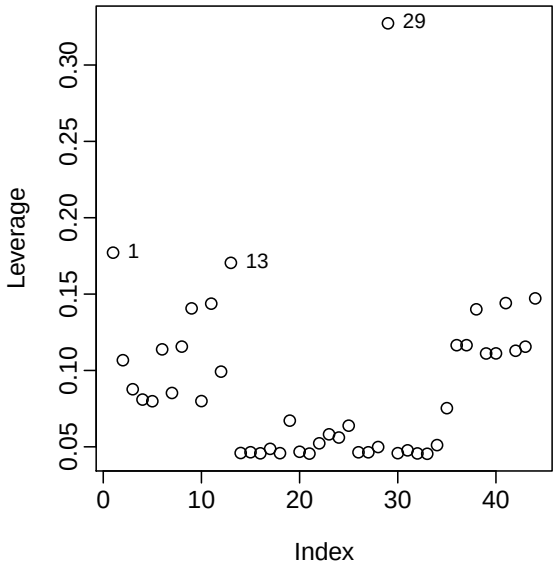
(A) `LBM` plotted against `Height`.



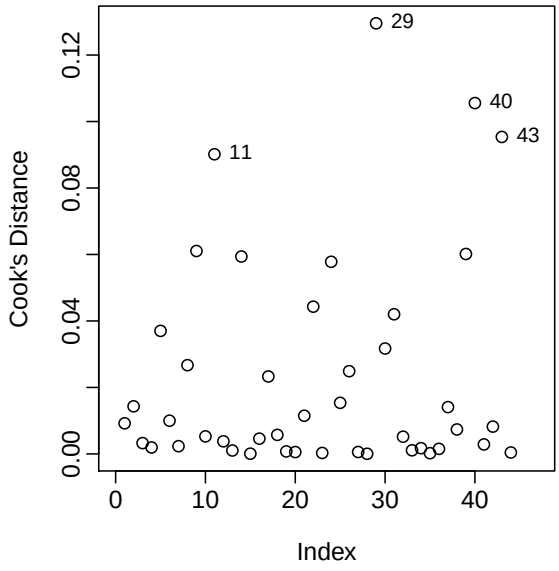
(B) Studentised residuals plotted against fitted values.



(C) Normal quantile-quantile plot of studentised residuals.



(D) Leverage plotted against data index.



(E) Cook's distance plotted against data index.

2. (a) [5 marks] Let $Y_1, \ldots, Y_n$ be discrete random variables distributed according to an exponential family (EF) distribution. State the probability mass function of Y_i in terms of the canonical parameter θ_i, the dispersion parameter ϕ, and functions $\kappa(\theta_i)$ and $c(y_i; \phi)$. Prove that $E(Y_i) = \kappa'(\theta_i)$ and $V(Y_i) = \phi\kappa''(\theta_i)$.

   (b) [2 marks] Let $\phi = 1$, let $\boldsymbol{x}_i$ be a $1 \times p$ vector of explanatory variables, let $\boldsymbol{\beta}$ be a $p \times 1$ vector of parameters, and let $\eta_i = g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, where $\mu_i = E(Y_i)$. What is the relationship between g and κ if g is a canonical link function?

   (c) [3 marks] Show that if Y_i has a Poisson(λ) distribution, then its PDF can be written in the EF form. Determine θ_i and ϕ. Find the mean μ_i, the variance function $V(\mu_i)$, and the canonical link function.

   (d) [12 marks] A study was carried out to assess the effect of newly installed speed cameras on the number of road accidents. The response is the number of accidents, in a given year, at a particular location. Four locations were considered, two with speed cameras (`cameras` $= 1$) and two without (`cameras` $= 0$), and the number of accidents per year was recorded for a 10 year period (`time` $= 0, 1, \ldots, 9$).

      (i) Show how to set up a generalised linear model for this data, specifying the distribution, mean, canonical link function and linear predictor. A Poisson GLM was fit to the data and the following output was obtained:

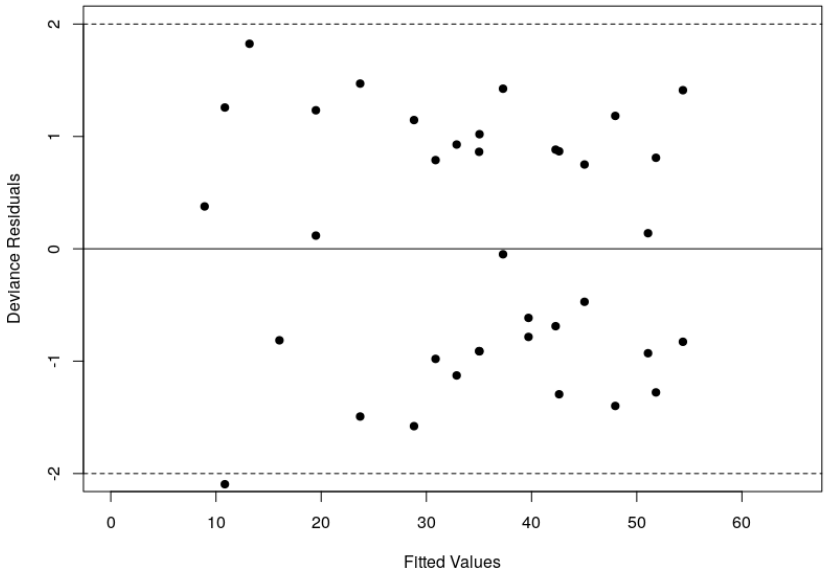
```
  Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.05905    0.06992  58.056  < 2e-16 ***
I(time)         -0.06291    0.01218  -5.164 2.42e-07 ***
cameras          0.08434    0.10720   0.787    0.431
I(time):cameras -0.13262    0.02095  -6.330 2.45e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 311.895  on 39  degrees of freedom
Residual deviance:  57.286  on 36  degrees of freedom
```

      (ii) Calculate the effect of `cameras` on the number of accidents and comment on this.

      (iii) Define a goodness of fit test that can be used given the information above.

A plot of the standardised deviance residuals against the fitted values is given below.



(iv) Define the standardised deviance residuals and comment on the plot with respect to the suitability of the model.

3. (a) [6 marks] Let $X_1, \ldots, X_n$ be independent and identically distributed real-valued random variables with unknown cumulative distribution function (cdf) F.
   (i) Define the empirical cdf F_n of the random sample $(X_1, \ldots, X_n)$.
   (ii) Show that $F_n(x)$ is an unbiased and consistent estimator of $F(x)$ for any $x \in \mathbb{R}$. [*You can use without proof any standard result from Probability.*]

   (b) [5 marks] Let $T_n = t(X_1, \ldots, X_n)$ be some estimator. Describe how to estimate $\mathrm{Var}(T_n)$ using the nonparametric bootstrap.

   (c) [6 marks] Assume that the mean $\mu := \mathbb{E}(X_i)$ is known, and $\mathbb{E}(X_i^4) < \infty$. Consider the estimator

$$T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

   Let T_n^* be a nonparametric bootstrap sample. Show that

$$\begin{aligned}\mathbb{E}_{F_n}(T_n^*) :=& \mathbb{E}(T_n^*|X_1, \ldots, X_n)\\ =& T_n\end{aligned}$$

   and

$$\begin{aligned}\mathrm{Var}_{F_n}(T_n^*) :=& \mathrm{Var}(T_n^*|X_1, \ldots, X_n)\\ =& \frac{\sum_{i=1}^{n}(X_i - \mu)^4}{n^2} - \frac{T_n^2}{n}.\end{aligned}$$

   (d) [5 marks] Deduce that the bootstrap estimators $\mathbb{E}_{F_n}(T_n^*)$ and $\mathrm{Var}_{F_n}(T_n^*)$ are consistent estimators of $\mathbb{E}(T_n)$ and $\mathrm{Var}(T_n)$ respectively.

4. Let $Z = (X, Y)$ be the combined sample of X_i, $i = 1 \ldots, n$, and Y_j, $j = 1, \ldots, m$. Assume that $X \sim F$ and $Y - \Delta \sim F$ for some arbitrary distribution F where Δ is the location shift.

   (a) [4 marks] Explain carefully how to test $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$ using the Wilcoxon test statistic W.

   (b) [4 marks] The table below shows the systolic blood pressure (mm Hg) of 12 patients, where 6 are in a treatment group and 6 are in a control group.

Control	122	130	117	132	114	128
Treated	120	118	123	102	98	125

   (i) Calculate the Wilcoxon rank statistic based on the difference of blood pressure between the treated and control patients.

   (ii) Test the hypothesis that the treatment has no effect on blood pressure using the normal approximation of the test statistic where:

$$\begin{aligned} E(W) &= \frac{m(n+m+1)}{2} \\ V(W) &= \frac{nm(n+m+1)}{12}. \end{aligned}$$

   (c) [10 marks] Consider the Lehmann-Hodges estimator

   (i) Define the Lehmann-Hodges estimator for a general test.

   (ii) Derive the Lehmann-Hodges estimator $\widehat{\Delta}$ for the Wilcoxon test using the normal approximation.

   (iii) Show that this estimator can also be written as:

$$\widehat{\Delta} = \text{median}\{Y_j - X_i, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}.$$

   (d) [4 marks] Consider the following output from R:

```
> wilcox.test(x,y,exact=T,conf.int = T)

Wilcoxon rank sum test

data:  x and y
W = 26, p-value = 0.2403
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -4 26
sample estimates:
difference in location
                   8.5
```

   (i) Explain why the p-value obtained here is different from what you would have got in your analysis in part (b).

   (ii) What changes would you need to make to the command line to get the same results as in part (b)?

   (iii) Explain fully how one might calculate the 95% confidence interval for $\widehat{\Delta}$.