THE UNIVERSITY OF AUCKLAND

EXAMINATION FOR BA BSc ETC 2000

STATISTICS

Advanced Statistical Modelling
Topics in Statistics C

(Time allowed: THREE hours)

NOTE: Attempt all FIVE questions. Each question is worth 20 marks.

1. Short answer questions.

(a) For each of the following model deficiencies identify a diagnostic plot that can be used to detect the deficiency and explain what pattern in the plot indicates the deficiency is present.

(i) A non-linear relationship between the response and the regressors.

(ii) The response has a non-Normal distribution.

(iii) Serial correlation between the observations.

(b) Explain what is meant by multicollinearity and how it can make selecting a model more difficult. How are variance inflation factors (VIF's) used to detect multicollinearity in a dataset?

(c) Consider a medical study that investigates the relationship between smoking and the occurrence of a particular disease. How would a "case-control" study be different from a "cohort" study? How does this affect the analysis of the data?

(d) The following table summarises the relationship between tonsil size and whether a child is a carrier of a certain virus (*Streptococcus pyogenes*).

	Tonsil Size		
	Not enlarged	Enlarged	Greatly enlarged
Carriers	19	29	24
Non-carriers	497	560	269

Calculate a 95% confidence interval for the odds ratio that compares the odds that a carrier has "Enlarged" or "Greatly enlarged" tonsils to the odds for a non-carrier. Note that if $Z \sim \mathrm{N}(0,1)$, then $\Pr(-1.96 \leq Z \leq 1.96) = 0.95$.

CONTINUED

2. The Scots are a strange race as evidenced by bagpipes, caber-tossing, haggis, and their passion for running up and down hills wearing kilts. Data were collected on the record winning times (`Time`), in minutes, for 35 hill races in Scotland. The distance travelled (`Distance`) and height climbed (`Climb`), both measured in kilometres, in each race were also recorded.

The data was used to fit the regression model:

$$\mathrm{E}(\text{Time}) = \beta_0 + \beta_1\ \text{Climb} + \beta_2\ \text{Distance} + \beta_3\ \text{Climb} \times \text{Distance}.$$

```
                Value Std. Error  t value Pr(>|t|)
  (Intercept) -0.3532   3.9122    -0.0903   0.9286
        Climb 29.7512  20.0096     1.4868   0.1472
     Distance  7.8864   0.7600    10.3765   0.0000
Climb:Distance  9.0975   2.3596     3.8555   0.0005

Residual standard error: 7.35 on 31 degrees of freedom
Multiple R-Squared: 0.9806
F-statistic: 521.1 on 3 and 31 degrees of freedom, the p-value is 0
```

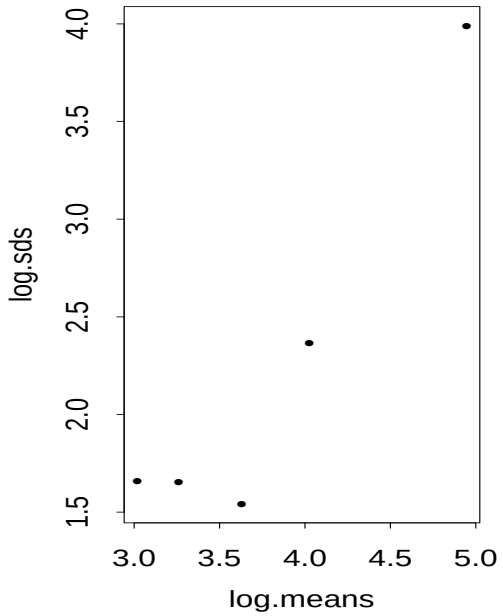
(a) Consider the following statement:

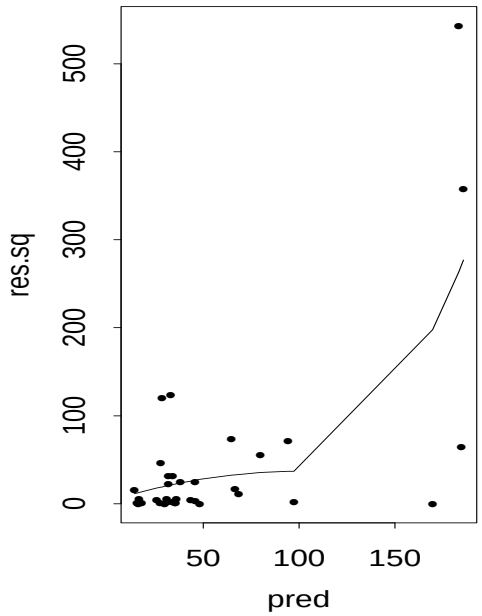
> This model estimates that the record winning time for a race should increase by 7.89 minutes, on average, for each increase of 1 kilometre in distance travelled provided that the height climbed remains the same.

Do you agree with this statement? Explain why or why not.

(b) The following diagnostic plots were obtained using the `funnel` command:

```
> funnel(race.lm1)
Slope: 1.27563894849827
```





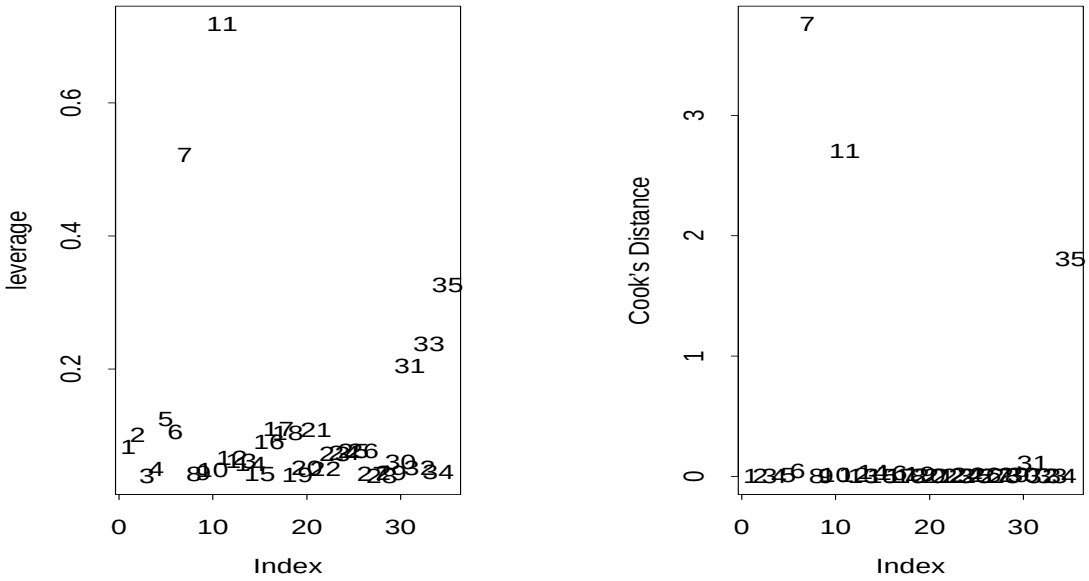
These plots indicate that a problem exists with the regression model.

(i) State the nature of the problem and explain how these plots indicate that problem is present.

(ii) One method of dealing with this problem is to transform the response. Based on the output, what type of transformation would you try? Explain your answer.

(iii) What other aspects of the regression model would be affected by a transformation of the response? What other remedies are available for the problem you identified in (i)?

(c) Consider the index plots of Leverage (h_{ii}'s) and Cook's Distance for the fitted model.



(i) Explain what a large value of leverage (h_{ii}) indicates about an observation and what a large value of Cook's distance indicates. In your answer make sure you clearly explain the difference in the information provided by these two diagnostics.

(ii) What do you learn from these two plots? Briefly explain what action (if any) you would take next.

(d) The leave one out diagnostics for observation 7 generated by influence.measures are:

	.Intercept.	Climb	Distance	Climb.Distance	dffits	cov.ratio	cooks.d	hats
7	1.040	-0.277	-2.511	2.250	5.126	0.224	3.758	0.521

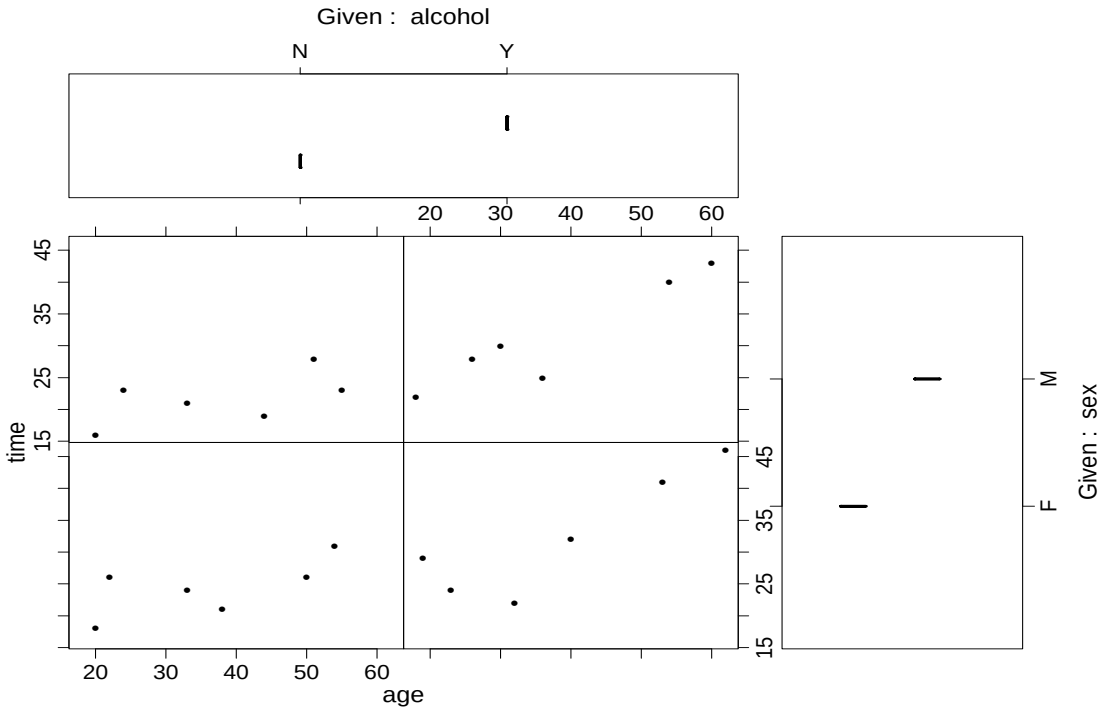
(i) What does the `dffits` value of 5.126 indicate?

(ii) What does the value for `Distance`, -2.511, indicate?

3. An experiment was conducted to investigate the effect of alcohol consumption on the times taken (in seconds) to complete a simple task. The age and gender of the participants were recorded because it was believed that these might affect the times. The data, a coplot, and some output from a regression analysis are given below.

	Alcohol Consumption			
	No		Yes	
	Time	Age	Time	Age
Male	16	20	22	18
	23	24	28	26
	21	33	30	30
	19	44	25	36
	28	51	40	54
	23	55	43	60

	Alcohol Consumption			
	No		Yes	
	Time	Age	Time	Age
Female	18	20	29	19
	26	22	24	23
	24	33	32	40
	21	38	22	32
	26	50	41	53
	31	54	46	62




```
Coefficients:
                  Value Std. Error  t value Pr(>|t|)
    (Intercept)  16.4133   4.8211     3.4045   0.0036
            sex  -1.1442   6.8745    -0.1664   0.8699
        alcohol  -3.0095   6.4582    -0.4660   0.6475
            age   0.2190   0.1257     1.7428   0.1005
    sex:alcohol   1.2187   9.1857     0.1327   0.8961
        sex:age  -0.0499   0.1754    -0.2845   0.7797
    alcohol:age   0.2770   0.1634     1.6956   0.1093
sex:alcohol:age   0.0322   0.2306     0.1395   0.8908

Residual standard error: 3.943 on 16 degrees of freedom
Multiple R-Squared: 0.8317
F-statistic: 11.29 on 7 and 16 degrees of freedom, the p-value is 3.822e-05
```

(a) What information can be gained from each of the following lines of the output.

(i) `Multiple R-Squared: 0.8317`

(ii) `F-statistic: 11.29 on 7 and 16 degrees of ....`

(b) The output suggests that a simpler model may be adequate. What aspect(s) of the output indicates this? Briefly, explain how you would identify a simpler model in this situation?

(c) The person conducting this experiment decided to use a model that contained `age`, `alcohol`, and the `alcohol:age` interaction.

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  16.1171   3.2440     4.9684   0.0001
    alcohol  -2.7019   4.3372    -0.6230   0.5404
        age   0.1860   0.0827     2.2492   0.0359
alcohol:age   0.3019   0.1088     2.7738   0.0117

Residual standard error: 3.728 on 20 degrees of freedom
Multiple R-Squared: 0.8119
F-statistic: 28.77 on 3 and 20 degrees of freedom, the p-value is 1.87e-07

> dummy.coef(alcohol.fit2)
$"(Intercept)":
 (Intercept)
    16.11713

$alcohol:
 NO     YES
 0 -2.701907

$age:
       age
 0.1860236

$"alcohol:age":
 NOage   YESage
   0  0.3018734
```

Assuming this model is adequate, what does it indicate about the way time depends on age, gender, and alcohol consumption? Your answer should consist of a short paragraph and should pay particular attention to the effect of alcohol consumption.

(d) Do you think the model given in (c) is reasonable? Provide support for your response using the coplot of the data and relevant parts of the computer output.

4. The Los Angeles Heart Study, supervised by John Chapman, collected information regarding risk factors associated with coronary heart disease. Part of these data are measurements made on 200 men of the following variables:

Age in years (`age`).

Systolic blood pressure in mm of Hg (`sbp`).

Diastolic blood pressure in mm of Hg (`dbp`).

Cholesterol in mg per dl (`cho`).

Height in inches (`ht`).

Weight in pounds (`wt`).

Coronary incident in the last 10 years (`cor`= 1 if an incident had occurred; `cor`= 0 otherwise)

A logistic regression model that contains all these regressors gives the following Splus output:

```
                  Value  Std. Error    t value
(Intercept) -4.517319021 7.479489249 -0.6039609
        age  0.045899976 0.023529058  1.9507783
        sbp  0.006855721 0.020194525  0.3394842
        dbp -0.006936751 0.038343820 -0.1809092
        cho  0.006306448 0.003631292  1.7366953
         ht -0.074001542 0.106189623 -0.6968811
         wt  0.020141537 0.009868974  2.0408948

    Null Deviance: 154.5547 on 199 degrees of freedom
Residual Deviance: 134.8515 on 193 degrees of freedom
```

The step.glm procedure in Splus identifies the model that contains `age`, `cho`, and `wt` as having the lowest value of AIC. The output for this model is:

```
                  Value  Std. Error   t value
(Intercept) -9.255888319 2.070916857 -4.469464
        age  0.053003641 0.020821917  2.545570
        cho  0.006517925 0.003587892  1.816645
         wt  0.017538629 0.008270296  2.120677

    Null Deviance: 154.5547 on 199 degrees of freedom
Residual Deviance: 135.5233 on 196 degrees of freedom
```

(a) A χ^2 test can be used to evaluate whether or not it is sensible to drop `sbp`, `dbp`, and `ht` from the model. The P-value for this test is 0.88.

(i) Write down the null hypothesis, H_0, for this test.

(ii) Find the value for the test statistic, χ^2_o.

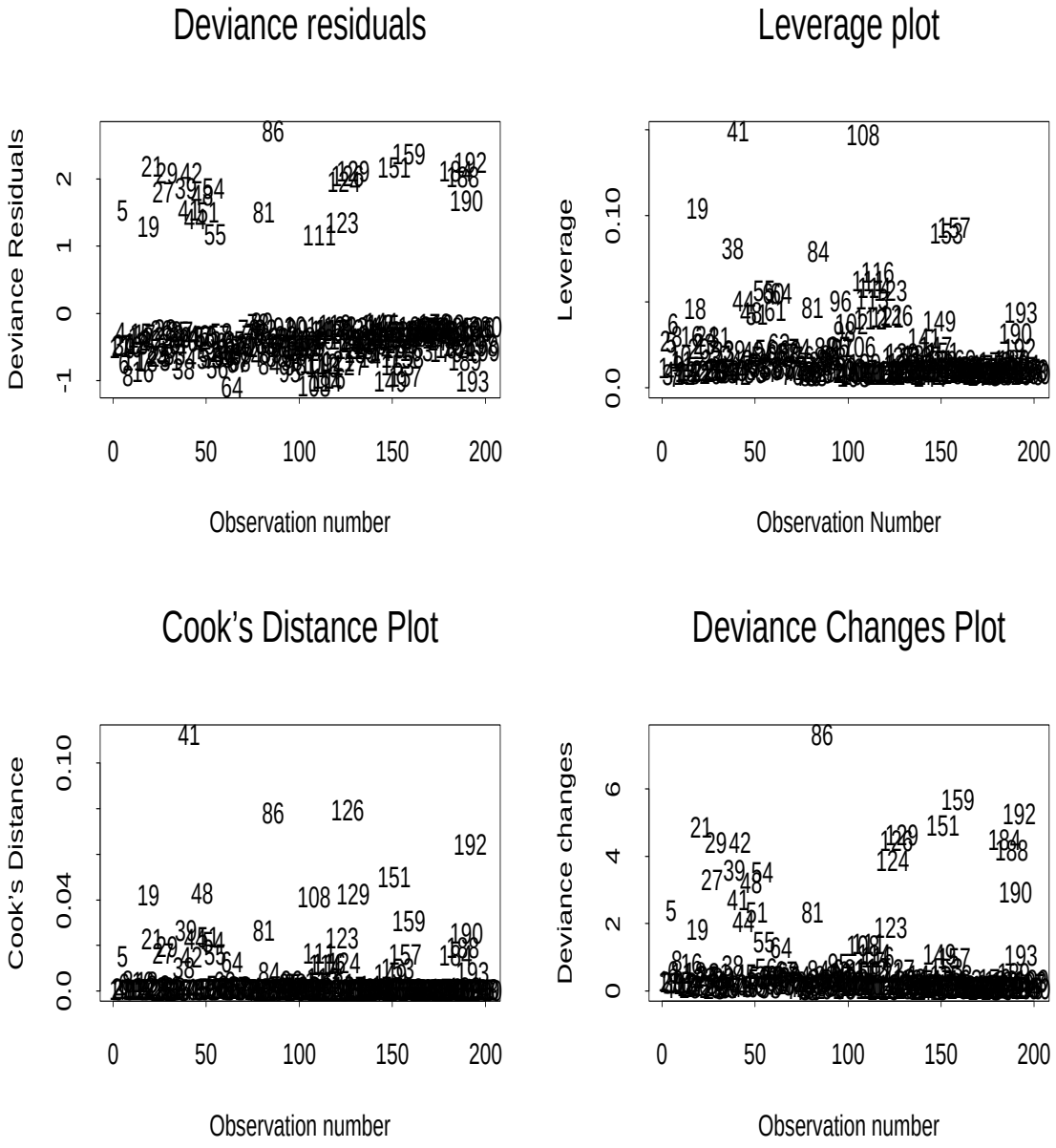
(iii) Write down the formula used to calculate the P-value.

(iv) What does the observed P-value of 0.88 indicate about dropping `sbp`, `dbp`, and `ht` from the model?

(b) Use the reduced model (the one that just contains `age`, `cho`, and `wt`) to explain the effect of age on the odds of a coronary incident. Provide an interval estimate in addition to a point estimate.

(c) Use the reduced model to estimate the probability of a coronary incident for a 40 year old man who weighs 180 pounds and has a cholesterol level of 300 mg per dl.

(d) The following set of diagnostic plots was produced for the reduced model.



Explain what problem(s) each of these plots is used to detect. Does this set of plots indicate any problems with the reduced model?

5. The following contingency table contains consumer preference data collected by the manufacturers of a particular laundry detergent (Brand M). A sample of 1008 consumers was cross classified according to:

the softness of laundry water used (`water`),

whether the consumer was a previous user of Brand M (`user`),

the temperature of laundry water used (`temp`), and

whether the consumer preferred Brand M or Brand X (a competitor) in a blind trial (`pref`).

		Previous user of M		Previous non-user of M	
Water softness	Brand preference	High temperature	Low temperature	High temperature	Low temperature
soft	X	19	57	29	63
	M	29	49	27	53
medium	X	23	47	33	66
	M	47	55	23	50
hard	X	24	37	42	68
	M	43	52	30	42

The step.glm function in Splus selected the following log-linear model:

```
Terms added sequentially (first to last)
           Df Deviance Resid. Df Resid. Dev   Pr(Chi)
      NULL                    23   118.6269
     water  2  0.50148        21   118.1255 0.7782248
      pref  1  0.06349        20   118.0620 0.8010584
      user  1  1.92125        19   116.1407 0.1657194
      temp  1 73.21206        18    42.9287 0.0000000
 pref:user  1 20.58147        17    22.3472 0.0000057
 pref:temp  1  4.36160        16    17.9856 0.0367577
water:temp  2  6.09910        14    11.8865 0.0473801
```

(a) Produce the association graph for this model. Using your association graph, give an example of each of the following types of relationships or state that none is present.

(i) Two factors that are independent.

(ii) Two factors that are directly related.

(iii) Two factors that are conditionally independent.

(b) Explain what is meant by conditional independence. If you identified an example of conditional independence in (a), use that example to illustrate your explanation, otherwise make up an example.

(c) Would it be sensible to collapse the 4-way table to:

(i) a 2-way table involving `pref` and `user`?

(ii) a 2-way table involving `pref` and `water`?

In each case if collapsing to the 2-way table is not sensible, indicate how you would investigate the relationship between those two factors.

(d) These data could be analysed using logistic regression to focus on how the results of the blind preference trial were related to the other factors.

(i) Explain how you would analyse the data using logistic regression. (What probability is being modelled? What is used as the response? What are the explanatory variables? ...)

(ii) Given the output from the Poisson regression, what regressors would you expect to be needed in the logistic regression model? Explain.