THE UNIVERSITY OF AUCKLAND

EXAMINATION FOR BA BSc ETC 1999

STATISTICS

Advanced Statistical Modelling
Topic in Statistics C

(Time allowed: THREE hours)

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

1. (a) Explain what is meant when an observation is described as having "high leverage". Consider a simple linear regression model, $Y = \beta_0 + \beta_1 X + \epsilon$. Draw simple scatter plots that clearly illustrate:

(i) an observation that has high leverage but is not influential,

(ii) an observation that has high leverage and is influential. (4 marks)

(b) Consider a multiple regression model, $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \epsilon$. Why is Mallow's Cp statistic more useful than multiple R-squared (R^2) in selecting a subset model? (4 marks)

(c) Leave one out diagnostics are used to assess the influence that an observation has on the fitted regression model. Explain what the following occurrences would indicate:

(i) one observation has a much larger value of Cook's distance than any other observation,

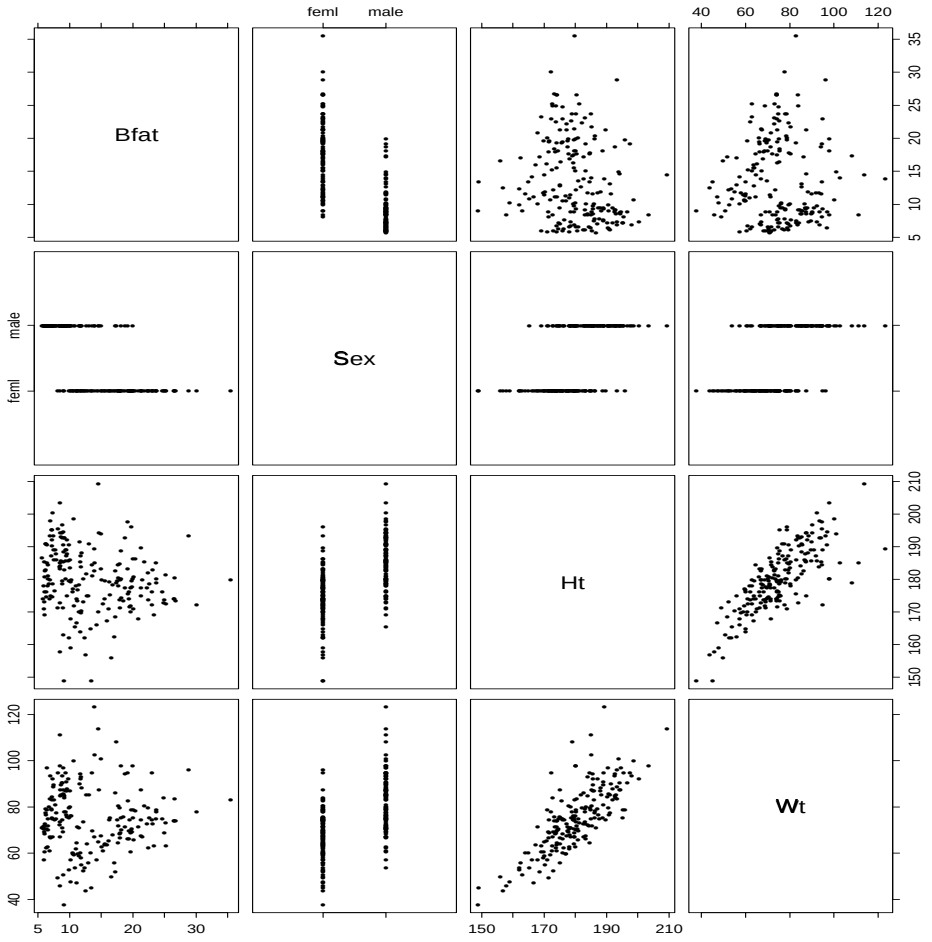
(ii) an observation has a value of DFFITS of 1.32. (4 marks)

(d) For what purpose is a constructed variable plot useful? Briefly explain how such a plot is used. Note: you are being asked to explain how to interpret a constructed variable plot, **not** how to create a constructed variable plot. (4 marks)

(e) What does the "dissimilarity index" measure? Describe a situation where it can be useful. (4 marks)

CONTINUED

2. Data consisting of measurements on 102 male and 100 female athletes was collected by the Australian Institute of Sport. Part of this data was used to fit a regression model that relates percent body fat (Bfat) to three explanatory variables. The explanatory variables used were Sex (male or female), Ht (height in cm), and Wt (weight in kg).



The following model was obtained using S-plus:

```
> sport.fit1<-lm(log(Bfat) ~ Sex + Ht + Wt + Sex:Ht, data = sport.df2)
> summary(sport.fit1,correlation=F)

               Value Std. Error  t value Pr(>|t|)
(Intercept)   1.6741   0.5074     3.2996   0.0011
        Sex   1.3026   0.7054     1.8466   0.0663
         Ht  -0.0017   0.0032    -0.5144   0.6075
         Wt   0.0215   0.0018    11.6319   0.0000
     Sex:Ht  -0.0122   0.0039    -3.1274   0.0020

Residual standard error: 0.2228 on 197 degrees of freedom
Multiple R-Squared: 0.7621
F-statistic: 157.7 on 4 and 197 degrees of freedom, the p-value is 0
```

```
> dummy.coef(sport.fit1)
$"(Intercept)":
 (Intercept)
    1.674104

$Sex:
 female     male
      0 1.302641

$Ht:
           Ht
 -0.001658649

$Wt:
         Wt
 0.02151617

$"Sex:Ht":
 femaleHt      maleHt
        0 -0.01222963
```

(a) Briefly summarise what you can learn about the data from the pairs plot. (4 marks)

(b) Explain clearly what each of the following lines from the S-plus output indicates

(i) `Multiple R-Squared:  0.7621`

(ii) `F-statistic:  157.7 on 4 and 197 degrees of ....` (3 marks)

(c) Does the output indicate the fitted model can/should be simplified? Explain. (4 marks)

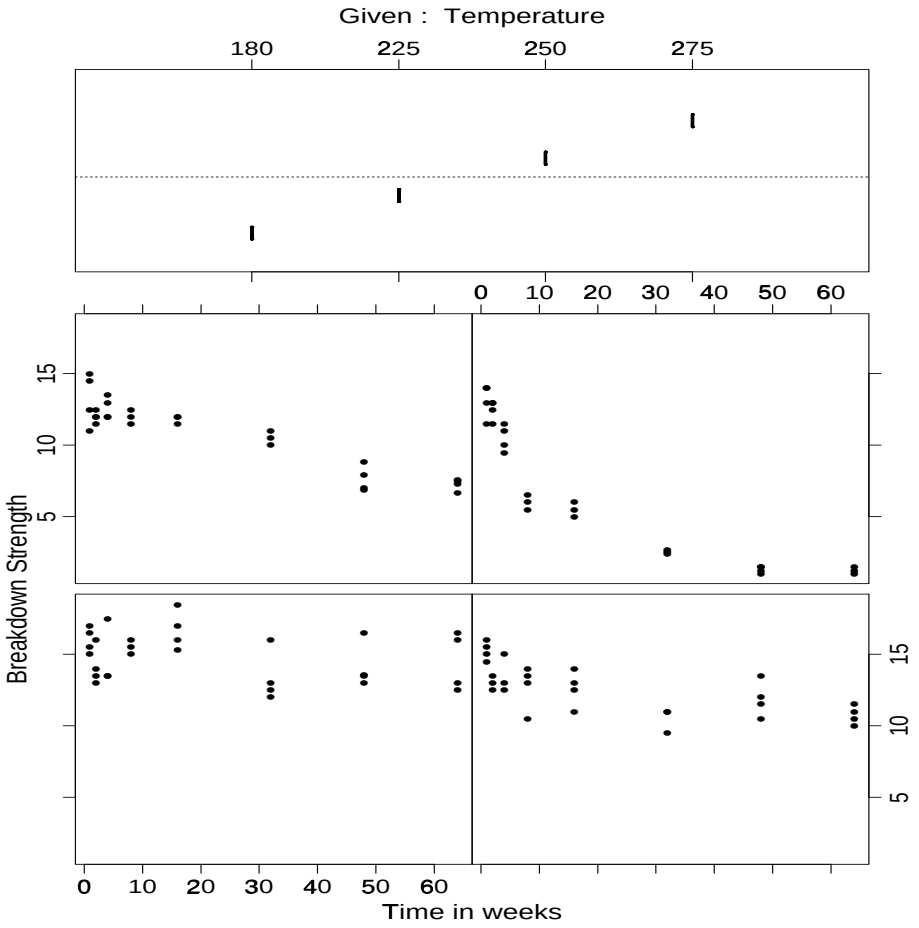
(d) Explain how the explanatory variables are related to the percent body fat of the athletes using the fitted model. Specifically comment on differences between female and male athletes. (5 marks)

(e) Suggest a plot(s) that you could use in a report to help commumicate your response to part (d). That is, you want a plot that will clarify the key relationships between percent body fat and the regressors. Justify your choice. (4 marks)

3. Part 1: Electrical Insulation Data.

The deterioration of samples of electrical insulation over time was studied. The samples of insulation were stored at elevated temperatures for varying periods of time before their insulating ability was measured.

The response variable is dielectric breakdown strength (kilo-volts), and the predictor variables are storage time (weeks) and temperature (°C). Plots of breakdown strength versus storage time for each of the four temperatures used in the study are given below.



(a) Summarise what you can learn about the way breakdown strength is related to storage time and temperature from these plots. (5 marks)

(b) Suppose that the person in charge of this study has decided to fit a regression model that has breakdown strength (strength) as the response and uses time (time) and temperature (temp) as numeric regressors. The S-plus code for the model she plans to fit is:

```
> model<-lm(strength~time + temp)
```

Given the trellis plot above, do you anticipate any problems with this model? If so, identify the features of the trellis plot that indicate this model may not be adequate and suggest possible remedies. If not, explain why this model is compatible with the trellis plot. (5 marks)

Part 2: Dolphin Activity Data.
Groups of dolphins were observed off the coast of Iceland near Keflavik in 1998. When a group was observed its main activity (travelling quickly, feeding, or socialising) and the time of day were recorded. The counts are given by the following table:

	Activity		
Period	Travel	Feed	Social
Morning	6	28	38
Afternoon	20	4	14
Evening	13	56	10

(c) A generalised linear model was fitted to the data.

```
> Activity<-rep(c("Travel","Feed","Social"),3)
> Period<-rep(c("Morning","Afternoon","Evening"),c(3,3,3))
> Groups<-c(6,28,38,20,4,14,13,56,10)
> dolphin.df<-data.frame(Activity,Period,Groups)
> dolphin.fit1<-glm(Groups ~ .^2, family = poisson, data = dolphin.df)
> anova(dolphin.fit1,test="Chi")
Terms added sequentially (first to last)
                Df Deviance Resid. Df Resid. Dev      Pr(Chi)
            NULL                    8   102.1247
       Activity  2 19.42882          6    82.6958 0.0000604066
         Period  2 16.56430          4    66.1315 0.0002529923
Activity:Period  4 66.13153          0     0.0000 0.0000000000
```

Explain why this analysis indicates that there is strong evidence that the behaviour pattern of dolphins is different at different times of the day? (2 marks)

(d) Estimate the odds that a group of dolphins observed in the morning will be socialising using the data from the contingency table. (2 marks)

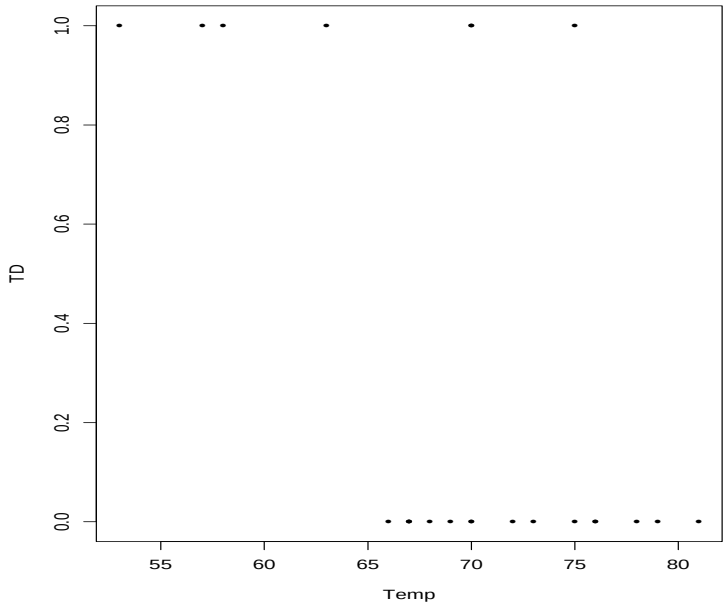
(e) Estimate the odds ratio that compares the odds that a group of dolphins observed in the morning will be socialising to the odds that a group of dolphins observed in the evening will be socialising. (2 marks)

(f) Find a 95% confidence interval for the odds ratio that you estimated in part (c). Note that if $Z \sim \mathrm{N}(0,1)$, then $\Pr(-1.96 \leq Z \leq 1.96) = 0.95$. (4 marks)

4. On 29 January, 1986 the Space Shuttle Challenger was destroyed by a massive explosion. The cause of the explosion was traced to the failure of a rubber seal called an O-ring. The temperature at the time of launch was 31°F. After the accident NASA's failure to anticipate O-ring problems at such low temperatures was criticised.

For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the following table shows the temperature in °F (Temp) and whether at least one primary O-ring suffered thermal distress (TD= 1 yes, TD= 0 no).

Flight	Temp	TD	Flight	Temp	TD	Flight	Temp	TD
1	66	0	9	57	1	17	70	0
2	70	1	10	63	1	18	81	0
3	69	0	11	70	1	19	76	0
4	68	0	12	78	0	20	79	0
5	67	0	13	67	0	21	75	1
6	72	0	14	53	1	22	76	0
7	73	0	15	67	0	23	58	1
8	70	0	16	75	0			



A logistic regression model was constructed to relate the probability of thermal distress to temperature using S-plus:

```
> challenger.fit<-glm(TD~Temp,family=binomial,data=challenger.df)
> summary(challenger.fit)
Coefficients:
                 Value Std. Error   t value
(Intercept)  15.0422911  7.3366528  2.050293
       Temp  -0.2321537  0.1076141 -2.157279

    Null Deviance: 28.26715 on 22 degrees of freedom
Residual Deviance: 20.31519 on 21 degrees of freedom
```

(a) Explain how you would perform a Chi-square test to test the hypothesis that the coefficient for `temp` is 0. Explain what you would use as a test statistic and how you would calculate the P-value (you will not be able to calculate the actual P-value – just explain how you would do it). (3 marks)

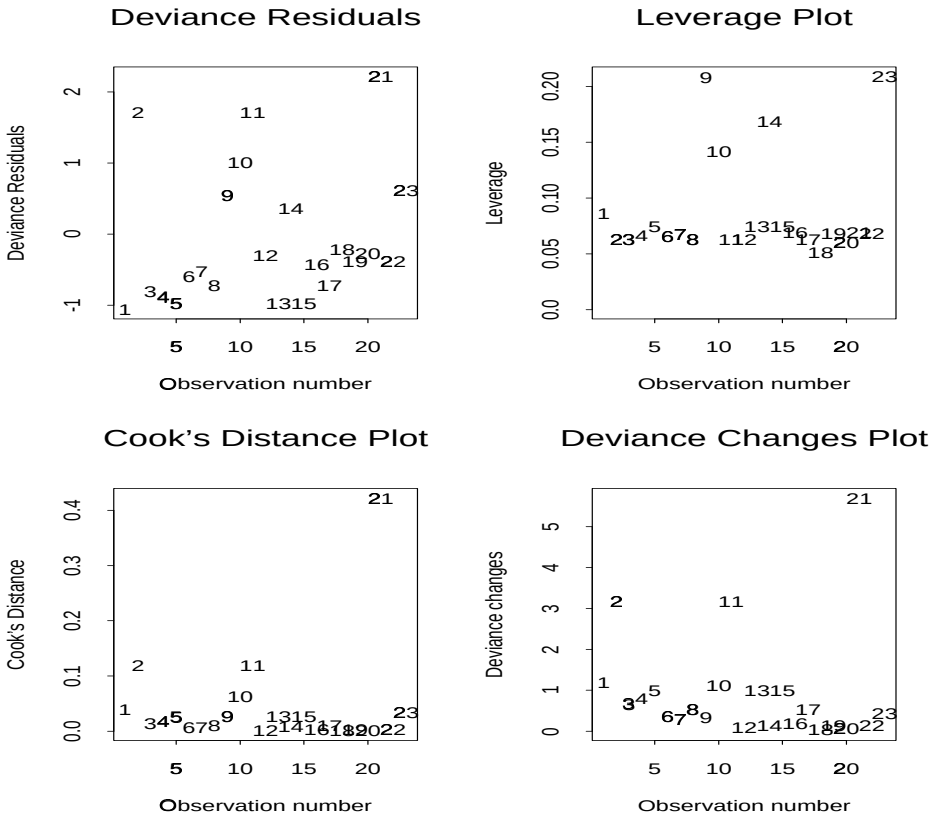
(b) Let π = probability of thermal distress. Write down the fitted model in
(i) the logit form
(ii) the logistic form (3 marks)

(c) Use the fitted model to explain the effect of temperature on
(i) logit(π)
(ii) the odds of thermal distress (4 marks)

(d) Use the fitted model to predict the probability of thermal distress at 31°F (the temperature for the Challenger launch). Do you have any reservations about the validity of this prediction? (3 marks)

(e) At what temperature would the fitted model predict that the probability of thermal distress would be 90%? (4 marks)

(f) Do the following diagnostic plots indicate any problems with the fitted model? If a problem is indicated briefly explain what action should be taken. (3 marks)



5. A study of possible risk factors for coronary heart disease was conducted at a car factory in Czechoslovakia. Data was collected on all 1841 men employed at the factory. A 6-way contingency table was constructed that cross-classified the data according to the following factors:

A: Smoking (yes, no)

B: Strenuous mental work (yes, no)

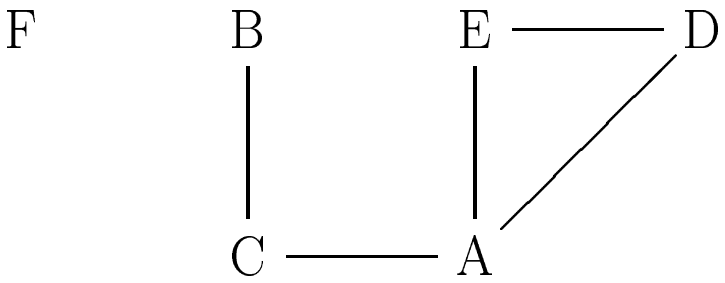
C: Strenuous physical work (yes, no)

D: Systolic blood pressure (< 140, > 140)

E: Ratio of beta and alpha lipoproteins (< 3, > 3)

F: Family history of heart disease (yes, no)

A initial analysis of the data produced the following association graph:



(a) Explain the procedure you would use to produce an association graph such as this one. Assume the data is in an S-plus data frame and outline how you would analyse it. Note: you do not need to give S-plus code - just indicate what model(s) you would fit and how you would use the results to produce the association graph. (5 marks)

(b) From the association graph give an example of:

(i) two factors that are independent.

(ii) two factors that are conditionally independent.

(iii) two factors that are directly related. (3 marks)

(c) Clearly explain what conditional independence means and how it is different from independence. Use your example of conditional independence from part (b) to illustrate your explanation. (4 marks)

(d) In order to investigate the relationship between "Strenuous mental work" (B) and "Strenuous physical work" (C) would it be sensible to collapse the table on all other factors and just examine the 2-way table involving B and C? Explain. (4 marks)

(e) Suppose that we decide to focus on men who were smokers. Use the association graph to determine how the remaining factors (B, C, D, E, and F) are related to each other if we restrict our attention to smokers. (4 marks)