THE UNIVERSITY OF AUCKLAND **475.330 SC** **475.794 SC**

EXAMINATION FOR BA BSc ETC 1998

STATISTICS

Advanced Statistical Modelling
Topic in Statistics C

(Time allowed: THREE hours)

NOTE: Attempt all SIX questions.

1. (a) List the assumptions that are made for the ordinary (Normal) regression model. (2 marks)

(b) List the 3 aspects of the ordinary regression model that are affected by transforming the response. (2 marks)

(c) What problem can be remedied by using weighted least squares to estimate the regression coefficients? What should the weights be proportional to? (2 marks)

(d) What is a partial regression plot used for? Briefly, describe how a partial regression plot is constructed (it is not necessary to give S-plus commands). (3 marks)

(e) One diagnostic procedure covered in this course involved plotting L versus p, where

$$L = \frac{n}{2} \log RSS^{(p)} + (1-p) \sum_{i=1}^{n} \log y_i \ .$$

What do p and $RSS^{(p)}$ stand for in the definition of L? What is this plot used for? (3 marks)

2. An experiment was conducted to assess the response of flour beetles to gaseous CS_2 (carbon disulfide). In the experiment CS_2 was added to flasks in which cloth cages containing about 30 beetles (each) were suspended. The number of dead beetles (y) out of the total number (n) was recorded for each flask. The entire experiment was replicated using the same concentrations for CS_2.

Concentration	Replicate 1		Replicate 2	
of CS_2 (mg/l)	y	n	y	n
49.06	2	29	4	30
52.99	7	30	6	30
56.91	9	28	9	34
60.84	14	27	14	29
64.76	23	30	29	33
68.69	29	31	24	28
72.61	29	30	32	32
76.54	29	29	31	31

The following S-plus commands were used to analyse this data:

```
> dead<-c(2,7,9,14,23,29,29,29,4,6,9,14,29,24,32,31)
> n<-c(29,30,28,27,30,31,30,29,30,30,34,29,33,28,32,31)
> cs2<-c(49.06,52.99,56.91,60.84,64.76,68.69,72.61,76.54,
+ 49.06,52.99,56.91,60.84,64.76,68.69,72.61,76.54)
> reps<-rep(c(1,2),c(8,8))
> reps<-factor(reps)
> beetles.glm2<-glm(dead/n~cs2*reps,family=binomial,weights=n)
```

(a) The `glm` function in S-plus can be used to fit logistic regression models to this data. Briefly, explain (drawing sketches may be useful) how the following three models are different. (4 marks)

(i) `glm(dead/n~cs2,family=binomial,weights=n)`

(ii) `glm(dead/n~cs2+reps,family=binomial,weights=n)`

(iii) `glm(dead/n~cs2*reps,family=binomial,weights=n)`

(b) For model (iii), the following Analysis of Deviance Table is obtained:

```
> beetles.glm2<-glm(dead/n~cs2*reps,family=binomial,weights=n)
> anova(beetles.glm2,test="Chi")
Analysis of Deviance Table
Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev   Pr(Chi)
    NULL                    15   289.1413
     cs2  1 276.6360        14    12.5053 0.0000000
    reps  1   0.0003        13    12.5050 0.9868596
cs2:reps  1   0.0038        12    12.5012 0.9509855
```

For each P-value in this table carefully state the hypothesis being tested. (4 marks)

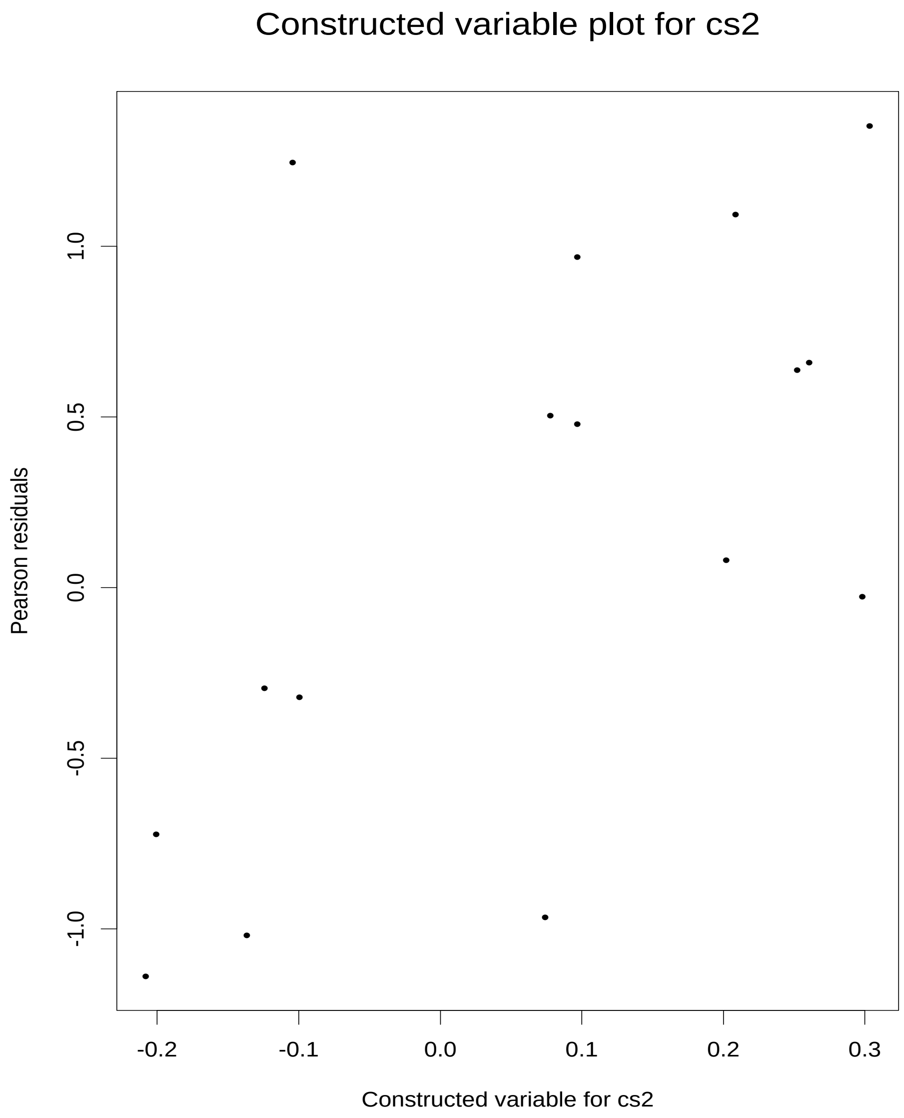
(c) Explain how the table in (b) leads to the conclusion that only `cs2` is needed in the model. (4 marks)

(d) For the model that just contains `cs2`, S-plus produces the following output:

```
> beetles.glm<-glm(dead/n~cs2,family=binomial,weights=n)
> summary(beetles.glm)
Coefficients:
                  Value Std. Error   t value
(Intercept) -14.8084459 1.28933576 -11.48533
        cs2   0.2491705 0.02137754  11.65571

    Null Deviance: 289.1413 on 15 degrees of freedom
Residual Deviance: 12.50526 on 14 degrees of freedom
```

Write down the logistic form of the fitted model for π = probability of death. Use this model to estimate π for a CS_2 concentration of 70 mg/l. (4 marks)

(e) A constructed variable plot for the model in (d) is given below.




```
Coefficients:
               Value Std. Error t value Pr(>|t|)
(Intercept) -0.0131  0.1818    -0.0719  0.9437
z.construct  2.7619  0.9670     2.8562  0.0127
```

Briefly (in 1 or 2 sentences) explain what a constructed variable plot is used for and how it works. What does this plot indicate about the current model? (4 marks)

3. The following data was extracted from applications to graduate studies for a number of different departments at the University of California at Berkeley. The applications were cross-classified by department, gender of applicant, and result (admitted or rejected).

	Male		Female	
Dept.	Admitted	Rejected	Admitted	Rejected
A	353	207	17	8
B	120	205	202	391
C	138	279	131	244
D	53	138	94	299
E	22	351	24	317

This data was analysed using the following S-plus commands:

```
> grad2.glm<-glm(counts~dept*gender*result,family=poisson,data=grad2.df)
> anova(grad2.glm,test="Chi")

Terms added sequentially (first to last)
                   Df Deviance Resid. Df Resid. Dev   Pr(Chi)
              NULL                    19   1879.776
              dept  4 111.5979        15   1768.178 0.0000000
            gender  1   5.3787        14   1762.799 0.0203836
            result  1 469.9027        13   1292.897 0.0000000
       dept:gender  4 753.4387         9    539.458 0.0000000
       dept:result  4 536.7766         5      2.681 0.0000000
     gender:result  1   0.1251         4      2.556 0.7236013
dept:gender:result  4   2.5564         0      0.000 0.6345607
```

(a) What model would you select for this data? Sketch the association graph for the model that you have selected. (4 marks)

(b) Are gender and result independent, conditionally independent, or related? Explain what this indicates about the relationship between the gender of an applicant and their chances of being accepted for graduate studies. (4 marks)

(c) If we are specifically interested in the relationship between gender and result, would it be sensible to collapse the contingency table for the above data into a 2-way table involving just gender and result? Explain. (4 marks)

(d) If we are specifically interested in the relationship between department and result would it be sensible to collapse the contingency table for the above data into a 2-way table involving just department and result? Explain. (4 marks)

4. The follow data was extracted from a study of maternal drinking (drinking during pregnancy) and congenital malformations in infants. A random sample of 32,574 pregnant women was selected. After the first 3 months of pregnancy the women in the sample completed a questionnaire about their alcohol consumption (drinks per day on average) over the first 3 months of pregnancy. Following childbirth, observations were recorded on the presence or absence of congenital malformations.

Alcohol Consumption	Malformation Present	Absent
< 1	86	31,530
≥ 1	7	951

(a) Calculate the odds ratio for the presence of congenital malformations comparing the ≥ 1 drink per day group with the < 1 drink per day group. Explain in words what this indicates about the relationship between maternal drinking and the odds of congenital malformations in infants. (4 marks)

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

(c) Is it possible to estimate the relative risk of congenital malformations given maternal alcohol consumption for this data? If so, calculate the estimated relative risk. If not, explain why it cannot be done for this data. (4 marks)

5. An experiment was carried to investigate factors that affect the germination of the seeds of a type of Lupin plant (*Lupinous polyphyllus*). Batches of 100 seeds each were stored at each of 3 different temperatures (21°C, 42°C, 62°C) and each of 3 different moisture levels (low, medium, high). Germination tests were carried out at 2 different temperatures (11°C, 21°C). The following table indicates the number of seeds (out of 100) that germinated for each combination of factors.

Germination temperature	Moisture level	Storage Temperature 21°C	42°C	62°C
11°C	low	98	96	62
11°C	medium	94	79	3
11°C	high	92	41	1
21°C	low	94	93	65
21°C	medium	94	71	2
21°C	high	91	30	1

(a) The following is the Analysis of Deviance table for the model that contains the three factors and the s.temp:moisture interaction:

```
Terms added sequentially (first to last)
                Df Deviance Resid. Df Resid. Dev    Pr(Chi)
           NULL                    17   1230.860
         s.temp  2 734.9062        15    495.954 0.00000000
       moisture  2 429.5838        13     66.370 0.00000000
         g.temp  1   3.1801        12     63.190 0.07453973
s.temp:moisture  4  58.7606         8      4.430 0.00000000
```

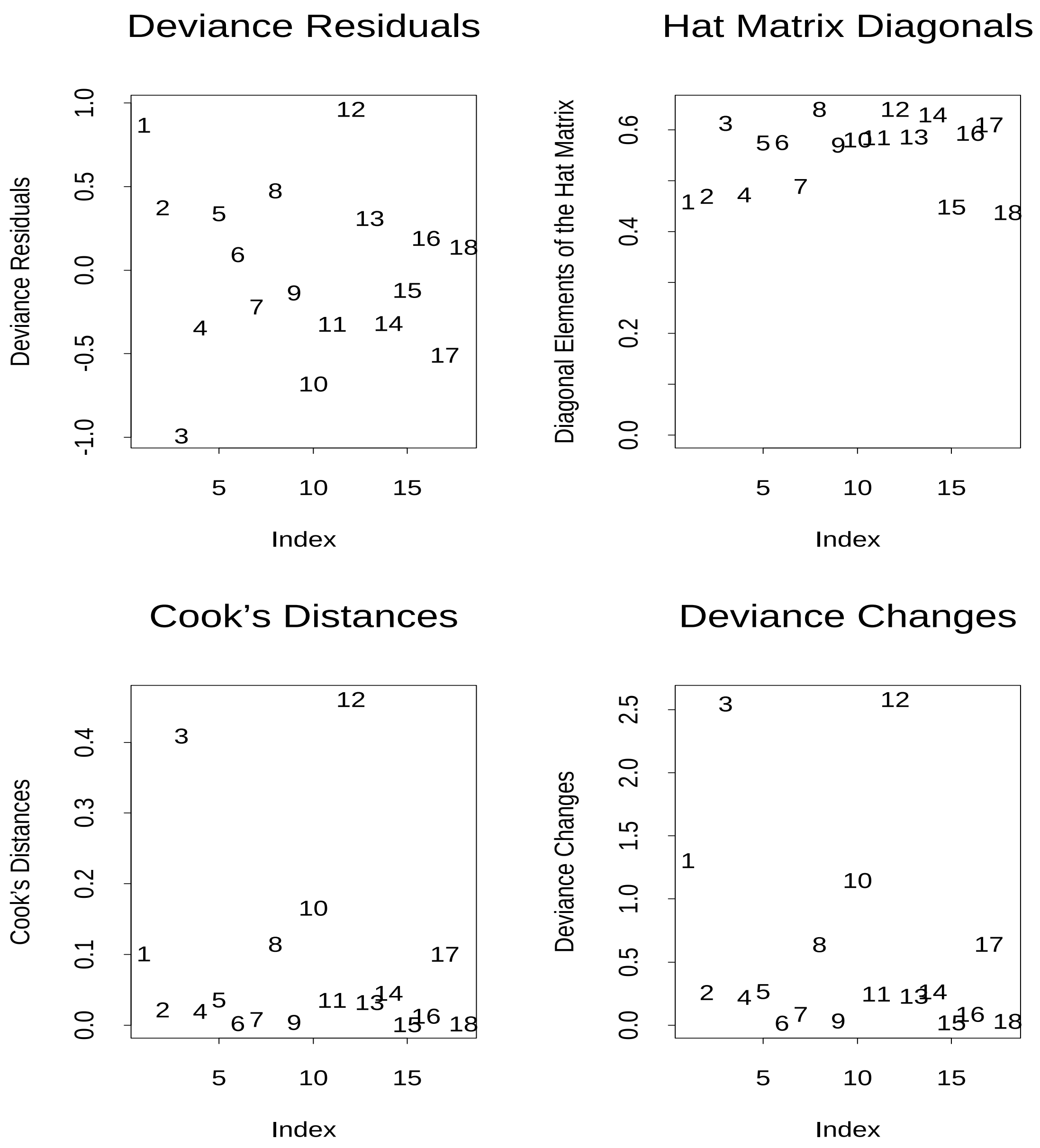
Explain how you would conduct a hypothesis test to determine if this model is adequate. Clearly state what hypothesis is being tested, what you would use for your test statistic, and how you would calculate the P-value. (4 marks)

(b) The following is the output from `dummy.coef` for the model in (a):

```
> dummy.coef(germ.glm3)
$"(Intercept)":
 (Intercept)
    2.522421

$s.temp:
 21         42         62
  0 -2.984096 -6.988644

$moisture:
 high       low      med
    0 0.8026329 0.375736

$g.temp:
 11         21
  0 -0.2764638

$"s.temp:moisture":
 21high 42high 62high 21low   42low    62low 21med    42med     62med
      0      0      0     0 2.64961 4.358132     0 1.327562 0.5561055
```

Find the estimated probability of germination for seeds that are stored at 42°C at the medium moisture level and then were germinated at 11°C. (4 marks)

(c) The following are a some diagnostic plots for the model in (a) and (b). Briefly, explain what these plots indicate about the fitted model. (4 marks)



(d) The DFFITS value for observation 12 is 4.286 and its DFBETA value for `g.temp` is 1.865. What do these values indicate? Note that observation 12 corresponds to seeds stored at 62°C and low moisture and germinated at 21°C. (4 marks)

6. An experiment was conducted to model oxygen uptake (OXY) as a function of five explanatory variables (all measured in milligrams per litre):

BOD : biological oxygen demand

TKN : total Kjedahl nitrogen

TS : total solids

TVS : total volatile solids

COD : chemical oxygen demand

Dairy wastes were kept suspended in water in a laboratory and were sampled every few days. The level of oxygen uptake and the levels of the 5 explanatory variables were measured for each sample. The data in the following table is in the order that the samples were taken.

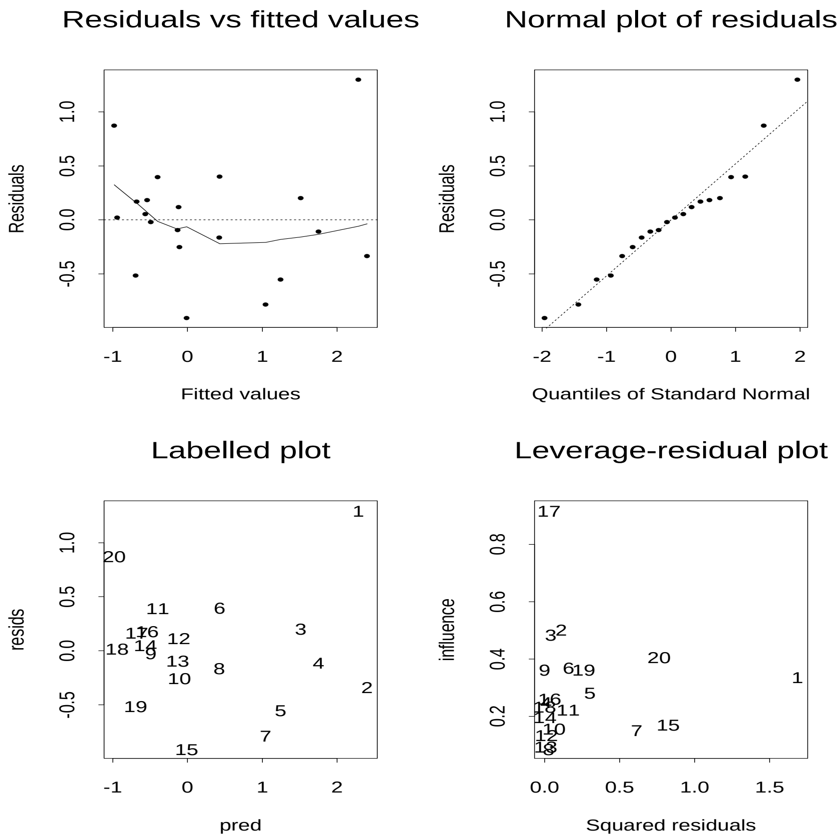
BOD	TKN	TS	TVS	COD	OXY
1125	232	7160	85.9	8905	36.0
920	268	8804	86.5	7388	7.9
835	271	8108	85.2	5348	5.6
1000	237	6370	83.8	8056	5.2
1150	192	6441	82.1	6960	2.0
990	202	5154	79.2	5690	2.3
840	184	5896	81.2	6932	1.3
650	200	5336	80.6	5400	1.3
640	180	5041	78.4	3177	0.6
583	165	5012	79.3	4461	0.7
570	151	4825	78.7	3901	1.0
570	171	4391	78.0	5002	1.0
510	243	4320	72.3	4665	0.8
555	147	3709	74.9	4642	0.6
460	286	3969	74.4	4840	0.4
275	198	3558	72.5	4479	0.7
510	196	4361	57.7	4200	0.6
165	210	3301	71.8	3410	0.4
244	327	2964	72.5	3360	0.3
79	334	2777	71.9	2599	0.9

(a) It was decided to model log(OXY) as a function of all 5 regressors. The following S-plus output is for the fitted model:

```
> summary(oxy.lm)

Call: lm(formula = log(OXY) ~ ., data = oxygen.df)
Residuals:
     Min      1Q    Median     3Q   Max
 -0.9076 -0.2727 0.0009278 0.1892 1.299

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -4.9683  2.1096    -2.3551  0.0336
        BOD  0.0000  0.0012    -0.0359  0.9719
        TKN  0.0030  0.0029     1.0338  0.3188
         TS  0.0003  0.0002     1.6659  0.1179
        TVS  0.0183  0.0323     0.5645  0.5814
        COD  0.0003  0.0002     1.9224  0.0751

Residual standard error: 0.6046 on 14 degrees of freedom
Multiple R-Squared: 0.8094
F-statistic: 11.89 on 5 and 14 degrees of freedom, the p-value is 0.0001238
```

What do you conclude from the S-plus output for the fitted model (make sure you comment on the t-tests for the model coefficients, the overall F-test, and the R^2 statistic)? (6 marks)

(b) Some diagnostic plots for the fitted model in (a) are given below. Comment on the suitability of this model. Identify any unusual observations and explain what makes them unusual. (4 marks)



(c) The following variance inflation factors were obtained for the model in (b). Note that these are ordered: BOD, TKN, TS, TVS, COD.

```
> VIF
[1] 7.134815 1.298440 4.455225 2.437709 4.366239
```

What do these tell us about the regressors? (4 marks)

(d) What problem can occur with data (such as this data) that are collected over time? What plot would you use to check for this problem? How does this plot indicate the problem is present (compare the appearance of the plot when the problem is present to the appearance when the problem is not present)? (4 marks)

(e) The S-plus output from `all.poss.regs` and a plot of the Mallow's C_p statistic are given below. Use these to identify a suitable model(s) for this data. Justify your choice(s). (6 marks)

```
> all.poss.regs(X,lOXY,nbest=4)
          rssp sigma2 adjRsq     Cp BOD TKN TS TVS COD
1(#1)   8.149  0.453  0.680  6.295   0   0  1   0   0
1(#2)   8.252  0.458  0.676  6.575   0   0  0   0   1
1(#3)  10.783  0.599  0.576 13.502   1   0  0   0   0
1(#4)  13.278  0.738  0.478 20.327   0   0  0   1   0
2(#1)   5.753  0.338  0.760  1.739   0   0  1   0   1
2(#2)   7.045  0.414  0.707  5.273   0   0  0   1   1
2(#3)   7.628  0.449  0.682  6.870   0   1  0   0   1
2(#4)   7.633  0.449  0.682  6.884   1   0  1   0   0
3(#1)   5.234  0.327  0.768  2.319   0   1  1   0   1
3(#2)   5.638  0.352  0.751  3.424   0   0  1   1   1
3(#3)   5.643  0.353  0.750  3.439   1   0  1   0   1
3(#4)   6.457  0.404  0.714  5.665   0   1  0   1   1
4(#1)   5.118  0.341  0.759  4.001   0   1  1   1   1
4(#2)   5.234  0.349  0.753  4.319   1   1  1   0   1
4(#3)   5.508  0.367  0.740  5.069   1   0  1   1   1
4(#4)   6.132  0.409  0.711  6.775   1   1  0   1   1
5(#1)   5.117  0.366  0.741  6.000   1   1  1   1   1
```

Mallow's Cp Plot