THE UNIVERSITY OF AUCKLAND **475.330SC**

EXAMINATION FOR BA BSc ETC 1997

STATISTICS

Advanced Statistical Modelling

(Time allowed: THREE hours)

NOTE: Attempt all FIVE questions. All are worth equal marks.

1. (a) What assumptions are made about the measurement errors in (normal) regression? What plots would you do to check these assumptions? (5 marks)

(b) Name three useful criteria which can be used for selecting a subset of regression variables. (3 marks)

(c) Compute the odds ratio for the following 2×2 table which gives the results of car accidents in Florida. (2 marks)

	Injury	
Seat Belt	Fatal	Non-fatal
No	1601	162527
Yes	510	412368

(d) What property of the odds ratio makes it useful when used with retrospective data. (2 marks)

(e) Suppose a GLM has been fitted and a residual deviance of 13.3 is obtained with 14 degrees of freedom. Is the fit satisfactory? Explain. (2 marks)

(f) Suppose that a logistic model with linear predictor $\beta_0 + \beta_1 x_{i1}$ has been fit to some data and the estimates $\widehat{\beta}_0 = -2$ and $\widehat{\beta}_1 = 1$ have been obtained. Compute the estimated probability of success when $x = 2$. (3 marks)

(g) Write down the assumed response distribution and link function which are used when *fitting* log-linear models. (3 marks)

2. In 1973 the members of OPEC decided to use their near-monopoly control of the world's oil to force its price up. They did this by reducing the supply to oil-importing countries by placing an *embargo* on export. Major upheavals occurred in the importing countries as a result and measures were taken to reduce oil consumption.

One of the main steps taken was an attempt to reduce highway driving speeds. A side effect of this reduction was a major drop in the road toll. The following table shows data for some road toll related variables (taken from *Historical Statistics of the United States*).

YEAR	*RATE*	*SPEED*	*EMB*
60	5.3	16	0
61	5.2	18	0
62	5.3	21	0
63	5.4	29	0
64	5.6	32	0
65	5.5	34	0
66	5.7	40	0
67	5.5	44	0
68	5.4	45	0
69	5.2	46	0
70	4.9	47	0
71	4.7	50	0
72	4.4	50	0
73	4.2	50	0
74	3.6	21	1
75	3.5	21	1

RATE gives death rate per 10^8 vehicle miles, *SPEED* the percentage of drivers at speeds greater than 60 mph and *EMB* is an indicator variable showing which years were affected by the embargo.

The following results were obtained when a regression model was fitted which predicts *RATE* from the other variables. The result of fitting this model is

```
               Value Std. Error  t value Pr(>|t|)
(Intercept)  23.7410   1.5571    15.2473   0.0000
      SPEED   0.0899   0.0098     9.2182   0.0000
       YEAR  -0.3297   0.0286   -11.5205   0.0000
        EMB   2.4874   0.3944     6.3071   0.0000
```

The R^2 value for the regression is 0.98.

The model suggests that the introduction of the oil embargo *increased* the death rate by 2.5. This does not seem to agree with the data values.

(a) Set down a causal diagram which specifies the how you think the variables affect one another. Describe why you set down this particular diagram. (6 marks)

(b) The following regression results were obtained for a variety of models. Redraw the causal diagram, and use the regression results to fill in coefficients and to delete unneeded links. (6 marks)

EMB ~ SPEED
```
                 Value Std. Error t value Pr(>|t|)
(Intercept)  0.5386  0.2409     2.2355  0.0422
      SPEED -0.0117  0.0065    -1.8185  0.0904
```
EMB ~ YEAR
```
                 Value Std. Error t value Pr(>|t|)
(Intercept) -2.6544  1.0623    -2.4987  0.0255
       YEAR  0.0412  0.0157     2.6224  0.0201
```
EMB ~ SPEED + YEAR
```
                  Value Std. Error  t value Pr(>|t|)
(Intercept)  -3.7794   0.3167   -11.9326   0.0000
      SPEED  -0.0238   0.0019   -12.5757   0.0000
       YEAR   0.0703   0.0051    13.9087   0.0000
```
SPEED ~ EMB
```
                  Value Std. Error  t value Pr(>|t|)
(Intercept)  37.2857   3.1662    11.7762   0.0000
        EMB -16.2857   8.9553    -1.8185   0.0904
```
SPEED ~ YEAR
```
                  Value Std. Error  t value Pr(>|t|)
(Intercept) -47.3382  42.9697    -1.1017   0.2892
       YEAR   1.2235   0.6351     1.9265   0.0746
```
SPEED ~ EMB + YEAR
```
                   Value Std. Error   t value  Pr(>|t|)
(Intercept) -150.5476   14.7781   -10.1872    0.0000
        EMB  -38.8822    3.0919   -12.5757    0.0000
       YEAR    2.8246    0.2218    12.7336    0.0000
```
YEAR ~ EMB
```
                 Value Std. Error t value Pr(>|t|)
(Intercept) 66.5000  1.0785    61.6570  0.0000
        EMB  8.0000  3.0506     2.6224  0.0201
```
YEAR ~ SPEED
```
                 Value Std. Error t value Pr(>|t|)
(Intercept) 61.4629  3.3196    18.5150  0.0000
      SPEED  0.1713  0.0889     1.9265  0.0746
```
YEAR ~ EMB + SPEED
```
                 Value Std. Error t value Pr(>|t|)
(Intercept) 54.2793  1.0070    53.9020  0.0000
        EMB 13.3378  0.9590    13.9087  0.0000
      SPEED  0.3278  0.0257    12.7336  0.0000
```

(c) Use the causal diagram to determine the full effect of embargo on death rates. Does this agree with the data? (4 marks)

(d) Would it make sense to carry out a Durbin-Watson test in this case? Explain. (4 marks)

3. The data in the table below are from an experimental piggery arranged individual feeding of six pigs in each of five pens. From each of five litters, six young pigs were selected and allotted to one of the pens. Three different feeding treatments denoted by A, B and C were used and each given to one male and one female in each pen. The pigs were individually weighed each week for 16 weeks. For each pig, the growth rate in pounds per week was calculated as the slope of a line fitted by least squares. This is denoted by y in the following table; the weight at the beginning of the experiment is denoted by x.

Pen	Variable	Food A Male	Food A Female	Food B Male	Food B Female	Food C Male	Food C Female
1	y	9.52	9.94	8.51	10.00	9.11	9.75
	x	38	48	39	48	48	48
2	y	8.21	9.48	9.95	9.24	8.50	8.66
	x	35	32	38	32	37	28
3	y	9.32	9.32	8.43	9.34	8.90	7.63
	x	41	35	46	41	42	33
4	y	10.56	10.90	8.86	9.68	9.51	10.37
	x	48	46	40	46	42	50
5	y	10.42	8.82	9.20	9.67	8.76	8.57
	x	43	32	40	37	40	30

The data are set up so that the the first level of the `sex` factor is that for "Male".

(a) The following two ANOVA tables were generated from the data.

```
> anova(lm(y~x+food+sex+pen))

          Df  Sum Sq  Mean Sq       F     Pr(>F)
x          1   5.559   5.5591  23.760  8.066e-05
food       2   2.317   1.1584   4.951  1.731e-02
sex        1   1.163   1.1628   4.970  3.684e-02
pen        4   2.393   0.5983   2.557  6.881e-02
Residual  21   4.913   0.2340

> anova(lm(y~x+pen+sex+food))

          Df  Sum Sq  Mean Sq       F     Pr(>F)
x          1   5.559   5.5591  23.760  8.066e-05
pen        4   2.256   0.5640   2.411  8.142e-02
sex        1   1.277   1.2773   5.459  2.945e-02
food       2   2.340   1.1698   5.000  1.675e-02
Residual  21   4.913   0.2340
```

Explain why the p-values for `pen` differ in the two tables. (4 marks)

(b) Why is it the second of the ANOVA tables above which is appropriate for testing whether there is a difference in the feeding treatments? (4 marks)

(c) The ANOVA table indicates that there is a significant difference between the treatments. Using the following `summary` output, describe the nature of the differences between feeding treatments. (4 marks)

```
> summary(lm(y~x+pen+sex+food), cor=F)

Call:
lm(formula = y ~ x + pen + sex + food)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07576 -0.28104  0.06541  0.26855  0.83356

Coefficients:
            Estimate Std.Error t Value Pr(>|t|)
(Intercept)   5.5369    1.0518  5.2645   0.0000
x             0.0913    0.0224  4.0709   0.0005
pen2          0.5542    0.3751  1.4777   0.1543
pen3         -0.1768    0.3023 -0.5846   0.5650
pen4          0.4627    0.2795  1.6555   0.1127
pen5          0.4833    0.3299  1.4650   0.1577
sex           0.4293    0.1826  2.3510   0.0286
foodB        -0.4431    0.2173 -2.0397   0.0542
foodC        -0.6730    0.2163 -3.1112   0.0053

Residual standard error: 0.4837 on 21 degrees of freedom
Multiple R-Squared: 0.6994
F-statistic: 6.108 on 8 and 21 degrees of freedom,
    the p-value is 0.0004092
```

(d) Predict the growth rate for a male pig which is kept in pen 1, given food B and which starts with an initial weight of $x = 45$. (4 marks)

(e) What important assumptions about the form of the model are being made in the analyses above. How would you check these assumptions. (4 marks)

4. Minor faults occur irregularly in an industrial process and, as an aid to their diagnosis, the following experiment was done. Batches of raw material were selected and each batch was divided into two equal sections: for each batch, one of the sections was processed by the standard method and the other by a slightly modified process, in which the temperature at one stage is reduced. Before processing, a purity index was measured for the whole batch of material. For the product from each section of material it as recorded whether minor faults did (F) or did not (NF) occur. Results for 22 batches are given in the following table.

Purity Index	Standard Process	Modified Process	Purity Index	Standard Process	Modified Process
7.2	NF	NF	6.5	NF	F
6.3	F	NF	4.9	F	F
8.5	F	NF	5.3	F	NF
7.1	NF	F	7.1	NF	F
8.2	F	NF	8.4	F	NF
4.6	F	NF	8.5	NF	F
8.5	NF	NF	6.6	F	NF
6.9	F	F	9.1	NF	NF
8.0	NF	NF	7.1	F	NF
8.0	F	NF	7.5	NF	F
9.1	NF	NF	8.3	NF	NF

F : Faults occur. NF : No faults occur.

Throughout the analysis "success" is to be understood as the presence of a fault.

(a) Indicate how you would set up this data for computer analysis. You do not need to write any computer input here; just show how the data could be written in a form suitable for processing by the `read.table` function in S-PLUS. (5 marks)

(b) The regression results below were all obtained by using S-PLUS model fitting statements of the form; `summary(glm(`*model*`, family=binomial), corr=FALSE)`. For each of these four models, give a brief description (in words) of the model being fitted and comment on the estimated coefficients and standard errors. (5 marks)

Model 1 : `y ~ process`

```
Coefficients:
               Value  Std.error  z value  P(>|z|)
(Intercept)   0.0000     0.4264   0.0000    1.000
processMod   -0.7621     0.6254  -1.2187    0.223

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.53428 on 43 degrees of freedom

Residual deviance: 58.0201 on 42 degrees of freedom
```

Model 2 : `y ~ purity`

```
Coefficients:
               Value  Std.error  z value  P(>|z|)
(Intercept)   3.8639     2.0378   1.8961   0.0579
purity       -0.5795     0.2761  -2.0992   0.0358

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.53428 on 43 degrees of freedom

Residual deviance: 54.52249 on 42 degrees of freedom
```

Model 3 : `y ~ process + purity`

```
Coefficients:
               Value  Std.error  z value  P(>|z|)
(Intercept)   4.4608     2.1585   2.0666   0.0388
processMod   -0.8645     0.6716  -1.2873   0.1980
purity       -0.6042     0.2837  -2.1296   0.0332

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.53428 on 43 degrees of freedom

Residual deviance: 52.81186 on 41 degrees of freedom
```

Model 4 : `y ~ process + purity + process:purity`

```
Coefficients:
                     Value  Std.error  z value  P(>|z|)
(Intercept)         6.4329     3.5092   1.8331   0.0668
processMod         -4.3379     4.4385  -0.9773   0.3284
purity             -0.8684     0.4641  -1.8709   0.0614
processMod:purity   0.4744     0.5956   0.7965   0.4258

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59.53428 on 43 degrees of freedom

Residual deviance: 52.14955 on 40 degrees of freedom
```

(c) Which process gives the higher probability of faults? Explain. (5 marks)

(d) Use model 3 to predict the probability of a fault during the standard process when the input has a purity index of 8. (5 marks)

5. The data in the table below relate to an investigation into satisfaction with housing conditions in Copenhagen. A total of 1681 residents from selected areas living in rented homes built between 1960 and 1968 were questioned in their satisfaction, the degree of contact with other residents and their feeling of influence on housing management. The purpose of the investigation was to study the relationship between these factors and the type of housing.

Contact		*Low*			*High*		
Satisfaction		*L*	*M*	*H*	*L*	*M*	*H*
Housing	**Influence**						
	L	21	21	28	14	19	37
Tower Blocks	*M*	34	22	36	17	23	40
	H	10	11	36	3	5	23
	L	13	9	10	20	23	20
Houses	*M*	8	8	12	10	22	24
	H	6	7	9	7	10	21

The data were entered into S-PLUS and the following variables were created for use in the analysis. created:

`C` – Level of contact with other residents.
`S` – Level of satisfaction.
`I` – Level of influence on management.
`H` – Housing type.
`Y` – Response count.

The data were used to fit a variety of log-linear models to the data.
[*Note: there is a Chi-square table at the end of this question.*]

(a) As part of the investigation, a sequence of models was fitted to the the data subset consisting of the containing just the *houses*, the following residual degrees of freedom and deviances obtained.

Fitted Model	Resid. D.f.	Resid. Deviance
`Y ~ C + S + I`	12	11.56
`Y ~ C + S + I + C:S`	10	9.12
`Y ~ C + S + I + C:I`	10	11.36
`Y ~ C + S + I + S:I`	8	3.93
`Y ~ C + S + I + C:S + C:I`	8	8.93
`Y ~ C + S + I + C:S + S:I`	6	1.48
`Y ~ C + S + I + C:I + S:I`	6	3.74
`Y ~ C + S + I + C:I + C:S + S:I`	4	1.20

Which model would you select as appropriate for modelling this data subset? Explain. (5 marks)

(b) How would you interpret the model which you have chosen in part (a). (5 marks)

(c) The following results were obtained for for the data subset consisting of *tower block* residents.

Fitted Model	Resid. D.f.	Resid. Deviance
`Y ~ C + S + I`	12	28.80
`Y ~ C + S + I + C:S`	10	22.06
`Y ~ C + S + I + C:I`	10	23.78
`Y ~ C + S + I + S:I`	8	14.32
`Y ~ C + S + I + C:S + C:I`	8	17.04
`Y ~ C + S + I + C:S + S:I`	6	7.58
`Y ~ C + S + I + C:I + S:I`	6	9.30
`Y ~ C + S + I + C:I + C:S + S:I`	4	0.57

Which model would you select as appropriate for modelling this data subset? Explain. (5 marks)

(d) How would you interpret the model which you have chosen in part (c). (5 marks)

Chi-square distribution

For fixed *prob* and *df*, the tabulated value is the number $\chi^2 = \chi^2_{df}(prob)$ such that for $\boldsymbol{\chi}^2 \sim$ Chi-square(df), $pr(\boldsymbol{\chi}^2 \geq \chi^2) = prob$.

[e.g. For $prob = 0.10$ and $df = 13$, $\chi^2_{13}(0.10) = 19.81$]

prob

df	.95	.20	.15	.10	.05	.025	.01	.005	.001
1	0.004	1.642	2.072	2.706	3.841	5.024	6.635	7.879	10.83
2	0.103	3.219	3.794	4.605	5.991	7.378	9.210	10.60	13.82
3	0.352	4.642	5.317	6.251	7.815	9.348	11.34	12.84	16.27
4	0.711	5.989	6.745	7.779	9.488	11.14	13.28	14.86	18.47
5	1.145	7.289	8.115	9.236	11.07	12.83	15.09	16.75	20.52
6	1.635	8.558	9.446	10.64	12.59	14.45	16.81	18.55	22.46
7	2.167	9.803	10.75	12.02	14.07	16.01	18.48	20.28	24.32
8	2.733	11.03	12.03	13.36	15.51	17.53	20.09	21.95	26.12
9	3.325	12.24	13.29	14.68	16.92	19.02	21.67	23.59	27.88
10	3.940	13.44	14.53	15.99	18.31	20.48	23.21	25.19	29.59
11	4.575	14.63	15.77	17.28	19.68	21.92	24.72	26.76	31.26
12	5.226	15.81	16.99	18.55	21.03	23.34	26.22	28.30	32.91
13	5.892	16.98	18.20	19.81	22.36	24.74	27.69	29.82	34.53
14	6.571	18.15	19.41	21.06	23.68	26.12	29.14	31.32	36.12
15	7.261	19.31	20.60	22.31	25.00	27.49	30.58	32.80	37.70
16	7.962	20.47	21.79	23.54	26.30	28.85	32.00	34.27	39.25
17	8.672	21.61	22.98	24.77	27.59	30.19	33.41	35.72	40.79
18	9.390	22.76	24.16	25.99	28.87	31.53	34.81	37.16	42.31
19	10.12	23.90	25.33	27.20	30.14	32.85	36.19	38.58	43.82
20	10.85	25.04	26.50	28.41	31.41	34.17	37.57	40.00	45.31
21	11.59	26.17	27.66	29.62	32.67	35.48	38.93	41.40	46.80
22	12.34	27.30	28.82	30.81	33.92	36.78	40.29	42.80	48.27
23	13.09	28.43	29.98	32.01	35.17	38.08	41.64	44.18	49.73
24	13.85	29.55	31.13	33.20	36.42	39.36	42.98	45.56	51.18
25	14.61	30.68	32.28	34.38	37.65	40.65	44.31	46.93	52.62
26	15.38	31.79	33.43	35.56	38.89	41.92	45.64	48.29	54.05
27	16.15	32.91	34.57	36.74	40.11	43.19	46.96	49.64	55.48
28	16.93	34.03	35.71	37.92	41.34	44.46	48.28	50.99	56.89
29	17.71	35.14	36.85	39.09	42.56	45.72	49.59	52.34	58.30
30	18.49	36.25	37.99	40.26	43.77	46.98	50.89	53.67	59.70
45	30.61	52.73	54.81	57.51	61.66	65.41	69.96	73.17	80.08
50	34.76	58.16	60.35	63.17	67.50	71.42	76.15	79.49	86.66
60	43.19	68.97	71.34	74.40	79.08	83.30	88.38	91.95	99.61
70	51.74	79.71	82.26	85.53	90.53	95.02	100.4	104.2	112.3
80	60.39	90.41	93.11	96.58	101.9	106.6	112.3	116.3	124.8
90	69.13	101.1	103.9	107.6	113.1	118.1	124.1	128.3	137.2
100	77.93	111.7	114.7	118.5	124.3	129.6	135.8	140.2	149.4