STATS 330
STATS 794

THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2001
Campus: City

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) Leave one out diagnostics can be used to assess the influence that each observation has on a 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 Deviance Changes of 4.6.

(iii) An observation has a value of DFBETA of 1.08 for a particular explanatory variable. (5 marks)

(b) The "dissimilarity index" is sometimes used when selecting a model for contingency table data.

(i) Describe a situation where the dissimilarity index would be useful.

(ii) What does the dissimilarity index measure and how is it used? (5 marks)

CONTINUED

(c) The following data comes from a case-control study investigating the occurrence of esophageal cancer. A sample of 200 cases were matched with 775 controls. The following table classifies the cases and controls by their annual alcohol consumption.

	Annual Alcohol consumption	
	≥ 80 g	< 80 g
Cases	96	104
Controls	109	666

Create a 95% confidence interval for the odds ratio that compares the odds of esophageal cancer for alcohol consumption ≥ 80 g to that for alcohol consumption < 80 g. Explain what this interval indicates. Note that if $Z \sim \mathrm{N}(0,1)$, then $\Pr(-1.96 \leq Z \leq 1.96) = 0.95$. (5 marks)

(d) An experiment was conducted to test the compressive strength of an alloy fastener used in the construction of aircraft. Samples of fasteners were tested at ten different pressure loads increasing from 2500 psi to 4300 psi in steps of 200 psi. The number of fasteners failing at each load was recorded.

Load	Number tested	Number failing
2500	50	10
2700	70	17
2900	100	30
3100	60	21
3300	40	18
3500	85	43
3700	90	54
3900	50	33
4100	80	60
4300	65	51

A logistic regression model was fitted and the following output obtained:

```
Coefficients:
                  Value   Std. Error   t value
(Intercept) -5.339711512 0.5456929729 -9.785194
       load  0.001548434 0.0001575379  9.828962
```

(i) Write down the fitted model in both its logistic form and in its logit form.

(ii) Use the fitted model to estimate the load at which 50% of fasteners fail.

(5 marks)

2. Criminologists are interested in identifying predictors of crime rates. Data for the following variables were obtained for 47 states of the USA for 1960:

M: percentage of males aged 14-24

Ed: mean years of schooling

Po1: police expenditure in 1960

Po2: police expenditure in 1959

Pop: state population

U1: unemployment rate of urban males 14-24

U2: unemployment rate of urban males 35-39

GDP: gross domestic product per head

Ineq: income inequality

Prob: probability of imprisonment

Time: average time served in state prisons

Crime: crime rate

(a) A regression model that models log(Crime) using the remaining variables as predictors was fitted using S-plus:

```
Call: lm(formula = log(Crime) ~ M + Ed + Po1 + Po2 + Pop + U1 + U2
                  + GDP + Ineq + Prob + Time, data = crime.df)

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -0.3123  1.2317    -0.2535  0.8014
          M  0.0131  0.0039     3.3665  0.0019
         Ed  0.0199  0.0058     3.4342  0.0015
        Po1  0.0153  0.0103     1.4846  0.1466
        Po2 -0.0064  0.0111    -0.5727  0.5705
        Pop -0.0008  0.0013    -0.6536  0.5177
         U1 -0.0058  0.0035    -1.6462  0.1087
         U2  0.0197  0.0083     2.3853  0.0226
        GDP  0.0017  0.0010     1.6082  0.1168
       Ineq  0.0090  0.0020     4.4612  0.0001
       Prob -4.7143  2.1127    -2.2314  0.0322
       Time -0.0058  0.0070    -0.8266  0.4140

Residual standard error: 0.2212 on 35 degrees of freedom
Multiple R-Squared: 0.7798
F-statistic: 11.26 on 11 and 35 degrees of freedom,
                                        the p-value is 1.589e-08
```

What do you conclude about the suitability of this model based on this output (comment on the t-tests for the model coefficients, the overall F-test, and the R^2 statistic)? (5 marks)

(b) For the model in part (a) log(Crime) was used as the response. Name three assumptions of the regression model that are affected by transforming the response. For each assumption, identify a diagnostic plot that can be used to detect problems with that assumption and explain what type of pattern in the plot indicates a problem is present. (5 marks)

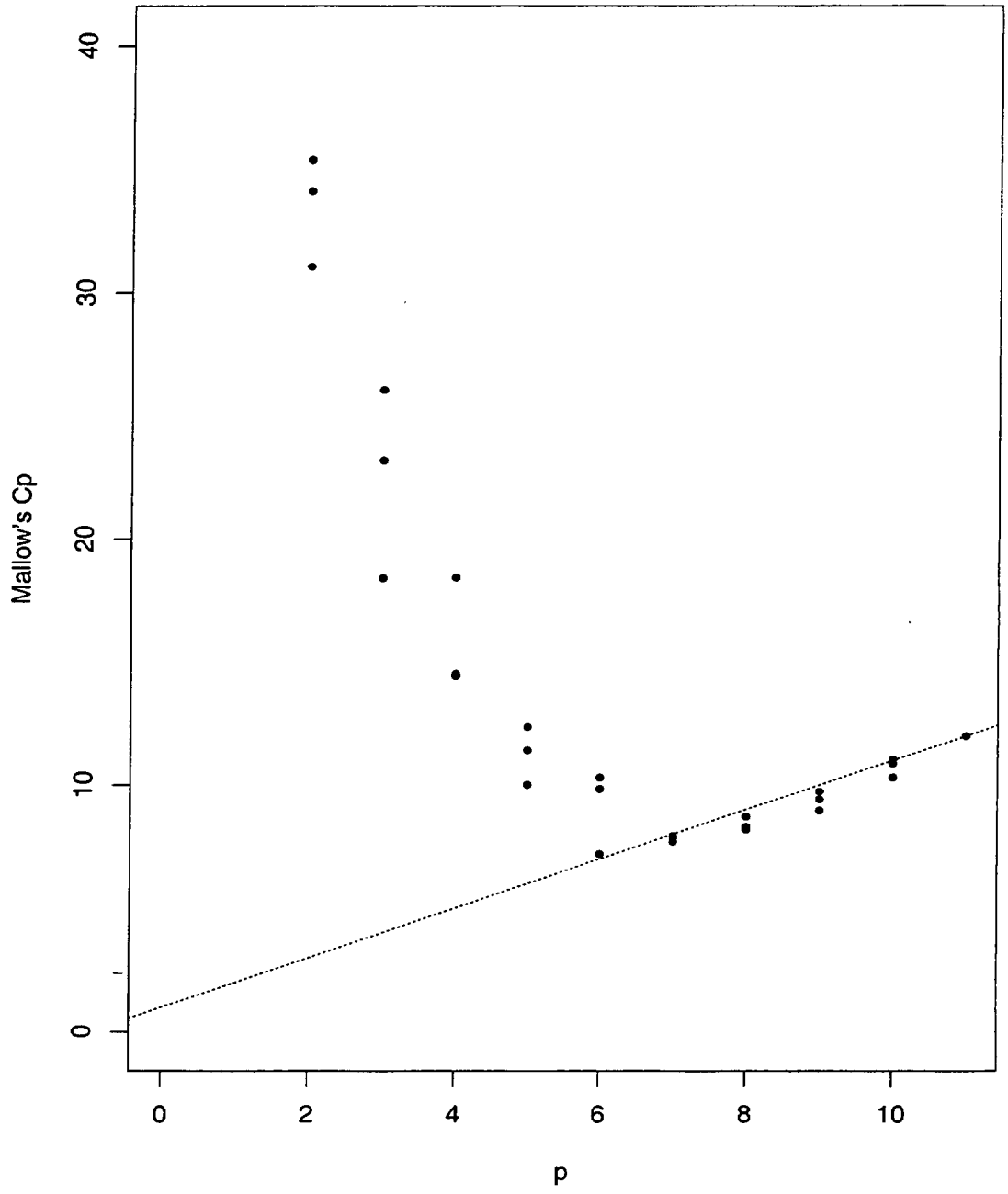
(c) Variance inflation factors (VIF's) are used to detect multicollinearity. Explain what is meant by multicollinearity. What impact does multicollinearity have on model selection? The VIF's for this data set are:

M	Ed	Po1	Po2	Pop	U1	U2	GDP	Ineq	Prob	Time
2.25	3.94	88.51	90.83	2.15	3.80	4.58	9.44	6.07	2.17	2.30

What do these VIF's indicate? (5 marks)

(d) The S-plus output from `all.poss.regs` and a plot of the Mallow's C_p statistic are given below.

	rssp	sigma2	adjRsq	Cp	M	Ed	Po1	Po2	Pop	U1	U2	GDP	Ineq	Prob	Time
1(#1)	362.763	8.061	0.461	47.524	0	0	1	0	0	0	0	0	0	0	0
1(#2)	382.230	8.494	0.432	52.382	0	0	0	1	0	0	0	0	0	0	0
1(#3)	554.078	12.313	0.177	95.265	0	0	0	0	0	0	0	1	0	0	0
2(#1)	288.781	6.563	0.561	31.062	0	0	1	0	0	0	0	0	1	0	0
2(#2)	301.088	6.843	0.543	34.134	1	0	1	0	0	0	0	0	0	0	0
2(#3)	306.214	6.959	0.535	35.413	0	0	0	1	0	0	0	0	1	0	0
3(#1)	230.076	5.351	0.642	18.413	0	1	1	0	0	0	0	0	1	0	0
3(#2)	249.225	5.796	0.613	23.192	0	1	0	1	0	0	0	0	1	0	0
3(#3)	260.712	6.063	0.595	26.058	0	0	1	0	0	0	0	0	1	1	0
4(#1)	206.135	4.908	0.672	14.439	1	1	1	0	0	0	0	0	1	0	0
4(#2)	206.578	4.919	0.671	14.549	0	1	1	0	0	0	0	0	1	1	0
4(#3)	222.122	5.289	0.646	18.428	0	1	1	0	0	0	0	1	1	0	0
5(#1)	180.329	4.398	0.706	9.999	1	1	1	0	0	0	0	0	1	1	0
5(#2)	186.036	4.537	0.697	11.424	1	1	1	0	0	0	1	0	1	0	0
5(#3)	189.797	4.629	0.691	12.362	1	1	1	0	0	0	0	1	1	0	0
6(#1)	161.106	4.028	0.731	7.202	1	1	1	0	0	0	1	0	1	1	0
6(#2)	171.680	4.292	0.713	9.841	1	1	1	0	0	0	0	1	1	1	0
6(#3)	173.511	4.338	0.710	10.298	1	1	1	0	0	0	1	1	1	0	0
7(#1)	155.115	3.977	0.734	7.707	1	1	1	0	0	0	1	1	1	1	0
7(#2)	155.623	3.990	0.733	7.834	1	1	1	0	0	1	1	0	1	1	0
7(#3)	155.974	3.999	0.733	7.922	1	1	1	0	1	0	1	0	1	1	0
8(#1)	149.000	3.921	0.738	8.182	1	1	1	0	1	1	1	0	1	1	0
8(#2)	149.385	3.931	0.737	8.277	1	1	1	0	1	0	1	1	1	1	0
8(#3)	151.107	3.976	0.734	8.707	1	1	1	0	0	1	1	1	1	1	0
9(#1)	144.104	3.895	0.740	8.960	1	1	1	0	1	1	1	1	1	1	0
9(#2)	145.954	3.945	0.736	9.421	1	1	1	1	1	1	1	0	1	1	0
9(#3)	147.219	3.979	0.734	9.737	1	1	1	1	1	0	1	1	1	1	0
10(#1)	141.431	3.929	0.737	10.293	1	1	1	1	1	1	1	1	1	1	0
10(#2)	143.717	3.992	0.733	10.863	1	1	1	0	1	1	1	1	1	1	1
10(#3)	144.394	4.011	0.732	11.032	1	1	1	1	0	1	1	1	1	1	1
11(#1)	140.258	4.007	0.732	12.000	1	1	1	1	1	1	1	1	1	1	1



Explain how the Mallow's C_p plot can be used to identify suitable subset models. Which variables do you think should be included in the regression model for this data? Justify your answer. (5 marks)

3. The data presented in the following table were collected as part of a study that investigated preterm infants born in three different hospitals. The birth weight in kilograms (wt), the gestation age in weeks (gest), and the hospital of birth (hosp) was recorded for 40 preterm infants.

wt	gest	hosp	wt	gest	hosp	wt	gest	hosp
1.4	30	A	1.0	27	A	1.0	30	A
0.9	27	B	1.8	35	B	0.9	28	A
1.2	33	A	1.4	36	C	1.0	31	A
1.1	29	C	1.2	34	A	1.6	31	B
1.3	35	A	1.1	28	B	1.6	33	B
0.8	27	B	1.2	30	B	1.7	34	B
1.0	32	A	1.0	29	C	1.6	35	C
0.7	26	A	1.4	33	C	1.2	28	A
1.2	30	C	0.9	28	A	1.5	30	B
0.8	28	A	1.0	28	C	1.8	34	B
1.5	32	B	1.9	36	B	1.5	34	C
1.3	31	A	1.3	29	B	1.2	30	A
1.4	32	C	1.7	35	C	1.2	32	C
1.5	33	B						

(a) For this data treat wt as the response, and gest and hosp as explanatory variables. Consider the regression models that are produced by the following S-plus commands:

(i) `lm(wt~hosp)`

(ii) `lm(wt~gest)`

(iii) `lm(wt~hosp+gest)`

(iv) `lm(wt~hosp+gest+hosp:gest)`

For each of these models, explain how wt is related to gest and hosp for that model. Make sure that your explanations clearly indicate how the models are different from each other. (5 marks)

(b) The variable hosp is a factor with 3 levels and thus two dummy variables are needed in the regression model to account for differences between hospitals. Assume that the dummy variables are defined as: $I_B = 1$ for hospital B and 0 otherwise, and $I_C = 1$ for hospital C and 0 otherwise. Write out the theoretical form ($E(\text{wt}) = \beta_0 + \ldots$) of the model (iv) from part (a). Clearly explain what each of the coefficients (β's) in this model represent. (5 marks)

(c) Model (iv) from part (a) was fitted using S-plus and the output from the anova function is:

```
bwt.fit<-lm(wt~hosp+gest+hosp:gest)
anova(bwt.fit)

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value      Pr(F)
hosp       2  1.024017 0.512009  28.7047 0.00000005
gest       1  1.825926 1.825926 102.3669 0.00000000
hosp:gest  2  0.187596 0.093798   5.2586 0.01023695
Residuals 34  0.606461 0.017837
```

Explain what hypothesis is being tested by each line in this table. Also explain why this table indicate that model (iv) should be preferred to the other models in part (a). (5 marks)

(d) The output from the dummy.coef function for the fitted model from (c) is:

```
> dummy.coef(bwt.fit)
$"(Intercept)":
 (Intercept)
  -0.3927785

$hosp: -
 A         B          C
 0 -1.567158 -0.7432735

$gest:
       gest
 0.04876204

$"hosp:gest":
 Agest      Bgest      Cgest
     0 0.05975531 0.02771551
```

Compare the birth weights of preterm infants between the three hospitals taking into account gestation age. Is there any indications that birth weights are higher at a particular hospital (if so which one)? (5 marks)

4. The following data comes from a study to investigate the effects of taking acetylaminoflourine (AAF). A large number of mice had AAF added to their diets over a prolonged period. Various concentrations of AAF (dose) were investigated as well as various periods of exposure (months). The mice were examined for liver cancer and the proportion developing cancer for each different combination of dose and months were recorded.

Months	Dose (parts per 10,000)							
on diet	0	.30	.35	.45	.60	.75	1.00	1.50
9	0/199	1/147	1/76	0/52	0/345	0/186	1/168	1/169
12	0/164	1/151	2/27	1/14	2/283	0/153	3/149	2/152
15	0/115	1/75	1/35	0/20	3/203	1/109	5/99	1/100
18	6/155	34/2014	20/1102	15/550	13/411	17/382	19/213	24/211
24	20/762	164/2109	128/1361	98/888	118/758	118/587	76/297	126/314

(a) Consider the S-plus output for a logistic regression model that uses months and dose to predict the probability of liver cancer:

```
                Value Std. Error   t value
(Intercept) -9.4719990 0.28668227 -33.04006
     months  0.2730741 0.01198792  22.77910
       dose  1.7819600 0.08274694  21.53506

    Null Deviance: 1208.208 on 39 degrees of freedom
Residual Deviance: 64.36696 on 37 degrees of freedom
```

This output can be used to test (1) H_0:$\beta_{\text{months}} = \beta_{\text{dose}} = 0$ which results in a very small p-value and (2) H_0: "the model is adequate" which results in a p-vaule of 0.003.

(i) For each test explain how the p-value was obtained. You should give the expression for the test statistic, the value of the test statistic, and the expression used to calculate the p-value.

(ii) What do the results of these two tests indicate about this model? (5 marks)

(b) Consider an alternative model that uses months, dose and and dose^2 as regressors:

```
                Value Std. Error    t value
(Intercept) -9.894010 0.30032148 -32.944731
     months  0.270810 0.01191559  22.727368
       dose  3.418067 0.35105469   9.736566
  I(dose^2) -0.984069 0.20367812  -4.831491

    Null Deviance: 1208.208 on 39 degrees of freedom
Residual Deviance: 40.10767 on 36 degrees of freedom
```

State precisely what hypothesis is being tested by the "t value" for I(dose^2). What does the observed value of -4.831491 indicate about this hypothesis? Outline a χ^2 test that could be used to test the same hypothesis. As in (a) you should give the expression for the test statistic, the value of the test statistic, and the expression used to calculate the p-value. (5 marks)

CONTINUED

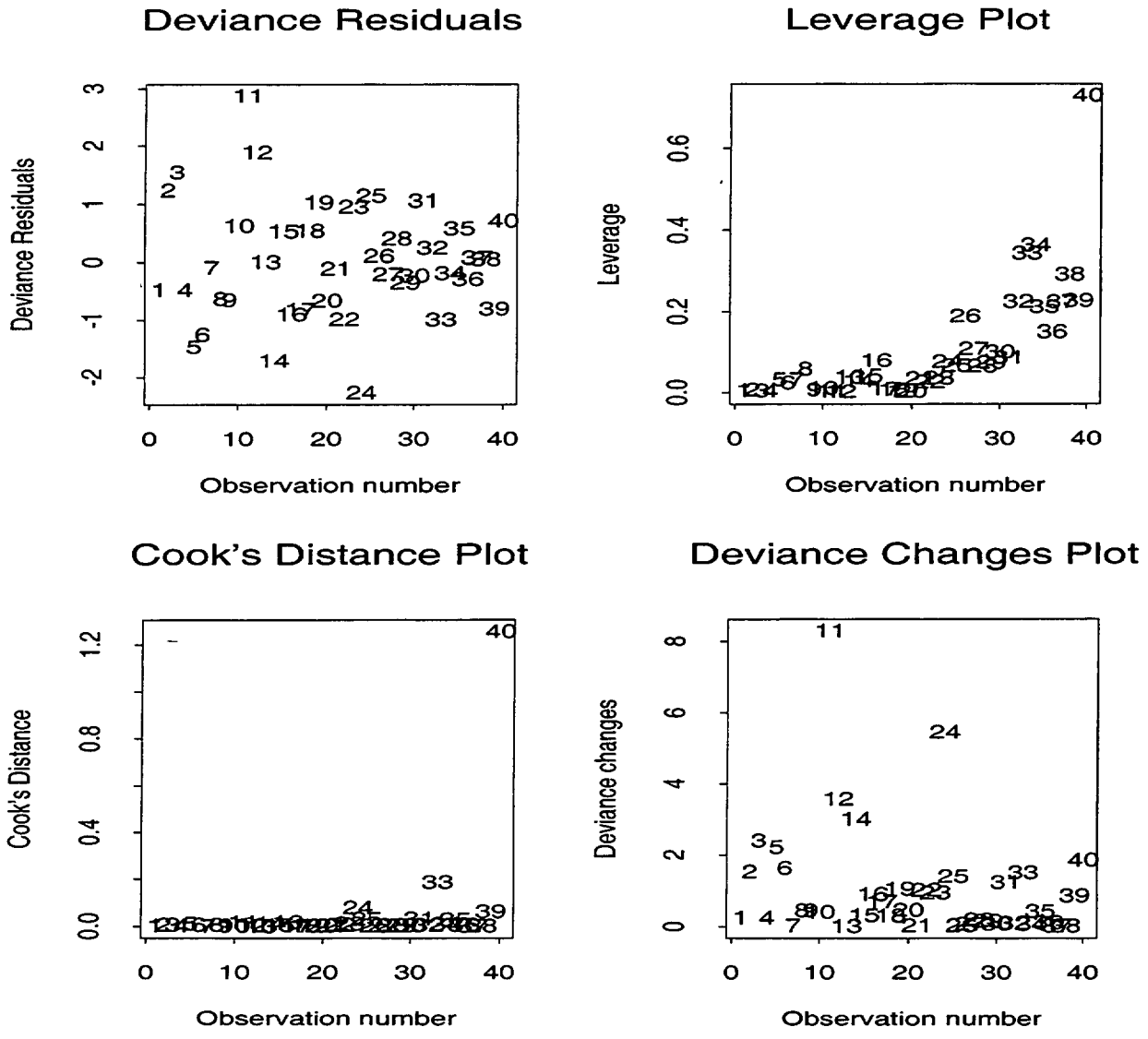
(c) Use the model from (b) to estimate the following:

(i) The effect on the odds of cancer if dose is fixed and months is increased from 12 to 24 months.

(ii) The effect on the odds of cancer if months is fixed and dose is increased from 0.5 to 1.5 parts per 10,000.

(5 marks)

(d) The following set of diagnostic plots was produced for the model from part (b).



Explain what problem(s) each of these plots is used to detect. Does this set of plots indicate any problems with the model? If a problem is indicated, briefly describe an appropriate course of action. (5 marks)

5. The following data cross-classify breast cancer patients according to five factors:

C: Centre where patient was diagnosed: Tokyo, Boston or Glamorgan.

A: Patient's age at time of diagnosis: < 50, $50 - 69$ or > 69.

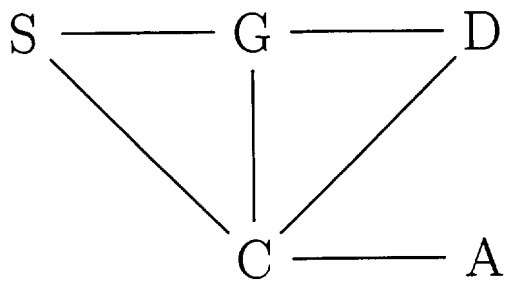
S: Three year survival: yes or no.

D: Degree of chronic inflammation: minimal or severe.

G: Grade of tumour: malignant or benign.

			Degree of Inflammation (D)			
			minimal		severe	
Diagnostic	Age	Three Year	Grade of Tumour (G)			
Centre (C)	(A)	Survival (S)	malignant	benign	malignant	benign
Tokyo	< 50	No	9	7	4	3
		Yes	26	68	25	9
	$50 - 69$	No	9	9	11	2
		Yes	20	46	18	5
	> 69	No	2	3	1	0
		Yes	1	6	5	1
Boston	< 50	No	6	7	6	0
		Yes	11	24	4	0
	$50 - 69$	No	8	20	3	2
		Yes	18	58	10	3
	> 69	No	9	18	3	0
		Yes	15	26	1	1
Glamorgan	< 50	No	16	7	3	0
		Yes	16	20	8	1
	$50 - 69$	No	14	12	3	0
		Yes	27	39	10	4
	> 69	No	3	7	3	0
		Yes	12	11	4	1

(a) An initial analysis of the data produced the following association graph:

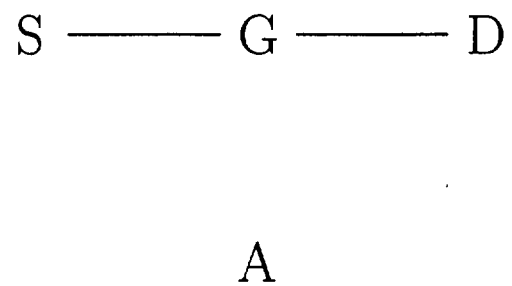


Give a brief explanation of how the data would be analysed to produce this association graph – describe the type of regression you would use, how you would select a model and how you would use that model to produce the association graph. (5 marks)

(b) Use the association graph from part (a) to determine how three year survival, S, is related (independent, directly related or conditionally independent) to each of the other factors. In cases of conditional independence indicate what factors must be fixed (conditioned on). (5 marks)

(c) Explain what is meant by conditional independence. Your explanation should clearly explain the difference between conditional independence and independence. (5 marks)

(d) Suppose we are particularly interested in the Glamorgan treatment centre. If only the data from Glamorgan is considered then the association graph for the other four factors becomes:



Would it be sensible to collapse the 4-way table for the Glamorgan data to create:

(i) A 2-way table for S and G?

(ii) A 2-way table for S and D?

(iii) A 2-way table for S and A?

Explain how you decided whether it was sensible to collapse the table or not. (5 marks)