STATS 330 SC
STATS 794 SC

THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2002
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) Each of the following diagnostic plots is used to diagnosis potential problems with ordinary regression models (Normal errors). Identify the problem that each plot is used to detect. Also, for each plot describe the appearance of the plot when no problem is present and when the problem is present.

(i) A normal probability plot of the residuals.

(ii) A plot of residuals versus lagged residuals.

(5 marks)

(b) Multicollinearity is one problem that can occur in a dataset.

(i) What is meant by multicollinearity?

(ii) What diagnostic is used to detect multicollinearity? How does this diagnostic indicate multicollinearity is present?

(5 marks)

(c) Beetles were exposed to gaseous carbon disulphide and their mortality rates recorded after five hours:

Obs	Dose	Deaths	Total	Obs	Dose	Deaths	Total
1	1.6907	6	59	2	1.7242	13	60
3	1.7552	18	62	4	1.7842	28	56
5	1.8113	52	63	6	1.8389	53	59
7	1.8610	61	62	8	1.8839	60	80

A logistic regression model that relates the probability of death to the level of dose was estimated using *R*:

```
> beetle.fit<-glm(Deaths/Total~Dose,family=binomial, weights=Total)
> summary(beetle.fit)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -40.971      3.658  -11.20   <2e-16 ***
Dose          23.040      2.044   11.28   <2e-16 ***
```

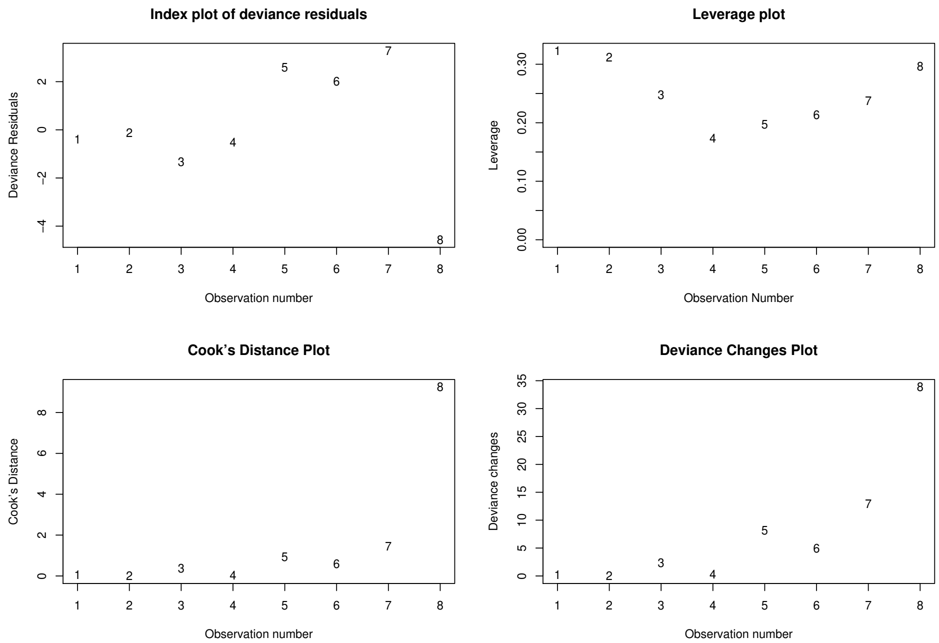
(i) Write down the fitted model in the logistic form ($\hat{\pi} = \ldots$) and in the logit form ($\operatorname{logit} \hat{\pi} = \ldots$).

(ii) According to the fitted model what effect does increasing the level of Dose by 0.1 units have on the odds of death?

(iii) What level of Dose corresponds to an estimated probability of death of 0.80?

(5 marks)

(d) The following diagnostics plots were obtained for the fitted model in part (c). Note that the observation numbers in the plots match those in the data table.



Do these plots indicate any problems with the fitted model? (For each plot state the problem that it is used detect and state whether or not that problem is present.) If you identify a problem briefly describe what action you would take to address it. (5 marks)

2. The data for this question relate to air pollution in 41 American cities. The variables measured were:

SO2:	annual mean SO_2 content (μg per m^3).
temp:	average annual temperature (°F).
manu:	number of manufacturing plants employing ≥ 20 workers.
pop:	population size (in thousands).
wind:	average annual wind speed (mph).
prep:	average annual precipitation (inches).
days:	average number of days with precipitation per year.

A regression model that predicts SO2 was obtained using R:

```
> usair.fit<-lm(SO2~.,data=usair.df)
> summary(usair.fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.29560   46.18128   2.345  0.02519 *
temp         -1.22725    0.60602  -2.025  0.05101 .
manu          0.06497    0.01823   3.565  0.00114 **
pop          -0.03910    0.01557  -2.512  0.01709 *
wind         -3.18212    1.85407  -1.716  0.09548 .
prep          0.48142    0.35518   1.355  0.18448
days         -0.03163    0.15711  -0.201  0.84170
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.84 on 33 degrees of freedom
Multiple R-Squared: 0.5304,Adjusted R-squared: 0.4451
F-statistic: 6.213 on 6 and 33 DF,  p-value: 0.0001929
```

(a) The multiple regression model for this example can be written as:

$$\mathsf{SO2} = \beta_0 + \beta_1\,\mathsf{temp} + \beta_2\,\mathsf{manu} + \beta_3\,\mathsf{pop} + \beta_4\,\mathsf{wind} + \beta_5\,\mathsf{prep} + \beta_6\,\mathsf{days} + \epsilon$$

where ϵ represents the error term.

(i) What three assumptions are made about the error terms in a standard regression model?

(ii) The estimated coefficients in the output given above were obtained using "least squares" estimation. Briefly explain what this means.

(5 marks)

(b) Consider the line for manu from the R output.

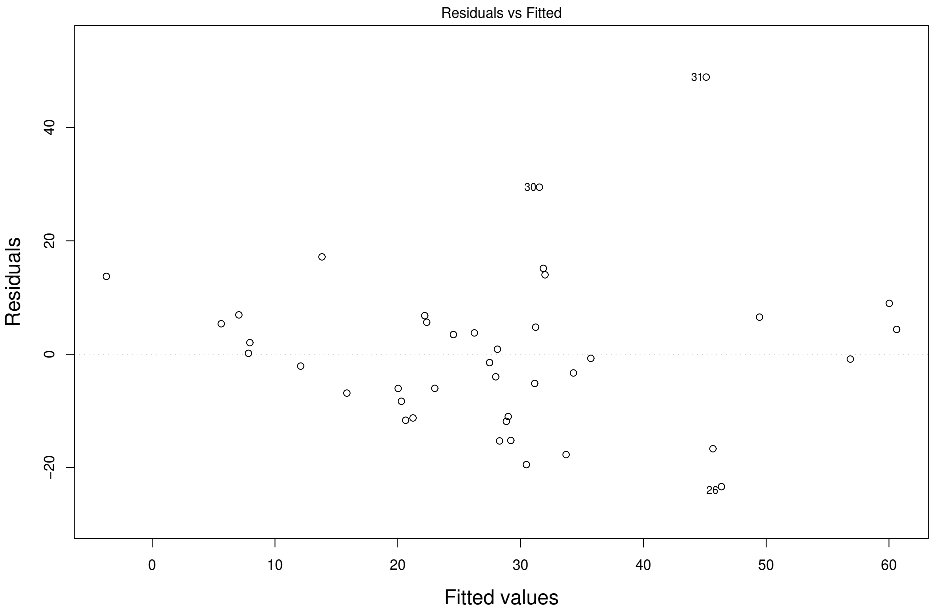
(i) What does the estimated coefficient for manu indicate about the effect manu has on the predicted value of SO2? Be precise in your explanation.

(ii) Explain how you would create a 95% confidence interval for the coefficient of manu? Note that you cannot actually produce this interval – just explain how you would do it.

(iii) What hypothesis is being tested by the P-value (0.00114) on this line (be precise)?

(5 marks)

(c) A plot of studentised residuals versus fitted values for the model from part(a) is given below:



(i) The studentised residuals are obtained by dividing the ordinary residuals by the square root of one minus the hat matrix diagonals: $r_i^* = r_i/\sqrt{1 - h_{ii}}$. What advantage do the studentised residuals have over the ordinary residuals?

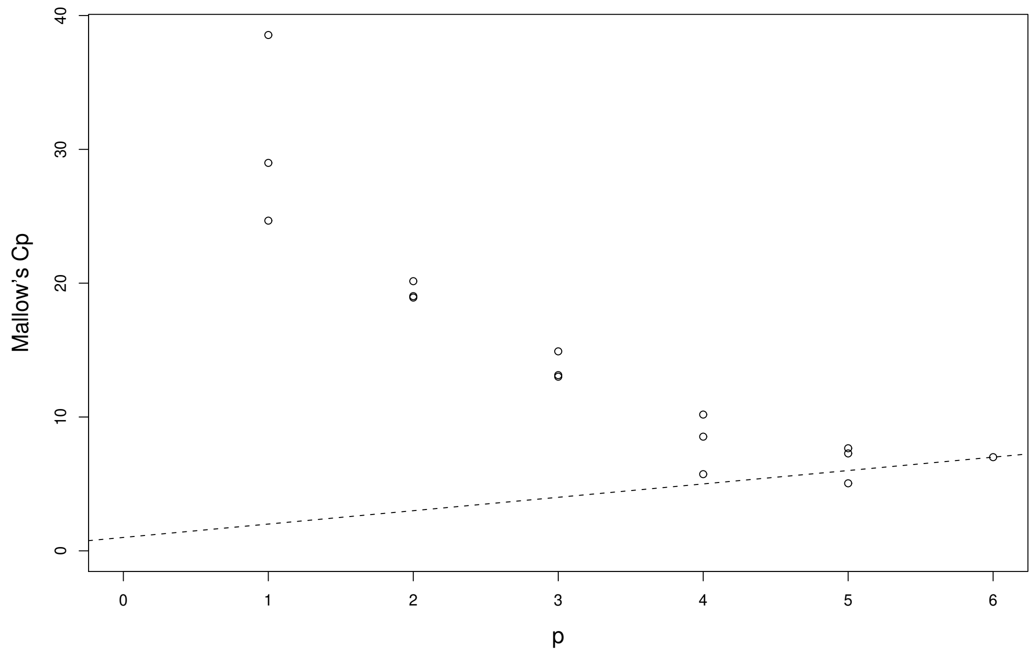
(ii) What model deficiency is clearly indicated by this plot? Make sure that you explicitly state what pattern in the plot indicates that this problem is present.

(iii) To correct the problem identified in (ii), it was decided that log(SO2) should be used as the response. Name the three assumptions of the regression model that are affected by transforming the response – one of these should be consistent with your answer to (ii).

(5 marks)

(d) The R output from `all.poss.regs` (using log(SO2) as the response) and a plot of Mallow's C_p statistic are given below:

```
    rssp sigma2 adjRsq     Cp temp manu pop wind prep days
1 12.198  0.321  0.275 24.673    1    0   0    0    0    0
1 13.066  0.344  0.224 28.989    0    0   0    0    0    1
1 14.987  0.394  0.110 38.546    0    1   0    0    0    0
2 10.640  0.288  0.351 18.925    1    0   0    0    0    1
2 10.661  0.288  0.349 19.030    1    0   0    0    1    0
2 10.887  0.294  0.336 20.150    1    1   0    0    0    0
3  9.049  0.251  0.432 13.011    1    1   0    1    0    0
3  9.073  0.252  0.431 13.129    1    0   0    1    1    0
3  9.430  0.262  0.409 14.906    1    0   0    1    0    1
4  7.183  0.205  0.537  5.729    1    1   0    1    1    0
4  7.746  0.221  0.500  8.531    1    1   0    1    0    1
4  8.078  0.231  0.479 10.182    1    1   1    1    0    0
5  6.644  0.195  0.559  5.047    1    1   1    1    1    0
5  7.093  0.209  0.529  7.279    1    1   1    1    0    1
5  7.170  0.211  0.524  7.666    1    1   0    1    1    1
6  6.634  0.201  0.546  7.000    1    1   1    1    1    1
```



Based on this output, which model or models would you select? Justify your choice(s). (5 marks)

3. A soft drink bottler is analyzing vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. This service activity including stocking the machine with beverage products and minor maintenance or housekeeping. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time are the number of cases of product stocked and the distance walked by the route driver. The engineer has collected 25 observations on delivery time (minutes), number of cases and distance walked (metres).

Time	Cases	Distance	Time	Cases	Distance	Time	Cases	Distance
16.68	7	171	11.50	3	67	12.03	3	104
14.88	4	24	13.75	6	46	18.11	7	101
8.00	2	34	17.83	7	64	79.24	30	445
21.50	5	184	40.33	16	210	21.00	10	66
13.50	4	78	19.75	6	141	24.00	9	137
29.00	10	237	15.35	6	61	19.00	7	40
9.50	3	11	35.10	17	235	17.90	10	43
52.32	26	247	18.75	9	137	19.83	8	194
10.75	4	46						

The engineer decided to fit a model that used Cases, Distance and the Cases:Distance interaction as regressors. Her reason for including the interaction term was that she felt that the impact that the distance walked by the route driver has on delivery time will increase as the number of cases increases.

```
> delivery.fit2<-lm(Time~Cases*Distance)
> summary(delivery.fit2)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.149924   1.402289   5.099 4.75e-05 ***
Cases          1.013252   0.191662   5.287 3.06e-05 ***
Distance       0.018963   0.011102   1.708 0.102358
Cases:Distance 0.002441   0.000575   4.246 0.000361 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.452 on 21 degrees of freedom
Multiple R-Squared: 0.9782,Adjusted R-squared: 0.9751
F-statistic: 313.8 on 3 and 21 DF,  p-value:   0
```

(a) Explain what each of the following parts of the regression output indicates:

(i) `Residual standard error: 2.452`

(ii) `Multiple R-Squared: 0.9782`

(iii) `F-statistic: 313.8 on 3 and 21 DF,  p-value:   0`

(5 marks)

(b) Does the fitted model support the engineer's contention that "the impact that the distance walked by the route driver has on delivery time will increase as the number of cases increases"? Justify your answer.

(5 marks)

(c) The `predict` command in R can be used to create two types of interval estimates: confidence intervals and prediction intervals. The following interval estimates were obtained for a delivery involving 25 cases that have to be transported 50 metres:

```
> new.df<-data.frame(list( Cases=25 , Distance=50))
> predict(delivery.fit2,new.df,interval="confidence",level=.95)
          fit      lwr      upr
[1,] 36.48094 29.43656 43.52532

> predict(delivery.fit2,new.df,interval="prediction",level=.95)
          fit      lwr      upr
[1,] 36.48094 27.78517 45.17671
```

Explain, in language that a non-statistics major can understand, what each of these intervals represent. Make sure that the difference between these two types of interval estimates is clear from your explanation.

(5 marks)

(d) (i) An index plot of Cook's Distance (not shown) indicates that observation 9 (this is the observation with `time` = 79.24) has a much larger value of Cook's Distance than any other observation. What does this indicate about observation 9?

(ii) The output from `influence.measures` for observation 9 is:

```
   dfb.1.  dfb.Case  dfb.Dist dfb.Cs.D  dffit  cov.r   cook.d   hat   inf
9  0.954    -1.310    -0.653   2.240    3.304  12.917  2.740   0.922   *
```

What do you deduce from (1) the value of the covariance ratio and (2) the values of the DFBETA's? Note that `dfb.Cs.D` represents the DFBETA diagnostic for the interaction term.

(5 marks)

4. The data in the following table summarises the results from an experiment that investigated the mortality rates of mice exposed to nitrous dioxide (NO_2). We wish to investigate how the probability of death is related to two explanatory variables: degree of exposure (dose) and exposure time (time). For each combination of dose and time listed in the table, the number of deaths (s) and the total number of mice exposed (n) is recorded.

dose	time	s	n
low	96.0	44	120
low	168.0	37	80
low	336.0	43	80
low	504.0	35	60
med	0.5	29	100
med	1.0	53	200
med	2.0	13	40
med	3.0	75	200
med	5.0	23	40

dose	time	s	n
med	7.0	152	280
med	14.0	55	80
med	24.0	98	140
med	48.0	121	160
high	0.5	52	120
high	1.0	62	120
high	1.5	61	120
high	2.0	86	120

The output from R for the fitted logistic model that uses dose and time as regressors is:

```
> m.fit1<-glm(s/n~dose+time,family=binomial, weights=n)
> summary(m.fit1)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.9936347  0.2110191  -4.709 2.49e-06 ***
dosemed      0.9464771  0.2111414   4.483 7.37e-06 ***
dosehigh     1.1646165  0.2293377   5.078 3.81e-07 ***
time         0.0035737  0.0007436   4.806 1.54e-06 ***
---
    Null deviance: 200.09  on 16  degrees of freedom
Residual deviance: 170.65  on 13  degrees of freedom
```

(a) There are two chi-square tests that are based on the values of the Null deviance and the Residual deviance given in this output.

Test 1: This test uses the difference between the Null deviance and the Residual deviance as the test statistic. The P-value of this test is given by:

```
> 1-pchisq(29.44,3)
[1] 1.809917e-06
```

Test 2: This test uses the Residual deviance as the test statistic. The P-value of this test is given by:

```
> 1-pchisq(170.65,13)
[1] 0
```

(i) State the null hypothesis that is being tested in each case.

(ii) What do you conclude about the fitted model from the results of these two tests?

(5 marks)

(b) It was thought that using log(time) instead of time might produce a better model. The output for this model (along with the P-values for the two chi-square tests) is:

```
> m.fit2<-glm(s/n~dose+log(time),family=binomial,weights=n)
> summary(m.fit2)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.03858    0.25613  -11.86   <2e-16 ***
dosemed      2.11992    0.20238   10.47   <2e-16 ***
dosehigh     3.16095    0.26889   11.76   <2e-16 ***
log(time)    0.54995    0.04365   12.60   <2e-16 ***
---
    Null deviance: 200.090  on 16  degrees of freedom
Residual deviance:  16.017  on 13  degrees of freedom

> 1-pchisq(184.073,3)
[1] 0

> 1-pchisq(16.017,13)
[1] 0.2482131
```

Does this output support using log(time) instead of time in the model? Justify your response.

(5 marks)

(c) Use the following output to produce a 95% confidence interval for the probability of death when dose is at the medium level and exposure time is 20. Note that if $Z \sim \mathrm{N}(0,\, 1)$, then $\Pr(-1.96 \leq Z \leq 1.96) = 0.95$.

```
> new.df<-data.frame(list(dose="med",time=20))
> predict(m.fit2,new.df,se=T)
$fit
[1] 0.7288546
$se.fit
[1] 0.08396623
```

(5 marks)

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

```
(Intercept):    -3.038579
dose:                 low      med     high
                 0.000000 2.119920 3.160955
log(time):          0.5499537
```

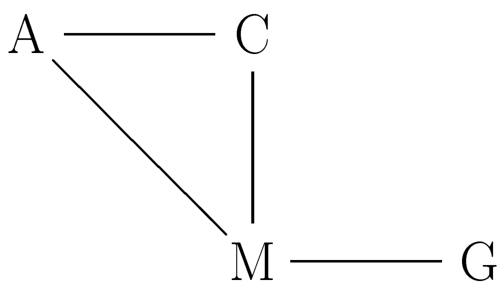
Use this output to estimate the odds ratio that compares the odds of death for dose = high with the odds of death for dose = med. Over what range of values for time is this comparison valid?
(5 marks)

CONTINUED

5. The data in the following table refers to a survey of students in their final year of high school from a non-urban area near Dayton Ohio USA. The survey asked the students whether they had ever used alcohol, cigarettes or marijuana. The given table also classifies the students by gender.

	Gender (G)	Female		Male	
		Marijuana Use (M)			
Alcohol Use (A)	Cigarette Use (C)	Yes	No	Yes	No
Yes	Yes	428	291	483	247
	No	15	237	29	219
No	Yes	1	18	2	25
	No	1	129	1	140

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



This graph was created by first fitting a Poisson regression model to the data and then writing down the association graph based on the identified model.

(i) The Poisson regression model for this data only contained main effects and 2-way interactions. Given this information, what model must have been identified in order to produce the given association graph?

(ii) Give an example of a second Poisson regression model that could have produced this association graph (note you are not restricted to main effects and 2-way interactions for this model).

(5 marks)

(b) What does the association graph indicate about the relationship between:

(i) gender (G) and marijuana use (M)

(ii) gender (G) and alcohol use (A).

In each case state the nature of the relationship (independent, conditionally independent or directly related) and then explain what that means in terms of this data.

(5 marks)

(c) If we collapse this table on alcohol use (A) and cigarette use (C) we get the following 2-way table:

	Marijuana Use	
Gender	Yes	No
Female	445	675
Male	515	631

Use this table to find a 95% confidence interval for the *odds ratio* that compares the odds of marijuana use for females to the odds for males. Note that if $Z \sim \mathrm{N}(0,\, 1)$, then $\Pr(-1.96 \leq Z \leq 1.96) = 0.95$.

(5 marks)

(d) The main objective of this study was to determine how marijuana use was related to alcohol use, to cigarette use, and to gender.

(i) Was it sensible to collapse the original table into the 2-way table given in part (b) to focus on the relationship between marijuana use and gender? Explain your response.

(ii) Explain how you would investigate the relationship between marijuana use and alcohol use and the relationship between marijuana use and cigarette use. Note that you are not required to actually investigate these relationships - just explain how you would go about it.

(5 marks)