THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2003
Campus: City

STATISTICS

Advanced Statistical Modeling
Topics in Statistics C

(Time allowed: THREE hours)

INSTRUCTIONS

SECTION A: Multiple Choice (40 marks)

- Answer **ALL 25** questions on the answer sheet provided.
- All questions have a single correct answer and carry the same mark value.
- If you give more than one answer to any question you will receive zero marks for that question.
- Incorrect answers are not penalised.

SECTION B (60 marks)

- Answer **ALL 3** questions. Each is worth 20 marks.

SECTION A

1. Suppose that we have a data set consisting of a continuous response variable Y, and two categorical explanatory variables X and Z, observed on each of 60 individuals. The variable X has 3 levels and Z has 2 levels. Which one of the following plots would be **most suitable** for portraying the relationships between the variables?

   (zz) A trellis plot with 3 panels corresponding to the levels of X, with each panel containing two boxplots, one for each level of Z.

   (1) A three-dimensional scatterplot.

   (1) A trellis plot consisting of 60 two-way tables.

   (1) A scatterplot of X versus Z, with the value of Y shown by a colour coding.

   (1) A barchart of Y, with colour coding indicating the levels of X and Z.

2. Suppose that we have a data set consisting of a continuous response variable Y, and two continuous explanatory variables X and Z. Which one of the following plots would be **unsuitable** for showing if the data were approximately planar?

   (zz) A normal plot of the residuals.

   (1) A coplot of Y versus X, conditioning on Z.

   (1) A coplot of Y versus Z, conditioning on X.

   (1) A 3-dimensional scatterplot.

   (1) A plot of residuals from the fitted model versus fitted values.

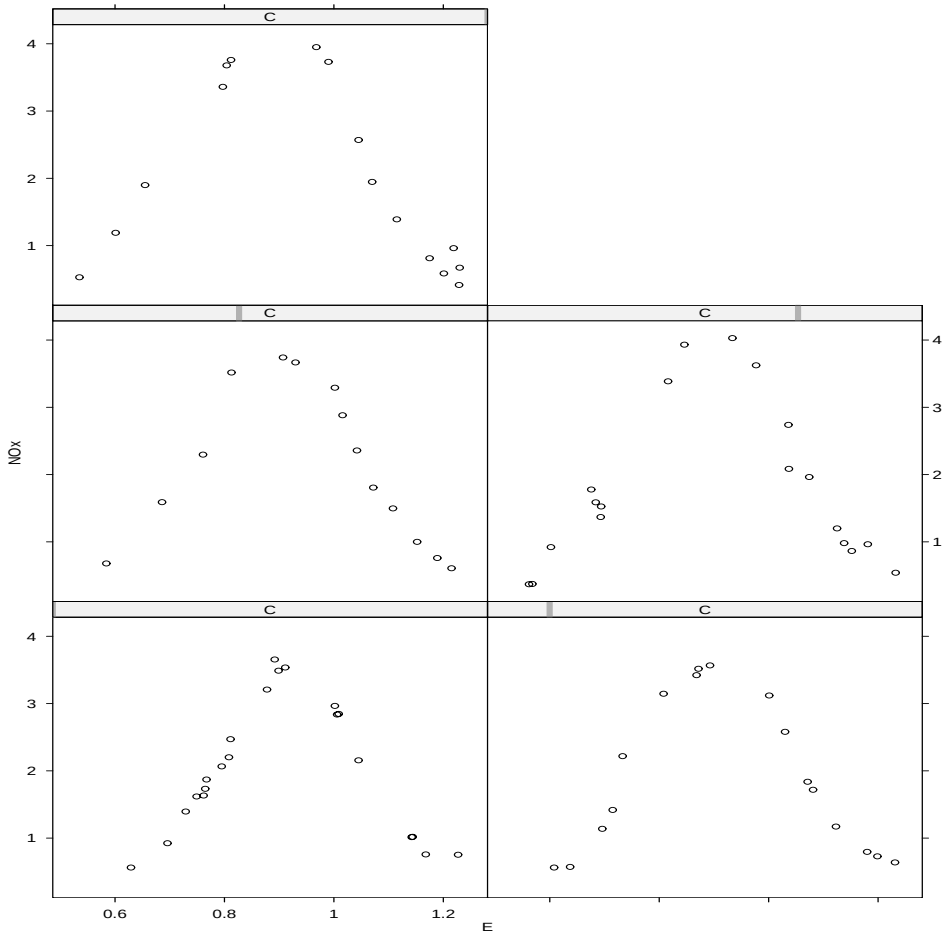

Figure 1: Trellis plot for Question 3.

3. The trellis plot in Figure 1 displays the relationship between two continuous variables NOx and E, and a categorical variable C having 5 levels. The plot was produced by the R code `xyplot(NOx~E|C)`. Which of the following statements is **FALSE**?

   (zz) The plots show that the relationship between NOx and E does not depend on the value of C.

   (1) There is a non-linear relationship between NOx and C.

   (1) The peak value of NOx occurs at about $E = 0.9$.

   (1) The higher the value of C, the lower the value of NOx.

   (1) There is roughly the same number of observations taken at each level of C.

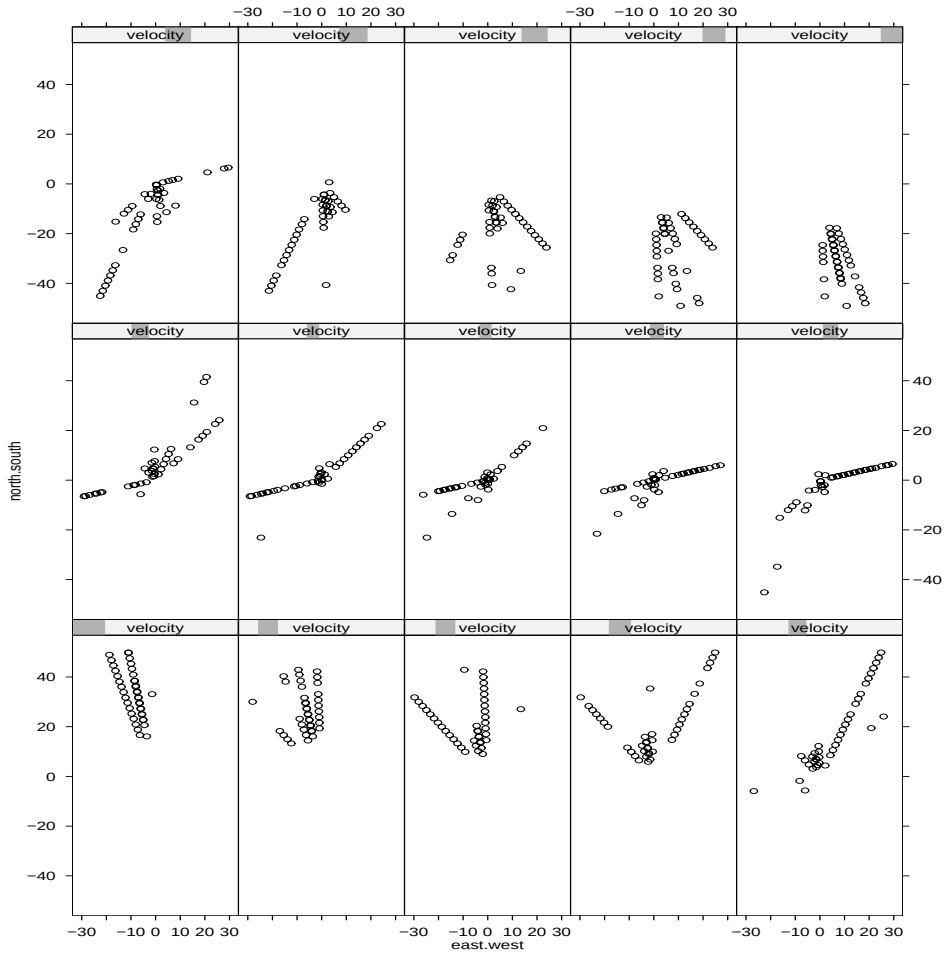

Figure 2: Trellis plot for Question 4.

4. The trellis plot in Figure 2 displays a data set consisting of measurements made on 323 stars in the galaxy NGC7531. On each star 3 measurements are made: (1) the star's position on an north/south axis on a small region of the celestial sphere, (2) the star's position on an east/west axis, and (3) the velocity of the star. The plot was produced with the R-code `xyplot(north.south~east.west|velocity)`. Which of the following statements is **TRUE**?

   (zz) The stars having the highest velocity are in the south-east quadrant of the region.

   (1) The stars having the highest velocity are in the north-west quadrant of the region.

   (1) The stars having the highest velocity are in the north-east quadrant of the region.

   (1) The stars having the highest velocity are in the south-west quadrant of the region.

   (1) The stars having the highest velocity are in the centre of the region.

   Note: Up = north, down = south, right = east, left = west.

5. Which one of the following statements is **TRUE**?

   (zz) If the residual sum of squares is zero, the R^2 must be one.

   (1) Adding an extra variable to a regression model increases the residual sum of squares.

   (1) If the residual sum of squares is one, the points all lie on a plane.

   (1) The adjusted R^2 is always bigger then the R^2.

   (1) To select a model, we pick the model with the smallest residual sum of squares.

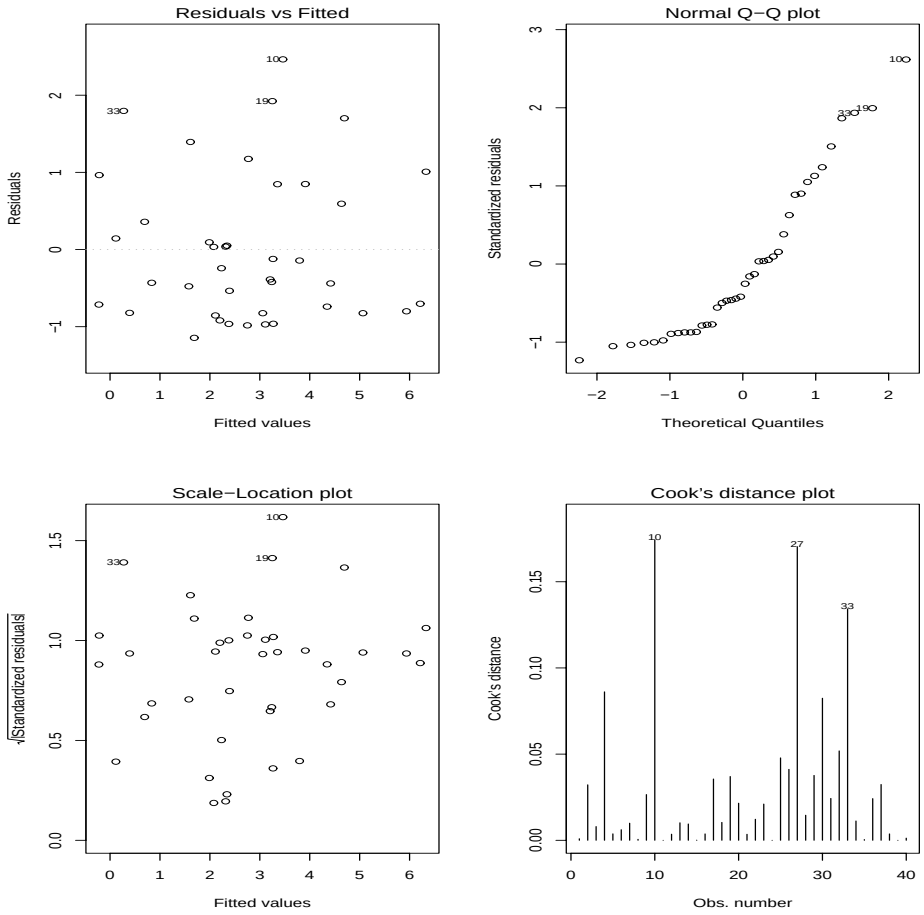

Figure 3: Diagnostic plots for Question 6.

6. Figure 3 shows some diagnostic plots obtained after fitting a regression. What, if anything, is wrong with the regression?

   (zz) The errors are not normally distributed.

   (1) There are outliers in the data.

   (1) The points are not scattered about a plane.

   (1) The error variances are not constant.

   (1) The plots do not indicate problems with the regression.

7. For the data in Question 6, what remedial action should we take?

   (zz) We should transform the response.

   (1) We should delete the outliers.

   (1) We should use logistic regression since the data are not normal.

   (1) We should use weighted least squares.

   (1) We need do nothing, there are no problems with the regression.

8. After fitting a regression, which has response Y and explanatory variables $X1$ and $X2$, we want to predict the value of Y for the values $X1 = 1.5$, $X2 = 30$. We get the the following R output:

```
> model1<-lm(y~X1+X2)
> predict(model1, data.frame(X1=1.5, X2=30), se.fit=T)
$fit
[1] 3.474069
$se.fit
[1] 0.2174274
$df
[1] 37
$residual.scale
[1] 0.821703
> qt(0.975,37)
[1] 2.026192
```

   Which of the following is **TRUE**?

   (zz) A 95% prediction interval for Y is $(1.7518, 5.1962)$.

   (1) A 95% confidence interval for the mean of Y is $(1.7518, 5.1962)$.

   (1) A 95% prediction interval for Y is $(3.0335, 3.9146)$.

   (1) The R^2 is 82%.

   (1) The estimate of σ is 0.2174.

9. In the petrol vapour data discussed in class, the response variable was `hc` and the explanatory variables were `t.temp`, `p.temp`, `t.vp` and `p.vp`. We want to choose a model for these data using the R code and output

```
>  vapour.lm<-lm(hc ~ p.temp + t.temp + p.vp + t.vp, data=vapour.df)
> all.poss.regs(vapour.lm)
      rssp sigma2 adjRsq      Cp     AIC     BIC p.temp t.temp p.vp t.vp
1 1513.347 12.404  0.820  83.366 206.366 212.023      0      0    1    0
1 1962.496 16.086  0.767 143.427 266.427 272.083      1      0    0    0
1 2364.791 19.384  0.719 197.222 320.222 325.879      0      0    0    1
```

```
2 1044.833  8.635  0.875  22.716 145.716 154.201      0      0    1    1
2 1089.397  9.003  0.869  28.675 151.675 160.160      1      0    1    0
2 1320.953 10.917  0.842  59.639 182.639 191.124      1      0    0    1
3  909.304  7.578  0.890   6.593 129.593 140.906      1      0    1    1
3  949.847  7.915  0.885  12.014 135.014 146.328      1      1    1    0
3 1040.640  8.672  0.874  24.155 147.155 158.469      0      1    1    1
4  889.913  7.478  0.892   6.000 129.000 143.142      1      1    1    1
```

Which model is most strongly indicated by this output?

(zz) `hc ~ p.temp + t.temp + p.vp + t.vp`

(1) `hc ~ p.temp + p.vp + t.vp`

(1) `hc ~ t.vp`

(1) `hc ~ t.temp + p.vp + t.vp`

(1) None of the above

10. In a machining process, the rate of metal removal is thought to depend on both the hardness of the material being machined and on the speed at which the machine operates. An experiment was set up to study the relationship between these variables, here called `rate`, `hardness` and `speed`. There were 3 levels of speed chosen: 1000 rpm, 1200 rpm and 1400 rpm, corresponding to the speed settings on the machine. Study the following R code and the resulting output and then pick the **best** interpretation from the alternatives below.

```
> model1<-lm(rate ~ speed + hardness)
> model2<-lm(rate ~ speed * hardness)

> summary(model1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.85158    7.70548  -5.431 0.000207 ***
speed1200     9.17001    2.08340   4.401 0.001061 **
speed1400    18.97645    2.10660   9.008 2.08e-06 ***
hardness      0.93928    0.05636  16.664 3.75e-09 ***

Residual standard error: 3.285 on 11 degrees of freedom
Multiple R-Squared: 0.9745,     Adjusted R-squared: 0.9675
F-statistic: 140.1 on 3 and 11 DF,  p-value: 4.809e-09

> summary(model2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -45.78282   18.62288  -2.458   0.0363 *
speed1200            0.47977   22.88317   0.021   0.9837
speed1400           33.60120   21.91644   1.533   0.1596
hardness             0.96858    0.13833   7.002 6.31e-05 ***
```

CONTINUED

```
speed1200:hardness   0.06283    0.16864   0.373   0.7181
speed1400:hardness  -0.10546    0.16062  -0.657   0.5279

Residual standard error: 3.311 on 9 degrees of freedom
Multiple R-Squared: 0.9788,     Adjusted R-squared: 0.967
F-statistic: 83.11 on 5 and 9 DF,  p-value: 2.963e-07


> anova(model1,model2)
Analysis of Variance Table
Model 1: rate ~ speed + hardness
Model 2: rate ~ speed * hardness
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     11 118.680
2      9  98.646  2    20.034 0.9139 0.4352
```

(zz) For a given hardness, the rate at 1400 rpm is about 9.8 units more the rate at 1200 rpm.

(1) As hardness increases, the rate goes down.

(1) For a given hardness, the rate at 1200 rpm is about 9.1 units less than at 1000 rpm.

(1) At a given hardness, the rate at 1400 rpm is about 33.6 units more than at 1000 rpm.

(1) The rate of increase in rate as hardness increases is more at 1400 rpm than at 1000 rpm.

11. In an experiment to study weight gain in rats, there were two explanatory factors: the level of protein in the feed (at two levels; high and low) and the source of protein (at three levels; beef, cereal and pork). The response was the weight gain of the rats. For each treatment combination, there were 10 rats. R code and some output are shown below:

```
> summary(lm(gain~source*level,data=rats.df))
Coefficients:
                       Estimate Std. Error   t value Pr(>|t|)
(Intercept)            100.00      4.632e+00    21.589  < 2e-16 ***
sourceCereal           -14.10      6.551e+00    -2.152  0.03585 *
sourcePork              -0.50      6.551e+00    -0.076  0.93944
levelLow               -20.80      6.551e+00    -3.175  0.00247 **
sourceCereal:levelLow   18.80      9.264e+00     2.029  0.04736 *
sourcePork:levelLow      0.00      9.264e+00 -3.29e-15  1.00000
```

The mean weight gain of the 10 rats having a low level of protein and cereal as the protein source is

(zz) 83.9

(1) 153.7

(1) 18.8

(1) 100.0

(1) 0.0

12. In an another experiment, this time to study weight gain in chickens, there were three explanatory factors: the level of protein in the feed (at three levels; low, medium and high), the source of protein (at two levels; groundnut and soybean), and the amount of fish solubles (at two levels, low and high). The response was the weight of the chickens at a fixed time after hatching. For each treatment combination, there were 2 chickens. R code and some output are shown below:

```
> anova(lm(chickweight~protein*protlevel*fish, data=chicks.df))
Analysis of Variance Table
Response: chickweight
                       Df  Sum Sq Mean Sq F value   Pr(>F)
protein                 1  373003  373003  3.7334 0.077286 .
protlevel               2  636519  318260  3.1854 0.077679 .
fish                    1 1423014 1423014 14.2429 0.002653 **
protein:protlevel       2  858702  429351  4.2974 0.039134 *
protein:fish            1    7073    7073  0.0708 0.794706
protlevel:fish          2  309421  154710  1.5485 0.252201
protein:protlevel:fish  2   50036   25018  0.2504 0.782453
Residuals              12 1198926   99911
```

The model indicated by this output is

(zz) `protein*protlevel + fish`

(1) `protein:protlevel + protein:fish`

(1) `protein + protlevel + fish`

(1) `protein*protlevel*fish`

(1) `protein + protlevel*fish`

13. In Question 12, which of the following best summarises the chosen model?

    (zz) The effect of changing the level of fish solubles does not depend on the level or source of protein.

    (1) The effect of changing the level of protein depends on the level of fish solubles.

    (1) The effect of changing the level of protein doesn't depend on the source of protein.

    (1) The effect of changing the source of protein doesn't depend on the level of protein.

    (1) The effect of changing the level of fish solubles depends on the level of protein.

14. Suppose that in a regression we have two explanatory factors A and B. We want to compare the two models `y ~ A` (Model 1) and `y ~ A * B` (Model 2), using the `anova` function in R. We obtain

```
> anova(model1,model2)
Analysis of Variance Table
Model 1: y ~ A
Model 2: y ~ A * B
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     57 15932.4
2     54 11586.0  3    4346.4 6.7526 0.0005971 ***
```

    In the output above, what hypothesis is being tested by the p-value 0.0005971?

    (zz) The factor B has no effect on the response.

    (1) There is no interaction between A and B.

    (1) The factor A has no effect on the response.

    (1) The model $A * B$ is a satisfactory model.

    (1) Both factor A and factor B are required in the model.

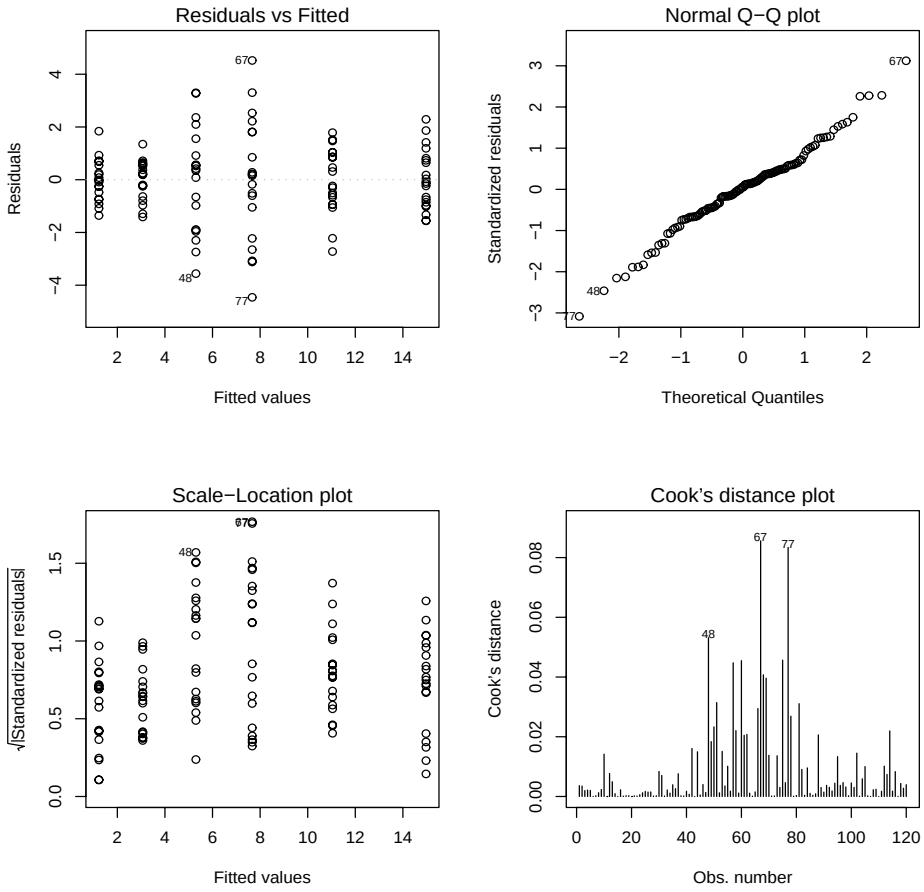

Figure 4: Diagnostic plots for Question 15.

15. When fitting a regression model to a continuous response Y using two categorical explanatory variables A and B, the diagnostic plots shown in Figure 4 were obtained. Each combination of factor levels was observed the same number of times. Which of the following statments is **TRUE**?

    (zz) The diagnostic plots show that the variances of the groups defined by the factor levels of A and B are not equal.

    (1) The diagnostic plots show that there are outliers in the data.

    (1) The diagnostic plots show that the data are not normally distributed.

    (1) The diagnostic plots show that some points have high leverage.

    (1) The diagnostic plots show that nothing is wrong.

16. The correct remedial action in Question 15 is

    (zz) Use weighted least squares.

    (1) Transform the explanatory variables to equalise the variances.

    (1) Remove the outliers.

    (1) Do nothing.

    (1) Transform the response variable to achieve a normal response.

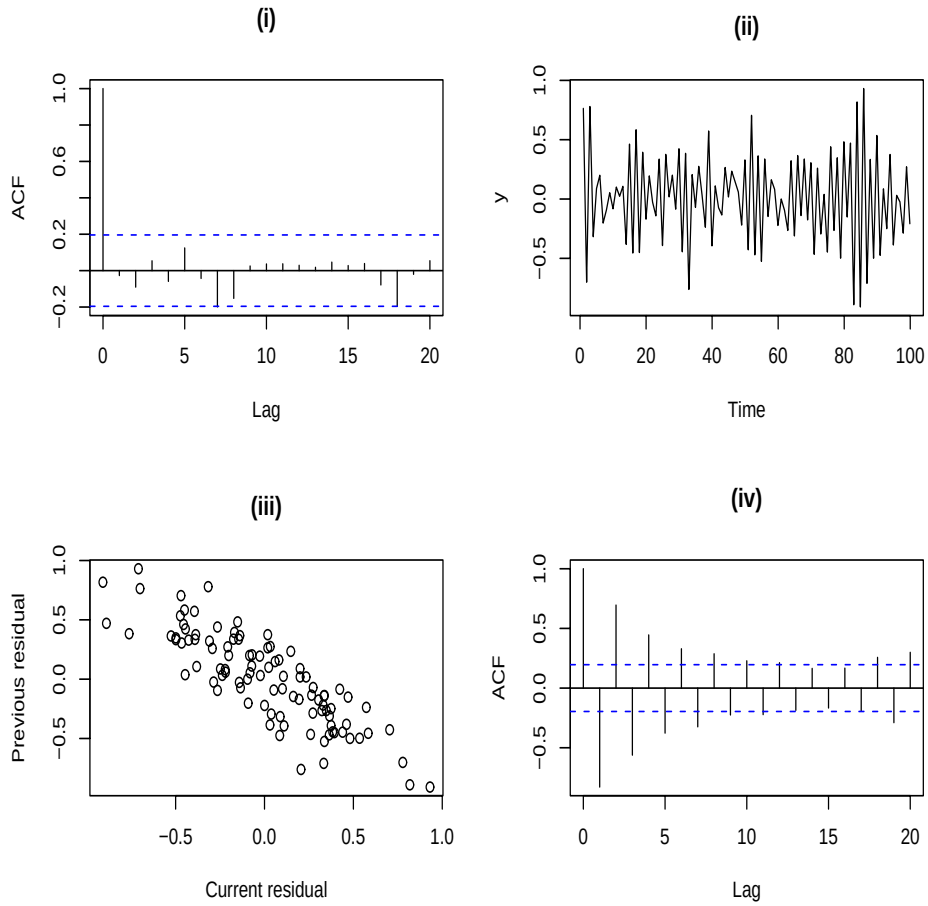

Figure 5: Diagnostic plots for Question 17.

17. Suppose that we have four separate regressions where the data are in time order. In the plots shown in Figure 5, plots (i) and (iv) are correlograms of residuals, plot (ii) is a time series plot of residuals (i.e. a plot of residuals versus time order) and plot (iii) a plot of residuals versus the residual from the previous observation. All the plots refer to separate regressions.

Which of the following statments is **TRUE**?

(zz) In plot (iii), the errors are strongly negatively autocorrelated.

(1) In plot(ii), the errors are strongly positively autocorrelated.

(1) In plot(i), there is evidence of negative autocorrelation in the errors.

(1) In plot (iv), the errors are strongly positively autocorrelated.

(1) In plot (iii), the plot is inconclusive, there is not enough data to decide if the errors are autocorrelated.

18. Suppose that we have a data set with a binary response Y, recording the occurence (1) or non-occurrence (0) of an event E, and a continuous explanatory variable X. We fit a logistic regression and the coefficient of X is estimated as 1.05. Which of the following is **TRUE**?

(zz) The effect of a unit increase in X is to increase our estimate of the odds of E occurring by a factor of 2.857.

(1) The effect of a unit increase in X is to increase our estimate of the odds of E occurring by a factor of 1.05.

(1) The effect of a unit increase in X is to increase our estimate of the log-odds of E occurring by a factor of 1.05.

(1) The effect of a unit increase in X is to increase our estimate of the probability that E occurs by a factor of 2.857.

(1) The effect of a unit increase in X is to increase our estimate of the log-odds of E occurring by 2.857.

19. The data for this question come from a study that investigated the effect of insulin on laboratory mice. The response was whether or not the mice had convulsions when given insulin. The investigators were interested in modelling how the proportion of mice with convulsions varied with the dose applied. They obtained the following data:

Dose (mg), x_i	Number with convulsions, s_i	Number of mice, n_i
3.4	0	33
5.2	5	32
7.0	11	38
8.5	14	37
10.5	18	40
13.0	21	37
18.0	23	31
21.0	30	37
28.0	27	30

Given that

$$\sum_{i=1}^{n}\{s_i \log(s_i/n_i) + (n_i - s_i)\log((n_i - s_i)/n_i)\} = -159.5074,$$

$$\max_{\alpha,\beta} \sum_{i=1}^{n}\{s_i(\alpha + \beta x_i) + n_i \log(1 + \exp(\alpha + \beta x_i))\} = -166.4280, \text{ and}$$

$$\max_{\alpha} \sum_{i=1}^{n}\{r_i\alpha + n_i \log(1 + \exp(\alpha))\} = -217.8824,$$

which of the following is **TRUE**?

(zz) The residual deviance of the logistic model is 13.8412.

(1) The residual deviance of the logistic model is 6.9206.

(1) The null deviance of the logistic model is 58.375.

(1) The null deviance of the logistic model is 116.75.

(1) The residual deviance of the logistic model is 58.375.

20. From the experiment in Question 19, the following output was obtained.

```
> model1<-glm(cbind(s,n-s)~x, family=binomial)
> predict( model1, data.frame(x=20),type="response")
[1] 0.8045766
> predict( model1, data.frame(x=20))
[1] 1.415148
```

Which of the following is **TRUE**?

(zz) The estimated odds of a convulsion when $x = 20$ is 4.117096.

(1) The estimated odds of a convulsion when $x = 20$ is 1.415148.

(1) The estimated log-odds of a convulsion when $x = 20$ is 0.8045766.

(1) The estimated probability of a convulsion when $x = 20$ is 0.1954234.

(1) The estimated odds of a convulsion when $x = 20$ is 2.235750.

21. In a logistic regression for ungrouped data, with 5 explanatory variables, the residual deviance was 49.455 on 75 degrees of freedom, while the null deviance was 83.234 on 80 degrees of freedom. The following output was obtained:

```
> 1-pchisq(83.234, 80)
[1] 0.3802344
> 1-pchisq(49.455,75)
[1] 0.9900576
> 1-pchisq(33.779, 5)
[1] 2.63478e-06
```

Which of the following is **TRUE**?

(zz) Some of the explanatory variables should be retained in the model.

(1) All of the explanatory variables should be retained in the model.

(1) None of the explanatory variables should be retained in the model.

(1) The small residual deviance indicates that the model fits well.

(1) The residual deviance indicates that the model doesn't fit well.

22. In a contingency table, which of the following models expresses the idea that factors A and B are conditionally independent given factor C?

(zz) $A * C + B * C$

(1) $B * C$

(1) $A * B * C$

(1) $A * B$

(1) $A * B + C$

23. In the analysis of two-dimensional contingency tables with I rows and J columns and cell counts y_{ij}, the log-likelihood is

$$\sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} \log \pi_{ij},$$

where π_{ij} is the probability that an individual will be classified into the i, j cell of the table. Which of the following gives the log-likelihood of the maximal model?

(zz) Substituting the frequencies calculated from the table for the π_{ij}.

(1) Substituting the values maximising the log-likelihood under the independence model for the π_{ij}.

(1) Substituting the values minimising the log-likelihood under the independence model for the π_{ij}.

(1) Substituting the value $1/(IJ)$ for the π_{ij}.

(1) Substituting the values calculated from the logistic curve for the π_{ij}.

24. In a three-dimensional contingency table with factors A, B and C, all the two and three factor interactions are insignificant. Which one of the following best summarises the fitted model?

    (zz) The factors A, B and C are mutually independent.

    (1) The factors B and C are independent.

    (1) No interpretation in terms of conditional probability is possible.

    (1) Conditional on C, A and B are independent.

    (1) The factor A is not required in the model.

25. Suppose that we classify a sample from each of five populations on the basis of a categorical factor F having 10 levels. We arrange the data in a data frame with 50 observations and 3 variables `pop`, `F` and `count`, where the value of `pop` and `F` give the population and level, and `count` gives the number of individuals in that sample at that level. We want to test the hypothesis that all five populations have the same distribution of factor F, (i.e. that the conditional distributions of F given the population are identical) using the code `anova(model1, model2)`, where `model1` and `model2` are two models fitted using Posson regression and the data frame described above. What should `model1` and `model2` be?

    (zz) Model 1: `count ~ F + pop`, Model 2: `count ~ F*pop`.

    (1) Model 1: `count ~ F`, Model 2: `count ~ F + pop`.

    (1) Model 1: `count ~ pop`, Model 2: `count ~ F`.

    (1) Model 1: `count ~ 1`, Model 2: `count ~ F + pop`.

    (1) Model 1: `count ~ F`, Model 2: `count ~ F*pop`.

SECTION B

1. Suppose that we are fitting a regression where the response variable and the explanatory variables are all continuous.

   (a) Describe three diagnostic plots that you can use to check if a transformation of the explanatory variables will improve the fit. Which plot do you think gives you the best idea of whether or not to transform?

   (b) Describe a plot that you can use to decide if the response variable needs transforming. Discuss how to interpret the plot, and how the information in the plot can be used to select an appropriate transformation.

   (c) The yield of product in a certain chemical process is thought to depend on the concentration of the raw materials and the flow rate. To explore the relationship between these variables, observations on a chemical plant were made, resulting in a data set with 44 observations on each of three variables, `Yield`, `Conc` and `Flow`. An initial regression model `Yield` $\sim$ `Conc + Flow` was fitted, with the result

```
Call: lm(formula = Yield ~ Con + Flow, data = chem.df)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 112.976406  10.686292  10.572 2.80e-13 ***
Con          -5.186987   0.907002  -5.719 1.09e-06 ***
Flow         -0.006993   0.003495  -2.001   0.0521 .
Residual standard error: 3.08 on 41 degrees of freedom
Multiple R-Squared:  0.46,      Adjusted R-squared: 0.4337
F-statistic: 17.46 on 2 and 41 DF,  p-value: 3.265e-06
```

   The diagnostic plots described in part (a) were examined, and it was decided to fit a cubic polynomial in Flow:

```
Call: lm(formula = Yield ~ Con + poly(Flow, 3), data = chem.df)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    104.4106    11.1228   9.387 1.49e-11 ***
Con             -4.6253     0.9569  -4.833 2.12e-05 ***
poly(Flow, 3)1  -6.0147     2.6610  -2.260 0.029461 *
poly(Flow, 3)2   0.8144     3.1734   0.257 0.798820
poly(Flow, 3)3  10.9140     2.7354   3.990 0.000282 ***
Residual standard error: 2.645 on 39 degrees of freedom
Multiple R-Squared: 0.6213,     Adjusted R-squared: 0.5824
F-statistic: 15.99 on 4 and 39 DF,  p-value: 7.858e-08
```

   Has the inclusion of these extra terms improved the fit? Quote specific parts of the output to justify your answer.

   (d) In Figure 6, we show some diagnostic plots, and following these, some additional output relating to the cubic model. Do the plots and additional output suggest that further action should be taken to improve the fit of this

model? Refer to specific parts of the plots and output to justify your answer.

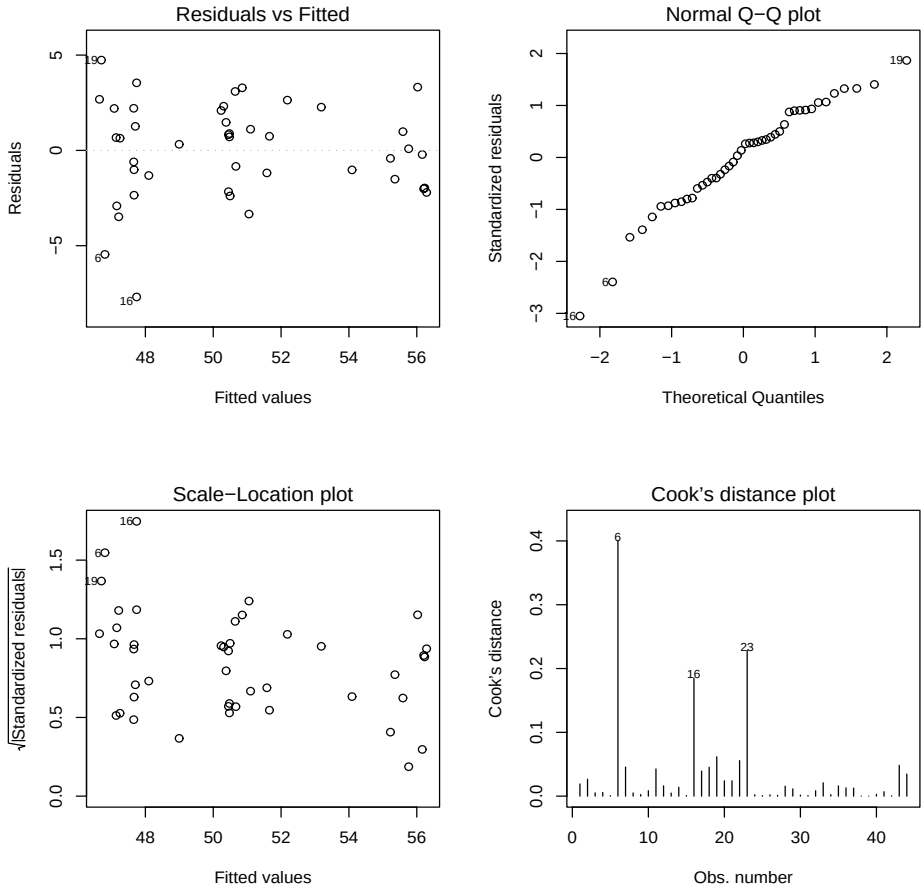

Figure 6: Diagnostic plots for Question B1 (d).

```
Influence measures of

        lm(formula = Yield ~ Con + poly(Flow, 3), data = newchem.df) :

     dfb.1.  dfb.Con dfb.p.F.3.1 dfb.p.F.3.2 dfb.p.F.3.3   dffit cov.r   cook.d    hat inf

5   1.13e-02 -0.01054    -0.04212     0.03451   -0.023535  0.0628 1.390 8.10e-04 0.1826   *
6  -1.59e-01  0.14345     0.89036    -0.77978    0.777012 -1.5128 0.694 4.01e-01 0.2590   *
8   2.08e-02 -0.01909    -0.09422     0.08302   -0.078304  0.1567 1.496 5.03e-03 0.2466   *
16  7.86e-01 -0.80639    -0.21157     0.79705   -0.003507 -1.0884 0.320 1.85e-01 0.0905   *
21  6.61e-03 -0.00274     0.17540     0.18823    0.153191  0.3442 1.405 2.41e-02 0.2303   *
23  2.87e-02 -0.03901    -0.49920    -0.55346   -0.604122 -1.0890 1.237 2.29e-01 0.3262   *
```

HINT: Recall that points are influential if:

i. Cook's D is more than $F_{4,39}(0.5) = 0.8026$,

ii. $|DFBETAS| > 3/\sqrt{n}$,

iii. $|DFFIT| > 3/\sqrt{p/(n-p)}$

iv. $|COVRATIO - 1| > 3p/n$,

v. $h > 3p/n$.

2. In an experiment to investigate the mortality (death) rates of two species of snails under different conditions, groups of snails were held for periods of 2, 3 or 4 weeks in carefully controlled conditions of temperature and relative humidity. At the end of the exposure time the snails were examined to see if they had survived.

   In all, there are four explanatory variables, all treated as factors:

   Species: Snail species A or B

   Exp: Exposure time (2,3 or 4 weeks)

   RH: Relative humidity (4 levels, 60.0%, 65.8%, 70.5%, 75.8%)

   Temp: Temperature, in degrees Celsius (3 levels, 10°, 15°, 20°)

   For each of the $2 \times 3 \times 4 \times 3 = 72$ possible level combinations of these four factors, 20 snails were examined, and the number (out of 20) of survivors recorded. The object of the investigation was to model the probability of death in terms of the explanatory variables.

   (a) This is a complicated experiment with four factors, with many possible interactions. Describe how you would go about selecting a simple model to represent the data. In your answer, you should mention specific R functions. How could you determine if your chosen model was adequate?

   (b) In fact, a simple model with no interactions fits the data quite well. Summary statistics for this model are shown below. Carefully discuss the effect of varying the exposure time on the odds of a snail dying.

```
Call:
glm(formula = cbind(Deaths, N - Deaths) ~ Species + factor(Exp) +
    factor(RH) + factor(Temp), family = binomial, data = snail.df)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.2147     0.3425 -12.307  < 2e-16 ***
SpeciesB         1.2895     0.1618   7.970 1.59e-15 ***
factor(Exp)3     2.2390     0.3030   7.390 1.47e-13 ***
factor(Exp)4     3.1841     0.2990  10.650  < 2e-16 ***
factor(RH)65.8  -0.6140     0.1979  -3.103  0.00192 **
factor(RH)70.5  -1.2561     0.2156  -5.825 5.72e-09 ***
factor(RH)75.8  -1.5872     0.2290  -6.931 4.19e-12 ***
factor(Temp)15   0.5756     0.1981   2.906  0.00366 **
factor(Temp)20   0.9444     0.1942   4.863 1.15e-06 ***

    Null deviance: 366.690  on 71  degrees of freedom
Residual deviance:  30.807  on 63  degrees of freedom
```

   (c) Standard diagnostics on this simple model suggest that instead of treating Exposure Time as a factor, we could treat it as a continuous variable, provided we include it as the reciprocal of time. Below are summary statistics for this model.

The additive model used in part (b) (Model 1 say) treats all four variables as factors, but the reciprocal model (Model 2 say) treats time as a continuous variable. Is Model 2 a reasonable alternative to Model 1? If Model 2 is used in place of Model 1, how does your answer in part (b) change?

Output for Model 2:

```
Call: glm(formula = cbind(Deaths, N - Deaths) ~ Species + I(1/Exp) +
    factor(RH) + factor(Temp), family = binomial, data = snail.df)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.1138     0.3808   5.551 2.84e-08 ***
SpeciesB         1.2931     0.1620   7.980 1.46e-15 ***
I(1/Exp)       -12.4662     1.0929 -11.407  < 2e-16 ***
factor(RH)65.8  -0.6145     0.1980  -3.104  0.00191 **
factor(RH)70.5  -1.2587     0.2159  -5.830 5.54e-09 ***
factor(RH)75.8  -1.5911     0.2293  -6.939 3.96e-12 ***
factor(Temp)15   0.5771     0.1983   2.910  0.00362 **
factor(Temp)20   0.9464     0.1944   4.868 1.13e-06 ***

Null deviance: 366.690  on 71  degrees of freedom
Residual deviance:  31.281  on 64  degrees of freedom

Analysis of Deviance Table
Model 1: cbind(Deaths, N - Deaths) ~ Species + I(1/Exp)
                + factor(RH) + factor(Temp)
Model 2: cbind(Deaths, N - Deaths) ~ Species + factor(Exp)
                + factor(RH) + factor(Temp)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        64    31.2810
2        63    30.8065  1   0.4745    0.4909
```

3. (a) In a three-way table with factors A, B and C, define (in terms of the table probabilities) what it means for A to be independent of B and C. How can this concept be expressed in terms of interactions?

   (b) Under what circumstances can a three-way table be collapsed over one of the factors?

   (c) The table below classifies 174 poliomyelitis cases in Des Moines, Iowa by age of subject, paralytic status and whether or not they had received the Salk vaccine.

		Paralysis	
Age	Vaccine	No	Yes
0-4	Yes	20	14
	No	10	24
5-9	Yes	15	12
	No	3	15
10-14	Yes	3	2
	No	3	2
15-19	Yes	7	4
	No	1	6
20-39	Yes	12	3
	No	7	5
40+	Yes	1	1
	No	3	2

A Poisson regression model was fitted to the table, with the following results:

```
> anova(glm(Count~ Paralysis*Vaccine*Age, family=poisson, data=Salk.df),
      test="Chisq")
Analysis of Deviance Table
                     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                    23    123.171
Paralysis             1    0.143        22    123.028     0.705
Vaccine               1    0.967        21    122.061     0.326
Age                   5   91.216        16     30.845 3.730e-18
Paralysis:Vaccine     1   14.229        15     16.616 1.619e-04
Paralysis:Age         5    7.996        10      8.620     0.156
Vaccine:Age           5    5.181         5      3.440     0.394
Paralysis:Vaccine:Age 5    3.440         0  1.554e-15     0.633
```

Does the relationship between Paralysis and Vaccine depend on age? Give a reason for your answer.

(d) Use the output below to decide if there is an association between paralysis and whether or not a person has been vaccinated. Is it legitimate to disregard age?

```
> anova(glm(Count~ Paralysis*Vaccine, family=poisson, data=Salk.df),
      test="Chisq")
Analysis of Deviance Table
                 Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                23    123.171
Paralysis         1    0.143        22    123.028     0.705
```

```
                  Df Deviance Resid. Df Resid. Dev      Pr(>F)
Vaccine            1    0.967        21    122.061       0.326
Paralysis:Vaccine  1   14.229        20    107.833 0.0001619
```
