THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2004
Campus: City

STATISTICS

Advanced Statistical Modeling
Topics in Statistics 3

(Time allowed: THREE hours)

INSTRUCTIONS

SECTION A: Multiple Choice (40 marks)

- Answer **ALL 20** 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 penalized.

SECTION B (60 marks)

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

SECTION A

1. A data set consist of measurements on three variables X, Y and Z. The variables X and Y are continuous and Z is categorical. Which of the following plots would you expect to give the **best picture** of the relationship between the variables?

   (1) A barchart, with bars corresponding to the frequencies of X and Y.

   (2) A coplot corresponding to the formula `Y ~ Z | X`.

   (3) A coplot corresponding to the formula `Y ~ X | Z`.

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

   (5) A trellis plot consisting of panels corresponding to values of X, and each panel containing a dot plot.

2. The coplot in Figure 1 displays the relationship between three continuous variables Y, $X1$ and $X2$. The plot was produced by the R code `coplot(Y~X1|X2)`. Which of the following statements is **TRUE**?

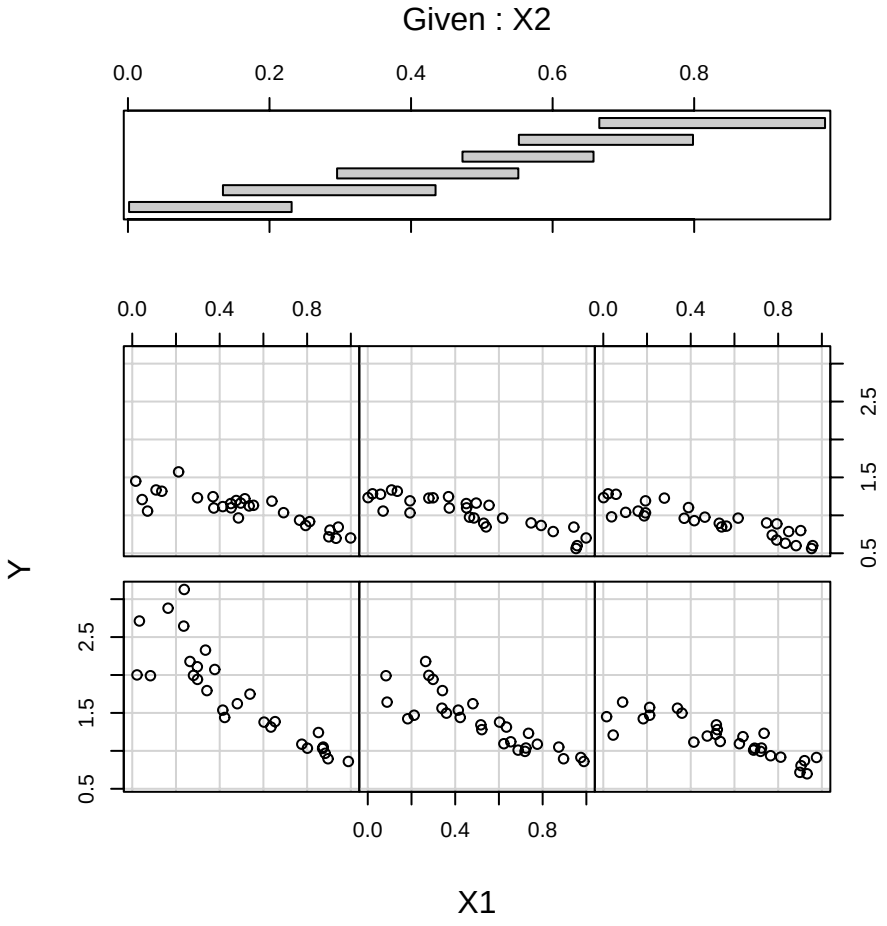

Figure 1: Coplot plot for Question 2.

(1) The plot shows that the data are not planar.

(2) The panels show linear relationships between the variables.

(3) A pairs plot would have been better for checking if the data are planar.

(4) A linear model fitting non-parallel lines would be appropriate for these data.

(5) The coplot is inappropriate because the explanatory variables are continuous.

3. An industrial experiment was run to explore the effect of two factors on the life of a machine tool. The factors were

   (a) `speed`: Cutting speed (125 rpm, 150 rpm, 175 rpm),

   (b) `angle`: Cutting angle (15 degrees, 20 degrees, 25 degrees).

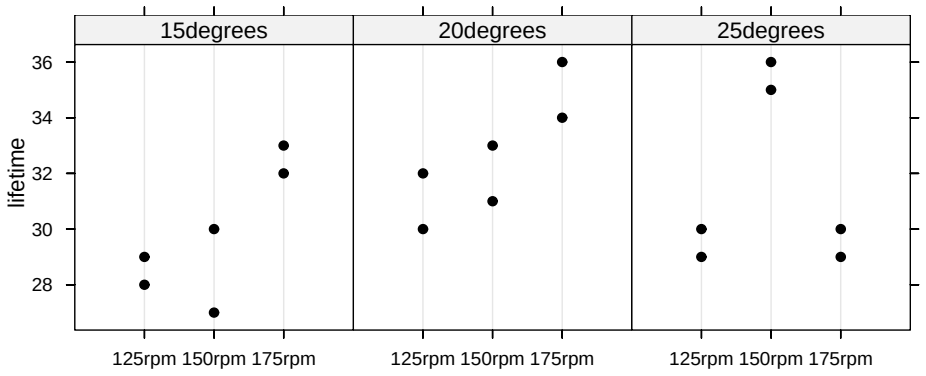

Figure 2: Trellis plot for Question 3.

The trellis plot in Figure 2 displays data from the experiment. Each treatment combination was observed twice. Which of the following statements is **TRUE**?

(1) Slow speeds result in longer lifetimes.

(2) The shortest lifetime is at 20 degrees.

(3) There is no evidence of a angle effect.

(4) The lifetimes for 15 degree angles are longer than for 20 degree angles.

(5) There is evidence of interaction in these data.

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

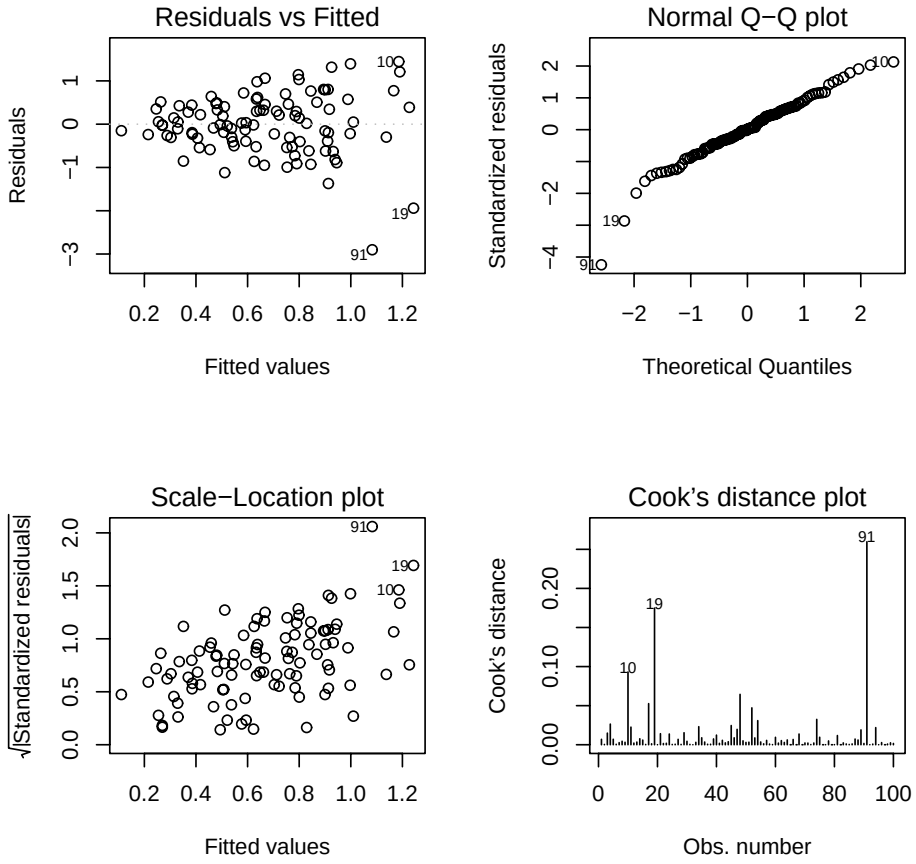

Figure 3: Diagnostic plots for Question 4.

(1) There are outliers in the data.

(2) The errors are not normally distributed.

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

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

(5) The error variances are not constant.

5. For the data in Question 4, what remedial action should we take?

   (1) No action is necessary as there are no problems with the regression.
   (2) We should use logistic regression since the data are not normal.
   (3) We should transform the response.
   (4) We should transform the independent variables.
   (5) We should delete the outliers.

6. Which of the following is **FALSE**?

   (1) The hat matrix diagonals always lie between 0 and 1.
   (2) The hat matrix diagonal measures how much leverage a point has.
   (3) The hat matrix diagonal measures how outlying a point is as far as the explanatory variables are concerned.
   (4) The hat matrix diagonal measures the extent to which a data point is an outlier.
   (5) The hat matrix diagonal measures the potential influence of the point on its fitted value.

7. The car data set was discussed in a tutorial. In this data set, various characteristics of several different makes of car were measured, along with the price of the cars. We want to choose a model for these data to use as an equation for predicting the price of other types of car. The following output was obtained:

```
> cars.lm<-lm(log(PRICE) ~  WEIGHT + CITY + DISP + COMP + HP +
                                   TORQ+ TRANS + CYL, data=cars.df)
> summary(cars.lm)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.732e+00  4.204e-01  18.391  < 2e-16 ***
WEIGHT       4.193e-04  7.225e-05   5.804 4.75e-08 ***
CITY        -2.126e-03  7.017e-03  -0.303  0.76238
DISP        -4.435e-04  5.724e-05  -7.748 2.42e-12 ***
COMP         1.425e-02  3.317e-02   0.430  0.66815
HP           6.314e-03  9.317e-04   6.777 3.95e-10 ***
TORQ         3.444e-03  1.383e-03   2.490  0.01403 *
TRANS       -9.478e-02  4.859e-02  -1.950  0.05329 .
CYL          8.123e-02  2.533e-02   3.207  0.00169 **
---
```

```
> all.poss.regs(cars.lm)
    rssp sigma2 adjRsq      Cp     AIC     BIC WEIGHT CITY DISP COMP HP TORQ TRANS CYL
1 10.615  0.078  0.843 153.481 289.481 292.408      0    0    0    0  1    0     0   0
1 14.956  0.110  0.778 271.052 407.052 409.979      0    0    0    0  0    1     0   0
1 21.903  0.161  0.675 459.190 595.190 598.117      0    0    0    0  0    0     0   1
2  8.618  0.064  0.871 101.391 237.391 243.246      1    0    0    0  1    0     0   0
2  9.771  0.072  0.854 132.612 268.612 274.467      0    0    1    0  1    0     0   0
2  9.948  0.074  0.851 137.417 273.417 279.272      0    0    1    0  0    1     0   0
3  5.496  0.041  0.917  18.844 154.844 163.625      1    0    1    0  1    0     0   0
3  7.618  0.057  0.885  76.314 212.314 221.096      0    0    1    0  0    1     0   1
3  7.697  0.057  0.884  78.462 214.462 223.244      1    0    0    0  1    1     0   0
4  5.193  0.039  0.921  12.647 148.647 160.356      1    0    1    0  1    0     0   1
4  5.291  0.040  0.920  15.301 151.301 163.010      1    0    1    0  1    0     1   0
4  5.348  0.040  0.919  16.845 152.845 164.554      1    0    1    0  1    1     0   0
5  4.914  0.037  0.925   7.094 143.094 157.730      1    0    1    0  1    1     0   1
5  5.000  0.038  0.924   9.414 145.414 160.051      1    0    1    0  1    0     1   1
5  5.179  0.039  0.921  14.270 150.270 164.906      1    0    1    1  1    0     0   1
6  4.773  0.036  0.926   5.266 141.266 158.829      1    0    1    0  1    1     1   1
6  4.908  0.037  0.924   8.919 144.919 162.483      1    1    1    0  1    1     0   1
6  4.911  0.037  0.924   8.992 144.992 162.556      1    0    1    1  1    1     0   1
7  4.767  0.037  0.926   7.092 143.092 163.583      1    0    1    1  1    1     1   1
7  4.770  0.037  0.926   7.185 143.185 163.675      1    1    1    0  1    1     1   1
7  4.904  0.038  0.924  10.804 146.804 167.295      1    1    1    1  1    1     0   1
8  4.763  0.037  0.926   9.000 145.000 168.418      1    1    1    1  1    1     1   1
```

Which model is most strongly indicated by this output?

(1) WEIGHT + DISP + HP

(2) CITY + COMP

(3) WEIGHT + DISP + HP + TORQ + TRANS + CYL

(4) WEIGHT + DISP + HP + TORQ + CYL

(5) WEIGHT + CITY + DISP + COMP + HP +TORQ + TRANS + CYL

8. In the education data studied in class, data were given on each of the 50 states of the United States. We fitted a model relating the reciprocal of spending on education per capita (`educ`) to the explanatory variables per capita income (`percap`) and number of residents under 18 per 1000 residents (`under18`). Use the output below to decide if some states are having an undue influence on the analysis.

```
> educ.lm<-lm(I(1/educ)~percap+under18, data=educ.df)
> influence.measures(educ.lm)
Influence measures of
         lm(formula = I(1/educ) ~ percap + under18, data = educ.df) :

      dfb.1_ dfb.prcp  dfb.un18    dffit cov.r   cook.d    hat
6  -0.016411  0.13616 -0.034595  0.17673 1.158 1.06e-02 0.0960
10 -0.076648  0.20169  0.026691  0.43607 0.707 5.60e-02 0.0257
29 -0.069398  0.02278  0.071737 -0.07930 1.195 2.14e-03 0.1103
44  0.218843  0.02846 -0.281892 -0.32528 1.241 3.56e-02 0.1690
50  0.407739 -0.25851 -0.384531 -0.42573 1.587 6.13e-02 0.3429
```

CONTINUED

Which of the following is **TRUE**?

(1) Point 50 is potentially influential since it has a large hat matrix diagonal.

(2) Point 10 is having no effect on the estimation of the standard errors.

(3) Point 6 is influential.

(4) Point 29 is having a big effect on the estimation of the coefficient for "under18".

(5) Point 44 is influential because it has a large Cook's distance.

HINT: Recall that points are influential if:

(a) Cook's D is more than $F_{3,47}(0.1) = 0.1940$,

(b) $|DFBETAS| > 3/\sqrt{n}$,

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

(d) $|COVRATIO - 1| > 3p/n$,

(e) $h > 3p/n$.

9. 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 $X1 = 16$, $X2 = 7.5$. We get the the following R output:

```
> q11.lm<-lm(y~X1 + X2)
> predict(q11.lm, data.frame(X1=16,X2=7.5), se=T)
$fit
[1] 34.15078
$se.fit
[1] 0.3367116
$df
[1] 97
$residual.scale
[1] 2.128866
> qt(0.975,97)
[1] 1.984723
```

Which of the following is **TRUE**?

(1) A 95% confidence interval for the mean of Y is $(29.926, 38.376)$.

(2) A 95% prediction interval for Y is $(29.873, 38.429)$.

(3) A 95% confidence interval for the mean of Y is $(29.873, 38.429)$.

(4) The estimate of σ is 1.459.

(5) A 95% prediction interval for Y is $(33.483, 34.819)$.

10. In a chemical process, the yield of the process is thought to depend on both the temperature and pressure at which the reaction takes place. An experiment was set up to study the relationship between these variables, here called `yield`, `temp` and `pressure`. There were 3 levels of temperature: Low, Medium and High, and three levels of pressure 200 psi, 215 psi and 230 psi. (Note in the output below, "Low" is the baseline level for temperature.) Each of the possible 9 treatment combinations was used twice in the experiment, generating 18 observations.

    The model `yield` $\sim$ `temp*pressure` was fitted to the data, and a fragment of the resulting output is shown below.

```
Coefficients:
                             Estimate
(Intercept)                    90.300
tempMedium                     -0.100
tempHigh                        0.300
pressure215.psi                 0.350
pressure230psi                  0.000
tempMedium:pressure215.psi      0.000
tempHigh:pressure215.psi       -0.100
tempMedium:pressure230psi      -0.200
tempHigh:pressure230psi        -0.350
```

    Which of the following is **TRUE**?

    (1) The mean yield at pressure 230 psi is estimated as 0.000.

    (2) The mean yield at medium temperature is estimated as 90.200.

    (3) The mean yield at high temperature and pressure 200 psi is estimated as 0.300.

    (4) The mean yield at medium temperature and pressure 215 psi is estimated as 90.550.

    (5) The mean yield at high temperature and pressure 230 psi is estimated as 90.100.

11. In the experiment described in Question 10, the following ANOVA table was obtained:

```
Response: yield
              Df  Sum Sq Mean Sq F value    Pr(>F)
temp           2 0.30111 0.15056  8.4687 0.0085392 **
pressure       2 0.76778 0.38389 21.5937 0.0003673 ***
temp:pressure  4 0.06889 0.01722  0.9687 0.4700058
Residuals      9 0.16000 0.01778
```

Which of the following is **FALSE**?

(1) There is strong evidence of interaction in these data.

(2) There is no effect of changing pressure from 200 psi to 230 psi.

(3) The effect of changing pressure from 215 psi to 230 psi is to decrease the yield by 0.350.

(4) The effect of changing temperature from medium to low is to increase the yield by 0.100.

(5) The effect of changing temperature from medium to high is to increase the yield by 0.400.

12. Suppose we want to fit a linear model to data consisting of a continuous response variable Y and covariates X, Z, A and B , where X and Z are continuous and A and B are factors.

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

(1) The model `Y ~ A*X + A*Z` fits separate planes though the data for each level of A.

(2) The model `Y ~ A*B*X*Z` fits separate planes though the data with the slopes and intercepts depending on the levels of A and B.

(3) The model `Y ~ A*X + Z` fits separate planes though the data for each level of A, with the coefficient of X being the same for each plane.

(4) The model `Y ~ A + X + Z` fits a single plane though the data.

(5) The model `Y ~ A*B + X + Z` fits separate planes though the data with the slopes depending on the levels of A and B.

13. In a Poisson regression of a count response Y on a continuous explanatory variable X, using the R code `glm(Y~X, family=poisson)`, the regression coefficient is 0.3. Which of the following is **TRUE**?

(1) The effect of a unit increase in X is to increase our estimate of the mean count by about 30%.

(2) The effect of a unit increase in X is to increase our estimate of the log-odds by approximately 1.35.

(3) The effect of a unit increase in X is to increase our estimate of the log-odds by about 35%.

(4) The effect of a unit increase in X is to increase our estimate of the the mean count by an amount 0.3.

(5) The effect of a unit increase in X is to increase our estimate of the mean count by about 35%.

14. An experiment has been done to test the effect of an insecticide on a certain species of beetle. For each of eight pre-specified dosages, a number (n) of beetles were exposed to the insecticide and the number dying (r) was recorded. The following data were obtained:

```
 logdose  r  n
  1.6907  5 59
  1.7242 11 60
  1.7552 31 62
  1.7842 37 56
  1.8113 48 63
  1.8369 55 59
  1.8610 61 62
  1.8839 59 60
```

A logistic model of the form

$$\log \frac{\pi}{1-\pi} = \alpha + \beta \times \text{logdose}$$

(where π is the probability a beetle will die) was fitted to the data with the following results:

```
Call: glm(formula = cbind(r, n - r) ~ logdose, family = binomial)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -57.375      5.018  -11.43   <2e-16 ***
logdose       32.588      2.834   11.50   <2e-16 ***
---
Null deviance: 253.1354  on 7  degrees of freedom
Residual deviance:   4.3618  on 6  degrees of freedom
AIC: 36.985
> 1-pchisq(4.3618,6)
[1] 0.6278396
```

Which of the following is **FALSE**?

(1) The output indicates that death is more likely with increasing dose.

(2) The model predicts that the odds of death for a beetle subjected to a logdose of 1.6907 is 0.1024.

(3) The null deviance indicates that the fit of the logistic model is poor.

(4) The model predicts that the probability of death for a beetle subjected to a logdose of 1.6907 is 0.0929.

(5) The data are grouped data.

15. In the beetle example in Question 14, the following additional output was obtained:

```
> predict(q18.glm, data.frame(logdose=1.8113), se=T)
$fit
[1] 1.65089
$se.fit
[1] 0.1747074
> qnorm(0.975)
[1] 1.959964
```

Which of the following is **FALSE**?

(1) The point estimate for the probability of death at log dose 1.8113 is 0.839.

(2) From the data given, it is not possible to calculate the confidence interval for the probability of death when the log dose is 1.8113.

(3) A 95% confidence interval for the odds of death at log dose 1.8113 is (3.701, 7.340).

(4) The point estimate for the log-odds of death at log dose 1.8113 is 1.651.

(5) A 95% confidence intervlal for the log-odds of death at log dose 1.8113 is (1.308, 1.993).

16. The beetle experiment in Questions 14 and 15 was repeated, using a modified form of the insecticide (treatment 1) as well as the previous form (treatment2), but keeping the same dosages. We now have data

logdose	r	n	treat
1.6907	7	59	1
1.7242	17	60	1
1.7552	21	62	1
1.7842	41	56	1
1.8113	52	63	1
1.8369	57	59	1
1.8610	61	62	1
1.8839	59	60	1
1.6907	9	64	2
1.7242	21	59	2
1.7552	41	63	2
1.7842	48	64	2
1.8113	55	59	2
1.8369	54	58	2
1.8610	57	57	2
1.8839	64	64	2

We obtained the following "anova".

```
Analysis of Deviance Table
Model: binomial, link: logit
Response: cbind(r, n - r)
Terms added sequentially (first to last)
              Df   Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                15     511.23
logdose        1     487.62        14      23.61 4.694e-108
treat          1       9.39        13      14.22  2.184e-03
logdose:treat  1  3.073e-03        12      14.21       0.96
> 1-pchisq(14.21,12)
[1] 0.2874996
```

Which of the following is **TRUE**?

(1) All of the main effects and interactions should be retained in the model.

(2) If we include `treat` as a variable, we don't need `logdose`.

(3) The test statistic 3.073e-03 is comparing the full model to the null model.

(4) On the log-odds scale, the effect of changing from treatment 1 to treatment 2 is different at different dosages.

(5) The model `cbind(r, n - r)` $\sim$ `logdose + treat` seems to fit well.

17. In a logistic regression for ungrouped data, with 3 explanatory variables, the residual deviance was 134.986 on 96 degrees of freedom, while the null deviance was 137.989 on 99 degrees of freedom. The following additional output was obtained:

```
> 1-pchisq(137.989,99)
[1] 0.005884545
>  1-pchisq(134.986,96)
[1] 0.005384848
>  1-pchisq(3.003,3)
[1] 0.3911629
```

Which of the following is most strongly indicated?

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

(2) The small residual deviance p-value indicates that the model fits well.

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

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

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

18. In the Florida murder data discussed in class, defendants convicted of murder were classified according to three factors; namely race of the defendant (`defendant`), race of the victim (`victim`), and whether or not the death penalty was imposed (`dp`). The following anova was obtained:

```
> murder.glm<-glm(count~defendant*dp*victim,family=poisson,
                        data=murder.df)
> anova(murder.glm, test="Chisq")
Analysis of Deviance Table.
Model: poisson, link: log, Response: count
Terms added sequentially (first to last)
                    Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                    7     774.73
defendant            1     0.22         6     774.51      0.64
dp                   1   443.51         5     331.00 1.861e-98
victim               1    64.10         4     266.90 1.183e-15
defendant:dp         1     0.43         3     266.47      0.51
defendant:victim     1   254.15         2      12.32 3.230e-57
dp:victim            1    12.30         1       0.02 4.535e-04
defendant:dp:victim  1     0.02         0  3.997e-15      0.89
> summary(murder.glm)
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)              2.5649     0.2774   9.248  < 2e-16 ***
defendantw              -2.5649     1.0377  -2.472  0.01345 *
dpno                     2.7081     0.2864   9.454  < 2e-16 ***
victimw                  0.5705     0.3470   1.644  0.10012
defendantw:dpno          0.2364     1.0652   0.222  0.82438
defendantw:victimw       3.0930     1.0705   2.889  0.00386 **
dpno:victimw            -1.1896     0.3675  -3.237  0.00121 **
defendantw:dpno:victimw  0.1613     1.1032   0.146  0.88375
---
Null deviance: 7.7473e+02  on 7  degrees of freedom
Residual deviance: 3.9968e-15  on 0  degrees of freedom
```

Which of the following independence models is indicated by this output?

(1) Victim's race and death penalty are conditionally independent, given defendant's race.

(2) Defendant's race and death penalty are conditionally independent, given victim's race.

(3) The association between defendant's race and death penalty is the same for black and white victims.

(4) All three factors are independent.

(5) Defendant's race and death penalty are independent of victim's race.

CONTINUED

19. In a three-dimensional contingency table with factors A, B and C, we want to test the hypothesis that factor A is independent of factors B and C, using an R statement of the form, `anova(model1, model2)`. What should the formulas defining model 1 and model 2 be? The R vector `count` contains the cell counts.

    (1) Model 1: `count ~ A + B + C,` Model 2: `count ~A*B*C.`
    (2) Model 1: `count ~ A*B + A*C,` Model 2: `count ~A + B*C.`
    (3) Model 1: `count ~ 1,` Model 2: `count ~A*B*C.`
    (4) Model 1: `count ~ A + B*C,` Model 2: `count ~A*B*C.`
    (5) Model 1: `count ~ A + B + C,` Model 2: `count ~A + B*C.`

20. The data below are from a famous data set, which records the number of deaths per year from horse kicks in corps of the Prussian army. The data have been grouped for your convenience:

Number of deaths	0	1	2	3	4
Counts	109	65	22	3	1

    We want to check if these data follow a Poisson distribution. We get the following R output:

```
> ndeaths<-c(0,1,2,3,4)
> counts<-c(109,65,22,3,1)
> mean.deaths<-sum(ndeaths*counts)/sum(counts)
> poiss.probs<-dpois(0:4,mean.deaths)
> logL1<-sum(counts *log(pois.probs))
> logL1
[1] -206.1067
> freqs<-counts/sum(counts)
> logL2<-sum(counts *log(freqs))
> logL2
[1] -205.6726
> logL3<-sum(counts *log(1/5))
> logL3
[1] -321.8876
> 1-pchisq(0.8682,3)
[1] 0.8330943
```

    Which of the following is **FALSE**?

    (1) The residual deviance for the Poisson model is 0.8682.
    (2) The test statistic for testing the hypothesis that the true distribution is Poisson is 232.43 - 0.87 = 231.56.
    (3) The Poisson model fits the data well.
    (4) The test statistic for testing the hypothesis that the true distribution is uniform (i.e. the probabilities of 0,1,2,3,4 deaths are all the same) is 232.43.
    (5) The null deviance for the Poisson model is 232.43.

CONTINUED

SECTION B

1. The data for this question are taken from an experiment which investigated the ascorbic acid content of cabbages. Three variables are thought to influence the ascorbic acid content:

   - The size of the cabbage head (measured by the head weight, variable `HeadWt`)
   - the genetic line or cultivar (there were two lines, 39 and 52, recorded as the factor `Line`)
   - the planting date (three dates were considered, labeled as 16, 20 and 21, recorded as the factor `Date`).

   Each row in the data set refers to an individual cabbage, and there are 10 observations for each of the six factor level combinations, for a total of 60 observations. The response variable is `Ascorbic`, the acsorbic acid content of the cabbage head.

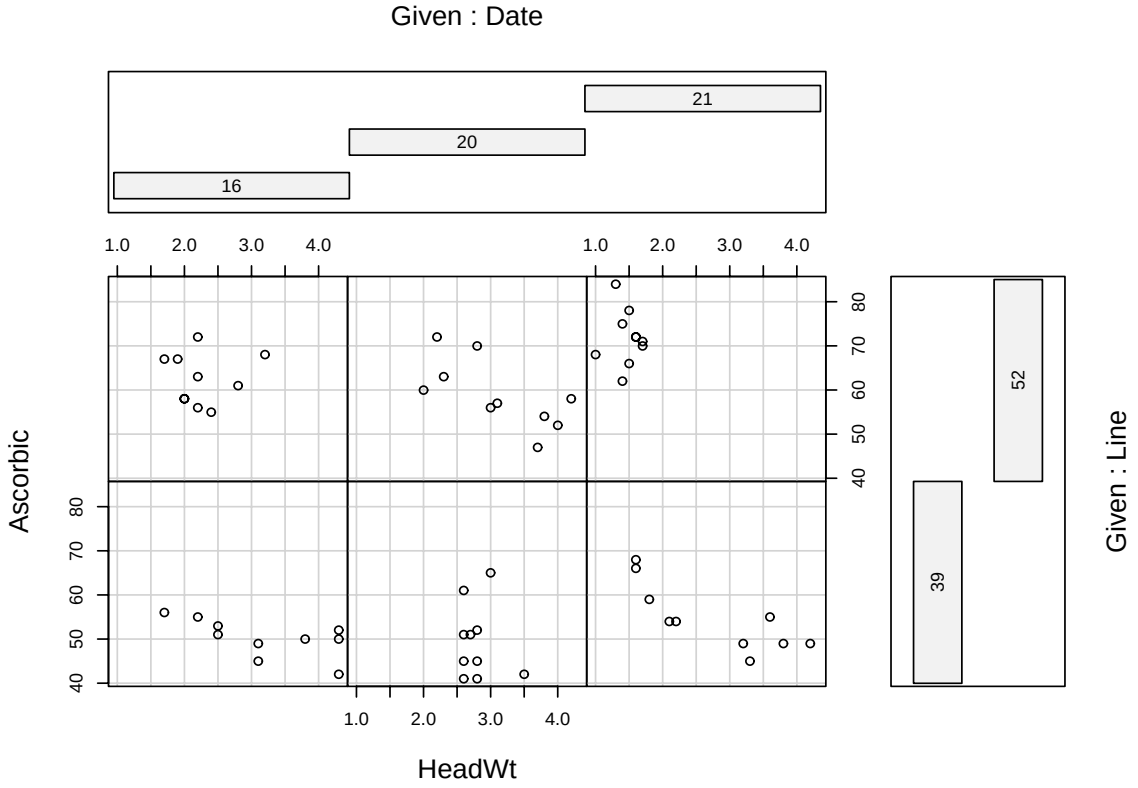

Figure 4: Coplot for Question B1.

   (a) In Figure 4, we show a coplot of the data. Based on this coplot only, I would try a model of the form `Ascorbic` $\sim$ `HeadWt + Date*Line`. Do you agree with this choice? What aspects of the plot are you basing your decision on?

(b) A model was fitted, producing the output below.

```
> cabbage.lm<-lm(Ascorbic~HeadWt*Date*Line, data=cabbage.df)
> anova(cabbage.lm)
Analysis of Variance Table Response: Ascorbic
                 Df  Sum Sq Mean Sq F value    Pr(>F)
HeadWt            1 2630.53 2630.53 68.3538 8.704e-11 ***
Date              2  101.17   50.58  1.3144    0.2781
Line              1 1303.36 1303.36 33.8675 4.724e-07 ***
HeadWt:Date       2  116.50   58.25  1.5136    0.2304
HeadWt:Line       1   15.95   15.95  0.4145    0.5227
Date:Line         2    1.32    0.66  0.0171    0.9830
HeadWt:Date:Line  2   24.78   12.39  0.3220    0.7263
Residuals        48 1847.24   38.48
```

Does this model confirm my choice in part (a)? Give reasons, referring to specific parts of the output.

(c) The diagnostic plots in Figure 5 on the next page refer to the model fitted in part (b). Point 30 seems to have a large Cook's distance. The following diagnostics on this point were also obtained.

```
       dfb.1_        dfb.HdWt      dfb.Dt20      dfb.Dt21      dfb.Ln52
-1.937071e-15    1.151518e-15  3.205382e+00  7.941197e-16  1.478540e-15
   dfb.HdW:D20     dfb.HdW:D21    dfb.HW:L52   dfb.D20:L52   dfb.D21:L52
-3.388566e+00   -9.064311e-16 -1.576766e-15 -2.729959e+00 -6.087945e-16
dfb.HW:D20:L52  dfb.HW:D21:L52         dffit         cov.r        cook.d
 2.788916e+00   4 .186369e-16 -3.773335e+00  2.680588e+00  1.126440e+00
          hat
 8.000000e-01
```

Can you explain why the point is so influential? Can you identify the point on the coplot (Figure 4)?

NOTE: A copy of the coplot in Figure 4 is attached to the back of the examination paper. Detach the plot and mark the point you think is point 30 on the plot. Hand in the plot along with your answer.

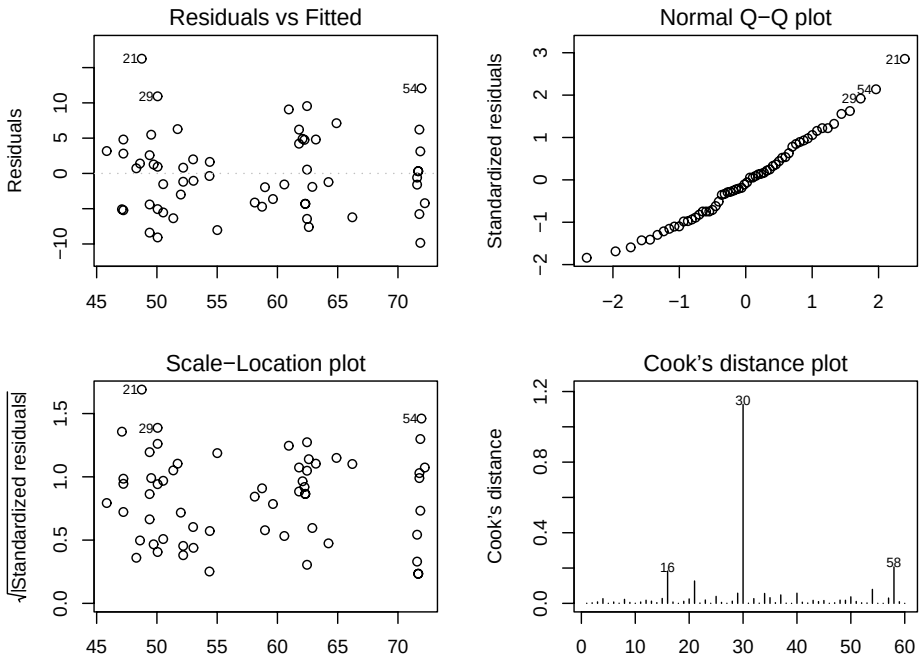

Figure 5: Diagnostic plots for Question B1(c).

(d) The model `Ascorbic HeadWt+Date*Line` was fitted to the same data(i.e. no points were deleted), The diagnostic plots are now satisfactory. Can you suggest why?

(e) The following output from the fit in part (d) was obtained:

```
> cabbage2.lm<-lm(Ascorbic~HeadWt+Date*Line, data=cabbage.df)
> summary(cabbage2.lm)
Call:
lm(formula = Ascorbic ~ HeadWt + Date * Line, data = cabbage.df)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     64.618      4.305  15.011  < 2e-16 ***
HeadWt          -4.503      1.210  -3.721 0.000481 ***
Date20          -2.611      2.768  -0.943 0.349899
Date21           2.519      2.781   0.906 0.369254
Line52           8.058      2.948   2.733 0.008510 **
Date20:Line52    2.838      4.138   0.686 0.495742
Date21:Line52    3.224      3.884   0.830 0.410209
---
Residual standard error: 6.105 on 53 degrees of freedom
Multiple R-Squared: 0.6731,     Adjusted R-squared: 0.636
F-statistic: 18.18 on 6 and 53 DF,  p-value: 2.502e-11
```

Are the two factors having an effect on the ascorbic acid content? If so, how? Does the effect depend on the head size?

2. Framingham is an industrial town located approximately 30 km from Boston. In 1948 a study was begun with the aim of identifying factors that are related to the occurrence of coronary heart disease (CHD). At the start of the study, a large proportion of the town's inhabitants were examined for the presence of CHD. Measurements were made on a number of potential risk factors. The individuals who were found to be free of CHD at that time were followed up for twelve years and those who developed CHD during that period were identified. The following dataset was extracted from that data and relates the proportions developing CHD to the initial serum cholesterol level (mg per 100 ml) of these individuals cross classified by age and sex.

```
> frame.df
   sex   age serum num tot
1    M 30-49     A  13 340
2    M 30-49     B  18 408
3    M 30-49     C  40 421
4    M 30-49     D  57 362
5    M 50-62     A  13 123
6    M 50-62     B  33 176
7    M 50-62     C  35 174
8    M 50-62     D  49 183
9    F 30-49     A   6 542
10   F 30-49     B   5 552
11   F 30-49     C  10 412
12   F 30-49     D  18 357
13   F 50-62     A   9  58
14   F 50-62     B  12 135
15   F 50-62     C  21 218
16   F 50-62     D  48 395
```

The variables `num` and `tot` give the number having heart disease and the total number in the study for each of the 16 risk factor combinations.

(a) Describe the model or models you would initially try to fit to these data, with a view to identifying the main risk factors for CHD. Describe briefly how you would check if your model fits well, and how you might find a simpler model.

(b) The the plots in Figure 6 refer to fitting a particular model. Can you identify any risk factor combinations for which the model fits poorly? Can you offer any explanation for this?

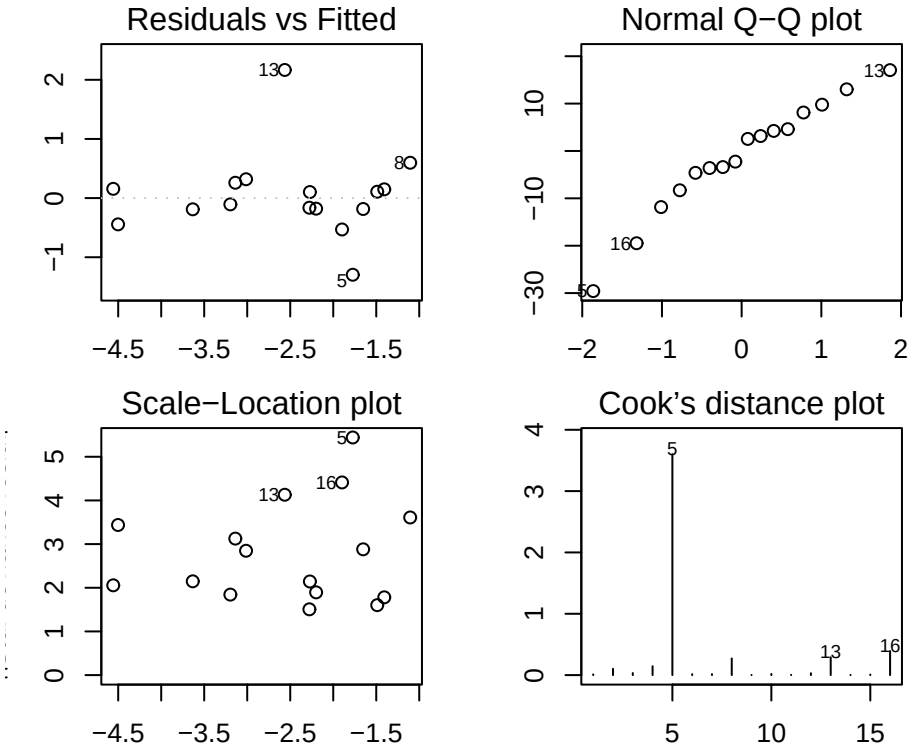

Figure 6: Diagnostic plots for Question B2(b).

(c) After some remedial action, The following output was obtained:

```
> frame.glm<-glm(cbind(num,tot-num)~  sex * age + age * serum,
                                  family=binomial, data=frame.df)
> summary(frame.glm)
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.55797    0.26742 -17.044  < 2e-16 ***
sexM             1.36270    0.18801   7.248 4.23e-13 ***
age50-62         2.23894    0.35980   6.223 4.89e-10 ***
serumB           0.05638    0.31516   0.179 0.858023
serumC           0.92448    0.27611   3.348 0.000813 ***
serumD           1.54452    0.26427   5.844 5.08e-09 ***
sexM:age50-62   -0.49769    0.24189  -2.057 0.039641 *
age50-62:serumB -0.06752    0.42424  -0.159 0.873549
age50-62:serumC -0.84739    0.38824  -2.183 0.029065 *
age50-62:serumD -1.15523    0.36796  -3.140 0.001692 **
---
Null deviance: 301.2147  on 15  degrees of freedom
Residual deviance:   1.5435  on  6  degrees of freedom
AIC: 96.225
> round(100*predict(frame.glm, type="response"),1)
 [1]  3.9  4.2  9.4 16.1 18.9 18.8 20.2 25.6  1.0  1.1
[11]  2.6  4.7  9.0  8.9  9.6 12.7
```

Discuss the effect of the various risk factors on the probability of heart disease. Which is the most important risk factor?

3. (a) In a 2×2 table with factors A and B, each at two levels, we can measure association between the factors using the odds ratio. Give a definition of the odds ratio and discuss its connection with the concept of independence.

   (b) In a $2 \times 2 \times 2$ table, with factors A, B and C, each at two levels, we can consider the odds ratios in the separate $A \times B$ tables corresponding to the two levels of C. Under what circumstances will these be the same? What is the name given to the situation when they are different? What implications does this have for an analysis?

   (c) The table below classifies admission applications to the Berkeley graduate school, classified according to (a) the success or failure of the application, (b) the gender of the applicant, and (c) the department.

		Gender	
Dept	Admit	Male	Female
A	Yes	353	17
	No	207	8
B	Yes	120	202
	No	205	391
C	Yes	138	131
	No	279	244
D	Yes	53	94
	No	138	299
E	Yes	22	24
	No	351	317

Below is a data frame obtained by collapsing the above table over departments. Ignoring departments, and using the output below, are success and gender independent? Give a reason for your answer.

```
> berk.df
  Count Gender Admit
1  1278      F    No
2  1493      M    No
3   557      F   Yes
4  1198      M   Yes
> berk.glm<-glm(Count~Gender*Admit, family=poisson, data=berk.df)
> anova(berk.glm, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: Count
Terms added sequentially (first to last)
             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                             3     486.35
Gender        1   162.87         2     323.48 2.665e-37
Admit         1   230.03         1      93.45 5.879e-52
Gender:Admit  1    93.45         0  3.393e-13 4.167e-22
```

(d) Some further output is shown below. Does this modify the conclusion you reached in part(c)?

NOTE: The data frame `berkeley.df` contains the data in the table shown on the previous page, and contains variables `Count`, `Admit`, `Gender` and `Dept`. The first few lines of this data frame are shown in the output.

```
> berkeley.df
   Count Admit Gender Dept
1    353   Yes      M    A
2    207    No      M    A
3     17   Yes      F    A
4      8    No      F    A
5    120   Yes      M    B
..............
20 lines in all

> berkeley.glm<-glm(Count~Dept*Gender*Admit, family=poisson,
                                    data=berkeley.df)
> anova(berkeley.glm, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: Count
Terms added sequentially (first to last)
                  Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                 19    1879.78
Dept               4   111.60        15    1768.18  3.320e-23
Gender             1     5.38        14    1762.80       0.02
Admit              1   469.90        13    1292.90 3.365e-104
Dept:Gender        4   753.44         9     539.46 9.333e-162
Dept:Admit         4   536.78         5       2.68 7.427e-115
Gender:Admit       1     0.13         4       2.56       0.72
Dept:Gender:Admit  4     2.56         0  1.936e-13       0.63
```

ATTACHMENT FOLLOWS

COPY OF COPLOT FOR QUESTION B1(c)