THE UNIVERSITY OF AUCKLAND

SECOND SEMESTER, 2005
Campus: City

STATISTICS

Advanced Statistical Modeling

(Time allowed: THREE hours)

INSTRUCTIONS

SECTION A: Multiple Choice (60 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 penalized.

SECTION B (40 marks)

- Answer **2 out of 3** questions. Each is worth 20 marks. Thus, there are 100 marks in total for both parts.

CONTINUED

SECTION A

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

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

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

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

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

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

2. The CO2 uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO2 concentration. Half the plants of each type were chilled overnight before the experiment was conducted.

   The data were assembled into a data frame with the following variables:

   Type: a factor with levels "Quebec" and "Mississippi" giving the origin of the plant;

   Treatment: a factor with levels "nonchilled" and "chilled";

   conc: a continuous variable representing ambient carbon dioxide concentrations (mL/L);

   uptake: a continuous variable representing carbon dioxide uptake rates (umol/m^2 sec).

   Which of the following R commands would produce the most informative graph?

   (zz) `xyplot(uptake~conc|Type*Treatment, data=CO2)`

   (1) `coplot(Type~Treatment|uptake*conc, data=CO2)`

   (1) `xyplot(Type~Treatment|uptake*conc, data=CO2)`

   (1) `bwplot(Type~Treatment|uptake*conc, data=CO2)`

   (1) `plot(uptake,conc)`

3. A Trellis display of the data in Question 2 is shown in Figure 1.

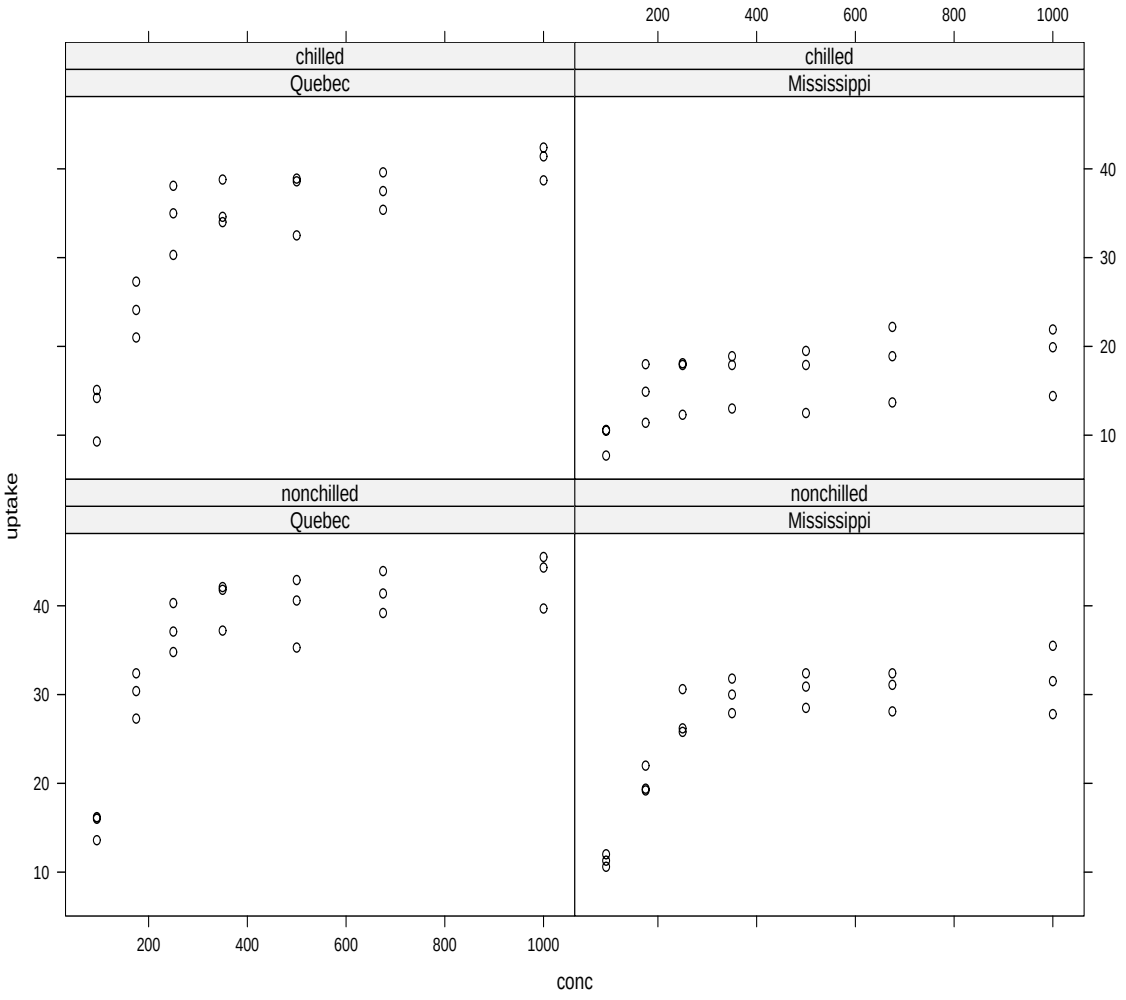

Figure 1: Trellis plot for Question 3.

Which of the following is **FALSE**?

(zz) For a fixed level of type, chilling tends to increase the uptake.

(1) There is an increasing relationship between `conc` and `uptake`.

(1) For large concentrations, the uptake doesn't change much.

(1) Conditional on treatment, uptake tends to be lower for Mississippi.

(1) As concentration changes, uptake changes more rapidly at low concentrations.

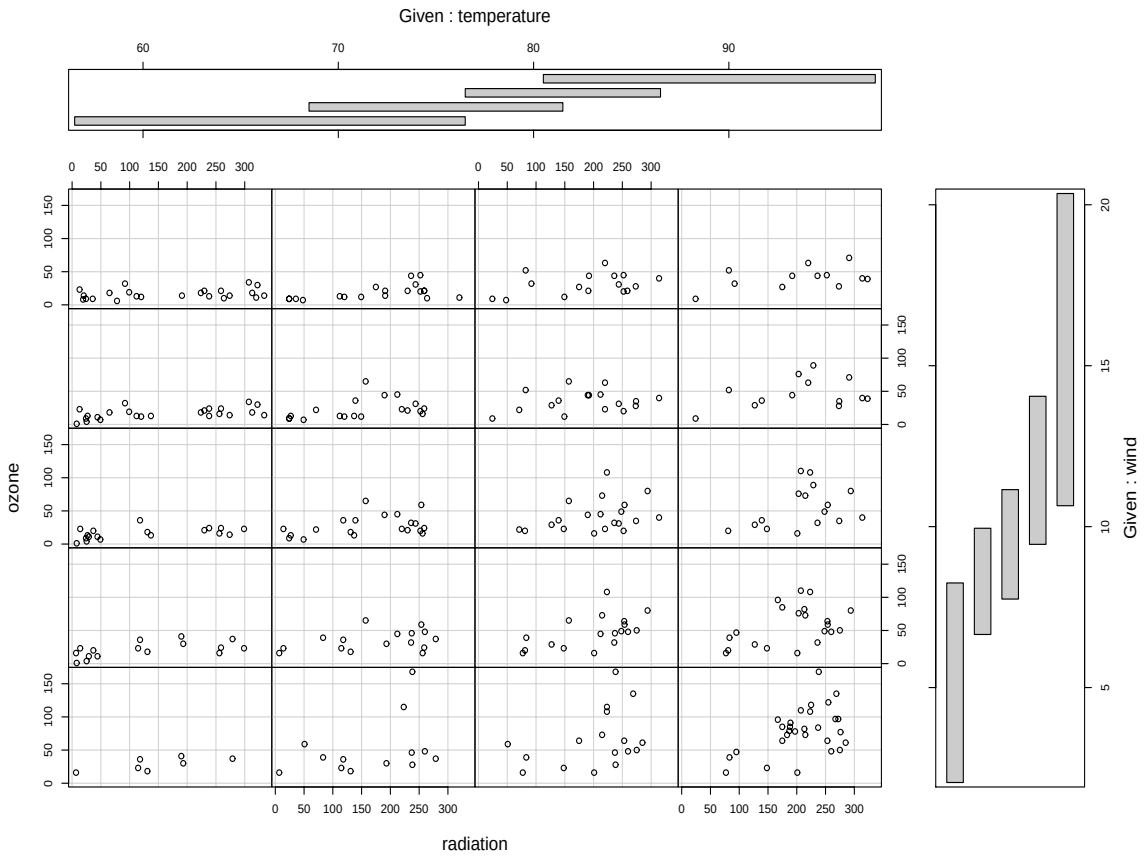

Figure 2: Trellis plot for Question 4.

4. This question relates to a study of air pollution in New York state. Four continuous atmospheric variables (ozone concentration, temperature, solar radiation and wind strength) were measured daily for 110 days. A coplot of the data is shown in Figure 2. Which of the following is **FALSE**?

   (zz) The relationship between ozone concentration and radiation doesn't depend on temperature.

   (1) At high temperatures, there is a increasing relationship between ozone concentration and radiation.

   (1) The highest ozone concentrations seem to be when temperature is high and wind is low.

   (1) When temperature is low and wind is high, there is not a strong relationship between ozone and radiation.

   (1) For a fixed amount of radiation, ozone is more variable when the wind is low and the temperature is high.

5. In a regression analysis, which of the following is **FALSE**?

   (zz) If the residual sum of squares is a small number, the fit must be good.

   (1) In linear regression, the "analysis of variance identity" expresses the "total sum of squares" as the sum of the "regression sum of squares" and the "residual sum of squares".

   (1) The residual sum of squares is zero if and only if the R^2 is 1.

   (1) If all of the estimated regression coefficients other than the constant term are zero, the regression sum of squares is zero.

   (1) If we use R^2 as a goodness-of-fit index, the bigger R^2 is, the better the fit.

6. Which of the following plots might be useful in diagnosing possible non-independence in a regression analysis where the data were collected sequentially in time?

   (zz) A plot of residuals versus previous residuals.

   (1) A plot of residuals versus fitted values.

   (1) A normal plot.

   (1) A gam plot.

   (1) A plot of Cook's distances.

7. In the course we discussed several types of influence diagnostics. Which of the following statements about these diagnostics is **FALSE**?

   (zz) The hat matrix diagonal measures the influence of a data point.

   (1) Cook's distances measure the change in the regression coefficients when a point is deleted.

   (1) The DFBETAS diagnostics measure the change in individual regression coefficients when a point is deleted.

   (1) The COVRATIO measures the change in the standard errors when a point is deleted.

   (1) The DFFITS measures the change in predicted values when a point is deleted.

8. The data for this question are taken from the Los Angeles heart Study. The following variables were measured on 60 men:

   wt: Weight in pounds.

   age: Age in years.

   sbp: Systolic blood pressure in mm of mercury.

   chl: Cholesterol level in mg per dl.

   ht: Height in inches.

   A regression model with **sbp** as response was fitted with the following results:

```
Call:
lm(formula = sbp ~ age + chl + ht + wt, data = heartstudy.df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.58519   66.09363   0.675  0.50277
age          0.53218    0.19459   2.735  0.00838 **
chl          0.03285    0.03508   0.936  0.35318
ht           0.13451    0.96248   0.140  0.88936
wt           0.20222    0.09360   2.160  0.03512 *

Residual standard error: 15.41 on 55 degrees of freedom
Multiple R-Squared: 0.2265,     Adjusted R-squared: 0.1702
F-statistic: 4.026 on 4 and 55 DF,  p-value: 0.006194
```

   which of the following is **FALSE**?

   (zz) The output strongly suggests that every additional unit of cholesterol increases blood pressure by about .03 mm, assuming the other variables don't change.

   (1) Every additional year of age increases mean blood pressure by about 0.53 mm, assuming the other variables don't change.

   (1) We can't expect to get accurate predictions from this model.

   (1) Adding height to the model doesn't improve the model.

   (1) At least some of the explanatory variables are related to the response.

9. Some diagnostic plots for the data in Question 8 are shown in Figure 3. Which of the following faults in the model is **not** indicated by this output?

   (zz) There is a strong suggestion that the variances are not equal.

   (1) There is a strong suggestion that the data are not planar.

   (1) Point 5 seems to be an outlier.

   (1) Age might benefit from a transformation.

   (1) There seems to be no reason to transform weight.

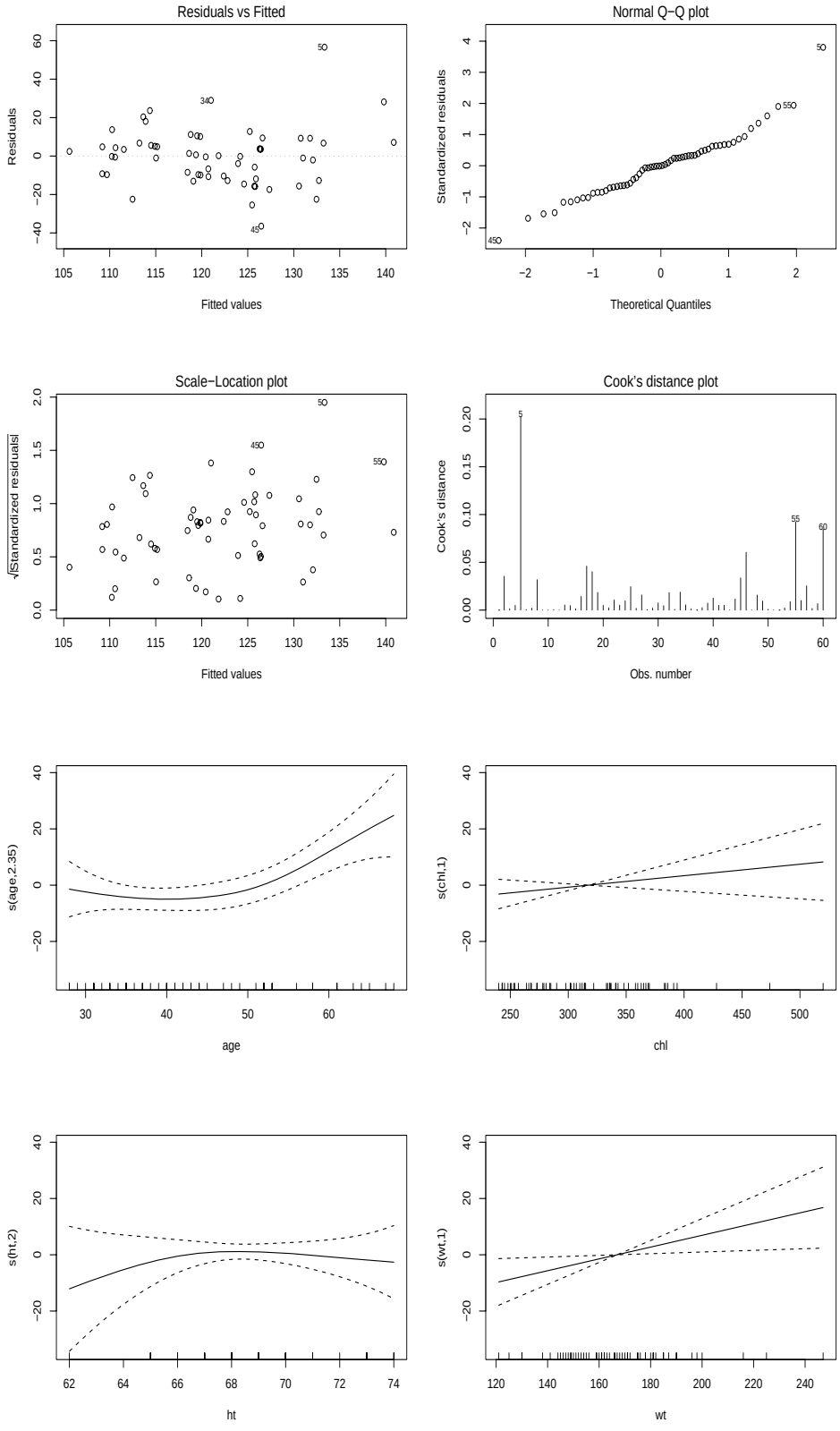

Figure 3: Diagnostic plots for Question 9.

10. In the course we discussed the concept of collinearity. Which of the following statements concerning a regression of a continuous response variable Y on 2 continuous explanatory variables X and Z is **FALSE**?

    (zz) The bigger the correlation between X and Z, the smaller the standard errors of the regression coefficients.

    (1) The variance inflation factors are the diagonal elements of the inverse of the correlation matrix of X and Z.

    (1) The bigger the error variance, the bigger the standard errors of the regression coefficients.

    (1) If X and Z are highly correlated, they are likely to have non-significant p-values in the regression summary.

    (1) The variance inflation factor for X is $1/(1-r^2)$ where r is the correlation between X and Z.

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

    - 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`).

    In the data set, there are 10 observations for each of the six factor level combinations, for a total of 60 observations. The response variable is `Ascorbic`, the ascorbic acid content of the cabbage head. The following table of means was obtained:

```
         Line
date    39   52
16    50.3 62.5
20    49.4 58.9
21    54.8 71.8
```

    Which of the following is **FALSE**?

    (zz) The main effect for Line=52 is -12.2.

    (1) The baseline mean is 50.3.

    (1) The row effect for date=20 is -0.9.

    (1) The interaction for date= 16, line=52 is 0.

    (1) The interaction for date= 21, line=52 is 4.8.

12. In addition the variables Ascorbic, Line and Date, the weight of the cabbage heads (measured by the variable HeadWt) was also measured. An analysis of variance including this covariate was performed on the data in Question 11, with the following results:

```
> cabbage.lm<-lm(Ascorbic~factor(Date)*factor(Line)*HeadWt,
                                                    data=cabbage.df)
> cabbage2.lm<-lm(Ascorbic~factor(Date)*factor(Line) + HeadWt,
                                                    data=cabbage.df)
> anova(cabbage2.lm, cabbage.lm)
Analysis of Variance Table

Model 1: Ascorbic ~ factor(Date) * factor(Line) + HeadWt
Model 2: Ascorbic ~ factor(Date) * factor(Line) * HeadWt
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     53 1975.05
2     48 1847.24  5    127.82 0.6643 0.6523

> anova(cabbage2.lm)
Analysis of Variance Table

Response: Ascorbic
                          Df  Sum Sq Mean Sq F value    Pr(>F)
factor(Date)               2  909.30  454.65 12.2004 4.381e-05 ***
factor(Line)               1 2496.15 2496.15 66.9835 5.687e-11 ***
HeadWt                     1  629.61  629.61 16.8955 0.0001379 ***
factor(Date):factor(Line)  2   30.73   15.37  0.4124 0.6641800
Residuals                 53 1975.05   37.27
```

Which of the following is **FALSE**?

(zz) No submodel of the model `Ascorbic~factor(Date) * factor(Line) + HeadWt` seems appropriate.

(1) There is no evidence that the factors Date and Line interact.

(1) An estimate of the error standard deviation is 6.10.

(1) The model `Ascorbic~factor(Date) * factor(Line) * HeadWt` does not seem appropriate.

(1) The model `Ascorbic~factor(Date) * factor(Line) + HeadWt` fits 6 parallel lines.

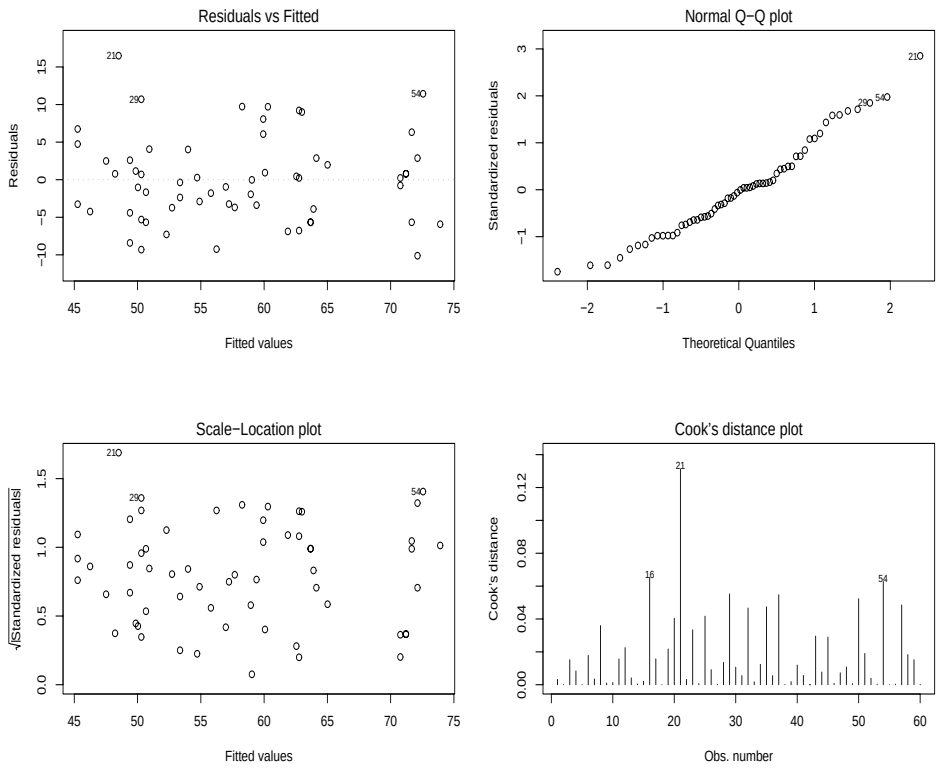

Figure 4: Diagnostic plots for Question 13.

13. The plots in Figure 4 are diagnostic plots from fitting the model `Ascorbic factor(Date) * factor(Line) + HeadWt` to the cabbage data. Which of the following is **TRUE**? The following output may be helpful:

```
qf(0.1,7, 53)
[1] 0.3968593
> qf(0.1,5,7)
[1] 0.2969210
> max(abs(rnorm(60)))
[1] 2.934225
```

(zz) The plots do not suggest any violation of the regression assumptions.

(1) The plots suggest that point 21 is an outlier.

(1) The plots suggest that point 21 is a high-leverage point.

(1) The plots suggest that the data are not planar.

(1) The plots suggest that the data are not independent.

14. In the standard logistic regression model, which of the following is a linear function of the covariates?

    (zz) The log-odds of a "success".

    (1) The odds of a "success".

    (1) The probability of a "success".

    (1) The mean of the response.

    (1) The log of the mean of the response.

15. In strongly grouped binomial data, which of the following is **TRUE**?

    (zz) The residual deviance provides a test of "goodness of fit".

    (1) The residual deviance has a normal distribution.

    (1) The Pearson residuals are useless for diagnostic purposes.

    (1) The only use of the diagnostic plots is to indicate influential points.

    (1) The Pearson residuals are the same as the deviance residuals.

16. In a casino, there are 3 roulette wheels: A, B and C. When you place a bet on black, a fair wheel should come up black with a probability of 9/19. You are suspicious that the wheels may not be fair, and make 100 bets on each wheel, betting each time on black. You assemble your data into a data frame:

```
> casino.df
   r   n wheel
1 44 100     A
2 54 100     B
3 56 100     C
```

    The variable `n` is the number of spins of the wheel, and `r` is the number of spins resulting in black. You analyse the data in R, getting the following output:

```
Call:
glm(formula = cbind(r, n - r) ~ wheel, family = binomial, data = casino.df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2412     0.2015  -1.197   0.2313
wheelB        0.4015     0.2843   1.412   0.1579
wheelC        0.4823     0.2849   1.693   0.0905 .

Null deviance: 3.3143e+00  on 2  degrees of freedom
Residual deviance: 1.7764e-15  on 0  degrees of freedom
> 1-pchisq(3.3143,2)
[1] 0.1906816
```

Which of the following is **FALSE**?

(zz) From this output, we can conclude that the wheels are all fair.

(1) The model fitted is a saturated (maximal) model.

(1) The null deviance is testing the hypothesis that the wheels have the same probability of coming up black.

(1) There is no evidence that the wheels have different probabilities of coming up black.

(1) The estimated probability of coming up black for wheel A is 0.44.

17. Which of the following is **TRUE**?

(zz) A 95% confidence interval for the log-odds of wheel A coming up black is -0.2412 +/- 1.96*0.2015.

(1) A 95% confidence interval for the log-odds of wheel B coming up black is 0.4015 +/- 1.96*0.2843.

(1) A 95% confidence interval for the probability of wheel B coming up black is 0.4015 +/- 1.96*0.2843.

(1) The p-value of 0.0905 relates to the hypothesis that wheel C is fair.

(1) The p-value of 0.1579 relates to the hypothesis that wheels B and C have the same probability of coming up black.

18. The data in Table 1 come from a study which investigated the effect of insulin on laboratory mice. Mice were injected with different doses of insulin. For each level of dose, the response observed was the number of mice having convulsions, and the number receiving the dose. The investigators were interested in modelling how the proportion of mice suffering convulsions changed with the dose. Suppose that

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

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

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

Table 1. Data for Question 18

Dose (mg), x_i	Number convulsing, 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

CONTINUED

Which of the following is **TRUE**?

(zz) To test the goodness of fit of the logistic model, we calculate the p-value corresponding to 13.841, using the Chi-square distribution with 7 degrees of freedom.

(1) To test that the probability of convulsion does not depend on the dose, we calculate the p-value corresponding to 217.8824, using the Chi-square distribution with 7 degrees of freedom.

(1) To test that the probability of convulsion does not depend on the dose, we calculate the p-value corresponding to 217.8824, using the Chi-square distribution with 8 degrees of freedom.

(1) The deviance of the saturated model is 2×159.5074.

(1) The deviance of the null model is 2×217.8824.

19. In class, we studied some data on lizards. There were two types of lizard, (Grahami and Opalinus), and we modelled the probability that a lizard observed at a particular site would be a Grahami. This probability depended on three factors, time of day (early, mid or late), the perch height (high or low) and the perch dimension (long or short). For each of these 12 factor combinations, we counted the number of lizards observed, and the number that were Grahami. A logistic regression model without interactions was found to be suitable, and gave the following results:

```
 Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.49466    0.28809   5.188 2.12e-07 ***
timelate    -1.05278    0.28026  -3.756 0.000172 ***
timemid      0.04003    0.23971   0.167 0.867384
lengthshort  0.67630    0.20588   3.285 0.001020 **
heightlow   -0.83011    0.23204  -3.578 0.000347 ***
---
Null deviance: 54.0430  on 11  degrees of freedom
Residual deviance:  9.8815  on  7  degrees of freedom
```

Which of the following interpretations is **FALSE**?

(zz) The probability that a lizard observed early in the day on a high, long perch is a Grahami is 0.4019.

(1) The log-odds of a lizard on a low perch being a Grahami is 0.83011 less than the log-odds of a lizard on a high perch being a Grahami.

(1) Being on a short perch increases the odds of being a Grahami by almost a factor of 2.

(1) Lizards observed early or in the middle of the day are about equally likely to be Grahami.

(1) Lizards observed late in the day less likely to be Grahami.

20. In a Poisson regression, a variable X has a regression coefficient of 0.5. Which is the correct interpretation?

(zz) If the other explanatory variables are held fixed, a unit increase in X is associated with a 65% increase the mean response.

(1) If the other explanatory variables are held fixed, a unit increase in X is associated with an increase of 0.5 in the mean response.

(1) Averaged over the other variables, a unit increase in X is associated with an increase of 0.5 in the mean response.

(1) Averaged over the other variables, a unit increase in X is associated with an increase of 0.5 in the log-odds ratio.

(1) If the other explanatory variables are held fixed, a unit decrease in X is associated with a 65% decrease the mean response.

21. The data in Table 2 arose when classifying 577,006 individuals involved in motor accidents in Florida. The accidents are classified by (a) whether a seat belt was used or not, and (b) whether the accident was fatal or not.

Tabke 2. Data for Question 21.

	Type of accident	
Seat belt	Fatal	Non-fatal
Not Worn	1601	162527
Worn	510	412368

Some R output is shown below:

```
> injury.df<-data.frame(expand.grid(belt=c("Not worn","Worn"),
  type=c("Fatal", "Non-fatal")), count=c(1601,510,162527,412368))
> injury.glm<-glm(count~belt*type, family=poisson, data=injury.df)
> y<-injury.df[,4]
> sum(y*log(y/sum(y)))
[1] -357463.9

> summary(injury.glm)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            7.37838    0.02499  295.23   <2e-16 ***
beltWorn              -1.14397    0.05085  -22.50   <2e-16 ***
typeNon-fatal          4.62022    0.02511  183.96   <2e-16 ***
beltWorn:typeNon-fatal 2.07505    0.05093   40.74   <2e-16 ***
```

Which of the following is **FALSE**?

(zz) The null deviance for this model is 0.

(1) The log-odds for this table is 2.07505.

(1) A 95% confidence interval for the odds ratio for this table is (7.208, 8.801).

(1) There is strong evidence of a relationship between wearing a seat belt and the accident being fatal.

(1) The maximum value of the log-likelihood $\sum y_i \log(\pi_i)$ for the maximal model is -357463.9.

22. Suppose in addition to the two factors in Question 21, we measured a third factor "Ejected from car". A slightly different set of data is shown in Table 3.

Table 3. Data for Question 22.

		Type of accident	
Seat belt	Ejected	Fatal	Non-fatal
Not Worn	Yes	497	4,624
Not Worn	No	1,008	157,342
Worn	Yes	14	1,105
Worn	No	483	411,111

An analysis of these data was performed, resulting in the following (edited) R output:

```
> injury2.df<-data.frame(expand.grid(ejected=c("Yes","No"),
+ belt=c("Not Worn","Worn"), type=c("Fatal", "Non-Fatal")),
+ count=c(497,1008, 14,483,4624,157342,1105,411111))
> injury2.glm<-glm(count~belt*ejected*type, family=poisson, data=injury2.df)
> summary(injury2.glm)
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      6.20859    0.04486 138.411  < 2e-16 ***
beltWorn                        -3.56953    0.27100 -13.172  < 2e-16 ***
ejectedNo                        0.70713    0.05481  12.902  < 2e-16 ***
typeNon-Fatal                    2.23043    0.04721  47.250  < 2e-16 ***
beltWorn:ejectedNo               2.83383    0.27659  10.246  < 2e-16 ***
beltWorn:typeNon-Fatal           2.13812    0.27306   7.830 4.87e-15 ***
ejectedNo:typeNon-Fatal          2.82003    0.05680  49.644  < 2e-16 ***
beltWorn:ejectedNo:typeNon-Fatal -0.44197    0.27863  -1.586    0.113
```

```
> anova(injury2.glm, test="Chisq")
                  Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                                 7    1624865
belt               1   111458         6    1513407          0
ejected            1   729871         5     783536          0
type               1   772092         4      11444          0
belt:ejected       1     7877         3       3568          0
belt:type          1     1887         2       1680          0
ejected:type       1     1678         1          3          0
belt:ejected:type  1        3         0 -2.492e-11  9.115e-02
```

Which model is indicated by this output?

(zz) The homogeneous association model.

(1) A model where all factors are independent.

(1) A model where type of accident is independent of being ejected, given seat belt use.

(1) A model where type of accident is independent of seat belt use, given "Ejected from car".

(1) None of the models mentioned are indicated.

23. A suitable model was fitted to the data. A fragment of computer output relating to this model is:

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
beltWorn                 -3.15645    0.06239  -50.59   <2e-16 ***
ejectedNo                 0.72784    0.05345   13.62   <2e-16 ***
typeNon-Fatal             2.24583    0.04650   48.30   <2e-16 ***
beltWorn:ejectedNo        2.39964    0.03334   71.97   <2e-16 ***
beltWorn:typeNon-Fatal    1.71732    0.05402   31.79   <2e-16 ***
ejectedNo:typeNon-Fatal   2.79779    0.05526   50.63   <2e-16 ***
```

Which of the following is **TRUE**?

(zz) A 95% confidence interval for the conditional odds ratio between "type of accident" and "seat belt use" given "Ejected from car" is (5.010, 6.192).

(1) A 95% confidence interval for the conditional log odds ratio between "type of accident" and "seat belt use" given "Ejected from car" is (2.689, 2.906).

(1) A 95% confidence interval for the conditional odds ratio between "type of accident" and "seat belt use" given "Ejected from car" is (14.724, 18.285).

(1) A 95% confidence interval for the conditional log odds ratio between "type of accident" and "seat belt use" given "Ejected from car" is (0.623, 0.832).

(1) All the other four answers are wrong.

24. Suppose we have a contingency table of data classified according to three factors A, B and C. In Figure 5, graphs are shown representing two models for this table (both have no 3-factor interaction).

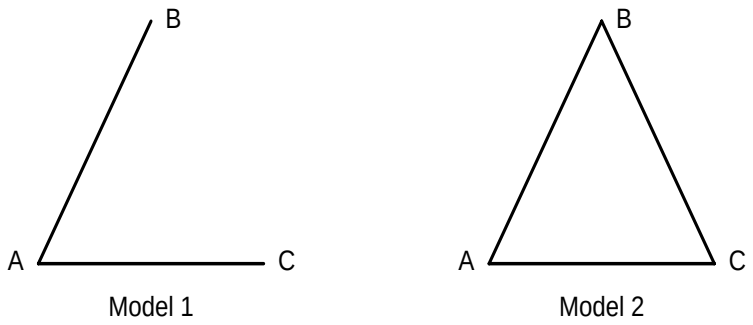

Figure 5: Models for for Question 24.

Which of the following is true?

(zz) If model 1 is true, the table can be collapsed over factor C.

(1) If model 1 is true, the table can be collapsed over factor A.

(1) If model 2 is true, the table can be collapsed over factor A.

(1) If model 2 is true, the table can be collapsed over factor B.

(1) If model 2 is true, the table can be collapsed over factor C.

25. The data in Table 4 refer to 6115 families in Saxony, all of which had 12 children. The table gives the number of families having 0,1,...,12 boys.

Table 4. Data for Question 25.

Number of boys	Number of families
0	3
1	24
2	104
3	286
4	670
5	1033
6	1343
7	1112
8	829
9	478
10	181
11	45
12	7

The investigators wanted to see if these data followed a binomial distribution. An analysis using R produced the following output:

```
> count<-c(3,24,104,286,670,1033,1343,1112,829,478,181,45,7)
> x<-0:12
> phat<-sum(x*count)/(12*sum(count))
> phat
[1] 0.519215
> logL1<-sum(count*log(count/sum(count)))
> logL1
[1] -12485.67
> bin.probs<-dbinom(x,12,phat)
> logL2<-sum(count*log(bin.probs))
> logL2
[1] -12534.17
> logL3<-sum(count*log(1/13))
> logL3
[1] -15684.67
> 1-pchisq(6398,11)
[1] 0
> 1-pchisq(97,11)
[1] 0
> 1-pchisq(107,11)
[1] 0
```

Note that the R function `dbinom` evaluates the binomial probability function, given by $Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$. Which if the following is **FALSE**?

(zz) The binomial model is a good fit to these data.

(1) The null deviance (corresponding to a model that says getting 0,1, ..., 12 boys are all equally likely) is about 6398.

(1) The residual deviance for the binomial model is about 97.

(1) The binomial model seems more plausible than the null model.

(1) Almost 52% of the children are boys.

SECTION B

1. (a) In class we discused three methods of selecting subsets of variables in linear regression: using anova tables, stepwise regression, and using all possible regressions. Briefly describe the differences between them, and any disadvantages that each might have. [6 marks]

   (b) Briefly describe and compare the different criteria used to select models in the all possible regressions approach.[7 marks]

   (c) In a tutorial, we looked at a set of data consisting of various measurements on 138 cars that were taken from Road and Track's "The Complete '99 Car Buyer's Guide". The variables in this data set are:

   CITY: mileage (miles per gallon) in city driving, (response);

   PRICE: price in dollars (US);

   WEIGHT: weight in pounds;

   DISP: displacement in cubic centimetres;

   COMP: compression ratio as value to 1;

   HP: horsepower at 6300 rpm;

   TORQ: torque at 5200 rpm;

   TRANS: transmission (1 = automatic, 0 = manual);

   CYL: number of cylinders.

   We want to develop a model which explains the variable CITY in terms of the other variables. Previous experience suggests that a model where the reciprocal of CITY (i.e the variable 1/CITY) is expressed as a linear function of the other variables will be satisfactory. Also assume that any large outliers have been removed from the data set. Use the output below to suggest a suitable subset of variables to include in the model. Your answer should include a justification for your chosen model. Also include in your answer a a comment on the effects of overfitting or underfitting the model. [7 marks]

```
> cars.lm<-lm(I(1/CITY)~ PRICE + WEIGHT + DISP + COMP + HP +
                    TORQ + TRANS + CYL, data = cars.df)
    rssp sigma2 adjRsq   Cp    AIC    BIC     CV PRICE WEIGHT DISP COMP HP TORQ TRANS CYL
1 2920.63 21.63  0.86 90.37 227.37 233.21 298.50     0      0    0    0  1    0     0   0
2 2070.52 15.45  0.90 27.36 164.36 173.12 219.42     1      0    0    0  0    1     0   0
3 1865.19 14.02  0.91 13.65 150.65 162.33 193.68     1      0    1    0  1    0     0   0
4 1732.08 13.12  0.91  5.47 142.47 157.07 193.61     1      1    1    0  1    0     0   0
5 1715.60 13.09  0.91  6.21 143.21 160.73 196.97     1      1    1    1  1    0     0   0
6 1688.66 12.99  0.91  6.15 143.15 163.59 196.73     1      1    1    1  1    1     0   0
7 1673.86 12.97  0.92  7.02 144.02 167.38 197.28     1      1    1    1  1    1     1   0
8 1673.58 13.07  0.91  9.00 146.00 172.28 199.27     1      1    1    1  1    1     1   1
```

2. (a) Carefully define the terms **null deviance** and **residual deviance** as applied to a logistic regression model for grouped data. In the same context, what is meant by a **saturated model**? Why must a saturated model have zero deviance?

[6 marks]

(b) The data below show the free-throw results obtained by the Los Angeles Lakers player Shaq O'Neal in 23 NBA playoff games in the year 2000. (For those unfamiliar with basketball, a free throw is when a player is allowed to take an unopposed shot at the basket from the free-throw line. Thus in game 1, O'Neal attempted 5 free throws, 4 of which were sucessful.)

```
> freethrows.df
     s  n game
1    4  5    1
2    5 11    2
3    5 14    3
4    5 12    4
5    2  7    5
6    7 10    6
7    6 14    7
8    9 15    8
9    4 12    9
10   1  4   10
11 13 27   11
12   5 17   12
13   6 12   13
14   8  9   14
15   7 12   15
16   3 10   16
17   8 12   17
18   1  6   18
19 18 39   19
20   3 13   20
21 10 17   21
22   1  6   22
23   3 12   23
```

Below is some R output from an analysis of these results.

```
> freethrows.glm<-glm(cbind(s,n-s)~game, family=binomial, data=freethrows.df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3863     1.1180   1.240   0.2150
game2        -1.5686     1.2715  -1.234   0.2173
game3        -1.9741     1.2494  -1.580   0.1141
game4        -1.7228     1.2621  -1.365   0.1722
game5        -2.3026     1.3964  -1.649   0.0992 .
game6        -0.5390     1.3138  -0.410   0.6816
game7        -1.6740     1.2416  -1.348   0.1776
game8        -0.9808     1.2360  -0.794   0.4275
game9        -2.0794     1.2748  -1.631   0.1028
game10       -2.4849     1.6073  -1.546   0.1221
game11       -1.4604     1.1825  -1.235   0.2168
game12       -2.2618     1.2383  -1.827   0.0678 .
game13       -1.3863     1.2583  -1.102   0.2706
game14        0.6931     1.5411   0.450   0.6529
game15       -1.0498     1.2621  -0.832   0.4055
game16       -2.2336     1.3138  -1.700   0.0891 .
game17       -0.6931     1.2748  -0.544   0.5866
game18       -2.9957     1.5652  -1.914   0.0556 .
game19       -1.5404     1.1633  -1.324   0.1854
game20       -2.5903     1.2974  -1.996   0.0459 *
game21       -1.0296     1.2218  -0.843   0.3994
game22       -2.9957     1.5652  -1.914   0.0556 .
game23       -2.4849     1.3017  -1.909   0.0563 .
---
```

```
Null deviance: 3.3376e+01  on 22  degrees of freedom
Residual deviance: 8.8818e-16  on  0  degrees of freedom
AIC: 109.17

> 1-pchisq(33.376, 22)
[1] 0.056777
```

Press reports of these games criticised O'Neal for his inconsistent performance. Is this justified? Give a reason for your answer. What statistical model are you using to arrive at your conclusion? [8 marks]

(c) What is meant by "over-dispersion" and "under-dispersion" in this context? Do you think either could apply here? What effect would it have on the analysis? [6 marks]

3. (a) Suppose we have a three-dimensional contingency table with factors A, B and C. Describe the connection between Poisson regression models and multinomial models for such tables, giving a full discussion how various types of independence are handled. [6 marks]

   (b) Describe how log-linear models can be represented by association graphs. Give an example of two models that have the same graph. [4 marks]

   (c) The data below arose from a consumer survey to test reaction to a new brand of laundry detergent, Brand X. Respondents were asked if they prefer brand X or a current brand M, if they use high or low temperature, and if they are previous users of M. The following data were obtained.

```
> detergent.df
  temp prev pref count
1 High  Yes    X    66
2  Low  Yes    X   141
3 High   No    X   104
4  Low   No    X   197
5 High  Yes    M   119
6  Low  Yes    M   156
7 High   No    M    80
8  Low   No    M   145

> anova(glm(count~temp*prev*pref, family=poisson, data=detergent), test="Chisq")

Analysis of Deviance Table
              Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                              7    103.619
temp           1   73.212         6     30.407  1.164e-17
prev           1    1.921         5     28.486      0.166
pref           1    0.063         4     28.423      0.801
temp:prev      1    1.253         3     27.170      0.263
temp:pref      1    4.362         2     22.808      0.037
prev:pref      1   20.020         1      2.788  7.664e-06
temp:prev:pref 1    2.788         0 -9.326e-15      0.095
> anova(glm(count~prev*pref, family=poisson, data=detergent), test="Chisq")

Analysis of Deviance Table
         Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         7    103.619
prev      1    1.921         6    101.698     0.166
pref      1    0.063         5    101.635     0.801
prev:pref 1   20.581         4     81.053 5.715e-06
```

   i. What log-linear model is indicated by this output? Draw the association graph corresponding to this model. [3 marks]

   ii. Discuss the relationship between the factors. Is there an association between brand preference (for X or M) and previous use of M? [4 marks]

   iii. Suppose we wanted to model the probability of preferring Brand X in terms of the other variables. What model could we use? [3 marks]